Computational Dissection of Composite Molecular Signatures and Transcriptional Modules

Ting Gong

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
The Bradley Department of Electrical and Computer Engineering

Jason J. Xuan, Chair

Chang-Tien Lu

Christopher L. Wyatt

Scott F. Midkiff

Yue J. Wang

December 14th, 2009
Arlington, Virginia

Keywords: Microarray, Latent Variable Modeling, Blind Source Separation, Tissue Heterogeneity Correction, Transcriptional Module, Gene Regulation

Computational Dissection of Composite Molecular Signatures and Transcriptional Modules

Ting Gong

ABSTRACT

This dissertation aims to develop a latent variable modeling framework with which to analyze gene expression profiling data for computational dissection of molecular signatures and transcriptional modules. The biological problem is formalizes as a source separation problem that permits the development of a unified latent variable modeling framework. The framework is implemented in different molecular levels with a top-down approach, *i.e.*, from tissue or cell mixtures to gene modules, and from gene modules to gene regulatory programs.

The first part of the dissertation is focused on extracting pure gene expression signals from tissue or cell mixtures. In molecular profiling of solid tumors, it is often needed to process composite gene expression data as obtained from heterogeneous tumor biopsies. The main goal of gene expression profiling is to identify the pure signatures of different cell types (such as cancer cells, stromal cells and inflammatory cells) and estimate the concentration of each cell type. In order to accomplish this, a new blind source separation method is developed, namely, nonnegative partially independent component analysis (nPICA), for tissue heterogeneity correction (THC). The nPICA method is specifically developed to extract nonnegative and partially uncorrelated signals from microarray data based on phenotypic up-regulated genes. The THC problem is formulated as a constrained optimization problem and solved with a learning algorithm based on geometrical and statistical principles. The nPICA method was tested on a series of numerical mixtures of microarray data sets to tackle cell and tissue heterogeneity, and accurately dissected mixtures into pure signals of component cell types.

The second part of the dissertation sought to identify gene modules from gene expression data to uncover important biological processes in different types of cells. Since biological processes are latent variables in gene expression data, the problem of gene module identification can be treated as a blind source separation problem as well. A new gene clustering approach, nonnegative independent component analysis (nICA), is

developed for gene module identification. The nICA approach is completed with an information-theoretic procedure for input sample selection and a novel stability analysis approach for proper dimension estimation. The nICA approach was applied to two different *Saccharomyces cerevisiae* microarray data sets and a muscle regeneration microarray data set for gene module identification. Experimental results showed that the nICA approach achieved an improved resolution for gene module identification when compared to conventional gene clustering methods. The gene modules identified by the nICA approach appear to be significantly enriched in functional annotations in terms of gene ontology (GO) categories.

The third part of the dissertation moves from gene module level down to DNA sequence level to identify gene regulatory programs by integrating gene expression data and protein-DNA binding data. A sparse hidden component model is first developed for this problem, taking into account a well-known biological principle, *i.e.*, a gene is most likely regulated by a few regulators. This is followed by the development of a novel computational approach, motif-guided sparse decomposition (mSD), in order to integrate the binding information and gene expression data. The mSD approach allows the identification of those potential binding sites actually used by the regulators of a gene module, but also uncovers the strengths of their control and their dynamic activities in a series of experiments. The mSD approach is first applied to simulation studies to demonstrate the feasibility of this approach, and subsequently applied to breast cancer cell lines for the identification of estrogen receptor (ER) signaling networks, thus leading to an understanding of the molecular mechanisms of antiestrogen resistance in breast cancer.

These computational approaches are primarily developed for analyzing high-throughput gene expression profiling data. Nevertheless, the proposed methods should be able to be extended to analyze other types of high-throughput data for biomedical research.

Acknowledgments

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

The first person is my advisor, Dr. Jason J. Xuan, for all his help, guidance, support, and encouragement. He brought me into the field of bioinformatics and computational biology, and provided significant guidance in many aspects of the research presented in this dissertation. All the research described in this dissertation is the result of our intensive discussions during the past five years. I also respect his insistence on precision in research work as well as in scientific writing. I also want to acknowledge his graciousness and flexibility as my research and dissertation advisor in giving me much freedom to pursue my academic goals.

I would like to especially thank Dr. Yue J. Wang as my steering committee member. Without his invaluable help and advice on my research, I could not have achieved so much progress over the last five years. I have learned much from him, and he is a role model for me in many ways. I appreciate his great insight about problems and remarkable knowledge in machine learning and mathematics.

I express my sincere gratitude to the other committee members, Dr. Chang-Tien Lu, Dr. Christopher L. Wyatt and Dr. Scott F. Midkiff. They have provided valuable discussions with me about many technical aspects in this dissertation, suggested improvements in my presentation, and shared their insightful feedback.

I also want to thank Dr. Huai Li for his generous support, guidance and resources designed to improve the quality of my research and publications during my graduate study. I am also very grateful to Dr. Robert Clark, Dr. Rebecca B. Riggins and Dr. Eric P. Hoffman as encouraging mentors to facilitate this inter-disciplinary research. As biologists, they understand the importance of computational modeling in this field and helped me narrow down biological problems important and relevant for computational modeling. They also allowed the use of various pioneering high-throughput data generated from their laboratories before they were published.

I shall thank all of my colleagues and lab mates with whom I have shared so much great moments during these years. I have benefitted greatly from discussions with them in the group meetings of discussing machine learning and bioinformatics literature. Many of them also sat in my preparation talks and offered useful suggestions about the presentation of my work. They make this period of time a unique experience in my life.

Finally, I would like to dedicate this dissertation to my parents, Mr. Bingchu Gong and Mrs. Yemin Jiao. Their dedication to my education shapes my life and values, and their selfless support allows me to accomplish goals in life. I could not have finished this dissertation without their support.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AP-1	Activator Protein
APC	Affinity Propagation Clustering
AUC	Area Under the ROC Curve
BiNGO	Biological Network Gene Ontology tool
BSS	Blind Source Separation
CAM	Convex Analysis of Mixtures
cDNA	complementary DeoxyriboNucleic Acid
CNS	Central Nerve System
COGRIM	Clustering of Genes into Regulons using Integrated Modeling
DEG	Differentially-Expressed Gene
DNA	DeoxyriboNucleic Acid
EM	Expectation Maximization
ER-	Estrogen Receptor negative
ER+	Estrogen Receptor positive
FP	False Positive
GF	Growth Factor
GO	Gene Ontology
ICA	Independent Component Analysis
IEG	Invariantly-Expressed Gene
i.i.d.	independently identically distributed
IPA	Ingenuity Pathway Analysis

ISG	Independence Support Gene
LCM	Laser Capture Microdissection
LDA	Linear Discriminant Analysis
LPP	Locality Preserving Projection
LPD	Latent Process Decomposition
MDL	Minimum Description Length
MM	Mis-Match
MPSS	Massively Parallel Signature Sequencing
mRNA	messenger RiboNucleic Acid
mSD	motif-guided Sparse Decomposition
NCA	Network Component Analysis
NFκB	Nuclear Factor κB
nICA	nonnegative Independent Component Analysis
nPICA	nonnegative Partially Independent Component Analysis
NMF	Nonnegative Matrix Factorization
PCA	Principal Component Analysis
PDF	Probability Density Function
PM	Perfect Match
PWM	Position Weight Matrix
ROC	Receiver Operating Characteristics
SAGE	Serial Analysis of Gene Expression
SCA	Sparse Component Analysis
SFNM	Standard Finite Normal Mixture

SNICA	Stochastic Non-negative Independent Component Analysis
SNR	Signal to Noise Ratio
SOM	Self-Organizing Maps
SP-1	Specificity Protein-1
SVD	Singular Vector Decomposition
SVM	Support Vector Machine
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
THC	Tissue Heterogeneity Correction
TSS	Transcription Start Site
VISDA	VIsual Statistical Data Analyzer
WGP	Well-Grounded Point

1 Introduction

The year 2003 marked the fiftieth anniversary year of the discovery of the double-helical structure of DNA, as well as the completion of a high-quality, comprehensive sequence of the human genome. This landmark event, which marked the beginning of the genomic era, has seen a steady stream of ever-larger and more complex genomic data sets from proteome to secretome generated by numerous new research strategies and experimental technologies [1]. To date, various public databases have provided astounding amount of genomic data that have literally transformed the study of virtually all life processes [2]. The genomic approach of technology development and large-scale generation of community resource data sets has introduced an important new dimension into biological and biomedical research [3].

Simultaneous quantification of the abundance levels of biomolecules on the genomic scale and follow-up data analyses provide great potential for the diagnosis and management of disease through this unprecedented comprehensive view of the molecular underpinnings of pathology [4]. The plethoric levels of biomolecules are tightly regulated to ensure the proper functions of the biological system. Abnormal variations at each level can correlate with many diseases. Some of the more successful applications, to name just a few, include cancer classification and prediction using gene expression arrays [5, 6], discovery of differentially expressed and functionally related genes, and inferring gene regulation networks using expression arrays [7-9] and/or ChIP-on-chip technology [10, 11].

With the implementation of new technology, large-scale methods for data generation are driving growth rapidly within both basic and biomedical research communities. As the amount and complexity of the data increases, and as the questions being addressed become more sophisticated, computational methods have become inseparable from modern biological research, and their importance can only increase [3]. New computational capabilities and methodologies will facilitate the analysis of experimental

data and stimulate the development of experimental approaches to test hypotheses. The resulting experimental data will, in turn, be used to refine models that will improve overall understanding and increase the opportunity for their application to disease. The areas of computational biology critical to the current and future of genomics research include [3]:

- New principled approaches to solving problems regarding the identification of different features in a DNA sequence, the analysis of gene expression and regulation, the elucidation of protein structure and protein–protein interactions, the determination of the relationship between genotype and phenotype, and the identification of genetic variation patterns in populations and biological processes that produced those patterns;
- Methods to elucidate the genetic and environmental influences and behaviors of gene–environment interactions in the context of health and disease;
- Improved database technologies to facilitate the integration and visualization of different data or information types, for example, information about pathways, protein structure, gene variation, chemical inhibition and clinical information;
- Improved knowledge management systems and the standardization of data sets to allow the coalescence of knowledge across disciplines.

We use gene expression array data throughout this dissertation to illustrate our data analysis schemes related to the aforementioned first area. Meanwhile, the proposed techniques can also be applied to data acquired through other platforms.

1.1 Microarray Transcription Profiling Technology

The goal of this dissertation is to develop computational methods for dissecting composite molecular signatures and inferring gene regulatory modules from large-scale genomic data sets. In this section we will give a general albeit basic overview about the biological knowledge and technology for data generation. This overview is by no means complete, but serves the purpose of providing some background for the discussions in subsequent chapters.

The protein information is encoded in DNAs (DeoxyriboNucleic Acid). DNA consists of a long, double-helix shaped polymer of nucleotides with base pairs of purines and pyrimidines discovered by Watson and Crick [12]. The sequence of nucleotides determines individual hereditary characteristics. Each position along a DNA is filled with one of the four bases: adenine (A), thymine (T), cytocine (C) and guanine (G). The bases along the two strands of a DNA are joined by hydrogen bonds between the complementary bases adenine and thymine or cytosine and guanine. The protein synthesis procedure of most contemporary organisms follows the central dogma [13]: DNA source → RNA template → protein product. The sequence information in a DNA (A, T, C, G) is first transcribed into another type of polymer called messenger Ribonucleic Acid (mRNA). By means of the family of transfer RNA molecules, an mRNA molecule is enabled to be translated into the sequence of amino acids in the protein.

The analysis of gene expression by microarray technology is based on the pairing of complementary nucleic acid molecules. In this technique, a collection of microscopic DNA spots, called probes, commonly representing single genes or transcripts are immobilized on a solid surface in an order and used to detect the concentration of the corresponding complementary RNA sequences, called targets, present in a sample of interest [14]. The advancements made in attaching or synthesizing nucleic acid sequences to solid supports and robotics have allowed tens of thousands of transcript species detected and quantified simultaneously.

There are two types of microarray gene expression profiling techniques, *i.e.* sequencing-based and hybridization-based [15]. For the sequencing-based profiling, the strategy is to analyze the level of gene expression in a sample by counting the number of individual mRNA molecules produced by each gene. The representative example are serial analysis of gene expression (SAGE) [16] and massively parallel signature sequencing (MPSS) [17]. There are three different types of hybridization-based microarray in common use: spotted cDNAs (complementary DNAs), spotted oligonucleotides and Affymetrix arrays [18]. These three technology platforms quantify the sample RNA based on fluorescence signal intensity. The first two techniques are quite similar except that the former exploits sets of plasmids of specific cDNAs as a probe in

gridded liquid aliquots, and the latter exploits synthetic oligonucleotides as a probe built on liquid handling glass slides. These two techniques differ significantly from the third in many aspects such as the hybridization method and the chip design. Specifically, for Affymetrix GeneChip (Figure 1.1), each gene is typically represented by 11-20 pairs of shorter oligonucleotide probes. The first component of these pairs is referred to as a Perfect Match (PM) probe and is designed to hybridize only with transcripts from the target gene, while the second component is referred to as a Mis-match (MM) probe and is designed to measure the noise introduced by non-specific hybridization. Overall, the PM/MM design is used for identification and subtraction of nonspecific hybridization and background signals. In this manner, Affymetrix arrays provide multiple measurements, a series of independent or semi-independent oligonucleotides queries of each RNA in solution (the probe set), that allow robust measures of gene expression. The increasing use of Affymetrix microarrays, along with the emergence of this technology as a potential endpoint in clinical trials, has led to requests to develop novel computational methods in bioinformatics communities and best practices in data generation and analysis.

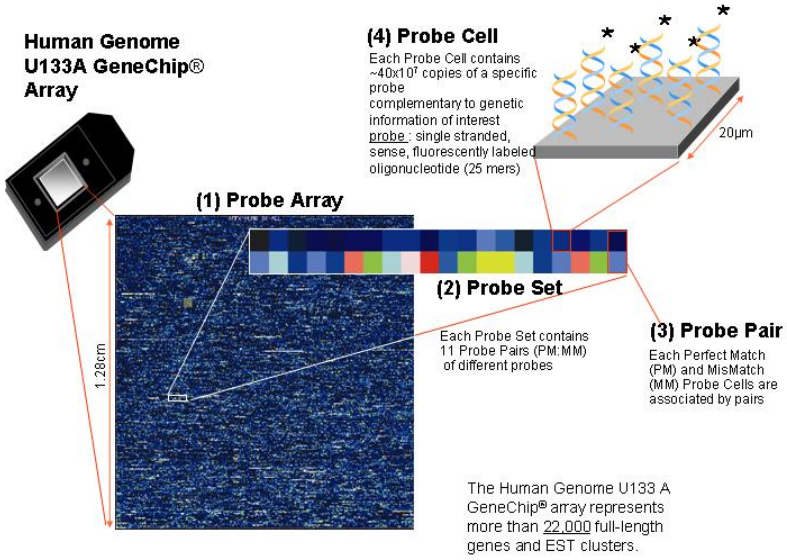


Figure 1.1: Schematic representative Affymetrix expression arrays principle (Human Genome U133A in this case). For additional information on GeneChip technology and array design, please refer to the Affymetrix Web site: **http://www.affymetrix.com/index.affx**.

Another type of data encountered in Chapter 4 is the ChIP-on-chip data. The technical platforms to conduct ChIP-on-chip experiments are also DNA microarrays, *i.e.,* the ChIP-

on-chip combines chromatin immunoprecipitation ("ChIP") with microarray technology ("chip"). ChIP-on-chip data are mainly used to investigate interactions between proteins and DNA *in vivo*, which allows the identification of DNA binding sites of proteins on a genome-wide basis [19].

In order to describe computational framework and strategy employed in the following chapters, terminologies are presented related to regulatory networks at a conceptual level.

- **Transcription factor**: Any of various proteins that bind to DNA and play a role in the regulation of gene expression by promoting transcription. (Source from: Merriam-Webster's Medical Dictionary)
- **Motif**: A distinctive usually recurrent molecular sequence (as of amino acids or base pairs) or structural elements (as of secondary protein structures) such as a simple protein motif consisting of two alpha helices. (Definition from: Human Genome Project Information)
- **Promoter region and binding site**: The part of a gene that contains the information to turn the gene on or off. The process of transcription is initiated at the promoter. A binding site in a DNA chain at which RNA polymerase binds to initiate transcription of messenger RNA by one or more nearby structural genes. (Source from: Merriam-Webster's Medical Dictionary)

The next section provides a discussion of the research motivation along with several research topics related to the analysis of microarray data associated with the consideration of unique data characteristics in gene expression profiling.

1.2 Research Motivation

Along with many other powerful genomics tools to probe the components and behaviors of biological systems, microarrays are routinely used to assess mRNA transcript levels on a genome-wide scale. As their use and acceptance increases, scientists and researchers are spending and will spend significant research efforts in developing appropriate methods for data generation and interpretation, with important questions to be asked [20].

As experimental technologies improve and more high-throughput technologies are

developed, the time and effort spent on data acquisition and collection is reduced while the analysis, extraction and processing of information becomes important parts of biological research. Information and engineering sciences are expected to play important roles in the development of this "new" biology due to their expertise in processing large amount of information. This trend has already become prominent in the field of bioinformatics and systems biology. A large amount of data covering different levels and aspects of the biological system have been accumulated, including DNA sequences, structures of proteins or other molecules, mRNA and protein expressions, molecular interactions, protein modifications and localizations, metabolic substrates fluxes, and many others. The staggering volume of molecular data resulting from the rapid adoption of sophisticated techniques of molecular biology has underscored that computational analysis, as a key link, will bridge the generation of biological data and the formulation of new hypotheses [4].

The emerging analytical tasks, such as quantifying complex biological systems, modeling or simulating complex systems, extracting statistically significant patterns from data, and so on, create tremendous opportunities for computational scientists to contribute in biological science. Despite the combined efforts of biologists, computational scientists, statisticians and software engineers, there is no '*one-size-fits-all'* solution to the analysis and interpretation of genome-wide expression data [21]. This dissertation presents one of the many works that attempt to understand composite molecular signatures and gene regulation mechanism by building computational models from large-scale genomic data. Due to the complexity of many biological systems, large-scale transcriptional profiling and insufficient data, our current progresses in this field are still preliminary. The research in this dissertation is viewed as an effort to tackle several important biological problems with novel computational methods in a machine learning framework.

1.3 Problem Statement

Data analyses of Affymetrix oligonucleotide microarrays enclose two relatively distinct steps: the development of a normalized 'signal' for each transcript on each microarray and the subsequent statistical analysis of differences in signals between

different arrays. The first step involves a series of low-level analyses, e.g., background correction, normalization and summarization [22, 23]. The second step is the application of computational and statistical methods to identify subsets of interest from the assembled array data [18]. Several issues related to the second step will be discussed throughout this dissertation.

As mentioned in the previous section, recent development of high throughput microarray transcriptional profiling techniques makes data acquisition less of a challenge. The primary challenge lies in the analytical side imposed by the noisy nature of microarray data and the so-called "small *N*, large *p*" phenomenon, *i.e.*, thousands of variables (genes, denoted as *p*) in the presence of only a few observations (samples, denoted as *N*) [24]. High throughput data analysis has raised a number of statistical and computational questions in traditional areas, such as image processing, machine learning, pattern recognition, discriminative analysis, multiple testing and Bayesian statistics [6, 9, 21, 25]. Computers and sophisticated tools are available to facilitate data analysis, but the methods that are used to analyze the data can have a profound influence on the interpretation of data, which is far from satisfactory.

Appropriate techniques for data analysis will be chosen depending both on the characteristic of data and the goal of the experiment. This dissertation summarizes our initial effort to address some of the important themes in microarray data analysis, including tissue heterogeneity correction, hidden variable modeling of latent biological processes and sparse decomposition of gene expression data to unravel transcriptional regulatory programs by data integration.

1.3.1 Tissue Heterogeneity Correction

Tissue heterogeneity is a major confounding variable in most microarray experiments [18]. In inbred mice, tissue heterogeneity can (or at least partially) be solved through normalization by using whole organs. In human experiments, however, this is seldom possible, and particularly not in clinical settings; the limited amount of human tissue that is available aggravates heterogeneity. The mixed cell populations result in a tissue heterogeneity problem encountered in all solid tissue, tumor biopsies [18] and in blood samples [26] as well. Transcriptional profiling of solid tumors is complicated by the fact

that they may contain various amounts of infiltrating tissue, such as stroma, endothelial or lymphoid cells [27]. Indeed, a recent study showed that variation as a result of tissue variability in human muscle biopsies often exceeded inter-individual variability [28]. The drawbacks of isolation of purified tumor cells by laser-capture microdissection (LCM) [29] lie in extremely low RNA yields, which requires substantial target amplification and can introduce significant bias into data analysis; thus it may not be suitable for robust, reproducible genome-wide profiling. One potential solution to the tissue heterogeneity problem is to use bioinformatics or computational methods. If a computational algorithm can be trained to distinguish the unique expression profile of each individual cell type within the mixed tissue samples, then it is possible to subtract them from future single mixed profile, obtaining a set of cell/tissue-specific expression profiles. This can be easily performed on tumor biopsies, in which the main cells of interest are tumor and contaminating normal cells.

Although there are only a few published examples so far [30, 31], it is evident that the problem can be simplified by incorporating partial *prior* biological knowledge. We will discuss in detail our unsupervised algorithm - *in silico* tissue heterogeneity correction (THC), which provides a theoretical framework to decompose mRNA expression data from mixed cell populations of targeted differential phenotypes. The algorithm is rooted from a US Patent [32] and developed based on a linear model of mRNA expression levels from phenotype-specific cells, which can be estimated by nonnegative independent component analysis (nICA) with a statistical selection of cross-phenotype independence support genes [33, 34].

1.3.2 nICA Modeling of Latent Process and Gene Module Composite

Another major challenge in computational biology is to elucidate the organization of genetic networks and their functions manifested in biological systems. Genes and gene products do not function independently, but function in complex, interconnected pathways, networks and molecular systems that, taken together, contribute to the driving of cells, tissues, organs and organisms. Therefore, it is crucial to develop a system level of understanding of gene interactions so as to elucidate how biological systems function. Yet these systems are far more complex than any system that molecular biology, genetics

or genomics have met to date [4]. To tackle this problem, a more practical approach is to focus on gene modules that are the basic building blocks of complex systems. Based on the extraction of a more interpretable characterization of transcriptional changes from gene modules, we aim to decipher the dynamics of molecular mechanisms underlying complex biological systems.

The study of patterns of gene expression alterations across many samples can serve as a powerful tool of molecular profiling to represent sets of cellular responses, phenotypes, or conditions. In general, pattern clustering provides a high-level overview of gene expression data and thus often serves as an initial step in a gene expression profiling study [21]. Simpler methods to identify genes of potential interest are to search for those that are consistently either up- or down-regulated across similar conditions, assuming that the expression of genes of interest will change across conditions concordantly. To this end, a simple statistical analysis of gene expression levels, such as using ANOVA [35] or t-test [36], will be adequate. However, identifying patterns of gene expression and grouping genes into modules in temporal gene expression profiling experiments might provide much greater insight into biological function and relevance in response to an external perturbation. As such, several statistical techniques have emerged for characterizing similar temporal gene expression profiles [37, 38]. These statistical techniques include many clustering techniques, such as *k*-means clustering and self-organizing maps [39, 40], designed for finding gene groups within the data. What these methods have in common is that (1) they simplify the data set, ideally in ways that convey implicit information about data structure and genes' relationship in functional pathways, and (2) they are considered 'unsupervised', in terms of that clusters are solely derived from data rather than from any *prior* knowledge. Such clustering techniques are based on the assumption that genes with closely related expression patterns are likely controlled by similar regulatory programs. Clusters of co-expressed genes have been identified in higher eukaryotes [41], but as the organization of regulatory elements is far more complex than in yeast, the usefulness of using such clustering methods to identify gene modules with common motifs in mammalian promoters remains unclear.

An alternative approach to identify biologically meaningful gene modules from microarray gene expression data is to use blind source separation (BSS) techniques [42-

45]. BSS attempts to separate a mixture of observations (e.g., gene expression profiles) into their different sources, using approaches with different statistical or geometrical constraints. Most BSS approaches deal with linear mixtures for source separation although nonlinear mixtures can provide a more realistic model for many applications [43, 46]. In the context of microarray data, "sources" may correspond to specific cellular responses or regulatory programs of co-regulated genes. The strong assumption of one of the BSS methods, independent component analysis (ICA), is the statistical independence between sources. Biologically, however, it is more plausible to assume that the independence holds only for those genes that actively participate in biological processes. Therefore, one needs to make further assumptions to constrain the ICA model for gene module discovery and propose to use nonnegative ICA (nICA) (Figure 1.2) for gene module identification [47-49]. In principle, nICA exploiting the non-negative nature of molecular expressions can be thought as a projection method where the expression levels are projected onto some new non-negative bases (*i.e.,* components) with minimum statistical dependence. We believe that the nICA approach is better positioned for gene module identification since it reflects the biological reality more closely.

In addition to the problem formulation and algorithm implementation, there are several other sub-problems in the development of the nICA approach for gene module identification. First, we need to select the most informative observations from available observations in order to reduce the computational complexity and avoid the possible singularity problem; we propose an information-theoretic approach to help select informative observations. Second, we need to determine the number of components to properly reflect the number of underlying biological processes; we propose to use a stability analysis approach to help determine the number of components. The experimental results on several yeast data sets and a muscle regeneration data set have demonstrated the advantage of the nICA approach over several conventional BSS approaches.

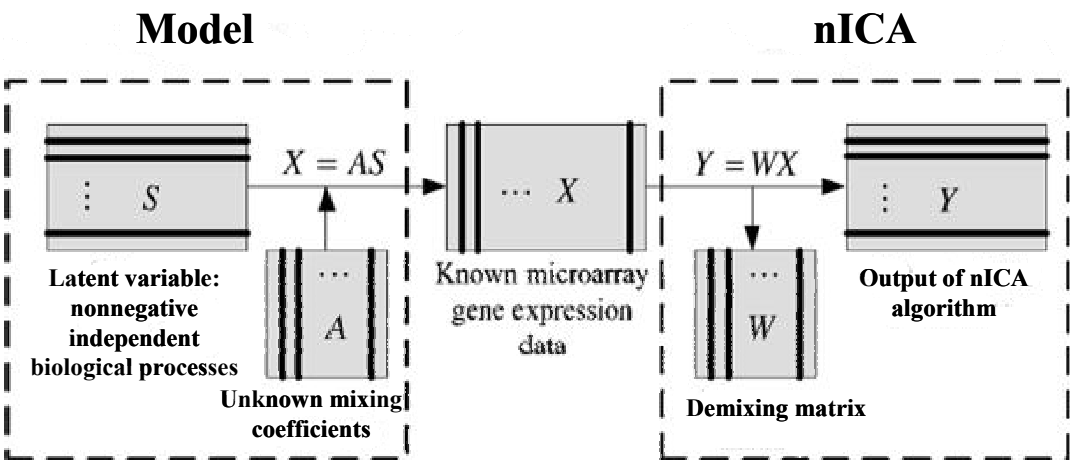


Figure 1.2 A black diagram of the nICA framework in application to microarray data analysis.

1.3.3 Motif-guided Sparse Decomposition to Unravel Transcriptional Regulatory Programs

In order to understand how genes function under different conditions with internal and/or external stimuli, it is necessary to know the fundamental processes that regulate genes to fulfill biological functions. Transcriptional regulatory processes constitute a network of genes, regulators and their interactions that form gene regulatory networks. The final goal of this dissertation is to develop a computational method to reconstruct gene regulatory networks from multiple data sources. In particular, we focus on the gene regulation mechanisms that can be revealed by physical (or molecular) interactions such as protein-DNA bindings. Consequently, we aim to reconstruct transcriptional regulatory networks for elucidating how the expression of the entire genome is controlled by a relatively small number of transcription factors.

As is known, detailed analysis of binding sites for known transcription factors within the promoters of co-regulated genes can help understand gene regulatory networks within the cell. A recent technical development that can identify genes regulated by particular transcription factors has emerged using chromatin immunoprecipitation (ChIP) and hybridization to microarrays (ChIP-on-chip) as a complementary tool to study protein-DNA interactions [50]. This allows the identification of all sites within the genome at which specific transcription factors bind [11, 51]. Overall, this approach may eventually

lead to the generation of accurate and comprehensive maps of transcriptional regulatory mechanisms. Nevertheless, most ChIP-on-chip analysis methods use conservative approaches aimed at minimizing false-positive transcription factor targets, which may not be able to detect complex regulatory interactions between transcription factors and their target genes. It is needed to develop computational approaches with improved sensitivity in detecting complex binding events from ChIP-on-chip data.

Apart from this, the major bottleneck currently in microarray data analyses lies in the integration of multiple biological data sources [52]. In particular, the integration of partially complementary information is not yet well established for the reverse engineering of gene regulatory networks. Especially, gene expression data alone do not provide sufficient information about gene regulation mechanisms. It is necessary to incorporate other types of data for reliable reconstruction of regulatory networks. Therefore, we need to develop integration strategies to extract information from multiple sources and construct regulatory networks that fit well to multiple data sources.

In this dissertation, we aim to tackle the problem of transcriptional module identification, which essentially requires finding sets of transcription factor binding sites (TFBS) that co-occur in promoter regions of genes with a common expression pattern. We have developed an algorithm that takes into account the coordination between gene expression profiles and motif information so as to find a balance point of "co-expression" and "co-regulation". In order to learn the membership of transcriptional modules, we model a transcriptional regulatory network in the form of a bipartite graph [53] (Figure 1.3), with graph vertices representing genes or transcription factors (TFs) and edges representing transcriptional relationships between TFs and their target genes. Since the activities of these TFs are usually unobserved owing to post-translational modifications, they are treated as hidden or latent variables in the statistical model. Specifically, we propose to integrate motif information and expression data in a two-step approach as follows. (1) We use motif information to guide finding co-regulated gene patterns under the regulation of the common TFs. Since transcriptional regulation is influenced by TF activities, we use these genes expression levels to approximate TF activities. (2) Because the expression of the entire genome is controlled by a relatively small number of transcription factors, a reasonable assumption for gene regulatory networks is that they

are likely partially connected, *i.e.*, TF-gene connections are sparse. We then utilize a sparse component analysis (SCA) method [54] to decompose gene expression data hence to recover the TF-gene connectivity information. This two-step approach will be termed as motif-guided sparse decomposition (mSD) method in this dissertation [55].

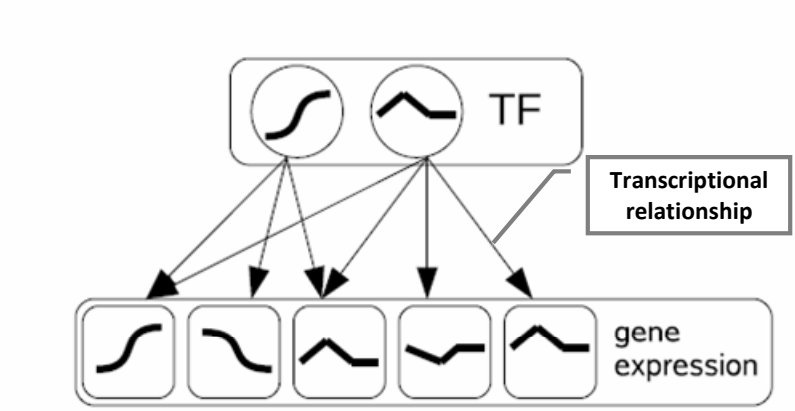


Figure 1.3: A bipartite graph representation of a transcriptional network. A small number of transcription factors (TFs), represented by circles, regulate a large number of genes (represented by squares) by binding to their promoter regions (Figure source: [51]).

1.4 Summary of Contributions

In the context of the research topics discussed above, we summarize the main contributions of this dissertation in this section as well as both the chapters in which they are covered and relevant publications.

- Microarray profiling of solid tumor tissues reflects gene expression corresponding to malignant cells as well as many different types of contaminating normal cells. Given the necessity for understanding complex biological processes such as development and carcinogenesis within the context of intact tissues, we apply nICA theory, together with statistically-principled geometrical selection of cross-phenotype independence support genes (ISGs) and an ensemble study of underlying phenotypes, to computationally correct tissue heterogeneity in gene expression profiling. We assess the feasibility of our proposed algorithms – nonnegative partially independent component analysis (nPICA) on a series of microarray data sets mixed *in silico* and the results show that computational

dissection can resolve the problem of tissue contamination, revealing novel cancer-specific gene expression. Through tissue heterogeneity correction, this computational decomposition method has the potential to increase the sensitivity of predictive signatures for differential gene expression experiments performed on complex tissues. We anticipate that this approach offers substantial utility and should be broadly applicable to identifying gene expression changes in tissues composed of multiple cell types. (Chapter 2)

- It is important is to identify gene modules in order to comprehend the mechanism of genome-wide gene regulation. In this dissertation, we apply a new gene clustering approach – non-negative independent component analysis (nICA) – for microarray gene expression data analysis. In conjunction with the nICA modeling, VIsual Statistical Data Analyzer (VISDA) is used to group genes into modules in the latent variable space. As a promising methodology, nICA can cluster genes into modules that are potentially co-regulated by a group of transcription factors. Compared with traditional matrix decomposition clustering methods, our method deploy more effectively in the identification of groups of genes that share a similar biological function or regulation mechanism. In the nICA approach, we also develop an information-theoretic procedure for input sample selection and a novel stability analysis approach for proper dimension estimation. (Chapter 3)

  Related publications are listed as follows:

  - T. Gong, J. Xuan, C. Wang, H. Li, E. P. Hoffman, R. Clarke and Y. Wang, (2007). "Gene Module Identification from Microarray Data Using Nonnegative Independent Component Analysis". **Gene Regulation and Systems Biology** 1: 349-363.
  - T. Gong., J. Xuan, Y. Zhu, H. Li, R. Clarke, E. P. Hoffman and Y. Wang, (2006). "Composite Gene Module Discovery Using Non-negative Independent Component Analysis," Proc. IEEE*/NLM Life Science Systems and Applications Workshop*, pp. 1-2, Bethesda, MD.
  - T. Gong., Y. Zhu, J. Xuan, H. Li, R. Clarke, E. P. Hoffman and Y. Wang, (2006). "Latent Variable and nICA Modeling of Pathway Gene Module Composite," *Proc. Intl Conference of the IEEE Engineering in*

Medicine and Biology Society, pp. 5872-5875, New York City, New York.

- Transcriptional regulation occurs when certain transcription factors (TFs) bind to the DNA at binding sites (TFBSs) and affect the transcription of regulated genes. Unraveling the transcriptional regulatory program, *i.e.,* detecting groups of cooperative transcription factors and co-regulated genes, is a fundamental goal of computational biology, yet still remains a challenge. We develop a new approach, namely motif-guided sparse decomposition (mSD), for transcriptional program identification. The mSD approach combines the motif information and gene expression data with an emphasis on the interplay of co-expression and co-regulation. Motif information is initially used to define potential target genes, providing a *prior* knowledge of the regulatory network topology. A sparse latent variable model is then used to integrate gene expression data to identify which of the potential target genes are actually activated by transcription factors in the context of experimental conditions. At the same time, the contribution of each data type is measured in a quantitative manner by maximizing conditional likelihoods of expression similarity and motif similarity simultaneously. The results validate that the mSD approach is an effective method to uncover the hidden transcriptional regulatory mechanisms that can facilitate the discovery of mechanisms of transcriptional regulation. (Chapter 4)

  Related publications are listed as follows:

  - T. Gong, J. Xuan, R. B. Riggins, H. Li, E. P. Hoffman, R. Clarke and Y. Wang, "Motif-guided Sparse Decomposition of Gene Expression Data for Regulatory Module Identification," submitted to **Bioinformatics**, 2009.
  - T. Gong, J. Xuan, L. Chen, R. B. Riggins, Y. Wang, E. P. Hoffman, and R. Clarke, (2008). "Sparse Decomposition of Gene Expression Data to Infer Transcriptional Modules Guided by Motif Information". *Proc. Intl Symposium on Bioinformatics Research and Applications*, pp. 244-255, Atlanta, Georgia.

– T. Gong, J. Xuan, R. B. Riggins, Y. Wang, E. P. Hoffman, and R. Clarke, (2008). "Exploring Transcriptional Modules by Integrative Gene Clustering Guided by Transcription Factor Binding Information". *Proc. Intl Conf. on Bioinformatics & Computational Biology*, pp. 191-197, Las Vegas, Nevada.

1.5 Outline of Dissertation

In this dissertation, we aim to present a unified framework of latent variable modeling to address three important problems in computational biology: (1) tissue heterogeneity correction, (2) gene module identification and (3) regulatory network reconstruction from high-throughput data by accounting for underlying biological and network constraints. Following the problems statement in Section 1.3, we structure the remaining parts of the dissertation as follows:

Chapter 2 lays out a framework for computational dissection of tissue heterogeneity. It first states the objective of the modeling framework, and then briefly reviews current existing approaches for tissue heterogeneity correction. Specifically, we will describe a novel approach, non-negative partially-independent component analysis (nPICA), based on a latent variable model for tissue heterogeneity correction in detail. The approach includes a supervised selection of independence-support genes (ISGs) to validate the feasibility of nPICA; a unsupervised convex analysis-based blind separation of non-negative yet dependent sources for tissue heterogeneity correction in microarray gene expression studies; and an experimental design for computational correction of expression profiles to reduce tissue heterogeneity in a series of cancer cell lines and its impact on diagnostic prediction.

Chapter 3 introduces a complete approach for latent variable and nICA modeling of gene modules and biological processes. In conjunction with the nICA modeling, VIsual Statistical Data Analyzer (VISDA) [56] is used to group genes into modules in the latent variable space. We introduce two different ICA models to elucidate the composite gene modules. Following the second model, we describe each element of the proposed approach at a detailed level, including an information-theoretic procedure for input

sample selection and a novel stability analysis approach for proper dimension estimation. We also provide comparison results of the proposed approach with several benchmark methods, *i.e.*, matrix-decomposition-based clustering methods and a soft clustering approach, to evaluate the performance. Application of the proposed algorithm on a microarray muscle regeneration data set is performed and the resulting experimental results are presented to demonstrate the great utility of the proposed method in biomedical research.

Chapter 4 addresses a latent variable model inferring regulatory modules from multiple biological sources. It first states the problem and the assumption associated with the proposed computational method – motif-guided sparse decomposition (mSD). The mSD approach is then implemented as a two-step algorithm consisting of (1) transcription factor activity estimation and (2) regulation strength estimation. Specifically, a motif-guided clustering method is first developed to extract the genes that common TFs regulate, with which to infer activity levels of unobserved regulators that control them; sparse component analysis is then followed to estimate the regulation strength, with which to identify the target genes of transcription factors. The mSD approach has been tested for its improved performance in finding regulatory modules using simulated *Saccharomyces cerevisiae* data and yeast cell cycle data. As a result, the mSD approach has indeed revealed functionally distinct co-expressed and co-regulated gene modules, enriched with biologically validated transcription factors and their target genes. Finally, we demonstrate the efficacy of the mSD approach on real breast cancer cell line data, uncovering several important gene regulatory modules related to endocrine therapy of breast cancer.

Chapter 5 draws the conclusion about this dissertation research. We summarize the original contributions of the research and then point out limitations of current approaches and possible extensions for future work.

2 In Silico Dissection of Tissue Heterogeneity in Gene Expression Profiling

2.1 Introduction

The advent of the technology of high-throughput transcriptional profiling using DNA microarrays has enabled the measurement of genome-wide regulatory changes in distinct circumstances, which may completely change biology - perhaps more than the advent of molecular biology in the 1970s [57]. Molecular analysis of cells in their native tissue environment provides the most accurate picture of the *in vivo* disease state [58]. However, due to the complicated structures of tissues and cellular environments, composed of large numbers of disparate yet interacting cell populations, accomplishing this goal becomes difficult. By grinding up a piece of tissue and applying the extracted molecules to a panel of assays, the resulting arrayed expression profiling may contain only a fraction of the total cell subpopulation of interest [59].

2.1.1 Background and Significance

Gene expression profiling by microarrays often represent heterogeneous tissues with distinct cellular compartments reflect weighted averages of expression levels within different cellular populations. As a consequence, the presence of these cells could mask the detection of genetic and gene expression alterations in the tissues or cells. Differential regulation of genes associated to inter-cellular changes can be hard to distinguish or even disappear entirely due to the change of compartment size; conversely, genuine regulation within a given cell type associated with different cell states may be swamped by the changes in the abundance of cellular compartments [60]. In addition to the measured

variable(s) of interest, there will tend to be sources of signal due to factors that are unknown, unmeasured, or too complicated to capture through simple models. We anticipate that failure to incorporate these sources of heterogeneity into an analysis can have widespread and detrimental effects on the study. Not only can this reduce power or induce unwanted connections across genes, but it can also introduce sources of spurious signal to many genes. Recently, Lamb *et al*. [61] proposed the '*Connectivity Map*' to identify functional connections between cancer subtypes, genetic background, and drug action. Lamb *et al*. noted: "*expression heterogeneity (e.g., due to cell type and batch effects) presented a major hurdle for extracting relevant biological signal from the 'Connectivity Map'*" [61].

Currently two main approaches have been taken to deal with this problem: *global survey* and *microdissection* [58]. The global survey approach assumed that expression profiles can be compared to profiles from purified cell types to identify a subset of genes with expression patterns that are specific to the purified cells, which can be extracted from the RNA directly from a heterogeneous piece of tissue (Figure 2.1). An earlier study [62] using grossly dissected breast-cancer specimens has demonstrated in this way the ability to circumvent the problem of sample heterogeneity. A subset of genes with expression patterns that are specific to the tumor cells can be identified by comparing expression profiles from whole solid tumors with profiles from potential unchanged infiltrating cell types, such as lymphocytes or endothelial cells [27]. Subsequent data analysis and sample clustering can then be carried out only on this 'intrinsic gene subset'. With this approach, Perou and colleagues used expression patterns to define clinically significant subtypes of human breast cancer [62, 63]. However, disentangling the dynamics of cell populations requires precise identification of cell types [64], ideally based on detailed measurements of molecular markers specific to each cell type. Identification of such markers is not trivial; especially from complicated gene expression profiles obtained in metazoan organisms. In contrast to the global survey approach, the microdissection method uses specialized technology to separate and isolate the cancer cells (or other tissue cell subpopulations of interest) directly from the tissue. The molecular changes in the isolated subpopulations can then be analyzed individually. Recently, Staub *et al.* [65] investigated three compartments of colorectal tumors, the

invasion front, the inner tumor mass, and surrounding normal epithelial tissue by microdissection and microarray-based expression profiling. They concluded that in both tumor compartments, many genes were differentially expressed when compared to normal epithelium. The sets of significantly deregulated genes in both compartments overlapped to a large extent as expected and revealed various known and novel pathways that could have contributed to tumorigenesis. On the contrary, cells from the invasion front and inner tumor mass did not show significant differences in their expression profiles, neither on the single gene level nor on the pathway level. All those results have implications for necessity of tissue heterogeneity correction to better understand tumor profiling.

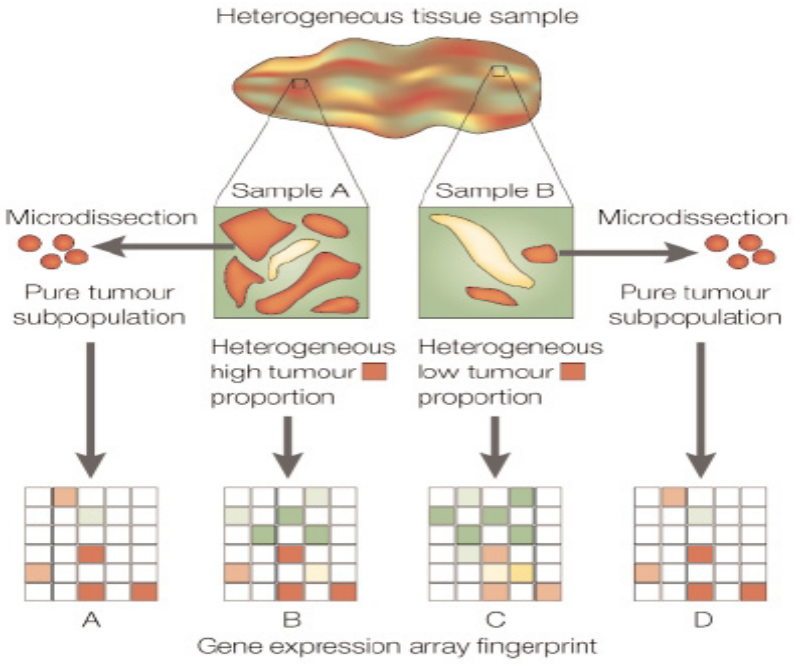


Figure 2.1: Global survey *versus* microdissection approaches to gene profiling from heterogeneous tissue specimens. (Figure source: [56])

The two methods of direct tissue sampling — global survey and microdissection — have their advantages and disadvantages. The global approach begins with a higher amount of starting material and is much less labor intensive. However, the disadvantage of this approach is that the actual proportion of the diseased cell subpopulation is unknown and thus not suitable for accurate gene expression profiling. A low-abundance yet interested mRNA type might be overwhelmed by the contaminating higher-

abundance species. To overcome these uncertainties, powerful analytical algorithms are needed to gauge the relative abundance of unknown tissue cell subpopulations within the tissue sample. The advantage of microdissection is that it focuses directly on the disease subpopulation to acquire pure tissue cell subpopulations, but it has difficulty in obtaining sufficient quantities of purified material to perform robust, reproducible genome-wide profiling [60]. A certified pathologist is required to select the cells to be microdissected under microscope, and the microdissection itself requires specialized technology and training. The critical niche in cancer research that can be fulfilled by microdissection is to profile the transition stages from normal cells, through carcinoma *in situ*, to invasive cancer. This transition takes place in microscopic regions of the tissue and cannot be adequately studied by the global survey approach.

2.1.2 Latent Variable Model Framework

Since computational dissection does not require microdissection of all samples or change of routine biological protocols, several authors have tried to answer whether it is possible to decompose DNA microarray data from a cell population to survey the proportions of different cell types, by treating specific transcriptional patterns in DNA microarray data as cell-type specific markers through computational methods. In order to answer this question, it is helpful to describe these methods in a simplified two-dimensional yet well-defined ***latent variable model framework*** [32]. Derived from two random biopsies of the same tissue (or two randomly divided tissue samples of one biopsy), we construct a two-dimensional space, $\mathbf{s}(i) = [s_{\text{cancer}}(i), s_{\text{stromal}}(i)]^T$, with each dimension denoting the specific expression signals for disparate cells respectively. Furthermore, we use the following linear latent model to represent original expression signals as a mixture of each compartment signal:

$$\begin{bmatrix} x_{\text{biopsy1}}(i) \\ x_{\text{biopsy2}}(i) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} s_{\text{cancer}}(i) \\ s_{\text{stromal}}(i) \end{bmatrix} . \tag{2.1}$$

The Eq. (2.1) describes how the observed gene expression profiles $\mathbf{x}(i)=[x_{\text{biopsy1}}(i), x_{\text{bopsy2}}(i)]^T$ are generated by a process of mixing the latent gene expression sources, where i is the gene index and the two heterogeneous specimens differ in the proportion of

different cell populations specified by mixing matrix **A,** the first item on the right side of Eq. (2.1).

Lu *et al.* are the pioneers in applying a simulated annealing-based algorithm to identify the proportions of cells using Eq. (2.1) [30]. By knowing the typical mRNA expression patterns of yeast cells in specific cell cycle phases (G1, S, G2, M, and M/G1), the expression data of asynchronous cells can be modeled as the weighted linear combination of expression data from cells in each synchronized population. The proportions of cells in each phase of the cell cycle are informed by the estimation of numerical weights from linear latent model as in Eq. (2.1). Because 696 identified genes exhibiting cell cycle-dependent changes in mRNA expression levels are known for each of the phases of the cell cycle [66], this provides a straightforward set of equations to solve by standard methods, with 696 equations with only 5 unknowns (the cell fractions). Following this line, Wang *et al*. [60] applied and extended of this strategy to decompose mouse mammary tissue and used the residuals of their fit to separate the differential expression due to changes in tissue composition. Very recently, Abbas *et al*. [26] applied microarray decomposition to measure proportions of cell types in blood samples and employed the results to study immune disease.

Another line of approaches is *in silico* dissection of cell-type-associated patterns by Stuart *et al.* [31]. They conducted a regression-based informatics approach to identify cell-type-specific patterns of gene expression in prostate cancer. Eighty-eight tissue samples from forty-one subjects undergoing prostatectomy for clinically early stage localized prostate carcinoma were independently scored by a panel of four pathologists for fractional composition of the four cell types: tumor, stroma, benign prostatic hypertrophy (BPH), and dilated gland cells. Due to the relative paucity of dilated gland cells in the samples (median proportion = 5%), the proportions of the other three cell types were then linked *in silico* to gene expression levels determined by microarray analysis, revealing unique cell-specific profiles. To assign gene expression to particular cell types within tumor specimens, a linear model was constructed in which it was assumed that the contribution to gene expression of any one cell type depends only on the proportion of that cell type and its corresponding characteristic cell-type expression level, s_{ij}, but not on the proportions of other cell types present. In Eq. (2.2), the average

expression level x_{kj} of gene j in a sample k is the average of cell type expectations, s_{ij}, weighted by cell type fractions a_{ki}:

$$x_{kj} = \sum_i a_{ki} s_{ij} + \varepsilon_{kj} \quad . \tag{2.2}$$

What we seek in the above model is an algorithm to recover the source profiles from their observed mixtures where the source signal $\mathbf{s}(i)$ and the mixing matrix $\mathbf{A}$ are both unknown. In this dissertation, we first investigate a supervised selection of independence-support genes (ISGs) to help identify typical patterns of distinct tumor compartments, which is the key component of nonnegative partially-independent component analysis (nPICA). Next, in an unsupervised mode we exploit convex analysis of mixtures (CAM) to identify ISGs or its counterpart, namely invariantly-expressed genes (IEGs), for tissue heterogeneity correction. In Section 2.2, we first formulate the problem as a linear instantaneous mixing model. Then in Section 2.3 and 2.4, we describe these approaches in detail. A performance analysis, including a comparison between the proposed algorithm and several alternatives, is also provided in Section 2.4. Finally, conclusions are presented in Section 2.5.

2.2 Problem Formulation

In transcriptional profiling of solid tumors, it is often necessary to process mixture data obtained by biopsies. The processing aims at identifying the pure components of the materials and estimating the concentration of each component. These objectives are formalized as a source separation problem, where the linear instantaneous mixture model holds. We first introduce the general blind source separation problem and then introduce the assumptions necessary to arrive at our problem statement.

The linear instantaneous mixing model is expressed as:

$$\tilde{\mathbf{x}}(i) = \tilde{\mathbf{A}}\mathbf{s}(i) + \tilde{\mathbf{n}}(i) = [\tilde{\mathbf{a}}_1, \cdots, \tilde{\mathbf{a}}_d]\mathbf{s}(i) + \tilde{\mathbf{n}}(i), \tag{2.3}$$

where $\mathbf{s}(i) = [s_1(i), \cdots, s_d(i)]^T$ denotes the $d \times 1$ vector of nonnegative source signals corresponding to gene i, $\tilde{\mathbf{x}}(i) = [\tilde{x}_1(i), \cdots, \tilde{x}_m(i)]^T$ the $m \times 1$ $(d \le m)$ vector of measured transcriptional profiles of gene i, and $\{\tilde{\mathbf{a}}_t\}_{t=1}^d$ unknown mixing coefficients that can be termed the *spatial signature* of the sources [67]. Concatenating N genes yields the

following matrix notation:

$$\tilde{\mathbf{X}} = \tilde{\mathbf{A}}\mathbf{S} + \tilde{\mathbf{N}} \ , \tag{2.4}$$

where the matrices $\tilde{\mathbf{X}} \in \mathbf{R}^{m \times N}, \mathbf{S} \in \mathbf{R}_{+}^{d \times N}$, and $\tilde{\mathbf{N}} \in \mathbf{R}^{m \times N}$. We use the *tilde* notation to denote the variables prior to dimension-reduction preprocessing.

We first assume that $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{S}}$ are full-rank, so that $\tilde{\mathbf{X}}$ lies in the *d*-dimensional column space of $\tilde{\mathbf{A}}$. (Note that the assumption that $\tilde{\mathbf{A}}$ is full-rank is equivalent to assuming that *d* sources have linearly independent spatial signatures.) Therefore, we can transform Eq. (2.4), with no loss of information, into a problem of lower dimension. One approach is to decompose $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathbf{T}}$ *via* singular vector decomposition (SVD) [68] to get $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathbf{T}} = \mathbf{U}\boldsymbol{\Sigma}^{\mathbf{2}}\mathbf{U}^{\mathbf{T}}$, where $\mathbf{U}$ is a $m \times d$ orthogonal matrix whose columns are the eigenvectors of $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathbf{T}}$ and Σ is a diagonal $d \times d$ matrix with eigenvalues. We use $\mathbf{U}$ to produce the '*un-tilded*' equation as follows:

$$\mathbf{X} \equiv \mathbf{U}^{\mathbf{T}}\tilde{\mathbf{X}} = \mathbf{A}\mathbf{S} \ , \tag{2.5}$$

where $\mathbf{A} \equiv \mathbf{U}^{\mathbf{T}}\tilde{\mathbf{A}}$ is a $d \times d$ full-rank matrix. Through this way, we can change the linear instantaneous mixing model into an exactly-determined case.

To achieve the separation, any prior knowledge and assumption about the mixing process and source signals should be taken into account since this inverse problem is ill-posed, in the sense that the solution is not unique [69]. Principal component analysis (PCA), which is the most popular approach for the analysis of multivariate data, assumes that the signals to reconstruct are mutually uncorrelated, but this orthogonality constraint does not ensure either the uniqueness or the non-negativity of the solution. A more constraining assumption used for source separation is the mutual independence of sources leading to the independent component analysis (ICA) concept [70], for which many algorithms has been developed. Assuming the mutual independence of sources with a non-Gaussian distribution yields a solution that is unique (up to order and scale indeterminacies), but it does not ensure explicitly the non-negativity of both sources and mixing coefficients. Clearly, if the nonnegative source signals are mutually statistically independent, they can be separated successfully by ICA methods and their non-negativity will be ensured implicitly. Our algorithm is also based on the identification of non-Gaussian components of phenotype-specific pattern in a sample space under the

assumption that Gaussian distributions represent noise. The identification of non-Gaussian, typically super-Gaussian, is biologically relevant in an expression profiling situation as most genes, e.g., housekeeping genes, are not expected to change at a given physiological/pathological transition, and thus conform to a Gaussian distribution. Only the genes that constitute the physiological/pathological state will change and thus produce super-Gaussian distributions [43, 44]. But when source signals are not mutually independent in the entire gene space, the non-negativity information should be considered instead.

Summarizing the previous derivations and assumptions, we arrive at the following problem statement with which we will use to derive our learning algorithm.

Problem Statement: Let $\mathbf{X} = \mathbf{AS}$ where $\mathbf{S} \in \mathbf{R}_{+}^{d \times N}$. Given **X**, find the de-mixing matrix $\mathbf{W} = \mathbf{PDA}^{-1}$ and source matrix **S**, where individual rows of **W** are a rescaling and permutation of those of the inverse mixing matrix **A**, **P** a permutation matrix and **D** a scaling matrix.

2.3 Supervised Nonnegative Partially-Independent Component Analysis (nPICA)

2.3.1 Supervised ISG selection

It is not possible in general to directly estimate **A** and **S** using the entire gene space. Expression levels of most genes are expected to be unchanged over most circumstances across different phenotypes [71]. The basic intuition here is that if the subset of genes contain many unwanted dependent genes, there will be source estimates that maximize the independence (*i.e.*, as independent as possible [68]) but produce overshoot from the "true" source profiles, resulting in a large estimation error. Only those genes that correspond to phenotype-associated molecular signatures can be considered as statistically independent. Based on this perception, we define a *phenotype-specific pattern* as the union of phenotypic up-regulated independence-support gene (ISG) subsets [33, 72].

Hence, developing mathematical criteria for selecting ISG indices is a key step

towards differentiating assorted phenotypes. Statistical independence requires high "unpredictability" among source profiles over ISGs, and the self-unpredictability of each gene contributes quantitatively to such statistical independence [33]. It is widely accepted that differentially-expressed genes (DEGs) across different phenotypes provide novel insights into altered underlying biological processes [73]; DEGs are often detected by fold-change. To assure the mutual differentiation across all phenotypes, we propose a *one-versus-each* extension of DEGs, termed as subset of ISG that can be defined as

$$\mathbb{S}_{\text{subISG}} = \bigcup_{j=1}^{d} \text{subISG}(j) = \bigcup_{j=1}^{d} \bigcap_{k \neq j}^{k \in d} \left\{ i \middle| \frac{\overline{s}_j(i)}{\overline{s}_k(i)} \geq \tau \text{ and } \sigma_j^2(i) \leq \gamma \right\}, \tag{2.6}$$

where τ and γ are pre-defined thresholds; $\overline{s}_j(i)$ and $\sigma_j^2(i)$ denote the sample mean and variance (*i.e.*, ensemble) of the normalized cell-type-associated gene expressions derived from a sufficient number of microdissected samples corresponding to phenotype *j,* $j = 1, \cdots, d$. Apparently, this subset of genes is statistically independent. In a practical and complete approximation of ISG set, genes with only moderate expressions, which often represent up-stream genes actually participated in various cascaded biological pathways, shall also be included:

$$\mathbb{S}_{\text{ISG}} = \mathbb{S}_{\text{subISG}} \bigcup \left\{ i \,|\, \left\| \overline{s}(i) \right\|_2 \leq \varepsilon \right\}, \tag{2.7}$$

where ε is a pre-defined small threshold. However, we will later show that it seems to often work reasonably well when using $\mathbb{S}_{\text{subISG}}$ alone.

To determine the optimal value of τ, we use '*minimum or no overshoot*' as the guiding criterion and use the ensemble subISG as both inputs and desired outputs in nPICA trials. The rationale behind this criterion is that, when the ensemble phenotypic patterns containing many dependent genes are used as both inputs and desired outputs, some level of '*overshoot*' at the output end of nPICA is theoretically expected. Starting initially with a relatively big value, τ gradually decreases to an optimal value that corresponds to the most complete subISG with '*minimum or no overshoot*' among all nPICA trials.

2.3.2 nICA Algorithm

After correct identification of ISGs, we will deploy a learning algorithm, non-negative independent component analysis (nICA) [74], to decompose $\mathbf{X}(\mathbb{S}_{\text{subISG}})$ so as to estimate **A** and **s**(i) (the i-th column in **S**).

If we define nonnegative well-grounded sources as

$$\begin{aligned} &p(s_k < \delta) > 0 \quad for\ \forall \delta > 0 \\ &p(s_k < 0) = 0 \quad k = 1, 2, ..., d \end{aligned} \quad , \tag{2.8}$$

i.e., if each source has non-zero probability density function (PDF) all the way down to the zero [48], then it has been proven [75] that we can find **y** = **Us** where **U** is a square orthonormal rotation and permutation matrix. It is equivalent to say that the elements y_i of **y** are a permutation of sources if and only if all y_i are nonnegative. We note that **y** = **Us** can be rewritten as **y** = **Wz** = **WVx** = **WVAs,** where **V** is a whitening matrix, **z** a pre-whitened observation vector and **W** an unknown orthonormal (rotation) matrix. Therefore, we can consider nICA as an iterative procedure with the following two steps: 1) removing the second order statistics by whitening; and 2) searching for a rotation matrix where all the data fit into the positive quadrant.

By defining the cost function J as:

$$\begin{aligned} &J(\mathbf{W}) = E\{\left\|\mathbf{z} - \mathbf{W}^T\mathbf{y}^+\right\|^2\} \\ &\mathbf{y} = \mathbf{W}\mathbf{z} \\ &y_i^+ = \max(0, y_i) \\ &\mathbf{y}^+ = (y_1^+, y_2^+, \cdots y_d^+) \end{aligned} \quad , \tag{2.9}$$

a learning algorithm to find the de-mixing matrix **W** can be summarized as follows [74]:

1) Pre-whitening the observed data **x**:

$$\mathbf{z}(i) = \mathbf{V}\mathbf{D}^{-1/2}\mathbf{V}^T\mathbf{x}(i), \tag{2.10}$$

where **V** is the orthogonal matrix of eigenvectors of the partial sample covariance matrix:

$$\Sigma_{\mathbf{x}(\mathbb{S}_{\text{subISG}})} = \frac{1}{N_{\mathbb{S}_{\text{subISG}}}} \sum_{i \in \mathbb{S}_{\text{subISG}}} \left(\mathbf{x}(i) - \overline{\mathbf{x}}(\mathbb{S}_{\text{subISG}})\right)\left(\mathbf{x}(i) - \overline{\mathbf{x}}(\mathbb{S}_{\text{subISG}})\right)^T, \tag{2.11}$$

with $\bar{\mathbf{x}}(\mathbb{S}_{\text{subISG}})$ being the sample mean of $\mathbf{X}(\mathbb{S}_{\text{subISG}})$ over $N_{\mathbb{S}_{\text{subISG}}}$ ISGs, and **D** is the diagonal matrix of corresponding eigenvalues. Note that to assure the non-negativity condition, we do not remove the mean $\bar{\mathbf{x}}(\mathbb{S}_{\text{subISG}})$ in the pre-whitening process.

2) Using a gradient descent algorithm to minimize the cost function J in Eq. (2.9):

$$\mathbf{W} = \mathbf{W} - \gamma \frac{\partial J}{\partial \mathbf{W}}. \tag{2.12}$$

3) Projecting the unconstraint gradient descent set onto a set of orthonormal vectors:

$$\mathbf{W} = (\tilde{\mathbf{W}}\tilde{\mathbf{W}}^{\mathrm{T}})^{-1/2}\tilde{\mathbf{W}}. \tag{2.13}$$

Lastly, the cell-type-associated gene expression profiles can be readily estimated via $\mathbf{s}(i) = \mathbf{W}\mathbf{z}(i)$.

2.3.3 Experimental Results

To test nPICA's ability in separating gene expression profiles of mixed differential phenotypes, we used public microarray gene expression data from the Gene Logic site. Source profiles were derived from central nerve systems (CNS) and liver, respectively. We performed ISG selection based on the scatter plot of source profiles and 324 ISG indices were subsequently used to define a valid nPICA model (Figure 2.2c). nPICA blindly and correctly recovered the true underlying source profiles from their observed mixtures (Figure 2.2d).

We used an iterative procedure to determine the optimal value of τ, with which the selected ensemble ISG indices define the most complete independent segments. Subsequent nPICA produces minimum '*overshoot*' when the ensemble source independent segments are used as both inputs and desired outputs. Figure 2.3 shows some intermediate results from such an iterative procedure. When our ensemble ISG indices are defined within independent segments, the '*overshoot*' is small (at least only few negative values appear in the estimates). However, when the ISG set contains even small part of dependent segments (*i.e.*, housekeeping genes), there will be severe '*overshoot'*.

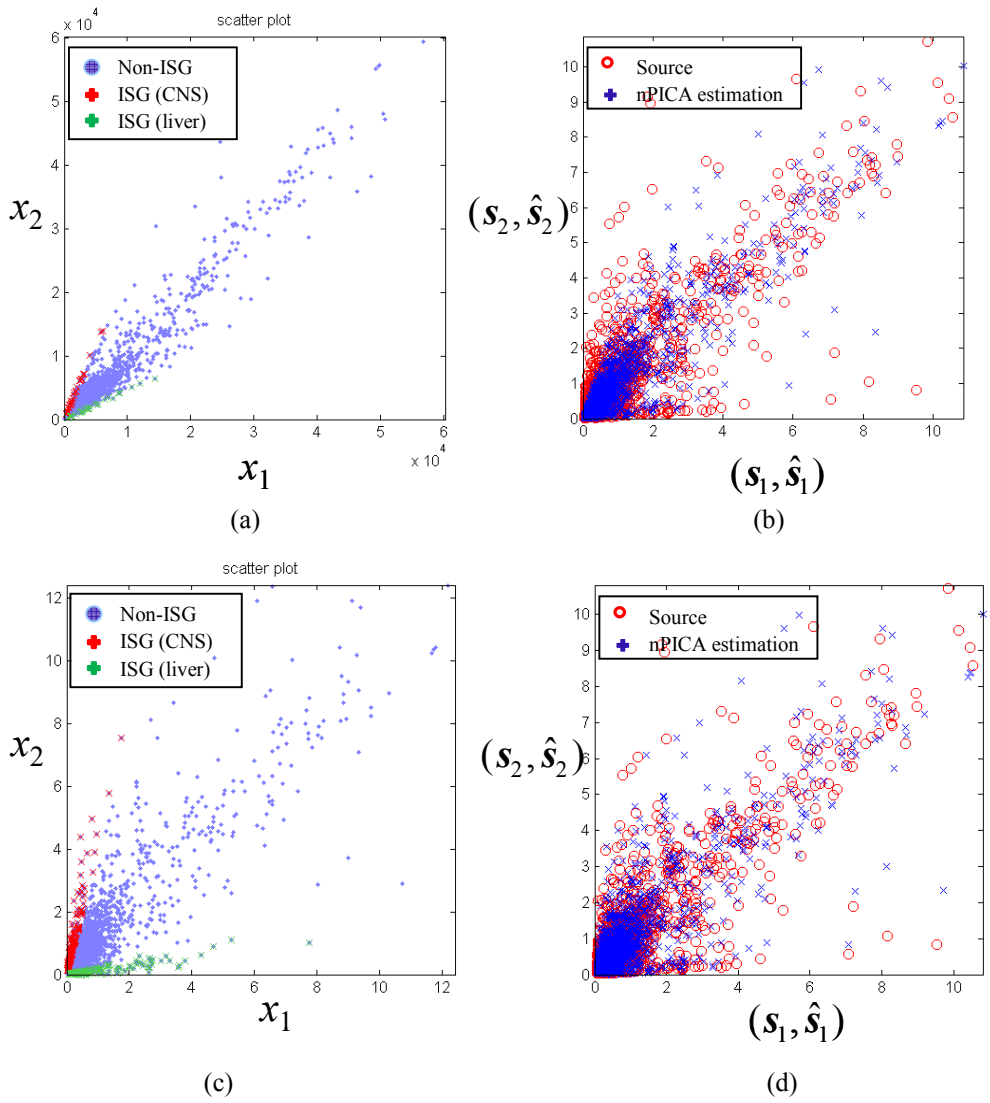


Figure 2.2: (a) Initial ISG selection based on the scatter plot of source-enriched observations. 64 ISGs were selected. The underlying sources are the expression profiles of CNS and liver cell lines. (b) Overlaid scatter plots of recovered interim (1st iteration) and true sources. (c) Interim ISG selection based on the scatter plot of recovered interim sources. 324 ISGs were selected. (d) Overlaid scatter plots of recovered interim (5th iteration) and true sources.

Since the true mixing matrix is generally unknown or inaccurate, we use the correlation coefficients between the profiles of ensemble subISG as a more direct measure of agreement to assess the performance of our algorithm. The correlation coefficient is calculated over the correct ensemble ISG indices, and given by

$$\rho(\hat{\mathbf{s}},\mathbf{s}) = \frac{\sum_{i\in \text{subISG}} \left(\hat{\mathbf{s}}(i) - \bar{\hat{\mathbf{s}}}\right)\left(\mathbf{s}(i) - \bar{\mathbf{s}}\right)}{\sqrt{\sum_{i\in \text{subISG}} \left(\hat{\mathbf{s}}(i) - \bar{\hat{\mathbf{s}}}\right)^2 \sum_{i\in \text{subISG}} \left(\mathbf{s}(i) - \bar{\mathbf{s}}\right)^2}}, \tag{2.14}$$

where $\hat{\mathbf{s}}$ and $\bar{\hat{\mathbf{s}}}$ are our source estimation and its mean, respectively. The correlation coefficients between the true and estimated sources for CNS and liver within this ISG subset are 0.9934 and 0.9960, respectively, while the true $\mathbf{A}$ and estimated mixing matrices $\hat{\mathbf{A}}$ are given below:

$$\mathbf{A} = \begin{bmatrix} 0.25 & 0.75 \\ 0.75 & 0.25 \end{bmatrix}, \ \hat{\mathbf{A}} = \begin{bmatrix} 0.2060 & 0.7940 \\ 0.5746 & 0.4254 \end{bmatrix}.$$

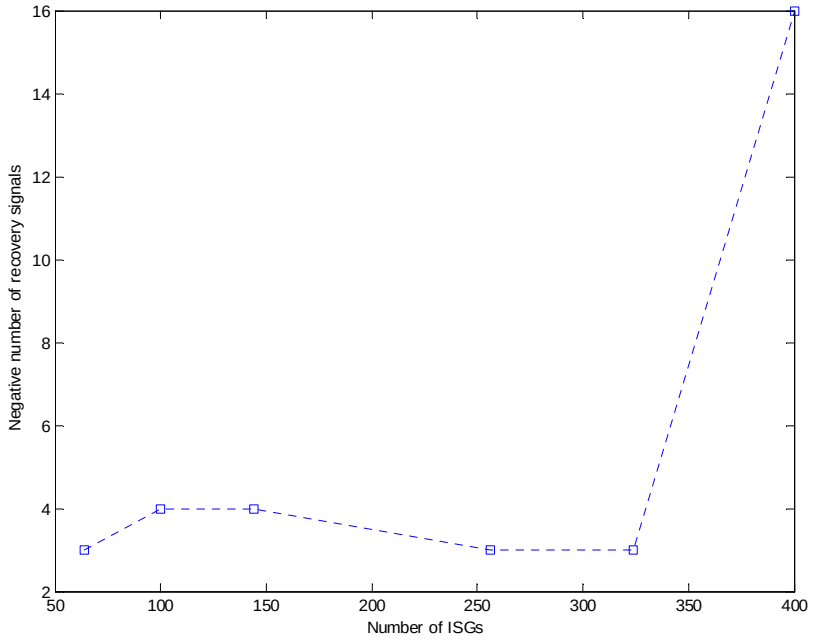


Figure 2.3: Iterative ISG selection procedures based on non-negativity constraint

We then explored the capability of nPICA to separate real composite profiles of MCF-7 breast cancer cell and Hs27 human diploid fibroblasts cell. We mixed samples and hybridized to DNA microarray to simulate the real situation with estrogen receptor positive (ER+) MCF-7 cells mixed with normal fibroblast cells (Figure 2.4(a)). The figure shows a successful decomposition of the composite profiles of MCF-7 and Hs27 mixtures, indicated by the close match between the scatter plots of recovered and true source profiles (Figure 2.4(d)). The correlation coefficients between the true and

estimated sources for MCF-7 and Hs27 are 0.9942 and 0.9823, respectively, while the true **A** and estimated mixing matrices $\hat{\mathbf{A}}$ are provided as follows:

$$\mathbf{A} = \begin{bmatrix} 0.25 & 0.75 \\ 0.75 & 0.25 \end{bmatrix}, \ \hat{\mathbf{A}} = \begin{bmatrix} 0.3270 & 0.6730 \\ 0.8967 & 0.1033 \end{bmatrix}.$$

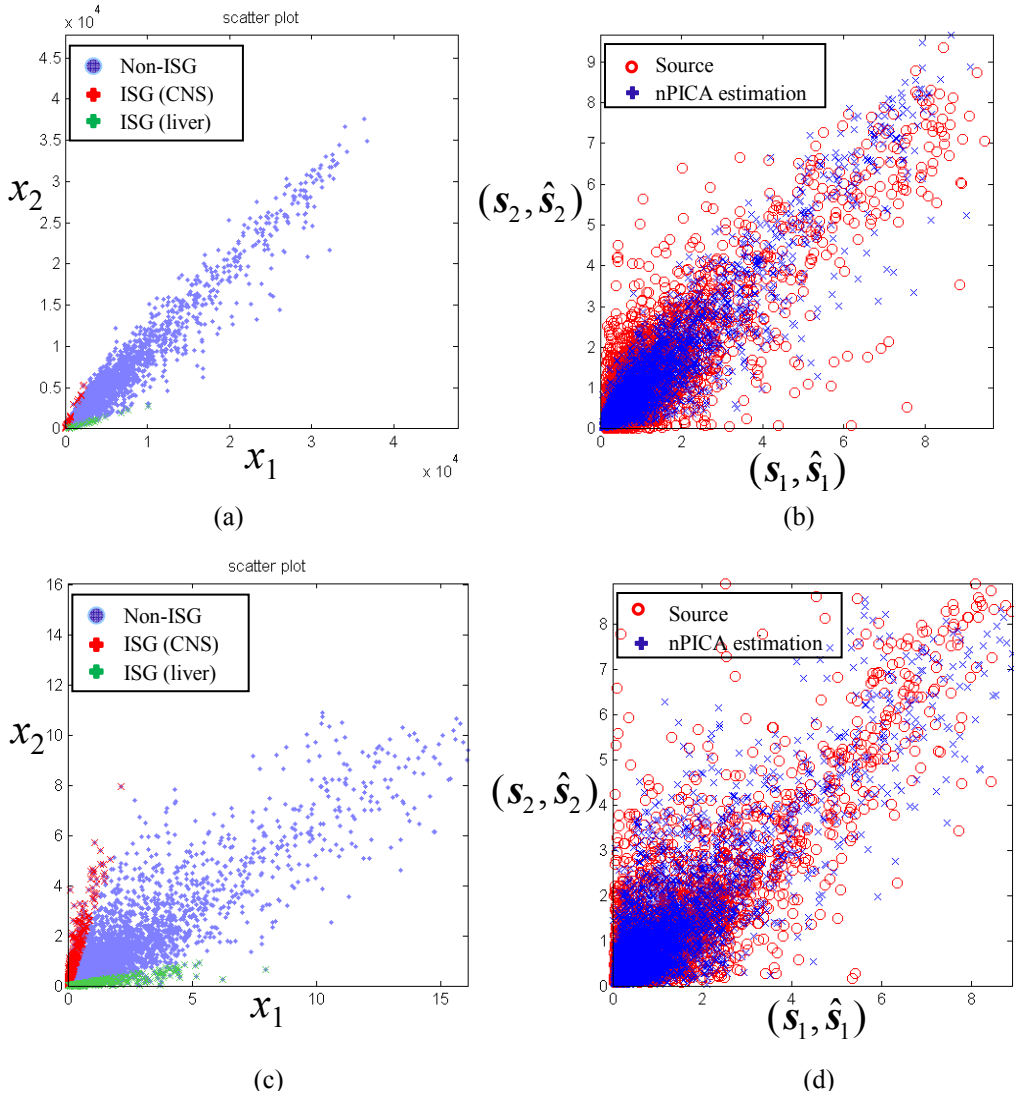


Figure 2.4: (a) Initial ISG selection based on the scatter plot of source-enriched observations. 64 ISGs were selected. The underlying sources are the expression profiles of MCF-7 and Hs27 cell lines. (b) Overlaid scatter plots of recovered interim (1st iteration) and true sources. (c) Interim ISG selection based on the scatter plot of recovered interim sources. 784 ISGs were selected. (d) Overlaid scatter plots of recovered interim (10th iteration) and true sources.

Finally, we explored the capability of nPICA to separate more complex composite profiles of MCF-7 and MDA-MB-231 breast cancer cells, and Hs27 (fibroblasts) cells. There are several phenotypic differences between these cell lines; MDA-MB-231 cells are metastatic, highly invasive and estrogen receptor negative (ER-); MCF-7 cells are ER+. Gene expression data were mixed *in silico*. Figure 2.5 shows a successful decomposition of the composite profiles of MCF-7, MDA-MB-231 and Hs27 mixtures, indicated by the side-by-side overlapping between 3D scatter plots of the recovered and true source profiles. The correlation coefficients between the true and estimated sources for MCF-7, MDA-MB-231 and Hs27 are 0.9822, 0.9983 and 0.9917, respectively.

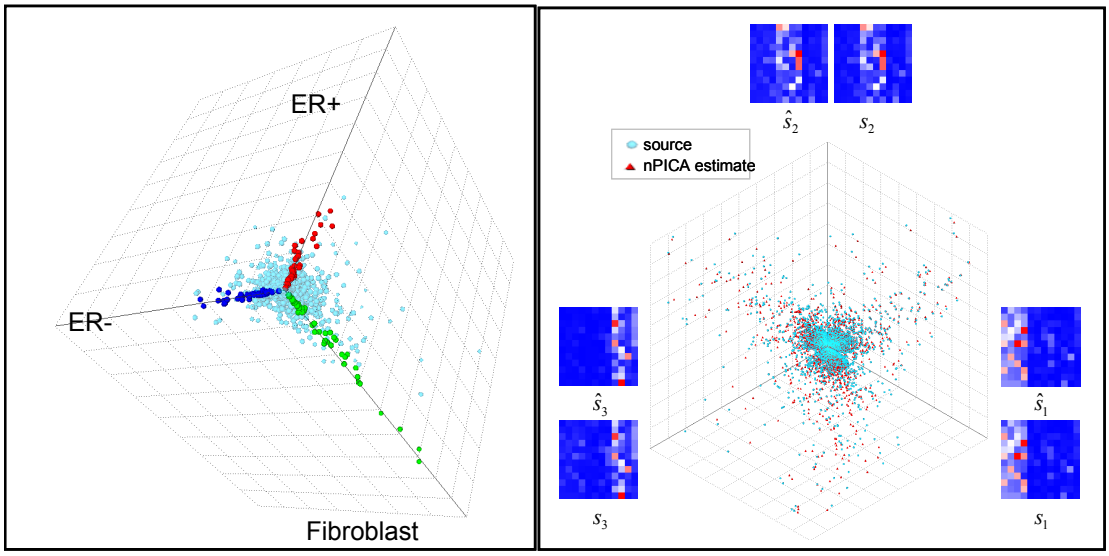


Figure 2.5: Results of tissue heterogeneity correction for ER+, ER- and fibroblast by nPICA. (Left panel): subISGs highlighted in the 3D source plot. (Right panel): The scatter plot of overlaid true sources and decomposed profiles are generated from selected ISGs. Heatmaps show the subISGs patterns in original sources and estimations.

In all these three experiments we applied the proposed nPICA model, defined over the subset of the selected ensemble ISG indices, to estimate first the de-mixing matrix $\mathbf{W}$ and independent source segments, and subsequently recover the entire source profiles $\hat{\mathbf{s}}(i)$ over all genes. The results demonstrated that the selected ensemble subISGs is sufficient to assure the accurate recovery of source profiles by nPICA. Next, we will further develop an unsupervised mode of nPICA for tissue heterogeneity correction.

2.4 Unsupervised Nonnegative Partially-Independent Component Analysis

2.4.1 Geometric Principles of the Problem

In order to establish the theoretical basis of the method, we will consider the following assumption besides the positivity constraint on **S**: for each source, there is at least one value of the acquisition variable for which this source is uniquely present, to the exclusion of all other sources. Such sources are said to be partially uncorrelated (or partially orthogonal) since their subparts exist that are uncorrelated. More formally, the source matrix **S** is assumed to satisfy the following condition:

Assumption 1 (Strictly Well-Grounded points): For each $i \in \{1,2,\cdots,d\}$ there exist an $j_i \in \{1,2,\cdots,N\}$ such that $s_{i,j_i} > 0$ and $s_{k,j_i} = 0$ $(k = 1,\cdots,i-1,i+1,\cdots,d)$.

Clearly, this assumption does not require fully orthogonal source signals, since orthogonality is required only for the subspectra defined by the subscripts $j_i,\ (i = 1,2,\cdots,d)$. Note that $j_i,\ (i = 1,2,\cdots,d)$ are not known and need to be computed.

It should not escape of our notice that we can equivalently view the well-grounded points (WGPs) of **S** as the intersection of *d* hyperplanes. We will use $\mathbb{C}_{\mathbf{s}}$ to denote those points. Recall that a hyperplane in $\mathbf{R}^d$ is defined as [76]

$$H(\boldsymbol{\alpha},\beta) \equiv \left(\mathbf{x} \in \mathbf{R}^d : \begin{matrix} \boldsymbol{\alpha}^T\mathbf{x} = \beta & \boldsymbol{\alpha} \in \mathbf{R}^d \\ \boldsymbol{\alpha} \neq \mathbf{0} & \beta \in \mathbf{R} \end{matrix} \right) . \tag{2.15}$$

The vector $\boldsymbol{\alpha}$ is commonly referred to as the normal vector. Letting $\mathbf{e}_i$ denote the *i*-th unit vector, we have that

$$\mathbb{C}_{\mathbf{s}} = \bigcap_{i=1}^{d} \{\mathbf{s} : \mathbf{e}_i^T\mathbf{s} = 0\} \ . \tag{2.16}$$

In other words, the points in the source signal space are just enclosed by the hyperplanes $\{H(\mathbf{e}_i, 0),\ i = 1,\cdots,d\}$.

Based on this perception, for nonsingular **A**, we have several important properties relating hyperplanes to linear transformations as follows [72].

Property 1: A nonsingular $d \times d$ matrix $\mathbf{A}$ maps a hyperplane into another hyperplane *via*

$$H(\boldsymbol{\alpha}, \beta) \xrightarrow{\mathbf{A}} H(\mathbf{A}^{-T}\boldsymbol{\alpha}, \beta). \tag{2.17}$$

Proof: Suppose $\mathbf{x} \in H(\boldsymbol{\alpha}, \beta)$ and $\mathbf{y} = \mathbf{A}\mathbf{x}$. Then

$$(\boldsymbol{\alpha}^T \mathbf{A}^{-1})\mathbf{y} = (\boldsymbol{\alpha}^T \mathbf{A}^{-1})\mathbf{A}\mathbf{x} = \boldsymbol{\alpha}^T \mathbf{x} = \beta . \qquad \square$$

Corollary 1: Let $\mathbb{C}_{\mathbf{s}}$ be as in (2.16), $\mathbb{C}_{\mathbf{x}} = \mathbf{A}\mathbb{C}_{\mathbf{s}}$, and $\mathbf{A}$ a nonsingular $d \times d$ matrix, then

$$\mathbb{C}_{\mathbf{x}} = \bigcap_{i=1}^{d} H(A^{-T}\mathbf{e}_i, 0) \quad , \tag{2.18}$$

so that $\mathbb{C}_{\mathbf{x}}$ defines a set of hyperplanes that enclose the mixtures and the transpose of normal vectors, *i.e.*, $\mathbf{e}_i^T \mathbf{A}^{-1}$ are the rows of $\mathbf{A}^{-1}$ (up to a scale factor). Thus, we can think of the columns of $\mathbf{X}$ as points in $\mathbf{R}^d$ defining a *d*-pyramid within the intersection of such hyperplanes. The matrix $\mathbf{A}$ then defines an invertible transformation from the first quadrant occupied by source $\mathbf{S}$ to the *d*-pyramid defined by mixture $\mathbf{X}$. Thus, an equivalent problem is to find the $d \times d$ matrix $\mathbf{A}^{-1}$ that maps the *d*-pyramid into the first quadrant (see Figure 2.6 for an illustration).

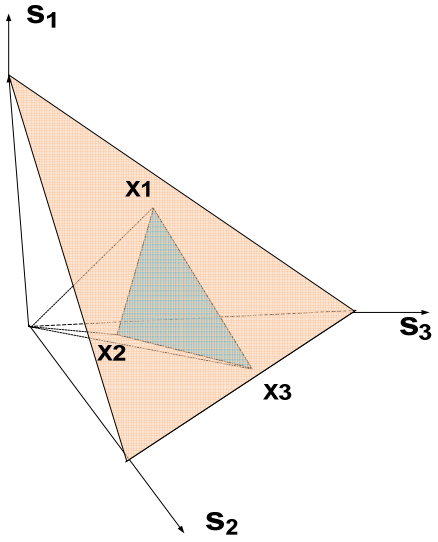


Figure 2.6: Illustration of the sources scatter plot **S** occupying in the first quadrant and the mixtures scatter plot **X** which is confined within a convex pyramid within the first quadrant. (d = 3 in this case)

Property 2: The lateral edges of *d*-pyramid formed by the observation scatter plot **X** are collinear to *d* column vectors of **A**.

Proof: Combining Eq. (2.3) and Eq. (2.5), we have:

$$\mathbf{x}(j) = \sum_{k=1}^{d} s_k(j)\mathbf{a}(k) \quad (j = 1, \cdots, N). \tag{2.19}$$

For the particular subscripts $j_i, (i = 1, 2, \cdots, d)$ described in Assumption 1, Eq. (2.19) becomes

$$\mathbf{x}(j_i) = s_i(j_i)\mathbf{a}(i) \quad (i = 1, \cdots, d), \tag{2.20}$$

since by Assumption 1, $s_k(j_i)$ is nonzero only if $k = i$. Then

$$\mathbf{a}(i) = \mathbf{x}(j_i) / s_i(j_i) \quad (i = 1, \cdots, d). \tag{2.21}$$

Using Eq. (2.21) to replace $\mathbf{a}(k)$ in Eq. (2.19), we can rewrite Eq. (2.19) as follows:

$$\mathbf{x}(j) = \sum_{k=1}^{d} \mathbf{x}(j_i) \frac{s_k(j)}{s_i(j_i)} \quad (j = 1, \cdots, N) \ . \qquad \square$$

From Property 2, we know that the lateral edges of *d*-pyramid formed by the observation scatter plot **X** are $\mathbf{x}(j_i), (i = 1, 2, \cdots, d)$. Every column of **A** is collinear to at least one column of **X**, so that theoretically we can estimate a column of **A** through observations as well. However, problems arise when the data are corrupted by experimental noise, e.g., the original/true edges of the *d*-pyramid of observations are blurred. This provides motivation to use an invariantly-expressed genes removal approach to select ISG subset, which will be discussed in the next section.

2.4.2 Complementary of ISG Subset – Invariantly-Expressed Genes (IEGs)

Scatter plot of observations and ISG concept suggest a new biologically plausible blind source separation mechanism for tissue heterogeneity correction: if the gene indices $\mathbb{S}_{\text{subISG}}$ of ISG subset or its counterpart ($\mathbb{S}_{\text{subIEG}} = \overline{\mathbb{S}}_{\text{subISG}}$), which is based on *normalized* observation **X** , can be identified, we could, in principle, estimate the de-mixing matrix **W** and source profiles **S** by performing nonnegative independent component analysis

(nICA) based on $\mathbf{X}(\mathbb{S}_{\text{subISG}})$. In this sense, we do not need those microdissected samples corresponding to phenotype *j* to help identify $\mathbb{S}_{\text{subISG}}$ any more, which we call unsupervised identification of ISGs. To apply this design, we first project **X** onto the standard simplex; the standard *n*-simplex (or unit *n*-simplex) (Figure 2.7(a)) is the subset of $\mathbf{R}^{n+1}$ given by

$$\Delta^n = \{(t_0, \cdots, t_n) \in \mathbf{R}^{n+1} \,\big|\, \textstyle\sum_i t_i = 1 \text{ and } t_i \geq 0 \text{ for all } i\}. \tag{2.22}$$

Thus, the perspective projection of data onto the standard simplex takes the general form of the following [77]:

$$\textstyle\sum_{j=1}^{d} x_j(i) = 1 \qquad (i = 1, \ldots, N), \tag{2.23}$$

which can be achieved by performing a simple cross-sample sum-based standardization for each sample *j*

$$x_j(i) = \frac{x_j^{org}(i)}{\sum_{k=1}^{d} x_k^{org}(i)} \qquad (i = 1, \ldots, N), \tag{2.24}$$

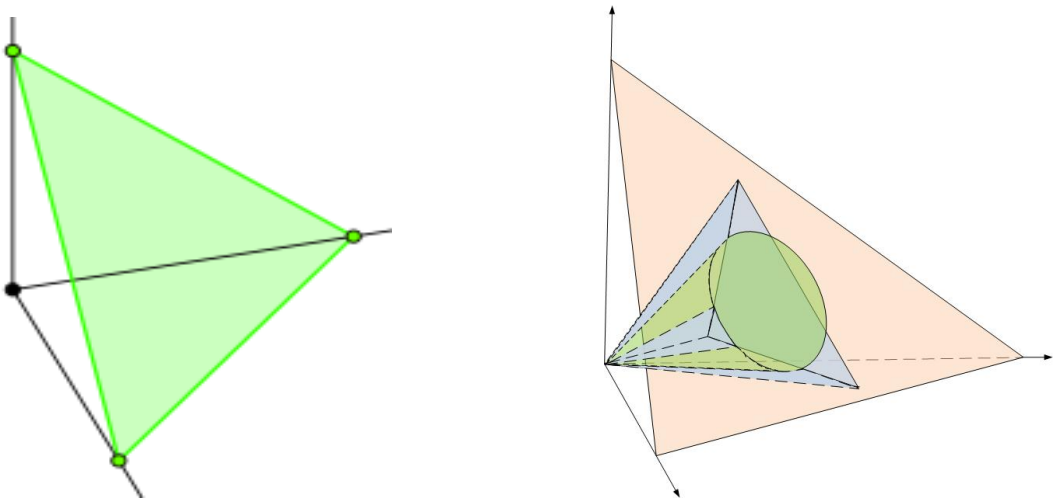
where $x_j^{org}(i)$ denotes the original value before projection.

Finally, the indices $\mathbb{S}_{\text{subISG}}$ are identified on the simplex hyperplane by using an IEG-removal procedure, so that we can exploit the nPICA algorithm based on the remaining genes. The overall scheme of the strategy is illustrated in Figure 2.7.

2.4.3 Experimental Design and Results

To ascertain that the IEG removal method can identify intrinsic phenotypic-up-regulated independence-support gene (ISG) subsets, we first evaluated the performance of our algorithm by a series of proof-of-concept experiments. We measured the accuracy of the method with several simulation experiments where known proportions of mRNA derived from differential phenotype cell lines were mixed *in silico*. Then we compared our approach with other similar algorithm(s) on noisy data to study the noise impact on the algorithm. Finally, experiments for tissue heterogeneity correction (THC) are

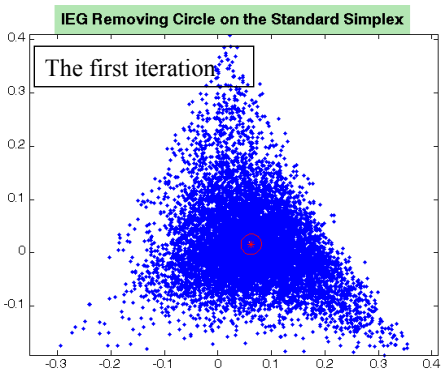
presented with application to real microarray data acquired from tumors, which demonstrates an improved predictive accuracy after THC.



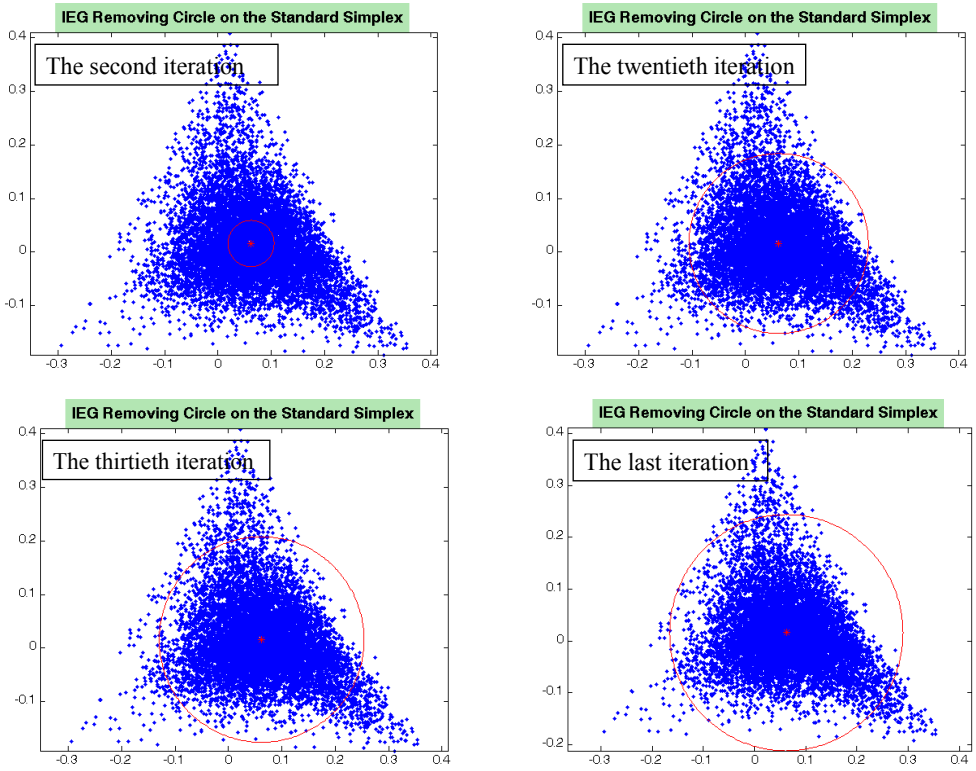
(a) (b)

Figure 2.7: (a) The standard 2-simplex in R^3 (Public domain image from Wikipedia: http://en.wikipedia.org/wiki/File:2D-simplex.svg); (b) An illustration of IEG removal scheme in the mixtures scatter plot **X** that is confined within a convex pyramid within the first quadrant. Different colors are used to depict different parts of genes, blue: $\mathbb{S}_{\text{subISG}}$, green: $\mathbb{S}_{\text{subIEG}}$. Through perspective projection, we first project all genes on the standard simplex, and then the indices $\mathbb{S}_{\text{subISG}}$ are identified on the simplex hyperplane by using an IEG removal procedure.

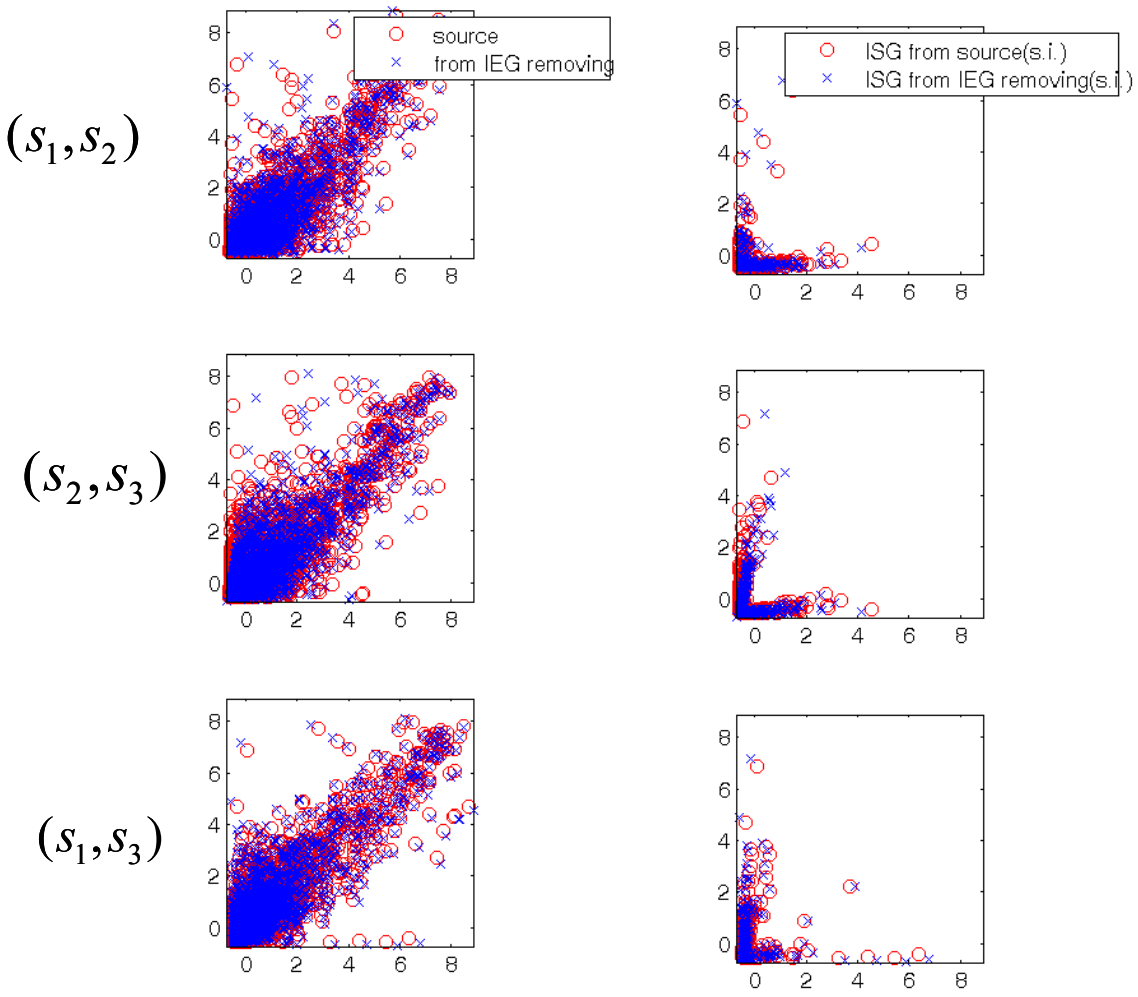
In particular, we applied the proposed THC method to real breast cancer cell line data. The three real microarray gene expression data sets consist of MCF-7, A1N4 and Hs27 cell lines. Figure 2.8 shows the following: (1) the real data distribution on the standard simplex after perspective projection; (2) the IEG removal procedure for ISG selection; and (3) superimposed pair-wise scatter plots of true and estimated source profiles, respectively. Figure 2.9 shows an example of the IEG removal procedure that identifies IEGs (within the cone) in 3-D scatter plots iteratively.


IEG Removing Circle on the Standard Simplex
The first iteration

(a)


IEG Removing Circle on the Standard Simplex
The second iteration
IEG Removing Circle on the Standard Simplex
The twentieth iteration
IEG Removing Circle on the Standard Simplex
The thirtieth iteration
IEG Removing Circle on the Standard Simplex
The last iteration

(b)



(c)

Figure 2.8: An IEG removal procedure on three-source real gene expression mixtures (MCF-7/A1N4/Hs27). (a) The real data distribution on the standard simplex after perspective projection; (b) iterative IEG removal procedure on the standard simplex. Within the red circle, the genes are categorized as IEGs; (c) superimposed pair-wise scatter plots of true and estimated source profiles. (Left panel: the entire gene space; right panel: ISGs only.) Red circles indicate the true sources. Blue crosses indicate the estimations.

2.4.3.1 ISG selection by IEG-removal in comparison with ISG selection from mixtures

We performed ISG selection based on IEG removal to select approximately equal number of ISGs per corners on the mixtures' simplex. Here, we compared our ISG selection scheme with the selection of ISGs directly from mixtures as a proof of concept.

Using various mixing matrices, we mixed the true sources to generate a number of mixtures with different proportions of sources. We identified ISGs through the IEG removal procedure or selected ISGs directly from mixtures by fold change. Then we compared the performance of blind source separation results from those two methods with the ground truth. The detailed comparison experiments are described as follows.

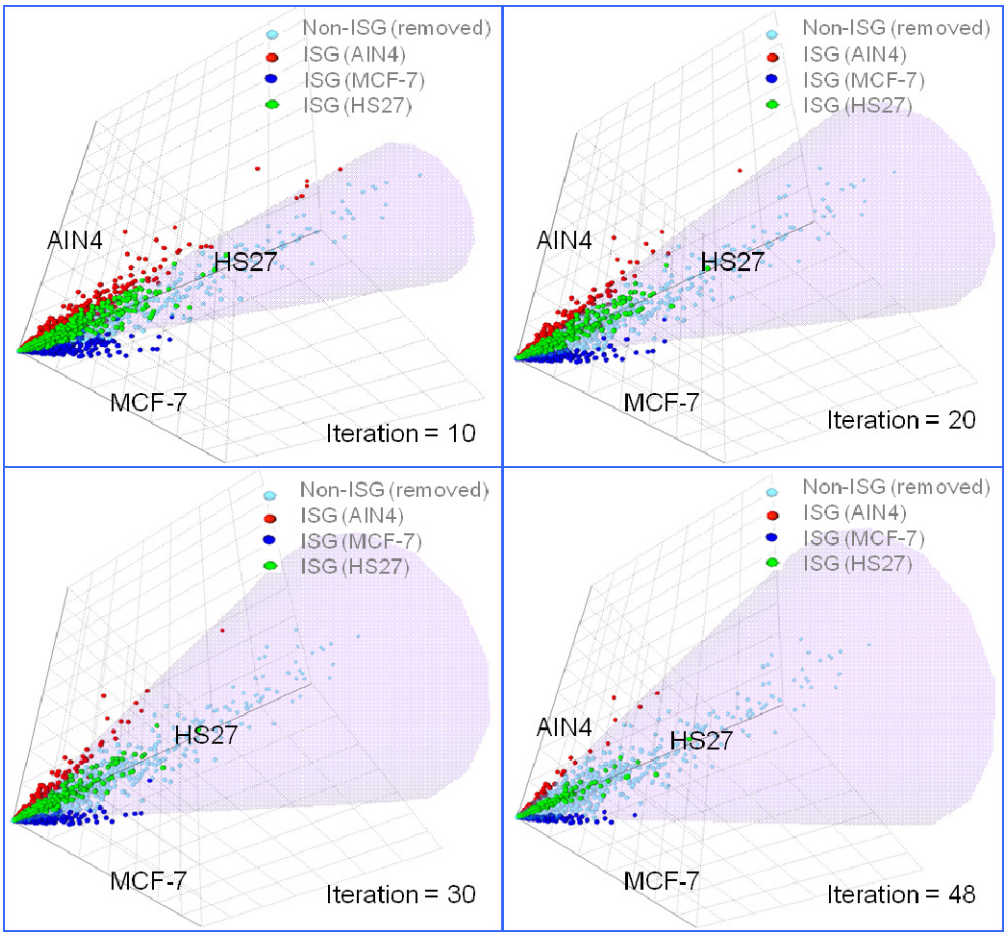


Figure 2.9: Iteratively identified $\mathbb{S}_{\text{subIEG}}$ (within the cone) on 3d scatter plot. The $\mathbb{S}_{\text{subISG}}$ is shown in different colors to depict different phenotypes.

Recall that the mixing matrix **A** defines an invertible transformation to map source **S** to a *d*-pyramid defined by mixture **X,** the mixing procedure is the rotation and bi-folding of the sources. As is well known, every rotation in three dimensions has an axis — a direction that is fixed by the rotation. Given a rotation matrix **R**, a vector **u** parallel to the rotation axis must satisfy

$$R\mathbf{u} = \mathbf{u}. \tag{2.25}$$

With a unit vector $\mathbf{u} = (u_x, u_y, u_z)$, where $u_x^2 + u_y^2 + u_z^2 = 1$, the matrix for a rotation by an angle of θ about an axis of the direction **u** is [78]:

$$R = \begin{bmatrix} u_x^2 + (1-u_x^2)c & u_x u_y(1-c) - u_z s & u_x u_z(1-c) + u_y s \\ u_x u_y(1-c) + u_z s & u_y^2 + (1-u_y^2)c & u_y u_z(1-c) - u_x s \\ u_x u_z(1-c) - u_y s & u_y u_z(1-c) + u_x s & u_z^2 + (1-u_z^2)c \end{bmatrix}, \tag{2.26}$$

where $c = \cos\theta, \quad s = \sin\theta.$

To validate the principle of IEG removal-based nPICA approach, we first evaluated the performance of IEG removal in a noise-free environment based on its ability to separate numerically mixed microarray gene expression data. The performance evaluation was conducted in three cases, *i.e.*, we generated the rotation matrix **R** around the axis of central line of the first quadrant with a rotation angle $\theta = \pi/6$, $\pi/4$ and $\pi/3$.

First, we used the following mixing matrix to mix three sources to make the sources bi-fold and rotate around $\theta = \pi/6$:

$$\mathbf{A} = \begin{bmatrix} 0.6271 & 0.0603 & 0.3481 \\ 0.3256 & 0.6250 & 0.0471 \\ 0.0267 & 0.3252 & 0.6137 \end{bmatrix}.$$

We projected the three sources and mixtures on the standard 2-simplex in the 3-D scatter plot shown in Figure 2.10 (a).

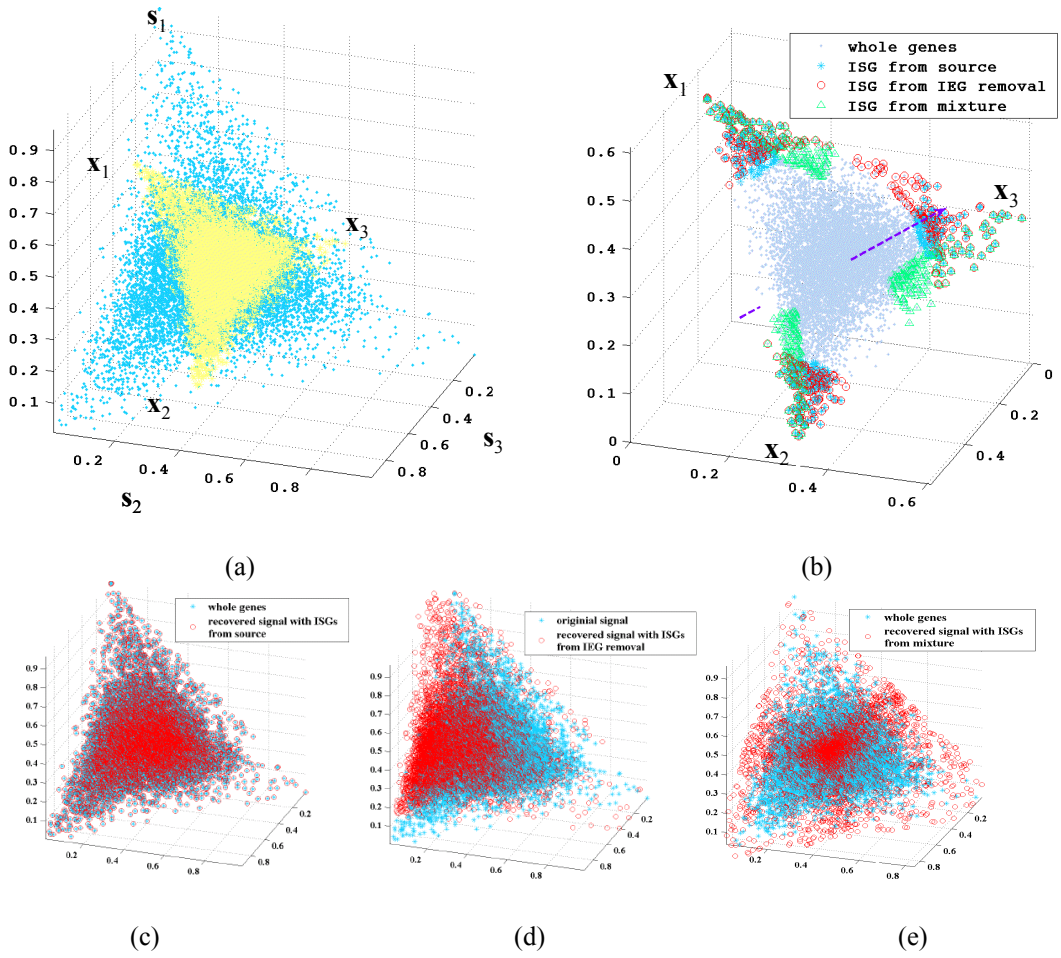


Figure 2.10: Experiments with the rotation angle $\theta = \pi/6$**.** (a) The superimposed 3-D scatter plot of true sources *vs.* the mixtures projected on the standard simplex. Blue dots: the projection of sources on 2-simplex; yellow dots: the projection of mixtures on 2-simplex. (b) Comparison of the contents of ISGs on the projection of 2-simplex of mixtures. Grey dots: all the genes in the mixed cell lines; Blue stars: true ISGs selected from sources using supervised mode; Red circles: ISGs selected using IEG-removal approach with unsupervised mode; Green triangles: ISGs selected from mixtures directly by *one-vs-each* fold change; Purple arrow: the rotation axis of the experiment. (c) (d) (e) are overlaid projections on the standard simplex between the true sources and the recovered signals. Blue stars: original signals; red circles: recovered signals. (c): Source estimation using ISGs identified from sources (number of ISGs = 484); (d): Source estimation using ISGs identified from IEG removal (number of ISGs = 480); (e): Source estimation using ISGs identified from mixtures (number of ISGs = 484).

Inasmuch as we already had the ground truth, we first identified the ISGs through *one-vs-each* extension of differentially expressed genes in the sources using supervised

mode discussed in Section 2.3 and selected 484 genes (161/161/162 for three different sources respectively) to define a valid nPICA model. The estimated mixing matrix is

$$\hat{\mathbf{A}} = \begin{bmatrix} 0.5039 & 0.1374 & 0.3587 \\ 0.2521 & 0.6792 & 0.0687 \\ 0.0111 & 0.4178 & 0.5711 \end{bmatrix}.$$

The correlation coefficients of the three estimations with regard to the ground-truth sources (ISGs only) are: $(\mathbf{s}_1, \hat{\mathbf{s}}_1)$: 0.99655; $(\mathbf{s}_2, \hat{\mathbf{s}}_2)$: 0.99725; $(\mathbf{s}_3, \hat{\mathbf{s}}_3)$: 0.99734. We assumed that the set of genes selected above is our **true ISGs**.

Then we performed IEG removal to identify ISGs or selected ISGs directly from mixtures through *one-vs-each* extension of differentially expressed genes with the (almost) same number of ISGs. The projection of true ISGs, ISGs from IEG removal and ISGs from mixtures on the standard 2-simplex of observations are presented in Figure 2.10 (b).

Besides correlation coefficients between true and estimated sources we introduced another measure to evaluate the statistical performance or accuracy of an algorithm, that is, the performance index defined as [68]:

$$E_1 = \sum_{i=1}^{d}\left(\sum_{j=1}^{d}\frac{|p_{ij}|}{\max_k |p_{ik}|} - 1\right) + \sum_{j=1}^{d}\left(\sum_{i=1}^{d}\frac{|p_{ij}|}{\max_k |p_{kj}|} - 1\right), \tag{2.27}$$

where p_{ij} is the ijth element of matrix $\mathbf{P} = \mathbf{WA}$, $\mathbf{W}$ is the estimated de-mixing matrix and $\mathbf{A}$ is the mixing matrix. The E_1 measure is invariant to the permutation and scaling, thus it measures how close matrix $\mathbf{W}$ is to the true de-mixing matrix $\mathbf{A}^{-1}$. E_1 is always nonnegative and the smaller the value of E_1 is, the better performance of the algorithm is.

We then compared true and estimated sources for MCF-7, A1N4 and Hs27 within their respective empirical ISG subsets identified by different methods using correlation coefficients as the first criterion. Due to the fact that the ISG subset identified from sources is the most accurate ISGs that we could find, we treated them as the ground truth. And further, we also compared correlation coefficients between true and recovered signals based on true ISGs. E1 is then used as the second criterion. The third criterion that is utilized for comparison includes the number and percentage of overlapped ISGs

between true ISGs and identified empirical ISGs. The detailed results are listed in Table 2.1.

Table 2.1 Comparison results of IEG-removal with ISGs selection from mixtures generated by the mixing matrix with rotation angle $\theta = \pi/6$, $\theta = \pi/4$ and $\theta = \pi/3$.

Rotation Angle $\theta = \pi/6$		
	IEG-Removal	ISGs from mixtures
Correlation coefficients based on empirical ISGs: $(\mathbf{s}_i, \hat{\mathbf{s}}_i)\ i = 1,2,3$	0.9986/0.9921/0.9799	0.9396/0.9852/0.9939
Correlation coefficients based on the true ISGs set: $(\mathbf{s}_i, \hat{\mathbf{s}}_i)\ i = 1,2,3$	0.9990/0.9922/0.9806	0.9303/0.9862/0.9952
E_1	1.327	1.3773
Number of ISGs for each mixtures	158/158/164 = 480	161/161/162 = 484
Number and Percentage of overlapped ISGs with the true ISGs	402 genes are also selected in the true ISGs set; 402/480 = 83.75%	180 genes are also selected in the true ISGs set; 180/484 = 37.19%
Rotation Angle $\theta = \pi/4$		
Correlation coefficients based on empirical ISGs: $(\mathbf{s}_i, \hat{\mathbf{s}}_i)\ i = 1,2,3$	0.9988/0.9941/0.9825	0.9055/0.9579/0.9940
Correlation coefficients based on the true ISGs set: $(\mathbf{s}_i, \hat{\mathbf{s}}_i)\ i = 1,2,3$	0.9991/0.9939/0.9836	0.8689/0.9679/0.9956
E_1	1.1636	1.7486
Number of ISGs for each mixtures	135/135/136 = 406	133/133/134 = 400
Number and Percentage of overlapped ISGs with the true ISGs	353 genes are also selected in the true ISGs set; 353/406= 86.95%	15 genes are also selected in the true ISGs set; 15/400 = 3.75%
Rotation Angle $\theta = \pi/3$		
Correlation coefficients based on empirical ISGs: $(\mathbf{s}_i, \hat{\mathbf{s}}_i)\ i = 1,2,3$	0.9988/0.9961/0.9883	0.9178/0.9408/0.9468
Correlation coefficients based on the true ISGs set: $(\mathbf{s}_i, \hat{\mathbf{s}}_i)\ i = 1,2,3$	0.9990/0.9957/0.9888	0.9349/0.9556/0.6777
E_1	1.0211	2.395
Number of ISGs for each mixtures	131/131/132 = 396	133/133/134 = 400

Number and Percentage of overlapped ISGs with the true ISGs	358 genes are also selected in the true ISGs set; 358/396= 90.40%	No genes are selected in the true ISGs set; 0/400 = 0%

Overlaid projections on the standard simplex between true sources and recovered signals (rotation angle $\theta = \pi/6$) using different approaches are presented in Figure 2.10 (c), (d) and (e).

Next, we used the following mixing matrix to mix three sources with the rotation around the central line of the first quadrant of $\theta = \pi/4$:

$$\mathbf{A} = \begin{bmatrix} 0.5753 & 0.0341 & 0.4388 \\ 0.4108 & 0.5771 & 0.0162 \\ 0.0164 & 0.4072 & 0.5543 \end{bmatrix}.$$

We still started with an identification of ISGs through *one-vs-each* extension of differentially expressed genes in the sources and selected 400 (133/133/134 for three sources) genes to define a valid nPICA model. The estimated mixing matrix is:

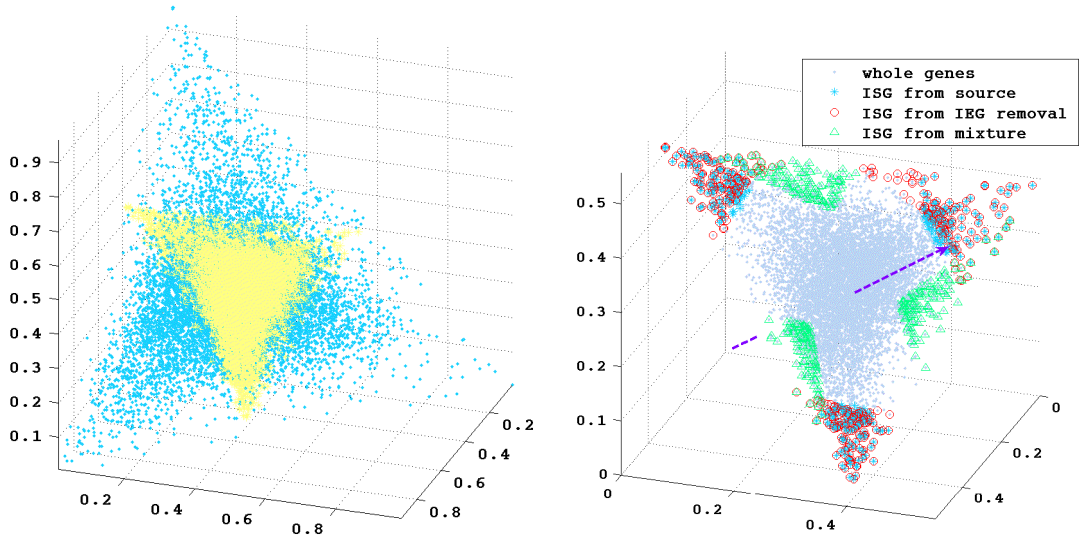
$$\hat{\mathbf{A}} = \begin{bmatrix} 0.4467 & 0.0854 & 0.4678 \\ 0.3119 & 0.6593 & 0.0287 \\ -0.0025 & 0.4952 & 0.5073 \end{bmatrix}.$$

The correlation coefficients of three estimations with regard to ground sources (ISG only) are: $(\mathbf{s}_1, \hat{\mathbf{s}}_1)$: 0.9979; $(\mathbf{s}_2, \hat{\mathbf{s}}_2)$: 0.9983; $(\mathbf{s}_3, \hat{\mathbf{s}}_3)$: 0.9964. We assume that these genes are our **true ISGs**.

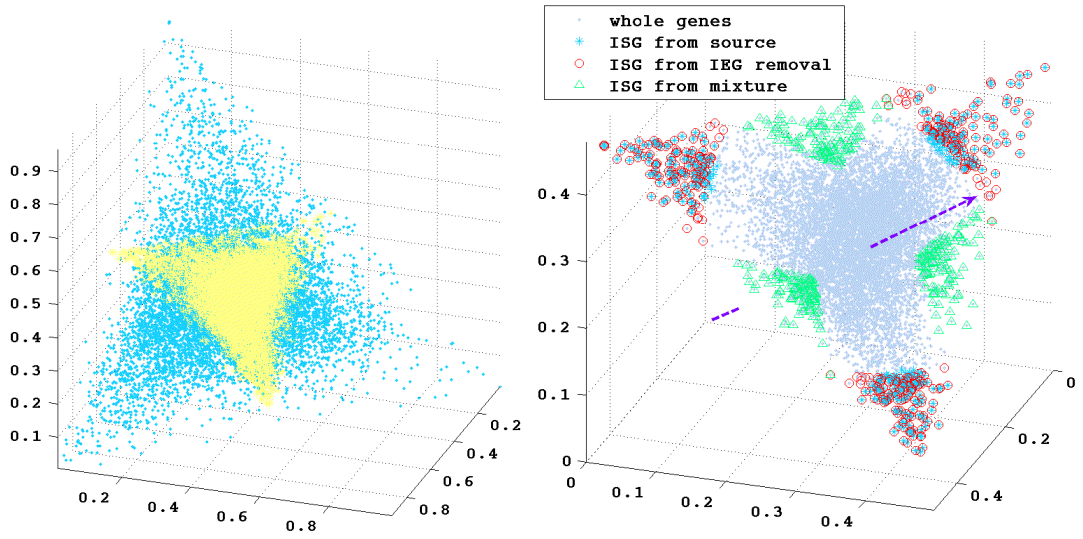
The superimposed projection of three sources and mixtures on the standard 2-simplex in 3-D scatter plot is shown in Figure 2.11(a) (left panel). The comparison of ISG contents on the projection of mixtures on the standard 2-simplex in 3-D scatter plot is presented in Figure 2.11 (a) (right panel). The detailed decomposition results are listed in Table 2.1. Overlaid projections on the standard simplex between true sources and recovered signals (rotation angle $\theta = \pi/4$) are presented in Appendix B (Figure B.1).

Finally, we applied the following mixing matrix to mix three sources with the rotation around the central line of the first quadrant with $\theta = \pi/3$:

$$\mathbf{A} = \begin{bmatrix} 0.4843 & 0.0286 & 0.5144 \\ 0.4770 & 0.5219 & 0.0079 \\ 0.0372 & 0.5056 & 0.4930 \end{bmatrix}.$$



(a)



(b)

Figure 2.11: Comparison of IEG removal with ISGs selection from mixtures using the mixing rotation angle $\theta = \pi/4$ (a) and $\theta = \pi/3$ (b). (Left column) The superimposed 3-D scatter plots of true sources *vs.* the mixtures projected on the standard simplex. Blue dots: the projection of sources on 2-simplex; yellow dots: the projection of mixtures on 2-simplex. (Right column) ISGs projected on the 2-simplex of mixtures. Grey dots: all the genes in the cell lines; Blue stars: true ISGs selected from sources using supervised mode; Red circles: ISGs selected from IEG removal approach using unsupervised mode; Green triangles: ISGs

selected from mixtures directly by *one-vs-each* fold change; Purple arrows: rotation axes of the experiments.

Similar to previous experiments, we defined **true ISGs** first through *one-vs-each* extension of differentially expressed genes in the sources. In this case, 400 (133/133/134 for three different sources) genes were identified to define a valid nPICA model. The estimated mixing matrix is:

$$\hat{\mathbf{A}} = \begin{bmatrix} 0.3699 & 0.0758 & 0.5544 \\ 0.3625 & 0.6154 & 0.0221 \\ 0.0218 & 0.5615 & 0.4166 \end{bmatrix},$$

The correlation coefficients of estimations with regard to ground sources (ISG only) are: $(\mathbf{s}_1, \hat{\mathbf{s}}_1)$: 0.9979; $(\mathbf{s}_2, \hat{\mathbf{s}}_2)$: 0.9983; $(\mathbf{s}_3, \hat{\mathbf{s}}_3)$: 0.9965. We assume that the set of genes consists of our **true ISGs**.

Besides demonstrating the overlying of true sources with mixtures on the standard simplex in 3-D plot (the left column of Figure 2.11(b)), we also highlighted the true ISGs, ISGs from IEG removal and ISGs from mixtures on the standard 2-simplex of the observations in the right column of Figure 2.11(b). The detailed results are listed in Table 2.1. Overlaid projections on the standard simplex between true and recovered sources (rotation angle $\theta = \pi/3$) are presented in Appendix B (Figure B.2).

We have evaluated the performance of the proposed method using three different criteria: correlation coefficients; E_1; and number and percentage of the overlapped ISGs with true ISGs. From all those results, we concluded that by removing invariantly-expressed genes (IEGs), we can identify the true independence-support genes (ISGs) correctly. The experiments were repeated for different rotation angles during the mixing procedures by different mixing matrices **A**. Similar results can be obtained by IEG removal, while selecting ISGs directly from mixture by *one-vs-each* fold change can only succeed when the rotation angle is less than $\pi/4$. This provides experimental evidence to support that ISG identification based on IEG removal is robust and accurate.

2.4.3.2 Comparison of IEG removal based nPICA with similar algorithm(s)

As mentioned previously, it is possible to separate a set of mixtures into component spectra using only the measurement of the mixtures' spectra provided that available number of linearly independent mixtures is equal or greater than the number of components [68]. This problem is generally known as blind source separation (BSS). Under the above instantaneous linear model in Eq. (2.5), tissue heterogeneity correction is a blind source separation problem of the so-called exactly-determined case that can be solved by nonnegative independent component analysis (nICA).

The technique of ICA was first used in 1982 for analyzing a problem pertaining to neurophysiology [79]. ICA assumes that pure components are statistically independent and that at most one is normally distributed. These two requirements seem to be most critical for the success of the BSS approach to extraction of pure components [70]. However, in our special blind source separation problem for tissue heterogeneity correction, the source signals are dependent in general. Therefore, the proportion of each cell type inferred from ICA is not accurate and the purified molecular patterns do not conform to realistic cell-type-specific patterns [34, 80]. Significant amount of efforts has been devoted to relaxing the statistical independence assumption. For example, there are some other approaches that work on dependent sources [81], but the underlying assumptions are not readily applicable to tissue heterogeneity correction for microarray data analysis yet.

Recently, the nonnegative matrix factorization (NMF) technique, a natural constraint to many real world problems is reflected to enable further understanding of microarray data because nonnegative sources are meaningful in gene expression profiling [82]. In NMF, the source matrix **S** and the mixing matrix **A,** as well as the observation matrix **X** are assumed to be **strictly nonnegative** where the sources may be dependent. Simulation experiments, however, show that an inevitable problem of the NMF approach is that it may not converge to a stationary point [83]. One reason is that the basis of the space that NMF projects may not be unique theoretically; therefore, separate runs of NMF lead to different results. Another reason may come from the algorithm itself, such that the cost function is sometimes stuck in a local minimum during its iteration. Because NMF does

not yield a unique decomposition result, additional constraints are needed. Considering gene expression profiles, if phenotypic molecular signatures should be indicative of on-going biological process in specific cells, then it may be expected that only a small number of genes are highly up- or down-regulated within a single process. Hence, an algorithm for blind decomposition of microarray expression data has been derived in [84] using a non-negativity constraint for observations, the mixing matrix and sources with an additional sparseness constraint concerning the encoding of source signals. Naturally, the authors minimized a contrast function that explicitly exploits sparseness rather than statistical independence among the pure components.

Stochastic nonnegative independent component analysis (SNICA) is another new Monte Carlo approach to blindly separate nonnegative well-grounded sources from their linear mixtures [85]. It has the similar assumptions as in our proposed approach, while the de-mixing procedure is based on a Metropolis type Monte Carlo search for least dependent components by minimizing a cost function of mutual information between recovered components and using their non-negativity as a **hard constraint**. Therefore, we compared our algorithm with SNICA using noisy data, where signal-to-noise ratios (SNRs) are set at different levels, beyond the idealized theoretical model used in the algorithm derivation.

In general, the observed d dimensional data $\mathbf{x} = [x_1, x_2 \ldots, x_d]^T$ may be modeled as a linear mixing by mixing matrix, **A**, ($d \times d$) with additive noise

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n} \,, \tag{2.28}$$

where $\mathbf{s} = [s_1, s_2 \ldots, s_d]^T$ are the sources and **n** is assumed to be a white Gaussian noise with variance σ^2 so that

$$\log P(\mathbf{x} \,|\, \mathbf{A}, \mathbf{s}) \propto -\frac{1}{2\sigma^2}(\mathbf{x} - \mathbf{A}\mathbf{s}) \,. \tag{2.29}$$

Given the above model, we further assume that **n** is independently identically distributed (i.i.d.) and that **n** has a mean of $\vec{\mathbf{0}}$ and a covariance matrix of $\Sigma_{\mathbf{n}}$. Then we define signal-to-noise ratio (SNR) of data as

$$\mathrm{SNR(dB)} = 10\log_{10}\left(\frac{P_{signal}}{P_{noise}}\right) = 10\log_{10}\left(\frac{\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{As}(i)\|^2}{trace(\Sigma_{\mathbf{n}})}\right), \tag{2.30}$$

where N is the number of genes in the data set.

To fully compare the two algorithms, we tested the performance of our algorithm and SNICA using real breast cancer cell lines, MCF-7, A1N4 and Hs27, mixed *in silico* on the noise-free condition and on the decreasing SNRs for 30dB/20dB/10dB cases. We deployed the following procedure. The dependent source estimation was run several times (*i.e.*, 50 times here) with different randomly generated mixing matrices, and two criteria were recorded for comparison: performance index E_1 and correlation coefficients between true and estimated sources within the subISG sets. We took the mean of E_1 and correlation coefficients over 50 runs and the comparison results of E_1 are presented in Table 2.2 and Figure 2.12. The correlation coefficients of subISGs between original sources with estimations are shown in Figure 2.13.

Table 2.2 The mean of E_1 comparison results of IEG removal based nPICA with SNICA. The smaller the value of E_1, the better of the performance of the algorithm is.

	NoiseFree	**30dB**	**20dB**	**10dB**
nPICA	**0.7816**	**0.8429**	**1.4727**	4.5423
SNICA	1.1711	1.8103	3.1648	5.9877

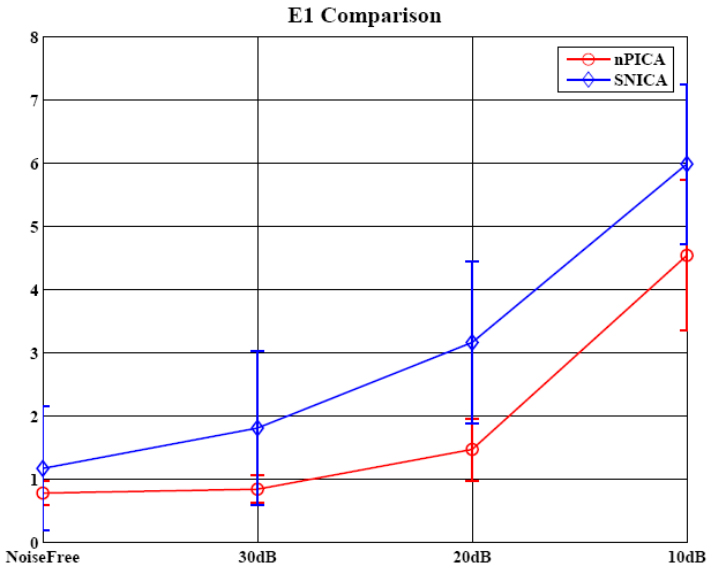
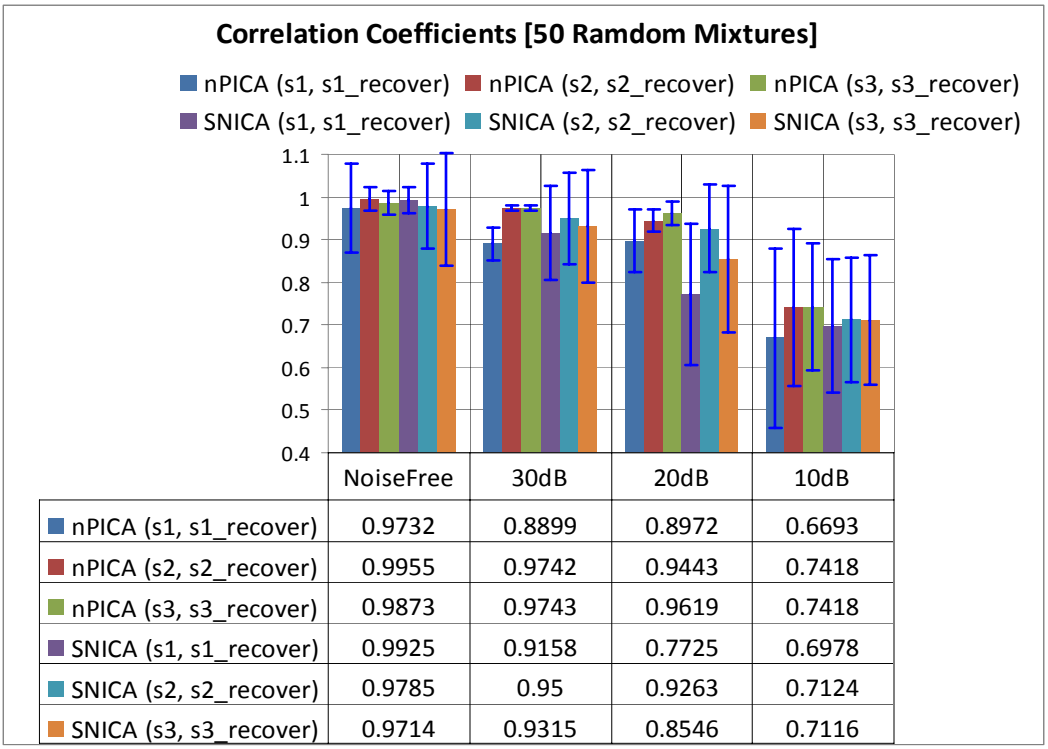


Figure 2.12: Comparison of the mean of E_1 between IEG removal based nPICA and SNICA with 50 random initializations of mixing matrixes. Error bars show the standard deviations of E_1.We compared at the noise-free and the signal-to-noise ratios (SNRs) at 30dB, 20dB and 10dB respectively.



	NoiseFree	30dB	20dB	10dB
nPICA (s1, s1_recover)	0.9732	0.8899	0.8972	0.6693
nPICA (s2, s2_recover)	0.9955	0.9742	0.9443	0.7418
nPICA (s3, s3_recover)	0.9873	0.9743	0.9619	0.7418
SNICA (s1, s1_recover)	0.9925	0.9158	0.7725	0.6978
SNICA (s2, s2_recover)	0.9785	0.95	0.9263	0.7124
SNICA (s3, s3_recover)	0.9714	0.9315	0.8546	0.7116

Figure 2.13: Comparison of correlation coefficients within the subISGs between IEG removal based nPICA and SNICA. The experiments ran 50 times with randomly generated mixing matrices **A**. The mean correlation coefficients of IEG removal based nPICA and SNICA are shown using different color bars in the figure. The error bars are the standard deviations of correlation coefficients over 50 runs. We compared at the noise-free case and the signal-to-noise ratios (SNRs) at 30dB, 20dB and 10dB respectively.

This study gives us useful insight into the performance of our proposed IEG removal based nPICA algorithm and that of SNICA. Results obtained here show that in the noise-free case, both algorithms can recover the dependent sources; our algorithm has a slightly better result. However, in SNICA, the de-mixing transformation is obtained as a sequence of random shears and rotations in search for the least dependent yet strictly nonnegative components [85]. When we applied SNICA to recover highly noisy mixtures, the strict non-negativity condition may be violated, thus, the performance of the SNICA algorithm degraded rapidly. Our proposed IEG removal-based nPICA algorithm is based on a cost

function whose minimum coincides with non-negativity, which can be achieved by a gradient algorithm under the whitening constraint. Therefore, it is more immune to noise, leading to an improved performance in recovering sources from noise corrupted signals as compared to SNICA.

2.4.3.3 Comparison of the performance of classification accuracy before and after tissue heterogeneity correction

DNA microarray technology has advanced our understanding of cancer by providing tumor gene expression signatures of different tumor types [86-88]. Specific gene expression signatures have been found to predict or elucidate treatment response [89], metastatic disease [90], and recurrence rate [91] that are associated with poor outcome in cancer patients. However, as discussed earlier, expression profiling studies of solid tumors generally include whole tumor sections consisting of tumor cells and the surrounding tissue microenvironment. We conjecture that the inclusion of extracellular matrix components and stromal cells, such as fibroblasts and immune response cells, influences the outcome of tumor profiling studies, because gene expression patterns are derived from both tumor cells and stromal cells. Therefore, in clinical settings, one of the criteria for inclusion of samples in study is the presence of more than 50% tumor cells in analyzed sections [92]. Even starting from a better tumor percentage of 60% to 70%, the discriminatory power of a predictor is clearly reduced [93].

To reduce tumor composition bias for a greater predictive accuracy and increase the number of samples that can be included for analysis, it is worthwhile to consider the effect of tumor composition on the performance of a predictive signature or ways of designing signatures that can take into account tumor samples with low tumor cell percentages. Toward this end, we designed a series of experiments to give some analytical results for the mixed samples that are consistent with our conjecture and some numerical simulations for the samples after tissue heterogeneity correction, aiming for an improved prediction accuracy of tumor tissue classification.

2.4.3.3.1 Microarray data description

We downloaded a microarray data set from NCBI (Gene Expression Omnibus GEO) GSE8671 [94]. It portrayed the characteristics of transcriptome profiles of human colorectal cancers. In order to characterize the molecular processes underlying the transformation of normal colonic epithelium to adenomatous tissue, the authors compared the transcriptomes of 32 prospectively collected adenomas with those of normal mucosa from the same individuals. Human Affymetrix Gene Chip U133 Plus 2.0 arrays were used to examine relative mRNA expression with 54,681 genes in total. Important differences emerged between expression profiles of normal and adenomatous tissues. The mRNA processing, microarray hybridization and analysis of the total 64 samples (32 pairs corresponding to 32 patients) was performed as described in [94]. We adopted this 32 patients data set in our study and treated their normal mucosa samples and adenomas samples as our purified source signals.

2.4.3.3.2 Generation of simulated mixture samples

We randomly generated 32 mixing matrices to mix normal and adenomas samples from 32 patients. The detailed mixing information is listed in the Table B.1. Linear discriminant analysis (LDA), which attempts to minimize the Bayes error by selecting the most discriminant feature vectors, is a popular method for feature extraction and dimensionality reduction in a supervised mode [95]. It is widely used in microarray data analysis [96, 97]. As an initial step to visualize the distribution of original pure samples and generated mixtures, we performed LDA to project the high-dimensional microarray data set into three dimensional space and the samples before and after mixing were plotted in Appendix B (Figure B.3).

2.4.3.3.3 Tissue heterogeneity correction by IEG removal based nPICA

We conducted the unsupervised IEG removal based nPICA procedure to decompose 64 mixture samples into 32 pairs of recovered signals corresponding to normal *vs*. adenomas tissues for 32 patients. The detailed results including E_1 for de-mixing matrix, correlation

coefficients of subISGs between pure and recovered sources and the numbers of subISGs selected for each trial of decompositions are listed in Table 2.3.

Table 2.3 IEG removal based nPICA results for 32 pairs of recovered signals corresponding to normal *vs*. adenomas tissues from 32 patients

patient #	Results		patient #	Results	
patient #1	E_1	0.8854	patient #17	E_1	0.6541
	CorrCoef (ISGs only)	0.9999/0.9998		CorrCoef (ISGs only)	0.9997/0.9998
	Number of ISGs	72/72		Number of ISGs	128/128
patient #2	E_1	0.9733	patient #18	E_1	0.363
	CorrCoef (ISGs only)	0.9997/0.9998		CorrCoef (ISGs only)	0.9982/0.9998
	Number of ISGs	242/242		Number of ISGs	288/288
patient #3	E_1	0.9597	patient #19	E_1	0.2649
	CorrCoef (ISGs only)	0.9999/0.9994		CorrCoef (ISGs only)	0.9988/0.9980
	Number of ISGs	72/72		Number of ISGs	242/242
patient #4	E_1	0.2163	patient #20	E_1	0.3271
	CorrCoef (ISGs only)	0.9982/0.9991		CorrCoef (ISGs only)	0.9989/0.9986
	Number of ISGs	72/72		Number of ISGs	200/200
patient #5	E_1	1.1219	patient #21	E_1	0.8049
	CorrCoef (ISGs only)	0.9998/0.9841		CorrCoef (ISGs only)	0.9792/0.9905
	Number of ISGs	72/72		Number of ISGs	72/72
patient #6	E_1	0.3694	patient #22	E_1	0.3219
	CorrCoef (ISGs only)	0.9983/0.9984		CorrCoef (ISGs only)	0.9999/1
	Number of ISGs	72/72		Number of ISGs	162/162
patient #7	E_1	0.2976	patient #23	E_1	0.532
	CorrCoef (ISGs only)	0.9935/0.9996		CorrCoef (ISGs only)	0.9997/0.9999
	Number of ISGs	200/200		Number of ISGs	288/288

patient #8	E_1	0.086	patient #24	E_1	0.4332
	CorrCoef (ISGs only)	0.9997/0.9997		CorrCoef (ISGs only)	0.9974/0.9990
	Number of ISGs	392/392		Number of ISGs	394/394
patient #9	E_1	0.605	patient #25	E_1	0.383
	CorrCoef (ISGs only)	0.9682/0.9983		CorrCoef (ISGs only)	0.9988/0.9472
	Number of ISGs	392/392		Number of ISGs	72/72
patient #10	E_1	0.2516	patient #26	E_1	1.3102
	CorrCoef (ISGs only)	0.9992/0.9997		CorrCoef (ISGs only)	0.9984/0.9714
	Number of ISGs	288/288		Number of ISGs	162/162
patient #11	E_1	0.4485	patient #27	E_1	1.3208
	CorrCoef (ISGs only)	0.9997/0.9914		CorrCoef (ISGs only)	0.9879/1
	Number of ISGs	288/288		Number of ISGs	72/72
patient #12	E_1	0.2722	patient #28	E_1	0.1907
	CorrCoef (ISGs only)	0.9955/0.9978		CorrCoef (ISGs only)	0.9993/1
	Number of ISGs	578/578		Number of ISGs	162/162
patient #13	E_1	0.2515	patient #29	E_1	0.3749
	CorrCoef (ISGs only)	0.9993/0.9996		CorrCoef (ISGs only)	0.9941/0.9987
	Number of ISGs	288/288		Number of ISGs	392/392
patient #14	E_1	1.2477	patient #30	E_1	0.2694
	CorrCoef (ISGs only)	0.9998/0.9983		CorrCoef (ISGs only)	0.9998/1
	Number of ISGs	128/128		Number of ISGs	200/200
patient #15	E1	0.1393	patient #31	E1	0.1178
	CorrCoef (ISGs only)	0.9985/0.9999		CorrCoef (ISGs only)	0.9998/0.9999
	Number of ISGs	50/50		Number of ISGs	392/392
patient #16	E1	0.37	patient #32	E1	0.3996
	CorrCoef (ISGs only)	0.9999/0.9999		CorrCoef (ISGs only)	0.9981/0.9970
	Number of ISGs	72/72		Number of ISGs	128/128

2.4.3.3.4 Classification comparison for the samples before and after tissue heterogeneity correction

Support vector machines (SVMs) [98] are widely used in bioinformatics [99, 100] and other applications of supervised learning. A support vector machine constructs a hyperplane that maximizes the margin between two classes of microarray samples consisting of colorectal adenomas cancer tissues and normal colonic epithelium tissues in n-dimensional space, where in this analysis n corresponds to the number of features (genes) selected by student's t-test.

The training set consists of original pure samples with 32 normal tissues plus 32 cancer tissues. SVM classifier was used with a linear kernel and default parameter setting: penalty parameter [98] $C = 1$ for binary classification. A 3-fold cross validation was then performed on the train set – original pure samples. In each fold, briefly, for each feature (a gene), a student's t-test was performed within the samples in the current two folds out of three to yield a p-value. Features were ranked by p-values, and the top 100 or less features that met the p-value cutoff, 0.05, were retained on the training data set, and we evaluated the performance of the classifier using the third fold with the number of features swept over from 1 to 100. The cross validation results are shown in Figure 2.14.

Then we did independent tests for mixtures before THC and recovered signals after THC. 100 top features were selected based on the p-values of student's t-test on the entire pure sample data set. The classifiers were built with the number of features from 1 to 100. Since we wanted to estimate the tissue contamination effects, we treated all the mixture samples as the cancer class. The sensitivity, false negative rate and overall classification accuracy on part of selected features in independent tests for mixtures and recovered signals, respectively, in noise free case are listed in Table 2.4. More figures of the performance analysis, including sensitivity, false negative rate and overall classification accuracy for the number of features from 1 to 100, are presented in Appendix B (Figure B.4).

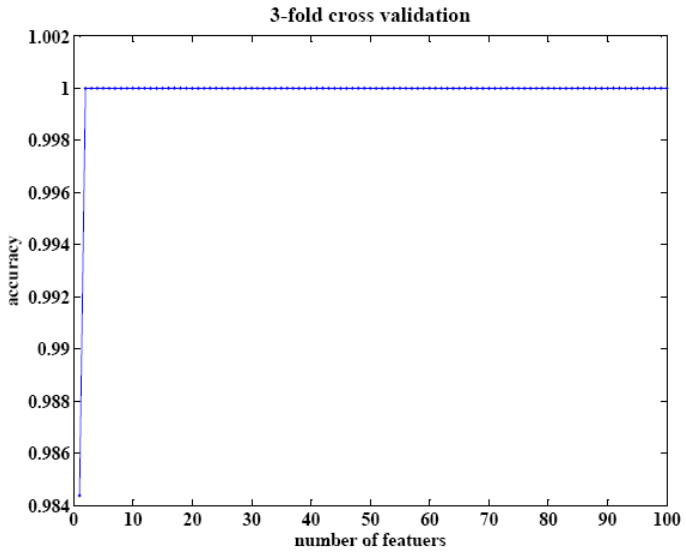


Figure 2.14: 3-fold cross validation results of SVM on the training set of original pure samples consisting of 32 normal tissues *vs*. 32 colorectal cancer tissues. We swept over the number of features from 1 to 100 to evaluate the classifier's performance.

To test the capability of our algorithm to recover useful signals from microarray measurements with the presence of noise for classification, we also examined classification accuracy in noisy models. We added noises to observations as described in Section 2.4.3.2 with SNRs at 40dB, 35dB, 30dB, 25dB, 20dB, 15dB and 10dB respectively. We listed the results of sensitivity, false negative rate and overall classification accuracy on the independent tests for recovered signals and mixtures in noisy models (feature number = 100) in Table 2.5**.** We reported the entire curves of sensitivity, false negative rate and overall classification accuracy in the independent tests for recovered signals and mixtures in noisy models in Appendix B (Figure B.5 and Figure B.6). The results show that, when varying SNRs from 40dB to 10dB, we can achieve a significantly better classification accuracy and lower false negative rate with the data after tissue heterogeneity correction (THC) than the mixture samples.

Table 2.4: The sensitivity, false negative rate and overall classification accuracy for the selected features on the independent tests for recovered signals and mixtures in noise-free case.

Feature Numbers (gene numbers)	Performance Measure	Before THC (mixtures)	After THC (recovered signals)
10	Sensitivity	0.5313	1
	False Negative Rate	0.4688	0
	Accuracy	0.5313	0.9844
30	Sensitivity	0.5156	1
	False Negative Rate	0.4844	0
	Accuracy	0.5156	0.9844
50	Sensitivity	0.5156	1
	False Negative Rate	0.4844	0
	Accuracy	0.5156	1
70	Sensitivity	0.5156	1
	False Negative Rate	0.4844	0
	Accuracy	0.5156	1
90	Sensitivity	0.5625	1
	False Negative Rate	0.4375	0
	Accuracy	0.5625	0.9844
100	Sensitivity	0.5625	1
	False Negative Rate	0.4375	0
	Accuracy	0.5625	0.9844

Although these classification results have already shown that normal tissues and adenomas tissues are quite well separated after THC and lead to a great improvement in classification accuracy, our goal was to investigate the THC effect in microarray experiments even in low signal-to-noise ratio situation with different biological noises and technical noises corrupting the sources. We decided to further add 5dB Gaussian noise on pure samples purposely, besides the observation noise, to simulate noisy sources and make the two classes of pure tissues inseparable. Next, we tested THC classification results as before by varying observation noise levels with SNRs from 40dB to 10dB. We listed the results of sensitivity, false negative rate and overall classification accuracy in the independent tests for mixtures and recovered signals in noisy source models (feature number = 100) in Table 2.6. The entire results are shown in Appendix B (Figure B.7 and Figure B.8). As is depicted in the figures, compared to the mixture tumor samples that

show very limited discriminatory power, recovered samples after THC have strong discriminatory power to discriminate tumor and stromal patterns.

Table 2.5: The sensitivity, false negative rate and overall classification accuracy in the independent tests for recovered signals and mixtures in noisy models (feature number = 100).

SNR (dB)		Before THC (mixtures)	After THC (recovered signals)
40	Sensitivity	0.5625	1
	False Negative Rate	0.4375	0
	Accuracy	0.5625	0.9844
35	Sensitivity	0.5625	1
	False Negative Rate	0.4375	0
	Accuracy	0.5625	1
30	Sensitivity	0.5625	1
	False Negative Rate	0.4375	0
	Accuracy	0.5625	0.9844
25	Sensitivity	0.5625	1
	False Negative Rate	0.4375	0
	Accuracy	0.5625	1
20	Sensitivity	0.5625	1
	False Negative Rate	0.4375	0
	Accuracy	0.5625	0.9844
15	Sensitivity	0.5469	0.9688
	False Negative Rate	0.4531	0.0313
	Accuracy	0.5469	0.9688
10	Sensitivity	0.5938	1
	False Negative Rate	0.4063	0
	Accuracy	0.5938	0.9844

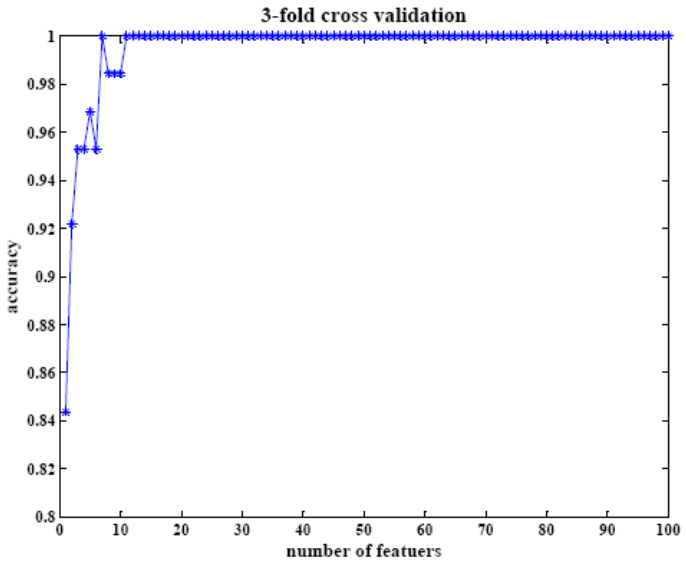


Figure 2.15: 3-fold cross validation results of SVM on the training set in noisy sources case. We swept over the number of features from 1 to 100 to evaluate the classifier's performance.

Table 2.6: The sensitivity, false negative rate and overall classification accuracy in the independent tests for recovered signals and mixtures in noisy sources models (feature number = 100).

SNR (dB)		Before THC (mixtures)	After THC (recovered signals)
Noise Free	Sensitivity	0.5625	1
	False Negative Rate	0.4375	0
	Accuracy	0.5625	0.9844
40	Sensitivity	0.5625	1
	False Negative Rate	0.4375	0
	Accuracy	0.5625	0.9844
35	Sensitivity	0.5625	1
	False Negative Rate	0.4375	0
	Accuracy	0.5625	0.9688
30	Sensitivity	0.5625	1
	False Negative Rate	0.4375	0
	Accuracy	0.5625	0.9688
25	Sensitivity	0.5469	1
	False Negative Rate	0.4531	0
	Accuracy	0.5469	0.9688
20	Sensitivity	0.5625	1
	False Negative Rate	0.4375	0
	Accuracy	0.5625	0.9844

15	Sensitivity	0.5781	1
	False Negative Rate	0.4219	0
	Accuracy	0.5781	0.9688
10	Sensitivity	0.5625	1
	False Negative Rate	0.4375	0
	Accuracy	0.5625	0.9688

2.5 Conclusion and Discussion

Although microarrays have now become established tools to study gene expression patterns in human cancer, it is often the case that tumor specimens naturally represent a mixture of tissues. We address this challenge in the context of composite phenotype separation by our proposed ISG-nPICA approach. Our experimental results demonstrate that ISG-nPICA as a technique for composite phenotype decomposition can rigorously characterize changes in mixed populations of cells. Our approach has several advantages over traditional method for quantification of cell species — laser capture microdissection (LCM) or other computational dissection methods. First, conventional method LCM deals with at most a handful of different cell types at one time, while expression analysis has the potential to simultaneously quantify a much greater number of cell types. Secondly, our approach integrates the partially independent information of a large number of genes to yield its results; this independency is important because it defines a complete characterization of necessary and sufficient conditions for separating composite phenotypes. ISG-nPICA currently focuses heavily on phenotype-specific cases, where each subISG term in Eq. (2.6) contains phenotypically up-regulated genes. We propose that computational extraction of a subISG complex from various phenotypic conditions may be important for developing nPICA-subISG based representations of regulatory modules. Thirdly, our method employs a cost function whose minimum coincides with non-negativity. Comparing to SNICA which relies on strict non-negativity constraint, our method mitigates the contribution of noise inherent in biological measurements and thus boosts robustness to the effects of tumor composition on the performance of a predictive signature. Currently, we only tackle the problem by assuming the underlying cell types

having two to three. Of course, expanding the arsenal of data on basis cell types would enable thorough analysis of such diverse kinds of samples.

3 Identification of Regulatory Gene Module Composites by Latent Process Decomposition (LPD)

3.1 Introduction

Genes often interact with each other to carry out cellular activities [101]. The set of genes tightly regulated in a specific cellular process can be considered as a process-specific transcriptional module [82, 101, 102]. It is important to identify such modules for understanding the organization and functions of gene groups associated with different experimental conditions, which may further help garner gene expression signatures associated with diseases.

A wide range of unsupervised methods have been proposed to identify gene transcriptional modules from microarray data. In general, approaches fall mainly into three categories:

1) Various conventional clustering techniques, such as hierarchical clustering [103], *k*-means [104] and self-organizing maps [105], are in common use for identifying meaningful subgroup genes exhibiting similar expression patterns. These approaches played a key role in gaining insights into the biological mechanisms associated with different physiological states. However, these basic clustering approaches are not well tuned for regulatory module identification due to two main factors

(1) A set of co-regulated genes may only co-express in a subset of experimental conditions [106, 107]. Many genes in the same functional pathway may not have similar expression profiles as measured by correlation statistics or other standard pair-wise expression similarity measure. This is especially true for pairs of genes that are not in the same region of a signaling pathway. These genes will not be

discoverable using those traditional clustering methods.

(2) Since many genes belong to multiple regulons [107], clustering the genes into one and only one group may also mask the interrelationships between genes that are assigned to different clusters but show local similarities in their expression patterns. Thus, biologists are more interested in finding the hidden regulatory patterns behind gene expression patterns, which strengthens the biological relevance of the grouped genes, *i.e.*, the genes are co-regulated to form transcriptional modules.

2) The second category of unsupervised method is model-based approaches. Typically, the procedures these approaches employ include first generating a model that explains the interactions among biological entities participating in genetic regulatory networks, and then training the parameters of the model on expression datasets [9, 108, 109]. Depending on the complexity of the model, one challenge of model-based approaches is the lack of sufficient data to train the parameters, and another challenge is the prohibitive complexity and computational load of training algorithms.

3) Recently, matrix decomposition methods have been introduced to uncover transcriptional modules from microarray data, including independent component analysis (ICA) [42-45] and non-negative matrix factorizations (NMF) [60, 110-112]. These methods treat microarray data as a mixture of unknown factors (or components) that may correspond to specific biological processes. Specifically, the level of any given mRNA expression is modeled as the net sum of a complex superposition of cooperating and/or counteracting biological processes. ICA is a statistical method for revealing independent hidden factors that underlie sets of random variables or observations. In the context of microarray data, these statistically independent hidden factors may correspond to putative biological processes or transcriptional modules. Clusters found by ICA have been shown to be directly associated with biological processes with common regulatory mechanisms [43, 44]. While the ICA model has demonstrably outperformed other linear representations of the data such as principal components analysis (PCA), a validation using explicit pathway and regulatory element information has also been performed by Teschendorff [45]. Although, ICA provides a framework for a more biologically relevant interpretation of genome-wide transcriptomic data, it is still problematic to directly apply

ICA to gene expression data due to its strong assumption of the independence of hidden variables in whole gene population. While a wealth of genes is constantly expressed [71, 113], biologically, it is more plausible to assume that the independence holds only for those genes that actively participating in biological processes. Therefore, we need to make further assumptions to constrain the ICA model for gene module identification.

The ICA model has been used to identify gene modules of co-regulated genes, their regulators and the regulation programs [42, 43, 114]. Actually, there are two ICA models within the framework of latent variable modeling.

- **ICA Model I**

  Let $x_j(i)$ be the expression level of gene *i*, $i = 1, \cdots, N$, in sample (sample-course) or at time (time-course) *j*, $j = 1, \cdots, M$; $a_j(k)$ be the activity profile of regulator or transcription factor (TF) *k*, $k = 1, \cdots, L$, in sample or at time *j*; and $s_k(i)$ be the regulation strength of regulator *k* on gene *i*. We assume $L \le M \ll N$ for assuring system identifiability. If in such a way that random variables $\{s_1(i), s_2(i), \cdots s_L(i)\}$ are statistically independent, we can write ICA model as [42, 114]:

$$\begin{bmatrix} x_1(i) \\ x_2(i) \\ \cdots \\ x_M(i) \end{bmatrix} = \begin{bmatrix} a_1(1) & a_1(2) & \cdots & a_1(L) \\ a_2(1) & a_2(2) & \cdots & a_2(L) \\ \cdots & \cdots & \cdots & \cdots \\ a_M(1) & a_M(2) & \cdots & a_M(L) \end{bmatrix} \begin{bmatrix} s_1(i) \\ s_2(i) \\ \cdots \\ s_L(i) \end{bmatrix}, \tag{3.1}$$

  which describes how the observed gene expressions are generated by a regulation program of mixing the latent strengths. In this model, the expression level $x_j(i)$ of gene *i* under 'condition' *j* is considered as the sum of the expression levels $a_j(k)$ of the regulators $k = 1, \cdots, L$ under 'condition' *j*, weighted by their contributions/influences $s_k(i)$ to the expression of gene *i*.

- **ICA Model II**

  ICA Model II differs from ICA Model I only in the interpretation of its variables. Let $x_j(i)$ still be the observed expression level of gene *i*, $i = 1, \cdots, N$ in sample (sample-course) or at time (time-course) *j*, $j = 1, \cdots, M$; $s_k(i)$ be the activation pattern of gene *i* within biological process *k*, $k = 1, \cdots, L$; and $a_j(k)$ be the participation degree of biological process *k* in sample or at time *j*. We also assume

$L \le M \ll N$ for assuring system identifiability. We further assume that the random variables $\{s_1(i), s_2(i), \cdots s_L(i)\}$ corresponding to distinct biological processes are statistically independent. Thus decomposition of the expression data matrix **X** using ICA Model II can also be characterized by Eq. (3.1). Specifically, it describes how the observed gene expressions are generated by a participation program of mixing the latent biological processes. In this model, the expression level $x_j(i)$ of gene *i* under 'condition' *j* is considered as the sum of the expression levels $s_k(i)$ of the gene *i* within biological process $k, k = 1, \cdots, L$ weighted by their contributions/influences $a_j(k)$ of the biological processes *k* to the observed expression of gene under 'condition' *j*.

One point that needs to be noted is that the difference of variable interpretation in Model I & II is subtle. In Model I, the activity profile $a_j(k)$ and regulation strength/influence $s_k(i)$ are explicit while the gene expression exclusively regulated by a particular regulator is implicit. In Model II, the regulator expression is implicit while the involvements of biological processes $a_j(k)$ and gene expression involved in a particular biological process $s_k(i)$ are explicit.

In this chapter, we follow Model II and propose to use nonnegative ICA (nICA) – a latent process decomposition (LPD) – for gene module identification [49, 115]. Each latent (hidden) process is defined as a set of functionally-related genes or transcriptional modules. Accordingly, in this model each sample, *i.e.* a gene, in the data set is represented as a combinatorial mixture over activation patterns of a finite set of latent processes, which are expected to correspond to various biological processes. We use the term *latent process*, rather than cluster, because a gene can have partial membership of several processes simultaneously, in contrast to many hard clustering partition approaches.

Since the expression levels (or ratios) of the latent independent biological processes should be nonnegative, non-negativity is a natural choice for this problem. nICA exploits the non-negativity constraint to enforce the independence/uncorrelation among biological processes in those participated genes. In principle, nICA can be thought of as a projection method with which the expression levels (or ratios) are projected onto some new

nonnegative components with least statistical dependence. We believe that nICA provides a better model of gene expression data than ICA does, hence, making it more appealing for gene module discovery.

Here, we describe a complete algorithm for the nICA approach and organize the chapter in the following way: In section 3.2, we introduce the algorithm which consists of the following components – input variable selection, stability-based dimension estimation, learning algorithm of nICA, and gene clustering by VIsual Statistical Data Analyzer (VISDA) [56]. In section 3.3, we demonstrate the effectiveness of the proposed approach for gene module identification using yeast and muscle regeneration datasets. The biological relevance of the identified gene modules is validated by functional annotation analysis. Compared with conventional soft clustering approaches to clustering gene expressions and matrix decomposition-based approaches, the proposed approach appears to have improved performance in finding biologically meaningful transcriptional modules with the lower time complexity. We finish with some discussions and conclusions in section 3.4.

3.2 Methods of Latent Process Decomposition

The block diagram of the proposed approach is outlined in Figure 3.1. As can be seen from the figure, there are four major steps in the approach: (1) an input variable selection procedure is first used for sample selection; (2) stability analysis is followed to determine the number of components; (3) nICA with a learning algorithm is then applied to recover the nonnegative independent components; and (4) gene clustering by VISDA is finally performed in the latent space to identify the gene modules. We provide a more detailed description of each step as follows.

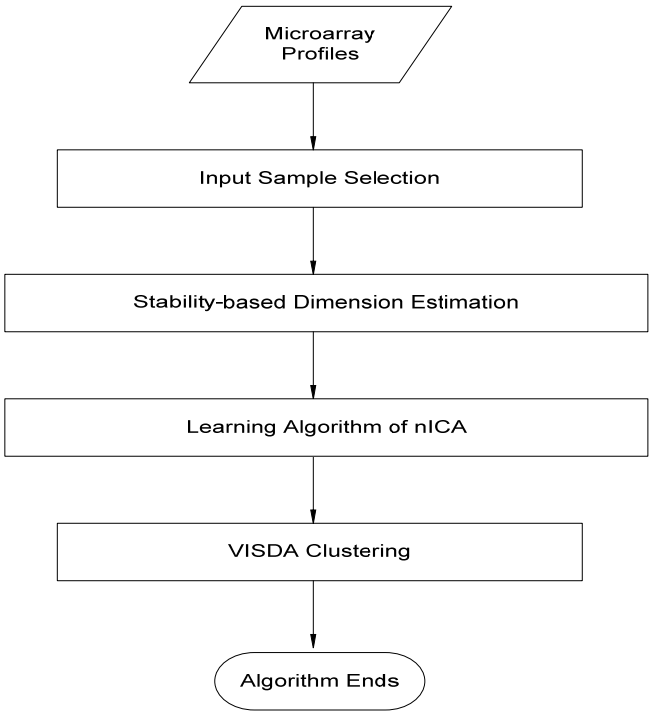


Figure 3.1: Flowchart of the proposed nICA approach for gene module identification.

3.2.1 Input Variable Selection

Input sample selection aims at selecting the most informative samples among the available samples for nICA decomposition. Without the proper selection, some computational problems such as increased computational complexity and degraded convergence may arise. Even worse, some samples may cause singularity problems for nICA decomposition. Principle components analysis (PCA), a variance-based dimension reduction technique, is often used for input sample selection. But PCA is not always effective for nICA decomposition since the variance of a sample is not necessarily related to its importance.

Here, we propose to use mutual information [116] to perform input sample selection. The objective is to select M' informative samples $(\boldsymbol{v}_1,...\boldsymbol{v}_{M'})$ from a set of M samples $(\boldsymbol{x}_1,...,\boldsymbol{x}_M)$, where $M > M'$. At each step of the algorithm, we choose a sample that is as statistically independent as possible [68] from the already selected samples $\boldsymbol{v}_j, j = 1,...,k-1$. In other words, $\boldsymbol{x}_l$ is the k-th selected sample (*i.e.*, $\boldsymbol{v}_k$) if the

following cost function $f(i,k-1)$ (defined as the sum of mutual information) is minimized when $i = l$:

$$f(i,k-1) = \sum_{j=1}^{k-1} MI(\boldsymbol{x}_i, \boldsymbol{v}_j) \quad for\ all\ \boldsymbol{x}_i \notin \{\boldsymbol{v}_j, j = 1,...,k-1\}, \tag{3.2}$$

where $MI(.,.)$ denotes the mutual information that is defined as [117]:

$$MI(\boldsymbol{x}_i, \boldsymbol{v}_j) = H(\boldsymbol{x}_i) + H(\boldsymbol{v}_j) - H(\boldsymbol{x}_i, \boldsymbol{v}_j). \tag{3.3}$$

In Eq. (3.3), $H(\boldsymbol{x}_i)$ ($H(\boldsymbol{v}_i)$) represents the entropy of a centered univariate random variable $\boldsymbol{x}_i$ ($\boldsymbol{v}_i$) and $H(\boldsymbol{x}_i, \boldsymbol{v}_j)$ represents the joint entropy of two centered univariate random variable $\boldsymbol{x}_i$ and $\boldsymbol{v}_j$ [118], which are computed in the following approximations respectively:

$$H(\boldsymbol{x}_i) = H(\boldsymbol{r}_i) + \log\sigma \approx \frac{1}{2}\log(2\pi e\sigma^2) - \frac{1}{48}(E\{\boldsymbol{r}_i^4\} - 3)^2 - \frac{1}{12}(E\{\boldsymbol{r}_i^3\})^2, \tag{3.4}$$

where $\boldsymbol{r}_i = \boldsymbol{x}_i/\sigma, \sigma^2 = \mathrm{var}(\boldsymbol{x}_i)$. And

$$\begin{aligned} H(\boldsymbol{r}_i, \boldsymbol{s}_i) \approx \log(2\pi e) &- \frac{1}{12}[(E\{\boldsymbol{r}_i^3\})^2 + (E\{\boldsymbol{s}_i^3\})^2 + 3(E\{\boldsymbol{r}_i^2\boldsymbol{s}_i\})^2 + 3(E\{\boldsymbol{r}_i\boldsymbol{s}_i^2\})^2] \\ &- \frac{1}{48}[(E\{\boldsymbol{r}_i^4\} - 3)^2 + (E\{\boldsymbol{s}_i^4\} - 3)^2 + 6(E\{\boldsymbol{r}_i^2\boldsymbol{s}_i^2\} - 1)^2 + 4(E\{\boldsymbol{r}_i^3\boldsymbol{s}_i\})^2 + 4(E\{\boldsymbol{r}_i\boldsymbol{s}_i^3\})^2] \end{aligned}, \tag{3.5}$$

where $(\boldsymbol{r}_i, \boldsymbol{s}_i)$ are whitened variables of $(\boldsymbol{x}_i, \boldsymbol{v}_j)$. Therefore, the selected subset $(\boldsymbol{v}_1,...\boldsymbol{v}_{M'})$ will contain the samples that are mutually "quite different" as a result of the m inimization of mutual information.

3.2.2 Stability-based Dimension Estimation

In practice, the number of independent components for nICA is often determined by the user's prior knowledge or obtained by PCA with a criterion of containing 95% of energy mainly to eliminate the noise effect [68]. However, it is often a difficult task in microarray data analysis to obtain a meaningful number of components by either the prior knowledge or PCA approach. When the number of components is incorrectly estimated, nICA will produce many possible false components for gene module identification. Hence, we propose to conduct stability analysis on gene expression data to estimate the

number of components. Figure 3.2 shows the proposed stability-based schema, namely "splitting by samples", for reliable dimension estimation of nICA [49, 119].

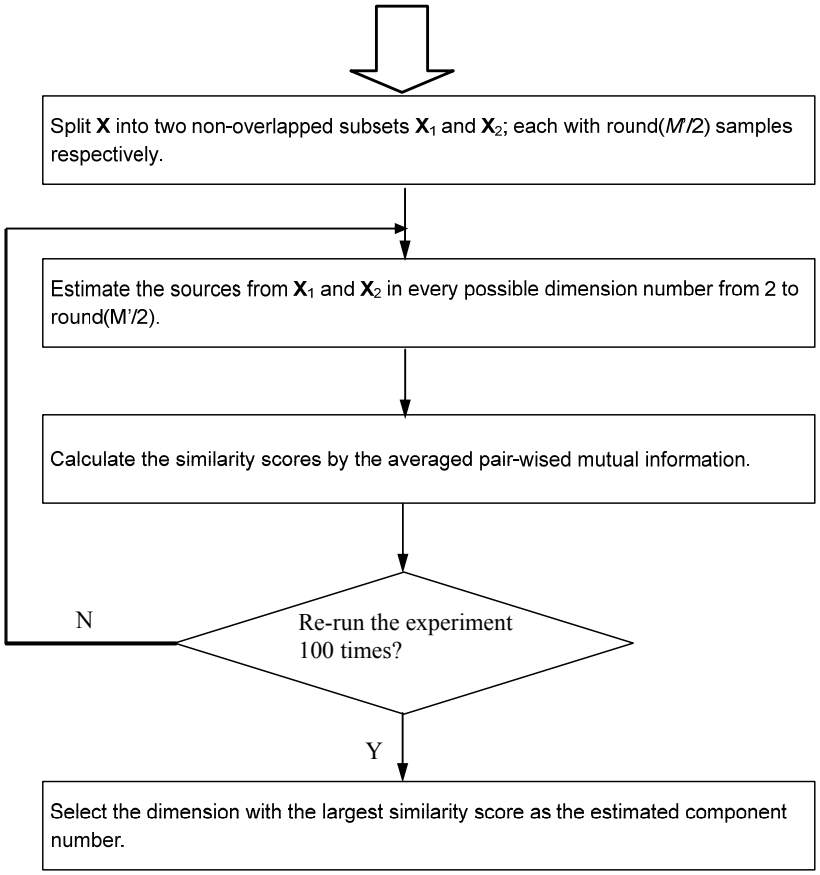


Figure 3.2: General schema of "splitting by samples" for dimension estimation

The basic idea of the stability-based approach is that the nICA results from two data subsets sampled from a common data set should be consistent. The consistency (or similarity) of the nICA results from two non-overlapped subsets reflects of the consistency between the assumed and underlying component numbers. More specifically, we split the samples into two non-overlapped subsets for nICA analysis and run the algorithm from $i = 2$ to the full dimension of the subset of samples. We believe that if the dimension estimation truly captures the underlying biological component number, the similarity score measured by mutual information between the components estimated from the two data subsets should give the best similarity score among all of the dimensions. When the estimated component number is not equal to the true number, the nICA results

will show a tendency of mismatched components being estimated, and a consequent decrease of similarity.

Due to the ambiguity of the scale in the nICA estimates, we need to normalize the estimated components and register them before calculating the similarity score. In our approach, we first normalize the estimated components to be unit-variance variables. We then perform the registration (or alignment) of two permutated versions of components via an information-theoretic approach. The exact way to align (or register) different pairs of components is by examining their mutual information. We calculate the similarity score after the alignment using averaged pair-wise mutual information:

$$Q = \frac{1}{\lfloor M'/2 \rfloor} \sum_{i=1}^{\lfloor M'/2 \rfloor} MI(\boldsymbol{s}_i^{(1)}, \boldsymbol{s}_i^{(2)}) \quad , \tag{3.6}$$

where $MI(.,.)$ denotes the mutual information estimate as defined in Eq. (3.3), $\lfloor \cdot \rfloor$ is the floor function, and $(\boldsymbol{s}_i^{(1)}, \boldsymbol{s}_i^{(2)})$ is the *i-th* aligned pair of the components estimated from two different subsets. In order to obtain a reliable estimation of the dimension number, stability tests are performed P times independently (in our experimental design, we re-run the algorithm P=100 times with random initialization), each time after a random shuffling to the order of samples. Finally, we choose the dimension with the largest similarity score averaged over P runs as the estimate of the component number.

3.2.3 Learning Algorithm of nICA

Although the nICA algorithm which consists of two major steps - pre-whitening and axis rotations - has been described in Section 2.3.2,a brief recap is provided to make this chapter complete. Specifically, the pre-whitening of the observed data can be achieved by

$$\mathbf{z}(i) = \mathbf{V}\mathbf{D}^{-1/2}\mathbf{V}^T\mathbf{x}(i), \tag{3.7}$$

where $\mathbf{V}$ is the orthogonal matrix of eigenvectors of the observation matrix, and $\mathbf{D}$ is the diagonal matrix of corresponding eigenvalues. To further correct the remaining orthonormal rotation ambiguity, after pre-whitening has succeeded in making the whitened data scatter plot orthogonal to each other, we should search for a rotation so that all the data fit into the first quadrant. The rotation matrix $\mathbf{W}$ can be constructed by minimizing the following cost function

$$J(\mathbf{W}) = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{z}(i) - \mathbf{W}^T g^+\left(\mathbf{W}\mathbf{z}(i)\right)\right)^2, \tag{3.8}$$

where $g^+(s)=\max(0, s)$ is the truncation nonlinearity function which is zero for negative $s<0$ and s otherwise. It can be shown that in the Stiefel manifold of rotation matrices, function (3.8) has no local minima, and it is a Lyapunov function for its gradient matrix flow [74]. Thus, the search for **W** is monotonically converging and is guaranteed to find a global yet unique solution. (The detailed steps of the algorithm are listed in the Eq. (2.10) to Eq. (2.13).)

3.2.4 Gene Clustering in the Latent Space by VISDA

After performing nICA, we obtain the independent components representing some distinct biological processes. In these putative biological processes, the genes showing relatively high or low expression levels are most interesting. The analysis of gene patterns that are significantly over- or under-expressed in the components may provide insights into the biological events associated with these latent processes. We first use a pre-screening procedure to single out these genes and then apply VISDA to analyze the gene patterns in the latent space. In the pre-screening procedure, we first sort the genes by their contributions (or loads) in each component, which creates a natural ordination in which genes are arranged based on their association with a given component. Then we select a subset of genes within each component, *i.e.*, the over-expressed genes or under-expressed genes according to the value of each gene in the component [43]. By taking the union of the selected genes from each component, we form a pool of genes that we believe are closely related to the biological processes revealed by nICA.

We then employ VISDA, a statistical model based clustering tool, to perform gene clustering on those selected genes in the latent space. Based on a hierarchical standard finite normal mixture (SFNM) model, VISDA captures the coherent structures in the latent space and performs top-down divisive clustering. The fitting process of the SFNM model is achieved by the Expectation Maximization (EM) algorithm [56] , which maximizes the likelihood function. For each cluster at a level of the hierarchy, VISDA uses five different projection methods (principle component analysis (PCA), PCA–projection pursuit (PCA–PPM), locality preserving projection (LPP), HC–KMC–SFNM–

DCA and affinity propagation clustering – discriminatory component analysis (APC–DCA) [56]) to visualize the sub-clusters within the clusters. The user chooses one of the projections that he/she thinks best reveals the data structures. On the chosen projection, the user initializes models with different number of clusters by clicking on the computer screen at the centers of the clusters. These two-dimensional (2-D) models will be refined by the EM algorithm and compete according to Minimum Description Length (MDL) criterion or human justification. The winning model in 2-D space will be transferred back to original data space to initialize the data model in that space. Then the EM algorithm in original data space will refine the model and obtain the partition of data at that level. At the top level, the whole dataset is split into several coarse clusters that may contain multiple functional modules; at lower levels, these coarse clusters are further decomposed into finer clusters, until no substructures can be found.

3.3 Results

We demonstrate the effectiveness of the proposed approach for identifying gene module using yeast and muscle regeneration datasets. The biological relevance of the identified gene modules is validated by functional similarity analysis. Compared with the conventional soft clustering and matrix-decomposition based approach, the proposed approach appears to have improved performance in finding biologically meaningful transcriptional modules.

The following three expression datasets are studied: Dataset 1 – budding yeast during cell cycle CLB2/CLN3 overactive stain [103], consisting of spotted array measurements of 6178 genes in 77 time points; Dataset 2 – yeast in various stressful conditions consisting of spotted array measurements of 6,152 genes in 173 experiments [120]; and Dataset 3 - a 27-time points muscle regeneration series *in vivo* murine regeneration with Affymetrix oligonucleotide array measurements of 7,570 genes [121]. For determining whether the proposed approach can uncover the gene modules from gene expression data in the latent space, we mainly used the Biological Network Gene Ontology tool (BiNGO) [122] to evaluate the enrichment functional annotations.

3.3.1 Yeast Cell Cycle Data

The yeast cell-cycle dataset was preprocessed to obtain log-ratios between red and green intensities, *i.e.*, $x_j(i) = \log_2(R_{ij}/G_{ij})$. Since the data set contains both positive and negative log-ratio values, we need to do data pre-treatment. We assume that distinct regulatory interactions are responsible for up-regulation versus down-regulation of gene expression. With the spirit of "divide and conquer", we split the data into two parts – positive and negative values corresponding to over-representative and under-representative gene sets respectively to fit the nICA model.

To prevent over-learning the dimension of the data was reduced using the input variable selection procedure described in Section 3.2 Methods. We used the mutual information quality index, $f(i,k-1)$ as in Eq. (3.2), to evaluate the observations for the most suitable number of inputs. Figure 3.3 shows the sum of the mutual information measured for all the input samples in the positive part of the data. As can be seen, there is an apparent increase at the dimension of 66. Therefore, we selected 65 samples for the positive part and 62 for the negative part (the figure is not shown here) for further nICA analysis. Secondly, we used the stability-based dimension estimation method to estimate the number of independent components. The stability analysis results are shown in Figure 3.4, and an apparent peak is obtained from the averaged pair-wise mutual information when the number of components is equal to 3. Third, we then applied the nICA learning algorithm to uncover the independent components. Finally, we obtained gene modules by gene clustering using VISDA in the latent space. The four most significant gene clusters are given in Table 3.1. We measured the biological significance of each cluster using BiNGO tool. The *p*-value of each cluster was calculated according to its overlap with the functional annotations in Gene Ontology (GO). Assume we found a module with *X* genes, in which there are *x* genes sharing one functional category *C*. In BiNGO, the hyper-geometric test uses the hyper-geometric distribution to calculate the probability of obtaining *x* or more of these genes belong to a functional category *C* shared by *n* of the *N* genes in the reference set as created above by chance. The obtained *p*-values have been corrected in order to control the type I error (false positive) rate [123] by Benjamini and Hochberg correction [124].

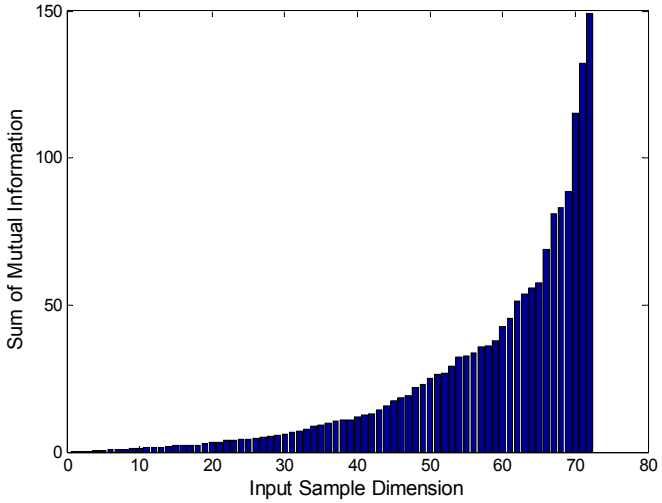


Figure 3.3: Input sample selection for the samples in the positive part of the cell cycle data set.

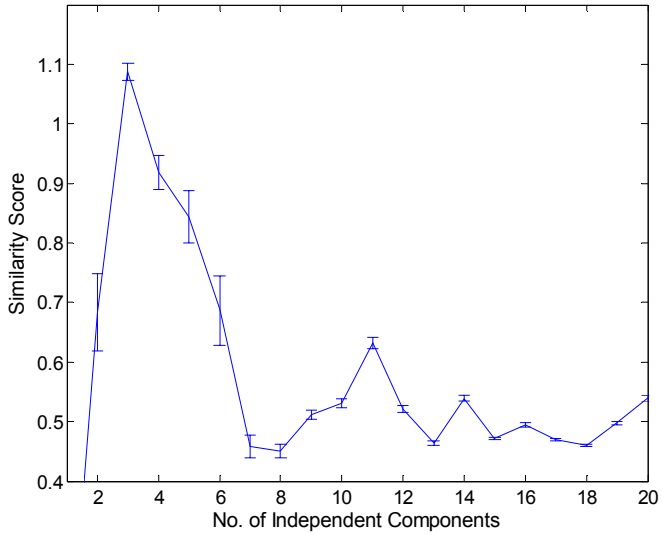


Figure 3.4: Stability analysis of the positive part of the cell cycle data. The average similarity score with error bars over 100 runs. The estimated underlying component number is three.

Table 3.1 The four most significant clusters from nICA for the cell cycle data set. Numbers in parentheses in the fifth column show the percentage of genes within the cluster that are presented in one of the functional categories. The numbers in the sixth column are presented in the similar way which corresponds to the total number within the whole genome set that are annotated with one of the special categories in GO system.

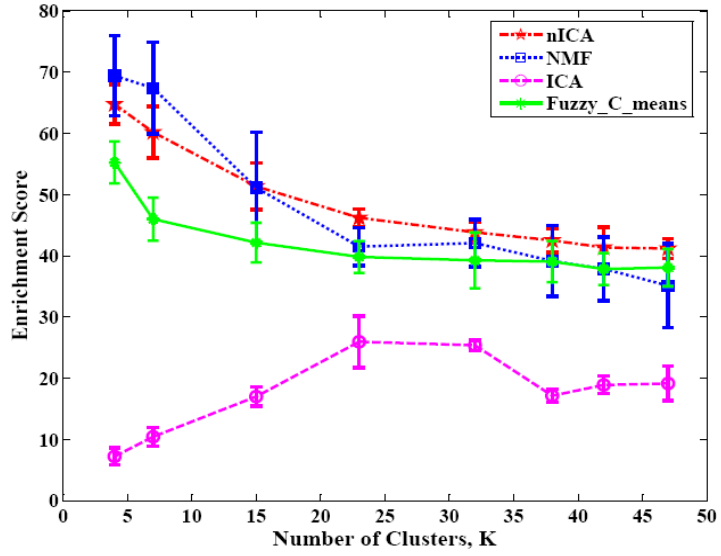
Cluster ID	GOID	GO term	p-value	cluster frequency	total frequency
1	6365	35S primary transcript processing	1.27E-27	26/102 (25.4%)	76/5638 (1.3%)
	42255	ribosome assembly	8.87E-18	18/102 (17.6%)	62/5638 (1.0%)
	42273	ribosomal large subunit biogenesis and assembly	1.00E-17	13/102 (12.7%)	23/5638 (0.4%)
	30490	processing of 20S pre-rRNA	2.53E-16	15/102 (14.7%)	43/5638 (0.7%)
	30489	processing of 27S pre-rRNA	8.10E-10	7/102 (6.8%)	13/5638 (0.2%)
2	6511	ubiquitin-dependent protein catabolic process	7.43E-06	11/86 (12.7%)	140/5638 (2.4%)
	19941	modification-dependent protein catabolic process	7.43E-06	11/86 (12.7%)	140/5638 (2.4%)
	51603	proteolysis involved in cellular protein catabolic process	8.52E-06	11/86 (12.7%)	142/5638 (2.5%)
7	7017	Microtubule-based process	1.38E-07	13/126 (10.3%)	95/5638 (1.6%)
	7067	mitosis	2.85E-06	13/126 (10.3%)	123/5638 (2.1%)
	16359	mitotic sister chromatid segregation	3.26E-06	9/126 (7.1%)	56/5638 (0.9%)
	7010	cytoskeleton organization and biogenesis	5.30E-06	17/126 (13.4%)	217/5638 (3.8%)
	7059	chromosome segregation	6.09E-06	12/126 (9.5%)	112/5638 (1.9%)
13	19941	modification-dependent protein catabolic process	2.69E-06	10/63 (15.8%)	140/5638 (2.4%)
	51603	proteolysis involved in cellular protein catabolic process	3.06E-06	10/63 (15.8%)	142/5638 (2.5%)
	43632	modification-dependent macromolecule catabolic process	4.18E-06	10/63 (15.8%)	147/5638 (2.6%)

3.3.2 Yeast Dataset

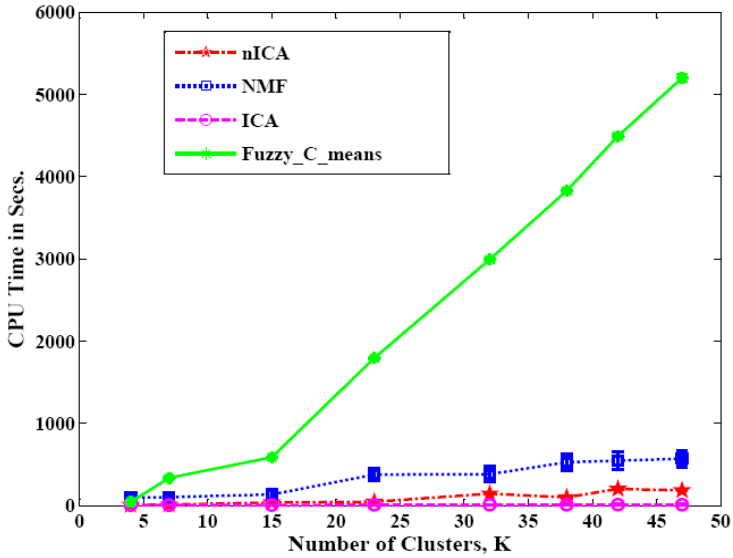
The yeast dataset, which exhibits highly coordinated metabolic fluctuations, gene expression patterns and cell division cycles, was cultured under diverse experimental conditions, temperature shocks, amino acid starvation, and progression into stationary phase [120]. This dataset has been extensively studied because of its importance in a variety of biotechnological applications. As in [43], we also used KNNimpute to fill in the missing values [125]. Due to the triviality of clustering environmental stress response (ESR) genes defined by [120], we eliminated them in our analysis. The final data set contains 5284 genes and 173 samples.

To objectively evaluate the clustering results from different methods, we used the ClusterJudge introduced in [126] to conduct a comparative study. As described in [126], ClusterJudge produced a table according to parsed annotation from SGD of *S. cerevisiae* genes (~6300) with GO attributes (~2000). If the gene i is known to possess attribute j, then in the table there is a 1 in position (i, j), otherwise, it will be a 0 indicating the lack of knowledge about whether gene i possesses attribute j. With this gene-attribute table, ClusterJudge further construct a contingency table. For each item in the contingency table, the number represents the mutual information as a sum of mutual information between each cluster–attribute_k pair (C_attr_k). Afterwards, it scored a partitioning as follows: (1) compute mutual information for the clustered data; (2) compute mutual information for a random clustering, repeating until a distribution of values is obtained; and (3) compute a z-score for real mutual information and the distribution of random values. A larger z-score indicates clustering results more significantly related to gene function. In this case study, we focus on comparing the result from the nICA approach, which is enforced by the non-negativity constraint, with that from conventional ICA approach [43] and NMF (another nonnegative matrix decomposition approach which assumes that both regulator expression and regulation strength are nonnegative) [60, 111]. Also, the nICA approach belongs to the category of soft clustering methods, which can assign a gene to several clusters. Hence, we compare it to another soft clustering algorithm – fuzzy c-means algorithm [112, 127].

We compared the clustering results of nICA, NMF, ICA and Fuzzy c-means from small to larger cluster numbers; the z-scores are shown in Figure 3.5(a). Figure 3.5(a) demonstrates that nICA consistently outperformed ICA with an average increase of z-score of 10. The NMF performed slightly better that nICA when the number of cluster is small, while nICA performed slightly better than NMF when the number of cluster becomes large. In our opinion, the overall performances of nICA and NMF are comparable. nICA also exceeds fuzzy c-means in terms of z-scores for every different cluster number. In order to compare the computational complexity, we collected the CPU times required by the methods, which are shown in Figure 3.5(b). The experiments were performed in MATLAB 7.0.4.365 (R14) on a Pentium D 3.4-GHz PC with 2.00-GB of RAM. Note that fuzzy c-means is computationally cumbersome, since the size of the membership matrix used by the algorithm grows as the product of the data set size and the number of membership classes [128], while the other three matrix-decomposition methods have much less computational complexity. Finally, in Table 3.2, we list five of the identified co-regulated gene groups that show significant enrichment in GO term categories.



(a)



(b)

Figure 3.5: Comparison of clustering results obtained by nICA (pentagram), NMF (square), ICA (circle) and fuzzy c-means (asterisk) respectively.

Table 3.2 The five most significant clusters from nICA for the yeast data set.

Cluster ID	GOID	GO term	p-value	cluster frequency	total frequency
13	4386	helicase activity	1.10E-10	11/ 47 (23.4%)	82/7288 (1.1%)
15	19752	carboxylic acid metabolic process	4.83E-15	17/28 (60.7%)	308/7288 (4.2%)
	6519	amino acid and derivative metabolic process	7.39E-15	15/28 (53.6%)	200/7288 (2.7%)
	6807	nitrogen compound metabolic process	1.49E-13	15/28 (53.6%)	244/7288 (3.3%)
	6144	purine base metabolic process	7.41E-12	7/28 (25.0%)	16/7288 (0.2%)
	103	sulfate assimilation	4.56E-11	6/28 (21.4%)	10/7288 (0.1%)
	6555	methionine metabolic process	1.56E-10	7/28 (25.0%)	23/7288 (0.3%)
16	6807	nitrogen compound metabolic process	1.04E-17	24/62 (38.7%)	244/7288 (3.3%)
	6519	amino acid and derivative metabolic process	1.99E-14	20/62 (32.3%)	200/7288 (2.7%)
18	32197	transposition, RNA-mediated	7.19E-11	13/57 (22.8%)	95/7288 (1.3%)
	3964	RNA-directed DNA polymerase activity	5.25E-10	10/57 (17.5%)	52/7288 (0.7%)
22	6091	generation of precursor metabolites and energy	2.34E-20	25/ 44 (56.8%)	336/7288 (4.6%)
	6119	oxidative phosphorylation	4.02E-17	13/ 44 (29.5%)	46/7288 (0.6%)
	6732	coenzyme metabolic process	1.55E-13	15/44 (34.1%)	135/7288 (1.9%)
	42775	organelle ATP synthesis coupled electron transport	1.91E-12	9/44 (20.5%)	25/7288 (0.3%)
	51186	cofactor metabolic process	4.28E-12	15/ 44 (34.1%)	168/7288 (2.3%)
	15980	energy derivation by oxidation of organic compounds	5.46E-12	18/44 (40.9%)	298/7288 (4.1%)
	6084	acetyl-CoA metabolic process	2.43E-11	8/44 (18.2%)	20/7288 (0.3%)
	9060	aerobic respiration	1.77E-10	11/ 44 (25.0%)	80/7288 (1.1%)

3.3.3 Muscle Regeneration Data

We further applied the nICA approach to a time course microarray data set from a profiling study of *in vivo* murine muscle regeneration. Staged skeletal muscle degeneration/regeneration was induced by injection of cardiotoxin (CTX) as described [121]. Mice were injected in gastrocnemius muscles of both sides, and then sacrificed at the following 27 time points: 0h(our), 12h, 1d(ay), 2d, 3d, 3.5d, 4d, 4.5d, 5d, 5.5d, 6d, 6.5d, 7d, 7.5d, 8d, 8.5d, 9d, 9.5d, 10d, 11d, 12d, 13d, 14d, 16d, 20d, 30d, and 40d [121]. Expression profiles were obtained with Affymetrix's U74Av2 and MAS5.0 summarization algorithm. As a preprocessing step, we used the last time point as the reference point and the expression matrix consists of log-ratios of the expression measurements with respect to the reference point. We then applied the nICA approach to the positive and negative parts respectively for gene module identification. As a result, we found 11 clusters from the positive part of the data and 9 clusters from the negative part, all with significant biological coherence. Several clusters showed expression patterns highly correlated with *MyoD1* gene (Figure 3.6 shows an example of the heatmap of cluster 8 from the positive part of the data). *MyoD1* has been widely studied for its important function in embryonic myogenesis and postnatal muscle regeneration. We also examined the biological relevance of these clusters. The results are shown in Table 3.3 and Table 3.4 (p-value less than 10^{-4} is the cut off for significance).

With the identified gene clusters, we further used the Ingenuity Pathway Analysis (IPA) [129] to assess their biological plausibility with respect to known information about gene regulatory networks, pathways and module functions. We found two clusters whose network functions are tightly related to skeletal and muscular system development (Figure 3.7). Moreover, cluster 13 found in the negative part contains *Rb1* and cluster 1 found in the negative part contains *MyoD1*, which are two novel downstream targets of *MyoD* [130] and one of muscle dystrophies - EDMD's prominent signature. By reason of the failure of the interactions between the nuclear envelope and *Rb* and *MyoD* at the point of myoblast exit from the cell cycle, it may lead to poorly coordinated phosphorylation and acetylation step.

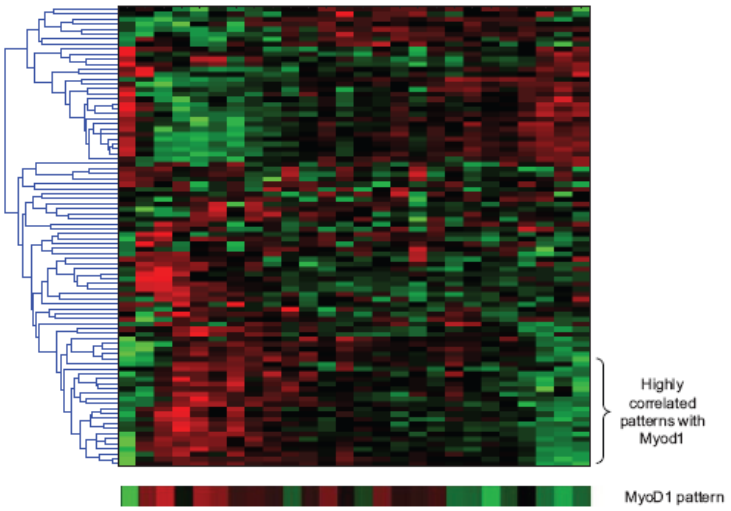


Figure 3.6: The heatmap of the cluster 8 from the positive part of muscle regeneration data, showing a highly correlated expression pattern with *MyoD1* gene.

Table 3.3 The five significant clusters from nICA for the muscle regeneration data set (positive part).

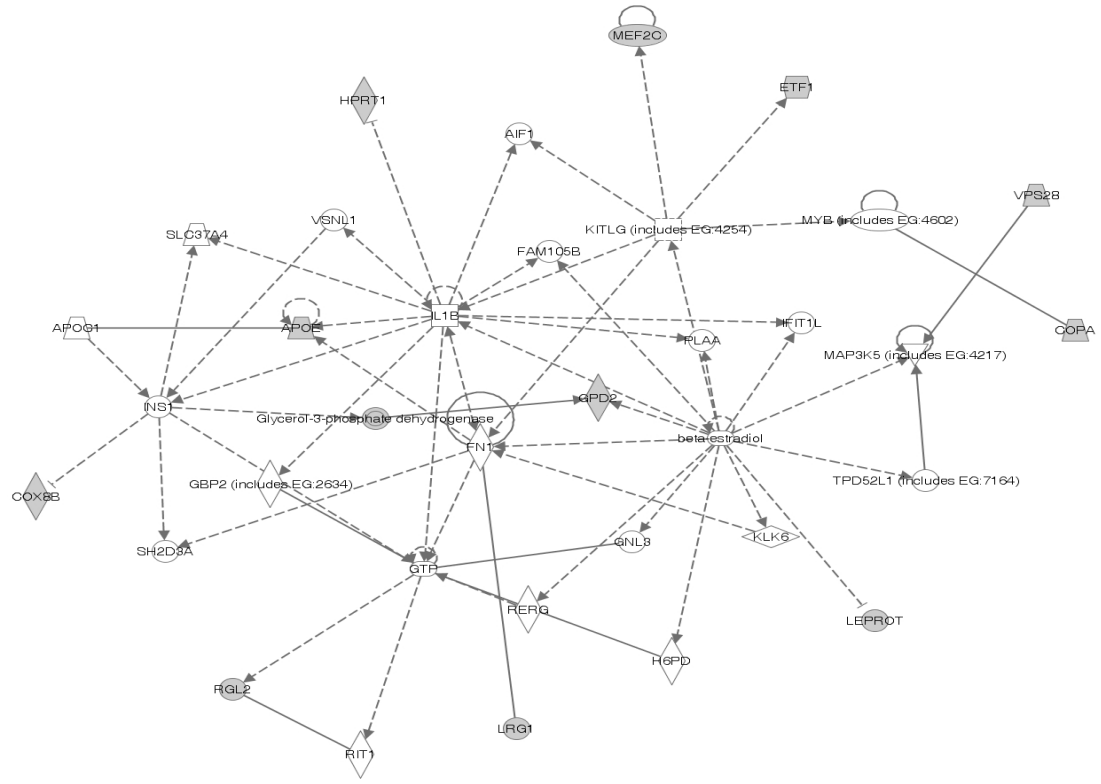
Cluster ID	GOID	GO_term	p-value	cluster frequency	total frequency
2	7156	homophilic cell adhesion	1.56E-16	22/235 (9.3%)	142/15873 (0.8%)
	16337	cell-cell adhesion	7.87E-12	22/235 (9.3%)	238/15873 (1.4%)
6	7399	nervous system development	1.34E-04	7/29 (24.1%)	651/15873 (4.1%)
	30900	forebrain development	2.12E-04	3/29 (10.3%)	64/15873 (0.4%)
	48731	system development	2.23E-04	7/29 (24.1%)	707/15873 (4.4%)
	48856	anatomical structure development	4.11E-04	11/29 (37.9%)	1962/15873 (12.3%)
8	6886	intracellular protein transport	1.19E-04	10/81 (12.3%)	463/15873 (2.9%)
	6091	generation of precursor metabolites and energy	1.82E-04	12/81 14.8%	688/15873 (4.3%)
11	6334	nucleosome assembly	7.26E-14	11/49 (22.4%)	121/15873 (0.7%)
	31497	chromatin assembly	2.68E-13	11/49 (22.4%)	136/15873 (0.8%)

	6325	establishment and/or maintenance of chromatin architecture	2.06E-09	11/49 (22.4%)	311/15873 (1.9%)
	6461	protein complex assembly	2.43E-09	11/49 (22.4%)	316/15873 (1.9%)
	6323	DNA packaging	2.77E-09	11/49 (22.4%)	320/15873 (2.0%)
	7001	chromosome organization and biogenesis (sensu Eukaryota)	3.10E-08	11/49 (22.4%)	404/15873 (2.5%)
15	6092	main pathway of carbohydrate metabolism	4.99E-10	13/170 (7.6%)	120/15873 (0.7%)
	6096	Glycolysis	2.59E-09	10/170 (5.8%)	68/15873 (0.4%)
	15980	energy derivation by oxidation of organic compounds	4.66E-09	14/170 (8.2%)	172/15873 (1.0%)
	44265	cellular macromolecule catabolic process	1.42E-08	18/170 (10.5%)	327/15873 (2.0%)
	46365	monosaccharide catabolic process	1.49E-08	10/170 (5.8%)	81/15873 (0.5%)
	19320	hexose catabolic process	1.49E-08	10/170 (5.8%)	81/15873 (0.5%)
	46164	alcohol catabolic process	1.89E-08	10/170 (5.8%)	83/15873 (0.5%)

Table 3.4 The nine significant clusters from nICA for the muscle regeneration data set (negative part).

Cluster ID	GOID	GO_term	p-value	Cluster frequency	total frequency
1	44238	primary metabolic process	3.29E-05	82/128 (64.0%)	7330/15873 (46.1%)
	19538	protein metabolic process	5.11E-05	48/128 (37.5%)	3504/15873 (22.0%)
2	6096	Glycolysis	3.99E-06	7/153 (4.5%)	68/15873 (0.4%)
	6007	glucose catabolic process	8.44E-06	7/153 (4.5%)	76/15873 (0.4%)
3	51641	cellular localization	3.69E-07	23/141 (16.3%)	780/15873 (4.9%)
	46907	intracellular transport	2.93E-06	21/141 (14.8%)	751/15873 (4.7%)
4	7156	homophilic cell adhesion	1.71E-20	22/155 (14.1%)	142/15873 (0.8%)

	16337	cell-cell adhesion	1.38E-15	22/155 (14.1%)	238/15873 (1.4%)
6	15992	proton transport	4.78E-05	5/83 (6.0%)	76/15873 (0.4%)
	6818	hydrogen transport	6.89E-05	5/83 (6.0%)	82/15873 (0.5%)
8	16043	cellular component organization and biogenesis	2.67E-11	63/199 (31.6%)	2146/15873 (13.5%)
	46907	intracellular transport	2.72E-10	33/199 (16.5%)	751/15873 (4.7%)
	51649	establishment of cellular localization	4.37E-10	33/199 (16.5%)	765/15873 (4.8%)
	6996	organelle organization and biogenesis	7.31E-07	36/199 (18.0%)	1197/15873 (7.5%)
	6886	intracellular protein transport	1.55E-06	20/199 (10.0%)	463/15873 (2.9%)
9	6457	protein folding	6.63E-07	19/331 (5.7%)	239/15873 (1.5%)
12	44260	cellular macromolecule metabolic process	2.48E-06	62/174 (35.6%)	3256/15873 (20.5%)
	44267	cellular protein metabolic process	7.48E-06	60/174 (34.4%)	3212/15873 (20.2%)
13	31497	chromatin assembly	2.43E-13	16/153 (10.4%)	136/15873 (0.8%)
	6334	nucleosome assembly	1.02E-11	14/153 (9.1%)	121/15873 (0.7%)
	6333	chromatin assembly or disassembly	1.23E-11	16/153 (10.4%)	175/15873 (1.1%)



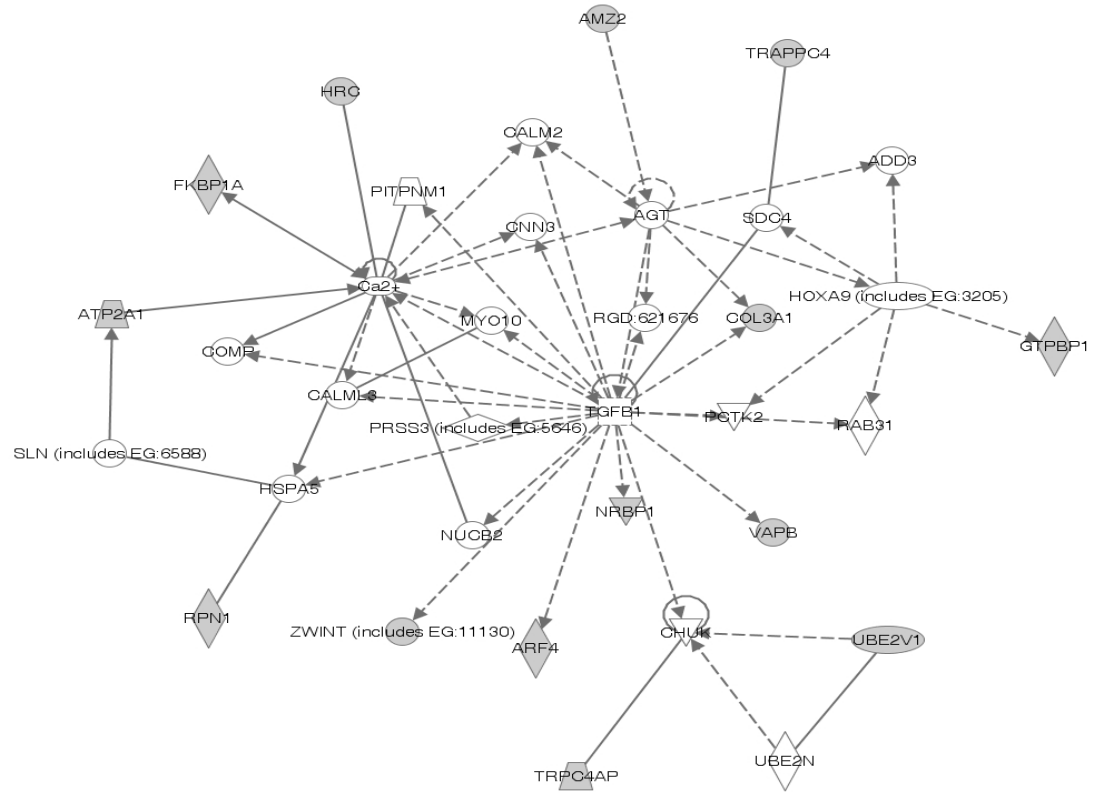


Figure 3.7: The first panel shows cluster 6 found in the positive part of the muscle regeneration data with the following functions: "post-translational modification", "cellular growth and proliferation", and "skeletal and muscular system development"; the second panel shows cluster 6 found in the negative part of the data with the main function of "skeletal and muscular system development". The analysis results were generated through the use of IPA (Ingenuity® Systems, www.ingenuity.com).

3.4 Conclusion

We present a new gene clustering approach, namely a nICA-based approach, for composite gene module discovery. A complete algorithm of the nICA approach has been developed with the following main components: (1) input sample selection, (2) stability-based dimension estimation, (3) nICA learning algorithm, and (4) gene clustering by VISDA. Specifically, we perform input variable selection to improve the quality of separation of the sources. We then develop a stability analysis procedure to determine the number of nonnegative independent components. We implement a learning algorithm of nICA with the non-negativity constraint. Finally, we use VISDA, a data visualization and clustering tool, to group the genes into modules in the latent variable space. By projecting the gene expression data onto nICA latent space, co-regulation structures of the modules can be revealed and highlighted. Using a pre-screening and VISDA clustering procedure, we can identify biological process enriched clusters with coherent functional annotations.

ICA is a compelling signal processing technique for revealing hidden factors from multivariate statistical data. Lee *et al.* [43] in 2003 analyzed the performance of different kinds of ICA algorithms for microarray data and showed that ICA outperforms PCA, *k*-means clustering, and the Plaid model. In order to suit the corresponding putative biological processes with the characteristic of the positive nature of gene expression levels, we did further studies to better understand and extract the underlying nonnegative independent biological expression profiles in microarray data. We conclude that nICA is quite effective in extracting nonnegative independent biological processes when compared with ICA. The experimental results on the yeast data sets have demonstrated the advantages of the nICA approach over conventional ICA-based approach. Another matrix-decomposition-based approach – NMF is a useful technique in approximating high-dimensional data as well. However, it has its own limitations [131]. First of all,

uniqueness and robust computations are missing in the NMF. Unlike nICA, there is no unique global minimum for the NMF, so the algorithms can only guarantee convergence to a local minimum, and many do not even guarantee that. Our experiments also confirmed that when the cluster number is small, NMF is a very powerful method to catch the underlining data structure. However, its performance decreases sharply when the data structures are complex and the number of clusters increases. The results also indicated that the performances of nICA are slightly better than fuzzy c-means clustering. There are several possible reasons to explain this. The fuzzy c-means approach to clustering suffers from several constrains that affect performance [132]. The main drawback is that it tends to give high membership values for the outlier points, which is unsuitable for noisy microarray data. Secondly, due to the influence (partial membership) of all the data members, the membership of a data point in a cluster depends directly on its membership values in other cluster centers, and this sometimes happens to produce unrealistic results. Further, the nICA approach has been applied to a muscle regeneration data set for novel gene module discovery. The results have shown that not only the identified gene modules are biologically significant and plausible, but novel downstream target genes can also be discovered by the nICA approach.

4 Motif-guided Sparse Decomposition of Gene Expression Data for Regulatory Module Identification

4.1 Introduction

Transcriptional gene regulation is a complex process that utilizes a network of interactions to jointly pattern gene expression [133]. The accurate identification of transcriptional modules or gene sub-networks involved in the regulation of critical biological processes remains a central problem in genomic research[134]. For cancer research, these sub-networks can help provide a signature of the disease that is potentially useful either for diagnosis or for suggesting novel targets for drug intervention. To date, the abundance of literature and databases that contain sequence information, gene expression profiling studies, and small scale biological experiments allows investigators to reconstruct gene regulatory networks so as to quantify the direct effects of transcription factors on gene expression.

4.1.1 Computational Approaches for Modeling Gene Regulatory Networks

Recently, the bioinformatics community has explored various computational approaches for transcriptional module identification [8, 9, 11, 135, 136]. These approaches can be classified into two major categories: the first category uses clustering methods to explore the similarity in gene expression pattern to form gene modules, whereas the second category uses projection methods to infer latent (hidden) components with which to group genes into modules.

There is a growing literature documenting attempts to reconstruct gene networks by applying clustering methods [127, 137] and their more sophisticated variants such as statistical regression [138] and Bayesian networks [109]. While this line of work is important to help formulate hypotheses regarding yet unexplored mechanisms, there are many limitations for using clustering methods to infer regulatory modules. One common challenge is to detect the interactions between transcription factors and their target genes based on gene expression data alone. For regulatory module identification, it is critical to distinguish 'co-regulation' from 'co-expression', and to understand the relationship between co-regulation and co-expression. Generally speaking, genes with highly homologous regulatory sequences (*i.e.*, co-regulation) should have a similar expression pattern (*i.e.*, co-expression), but the reverse is likely not true [139]. Therefore, traditional clustering analysis often returns clusters lacking shared regulatory sequences, thus making the biological relevance of these clusters relatively low for the identification of regulatory mechanisms.

The shortcoming of the methods in the first category is partially addressed by the methods in the second category. A group of projection methods from the second category (e.g., principle component analysis (PCA); independent component analysis (ICA); non-negative matrix factorization (NMF)) [43, 82, 140] have also been extensively applied for transcriptional module identification. These methods decompose gene expression data into components that are constrained to be mutually uncorrelated or independent, and then cluster genes based on their loading in the components. Since these methods do not cluster genes based on their expression similarity, they are better equipped to find co-regulated gene modules. However, one major difficulty using such approaches is that the components usually represent joint effects of many underlying transcription factors; in other words, the components do not correspond to individual known transcription factors (TFs), thus making the biological interpretation of the components very difficult.

To overcome the above-mentioned shortcomings, several integrative methods have been proposed that integrate TF-gene interaction data with gene expression data. For instance, network component analysis (NCA) has been recently developed to successfully estimate the TF activities of regulatory networks using both ChIP-on-chip and gene expression data [8]. Note that the NCA approach heavily relies on ChIP-on-chip data for

network connectivity information with which to define regulatory modules. Therefore, the NCA scheme is not readily applicable to many biological studies where adequate network connectivity information is not available, due to lack of complete ChIP-on-chip data. To deal with this difficulty, Sabatti and James [141] were the first to use motif information as initial network topology, and they subsequently adopted a Bayesian algorithm to reconstruct regulatory modules. Although theoretically elegant, the approach needs to estimate the posterior probability, a joint distribution of network topology and transcription factor activity. Even using the Gibbs sampling technique, it is a formidable task to estimate the joint distribution when the number of samples is limited.

4.1.2 Linear Latent Modeling for Gene Regulatory Networks

Considering an illustrative regulatory network depicted in Figure 4.1, the problem of interest is to reconstruct the latent signals TF_1, …, TF_n and the underlying regulatory structure from observations. We adopt a latent variable model that has been used by Liao *et al*. [8] and Kao *et al*. [142] to establish a link between gene expression data and motif information.

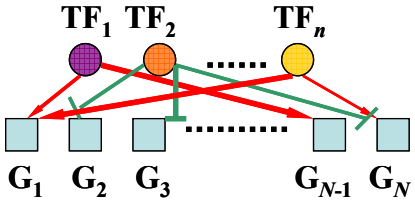


Figure 4.1: An illustrative gene regulatory network with n transcription factors (TFs) (Circle) and N regulated genes (Rectangle). Red arrows indicate TFs activate their target genes. Green arrows show that genes are repressed by TFs. We use different line widths to indicate different regulation strengths between TFs and target genes.

The central theme of the model is that gene expression measurements can be largely determined by the unknown activities of transcription factors acting on binding motifs, which is consistent with the ICA model I that we have described in Chapter 3. In particular, using log-ratios of gene expression measurements, a simplified yet biologically justified linear model can be formulated as follows [8] :

$$x_{pg} = \sum_t a_{pt} \cdot s_{tg} \quad \text{or} \quad \mathbf{X} = \mathbf{AS}, \tag{4.1}$$

where x_{pg} is defined as the logarithm of expression ratio of gene g between sample p and control sample, a_{pt} the activity level of TF t in sample p and s_{tg} the regulation strength of TF t onto gene g. A derivation of this approximate representation from a transcriptional network is described by Hill's equations in [8]. Specifically, the relationships between transcription factors' activities and their regulated genes' expression profiles are approximated by a log-linear model:

$$\frac{E_{pg}}{E_{0g}} = \prod_{t=1}^{T} \left(\frac{TFA_{pt}}{TFA_{0t}} \right)^{s_{tg}}, \tag{4.2}$$

where E_{pg} and E_{0g} are expression levels of gene g in sample p and control sample, respectively, and $x_{pg} = \ln(E_{pg}/E_{0g})$. TFA_{pt} and TFA_{0t} are transcription factor t's activities in sample p and control sample, respectively, and $a_{pt} = \ln(TFA_{pt}/TFA_{0t})$. s_{tg} is the regulation strength of transcription factor t on gene g.

This formulation (Eq. (4.1)) has of the following advantages. First, it enables a direct application of well-developed linear system theory to capture local accuracy and obtain computationally tractable results. Second, it permits a certain nonlinear relationship between TFs and their target genes. However, there is a shortcoming associated with this model that needs to be pointed out, namely, that the linear model of gene expression upon which NCA and our method rest, as an initial step to reverse-engineering transcriptional modules, does not account for the interaction between transcription factors.

In this chapter, we propose a novel approach, namely motif-guided sparse decomposition (mSD), to identify co-regulated transcriptional modules by integrating motif information and gene expression data. The mSD method is a Bayesian-principled method without the need to estimate the joint distribution. Instead, in Section 4.2 a two-step approach is used to first estimate transcription factor activity and then regulation strength on the target genes. A motif-guided clustering method is developed to help estimate transcription factor activity by taking into account both co-expression and co-regulation. A sparse decomposition step is followed to estimate the regulation strength of regulatory networks. To evaluate the performance of the proposed approach, we applied the mSD method to simulated and real yeast cell cycle data in Section 4.3, showing an improved performance in identifying three kinds of coherent modules associated with

known cell cycle transcription factors. We then applied our approach to an estrogen-dependent profiling study of breast cancer to recover condition-specific transcriptional modules related to estrogen signaling and actions. The results demonstrated that our approach can effectively find important condition-specific regulatory modules that are functionally relevant to estrogen signaling pathways. Finally, we gave some conclusions in Section 4.4. We expect that the proposed approach can make a useful contribution to the reconstruction of gene regulatory networks from diverse sources of genomic data.

4.2 Methods of mSD

The overall scheme of the proposed mSD approach is illustrated in Figure 4.2. We start with the extraction of motif information from upstream DNA sequences of genes, followed by a two-stage approach to integrate motif information and gene expression data for regulatory module identification. In the first stage, we use a motif-guided clustering method for transcription factor activity estimation by maximizing the motif support for co-expressed gene modules. In the second stage, we use a sparse decomposition method for regulation strength estimation to reinforce that the genes in a module are likely regulated by a few transcription factors. Finally, regulatory modules are reconstructed from the detected active regulators and their target genes that exhibit large regulation strengths. Next, we give a detailed description of each major component in the mSD approach.

Analogously to previous studies, we assume that the log-ratios of gene expression $\mathbf{X} \in \mathbf{R}^{m \times N}, (N >> 1)$ are expressed as a linear combination of log-ratios of TF activities ($\mathbf{A} \in \mathbf{R}^{m \times n}$) weighted by their regulation strength ($\mathbf{S} \in \mathbf{R}^{n \times N}$). Note that m is the number of samples, N the number of genes and n the number of TFs.

In general, the number of TFs is much smaller than the number of transcribed genes ($n << N$) and most genes are regulated only by a small number of TFs. Hence, the matrix **S** that describes the regulation strength between the TFs and their regulated genes is sparse. Further, the number of TFs (n) is usually greater than the number of samples (m), *i.e.*, $n > m$, such that Eq. (4.1) represents an underdetermined linear system (ULS). To

obtain a sparse solution to this ULS, we develop a two-stage approach to estimate transcription factor activity (**A**) and regulation strength (**S**) sequentially.

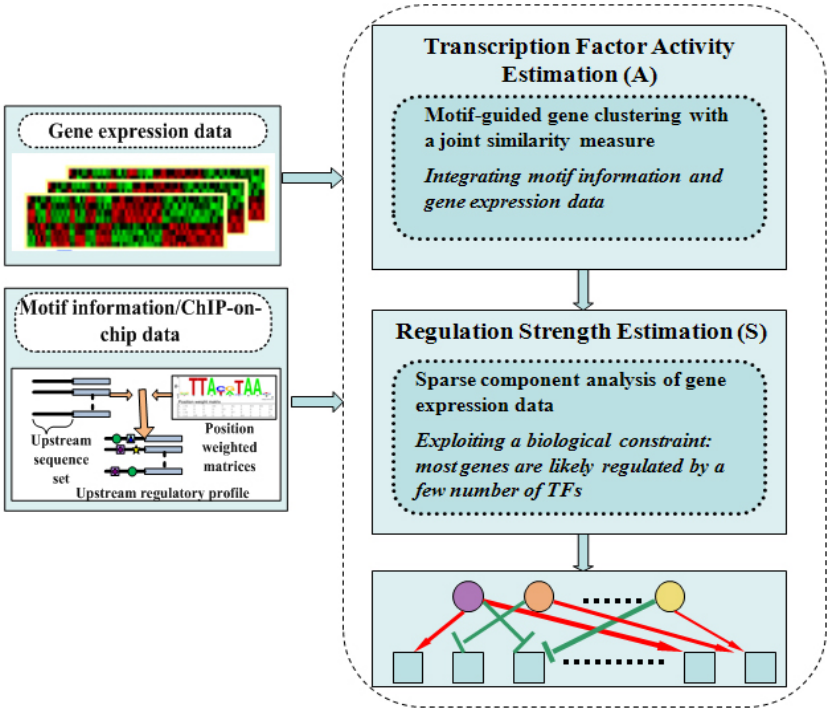


Figure 4.2: A block diagram of the motif-guided sparse decomposition (mSD) approach.

4.2.1 Transcription Factor Activity Estimation

In order to reliably estimate **A** matrix from Eq. (4.1), the sparsity property of the regulation strength matrix **S** is very important. To allow for a unique mathematical decomposition, some identifiability criteria need to be satisfied [54]: 1) The matrix $\mathbf{A} \in \mathbf{R}^{m \times n}$ has the property that any square $m \times m$ sub-matrix of **A** is nonsingular; 2) The matrix **S** is sparse of level $n - m + 1$, *i.e.,* each column of **S** has at least $n - m + 1$ zero elements.

In fact, the following theorem is the key to obtaining a reliable estimation of **A** [54].

Theorem 1: (Identifiability Conditions - Locally Very Sparse Representation): Assume that the number of transcription factors (TFs) is unknown and the following conditions are met: 1) Each TF has at least two strictly Well-Grounded Points (sWGPs),

which means that for each index $i = 1, \cdots, n$, there are at least two columns of **S**: $\mathbf{s}_{j_1}$ and $\mathbf{s}_{j_2}$ that have nonzero elements only in position *i* (so each TF is uniquely present at least twice); 2) Mutual non-collinearity of columns of **X**: $\mathbf{x}_k \neq c\mathbf{x}_q$ for any $c \in \mathbf{R}$, any $k = 1, \cdots, N$ and any $q = 1, \cdots, N$, $k \neq q$ for which $\mathbf{s}_k$ (the *k*-th column of **S**) has more than one nonzero element. Then **A** is uniquely determined by **X** except for left multiplication with a permutation and scaling matrix. For proofs of the theorem we refer to [54].

A generic approach for transcription factor activity estimation is to use a clustering method to find representative genes whose expression profiles, *i.e.*, columns of **X**, can be utilized to estimate **A** [54]. Many clustering techniques have been proposed to cluster gene expression data, such as *k*-means clustering [143] and self-organizing maps (SOM) [40], which are designed to find gene expression patterns by grouping the genes with similar expression profiles. Very recently, an affinity propagation clustering (APC) algorithm has been proposed for data clustering that showing an improved performance [144]. Based on an *ad hoc* pair-wise similarity function between data points, APC seeks to identify each cluster by one of its elements, the so-called *exemplar*. APC takes as input a collection of real-valued similarities between data points, where the similarity *s*(*i*, *k*) indicates how well data point *k* is suited to be the exemplar for data point *i*. The goal is to maximize the similarity *s*(*i*, *k*) or equivalently, to minimize the Euclidean distance [144], $d(i,k) = \|\mathbf{x}_i - \mathbf{x}_k\|^2$, where $\mathbf{x}_i$ and $\mathbf{x}_k$ are two column vectors of gene *i* and gene *k*, respectively, in **X**.

However, direct application of the APC clustering technique to gene expression data will only give rise to co-expressed gene clusters. To identify gene regulatory modules, we need a clustering technique to integrate motif information and gene expression data, aiming to find co-regulated gene clusters with co-expressed patterns. We here propose a motif-guided clustering method to find a group of genes that not only is of similar expression pattern but also shares a common set of binding motifs as much as possible.

4.2.1.1 Motif-guided gene clustering with a joint similarity measure

To incorporate motif information, we propose a new similarity measure, taking into account both expression similarity and motif binding similarity, for the APC clustering method. The motif information can be represented by a TF-gene binding strength matrix, $\mathbf{W} = [w(t,g)]$, considering a set of n TFs binding onto a set of N genes. Each element of **W**, *i.e.*, $w(t,g)$, denotes the binding strength of TF t onto gene g. As a common practice, the binding strength is usually approximated 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 [145]. Here, we should point out that we do not perform motif discovery as part of our learning procedure, but rather assume that we have a list of motifs for putative transcription factor binding sites (TFBSs) by searching a database of regulatory elements such as TRANSFAC [146]. Our learning algorithm only inputs validated transcription factor binding sites (TFBSs) or motifs that allow for a straight-forward biological interpretation, which facilitates biologists to decipher the function of genes being regulated under a given experimental condition. In this chapter, all human promoter DNA sequences were obtained from the UCSC Genome database [147]; in particular, upstream 5,000 base pair (bp) from the transcription start site (TSS) was obtained. With all vertebrate PWMs provided by the TRANSFAC 11.1 Professional Database [146], Match^{TM} [25] algorithm was used to generate a gene-motif binding strength matrix with the cut offs that minimize the false-positive rate. Each element of gene-motif binding strength $w(t,g)$ represents the binding strength of motif t in the promoter region of a gene g.

Given the binding strength of TF t onto gene i ($w(t,i)$) and that of TF t onto gene k ($w(t,k)$), the joint binding strength of TF t onto both gene i and gene k is proportional to $w(t,i) \times w(t,k)$, assuming that these two binding events are independent. Thus, for all possible TFs (*i.e.*, TF $t,\ t = 1,...,n$) binding onto gene i and gene k, it is reasonable to use the sum of their joint binding strengths to measure the likelihood of gene i and gene k being co-regulated by the possible set of TFs (*i.e.*, TF $t,\ t = 1,...,n$):

$$s_m(i,k) = \sum_{t=1}^{n} w(t,i) \times w(t,k) \ . \tag{4.3}$$

For motif-guided clustering, we propose the following pair-wise similarity measure to simultaneously consider the binding motif likelihood and gene expression similarity:

$$d(i,k) = -(1-\lambda)\left\|\mathbf{x}_i - \mathbf{x}_k\right\|^2 + \lambda s_m(i,k) \ , \tag{4.4}$$

where λ is a trade-off parameter that controls the contribution from two different information sources, *i.e.*, motif information and gene expression data. When incorporated into an APC clustering method, the first term in Eq. (4.4) is intended to find a group of genes with similar expression pattern while the second term to enforce the genes sharing a common set of TFs to a feasible extent. It is worth noting that the two terms in Eq. (4.4) can also be interpreted as, in a Bayesian framework, conditional likelihoods of expression similarity and motif similarity, respectively.

Given a set of transcription factors (TFs) and a set of genes as depicted in Figure 4.1, we want to infer transcriptional modules or more specifically, the relationship between gene *i* and gene *k*, $d(i,k),\ i,k = 1,...,N$. In a Bayesian framework, the above problem can be restated as follows: given gene expression data (*e*) and motif information (*m*), we try to maximize the posterior probability of $P(d(i,k)|e_{i,k}, m_{i,k})$ for a Bayesian solution to the relationship of gene *i* and gene *k*. Using Bayes rules, we have the following equation to work with:

$$P(d(i,k)|e_{i,k}, m_{i,k}) = P(e_{ik}, m_{ik}|d(i,k))P(d(i,k)) \ , \tag{4.5}$$

where e_{ik} represents the expression similarity between gene *i* and gene *k* and m_{ik} represents motif similarity between gene *i* and gene *k*. Our goal is to find the model $\mathcal{M}$ with the pair-wise gene relationships $d(i,k),\ i,k = 1,...,N$ that maximizes the posterior probability in Eq. (4.5). In practice, the Bayesian solution to the above problem can be obtained by maximizing the following log-likelihood score:

$$\ell(d(i,k)) = \log[P(e_{ik}, m_{ik}|d(i,k))] \ . \tag{4.6}$$

Without loss of generality, assuming conditional independence between the expression similarity and motif similarity for a given relationship between gene *i* and gene *k* leads to the following equation:

$$\begin{aligned}\text{maximize } \{\ell(d(i,k))\} &= \text{maximize}\{\log[P(e_{ik}\,|\,d(i,k))\,P(m_{ik}\,|\,d(i,k))]\} \\ &= \text{maximize}\{\log[P(e_{ik}\,|\,d(i,k))]\} + \text{maximize}\{\log[P(m_{ik}\,|\,d(i,k))]\}.\end{aligned} \tag{4.7}$$

For the first term in Eq. (4.7), we can model the expression levels of genes in a cluster as random samples drawn from a Gaussian distribution [92], and thus use Euclidean distance ($\|\mathbf{x}_i - \mathbf{x}_k\|^2$) to measure the expression similarity between two genes. For the second term in Eq. (4.7), since it is hard to obtain the exact distribution of motif similarity, we propose to use the joint binding strength of TF *t* onto both gene *i* and gene *k*, $s_m(i,k)$, as a measure of motif similarity. Therefore, we propose joint similarity measure in Eq. (4.4) to count the contribution from two different biological sources.

Ideally, the clustering result will generate a better representation of the transcription factor activity that underlies a co-regulated group of genes. However, both motif information and gene expression data are quite noisy due to the nature of binding motif being very short DNA sequence [88] and low signal-to-noise (SNR) ratio in gene expression measurements [148, 149]. The impact of the noises can be clearly observed in two extreme cases: (1) the gene cluster resulting from (noisy) motif information alone will show a noisy expression pattern; and (2) the cluster resulting from gene expression data alone will often gain little support in terms of being regulated by a shared set of motifs. Therefore, it is important to understand the contribution of each data source and assign its proper weight, *i.e.*, the trade-off parameter λ in Eq. (4.4), aiming to alleviate the noise impact. In the following section, we will design an entropy-based measure in conjunction with a non-uniformity measure to help find the optimal value for the trade-off parameter λ.

4.2.1.2 Determination of the trade-off parameter λ

To measure the relative contribution of motif information to gene clustering, we propose an entropy-based measure to capture the essence of that a regulatory module shall be regulated by a unique set of active transcription factors. For each gene cluster, an enrichment analysis will be first performed to identify the significant motifs associated with the genes in the cluster. Specifically, a hyper-geometric test is designed to calculate

the significance value (*i.e.*, *p*-value) of a motif (e.g., motif *t*) enriched in the cluster. The testing procedure can be described as follows. The null distribution is generated by randomly sampling the entire gene population (with *N* genes) as many times as possible (approximately 10,000 times) to form random gene clusters. Let us assume that the gene cluster *j* under examination consists of N_j genes in which N_b genes have the support of motif *t*, while in the entire gene population the total number of genes that contain the motif *t* in their promoters is N_B. For the randomly generated clusters (each with a size of N_j), we count the number of genes containing motif *t* in each cluster, denoted as i_r, to finally form the null distribution. Then, the *p*-value for motif *t* enriched in cluster *j* can be calculated as follows:

$$p_{jt} = P(i_r \geq N_b) = \sum_{i_r = N_b}^{\min(N_B, N_j)} \frac{\binom{N_B}{i_r}\binom{N - N_B}{N_j - i_r}}{\binom{N}{N_j}} \tag{4.8}$$

With the *p*-value for each motif's enrichment, we calculate the motif emission frequency [150] for all the motifs in each cluster. For a particular cluster index *j*, $j = 1,...,J$, a set of motif frequencies can be defined as $\boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2}, \ldots, \theta_{jn})$, where $\theta_{jt} = -\log_{10} p_{jt}$, $t = 1,...,n$, and p_{jt} is the *p*-value obtained from Eq. (4.8). We then normalize $\boldsymbol{\theta}_j$ by $\sum_{t=1}^{n} \theta_{jt} = 1$ to ensure that each element in $\boldsymbol{\theta}_j$ falls in the range of [0, 1]. Treating this motif occupancy as a random variable associated with an appropriate probability space, we can quantitatively measure the 'uncertainty' of motif occupancy in cluster *j,* from an information-theoretic point of view, by the following entropy definition [151]:

$$H(\boldsymbol{\theta}_j) = -\sum_{t=1}^{n} \theta_{jt} \log_2(\theta_{jt}) \,. \tag{4.9}$$

The entropy is then normalized to be in the range of [0, 1] as divided by the maximum entropy, $H_{\max}(\boldsymbol{\theta}_j)$), *i.e.*, $\tilde{H}(\boldsymbol{\theta}_j) = H(\boldsymbol{\theta}_j) / H_{\max}(\boldsymbol{\theta}_j)$; the maximum entropy is achieved when the motif occupancy is uniformly distributed. Summing over all the clusters, we can obtain the mean entropy to measure the overall 'uncertainty' of motif occupancy in the clusters as follows:

$$\bar{H} = \frac{1}{J}\sum_{j=1}^{J}\tilde{H}(\boldsymbol{\theta}_j) . \tag{4.10}$$

Conceptually, when motifs are randomly distributed (with an assumed uniform distribution) among the clusters, the mean entropy reaches its maximum; in contrast, when motifs are uniquely distributed for each cluster (*i.e.*, cluster-specific), the mean entropy reaches its minimum.

To measure the relative contribution of gene expression data to gene clustering, we adopt a non-uniformity measure [152] to characterize the co-expression nature of the genes in a cluster. The non-uniformity of expression pattern is measured as proportional to the variance of gene expression weighted by an appropriate weighting factor, as shown in the following equation:

$$NonU = \sum_{j=1}^{J}\frac{w_j\sigma_j^2}{\sigma_{\max}^2}, \tag{4.11}$$

where σ_j^2 is the variance of gene expression pattern for cluster j ($j = 1,...,J$), $\sigma_{\max}^2$ is the maximum variance for all clusters, and w_j is the weight of cluster j defined as the proportion of genes to the entire gene population.

By varying the trade-off parameter λ in Eq. (4.4), the APC clustering method will give us different clustering results. As noted earlier, there are noises in both motif information and gene expression data, which will have a profound impact on the clustering results. In particular, when λ is small the contribution from gene expression data dominates, which will give rise to gene clusters with small non-uniformity of expression pattern but large entropy of motif occupancy, *i.e.*, not cluster-specific; in contrast, when λ is large, the contribution from motif information dominates, leading to gene clusters with large non-uniformity but small entropy of motif occupancy, *i.e.*, cluster-specific. Therefore, it is important to find the optimal λ value to alleviate the noise impact on finding regulatory modules. In this dissertation, we propose to use the following cost function to combine the measure of motif occupancy (Eq. (4.10)) and expression pattern (Eq. (4.11)) as follows:

$$C(\lambda) = \bar{H} + NonU . \tag{4.12}$$

Theoretically, the cost function $C(\lambda)$ is a U-shaped function; when λ reaches its optimal value, the cost function $C(\lambda)$ reaches its minimum. In other words, by minimizing $C(\lambda)$

we can find the optimal value of λ to take advantage of both information sources, *i.e.*, motif information and gene expression data, while alleviating the noise impact on gene clustering.

4.2.2 Regulation Strength Estimation

Next, we use the sparse component analysis (SCA) approach [54] to exploit a well-known biological constraint that most genes are likely regulated by a few transcription factors, and then to estimate the regulation strength matrix **S**.

4.2.2.1 Sparseness measure

We first give the definition of sparseness in order to describe the algorithm for regulation strength estimation. In a conceptual manner, sparseness represents an important property of systems that are loosely connected. In numerical analysis a *sparse matrix* is defined as a matrix mostly with zeros. Numerous metrics of sparseness have been proposed and used in the literature to date, such as L_0 norm, Shannon Entropy and Kurtosis [153]. Such measures are mappings from $\mathbf{R}^n$ to $\mathbf{R}$ that quantify how much energy of a vector is packed into a few components. Hoyer [154] proposed a sparseness measure based on the relationship between L_1 norm and L_2 norm on a normalized scale. After some basic re-arrangement, Hoyer's measure is found to be equivalent to the following measure:

$$\frac{\sqrt{n}}{\sqrt{n}-1}-\frac{1}{\sqrt{n}-1}\frac{\|x\|_1}{\|x\|_2}\ , \tag{4.13}$$

which has the following three properties:

1) Eq. (4.13) is bounded between 0 and 1, *i.e.*, $0 \le \text{sparseness}(\mathbf{x}) \le 1$
2) sparseness($c\mathbf{x}$) = sparseness($\mathbf{x}$), *i.e.*, sparseness is scale-invariant
3) $\text{sparseness}(\alpha\mathbf{x}_1 + \beta\mathbf{x}_2) \le \max(\text{sparseness}(\mathbf{x}_1), \text{sparseness}(\mathbf{x}_2))$

   where α, β are arbitrary constants that satisfy: $\text{sgn}(\alpha\mathbf{x}_1(n)) = \text{sgn}(\beta\mathbf{x}_2(n))$. That is, the sparseness of a linear combination of two sparse signals is not larger than the sparseness of the original signals.

The proof of these properties can be found in Appendix A. With these desired properties of the sparseness measure we will provide an optimization algorithm in a projected "active subspace" for regulation strength estimation in the next section.

4.2.2.2 Inferring Regulation Strength Matrix by SCA

After estimating **A**, we employ the sparse component analysis (SCA) approach to estimate the regulation strength matrix **S**, describing the relationships between TFs and gene populations. The matrix **S** consists of weights of *n* TFs in *N* genes, which is a *sparse matrix*. Specifically, we have devised a projected "active subspace" algorithm for regulation strength estimation, which can be described as follows:

(1) Initialize source **S** with the matrix **W** obtained from section 4.2.1.1, which comes from either ChIP-on-chip data or TF-gene binding strength matrix searched from TRANSFAC [146].

Loop

(2) Iterate for every column of **S** (corresponding to each gene)

(a) If sparseness constraints on the current column of **S** (denote $\mathbf{s}_g$) apply, project $\mathbf{s}_g$ to be desired sparse by making its L_1 norm larger than a predefined sparseness threshold, while having the L_2 norm unchanged.

(b) In the projected space, roughly detect which TFs are "active"; the term "active" is used to refer to the TFs with "considerable nonzero" strengths.

(c) For the sake of discussion, we assume that the first q TFs, $\{s_{tg}\}, t = 1,...,q$, have been found to be *inactive*. Find the new estimation of $\mathbf{s}_g$ by minimizing the cost function $\sum_{t=1}^{q} s_{tg}^2$ subject to $\mathbf{x}_g = \mathbf{A}\mathbf{s}_g$.

Until converged

Notice that a major step in the above algorithm (*i.e.,* Step (2a)) requires a projection operator that enforces sparseness by explicitly setting both L_1 norm and L_2 norm. This operator, fortunately, has been found by Hoyer [154] to incorporate the sparseness constraint in the context of nonnegative matrix factorization (NMF). Note that we proved in section 4.2.2.1 that Hoyer's sparseness measure has several favorable properties for

our algorithm. Therefore, we use this projection operator in the SCA approach to find the closest sparse vector $\mathbf{s}_g$ (in the Euclidean sense) with a desired L_1 and L_2 norm. The cost function in Step (2c) is designed to minimize the regulation strength of "inactive" TFs, while letting the regulation strength of "active" TFs to change freely in order to fulfill the imposed constraint $\mathbf{x}_g = \mathbf{A}\mathbf{s}_g$. This can also be viewed as a form of projection into an active subspace [155], resulting in an elegant mathematical approach to obtain the solution to a Karush-Kuhn-Tucker (KKT) system.

For the sparse decomposition of gene expression data, the solution set of $\mathbf{x} = \mathbf{A}\mathbf{s}$, in variable $\mathbf{s}$, defines an affine set in $\mathbf{R}^n$. In fact, the cost function in step (2c) can be reformulated into a quadratic form: $f(\mathbf{s}) = \mathbf{s}^T\mathbf{H}\mathbf{s}$ with $\mathbf{H} = \begin{pmatrix} \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$ where $\mathbf{I}_q$ is the $q \times q$ identity matrix. When the cost function $f(\mathbf{s})$ is strictly convex for all feasible points, it has a unique local minimum that is also the global minimum. A sufficient condition to guarantee the strict convexity of $f(\mathbf{s})$ is for $\mathbf{H}$ to be positive definite [156]. The projection into active subspace finally leads to an elegant solution to our sparse decomposition problem, *i.e.*, the solution to the following Karush-Kuhn-Tucker (KKT) system [155]:

$$\begin{pmatrix} \mathbf{H} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{s} \\ \lambda \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{x} \end{pmatrix}. \tag{4.14}$$

where λ is the $n \times 1$ vector of Lagrange multipliers.

4.3 Results

4.3.1 Synthetic and Real Yeast Data

To validate the proposed integrative approach, we applied the mSD approach to synthetic and real yeast cell cycle data for regulatory module identification, and then compared its performance with those of other approaches such as FastNCA [157] and sparse decomposition [54]. For the synthetic data set, we adopted a network generator, *SynTReN* [158], to produce a benchmark gene expression data set based on a synthetic *S. cerevisiae* transcriptional regulatory network. Specifically, *SynTReN* generated 15 samples of expression data with a set of 345 genes in different conditions. The genome-

wide location data (*i.e.*, ChIP-on-chip data) [11] were then used to provide the binding information and integrated with gene expression for transcription factor activity estimation and regulation strength estimation.

By integrating known TF binding site information and gene expression data with the mSD approach, we can identify three different scenarios of co-regulated genes that are not only co-expressed but also share common regulatory elements in a condition-dependent way. The three scenarios are termed as "*condition-enabled*", "*condition-expanded*" and "*condition-combined*" in this chapter, which are discussed as follows:

Scenario 1 - *"condition-enabled"*: in this scenario, the TFs regulate some of their target genes in one condition but not in others. For example, the initial cluster associated with HAP1 obtained from ChIP-on-chip data alone shows a complex pattern of gene expression (Figure 4.3(a)); the mSD approach selected a subset of genes in the initial cluster by consulting with the expression data, showing a much coherent expression pattern (Figure 4.3(d)). Comparing with the ground truth from the simulation, the genes selected by mSD correspond to a major portion of the regulated genes by HAP1 (Figure 4.3(g)). Since the experimental condition of the available ChIP-on-chip data is not consistent with that of the gene expression data, not all the initial genes from ChIP-on-chip data are activated or transcribed. Instead, the genes are regulated in a "condition-enabled" way, *i.e.*, only a subset of genes being actually activated and transcribed with a coherent pattern of gene expression. This scenario indicates that the binding of HAP1 to its target genes might be influenced by different experimental conditions.

Scenario 2 - *"condition-expanded"*: in this second scenario, the target genes in one condition are further expanded to include more target genes in another condition. For example, the MIG1 ChIP-on-chip data give us only two target genes, YEL070W and HXT13, when the cut off *p*-value is relatively small (cut off *p*-value = 0.01) (Figure 4.3(b)). With the help of gene expression data, the mSD approach can help find more target genes (Figure 4.3(e) that are actually included in the simulation (*i.e.*, the ground truth) (Figure 4.3(h)). As the figure demonstrates, the mSD approach selected a subset of genes showing a highly coherent pattern of gene expression with expanded support from binding information (noting that the actual cut off *p*-value used to generate the gene expression data is relatively large).

Scenario 3 -*"condition-combined"*: as the third scenario, the target genes of a TF were identified as combined ones from different conditions. For each TF, we allocated the target genes by gathering genes with similar expression pattern and shared binding site. For example, the target genes of STE12 are shown in Figure 4.3(i). The initial cluster associated with STE12 obtained from ChIP-on-chip data (Figure 4.3(c)) shows a relatively simple expression pattern, but the actual target genes of STE12 shows a much complex expression pattern for the gene expression data (Figure 4.3(f)).

The complex expression pattern is supported by several biological studies. STE12 was reported to participate in the cell wall integrity signaling pathway [159], and to constitute a coordinated group with other TFs regulating genes involved in cell cycle control or regulation of telomere maintenance [160]. The mSD approach selected a combined subset of genes as shown in Figure 4.3(f); in addition to the genes backed up by both data sources, some genes are backed up by ChIP-on-chip data and the others are backed up by gene expression data. This scenario demonstrates that the mSD approach can obtain "condition-combined" target genes from both gene expression data and binding information.

To evaluate the performance of the mSD approach, we compared its performance with those of other similar methods, including FastNCA [157] and sparse decomposition (SD) [54]. The performances are measured by Receiver Operating Characteristics (ROC) analysis and the area under the ROC curve (AUC). The ROC curve measures the sensitivity and specificity of a method by calculating the true-positive (TP) rate against the false-positive (FP) rate. To generate a ROC curve, we first ranked the target genes for each TF according to their connection strengths in **S**, and then we calculated the true and false positive rates by running down the ranked gene list one at a time. To investigate the noise impact on the performances of mSD and FastNCA approaches, the binding information was obtained from the ChIP-on-chip data with different cut off *p*-values (*i.e.,* 0.01, 0.05 and 0.1); a large cut off *p*-value results in a high false positive rate in binding information, *i.e.*, a high noise level. In this experiment, we selected the following eleven well-known regulatory TFs: ARG80, DAL82, GCN4, GCR2, HAP1, MIG1, RGT1, RTG1, RTG3, STE12 and XBP1, to calculate the averaged TP rates and FP rates for ROC analysis. Figure 4.4 shows the ROC curves of three different approaches, while

Table 4.1 shows the AUCs of the ROC curves. (For more ROC analysis results, please refer to Figure C.1 and Table C.1 in the Addendum C to see the detailed performance for several individual transcription factors.) As can be seen from the figures and tables, the mSD approach outperforms other two methods in identifying co-regulated genes in all three cut off *p*-values. Surprisingly, the performance of FastNCA is worse than that of SD even though no binding information is used in the SD approach. It is worth noting that FastNCA largely depends on correct network topology, *i.e.*, assumed noiseless binding information in this case. When the noise level in binding information is relatively large, the performance of FastNCA degrades to an unacceptable degree. In contrast, the mSD approach finds a subset of target genes to reinforce the consistency between binding information and gene expression data, hence, to combat the noise impact from both binding information and gene expression data.

Table 4.1 AUCs of mSD, SD and FastNCA methods, respectively, under different cut off *p*-values

	mSD	SD	Fast NCA
cut off *p*-value = 0.1	**0.7160**	0.6912	0.5707
cut off *p*-value = 0.05	**0.7799**	0.6881	0.5891
cut off *p*-value = 0.01	**0.8024**	0.6801	0.5547

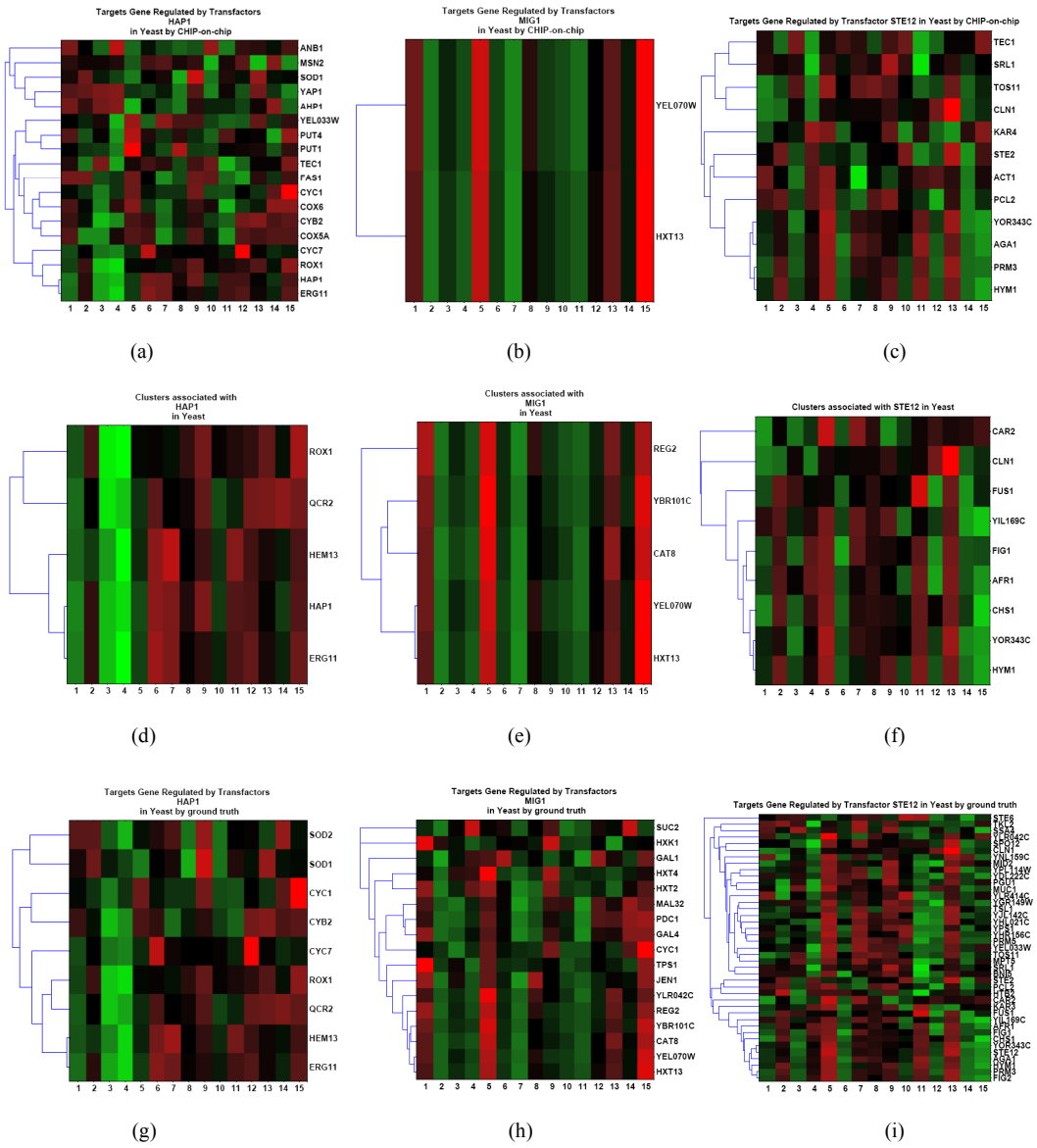
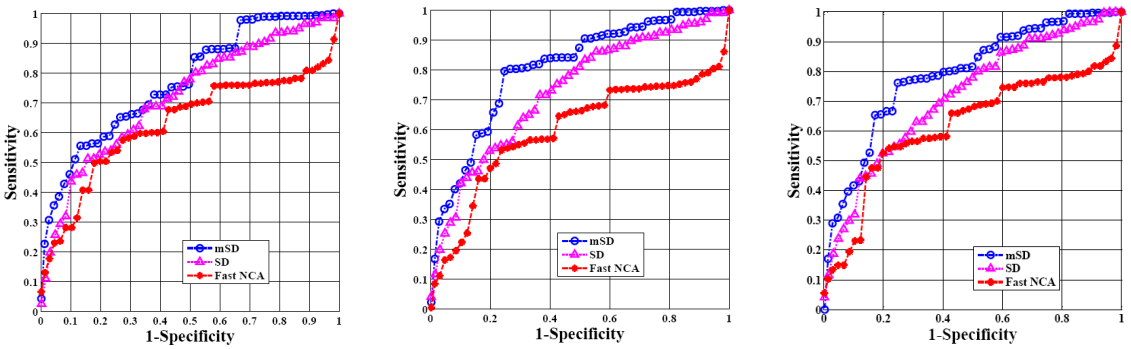


Figure 4.3: Gene clusters identified as co-regulated by HAP1 (left), MIG1 (middle) and STE12 (right), respectively. The first row: initial clusters from ChIP-on-chip data for HAP1 (a), MIG1 ((b) and STE12 (c), respectively; the second row: identified target genes of HAP1 (d), MIG1 (e) and STE12 (f), respectively; the third row: the ground truth of target genes regulated by HAP1 (g), MIG1 (h) and STE12 (i).



(a) cut off *p*-value = 0.1 (b) cut off *p*-value = 0.05 (c) cut off *p*-value = 0.01

Figure 4.4: Comparison of Receiver Operator Characteristic (ROC) curves for mSD and other methods (*i.e.*, SD and FastNCA) on simulation data. In this comparison study, three different cut off *p*-values (0.1, 0.05 and 0.01) have been applied to ChIP-on-chip data for investigating the noise impact on the performance.

To further evaluate our algorithm, we applied the mSD approach to a cell cycle data set obtained under the condition of arrest of a *cdc*15 temperature-sensitive mutant [66]. As a preprocessing step, we employed KNNimpute [125] to fill in missing values and then identified 800 cell cycle-related genes as the gene subpopulation to test the mSD approach. For the mSD approach, we set the trade-off parameter λ in Eq. (4.4) as 0.08 for this experiment, since the cost function $C(\lambda)$ (Eq. (4.12)) reached its minimum at $\lambda = 0.08$. The $C(\lambda)$ curve is demonstrated in the Figure 4.5. Since there is no ground truth of target genes available for this experiment, we used the functional enrichment of regulatory modules to compared the performance of mSD with that of another method, COGRIM [161]. COGRIM is derived from a Bayesian hierarchical model and implemented using the Gibbs sampling technique. COGRIM can help infer the activation or inhibition of TFs acting on their target genes, with an integration of microarray gene expression data, ChIP-on-chip data and motif information. The top GO enrichment *p*-values were transformed to negative logarithm values and averaged over all identified modules. The averaged enrichment score for the mSD method is 3.900, which is slightly better than the score for COGRIM (3.894), demonstrating that the mSD method can help identify functionally coherent gene clusters associated with specific TFs.

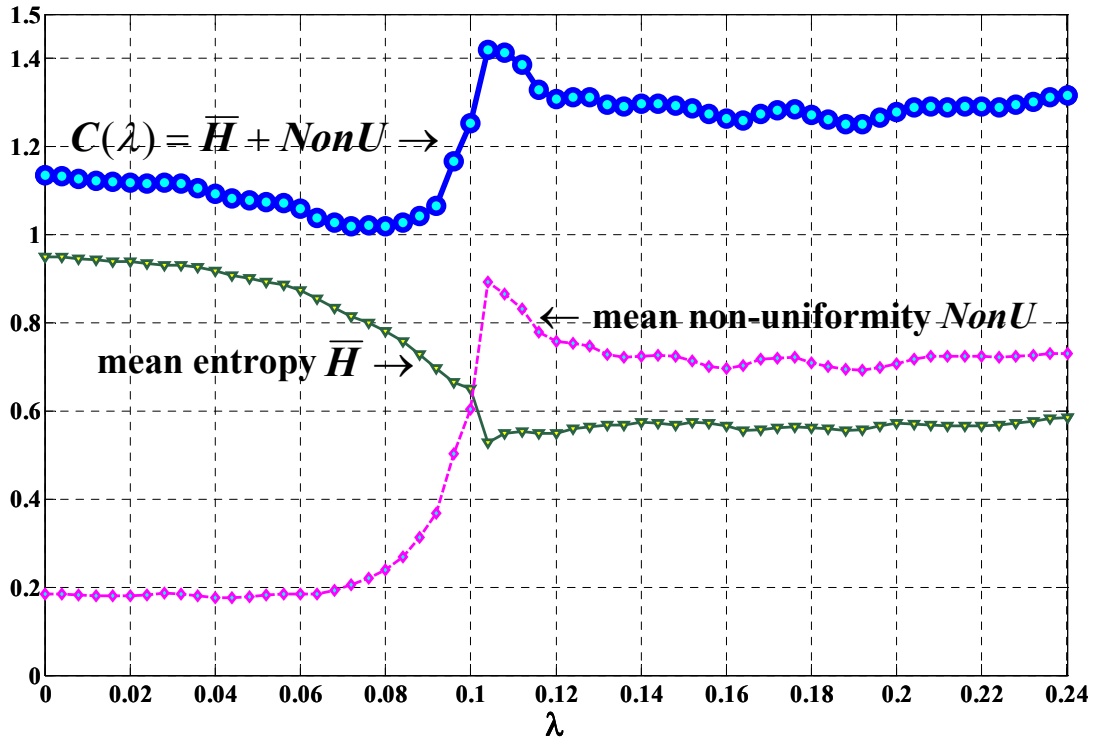


Figure 4.5: Determination of the trade-off parameter λ for yeast cell cycle data. Dark-green triangle: mean entropy of motif occupancy; magenta diamond: mean non-uniformity of gene expression pattern; blue circle: C(λ) that adds up mean entropy of motif occupancy and mean non-uniformity of gene expression pattern.

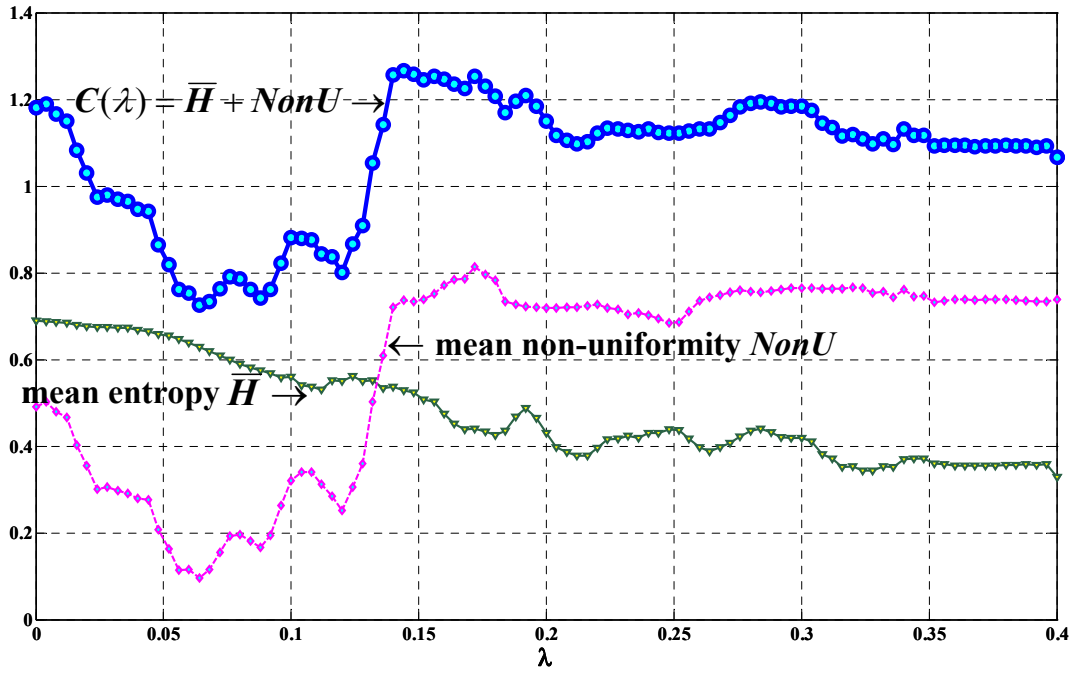
4.3.2 Breast Cancer Cell Line Data

We applied the mSD approach to breast cancer cell line data to help understand estrogen signaling and action in breast cancer cells. Greater than 70% of invasive breast cancers diagnosed each year in the U.S. manifest detectable levels of estrogen receptor alpha (ER, ER+). The most potent natural ligand for ER is 17β-estradiol, which is known to regulate the proliferation of breast cancer cells and alter their cytoarchitectural and phenotypic properties [162, 163]. Antiestrogen drugs such as Tamoxifen and Fulvestrant are widely used in the treatment of these breast cancers and produce a significant survival benefit for some patients. However, half of these cancers will recur and recurrent metastasis breast cancer remains an incurable disease. It is, therefore, clinically and biologically important to understand what transcriptional programs regulate these recurrence events [164, 165].

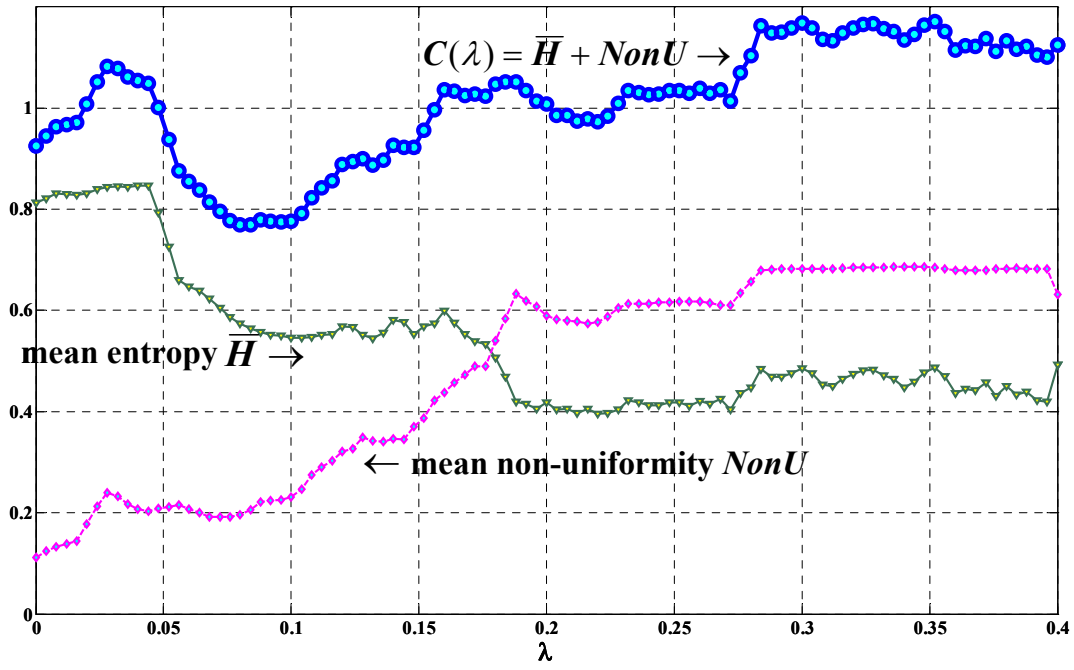
To gain insights into the transcriptional programs that drive tumor recurrence, we have collected and acquired breast cancer cell line data in estrogen-induced and estrogen-deprived conditions, respectively. The estrogen-induced data set is a time course microarray data set obtained from the ER+, estrogen-dependent breast cancer cell line-MCF-7, treated with 17β-estradiol (E2) [166]. The estrogen-deprived data set consists of a series of breast cancer variants that closely reflect clinical phenotypes of endocrine sensitive tumors [167]. The breast cancer variants are also derived from the MCF-7 cell line, including MIII cells and LCC1 cells; MIII cells were derived directly from MCF-7 and become estrogen independent and proliferate aggressively after six months of selection *in vivo* in ovariectomized athymic mice. LCC1 cells were derived from MIII following a further similar selection *in vivo*. Both cell lines remain ER+ and exhibit an estrogen-independent but antiestrogen sensitive phenotype.

In this experiment, we focused on 26 breast cancer and estrogen receptor (ER) related transcription factors. The motif information was obtained from TRANSFAC database [146] and ChIP-on-chip experiments [168]. We first identified a set of key transcription factors previously known to be involved in the Estrogen Receptor Signaling or breast cancer related and generated a list of 26 transcription factors (TFs). The complete list of these TFs is included in the Table C.2 of Appendix C.

For the mSD approach, we optimized the trade-off parameter λ in Eq. (4.4) by examining the cost function $C(\lambda)$ (Eq. (4.12)) (see Figure 4.6 for the detailed $C(\lambda)$ curves). Utilizing the mSD approach to integrate motif information and gene expression data, we identified several key regulatory networks associated with estrogen signaling. Figure 4.7 shows the activities of five transcription factors (*i.e.,* AP-1, ETF, ER, STAT and NFκB) in estrogen-induced and estrogen-deprived conditions, respectively, that show distinctive patterns of regulation. As we can see from Figure 4.7(a), these transcription factor activities clearly show different actions in response to estrogen induction. V$AP1_Q4_01 was activated within 1 hour after the estrogen was introduced; V$ETF_Q6 and V$ER_Q6 were also activated early, but showed a subsequent decrease in activity followed by a second activation event by 24 hours; V$STAT_Q6 exhibited a response to estrogen induction within 2 hours. This STAT activity estimation correlates well with previous findings that STATs are activated via the tyrosine phosphorylation



(a)



(b)

Figure 4.6: Determination of the trade-off parameter λ for breast cancer cell line data: (a) estrogen-induced condition and (b) estrogen-deprived condition. Dark-green triangle: mean entropy of motif occupancy; magenta diamond: mean non-uniformity of gene expression pattern; blue circle: $C(\lambda)$ that adds up mean entropy of motif occupancy and mean non-uniformity of gene expression pattern.

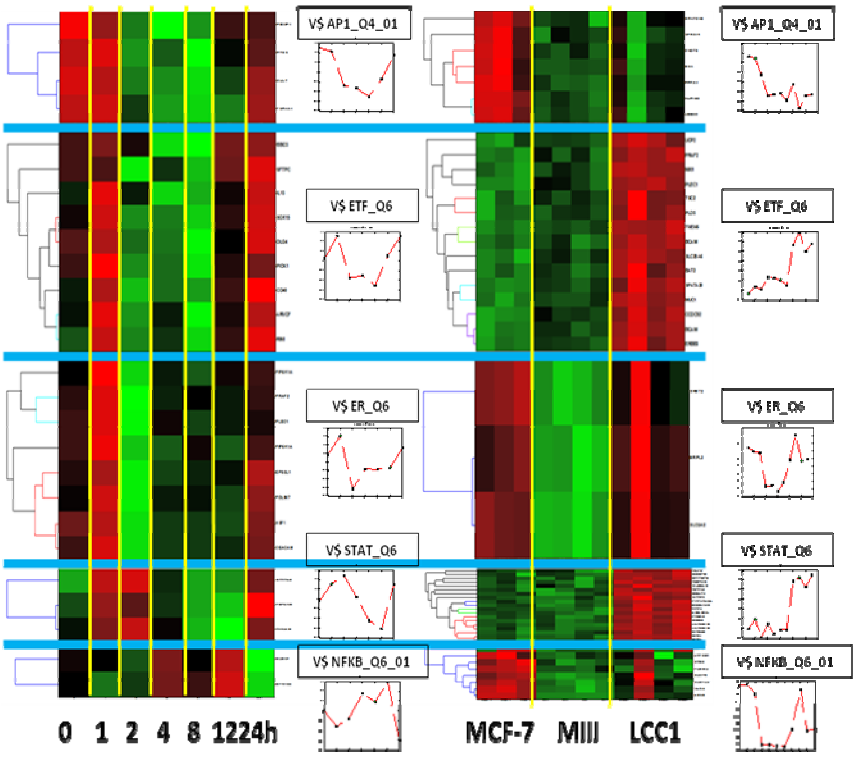


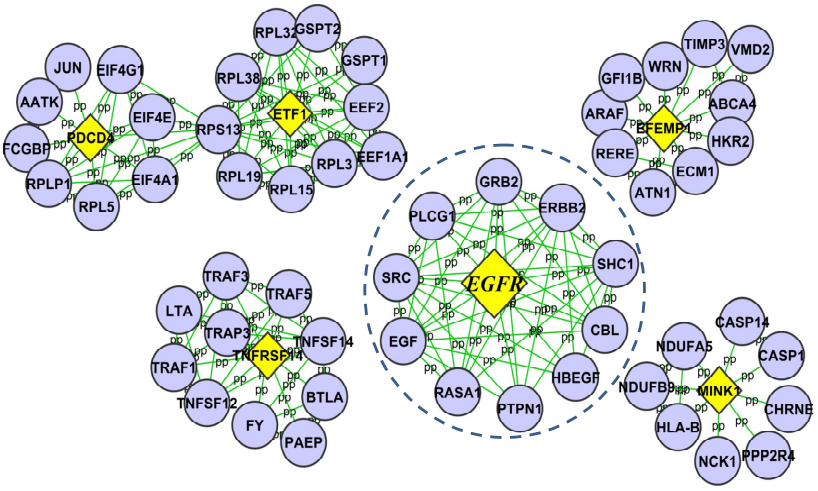
Figure 4.7: Transcription factor activity estimated by the mSD approach. (a) Estimated activities of the five transcription factors (AP-1, ETF, ER, STAT and NFκB) in estrogen-induced condition. In the expression pattern, columns represent samples in the time-course data and rows represent a group of target genes that are regulated by the TFs. The activity of each transcription factor is shown besides the expression pattern. (b) Estimated activities of the five transcription factor bind sites in estrogen-deprived condition.

cascade after ligand binding and stimulation of the cytokine receptor – kinase complex [169]. One of the mechanisms by which ER signaling occurs involves protein-protein interactions, e.g., activated estrogen receptors interact directly with transcription factors such as nuclear factor κB (NFκB), activator protein-1 (AP-1) and specificity protein-1 (SP-1), to activate gene transcription [170]. As shown in Figure 4.7 (a), an extended period of NFκB activation can be observed from 4 hours to 12 hours, which could be explained, at least in part, by such mechanism.

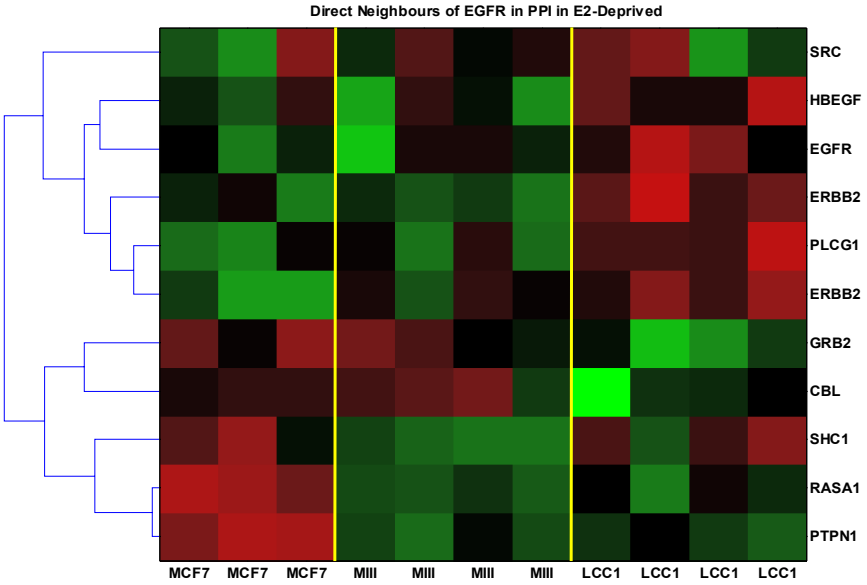
Figure 4.7 (b), on the other hand, shows the activities of these five transcription factors in the estrogen-deprived condition. Activation of ER can be clearly observed in LCC1 cells, along with activation of both ETF (V$ETF_Q6) and STAT (V$STAT_Q6), suggesting that the additional *in vivo* selection has led to further adaptations in ER

signaling in these cells. To understand the mechanisms behind this, we examined both transcript factor activity (**A**) and regulation strength (**S**) to gain some insights into condition-specific regulation programs, particularly, the program in the estrogen-deprived condition for ETF and STAT. For example, we examined the target genes of EGFR-regulating transcription factor ETF (HUGO gene symbol: TEAD2, V$ETF_Q6) to understand its regulation role in estrogen-deprived condition; ETF is known to stimulate EGFR transcription and might play a role in the overexpression of the cellular oncogene EGFR [171]. As expected, there is a large overlap between the identified ETF target gene sets in the two conditions, which are listed in Appendix C (Table C.3). These genes are enriched in the following Gene Ontology terms: 'cell adhesion'; 'cell cycle process'; 'negative regulation of progression through cell cycle'; 'regulation of kinase activity' and 'regulation of transferase activity and apoptosis'. Notably, EGFR itself is among the overlapped genes and the expression of EGFR is up-regulated in LCC1 cells. We then searched the STRING database [172] to find direct neighbors of EGFR in the protein-protein interaction (PPI) network [165]. The STRING database collects known and predicted protein-protein interactions where the interactions are either direct (physical) or indirect (functional) from literature mining, experiments and pre-computed prediction [172].

Figure 4.8(a) shows some of the putative ETF target genes and their PPI networks from the String Database, which notably includes EGFR and several direct neighbors of EGFR, e.g., CBL, RASA1, PTPN1, SHC1, HBEGF, SRC, ERBB2, GREB2 and PLCC1. Other ETF target genes and their PPI networks can be found in the Appendix C (Figure C.2). Figure 4.8(b) shows the gene expression pattern of EGFR and its direct neighbors under estrogen-deprived conditions. As we can see from the figure, the expression level of CBL was largely suppressed in the estrogen-deprived condition. Considering that CBL has been reported to promote the ubiquitination and degradation of activated EGF and platelet-derived growth factor (PDGF) receptors [173], we hypothesize that EGFR expression may be increased in LCC1 cells due to both the activation of ETF and the down-regulation of CBL. Studies to explore these predictions are currently in progress.



(a)



(b)

Figure 4.8: Identified target genes of EGFR-specific transcription factor (ETF) in estrogen-induced and estrogen-deprived conditions and their PPI sub-networks. (a) Yellow diamond: (part) target genes of ETF; purple circle: direct neighbors of the target genes from protein-protein interaction data. (b) Gene expression pattern of EGFR and its direct neighbors (obtained from protein-protein interaction data) in estrogen-deprived condition.

Overexpression and/or activation of the ErbB receptors (ErbB1 = EGFR) may also promote proliferation, motility, adhesion, differentiation [174]. Recent evidence has shown that increased growth factor (GF) signaling augments the ligand (estrogen)-independent activity of ER [175], which may partially explain the activity of ER (V$ER_Q6) in LCC1 cells as seen in Figure 4.7(b). In addition, the PLC-Gamma

(PLCG1) and the JAK-STAT pathways are known for their enhancement of transcription genes that regulate cell proliferation. This could contribute to the induced activity of STAT (V$STAT_Q6) (see Figure 4.7(b)), since one of the important signaling events activated by EGFR involves tyrosine phosphorylation of STAT. Stimulation of EGFR may induce Tyrosine phosphorylation of STAT1, STAT3 and STAT5, initiating complex formation of these STATs with JAK1 and JAK2. JAKs are essential mediators of the interaction between EGFR and the STATs, which then translocate to the nucleus to stimulate gene transcription [29, 176]. Importantly, we have recently shown that EGFR signaling through p130Cas and the tyrosine kinase c-Src leads to phosphorylation of STAT5B, and that this signal transduction pathway induces Tamoxifen resistance in MCF-7 breast cancer cells [177].

4.4 Conclusion and Discussion

In this chapter, we have proposed a new strategy, namely motif-guided sparse decomposition (mSD) of gene expression, to integrate motif information and gene expression data for regulatory module identification. Binding motif information is initially used to define potential target genes, providing a priori knowledge of the regulatory network topology. A sparse latent variable model is then used to integrate gene expression data to identify which of the potential target genes are actually activated by transcription factors. The mSD approach has been implemented by a two-step algorithm to perform (1) transcription factor activity estimation and (2) regulation strength estimation. In the first step, we start to integrate binding motif information and gene expression data with the goal to identify co-regulated gene clusters. Specifically, a motif-guided gene cluster method has been developed and used to find the gene clusters, based on a joint similarity measure from both gene expression data and motif information. To combat the impact of noise on gene clustering performance, the contribution of each data type to clustering is properly quantified and the optimal trade-off between data sources can be determined by minimizing a cost function taking into account the frequency of motif occupancy and non-uniformity of expression pattern. Subsequently, we exploit a well-know biological constraint, *i.e.*, most of genes are likely regulated by a few

transcription factors, to use a sparse decomposition method for regulation strength estimation.

Flexible network configuration and regulation strength estimation can be obtained in the proposed mSD approach. Unlike the NCA method [8] that assumes the network topology (derived from ChIP-on-chip data or motif information) known without error, we consider both network configuration and connection strength estimation as integrative components of the decomposition method. The use of prior knowledge of binding motif-information provides a solid starting point. As in Sabatti's work [141], we also incorporate a sparse constraint to achieve a biologically meaningful representation of regulatory networks. The experimental results on synthetic and real yeast data have demonstrated that our method can effectively identify the target genes of transcription factors. The application of mSD to breast cancer cell line data further revealed condition-specific regulatory modules, associated with estrogen signaling and action in breast cancer, which are consistent with known gene functions in this cellular context.

The current work represents an important step toward integrating available biological information for reconstruction of complex biological networks. This goal will be better accomplished by incorporating an analysis of the synergistic effect of regulators into the proposed method. Combinatorial analysis may help discover the complex interplay between different regulators in order to assemble a complete map of regulatory networks for complex biological systems.

5 Conclusion and Future Work

We conclude this dissertation with a summary of the contributions of the proposed research. We pinpoint some limitations of current research and remaining questions as well. Subsequently we propose several directions for future work to extend the current framework.

5.1 Summary of Contributions

The advent of the technology of high-throughput transcriptional profiling using DNA microarrays has enabled scientists to routinely measure genome-wide regulatory changes in distinct circumstances. In this dissertation, we present a unified framework of latent variable modeling to tackle several issues in gene expression profiling using microarray techniques: computational dissection of tissue samples for tissue heterogeneity correction; nICA modeling of gene module composite and latent processes; and motif-guided sparse decomposition for transcriptional regulatory programs. The main contributions of this dissertation are summarized in the following sections.

5.1.1 Computational Correction of Tissue Heterogeneity

One of the challenges in applying genomic analysis to cancer is related to multi-cellularity, which can confound the analysis of data from tissue samples that contain heterogeneous population of cells [4]. Most genomic techniques actually measure an average signal in a sample from a cell population. When analyzing a heterogeneous tissue sample, this problem is more prominent because the signals from different cell types are entangled; differential regulation of genes associated with changes in cell state can be hard to distinguish. This situation quickly becomes increasingly complex once we go

beyond a handful of cell types considered in a sample and start to consider more finely delineated populations, representing the true biological complexity of a sample.

In light of this challenge, we developed a computational method to estimate proportions of different cell types, recover cell-type-associated expression profiles, and identify phenotype-specific genes. The tissue heterogeneity correction problem was first formulated as a constrained linear instantaneous mixture model in Chapter 2 and solved by a learning algorithm based on geometrical and statistical principles. We then applied the nICA theory, together with statistical selection of cross-phenotype independence-support-genes (ISGs) [71] and an ensemble study of underlying phenotypes, to computationally correct tissue heterogeneity in gene expression profiling. Methodologically, we developed a novel computational approach, non-negativity constrained partially-independent component analysis (nPICA), to solve the tissue heterogeneity correction problem in a biologically plausible way.

Compared to other computational methods, our method has the unique characteristic of discovering patient-specific compartment patterns for different phenotypes. For example using a prostate cancer data set from University of California, San Diego (UCSD) [31], Stuart *et al.* took advantage of multiple samples from multiple patients and explored the patient-independent genome-wide expression patterns of different compartments of tumors, whereas we used multiple samples from the same patient to extract individual expression patterns that may facilitate customized clinical prognosis. Furthermore, a by-product and an important attribute of our method is independent-support-genes (ISGs); the sets of significantly deregulated genes in different compartments reveal various known and novel pathways that could have contributed to tumorigenesis. The selection of ensemble ISG indices defines the most complete independent segments and subsequent nPICA produces minimum "*overshoot*", which avoids relying on the strict non-negativity constraint. We have tested the nPICA method on a series of numerical mixtures of microarray data sets to tackle cell and tissue heterogeneity in noisy models, and demonstrated its accurate performance in dissecting mixtures into pure signals of component cell types. Lastly, in Chapter 2, we have investigated the impact of tumor composition on the performance of a predictive signature before and after dissecting the signature into different components. The

experimental results have shown that a significant improved predictive accuracy can be achieved after tissue heterogeneity correction.

5.1.2 Modeling of Gene Module Composite by Latent Process Decomposition (LPD)

The reconstruction of molecular processes that underlie a complex biological system, such as tumorigenesis, is a formidable challenge. Several complex diseases, such as cancers, are thought to progress through a stepwise process that involves many aberrations at the molecular level [178]. Even though gene expression level analysis in a complex disease has advanced the knowledge on diseases progression, differentiation and development, the biological processes that drive the diseases are far more complex and, for now, elusive. Another confounding factor associated with such studies is that aberrant gene expression patterns may be caused by some other processes independent of tumorigenesis: processes that are protective (e.g., immune response and inflammatory or cell infiltration) and processes that are normal on their own but taken advantage of by tumors to support their proliferation (e.g., cell division or angiogenesis). Considering these challenges, conventional clustering algorithms have many limitations, as we discussed in Chapter 3, for elucidating molecular mechanisms underlying the observed changes in gene expression, making the results often difficult to interpret biologically.

An alternative to conventional approaches to uncover biological processes using microarray data is to treat it as a blind source separation (BSS) problem. In the context of microarray data, "sources" may correspond to specific cellular responses or regulation programs for gene modules. In this dissertation, we have developed a novel gene clustering approach – nonnegative independent component analysis (nICA) followed by VIsual Statistical Data Analyzer (VISDA) – for biological process identification. nICA exploits the positive nature of molecular expression, and thus fits better to the reality of underlying putative biological processes. Genes that exhibit significant up-regulation or down-regulation within each component are grouped together for further clustering. With the assumption that each component is regulated by several transcription factors, we believe that the advantages of clustering transcriptional modules in the latent space lie in

detecting cellular processes under regulatory effects and revealing biological functions with the corresponding transcription factors. In the nICA approach, we also developed an information-theoretic procedure for input sample selection and a novel stability analysis approach for proper dimension estimation.

We applied nICA, together with VISDA, to *Saccharomyces cerevisiae* microarray data. Experimental results showed that the nICA approach achieved an improved resolution for gene module identification when compared to conventional gene clustering methods. Evidently, the gene modules identified by the nICA approach are significantly enriched in functional annotations in terms of gene ontology (GO) categories. We applied nICA in conjunction with VISDA onto muscle regeneration data as well. The results demonstrated that besides the identification of functionally enriched or co-regulated gene groups, nICA also identified more than one gene clusters significant for the same GO categories and hence revealed a higher level of biological complexity, even within coherent groups of genes.

5.1.3 Deciphering Transcriptional Regulatory Programs by Motif-Guided Sparse Decomposition (mSD)

Transcription regulation is a starting point for controlling a variety of biological processes that coordinate the expression of thousands of genes throughout their life cycle and response to external stimuli, such as nutrients or pheromones [160]. In eukaryotes, the regulation is realized by intricate regulatory gene networks that are mainly controlled by transcription factors (TFs), and different combinations of transcription factors may alternatively activate or repress gene expression. The problem of discovering co-regulated sets of genes and transcriptional sub-networks using large high-throughput data sources is an important research topic. Several papers addressing this issue have recently been published. For example, two early papers by Pilpel *et al.* [179] and Ihmels *et al.* [106] discussed the discovery of transcriptional modules by using gene expression data to refine a set of genes selected based on some other criteria (DNA motifs or functional categories). This usually results in subsets of genes that are activated by the cell under the same conditions. While this is an important first step, with the growing surge of

biological measurements the problem of integrating different types of genomic measurements has become an immediate challenge for elucidating regulatory events at the molecular level.

In our work on the motif-guided sparse decomposition (mSD) algorithm, we aim to tackle the problem of transcriptional module identification, which essentially requires finding sets of transcription factor binding sites (TFBS) that co-occur in promoter regions of genes with a similar expression pattern. In order to learn the membership of transcriptional modules, we proposed to combine motif information and expression data in a novel way: (1) using motif information to guide finding clusters of co-regulated gene and their patterns; (2) using a sparse component analysis (SCA) [54] method to further decompose regulated gene patterns to recover the TF-gene connectivity information. In the first step, clusters are defined based on a joint similarity measure of motif binding information and gene expression data in a Bayesian-principled way. This implies that our model allows us to obtain a balance point of co-regulation and co-expression for a gene module. In order to to alleviate the noise impact on finding regulatory modules from both data sources, the contribution of each data type to clustering is properly quantified in this step; in the second step, the use of prior knowledge of TF/motif-binding relationship severs us well in providing a reasonable starting point. We then incorporate a sparse model to achieve a biologically realistic representation of gene expression data with a properly defined sparseness measure and a sparse decomposition algorithm.

In this study, we tested our newly developed approach, *i.e.*, mSD, on both simulated and real *Saccharomyces cerevisiae* data. We reported several regulatory networks under specific conditions for *Saccharomyces cerevisiae* to demonstrate the dynamic nature of transcriptional mechanisms. Compared to a conventional Bayesian method for transcriptional module identification, our method achieved improved performance in the identification of transcription factors and their target genes. We then applied our method to breast cancer cell line data to help understand estrogen signaling in breast cancer. The mSD approach helped uncover condition-specific transcriptional modules that might have important implications in endocrine therapy of breast cancer.

5.2 Future extensions

There are many possible directions to extend the current framework of latent variable modeling. We discuss some of the primary extensions in this section. The extensions fall into two categories that aim to (1) overcome the current limitations and (2) broaden the types of data to be analyzed.

- **Improve the identification of ensemble independence support genes (ISGs)**

The expression heterogeneity within a primary tumor may result in mixture signature of expression pattern, which motivates us to investigate different compartments of tumors, e.g., the inner tumor mass, stromal cells and surrounding normal epithelial tissue by microarray-based expression profiling. Given the difficulty of the task, while the optimality of the proposed method may be data- or modality-dependent, we showed promising empirical outcomes on several *in silico* and real gene expression profiling studies of heterogeneous tissues. In reality, problems arise when data are corrupted by experimental noises or confounded by clinical variations. These will introduce a great degree of variability within the context of ISG selection. As we pointed out in Eq. (2.7), part of up-stream genes often exhibit very low expression levels across all the microarrays. Although they belong to the ISG set, in practice they are difficult to select since the genes would often be "randomly" allocated on the simplex hyperplane. We suggest reducing the noise impact by screening out those genes after normalization. If a subset of ISGs is obtained from an ensemble study, selection frequency may serve as another measure of reliability for ISG. If we select ISG indices from an ensemble study, and perform nPICA on individual samples using these ensemble ISG indices, we can test whether the ensemble ISG-nPICA is applicable to individual samples.

- **Broaden the application of ISG-nPICA on population-based studies to identify cell-type-associated molecular signatures**

In Chapter 2, we investigated the tissue heterogeneity problem by unsupervised multivariate data mining techniques, incorporating statistical properties and geometrical constraints. In particular, we focused on the exactly-determined case, *i.e.*, the number of samples is same as the number of cell types or sources. In population-based studies, there will be multiple heterogeneous biopsy samples from multiple patients, which constitute

an over-determined problem, *i.e*. the number of samples is larger than the number of sources. To apply the ISG-nPICA algorithm to population-based studies, we shall select *d* representative samples (exemplars) from the available samples to form an exactly-determined problem, assuming that the true number of underlying sources *d* is known before hand. A possible way to sample is to perform unsupervised sample clustering based on normalized global expression profiles. Therefore, over-determined problem requires additional analytic steps, and the complete analysis may consist of the following three steps: (1) unsupervised sample clustering to select "exemplars" — cluster centers; (2) ISG stability analysis as we discussed before; (3) ISG-nPICA analysis to extract cell-type-associated molecular signatures.

- **Include partial prior knowledge about TF-gene interactions in nICA modeling of biological processes**

nICA modeling of gene module composite is introduced in Chapter 3, in which each observation in the data set is represented as a combinatorial mixture over a finite set of latent processes. nICA modeling has several advantages over conventional clustering methods. For example, activities of latent processes are included in the model, which can facilitate biological interpretation. Several processes can be simultaneously combined to determine gene expression levels so that the relationship between processes is more transparent. However, the latent components, which are expected to correspond to biological processes, usually do not correspond exactly to transcription factors. We believe that prior knowledge about TF-gene interactions should be incorporated into the model, which will help determine more accurate TF-gene regulatory interactions. It is worth noting that such knowledge, albeit still partial, has already been proven useful for comprehensive identification of modules in yeast [8].

Besides, we used a stability-based method to detect the number of underlying biological processes. However, there is no theoretic proof demonstrating that the "correct" number of biological processes corresponds to the result from stability analysis. In order to unambiguously determine the number of processes, we may need an alternative model selection approach such as cross-validation to confirm the results from stability analysis [137].

- **Ameliorate the estimation of cluster numbers in motif-guided clustering of gene expression profile data**

For motif-guided clustering to find groups of genes co-regulated by common TFBSs, in the current implementation we empirically control the number of clusters to be equal to that of TFBSs. However, it is important to determine the number of clusters from gene expression data. Various methods have been proposed for cluster analysis to determine the number of clusters, but none has been proven robust across diverse data types [39]. Some methods use biological knowledge to determine the optimal number of clusters. For instance, ClusterJudge [126] chooses the number of clusters based on the strongest tendency to bring genes of similar function together. Other clustering analysis methods use information-theoretic criteria such as Bayes Information Criterion (BIC) [180] or statistical procedures [181] to estimate the number of clusters. In the future, it is our intention to enhance the motif-guided clustering method with a systematic approach to determine the number of gene clusters by integrating gene expression data and biological knowledge.

- **Enhance the modeling of transcriptional regulation networks by considering the cooperation of several TFs**

In Chapter 4, we developed the motif-guided sparse decomposition (mSD) approach to reconstruct regulatory networks through the decomposition of gene expression data into two matrices represented by TF activity and regulation strength respectively. The mSD approach combines gene expression data and prior knowledge about TF-gene interactions from motif information to infer the regulatory networks. The major limitation of our mSD regulatory model lies at its simplicity. The model is built upon several strong assumptions as described in Section 4.2.1, which exclude the interaction effect between TFs. But in higher eukaryotes, transcription factors co-operate as a functional complex to regulate gene expression [182]. To fully specify the combinatorial effect of multiple regulators, efficient and large-scale assays that can capture many configurations of multiple regulators together seem to be required, so that we can develop regulatory models of multiple regulators. One potential scheme is to apply a mixture of latent variable models, in which each component of the mixture corresponds to a TF complex composed of several TFs [53].

Another possible extension is to incorporate other types of data. There are vast amounts of data capturing different aspects of cellular processes. We hypothesize that genes participated in cancer initiation and progression will show dysregulated interactions with their molecular partners in several different molecular levels. To build a large-scale model of gene regulation, each type of data is important. As mentioned in Chapter 1, data fusion serves to not only reduce errors and noises from independent observations, but also constrain the model space from complementary sources. By developing genome-wide, mixed-interaction networks, instead of the individual protein-DNA interaction layer studied here, we can cover a far greater range of processes within the cell. In the long run, we envision building models that can account for more types of data in cancer research.

5.3 Conclusion

Nowadays, many examples in computational biology and bioinformatics communities have highlighted the power of gene expression profiling to deepen our understanding of biological phenotypes, transcriptional programs, etc. However, the ability to tease out information from microarray gene expression data and the attempt to derive system models that capture the dynamics of biological systems still remain challenging due to huge gene numbers, relative small sample numbers and significant noises in microarray data. In this dissertation, we present a unified framework for a latent variable model for tissue heterogeneity correction, gene module identification and regulatory network reconstruction. In particular, we developed a suite of statistically principled and biologically plausible algorithms for (1) nonnegative partially independent component analysis (nPICA) to extract the pure signals from biopsies, (2) nonnegative independent component analysis (nICA) in conjunction with VIsual Statistical Data Analyzer (VISDA) to de-correlate observations and group genes in the latent space, and (3) motif-guided sparse decomposition (mSD) to identify gene regulatory modules by integrating gene expression data and motif information. Through extensive experiments on the simulation and real microarray data, we have demonstrated that (1) new insights can be garnered from gene expression profiling experiments when expression patterns of

purified individual cell types are dissected; (2) the significant enrichment of gene annotations within clusters can be obtained by our proposed nICA approach; 3) functionally distinct gene regulatory networks can be revealed by the mSD approach, consisting of biologically validated transcription factors and their target genes. Finally, several remaining research problems have also been discussed for future study.

Appendix A. Proof of the Properties of Eq (4.13)

Proof: 1) Since the l_1, l_2 - norms are all equivalent on $\mathbf{R}^n$:

$$\|x\|_2 \le \|x\|_1 \le \sqrt{n}\,\|x\|_2 \tag{A.1}$$

We have

$$1 \le \frac{\|\boldsymbol{x}_1\|}{\|\boldsymbol{x}_2\|} \le \sqrt{n} \tag{A.2}$$

$$0 \le \frac{\sqrt{\boldsymbol{n}}}{\sqrt{\boldsymbol{n}}-1} - \frac{1}{\sqrt{\boldsymbol{n}}-1}\frac{\|\boldsymbol{x}\|_1}{\|\boldsymbol{x}\|_2} \le 1 \ . \tag{A.3}$$

2) Since any norm has the positive homogeneity or positive scalability, that is

$$\|a\mathbf{v}\| = |a|\|\mathbf{v}\|, \tag{A.4}$$

thus

$$\text{sparseness}(c\boldsymbol{x}) = \frac{\sqrt{\boldsymbol{n}}}{\sqrt{\boldsymbol{n}}-1} - \frac{1}{\sqrt{\boldsymbol{n}}-1}\frac{\|c\boldsymbol{x}\|_1}{\|c\boldsymbol{x}\|_2} = \frac{\sqrt{\boldsymbol{n}}}{\sqrt{\boldsymbol{n}}-1} - \frac{1}{\sqrt{\boldsymbol{n}}-1}\frac{\|\boldsymbol{x}\|_1}{\|\boldsymbol{x}\|_2} = \text{sparseness}(\boldsymbol{x}) . \tag{A.5}$$

3) $\text{sparseness}(\alpha\boldsymbol{x}_1 + \beta\boldsymbol{x}_2) \le \max(\text{sparseness}(\boldsymbol{x}_1), \text{sparseness}(\boldsymbol{x}_2))$ is equivalent to prove $\dfrac{\|\boldsymbol{x}_1 + c\boldsymbol{x}_2\|_1}{\|\boldsymbol{x}_1 + c\boldsymbol{x}_2\|_2} \ge \min\left\{\dfrac{\|\boldsymbol{x}_1\|_1}{\|\boldsymbol{x}_1\|_2}, \dfrac{\|\boldsymbol{x}_2\|_1}{\|\boldsymbol{x}_2\|_2}\right\}, c = \beta/\alpha$ because of Properity (2) above.

For the sake of proofing, we assume $\|\boldsymbol{x}_1\|_1 = \|\boldsymbol{x}_2\|_1$ (still because of Properity of (2), we are allowed to do such assumption), $\dfrac{\|\boldsymbol{x}_1\|_1}{\|\boldsymbol{x}_1\|_2} \le \dfrac{\|\boldsymbol{x}_2\|_1}{\|\boldsymbol{x}_2\|_2}$ and $\text{sgn}(\alpha\boldsymbol{x}_1(n)) = \text{sgn}(\beta\boldsymbol{x}_2(n))$, that is $\text{sgn}(\boldsymbol{x}_1(n)) = \text{sgn}(c\boldsymbol{x}_2(n))$, we have

$$\frac{\|\boldsymbol{x}_1 + c\boldsymbol{x}_2\|_1}{\|\boldsymbol{x}_1 + c\boldsymbol{x}_2\|_2} = \frac{\sum_{i=1}^{n}|x_1(i) + cx_2(i)|}{\sqrt{\sum_{i=1}^{n}x_1^2(i) + c^2\sum_{i=1}^{n}x_2^2(i) + 2c\sum_{i=1}^{n}x_1(i)x_2(i)}}$$

$$= \frac{\sum_{i=1}^{n}|x_1(i) + cx_2(i)|}{\|\boldsymbol{x}_1\|_2} \frac{1}{\sqrt{1 + c^2\sum_{i=1}^{n}\left[\frac{x_2(i)}{x_1(i)}\right]^2 + 2c\frac{\sum_{i=1}^{n}x_1(i)x_2(i)}{\sum_{i=1}^{n}x_1^2(i)}}}$$

$$\geq \frac{\sum_{i=1}^{n}|x_1(i) + cx_2(i)|}{\|\boldsymbol{x}_1\|_2} \frac{1}{\sqrt{1 + c^2 + 2c\frac{\sum_{i=1}^{n}x_1(i)x_2(i)}{\sum_{i=1}^{n}x_1^2(i)}}} \quad (\because \frac{x_2(i)}{x_1(i)} \leq 1)$$

$$\geq \frac{\sum_{i=1}^{n}|x_1(i) + cx_2(i)|}{\|\boldsymbol{x}_1\|_2} \frac{1}{\sqrt{1 + c^2 + c\frac{\sum_{i=1}^{n}[x_1^2(i) + x_2^2(i)]}{\sum_{i=1}^{n}x_1^2(i)}}} \quad (\because 2ab < a^2 + b^2)$$

$$\geq \frac{\sum_{i=1}^{n}|x_1(i) + cx_2(i)|}{\|\boldsymbol{x}_1\|_2} \frac{1}{\sqrt{1 + c^2 + 2c}} = \frac{\sum_{i=1}^{n}|x_1(i) + cx_2(i)|}{\|\boldsymbol{x}_1\|_2} \frac{1}{|1 + c|} \tag{A.6}$$

$\because \operatorname{sgn}(\boldsymbol{x}_1(n)) = \operatorname{sgn}(c\boldsymbol{x}_2(n))$

$$\therefore (\text{A.6}) = \frac{\sum_{i=1}^{n}|x_1(i)| + |cx_2(i)|}{\|\boldsymbol{x}_1\|_2} \frac{1}{|1 + c|} = \frac{|1 + c|}{\|\boldsymbol{x}_1\|_2} \frac{1}{|1 + c|} = \frac{\|\boldsymbol{x}_1\|_1}{\|\boldsymbol{x}_1\|_2} \qquad \square$$

Appendix B. Addendum of Empirical Results for Chapter 2

In this appendix, we include various empirical results which are not reported in the main text of Chapter 2.

They include overlaid projections on the standard simplex between the true and recovered signals (rotation angle $\theta = \pi/4$) (Figure B.1); overlaid projections on the standard simplex between the true and recovered signals (rotation angle $\theta = \pi/3$) (Figure B.2)**;** sample distributions after LDA dimension reduction (Figure B.3); results of independent tests for recovered signals and mixtures on noise free case (Figure B.4); results of independent tests for recovered signals and mixtures on noise cases with SNRs from 40dB to 15dB (Figure B.5); comparison results in independent tests for sensitivity, false negative rate and accuracy between recovered samples after THC and mixtures with 10dB Gaussian noise added (Figure B.6); results of classification results for mixtures and recovered signals on the noisy sources case (Figure B.7); comparison results in independent tests for sensitivity, false negative rate and accuracy between recovered samples after THC and mixtures on noisy sources with 40dB, 35dB, 30dB, 25dB, 20dB, 15dB and 10dB Gaussian observation noise added (Figure B.8) and mixing information of 64 mixture samples generated from 32 patients' pure microarray data (Table B.1).

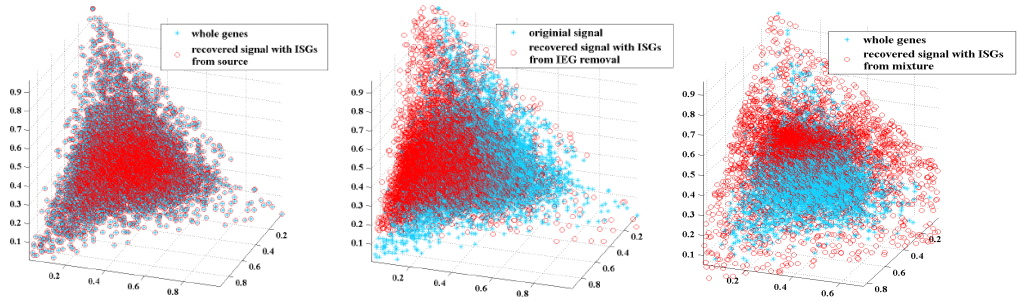


Figure B.1: Overlaid projections on the standard simplex between the true sources and the recovered signals (rotation angle $\theta = \pi/4$). Blue stars: original signals; red circles: recovered signals. Left panel: ISGs identified from sources (number of ISGs = 400); middle panel: ISGs identified from IEG removal (number of ISGs = 406); right panel: ISGs indentified from mixtures (number of ISGs = 400).

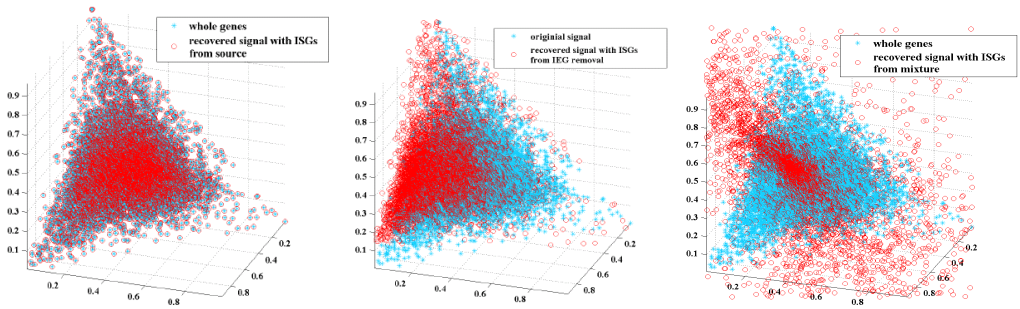


Figure B.2: Overlaid projections on the standard simplex between the true sources and the recovered signals (rotation angle $\theta = \pi/3$). Blue stars: original signal; red circles: recovered signals. Left panel: ISGs identified from sources (number of ISGs = 400); middle panel: ISGs identified from IEG removal (number of ISGs = 396); right panel: ISGs indentified from mixtures (number of ISGs = 400).

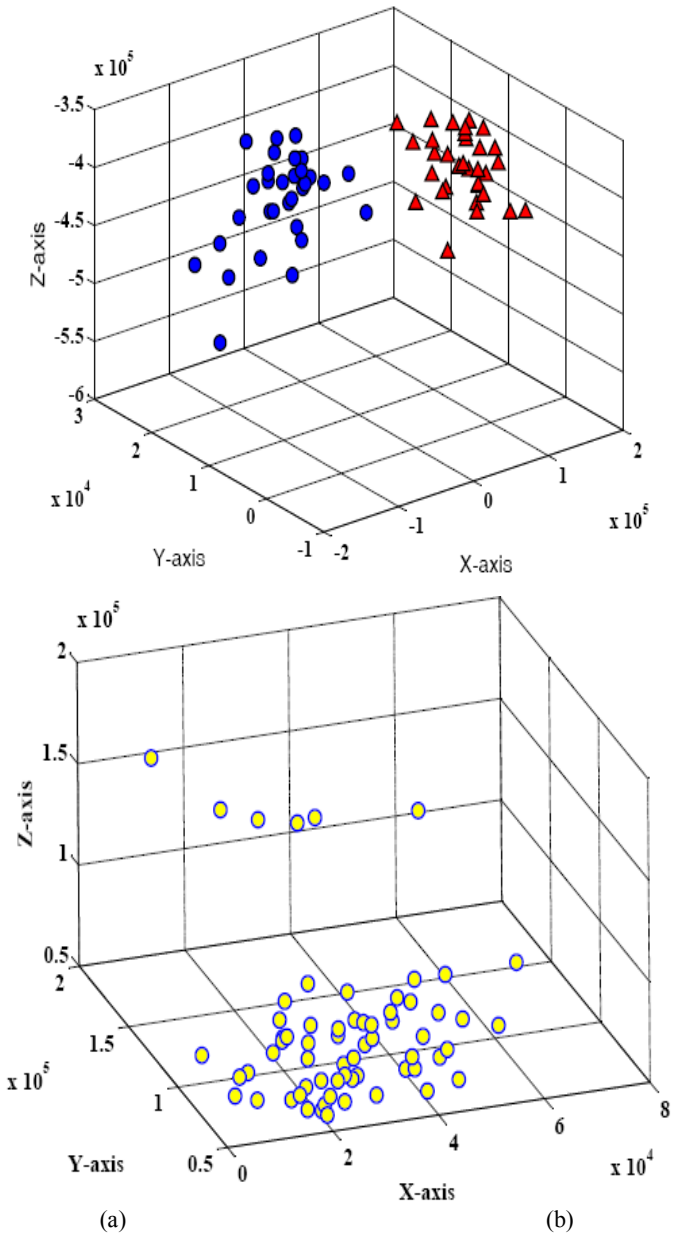
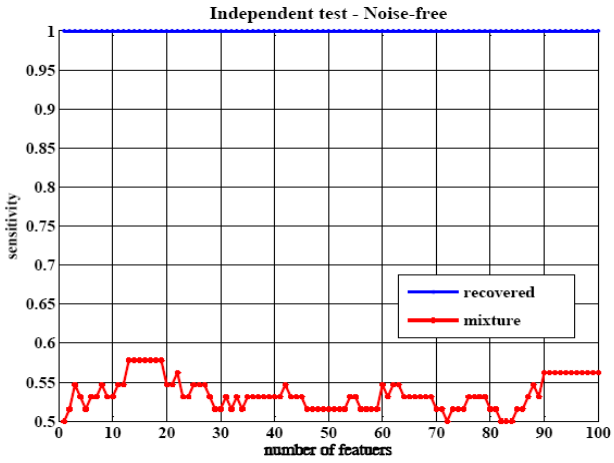
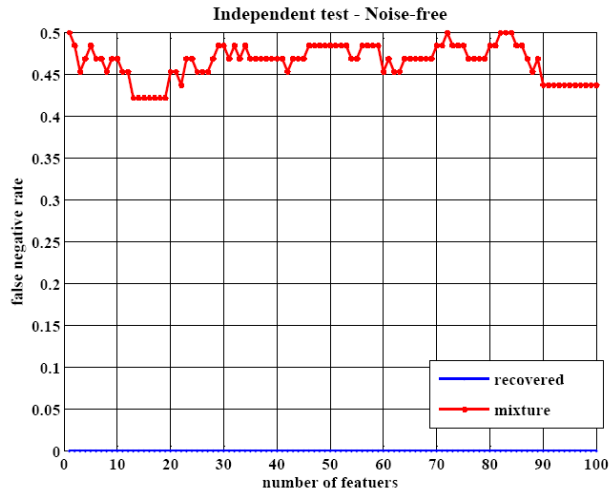


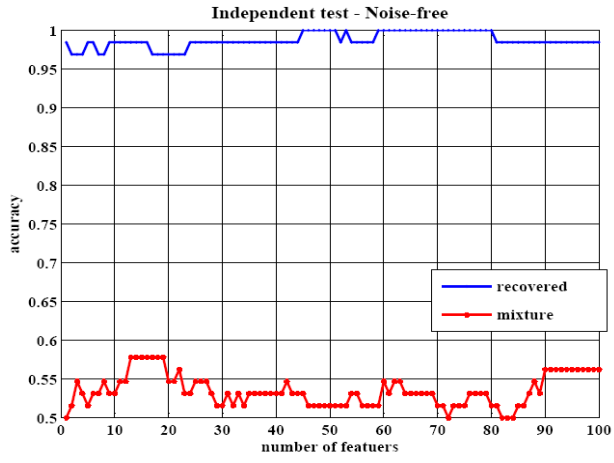
Figure B.3: Sample distributions after LDA dimension reduction. (a) The distribution of original samples in 3-D space using LDA. Blue circles: adenomas tissues; Red triangles: normal tissues; (b) the distribution of mixture samples in 3-D space using LDA. Yellow circle: 64 observations generated from 32 pairs of normal and adenomas tissues from 32 patients.



(a)



(b)



(c)

Figure B.4: Results of independent tests for recovered signals and mixtures on noise free case: (a) sensitivity curves; (b) false negative rate curves; (c) overall classification accuracy curves.

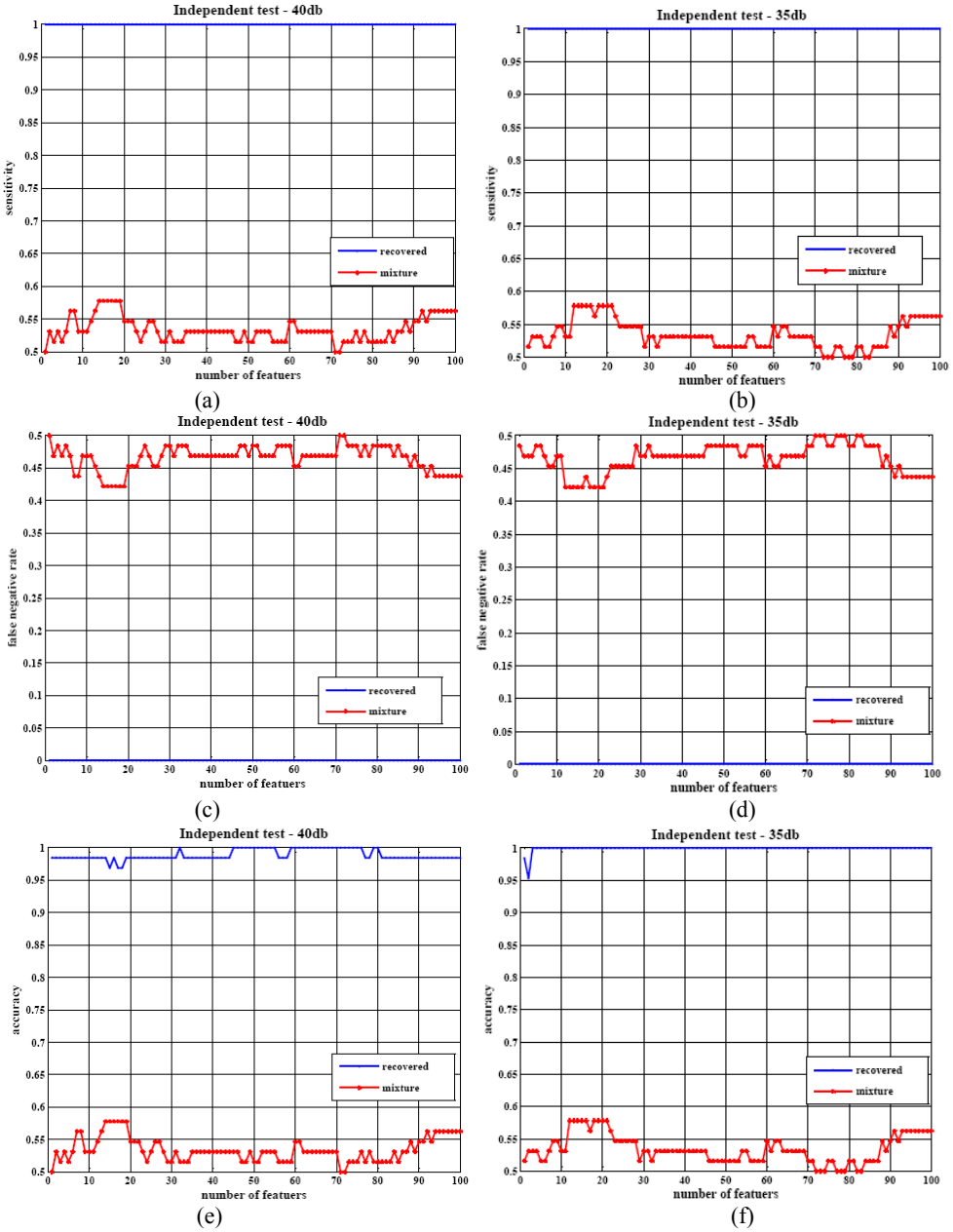


Figure B.5: Results of independent tests for recovered signals and mixtures on noise cases: (a) sensitivity curves with SNR = 40dB; (b) sensitivity with curves SNR = 35dB; (c) false negative rate curves with SNR = 40dB; (d) false negative rate curves with SNR = 35dB; (e) overall classification accuracy curves with SNR = 40dB; (f) overall classification accuracy curves with SNR = 35dB.

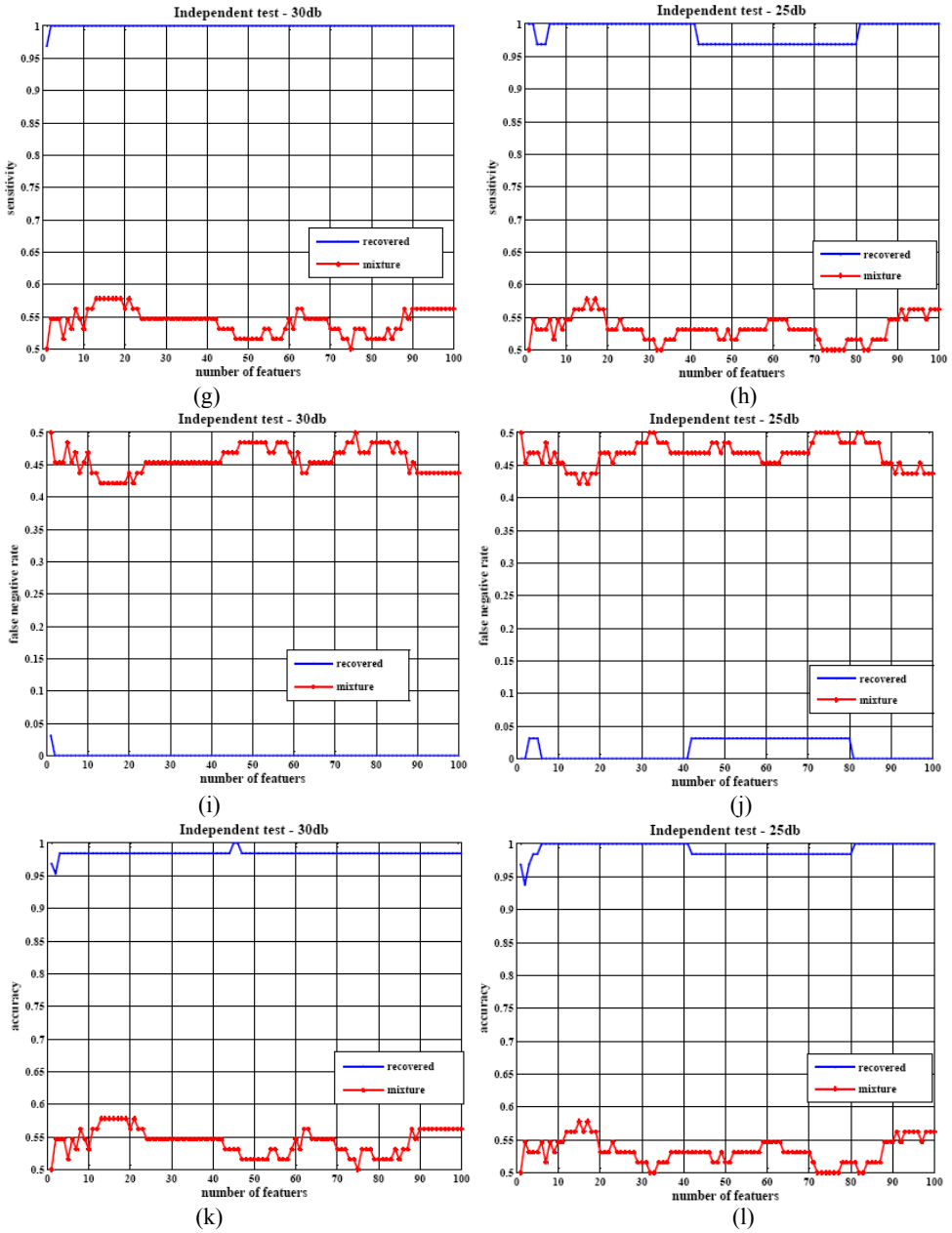


Figure B.5 (cont'd). Results of independent tests for recovered signals and mixtures on noise cases: (g) sensitivity curves with SNR = 30dB; (h) sensitivity curves with SNR = 25dB; (i) false negative rate curves with SNR = 30dB; (j) false negative rate curves with SNR = 25dB; (k) overall classification accuracy curves with SNR = 30dB; (l) overall classification accuracy curves with SNR = 25dB.

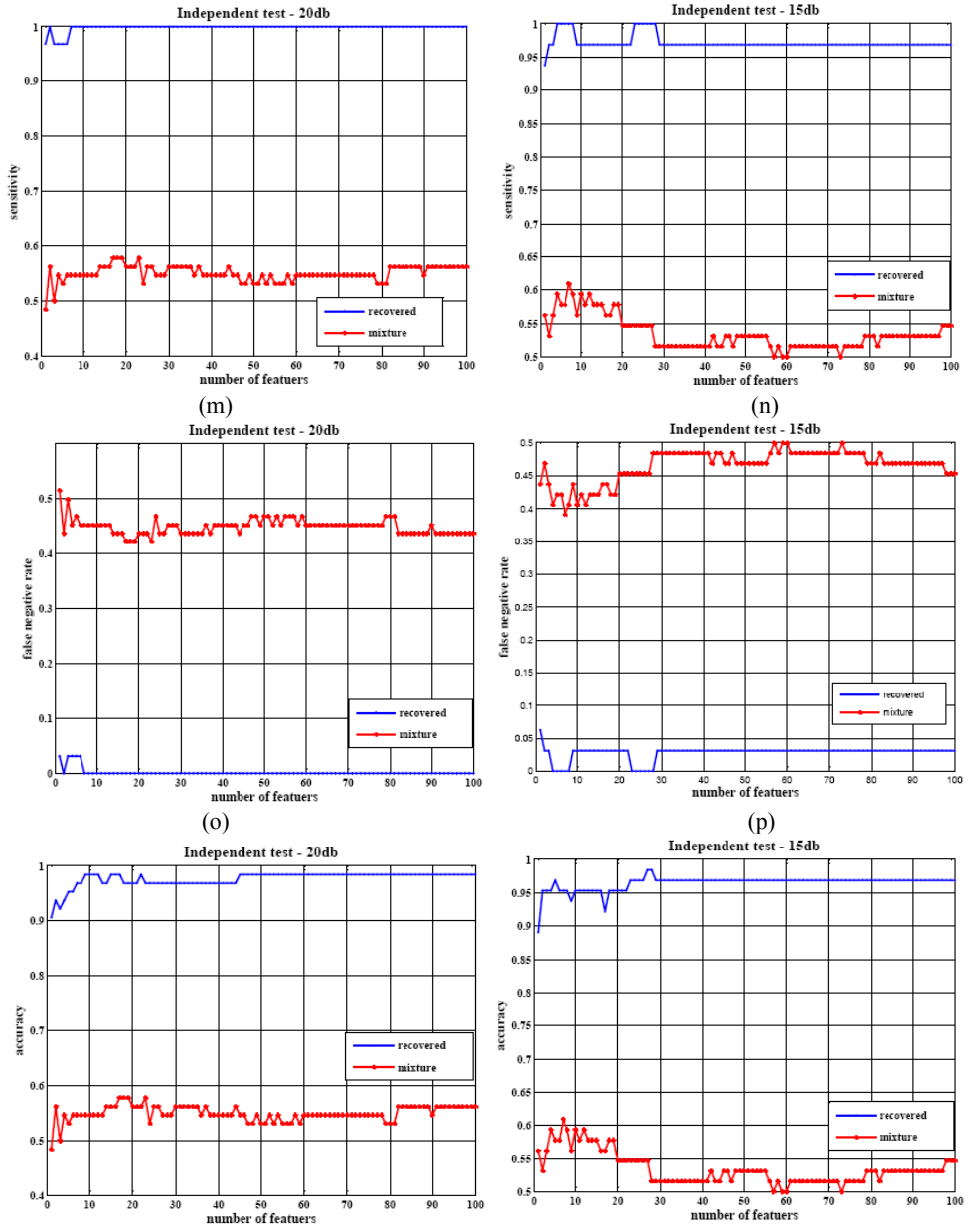
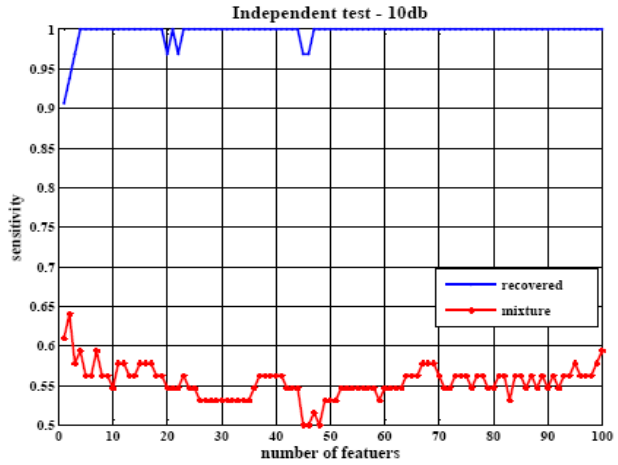
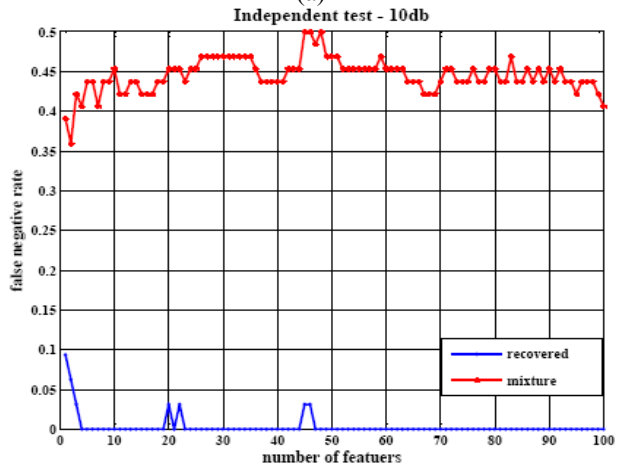


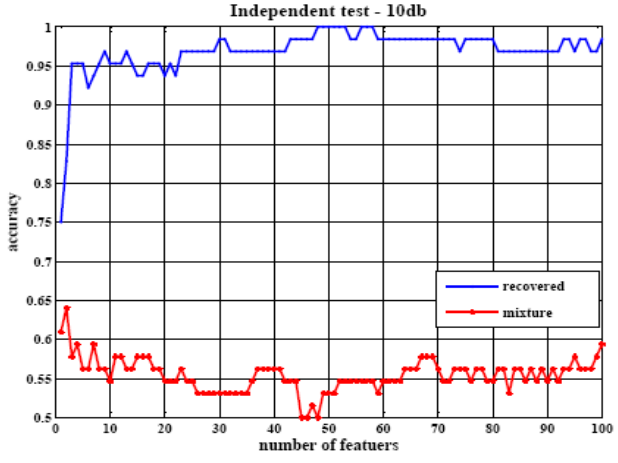
Figure B.5 (cont'd). Results of independent tests for recovered signals and mixtures on noise cases: (m) sensitivity curves with SNR = 20dB; (n) sensitivity with curves SNR = 15dB; (o) false negative rate curves with SNR = 20dB; (p) false negative rate curves with SNR = 15dB; (q) overall classification accuracy curves with SNR = 20dB; (r) overall classification accuracy curves with SNR = 15dB.



(a)



(b)



(c)

Figure B.6: Results of independent tests for recovered signals and mixtures with 10dB additive Gaussian noise: (a) sensitivity curves; (b) false negative rate curves; (c) overall classification accuracy curves**.**

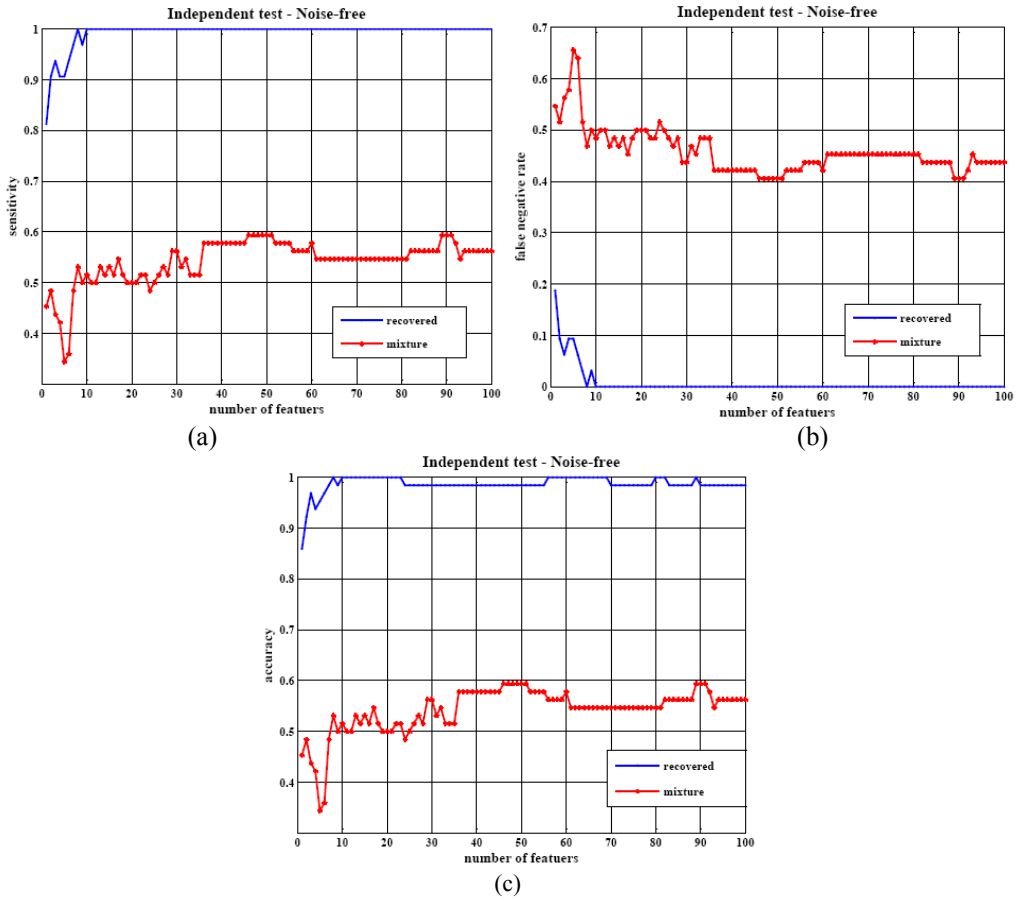


Figure B.7: Results of classification results for mixtures and recovered signals on the noisy sources case: (a) Independent test results for sensitivity curves on the noisy sources case without observation noise; (b) Independent test results for false negative rate curves on the noisy sources case without observation noise; (c) Independent test results for accuracy curves on the noisy sources case without observation noise.

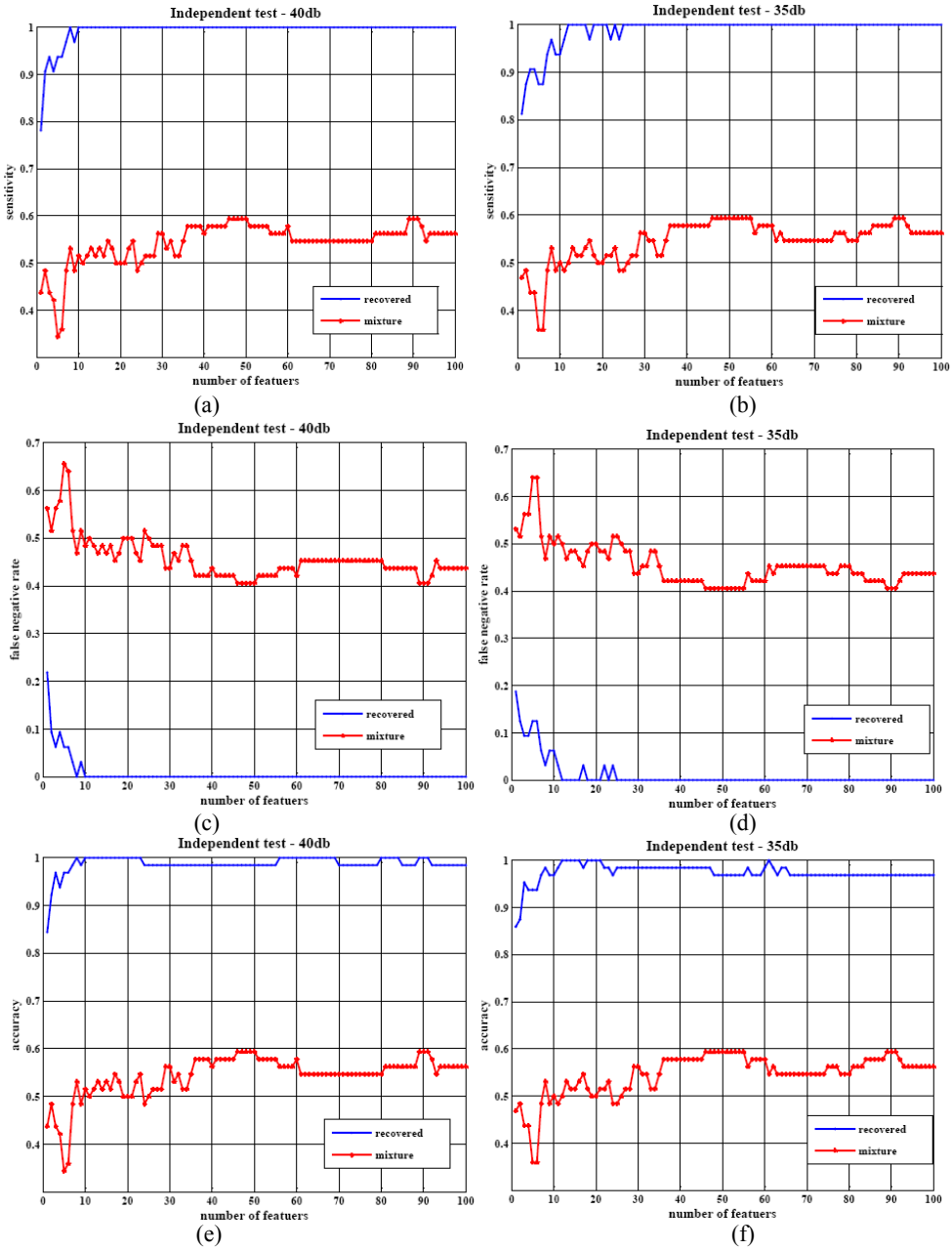


Figure B.8: Results of independent tests for recovered signals and mixtures on noisy sources with additive observation noises: (a) sensitivity curves with SNR = 40dB; (b) sensitivity curves with SNR = 35dB; (c) false negative rate curves with SNR = 40dB; (d) false negative rate curves with SNR = 35dB; (e) overall classification accuracy curves with SNR = 40dB; (f) overall classification accuracy curves with SNR = 35dB.

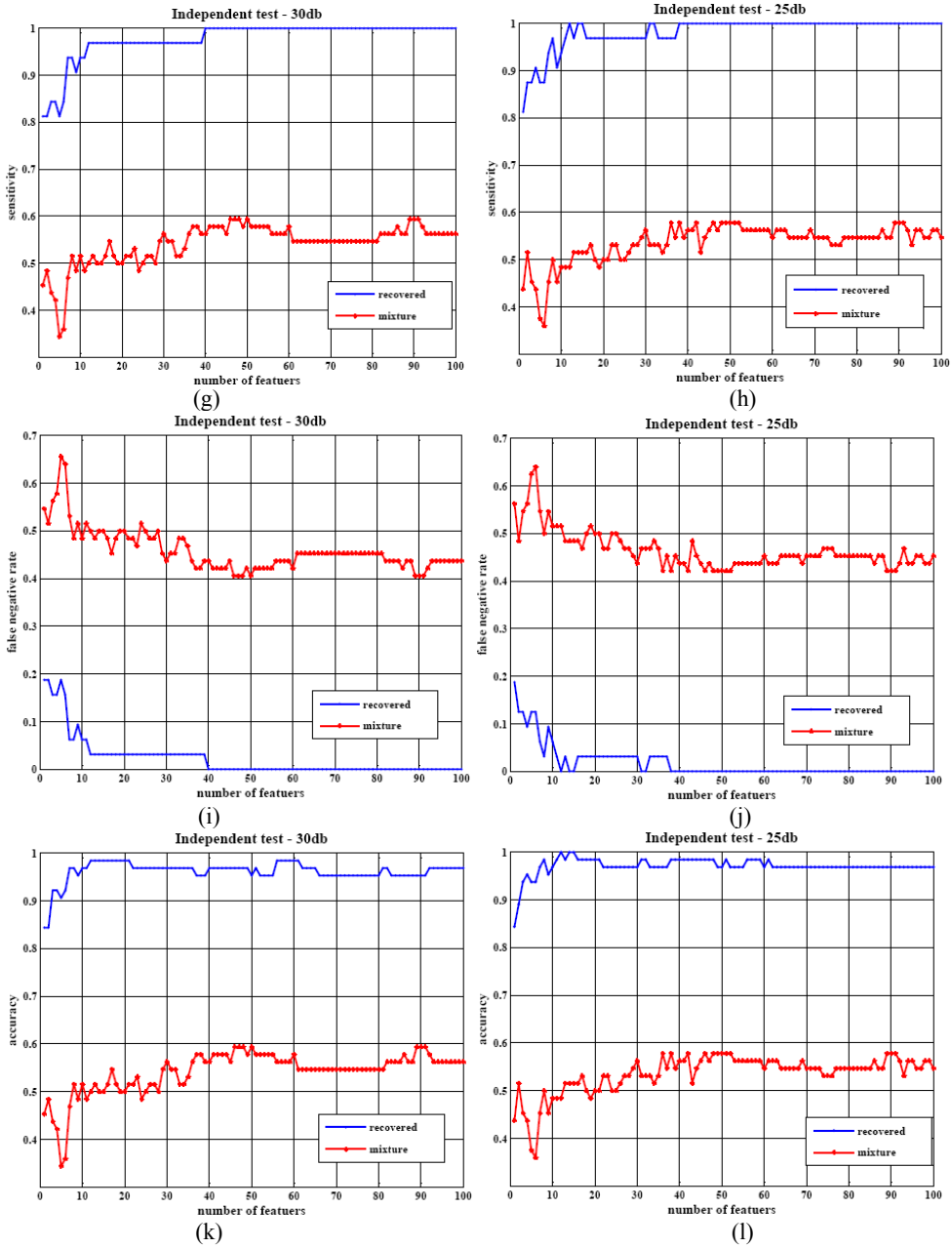


Figure B.8 (cont'd). Results of independent tests for recovered signals and mixtures on noisy sources with additive observation noises: (g) sensitivity curves with SNR = 30dB; (h) sensitivity curves with SNR = 25dB; (i) false negative rate curves with SNR = 30dB; (j) false negative rate curves with SNR = 25dB; (k) overall classification accuracy curves with SNR = 30dB; (l) overall classification accuracy curves with SNR = 25dB.

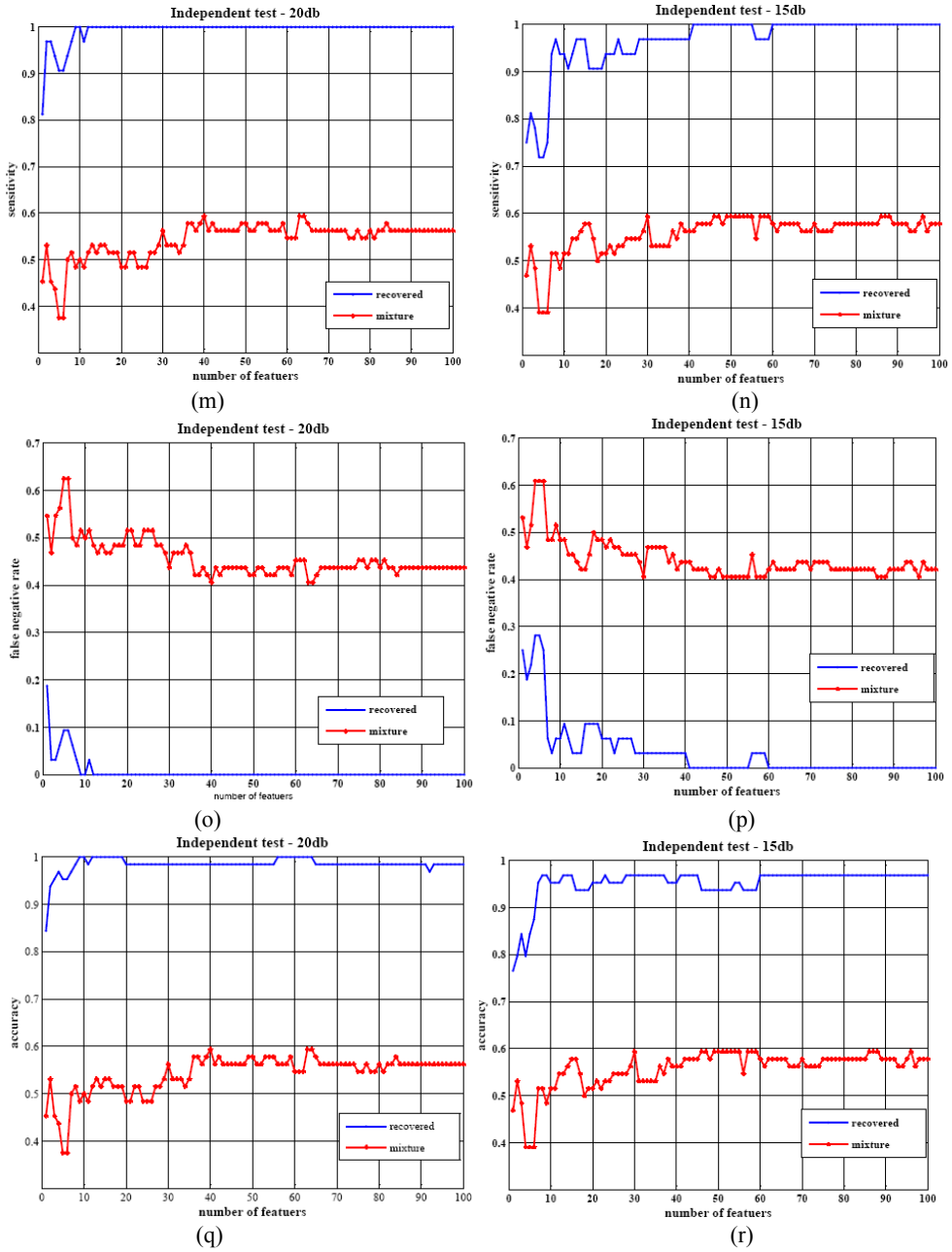
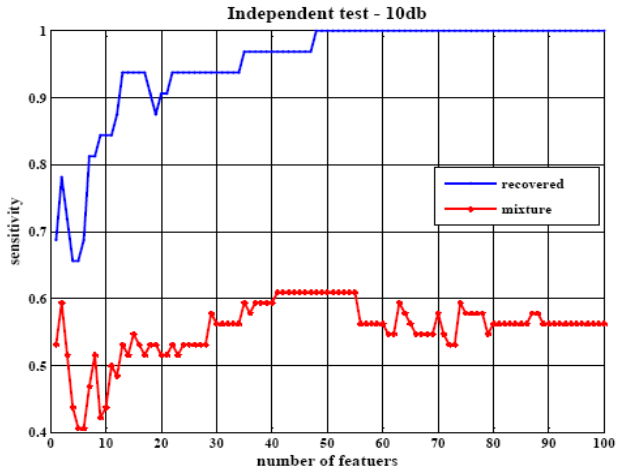
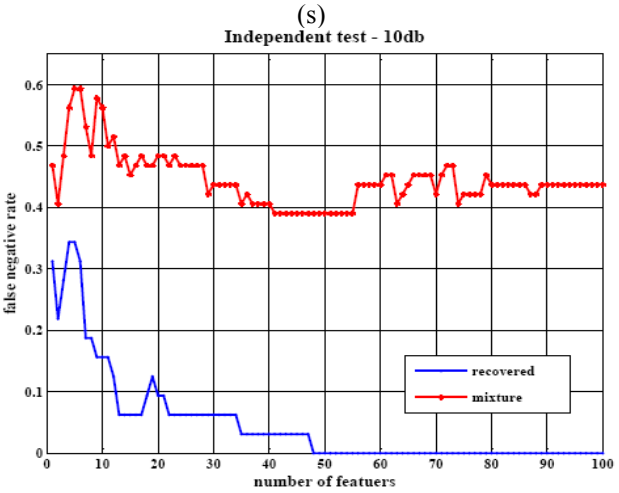


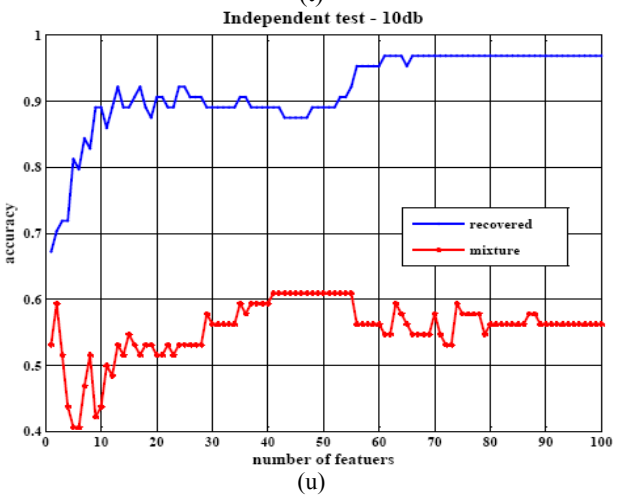
Figure B.8 (cont'd). Results of independent tests for recovered signals and mixtures on noisy sources with additive observation noises: (m) sensitivity curves with SNR = 20dB; (n) sensitivity curves with SNR = 15dB; (o) false negative rate curves with SNR = 20dB; (p) false negative rate curves with SNR = 15dB; (q) overall classification accuracy curves with SNR = 20dB; (r) overall classification accuracy curves with SNR = 15dB.



(s)



(t)



(u)

Figure B.8 (cont'd). Results of independent tests for recovered signals and mixtures noisy sources with 10dB additive Gaussian noise: (s) sensitivity curves; (t) false negative rate curves; (u) overall classification accuracy curves.

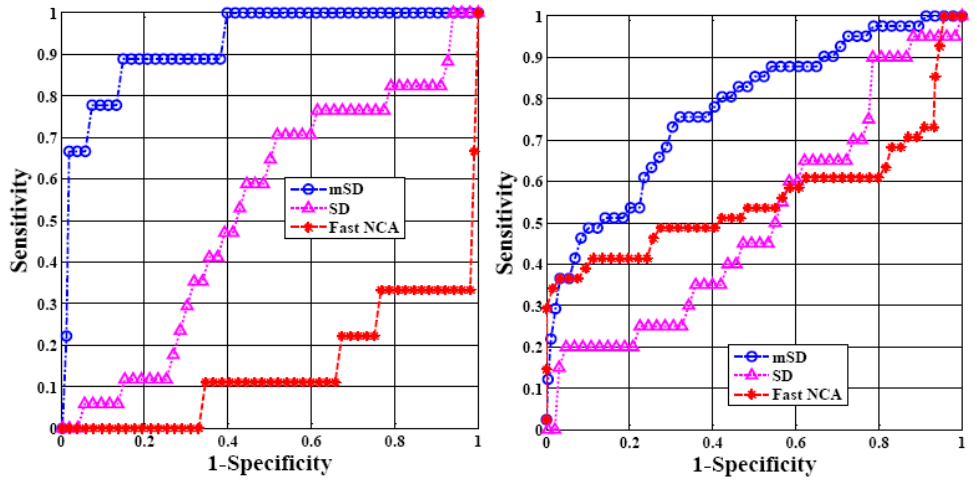
Table B.1 Mixing information of 64 mixture samples generated from 32 patients' original pure microarray data. The mixing ratios listed here are corresponding to normal tissues (N) *vs*. adenomas tissues (A).

patient #	mixing ratio (N *vs*. A)	mixture #	patient #	mixing ratio (N *vs*. A)	mixture #
patient #1	0.2614 vs. 0.7386	mixture 1	patient #17	0.3095 vs. 0.6905	mixture 33
	0.6578 vs. 0.3422	mixture 2		0.8114 vs. 0.1886	mixture 34
patient #2	0.4673 vs. 0.5327	mixture 3	patient #18	0.9054 vs. 0.0946	mixture 35
	0.9417 vs. 0.0583	mixture 4		0.4158 vs. 0.5842	mixture 36
patient #3	0.3004 vs. 0.6996	mixture 5	patient #19	0.4277 vs. 0.5723	mixture 37
	0.6252 vs. 0.3748	mixture 6		0.8550 vs. 0.1450	mixture 38
patient #4	0.5704 vs. 0.4296	mixture 7	patient #20	0.3197 vs. 0.6803	mixture 39
	0.1400 vs. 0.8600	mixture 8		0.5116 vs. 0.4884	mixture 40
patient #5	0.2752 vs. 0.7248	mixture 9	patient #21	0.3964 vs. 0.6036	mixture 41
	0.5405 vs. 0.4595	mixture 10		0.6656 vs. 0.3344	mixture 42
patient #6	0.7553 vs. 0.2447	mixture 11	patient #22	0.3503 vs. 0.6497	mixture 43
	0.3724 vs. 0.6276	mixture 12		0.6270 vs. 0.3730	mixture 44
patient #7	0.6576 vs. 0.3424	mixture 13	patient #23	0.7263 vs. 0.2737	mixture 45
	0.4521 vs. 0.5479	mixture 14		0.1462 vs. 0.8538	mixture 46
patient #8	0.2803 vs. 0.7197	mixture 15	patient #24	0.5380 vs. 0.4620	mixture 47
	0.5598 vs. 0.4402	mixture 16		0.3117 vs. 0.6883	mixture 48
patient #9	0.5386 vs. 0.4614	mixture 17	patient #25	0.3212 vs. 0.6788	mixture 49
	0.4622 vs. 0.5378	mixture 18		0.5683 vs. 0.4317	mixture 50
Patient #10	0.7231 vs. 0.2769	mixture 19	patient #26	0.2673 vs. 0.7327	mixture 51
	0.4680 vs. 0.5320	mixture 20		0.8152 vs. 0.1848	mixture 52
Patient #11	0.2468 vs. 0.7532	mixture 21	patient #27	0.5525 vs. 0.4475	mixture 53
	0.9227 vs. 0.0773	mixture 22		0.0531 vs. 0.9469	mixture 54
Patient #12	0.5900 vs. 0.4100	mixture 23	patient #28	0.5513 vs. 0.4487	mixture 55
	0.2482 vs. 0.7518	mixture 24		0.4962 vs. 0.5038	mixture 56
Patient #13	0.9697 vs. 0.0303	mixture 25	patient #29	0.4502 vs. 0.5498	mixture 57
	0.4158 vs. 0.5842	mixture 26		0.5358 vs. 0.4642	mixture 58
Patient #14	0.4763 vs. 0.5237	mixture 27	patient #30	0.4692 vs. 0.5308	mixture 59
	0.8421 vs. 0.1579	mixture 28		0.7058 vs. 0.2942	mixture 60
Patient #15	0.8587 vs. 0.1413	mixture 29	patient #31	0.3825 vs. 0.6175	mixture 61
	0.0014 vs. 0.9986	mixture 30		0.9041 vs. 0.0959	mixture 62
Patient #16	0.3261 vs. 0.6739	mixture 31	patient #32	0.5930 vs. 0.4070	mixture 63
	0.5059 vs. 0.4941	mixture 32		0.3318 vs. 0.6682	mixture 64

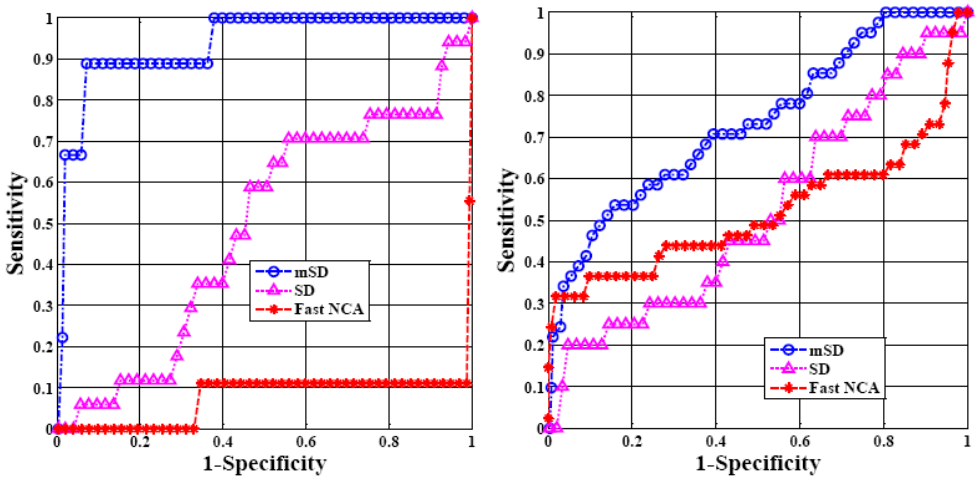
Appendix C. Addendum of Empirical Results for Chapter 4

In this appendix, we include various empirical results which are not reported in the main text of Chapter 4.

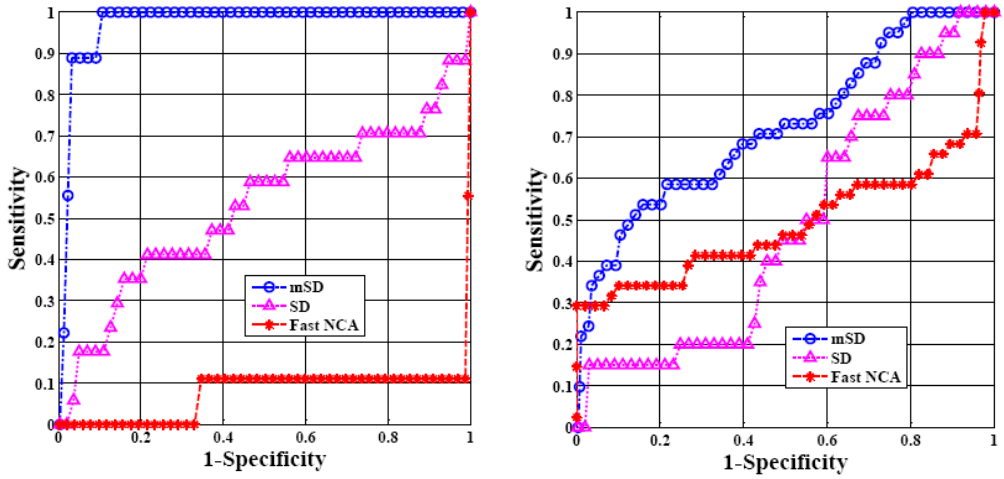
They include ROC curves for individual TF (Figure C.1); Part of PPI sub-networks of target genes of ETF identified in estrogen-induced and estrogen-deprived conditions (Figure C.2); detailed AUC Comparison of mSD, SD and FastNCA methods (Table C.1); a set of identified key transcription factors (Table C.2) and Target genes of ETF (Table C.3).



(a) cut-off *p*-value = 0.1



(b) cut-off *p*-value = 0.05



(c) cut-off *p*-value = 0.01

Figure C.1: Performance comparison of mSD, SD and FastNCA methods - ROC curves for the identified regulatory modules of HAP1 (left) and STE12 (right), respectively. Three different cut-off p-values (0.1,

0.05 and 0.01) have been applied to ChIP-on-chip data for investigating the noise impact on the performance.

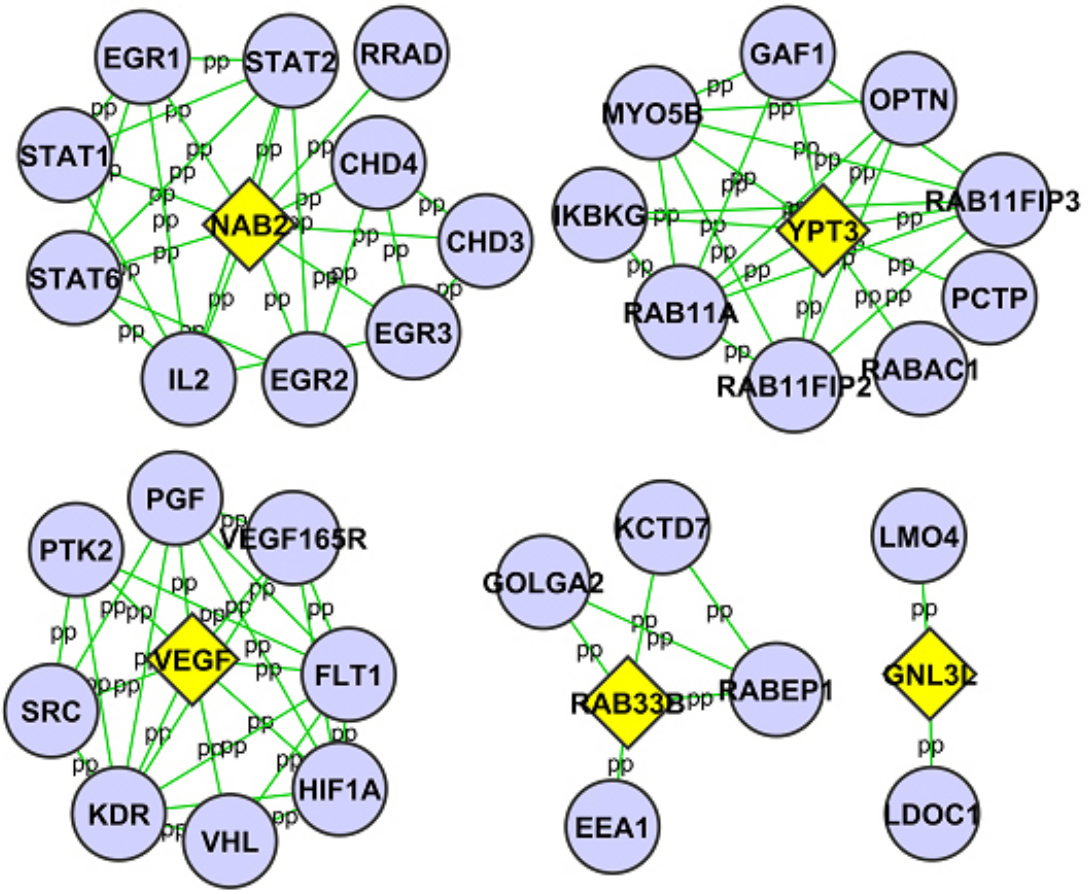


Figure C.2 More PPI sub-networks of target genes of ETF identified in estrogen-induced and estrogen-deprived conditions. Yellow diamond: target genes of ETF; purple circle: direct neighbors of the target genes as obtained from protein-protein interaction data.

In Figure C.2, we found several proteins such as YPT3, Rab33B related to Ras signal transduction pathway. Since Ras-dependent events appear to be activated as a consequence of EGFR mutations in cancer cells [183], it is possible that the aberrant function of Ras-related proteins may contribute to breast cancer development [184] by a network of proto-oncogene proteins controlling diverse signaling events that regulate cell growth and differentiation defined in Ras signal transduction pathway.

Table C.1 AUC Comparison of mSD, SD and FastNCA methods for 11 transcription factors in yeast synthetic data

Cut-off *p*-value	Method	ARG80	DAL82	GCN4	GCR2
***p* = 0.1**	**mSD**	0.5892	0.7057	0.8204	0.7827
	FastNCA	0.4634	0.4466	0.8129	0.6807
	SD	0.4984	0.9290	0.8730	0.6485
***p* = 0.05**	**mSD**	0.5311	0.6859	0.8344	0.7811
	FastNCA	0.4554	0.4418	0.8099	0.4196
	SD	0.5916	0.7607	0.8784	0.6341
***p* = 0.01**	**mSD**	0.7779	0.7105	0.8154	0.7938
	FastNCA	0.4618	0.4493	0.7974	0.4276
	SD	0.5255	0.7660	0.8613	0.6180

Cut-off *p*-value	Method	HAP1	MIG1	RGT1	RTG1
***p* = 0.1**	**mSD**	0.9238	0.8390	0.8180	0.3281
	FastNCA	0.1435	0.5531	0.3873	0.8297
	SD	0.5226	0.9022	0.7246	0.8180
***p* = 0.05**	**mSD**	0.9335	0.8251	0.8222	0.7382
	FastNCA	0.0795	0.5234	0.7175	0.8454
	SD	0.4858	0.8927	0.7422	0.8688
***p* = 0.01**	**mSD**	0.9738	0.8136	0.9058	0.7413
	FastNCA	0.0773	0.4725	0.3831	0.8612
	SD	0.5304	0.8801	0.7418	0.8730

Cut-off *p*-value	Method	RTG3	STE12	XBP1	Averaged AUC over TFs
***p* = 0.1**	**mSD**	0.4890	0.7487	0.8318	**0.7160**
	FastNCA	0.8612	0.5318	0.5672	0.5707
	SD	0.4483	0.5055	0.7326	0.6912
***p* = 0.05**	**mSD**	0.8454	0.7325	0.8497	**0.7799**
	FastNCA	0.8517	0.4960	0.8402	0.5891
	SD	0.4704	0.5211	0.7233	0.6881
***p* = 0.01**	**mSD**	0.8265	0.7234	0.7439	**0.8024**
	FastNCA	0.8644	0.4700	0.8370	0.5547
	SD	0.4735	0.4836	0.7279	0.6801

We first identified a set of key transcription factors previously known to be involved in the Estrogen Receptor Signaling: AP-1, CREB, ER, NFκB, STATs [185]; authentic *cis* binding sites in breast cancer cell lines: C/EBP, Forkhead [168]; or playing a role in overexpression in estrogen receptor (ER)-positive breast tumors: EGR-1[186], ETF [171], MYB[187], P53[188]. Meanwhile, we also included some motifs involved in cell cycle or apoptosis: MYCMAX[189], NFY[190], PBX1 [191]. Then for each identified TF, a PWM was chosen from the vertebrate non-redundant profiles (PWM) of TRANSFAC [146] database. Further motif information was obtained from ChIP-on-chip experiments [168], and a final list of 26 transcription factors is given in the following:

Table C.2 Identified key transcription factors utilized in the breast cancer cell line data

V$AP1_Q2_01	V$AP1_Q4_01	V$CREBP1CJUN_01	V$CEBP_Q3	V$CEBPA_01
V$CEBPGAMMA_Q6	V$CREB_02	V$CREB_Q3	V$CREB_Q2	V$NFKB_Q6_01
V$SP1_Q6	V$ER_Q6	V$ETF_Q6	V$MYCMAX_03	V$STAT_Q6
V$STAT_01	V$EGR1_01	V$FOXJ2_02	V$FOXP1_01	V$MYB_Q3
V$P53_02	V$PBX_Q3	V$PBX1_03	V$NFY_Q6_01	V$NFY_01
V$CEBPDELTA_Q6				

Table C.3 Target genes of ETF (V$ETF_Q6) in both E2-induced and ER-deprived conditions

Probe Set ID	GENE_SYMBOL	Gene Name
200646_S_AT	NUCB1	NUCLEOBINDIN 1
200690_AT	HSPA9	HEAT SHOCK 70KDA PROTEIN 9B (MORTALIN-2)
201373_AT	PLEC1	PLECTIN 1, INTERMEDIATE FILAMENT BINDING PROTEIN 500KDA
201573_S_AT	ETF1	EUKARYOTIC TRANSLATION TERMINATION FACTOR 1
201601_X_AT	IFITM1	INTERFERON INDUCED TRANSMEMBRANE PROTEIN 1 (9-27)
201753_S_AT	ADD3	ADDUCIN 3 (GAMMA)
201842_S_AT	EFEMP1	EGF-CONTAINING FIBULIN-LIKE EXTRACELLULAR MATRIX PROTEIN 1
201910_AT	FARP1	FERM, RHOGEF (ARHGEF) AND PLECKSTRIN DOMAIN PROTEIN 1 (CHONDROCYTE-DERIVED)
201984_S_AT	EGFR	EPIDERMAL GROWTH FACTOR RECEPTOR (ERYTHROBLASTIC LEUKEMIA VIRAL (V-ERB-B) ONCOGENE HOMOLOG, AVIAN)
202088_AT	SLC39A6	SOLUTE CARRIER FAMILY 39 (ZINC TRANSPORTER), MEMBER 6
202235_AT	SLC16A1	SOLUTE CARRIER FAMILY 16 (MONOCARBOXYLIC ACID TRANSPORTERS), MEMBER 1
202295_S_AT	CTSH	CATHEPSIN H
202304_AT	FNDC3A	FIBRONECTIN TYPE III DOMAIN CONTAINING 3A
202429_S_AT	PPP3CA	PROTEIN PHOSPHATASE 3 (FORMERLY 2B), CATALYTIC SUBUNIT, ALPHA ISOFORM (CALCINEURIN A ALPHA)
202602_S_AT	HTATSF1	HIV-1 TAT SPECIFIC FACTOR 1
202730_S_AT	PDCD4	PROGRAMMED CELL DEATH 4 (NEOPLASTIC

		TRANSFORMATION INHIBITOR)
202826_AT	SPINT1	SERINE PEPTIDASE INHIBITOR, KUNITZ TYPE 1
202979_S_AT	CREBZF	HCF-BINDING TRANSCRIPTION FACTOR ZHANGFEI
203079_S_AT	CUL2	CULLIN 2
203278_S_AT	PHF21A	PHD FINGER PROTEIN 21A
203358_S_AT	EZH2	ENHANCER OF ZESTE HOMOLOG 2 (DROSOPHILA)
203456_AT	PRAF2	PRA1 DOMAIN FAMILY, MEMBER 2
203493_S_AT	CEP57	CENTROSOMAL PROTEIN 57KDA
203607_AT	INPP5F	INOSITOL POLYPHOSPHATE-5-PHOSPHATASE F
203855_AT	WDR47	WD REPEAT DOMAIN 47
203869_AT	USP46	UBIQUITIN SPECIFIC PEPTIDASE 46
204129_AT	BCL9	B-CELL CLL/LYMPHOMA 9
204527_AT	MYO5A	MYOSIN VA (HEAVY POLYPEPTIDE 12, MYOXIN)
204629_AT	PARVB	PARVIN, BETA
204710_S_AT	WIPI2	WD REPEAT DOMAIN, PHOSPHOINOSITIDE INTERACTING 2
204989_S_AT	ITGB4	INTEGRIN, BETA 4
204995_AT	CDK5R1	CYCLIN-DEPENDENT KINASE 5, REGULATORY SUBUNIT 1 (P35)
205222_AT	EHHADH	ENOYL-COENZYME A, HYDRATASE/3-HYDROXYACYL COENZYME A DEHYDROGENASE
205258_AT	INHBB	INHIBIN, BETA B (ACTIVIN AB BETA POLYPEPTIDE)
206231_AT	KCNN1	POTASSIUM INTERMEDIATE/SMALL CONDUCTANCE CALCIUM-ACTIVATED CHANNEL, SUBFAMILY N, MEMBER 1
206574_S_AT	PTP4A3	PROTEIN TYROSINE PHOSPHATASE TYPE IVA, MEMBER 3
206604_AT	OVOL1	OVO-LIKE 1(DROSOPHILA)
207038_AT	SLC16A6	SOLUTE CARRIER FAMILY 16 (MONOCARBOXYLIC ACID TRANSPORTERS), MEMBER 6
207844_AT	IL13	INTERLEUKIN 13
208296_X_AT	TNFAIP8	TUMOR NECROSIS FACTOR, ALPHA-INDUCED PROTEIN 8
208754_S_AT	NAP1L1	NUCLEOSOME ASSEMBLY PROTEIN 1-LIKE 1
208876_S_AT	PAK2	P21 (CDKN1A)-ACTIVATED KINASE 2
209135_AT	ASPH	ASPARTATE BETA-HYDROXYLASE
209241_X_AT	MINK1	MISSHAPEN-LIKE KINASE 1 (ZEBRAFISH)
209288_S_AT	CDC42EP3	CDC42 EFFECTOR PROTEIN (RHO GTPASE BINDING) 3
209354_AT	TNFRSF14	TUMOR NECROSIS FACTOR RECEPTOR SUPERFAMILY, MEMBER 14 (HERPESVIRUS ENTRY MEDIATOR)
209736_AT	SOX13	SRY (SEX DETERMINING REGION Y)-BOX 13
209872_S_AT	PKP3	PLAKOPHILIN 3
209900_S_AT	SLC16A1	SOLUTE CARRIER FAMILY 16 (MONOCARBOXYLIC ACID TRANSPORTERS), MEMBER 1
209988_S_AT	ASCL1	ACHAETE-SCUTE COMPLEX-LIKE 1 (DROSOPHILA)
210184_AT	ITGAX	INTEGRIN, ALPHA X (COMPLEMENT COMPONENT 3 RECEPTOR 4 SUBUNIT)
210513_S_AT	VEGFA	VASCULAR ENDOTHELIAL GROWTH FACTOR
210854_X_AT	SLC6A8	SOLUTE CARRIER FAMILY 6 (NEUROTRANSMITTER TRANSPORTER, CREATINE), MEMBER 8
211097_S_AT	PBX2	PRE-B-CELL LEUKEMIA TRANSCRIPTION FACTOR 2
211527_X_AT	VEGFA	VASCULAR ENDOTHELIAL GROWTH FACTOR
212375_AT	EP400	TRINUCLEOTIDE REPEAT CONTAINING 12
212467_AT	DNAJC13	DNAJ (HSP40) HOMOLOG, SUBFAMILY C, MEMBER 13
212594_AT	PDCD4	PROGRAMMED CELL DEATH 4 (NEOPLASTIC TRANSFORMATION INHIBITOR)
212739_S_AT	NME4	NON-METASTATIC CELLS 4, PROTEIN EXPRESSED IN
212837_AT	KIAA0157	KIAA0157
212878_S_AT	KLC1	KINESIN 2
213051_AT	ZC3HAV1	ZINC FINGER CCCH-TYPE, ANTIVIRAL 1

213187_X_AT	FTLL1	FERRITIN, LIGHT POLYPEPTIDE
213271_S_AT	DOPEY1	DOPEY FAMILY MEMBER 1
213451_X_AT	TNXB	TENASCIN XB
213505_S_AT	SFRS14	SPLICING FACTOR, ARGININE/SERINE-RICH 14
213756_S_AT	HSF1	HEAT SHOCK TRANSCRIPTION FACTOR 1
213757_AT	EIF5A	EUKARYOTIC TRANSLATION INITIATION FACTOR 5A
213856_AT	CD47	CD47 ANTIGEN (RH-RELATED ANTIGEN, INTEGRIN-ASSOCIATED SIGNAL TRANSDUCER)
214095_AT	SHMT2	SERINE HYDROXYMETHYLTRANSFERASE 2 (MITOCHONDRIAL)
214437_S_AT	SHMT2	SERINE HYDROXYMETHYLTRANSFERASE 2 (MITOCHONDRIAL)
214697_S_AT	ROD1	ROD1 REGULATOR OF DIFFERENTIATION 1 (S. POMBE)
215735_S_AT	TSC2	TUBEROUS SCLEROSIS 2
216017_S_AT	NAB2	NGFI-A BINDING PROTEIN 2 (EGR1 BINDING PROTEIN 2)
216080_S_AT	FADS3	FATTY ACID DESATURASE 3
216237_S_AT	MCM5	MCM5 MINICHROMOSOME MAINTENANCE DEFICIENT 5, CELL DIVISION CYCLE 46 (S. CEREVISIAE)
217693_X_AT	LOC388335	SIMILAR TO RIKEN CDNA A730055C05 GENE
217928_S_AT	SAPS3	CHROMOSOME 11 OPEN READING FRAME 23
218807_AT	VAV3	VAV 3 ONCOGENE
218887_AT	MRPL2	MITOCHONDRIAL RIBOSOMAL PROTEIN L2
218889_AT	NOC3L	NUCLEOLAR COMPLEX ASSOCIATED 3 HOMOLOG (S. CEREVISIAE)
219829_AT	ITGB1BP2	INTEGRIN BETA 1 BINDING PROTEIN (MELUSIN) 2
220116_AT	KCNN2	POTASSIUM INTERMEDIATE/SMALL CONDUCTANCE CALCIUM-ACTIVATED CHANNEL, SUBFAMILY N, MEMBER 2
221014_S_AT	RAB33B	RAB33B, MEMBER RAS ONCOGENE FAMILY
221926_S_AT	IL17RC	INTERLEUKIN 17 RECEPTOR C
222071_S_AT	SLCO4C1	HYPOTHETICAL PROTEIN PRO2176
46947_AT	GNL3L	GUANINE NUCLEOTIDE BINDING PROTEIN-LIKE 3 (NUCLEOLAR)-LIKE

Bibliography

[1] D. Greenbaum, N. M. Luscombe, R. Jansen, J. Qian, and M. Gerstein, "Interrelating different types of genomic data, from proteome to secretome: 'oming in on function," *Genome Res*, vol. 11, pp. 1463-8, 2001.

[2] A. Hasman, "Description of a blockcourse in medical informatics," *Methods Inf Med*, vol. 28, pp. 239-42, 1989.

[3] F. S. Collins, E. D. Green, A. E. Guttmacher, and M. S. Guyer, "A vision for the future of genomics research," *Nature*, vol. 422, pp. 835-847, 2003.

[4] E. Segal, N. Friedman, N. Kaminski, A. Regev, and D. Koller, "From signatures to models: understanding cancer using microarrays," *Nat Genet*, vol. 37, pp. S38-S45, 2005.

[5] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nat Med*, vol. 7, pp. 673-679, 2001.

[6] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, vol. 286, pp. 531-537, 1999.

[7] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, pp. 5116-5121, 2001.

[8] 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," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp. 15522-15527, 2003.

[9] 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. 166-176, 2003.

[10] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J.-B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young, "Transcriptional regulatory code of a eukaryotic genome," *Nature*, vol. 431, pp. 99-104, 2004.

[11] 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, 2002.

[12] J. D. Watson and F. H. Crick, "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid," *Nature*, vol. 171, pp. 737-8, 1953.

[13] F. Crick, "Central dogma of molecular biology," *Nature*, vol. 227, pp. 561-3, 1970.

[14] L. Marchionni, R. F. Wilson, S. S. Marinopoulos, A. C. Wolff, G. Parmigiani, E. B. Bass, and S. N. Goodman, "Impact of gene expression profiling tests on breast cancer outcomes," *Evid Rep Technol Assess (Full Rep)*, pp. 1-105, 2007.

[15] F. Liu, T. K. Jenssen, J. Trimarchi, C. Punzo, C. L. Cepko, L. Ohno-Machado, E. Hovig, and W. P. Kuo, "Comparison of hybridization-based and sequencing-based gene expression technologies on biological replicates," *BMC Genomics*, vol. 8, pp. 153, 2007.

[16] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler, "Serial analysis of gene expression," *Science*, vol. 270, pp. 484-7, 1995.

[17] T. T. Torres, M. Metta, B. Ottenwalder, and C. Schlotterer, "Gene expression profiling by massively parallel sequencing," *Genome Res*, vol. 18, pp. 172-7, 2008.

[18] E. P. e. a. T. T. A. B. P. W. G. Hoffman, "Expression profiling - best practices for data generation and interpretation in clinical trials," *Nat Rev Genet*, vol. 5, pp. 229-237, 2004.

[19] M. J. Buck and J. D. Lieb, "ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments," *Genomics*, vol. 83, pp. 349-60, 2004.

[20] J. Herrero, J. M. Vaquerizas, F. Al-Shahrour, L. Conde, A. Mateos, J. S. R. Diaz-Uriarte, and J. Dopazo, "New challenges in gene expression data analysis and the extended GEPAS," *Nucl. Acids Res.*, vol. 32, pp. W485-491, 2004.

[21] D. K. Slonim, "From patterns to pathways: gene expression data analysis comes of age," *Nat Genet*, vol. 32, pp. 502-508, 2002.

[22] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, pp. 249-64, 2003.

[23] Y. H. Yang, M. J. Buckley, S. Dudoit, and T. P. Speed, "Comparison of Methods for Image Analysis on cDNA Microarray Data," *Journal of Computational & Graphical Statistics*, vol. 11, pp. 108-136, 2002.

[24] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, Jr., J. R. Marks, and J. R. Nevins, "Predicting the clinical status of human breast cancer by using gene expression profiles," *Proc Natl Acad Sci U S A*, vol. 98, pp. 11462-7, 2001.

[25] 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 Biotech*, vol. 21, pp. 1337-1342, 2003.

[26] A. R. Abbas, K. Wolslegel, D. Seshasayee, Z. Modrusan, and H. F. Clark, "Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus," *PLoS One*, vol. 4, pp. e6098, 2009.

[27] A. Schulze and J. Downward, "Navigating gene expression using microarrays [mdash] a technology review," *Nat Cell Biol*, vol. 3, pp. E190-E195, 2001.

[28] M. Bakay, Y.-W. Chen, R. Borup, P. Zhao, K. Nagaraju, and E. Hoffman, "Sources of variability and effect of experimental approach on expression profiling data interpretation," *BMC Bioinformatics*, vol. 3, pp. 4, 2002.

[29] M. R. Emmert-Buck, R. F. Bonner, P. D. Smith, R. F. Chuaqui, Z. Zhuang, S. R. Goldstein, R. A. Weiss, and L. A. Liotta, "Laser capture microdissection," *Science*, vol. 274, pp. 998-1001, 1996.

[30] P. Lu, A. Nakorchevskiy, and E. M. Marcotte, "Expression deconvolution: A reinterpretation of DNA microarray data reveals dynamic changes in cell populations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp. 10370-10375, 2003.

[31] R. O. Stuart, W. Wachsman, C. C. Berry, J. Wang-Rodriguez, L. Wasserman, I. Klacansky, D. Masys, K. Arden, S. Goodison, M. McClelland, Y. Wang, A. Sawyers, I. Kalcheva, D. Tarin, and D. Mercola, "In silico dissection of cell-type-associated patterns of gene expression in prostate cancer," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, pp. 615-620, 2004.

[32] Y. Wang, "Independent Component Imaging, US Patent #6,728,396, April 27, 2004," 2004.

[33] Y. Wang, J. Zhang, J. Khan, R. Clarke, and Z. Gu, "Partially-independent component analysis for tissue heterogeneity correction in microarray gene expression analysis," presented at Proc. IEEE Workshop on Machine Learning for Signal Processing, Toulouse, France, 2003.

[34] Y. Wang, J. Zhang, K. Huang, J. Khan, and Z. Szabo, "Independent component imaging of disease signatures," presented at Biomedical Imaging, 2002. Proceedings. 2002 IEEE International Symposium on, Washington DC, 2002.

[35] G. A. Churchill, "Using ANOVA to analyze microarray data," *Biotechniques*, vol. 37, pp. 173-5, 177, 2004.

[36] R. J. Fox and M. W. Dimmic, "A two-sample Bayesian t-test for microarray data," *BMC Bioinformatics*, vol. 7, pp. 126, 2006.

[37] J. Quackenbush, "Computational analysis of microarray data," *Nat Rev Genet*, vol. 2, pp. 418-427, 2001.

[38] J. T. Leek, E. Monsen, A. R. Dabney, and J. D. Storey, "EDGE: extraction and analysis of differential gene expression," *Bioinformatics*, vol. 22, pp. 507-8, 2006.

[39] 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, 1998.

[40] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, pp. 2907-2912, 1999.

[41] C. Niehrs and N. Pollet, "Synexpression groups in eukaryotes," *Nature*, vol. 402, pp. 483-487, 1999.

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

[43] S.-I. Lee and S. Batzoglou, "Application of independent component analysis to microarrays," *Genome Biology*, vol. 4, pp. R76, 2003.

[44] 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, pp. 290, 2006.

[45] A. E. Teschendorff, M. Journée, P. A. Absil, R. Sepulchre, and C. Caldas, "Elucidating the Altered Transcriptional Programs in Breast Cancer using Independent Component Analysis," *PLoS Computational Biology*, vol. 3, pp. e161, 2007.

[46] H. Li, Y. Sun, and M. Zhan, "The discovery of transcriptional modules by a two-stage matrix decomposition approach," *Bioinformatics*, vol. 23, pp. 473-9, 2007.

[47] G. Ting, Z. Yitan, X. Jianhua, L. Huai, R. Clarke, E. P. Hoffman, and W. Yue, "Latent Variable and nICA Modeling of Pathway Gene Module Composite," 2006.

[48] E. Oja and M. Plumbley, "Blind separation of positive sources by globally convergent gradient search," *Neural Computation*, vol. 16, pp. 1811-1825, 2004.

[49] 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.

[50] A. A. Margolin, T. Palomero, P. Sumazin, A. Califano, A. A. Ferrando, and G. Stolovitzky, "ChIP-on-chip significance analysis reveals large-scale binding and regulation by human transcription factor oncogenes," *Proc Natl Acad Sci U S A*, vol. 106, pp. 244-9, 2009.

[51] 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. 2306-2309, 2000.

[52] J. D. Hoheisel, "Microarray technology: beyond transcript profiling and genotype analysis," *Nat Rev Genet*, vol. 7, pp. 200-10, 2006.

[53] K. Lin and D. Husmeier, "Modelling transcriptional regulation with a mixture of factor analyzers and variational bayesian expectation maximization," *EURASIP journal on bioinformatics & systems biology*, 2009.

[54] P. Georgiev, F. Theis, and A. Cichocki, "Sparse component analysis and blind source separation of underdetermined mixtures," *Neural Networks, IEEE Transactions on*, vol. 16, pp. 992-996, 2005.

[55] T. Gong, J. Xuan, L. Chen, R. Riggins, Y. Wang, E. Hoffman, and R. Clarke, "Sparse Decomposition of Gene Expression Data to Infer Transcriptional Modules Guided by Motif Information," in *Bioinformatics Research and Applications*, 2008, pp. 244-255.

[56] Y. Zhu, H. Li, D. Miller, Z. Wang, J. Xuan, R. Clarke, E. Hoffman, and Y. Wang, "caBIGTM VISDA: Modeling, visualization, and discovery for cluster analysis of genomic data," *BMC Bioinformatics*, vol. 9, pp. 383, 2008.

[57] J. R. Nevins and A. Potti, "Mining gene expression profiles: expression signatures as cancer phenotypes," *Nat Rev Genet*, vol. 8, pp. 601-9, 2007.

[58] L. Liotta and E. Petricoin, "Molecular profiling of human cancer," *Nat Rev Genet*, vol. 1, pp. 48-56, 2000.

[59] W. B. Coleman and G. J. Tsongalis, *Molecular Pathology: The Molecular Basis of Human Disease*: Elsevier Science & Technology Books, 2009.

[60] M. Wang, S. R. Master, and L. A. Chodosh, "Computational expression deconvolution in a complex mammalian organ," *BMC Bioinformatics*, vol. 7, pp. 328, 2006.

[61] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J. P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander, and T. R. Golub, "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease," *Science*, vol. 313, pp. 1929-35, 2006.

[62] C. M. Perou, T. Sorlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lonning, A. L. Borresen-Dale, P. O. Brown, and D. Botstein, "Molecular portraits of human breast tumours," *Nature*, vol. 406, pp. 747-52, 2000.

[63] T. Sorlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C. M. Perou, P. E. Lonning, P. O. Brown, A. L. Borresen-Dale, and D. Botstein, "Repeated observation of breast tumor subtypes in independent gene expression data sets," *Proc Natl Acad Sci U S A*, vol. 100, pp. 8418-23, 2003.

[64] C. D. Stern and S. E. Fraser, "Tracing the lineage of tracing cell lineages," *Nat Cell Biol*, vol. 3, pp. E216-8, 2001.

[65] E. Staub, J. Groene, M. Heinze, D. Mennerich, S. Roepcke, I. Klaman, B. Hinzmann, E. Castanos-Velez, C. Pilarsky, B. Mann, T. Brummendorf, B. Weber, H. J. Buhr, and A. Rosenthal, "Genome-wide expression patterns of invasion front, inner tumor mass and surrounding normal epithelium of colorectal tumors," *Mol Cancer*, vol. 6, pp. 79, 2007.

[66] 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. 3273-3297, 1998.

[67] L. K. Hansen and X. Guanghan, "A hyperplane-based algorithm for the digital co-channel communications problem," *Information Theory, IEEE Transactions on*, vol. 43, pp. 1536-1548, 1997.

[68] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis* John Wiley & Sons, 2001.

[69] S. Moussaoui, D. Brie, A. Mohammad-Djafari, and C. Carteret, "Separation of Non-Negative Mixture of Non-Negative Sources Using a Bayesian Approach and MCMC Sampling," *Signal Processing, IEEE Transactions on*, vol. 54, pp. 4133-4145, 2006.

[70] P. Comon, "Independent component analysis, A new concept?," *Signal Processing*, vol. 36, pp. 287-314, 1994.

[71] Y. Wang, J. Lu, R. Lee, Z. Gu, and R. Clarke, "Iterative normalization of cDNA microarray data," *IEEE Trans Inf Technol Biomed*, vol. 6, pp. 29-37, 2002.

[72] Y. Wang, T. Gong, J. Xuan, E. P. Hoffman, and R. Clarke, "Unsupervised In Silico Dissection of Tissue Heterogeneity for Molecular Characterization of Cell Mixtures, Technical Report, CAM-nPICA TP09," Virginia Tech Oct. 18 2008.

[73] 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, 2008.

[74] E. Oja and M. Plumbley, "Blind separation of positive sources by globally convergent gradient search," *Neural Comput.*, vol. 16, pp. 1811-1825, 2004.

[75] M. Plumbley, "Conditions for nonnegative independent component analysis," *Signal Processing Letters, IEEE*, vol. 9, pp. 177-180, 2002.

[76] L. D. Berkovitz, *Convexity and Optimization in R[superscript n]*: Wiley-Interscience, 2001.

[77] L. Cinotti, J. P. Bazin, R. DiPaola, H. Susskind, and A. B. Brill, "Processing of Xe-127 regional pulmonary ventilation by factor analysis and compartmental modeling," *Medical Imaging, IEEE Transactions on*, vol. 10, pp. 437-444, 1991.

[78] A. Baker, *Matrix Groups: An Introduction to Lie Group Theory*: Springer, 2003.

[79] C. Jutten and A. Taleb, "Source Separation: From Dusk Till Dawn," presented at Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation, 2000.

[80] J. Xuan, R. Lee, Y. Zhu, A. L. Zwart, R. Clarke, and Y. Wang, "Tissue heterogeneity correction in gene expression mapping," *AACR Meeting Abstracts*, vol. 2004, pp. 378-c-379, 2004.

[81] F. Abrard and Y. Deville, "Blind separation of dependent sources using the "time-frequency ratio of mixtures" approach," presented at Signal Processing and Its Applications, 2003. Proceedings. Seventh International Symposium on, 2003.

[82] J. P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proc Natl Acad Sci U S A*, vol. 101, pp. 4164-9, 2004.

[83] C.-J. Lin, "On the Convergence of Multiplicative Update Algorithms for Nonnegative Matrix Factorization," *Neural Networks, IEEE Transactions on*, vol. 18, pp. 1589-1596, 2007.

[84] K. Stadlthanner, F. J. Theis, E. W. Lang, A. M. Tom, C. G. Puntonet, G. J. M, and rriz, "Hybridizing sparse component analysis with genetic algorithms for microarray analysis," *Neurocomput.*, vol. 71, pp. 2356-2376, 2008.

[85] S. A. Astakhov, H. Stogbauer, A. Kraskov, and P. Grassberger, "Monte Carlo algorithm for least dependent non-negative mixture decomposition," *Anal Chem*, vol. 78, pp. 1620-7, 2006.

[86] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M. Loda, E. S. Lander, and T. R. Golub, "Multiclass cancer diagnosis using tumor gene expression signatures," *Proc Natl Acad Sci U S A*, vol. 98, pp. 15149-54, 2001.

[87] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker, and M. Meyerson, "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proc Natl Acad Sci U S A*, vol. 98, pp. 13790-5, 2001.

[88] Y. P. Yu, D. Landsittel, L. Jing, J. Nelson, B. Ren, L. Liu, C. McDonald, R. Thomas, R. Dhir, S. Finkelstein, G. Michalopoulos, M. Becich, and J. H. Luo, "Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy," *J Clin Oncol*, vol. 22, pp. 2790-9, 2004.

[89] X. J. Ma, Z. Wang, P. D. Ryan, S. J. Isakoff, A. Barmettler, A. Fuller, B. Muir, G. Mohapatra, R. Salunga, J. T. Tuggle, Y. Tran, D. Tran, A. Tassin, P. Amon, W. Wang, W. Wang, E. Enright, K. Stecker, E. Estepa-Sabal, B. Smith, J. Younger, U. Balis, J. Michaelson, A. Bhan, K. Habin, T. M. Baer, J. Brugge, D. A. Haber, M. G. Erlander, and D. C. Sgroi, "A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen," *Cancer Cell*, vol. 5, pp. 607-16, 2004.

[90] S. Ramaswamy, K. N. Ross, E. S. Lander, and T. R. Golub, "A molecular signature of metastasis in primary solid tumors," *Nat Genet*, vol. 33, pp. 49-54, 2003.

[91] S. Paik, S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F. L. Baehner, M. G. Walker, D. Watson, T. Park, W. Hiller, E. R. Fisher, D. L. Wickerham, J. Bryant,

and N. Wolmark, "A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer," *N Engl J Med*, vol. 351, pp. 2817-26, 2004.

[92] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, pp. 530-6, 2002.

[93] P. Roepman, E. de Koning, D. van Leenen, R. A. de Weger, J. A. Kummer, P. J. Slootweg, and F. C. Holstege, "Dissection of a metastatic gene expression signature into distinct components," *Genome Biol*, vol. 7, pp. R117, 2006.

[94] J. Sabates-Bellver, L. G. Van der Flier, M. de Palo, E. Cattaneo, C. Maake, H. Rehrauer, E. Laczko, M. A. Kurowski, J. M. Bujnicki, M. Menigatti, J. Luz, T. V. Ranalli, V. Gomes, A. Pastorelli, R. Faggiani, M. Anti, J. Jiricny, H. Clevers, and G. Marra, "Transcriptome profile of human colorectal adenomas," *Mol Cancer Res*, vol. 5, pp. 1263-75, 2007.

[95] A. Martinez and A. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, pp. 228-233, 2001.

[96] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proc Natl Acad Sci U S A*, vol. 99, pp. 6567-72, 2002.

[97] L. Yijuan, T. Qi, N. Jennifer, L. Feng, and W. Yufeng, "Adaptive discriminant analysis for microarray-based classification," *ACM Trans. Knowl. Discov. Data*, vol. 2, pp. 1-20, 2008.

[98] C. Cortes and V. Vapnik, "Support Vector Networks," presented at Machine Learning, 1995.

[99] E. Byvatov and G. Schneider, "Support vector machine applications in bioinformatics," *Appl Bioinformatics*, vol. 2, pp. 67-77, 2003.

[100] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, pp. 906-14, 2000.

[101] E. Segal, N. Friedman, D. Koller, and A. Regev, "A module map showing conditional activity of expression modules in cancer," *Nat Genet*, vol. 36, pp. 1090-8, 2004.

[102] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. F. Armour, H. A. Bennett, E. Coffey, H. Dai, and Y. D. He, "Functional discovery via a compendium of expression profiles," *Cell*, vol. 102, pp. 109 - 126, 2000.

[103] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. C. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, and P. O. Brown, "Systematic variation in gene expression patterns in human cancer cell lines," *Nat Genet*, vol. 24, pp. 227-35, 2000.

[104] M. A. Beer and S. Tavazoie, "Predicting gene expression from sequence," *Cell*, vol. 117, pp. 185 - 198, 2004.

[105] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, and E. Dmitrovsky, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation," *Proc. Natl. Acad. Sci. USA*, vol. 96, pp. 2907-2912, 1999.

[106] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai, "Revealing modular organization in the yeast transcriptional network," *Nat Genet*, vol. 31, pp. 370-377, 2002.

[107] M. Kloster, C. Tang, and N. S. Wingreen, "Finding regulatory modules through large-scale gene-expression data analysis," *Bioinformatics*, vol. 21, pp. 1172-9, 2005.

[108] W. Pan, J. Lin, and C. Le, "Model-based cluster analysis of microarray gene-expression data," *Genome Biology*, vol. 3, pp. research0009.1 - research0009.8, 2002.

[109] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian Networks to Analyze Expression Data," *Journal of Computational Biology*, vol. 7, pp. 601-620, 2000.

[110] P. Carmona-Saez, R. D. Pascual-Marqui, F. Tirado, J. M. Carazo, and A. Pascual-Montano, "Biclustering of gene expression data by non-smooth non-negative matrix factorization," *BMC Bioinformatics*, vol. 7, 2006.

[111] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788-91, 1999.

[112] E. R. Dougherty, J. Barrera, M. Brun, S. Kim, R. M. Cesar, Y. Chen, M. Bittner, and J. M. Trent, "Inference from clustering with application to gene-expression microarrays," *J Comput Biol*, vol. 9, pp. 105-26, 2002.

[113] G. C. Tseng, M. K. Oh, L. Rohlin, J. C. Liao, and W. H. Wong, "Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects," *Nucleic Acids Res*, vol. 29, pp. 2549-57, 2001.

[114] G. Hori, M. Inoue, S. Nishimura, and H. Nakahara, "Blind gene classification: An ICA-based gene classification/clustering method," *RIKEN BSI BSIS Technical Report*, vol. No.02-5, 2002.

[115] T. Gong, Y. Zhu, J. Xuan, H. Li, R. Clarke, E. P. Hoffman, and Y. Wang, "Latent Variable and nICA Modeling of Pathway Gene Module Composite," presented at Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE, 2006.

[116] F. Vrins, J. A. Lee, M. Verleysen, V. Vigneron, and C. Jutten, "Improving independent component analysis performances by variable selection," presented at Neural Networks for Signal Processing, 2003. NNSP'03. 2003 IEEE 13th Workshop on, 2003.

[117] T. M. Cover and J. A. Thomas, *Elements of information theory*: John Wilsy and sons, 1991.

[118] F. R. Bach and M. I. Jordan, "Beyond independent components: trees and clusters," *The Journal of Machine Learning Research*, vol. 4, pp. 1205-1233, 2003.

[119] 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," presented at The International Conference on Bioinformatics & Computational Biology: 2007, Las Vegas, 2007.

[120] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown, "Genomic expression programs in the response of yeast cells to environmental changes," *Molecular Biology of the Cell*, vol. 11, pp. 4241-4257, 2000.

[121] 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, 2003.

[122] 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. 3448-3449, 2005.

[123] Y. Ge, S. Dudoit, and T. Speed, "Resampling-based multiple testing for microarray data analysis," *TEST*, vol. 12, pp. 1-77, 2003.

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

[125] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, pp. 520-525, 2001.

[126] F. D. Gibbons and F. P. Roth, "Judging the quality of gene expression-based clustering methods using gene annotation," *Genome Research*, vol. 12, pp. 1574-1581, 2002.

[127] D. Dembele and P. Kastner, "Fuzzy C-means method for clustering microarray data," *Bioinformatics*, vol. 19, pp. 973-80, 2003.

[128] J. F. Kolen and T. Hutcheson, "Reducing the time complexity of the fuzzy c-means algorithm," *Fuzzy Systems, IEEE Transactions on*, vol. 10, pp. 263-267, 2002.

[129] "Ingenuity® Systems."

[130] M. Bakay, Z. Wang, G. Melcon, L. Schiltz, J. Xuan, P. Zhao, V. Sartorelli, J. Seo, E. Pegoraro, C. Angelini, B. Shneiderman, D. Escolar, Y. W. Chen, S. T. Winokur, L. M. Pachman, C. Fan, R. Mandler, Y. Nevo, E. Gordon, Y. Zhu, Y. Dong, Y. Wang, and E. P. Hoffman, "Nuclear envelope dystrophies show a transcriptional fingerprint suggesting disruption of Rb-MyoD pathways in muscle regeneration," *Brain*, vol. 129, pp. 996-1013, 2006.

[131] L. Hogben, A. G. (Editor), R. B. (Editor), and R. M. (Editor), *Handbook of Linear Algebra*: Chapman & Hall, 2007.

[132] E. Cox, *Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration*: Elsevier Science, 2005.

[133] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe, "A census of human transcription factors: function, expression and evolution," *Nat Rev Genet*, vol. 10, pp. 252-63, 2009.

[134] E. Neame, "Gene networks: Network analysis gets dynamic," *Nat Rev Genet*, vol. 9, pp. 897-897, 2008.

[135] M. Clements, E. P. v. Someren, T. A. Knijnenburg, and M. J. T. Reinders, "Integration of Known Transcription Factor Binding Site Information and Gene Expression Data to Advance from Co-Expression to Co-Regulation," *Genomics, Proteomics & Bioinformatics* vol. 5, pp. 86-101, 2007.

[136] 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. 2005-2011, 2006.

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

[138] M. K. S. Yeung, J. Tegnér, and J. J. Collins, "Reverse engineering gene networks using singular value decomposition and robust regression," *Proceedings of the*

National Academy of Sciences of the United States of America, vol. 99, pp. 6163-6168, 2002.

[139] D. S. Latchman, "Transcription Factors as Potential Targets for Therapeutic Drugs," *Current Pharmaceutical Biotechnology*, vol. 1, pp. 57-61, 2000.

[140] K. Y. Yeung and W. L. Ruzzo, "Principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, pp. 763-774, 2001.

[141] C. Sabatti and G. M. James, "Bayesian sparse hidden components analysis for transcription regulation networks," *Bioinformatics*, vol. 22, pp. 739-46, 2006.

[142] X. J. Zhou, M.-C. J. Kao, H. Huang, A. Wong, J. Nunez-Iglesias, M. Primig, O. M. Aparicio, C. E. Finch, T. E. Morgan, and W. H. Wong, "Functional annotation and network reconstruction through cross-platform integration of microarray data," *Nat Biotech*, vol. 23, pp. 238-243, 2005.

[143] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," *Nat Genet*, vol. 22, pp. 281-285, 1999.

[144] C. M. Perou, T. Sorlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lonning, A.-L. Borresen-Dale, P. O. Brown, and D. Botstein, "Molecular portraits of human breast tumours," *Nature*, vol. 406, pp. 747-752, 2000.

[145] I. Ben-Gal, A. Shani, A. Gohr, J. Grau, S. Arviv, A. Shmilovici, S. Posch, and I. Grosse, "Identification of transcription factor binding sites with variable-order Bayesian networks," *Bioinformatics*, vol. 21, pp. 2657-2666, 2005.

[146] 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(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes," *Nucl. Acids Res.*, vol. 34, pp. D108-110, 2006.

[147] 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, 2003.

[148] Y. Tu, G. Stolovitzky, and U. Klein, "Quantitative noise analysis for gene expression microarray experiments," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, pp. 14031-14036, 2002.

[149] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, "Microarray data analysis: from disarray to consolidation and consensus," *Nat Rev Genet*, vol. 7, pp. 55-65, 2006.

[150] A. Kundaje, A. Kundaje, M. Middendorf, G. Feng, C. A. W. C. Wiggins, and C. A. L. C. Leslie, "Combining sequence and time series expression data to learn transcriptional modules," *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 2, pp. 194-202, 2005.

[151] T. M. Cover and J. A. Thomas, *Elements of Information Theory 2nd Edition*: Wiley-Interscience, 2006.

[152] M. D. Levine and A. M. Nazif, "Dynamic Measurement of Computer Generated Image Segmentations," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-7, pp. 155-164, 1985.

[153] M. K. Broadhead and L. A. Pflug, "Performance of some sparseness criterion blind deconvolution methods in the presence of noise," *J Acoust Soc Am*, vol. 107, pp. 885-93, 2000.

[154] P. O. Hoyer, "Non-negative Matrix Factorization with Sparseness Constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457-1469, 2004.

[155] A. Arash Ali, B.-Z. Massoud, and J. Christian, "A Fast Method for Sparse Component Analysis Based on Iterative Detection-Estimation," *AIP Conference Proceedings*, vol. 872, pp. 123-130, 2006.

[156] J. Nocedal and S. Wright, *Numerical Optimization*, 2nd ed. ed, 2006.

[157] C. Chang, Z. Ding, Y. S. Hung, and P. C. W. Fung, "Fast network component analysis (FastNCA) for gene regulatory network reconstruction from microarray data," *Bioinformatics*, vol. 24, pp. 1349-1358, 2008.

[158] 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, pp. 43, 2006.

[159] R. Garcia, C. Bermejo, C. Grau, R. Perez, J. M. Rodriguez-Pena, J. Francois, C. Nombela, and J. Arroyo, "The Global Transcriptional Response to Transient Cell Wall Damage in Saccharomyces cerevisiae and Its Regulation by the Cell Integrity Signaling Pathway," *J. Biol. Chem.*, vol. 279, pp. 15183-15195, 2004.

[160] H. G. Lee, H. S. Lee, S. H. Jeon, T. H. Chung, Y. S. Lim, and W. K. Huh, "High-resolution analysis of condition-specific regulatory modules in Saccharomyces cerevisiae," *Genome Biol*, vol. 9, pp. R2, 2008.

[161] G. Chen, S. Jensen, and C. Stoeckert, "Clustering of genes into regulons using integrated modeling-COGRIM," *Genome Biology*, vol. 8, pp. R4, 2007.

[162] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, and M. J. Thun, "Cancer statistics, 2009," *CA Cancer J Clin*, vol. 59, pp. 225-49, 2009.

[163] E. A. Musgrove and R. L. Sutherland, "Biological determinants of endocrine resistance in breast cancer," *Nat Rev Cancer*, vol. 9, pp. 631-43, 2009.

[164] R. Clarke, M. C. Liu, K. B. Bouker, Z. Gu, R. Y. Lee, Y. Zhu, T. C. Skaar, B. Gomez, K. O'Brien, Y. Wang, and L. A. Hilakivi-Clarke, "Antiestrogen resistance in breast cancer and the role of estrogen receptor signaling," *Oncogene*, vol. 22, pp. 7316-39, 2003.

[165] T. Gong, J. Xuan, R. B. Reggins, and R. Clarke, "A Systems Biology Approach to Identify Affected Regulatory and Signaling Circuits in Protein Interaction Networks," presented at 2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, Shanghai, China 2009.

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

[167] Y. Wang, J. Zhang, J. Khan, R. Clarke, and Z. Gu, "Partially-independent component analysis for tissue heterogeneity correction in microarray gene expression analysis," presented at Neural Networks for Signal Processing, 2003. NNSP'03. 2003 IEEE 13th Workshop 2003.

[168] 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. 1289-1297, 2006.

[169] K. Abell and C. J. Watson, "The Jak/Stat Pathway: A Novel Way to Regulate PI3K Activity," *Cell cycle*, vol. 4, pp. 4, 2005.

[170] J. G. Moggs and G. Orphanides, "Estrogen receptors: orchestrators of pleiotropic cellular responses " *EMBO reports*, vol. 2, pp. 7, 2001.

[171] R. Kageyama, G. T. Merlino, and I. Pastan, "A transcription factor active on the epidermal growth factor receptor gene," *Proc Natl Acad Sci U S A*, vol. 85, pp. 5016-20, 1988.

[172] L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. von Mering, "STRING 8--a

global view on proteins and their functional interactions in 630 organisms," *Nucleic Acids Res*, vol. 37, pp. D412-6, 2009.

[173] S. K. Muthuswamy, M. Gilman, and J. S. Brugge, "Controlled dimerization of ErbB receptors provides evidence for differential signaling by homo- and heterodimers," *Mol Cell Biol*, vol. 19, pp. 6845-57, 1999.

[174] Y. Yarden and M. X. Sliwkowski, "Untangling the ErbB signalling network," *Nat Rev Mol Cell Biol*, vol. 2, pp. 127-37, 2001.

[175] 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, 2004.

[176] P. Alvarez, P. Saenz, D. Arteta, A. Martinez, M. Pocovi, L. Simon, and P. Giraldo, "Transcriptional profiling of hematologic malignancies with a low-density DNA microarray," *Clin Chem*, vol. 53, pp. 259-67, 2007.

[177] R. B. Riggins, K. S. Thomas, H. Q. Ta, J. Wen, R. J. Davis, N. R. Schuh, S. S. Donelan, K. A. Owen, M. A. Gibson, M. A. Shupnik, C. M. Silva, S. J. Parsons, R. Clarke, and A. H. Bouton, "Physical and functional interactions between Cas and c-Src induce tamoxifen resistance of breast cancer cells through pathways involving epidermal growth factor receptor and signal transducer and activator of transcription 5b," *Cancer Res*, vol. 66, pp. 7007-15, 2006.

[178] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, pp. 57-70, 2000.

[179] Y. Pilpel, P. Sudarsanam, and G. M. Church, "Identifying regulatory networks by combinatorial analysis of promoter elements," *Nat Genet*, vol. 29, pp. 153-159, 2001.

[180] S. Aburatani, F. Sun, S. Saito, M. Honda, S. Kaneko, and K. Horimoto, "Gene systems network inferred from expression profiles in hepatocellular carcinogenesis by graphical gaussian model," *EURASIP J Bioinform Syst Biol*, pp. 47214, 2007.

[181] K. Horimoto and H. Toh, "Statistical estimation of cluster boundaries in gene expression profile data," *Bioinformatics*, vol. 17, pp. 1143-51, 2001.

[182] A. Remenyi, H. R. Scholer, and M. Wilmanns, "Combinatorial control of gene expression," *Nat Struct Mol Biol*, vol. 11, pp. 812-5, 2004.

[183] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander, and T. R. Golub, "The Connectivity Map: Using Gene-Expression

Signatures to Connect Small Molecules, Genes, and Disease," *Science*, vol. 313, pp. 1929-1935, 2006.

[184] G. J. Clark and C. J. Der, "Aberrant function of the Ras signal transduction pathway in human breast cancer," *Breast Cancer Research and Treatment*, vol. 35, 1995.

[185] L. Bjornstrom and M. Sjoberg, "Mechanisms of Estrogen Receptor Signaling: Convergence of Genomic and Nongenomic Actions on Target Genes," *Mol Endocrinol*, vol. 19, pp. 833-842, 2005.

[186] 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. 2007, pp. 349, 2008.

[187] A. Niida, A. Smith, S. Imoto, S. Tsutsumi, H. Aburatani, M. Zhang, and T. Akiyama, "Integrative bioinformatics analysis of transcriptional regulatory programs in breast cancer cells," *BMC Bioinformatics*, vol. 9, pp. 404, 2008.

[188] M. Gasco, S. Shami, and T. Crook, "The p53 pathway in breast cancer," *Breast Cancer Res*, vol. 4, pp. 70 - 76, 2002.

[189] P. Jansen-Durr, A. Meichle, P. Steiner, M. Pagano, K. Finke, J. Botz, J. Wessbecher, G. Draetta, and M. Eilers, "Differential modulation of cyclin gene expression by MYC," *Proc Natl Acad Sci U S A*, vol. 90, pp. 3685-9, 1993.

[190] J. Zwicker, F. C. Lucibello, L. A. Wolfraim, C. Gross, M. Truss, K. Engeland, and R. Muller, "Cell cycle regulation of the cyclin A, cdc25C and cdc2 genes is based on a common mechanism of transcriptional repression," *Embo J*, vol. 14, pp. 4514-22, 1995.

[191] D. A. Dedera, E. K. Waller, D. P. LeBrun, A. Sen-Majumdar, M. E. Stevens, G. S. Barsh, and M. L. Cleary, "Chimeric homeobox gene E2A-PBX1 induces proliferation, apoptosis, and malignant lymphomas in transgenic mice," *Cell*, vol. 74, pp. 833-43, 1993.