From network to pathway:

integrative network analysis of genomic data

Chen Wang

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
Electrical Engineering

Jason J. Xuan, Chair
Chang-Tien Lu
Christopher L. Wyatt
William T. Baumann
Yue J. Wang

June 14, 2011
Arlington, Virginia

From network to pathway: integrative network analysis of genomic data

Chen Wang

Abstract

The advent of various types of high-throughput genomic data has enabled researchers to investigate complex biological systems in a systemic way and started to shed light on the underlying molecular mechanisms in cancers. To analyze huge amounts of genomic data, effective statistical and machine learning tools are clearly needed; more importantly, integrative approaches are especially needed to combine different types of genomic data for a network or pathway view of biological systems. Motivated by such needs, we make efforts in this dissertation to develop integrative approaches for gene network and pathway inference. Specifically, we dissect the molecular pathway into two parts: protein-DNA interaction network and protein-protein interaction network. Several novel approaches are proposed to integrate gene expression data with various forms of biological knowledge, such as protein-DNA interaction and protein-protein interaction for reliable molecular network identification.

The first part of this dissertation seeks to infer condition-specific transcriptional regulatory network by integrating gene expression data and protein-DNA binding information. Protein-DNA binding information provides initial relationships between transcription factors (TFs) and their target genes, and this information is essential to derive biologically meaningful integrative algorithms. Based on the availability of this information, we discuss the inference task based on two different situations: (a) if protein-DNA binding information of multiple TF is available: based on the protein-DNA data of multiple TFs, which are derived from sequence analysis between DNA motifs and gene promoter regions, we can construct initial connection matrix and solve the network inference using a constraint least-squares approach named motif-guided network component analysis (mNCA). However, connection

matrix usually contains a considerable amount of false positives and false negatives that make inference results questionable. To circumvent this problem, we propose a knowledge based stability analysis (kSA) approach to test the conditional relevance of individual TFs, by checking the discrepancy of multiple estimations of transcription factor activity with respect to different perturbations on the connections. The rationale behind stability analysis is that the consistency of observed gene expression and true network connection shall remain stable after small perturbations are applied to initial connection matrix. With condition-specific TFs prioritized by kSA, we further propose to use multivariate regression to highlight condition-specific target genes. Through simulation studies comparing with several competing methods, we show that the proposed scheme are more sensitive to detect relevant TFs and target genes for network inference purpose. Experimentally, We have applied stability analysis to yeast cell cycle experiment and further to a series of anti-estrogen breast cancer studies. In both experiments not only biologically relevant regulators are highlighted, the condition-specific transcriptional regulatory networks are also constructed, which could provide further insights into the corresponding cellular mechanisms. (b) if only single TF's protein-DNA information is available: this happens when protein-DNA binding relationship of individual TF is measured through experiments. Since original mNCA requires a complete connection matrix to perform estimation, an incomplete knowledge of single TF is not applicable for such approach. Moreover, binding information derived from experiments could still be inconsistent with gene expression levels. To overcome these limitations, we propose a linear extraction scheme called regulatory component analysis (RCA), which can infer underlying regulation relationships, even with partial biological knowledge. Numerical simulations show significant improvement of RCA over other traditional methods to identify target genes, not only in low signal-to-noise-ratio situations and but also when the given biological knowledge is incomplete and inconsistent to data. Furthermore, biological experiments on Escherichia coli regulatory network inferences are performed to fairly compare traditional methods, where the effectiveness and superior performance of RCA are confirmed.

The second part of the dissertation moves from protein-DNA interaction network up to

protein-protein interaction network, to identify dys-regulated protein sub-networks by integrating gene expression data and protein-protein interaction information. Specifically, we propose a statistically principled method, namely Metropolis random walk on graph (MRWOG), to highlight condition-specific PPI sub-networks in a probabilistic way. The method is based on the Markov chain Monte Carlo (MCMC) theory to generate a series of samples that will eventually converge to some desired equilibrium distribution, and each sample indicates the selection of one particular sub-network during the process of Metropolis random walk. The central idea of MRWOG is built upon that the essentiality of one gene to be included in a sub-network depends on not only its expression but also its topological importance. Contrasted to most existing methods constructing sub-networks in a deterministic way and therefore lacking relevance score for gene node, MRWOG is capable of assessing the importance of each individual protein node in a global way, not only reflecting its individual association with clinical outcome but also indicating its topological role (hub, bridge) to connect other important proteins. Moreover, each protein node is associated with a sampling frequency score, which enables the statistical justification of each individual node and flexible scaling of sub-network results. Based on MRWOG approach, we further propose two strategies : one is bootstrapping used for assessing statistical confidence of detected sub-networks; the other is graphic division to separate a large sub-network to several smaller sub-networks for facilitating interpretations. MRWOG is easy to use with only two parameters need to be adjusted, one is beta value for performing random walk and another is Quantile level for calculating truncated posteriori mean. Through extensive simulations, we show that the proposed scheme is not sensitive to these two parameters in a relatively wide range. We also compare MRWOG with deterministic approaches for identifying sub-network and prioritizing topologically important proteins, in both cases MRWG outperforms existing methods in terms of both precision and recall. By utilizing MRWOG generated node/edge sampling frequency, which is actually posteriori mean of corresponding protein node/interaction edge, we illustrate that condition-specific nodes/interactions can be better prioritized than the schemes based on scores of individual node/interaction. Experimentally, we have applied

MRWOG to study yeast-stress condition first to and then breast cancer patient prognostics, where the sub-network analysis could lead to an understanding of the molecular mechanisms of antiestrogen resistance in breast cancer.

Finally, we conclude this dissertation with a summary of the original contributions, and the future work for deepening the theoretical justification of the proposed methods and broadening their potential biological applications such as cancer studies.

Acknowledgments

I would like to thank all the people who help and support me through Ph.D. study period.

First I want to thank my advisor, Dr. Jianhua J. Xuan, for his guidance, support and great help throughout the past five years. He always provided insightful guidance, accurately pinpointed my weakness and drawbacks, shared his experience and positive attitudes with me, and revised my papers with great patience. It is my big fortune to have his advices not only for this Ph.D. study, but also for my future career.

I would like to thank Dr. Yue J. Wang as my steering committee member, who inspired and guided me a lot from machine learning and mathematics during my dissertation preparation. I am also grateful to him as the director of CBIL for providing me with stimulating environment to conduct exciting research in the past five years.

I would like to express my sincere gratitude to the other committee members, Dr. Chang-Tien Lu, Dr. Christopher L. Wyatt and Dr. William T. Baumann for their insightful comments and great help. They provided me with valuable inputs in my preliminary examination, and all these suggestions appear to be very important for the formation of my final dissertation.

I appreciate our collaborators in Georgetown University, Children's National Medical Center and Johns Hopkins University. I would like to express my thanks to Dr. Huai Li, Dr. Robert Clark, Dr. Eric P. Hoffman, Dr. Ie-Ming Shih and Dr. Tian-Li Wang to facilitate various inter-disciplinary research topics. They provided me unique biological problems,

insightful interpretations and allowed the use of various pioneering high-throughput data generated from their laboratories.

I thank all of my colleagues and lab mates in CBIL. I enjoyed a lot for so many wonderful moments together. I have benefited greatly through the daily discussions with them. We discussed many inspiring ideas from machine learning to molecular biology. They also provided useful suggestions to improve my work. They made this period of time a unique experience in my life.

According to tradition, my dissertation ought to be dedicated to several special persons, who do not have appropriate background to read it, not mentioning to appreciate it. Instead of doing so, I just want to end this journey soon, and then come back to you, my beloved ones.

Contents

List of Figures

List of Tables

List of Abbreviations

AUC	Area under the ROC Curve
CDF	Cumulative Density Function
cDNA	complementary DeoxyriboNucleic Acid
ChIP	Chromatin ImmunoPrecipitation
DE	Differentially Expressed
DNA	DeoxyriboNucleic Acid
EE	Equally Expressed
ER+	Estrogen Receptor positive
FDR	False Discovery Rate
GG	Gamma-Gamma
GO	Gene Ontology
HPRD	Human Protein Reference Database
ICA	Independent Component Analysis
JADE	Joint Approximate Diagonalization of Eigenmatrices
MAP	Maximum A Posterior
mNCA	motif-directed Network Component Analysis
MRWOG	Metropolis Random Walk On Graph
mRNA	messenger RiboNucleic Acid
NCA	Network Component Analysis
PCA	Principal Component Analysis
PDF	Probability Density Function

PDI	Protein-DNA Interaction
Plier	Probe Logarithmic Intensity ERror
PPI	Protein-Protein Interaction
PWM	Position Weight Matrix
RC	Regulatory Component
RCA	Regulatory Component Analysis
RMA	Robust Multichip Average
ROC	Receiver Operating Characteristics
SNR	Signal to Noise Ratio
TF	Transcription Factor
TFA	Transcription Factor Activity

Chapter 1

Introduction

1.1 Research Motivation

From human genome sequencing (Lander, Linton et al. 2001) to microarray gene expression profiling (Hoheisel 2006), recent biotechnologies have revolutionized the traditional biology research and discovery. In the past, biologists need to narrow down to specific genes or proteins first and measure their activities individually. According to measured results, they have to generate and test biological hypothesis one by one, as the old-fashioned technique cannot profile the activities of multiple molecular targets simultaneously. Nowadays, the advanced genome technology is rapidly changing the situation - with a chip as small as a fingernail one can obtain the dynamic mRNA expression profiling of entire genome.

The wide adoption of high-throughput technology has accelerated the hypothesis generation process for biological research dramatically. At the same time, thousands of new data have been generated and made publicly available to all researchers. For example, The Cancer Genome Atlas (TCGA) project (TCGA-Research-Network 2008) initiated the effort to acquire the genomic characteristics of human tumors using multiple platforms, measuring genetic signals of copy number alternation/variation, Single-nucleotide polymorphism (SNP),

methylation, microRNA, mRNA, etc. The initial TCGA pilot research has been expanded to a large-scale project aiming to provide genomic measurements of 20 25 different tumor types. Having huge amount of genomic data, traditional statistical analysis tools and machine learning approaches, such as differential analysis, classification, clustering and feature selection, are adopted to generate individual gene based hypotheses. Based on each chip platform, these algorithms can potentially provide hundreds of genes or thousands of loci having statistically significant association with biological phenotype or clinical outcome. Hence, biological researchers are overwhelmed by numerous hypotheses (that could lead to potential discoveries) as flooded from high-throughput data analysis.

However, many of the computationally generated hypotheses cannot be verified with further biological validation experiments, and even some gene markers discovered in one cohort cannot be validated in another cohort (Ein-Dor, Kela et al. 2005; Allison, Cui et al. 2006). Why does this happen? First of all, biological data are very noisy; they are heavily affected by a series of experimental procedures such as tissue preparation, biological experiment protocols and hybridization. Second, biological systems by nature are heterogeneous and dynamic. The profiling measurement that we acquired is mixture of multiple cells at different stages. Although some external controls (e.g., synchronization of the cell cycle by applying certain chemical compound to arrest all the cells in G1 stage (Spellman, Sherlock et al. 1998) could be used to synchronize/initialize the biological process of interest, these controls may also introduce other types of artifacts, which is non-specific to biological conditions of interest. To limit the number of false discoveries caused by all these factors, continuous efforts have been made in the field statistical analysis by estimating and consequently controlling the false discovery rate (Storey and Tibshirani 2003), the probability that an algorithm claimed discovery is a false one. However, pure data driven statistical approaches, even with a stringent false discovery rate control, cannot provide immediate biologically interpretable results. Moreover, data driven statistical data analysis or machine learning approaches also suffer from the risk of over-fitting due to the curse-of dimensionality problem (Clarke, Ressom et al. 2008), i.e., the very high dimensionality of genes as compared to the small number of

available biological samples. As a result, the genes selected based on data-driven approaches using one data set are less likely to be reproduced in another cohort study.

The abovementioned difficulties have motivated researchers to integrate gene expression data with existing biological knowledge sources for modeling and analyzing genomic data. In the field of bioinformatics and systems biology, it is becoming a common practice to incorporate biological knowledge into the mathematical or statistical modeling process. The advantage of knowledge integration is that one can interpret the genes selected by computational approaches within biological context, like their potential interacting neighbors and biological functionality groups. It has been shown that with the gene set analysis or gene network analysis one can enlarge the overlap of the results from the analysis of different platforms for a same study (Manoli, Gretz et al. 2006; Chuang, Lee et al. 2007; Carro, Lim et al. 2010). Additionally, integration of knowledge can further decrease the model complexity and reduce the over-fitting risk, as the feasible parameter space is largely reduced by the biological constraints (Ideker, Dutkowski et al. 2011).

Having these benefits of biological knowledge integration, in this dissertation we intend to address several specific problems associated with integrative leaning approaches, which lead us to develop novel methods for transcriptional regulatory network inference, target gene identification and sub-network identification. We will describe these problems in detail in the following sections, starting with a brief introduction to biological background.

1.2 Biological Background

1.2.1 Central dogma of molecular biology

The central dogma of molecular biology is about the information transferring among DNA, RNA and protein, which are the key elements in molecular systems. DNA contains all types of genetic information of a cellular system, including coding regions that serve as the

blueprint for producing protein and non-coding regions that are believed to be responsible for controlling mechanisms of the system. RNA is the intermediate messenger to carry the protein coding information from DNA out of nucleic to plasma. According to RNA, different types of proteins are synthesized to perform numerous biological functions. Among them, some proteins called transcription factors (TFs) physically interact with DNA to initialize the transcription of RNA, some proteins synthesize RNA according to DNA coding region, and some proteins translate RNA information into the end product protein. Therefore, the levels of DNA, RNA and protein are also closely coupled. In each level, we can measure different signals to describe the genetic characteristics. For examples, in the DNA level, we can measure sequence mutations, copy number alternations, and epigenetic changes such as methylation status; in the RNA level, we can measure activity, i.e., the concentration of mRNA and microRNA; in the protein level, we can measure the amount, structure, and phosphorylation status of certain protein. Moreover, different physical interactions among them can also be measured, such as protein-DNA binding and protein-protein interaction. Overall, the cellular system has sophisticated hierarchical structure not only about intra-cellular controlling mechanism but also involving cell-cell signaling events, shown in Fig. 1.1. In the studies of this dissertation, we mainly focus on how to model the cellular system using data from mRNA microarray, as the number of available microarray data sets is sufficiently large for data modeling and analysis and its measuring technique is relatively mature.

1.2.2 Microarray technology and gene expression

The basic idea of microarray technique is to measure the concentration of certain molecular by utilizing the hybridization between the signature sequence of this molecular and another complementary DNA sequences (Allison, Cui et al. 2006; Hoheisel 2006). The commonly used microarray chip consists of a series of microscopic spots of DNA oligonucleotides, each of which contains specific DNA sequence known as probe (or reporter) to hybridize a corresponding cDNA or cRNA sample (also known as target) under well controlled experimental

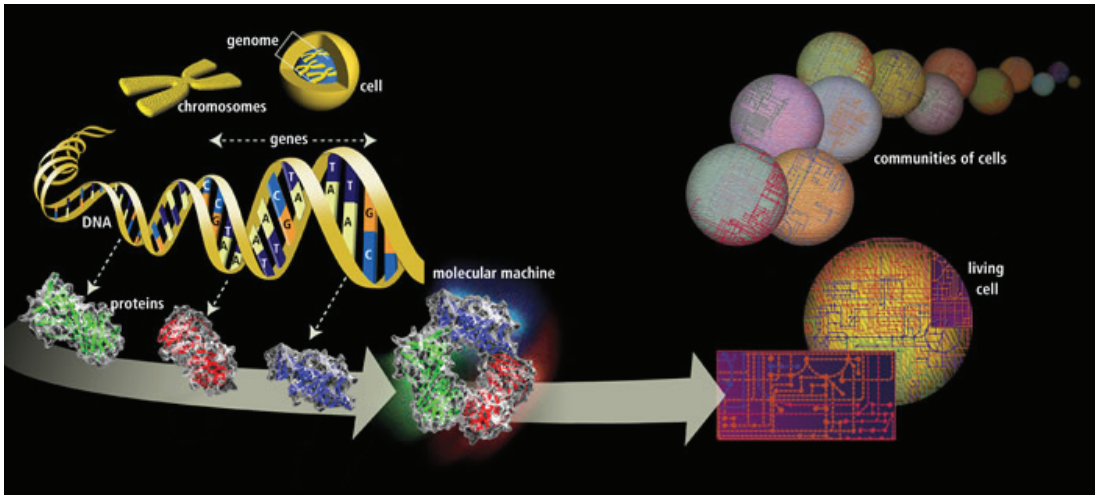


Figure 1.1: The complex genetic system: from the human genome and beyond. (`http://www.ornl.gov/sci/techresources/Human_Genome/publicat/primer2pager.pdf`)

conditions. Tens of thousands of probes are laid out in a hard surface such as glass and silicon using surface engineering. Using these probes, a microarray chip measures the activities of huge amounts of genetic targets in parallel. Besides acquiring mRNA gene expression, microarray technique is also applied to monitor other genetic signals such as single nucleotide polymorphism (SNP), methylation, etc. (Hoheisel 2006).

Taking the measurement of mRNA as an example, the microarray experimental procedures comprise of several steps as shown in Fig. 1.2: biological sample preparation to extract tissue of interest, purification to isolate mRNA, reverse transcriptase mRNA to cDNA, coupling to dye the color on cDNA, hybridizing the cDNA, washing away non-specific binding, and signal detection by a scanning machine. As multiple steps are involved in acquisition of microarray data, the data could be very noisy, shown in Fig. 1.3 as an example, and variation and difference in experimental operation make different microarray samples not comparable. As the results, normalization of microarray signals has become a necessary and essential step for any further data analysis. The purpose of normalization is to correct the experimental bias by adjusting multiple data according to common reference or the same distribution. Several

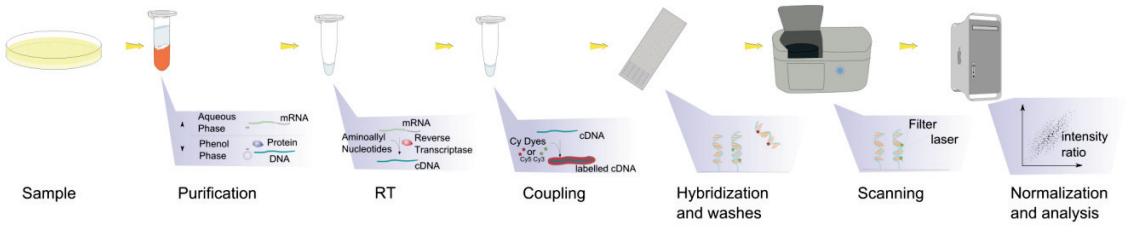


Figure 1.2: Multiple experimental procedures in acquiring microarray data. (`http://upload.wikimedia.org/wikipedia/en/e/e8/Microarray_exp_horizontal.svg`)

typical normalization algorithms have been regarded as standard approaches, such as MAS, RMA and PLIER (Quackenbush 2002).

After the normalization, gene expression data from multiple biological samples are usually organized as a numerical matrix, in which one dimension is corresponding to gene and another dimension to biological sample. Based on this data matrix and explicit biological information like condition or phenotype labels, several types of statistical analysis and machine learning methods such as differentially expressed gene analysis (Tusher, Tibshirani et al. 2001; Storey, Xiao et al. 2005), clustering and classification (Chuang, Lee et al. 2007) were also carried out to explore the data characteristics and pattern structures.

1.2.3 Various biological knowledge sources

With expression measurement of all the genes, it remains unclear how do they interplay with each other and contribute to phenotype difference, since the protein product of a single gene could participate in various cellular processes and collaborate with other molecules to perform different biological functions. Therefore, it is very important to understand relationships of genes in a cellular system. Tremendous efforts have been made both experimentally and computationally in order to complete the understanding of genetic systems from multiple molecular levels. Without naming all of biological knowledge sources, a brief introduction of protein-DNA interactions (PDIs), protein-protein interactions (PPIs) and

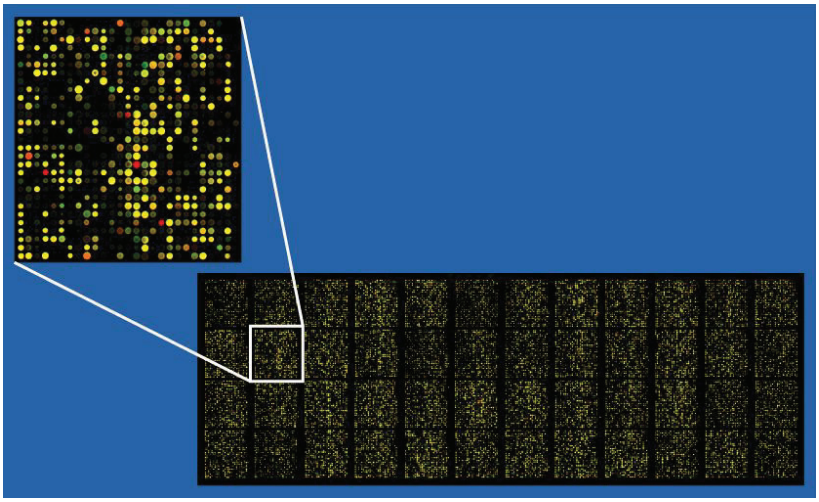


Figure 1.3: One example of a spotted oligo microarray chip with a zoom-in picture. (`http://upload.wikimedia.org/wikipedia/commons/0/0e/Microarray2.gif`)

annotation databases will be given in the following sections, as the necessary background of our proposed knowledge integrated approaches.

Protein-DNA interaction information

Transcription is the biochemistry step that certain DNA regions are transcribed to mRNA. It is initiated and controlled by transcription factors (TFs), a special type of proteins that can bind to DNA. The binding site of a TF is typically close to the gene's promoter region, and has certain short sequence pattern. Such specific sequence pattern is defined as a DNA sequence motif. When direct physical binding measurement is not available, such motif analysis is usually treated as an initial step to explore or discover the TF-DNA binding relationship.

Chromatin Immunoprecipitation (ChIP) is the way to investigate whether one protein is associated with certain DNA segment by immunoprecipitation, a technique precipitating the protein antigen using a specifically binding antibody. When ChIP is incorporated with DNA

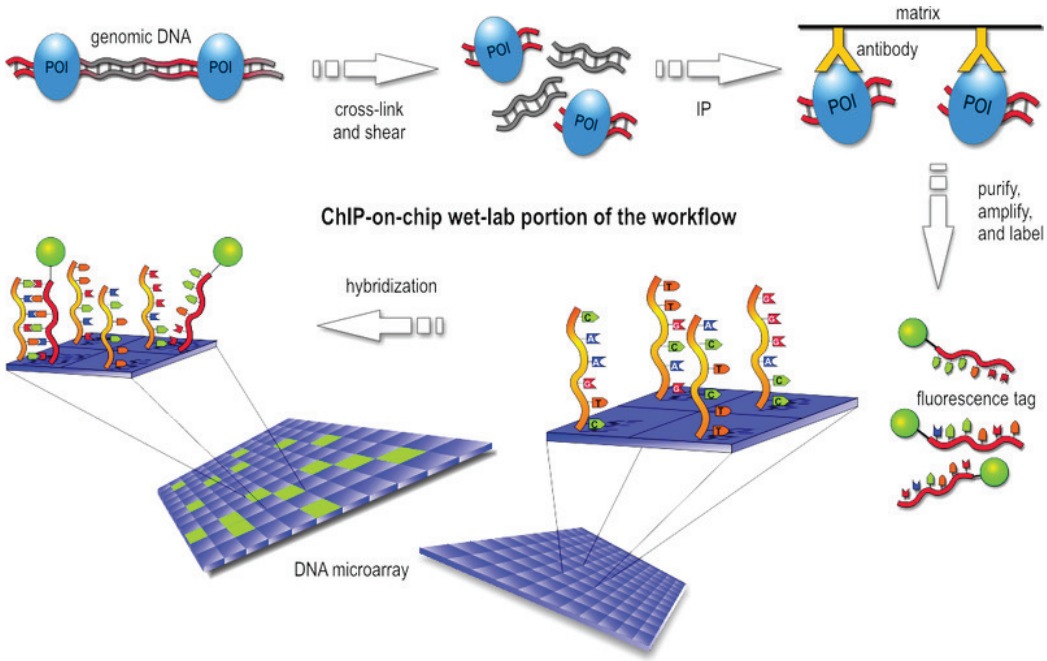


Figure 1.4: Workflow of ChIP-chip experiment.(`http://upload.wikimedia.org/wikipedia/en/8/8d/ChIP-on-chip_wet-lab.png`)

microarray technology (i.e., ChIP on chip or ChIP-chip), the high-throughput measurements of multiple DNA regions' binding events are acquired simultaneously. The workflow of ChIP-chip is shown in Fig. 1.4.

Similar to gene expression data, ChIP-chip data also need to be normalized to eliminate the systematic bias caused by experimental operations and unbalanced hybridizations. After normalization, binding peak detection is performed to locate DNA regions that are bound by the protein of interest. The ChIP-chip data are almost complete and available for the yeast model system (Lee, Rinaldi et al. 2002; Harbison, Gordon et al. 2004), and analysis of these data reveal complex regulation structures, shown in Fig. 1.5 as an example. However, only a few of ChIP-chip experiments are available for mouse and human studies.

ChIP-chip is specifically designed for certain known protein and its corresponding antibody is available for cross-link purpose. In many cases when the proteins actively regulating gene

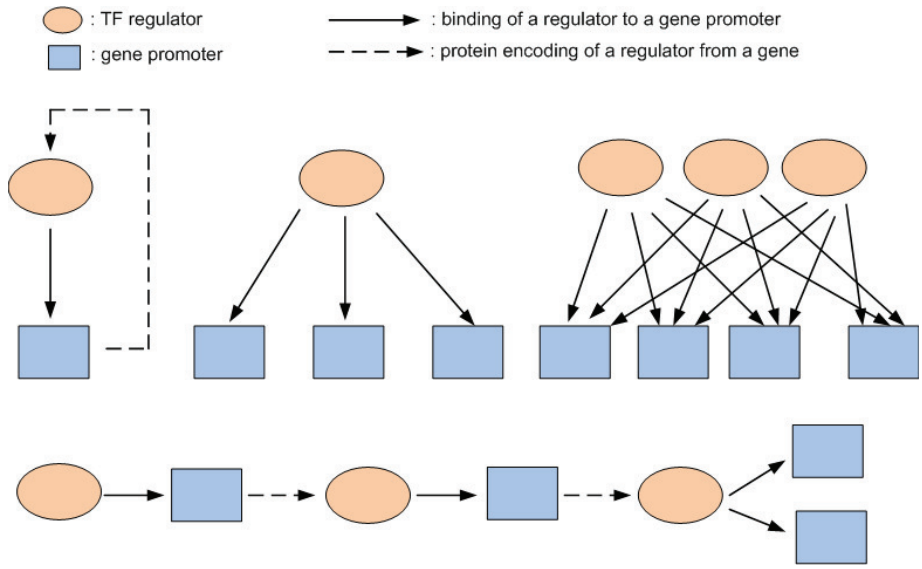


Figure 1.5: Different types of regulation structures revealed from analysis of yeast ChIP-chip experiments. (This figure is summarized from (Lee, Rinaldi et al. 2002))

expression are unknown, DNA sequence motif analysis becomes a practical way to shed some lights on the transcriptional regulation. The certain sequence segments recognized by TFs usually have some nucleotide patterns, and this sequence pattern is called binding motif (or simply called motif thereinafter). Motifs are relatively short sequence patterns (6 20 bp), and they are possibly distributed in a quite large range of one gene's promoter region (5,000 50,000 bp). There are several existing strategies to find motifs:

(a) Motif discovery based on sets of co-regulated genes: this is a statistical technique to find over-represented sequence pattern according to gene sets that are selected based on the existing biological knowledge or expression profile analysis. Due to some common features or expression patterns that these genes share, their promoter regions tend to be bound by the same set of TFs as compared to random selected DNA sequences.

(b) Conserved motif analysis by evolution clues: since most regulatory elements are in non-coding regions and show considerable variation in sequence even for the same TF, they are

not easily recognizable. However, binding sites are often preserved through evolution, thus the alignments of othologous regions from different genomes can help make the binding sites apparent. The conservation of motif can help filter out some irrelevant or false positive results from motif discovery methods.

(c) Discovery of Cis-regulatory module (CRM): Cis- is the prefix indicating on the same side, so Cis-regulatory module is referring to the regulatory motif organization and cooperation relationship within short distance. CRM contains multi-TFs that interact or cooperate with each other for either synergistic or antagonistic combinatorial effects. For most eukaryotic genes, the binding of a single TF is not sufficient to regulate transcription. Efficient regulation of downstream targets needs not only the binding of a TF itself, but also other co-operative TFs. For example, researchers have found that there are a number of other motifs, which are significantly enriched close to ERα's binding positions, and these motifs serve as the combinatorial regulatory candidates of ERα, an important TF in breast cancer (Carroll, Meyer et al. 2006).

DNA sequence motif is usually represented as a weighted position matrix, indicating the occurrence probability of "ACGT" in each position. The motif information has been collected and organized by some biological knowledge databases such as TRANSFAC (Matys, Kel-Margoulis et al. 2006).

Protein-protein interaction information

Proteins interact with each other to perform all types of molecular functions. The abnormal change in the collaboration of certain functionally important protein with its partners has been observed and associated with disease status (Taylor, Linding et al. 2009). Therefore, the global picture of PPI is extremely helpful to understand the underlying mechanisms. There are multiple ways to measure the protein interaction, such as Y2H (yeast two hybrid) and affinity purification-MS (Shoemaker and Panchenko 2007). Y2H utilizes the fact that eukaryotic transcription activators have at least two domains called binding domain (BD)

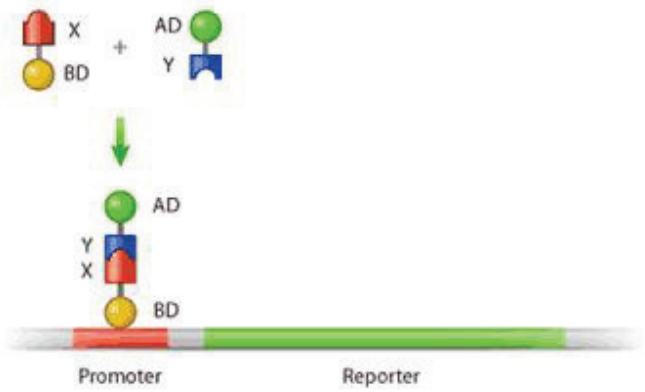


Figure 1.6: Yeast two-hybrid (Y2H) technique for measuring protein-protein interaction. If physical interaction occurs between proteins X and Y, their combination will lead to the activation of the reporter gene that is paired with its promoter. (This figure is from (Sobhanifar 2003))

and activating domain (AD), and the transcription of target gene will be activated only when BD is physically associated with AD domain. By designing the BD (bait) on one protein and AD on the other protein, we can detect whether these two proteins physically interact with each other by checking the transcription level of target gene (also called reporter gene). The most commonly used bait-prey combination is according to transcription factor GAL4 and its target LacZ, BD and AD are fused on two proteins first and according to the expression of reporter gene LacZ the interaction between these two proteins can be determined, as illustrated in Fig. 1.6.

As the protein interaction occur in the interfaces of proteins where the sequence structure matched, the PPI relationship can be computationally predicted based on this information (Shoemaker and Panchenko 2007). Some researchers also show that the interaction can be predicted through co-expression of genes, as the members of same protein complex tend to show consistent expression changes. The co-expression during the evolution could even serve as better evidence than sequence similarity to predict protein interaction. The collections of

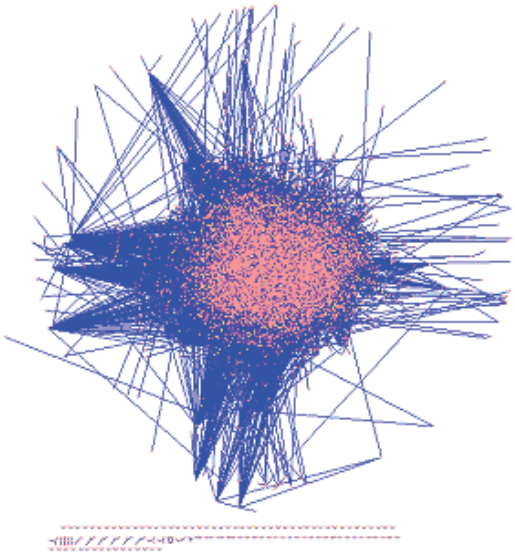


Figure 1.7: Protein interaction network from Human Protein Reference Database (HPRD) (Keshava Prasad, Goel et al. 2009). The network layout is performed using Cytoscape software (Smoot, Ono et al. 2011).

protein interactions derived from text-mining, co-expression and/or experimental evidence are also available in several databases such as STRING (Jensen, Kuhn et al. 2009). Fig. 1.7 shows a PPI network derived from Human Protein Reference Database (HPRD) (Keshava Prasad, Goel et al. 2009).

Biological knowledge based gene set

Gene Ontology (GO), which is the most widely used biological knowledge source, organizes the representation of gene and gene product attributes. Through GO, researchers can find one particular gene's function, cellular location and the biological processes that the gene might be involved in. Several popular websites and web tools are available for GO term annotation, such as AmiGO (Carbon, Ireland et al. 2009), DAVID (Huang da, Sherman et al. 2009) and GSEA (Subramanian, Tamayo et al. 2005). Fig. 1.8 shows an example of

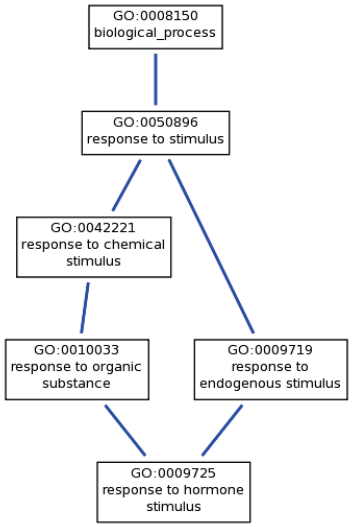


Figure 1.8: Graphic result of querying GO term "response to hormone stimulus" in AmiGO (Carbon, Ireland et al. 2009), which is a web tool to access ontology information.

querying GO term "response to hormone stimulus" in AmiGO (Carbon, Ireland et al. 2009), which is a web tool to access ontology information.

Another important knowledge database is Kyoto Encyclopedia of Genes and Genomes (KEGG: `http://www.genome.jp/kegg/`), which contains a collection of manually annotated molecular pathways. Besides GO terms and KEGG canonical pathways, there are also other biologically defined gene sets, as defined or organized according to biochemistry perturbations, chromosome arrange-ments, or other biological commonalities. These gene sets provide researchers a knowledge reference to check with for looking into a newly generated gene list. There are mainly two common practices to assess the relationship between an observed gene list from experiments and given knowledge gene sets. The first practice is to use over- or under-representative analysis, which is a statistical significance analysis aiming to investigate whether the occurrence of genes belong to certain category is significantly different from random occurrence. The second practice is called gene set enrichment analysis (Subramanian, Tamayo et al. 2005), which is usually based on a given gene ranking, to interrogate how the

ranking distribution of genes in a category is different from a randomly distributed way.

1.3 Problem Statement

1.3.1 Transcriptional regulatory network inference

Transcription factors are the special proteins that control and affect the rate/efficiency of downstream genes' mRNA transcription. Regarding TFs and their target genes as nodes in a graph, and the regulation relation from a TF to its target gene as an edge, this bipartite graph represents transcription regulatory network.

Inference of transcriptional regulatory networks would benefit the understanding of complex cell systems, especially how cells respond to different environmental changes. It can also help reveal the dys-regulation mechanism happening in cancer cells as several key TFs have well-known relationships with tumor progression (Nebert 2002; Libermann and Zerbini 2006). Although the underlying biochemistry regulation relationship from TF to target gene may be highly nonlinear, through log transform and Taylor expansion approximation the transcription regulatory network can be simplified as a linear latent variable model, where the activities of TFs are latent variables and regulatory impacts of TFs to genes are coefficients of the latent variables (Liao, Boscolo et al. 2003).

The biological fact that TF regulates gene through protein-DNA interaction (PDI) provides researchers further clues to tackle the inference problem. There are several information or data sources that provide us potential PDI relationship: DNA sequence motif, ChIP-chip experiment, and biological knowledge database. By incorporating the PDI information into a linear latent variable model, several methods have been proposed to solve the inference problem using regression (Alter and Golub 2004) or constraint regression (Liao, Boscolo et al. 2003; Nguyen and D'Haeseleer 2006). Among them, Network Component Analysis (NCA) is a prominent method, which has mathematically derived identifiability conditions

(Boscolo, Sabatti et al. 2005) and resulted in successful applications for several real biological problems (Brynildsen and Liao 2009; Ye, Galbraith et al. 2009). However, most of the existing methods including NCA do not consider the false positives in the given connections, as they assume that the given biological knowledge is perfect. This unrealistic assumption could degrade the performance of proposed methods greatly and even lead to misleading inference results. Moreover, most of currently available methods only focus on the study of simple organisms such as E. coli and yeast, where the regulation relationship is relatively simple and the PDI information is readily available with good quality. For mouse and human, since these species have much more complicated regulation mechanisms and less confident PDI information, the false-positives within biological knowledge cannot be simply ignored. Integrating gene expression data with in-consistent biological knowledge is a true challenge for the development of computational methods, since the inconsistency can degrade the performance of computational methods greatly or even fail the inference of transcriptional regulatory network completely.

In this dissertation, we address the regulatory network inference problem under two different circumstances, depending on the availability of protein-DNA information:

(a) When protein-DNA information of all TFs is available, we propose a knowledge based stability analysis approach to tackle this problem. By assuming that the consistent expression-knowledge relationship remains stable after small perturbations are applied to the knowledge, we can assign the knowledge-derived estimation (ie., TFA, regulatory strength, etc.) a stability score that is the average distance of multiple estimations upon different perturbations. Having the estimated activities of condition-specific TFs prioritized by stability analysis, we further propose to use multivariate regression to rank condition-specific target genes.

(b) When only protein-DNA information of a few or even single TF is available, it is not realistic to use decomposition or regression-based methods for transcriptional regulatory network reconstruction. We propose to a single TF knowledge guided approach named as regulatory component analysis (RCA), which explicitly captures the TF target genes that

are likely to be expressed in a coordinative manner. RCA is based on a linear extraction scheme and can be solved efficiently using generalized eigenvalue decomposition.

1.3.2 Dys-regulated protein interaction sub-network identification

Protein, the end product of gene transcription and translation processes, is responsible for every biological process and molecular function of living cells. The defects of proteins can lead to diseases such as muscular dystrophy (Khurana and Davies 2003), which is single gene mutation disease, and the re-wiring of the protein signaling network could contribute to the metastasis mechanisms (Chuang, Lee et al. 2007) and clinical outcome (Taylor, Linding et al. 2009) of breast cancer. Therefore, it would be very important to understand the function of protein and the biological relationship among proteins. With the advent of yeast2hybrid technique, protein-protein interactions (PPIs) can now be measured in a high-throughput way and protein-protein interaction data are readily available for several species.

PPI data alone can be utilized to predict proteins' molecular functions according to their adjacent neighbors (Sharan, Ulitsky et al. 2007). The topological characteristics of PPIs are also of interest for cancer studies (Taylor, Linding et al. 2009). In addition, PPI network structure has also been integrated with gene expression data to pinpoint the local changes of biological systems, in response to environmental change or disease status switch. Technically, such changes can be identified from a given overall PPI network by using sub-network identification approaches, aiming to extract certain sub-graph that undergoes significant changes between different biological conditions or disease phenotypes. Several methods have been proposed based on different criterion functions and optimization procedures. For example, p-values of the gene nodes were transformed to normal distributed z-scores and simulated annealing was used in (Ideker, Ozier et al. 2002) to find optimal sub-networks. Mutual information was used in (Chuang, Lee et al. 2007) to narrow down phenotype-specific sub-networks and three different levels of permutation tests were implemented to identify statistically significant sub-networks. In (Dittrich, Klau et al. 2008) a modified

prize-collecting tree was devised to solve the integer programming problem for sub-network identification, and a false discovery rate control was also incorporated into the scheme for significant sub-network identification. Clustering and graph cut techniques were combined in (Ulitsky and Shamir 2007) to find coherent sub-networks. However, there are two major limitations in these approaches. First, there is no importance ranking for the nodes in identified sub-networks, consequently hindering the interpretation of sub-network members. Second, all the protein interactions are treated equally and no importance ranking of edges is provided when sub-network identification is done, leaving the question of how these proteins interact within sub-network unanswered. Moreover, considering PPI data are very general and not conditionally specific, it would not be sufficient to use deterministic methods to solve the sub-network identification problem due to the high rate of false-positives in the PPI data.

With the awareness of above-mentioned problems, we propose a Metropolis sampling-based approach for sub-network identification in this proposal. Instead of searching for an optimal solution, we explore the entire available network using graphical random walk based metropolis sampling. By regarding random walk through the PPI network as the proposal function of metropolis sampling and casting the optimization problem as a distribution learning one, we avoid the strong and unrealistic assumption that every protein-protein interaction is plausible for the biological system under study. Attributed to the Markov chain Monte Carlo (MCMC) nature of Metropolis sampling, we can check to ensure the convergence of the sampling process, and obtain probability-like scores for the nodes and edges in the whole network, reflecting how likely this protein/interaction is involved in given biological condition. Based on proposed sampling approach, we further propose two strategies: one is bootstrapping used for assessing statistical confidence of detected sub-networks; the other is graphic division to separate a large sub-network to several smaller sub-networks for facilitating biological interpretations.

1.4 Summary of Contributions

In the context of the research topics discussed above, we summarize the main contributions of this dissertation:

Transcriptional regulatory network inference by motif-guided network component analysis (mNCA) and knowledge-based stability analysis (kSA)

For the transcriptional regulatory network inference problem where protein-DNA interaction information is complete, we propose to develop a motif-guided Network Component Analysis (mNCA) approach, in which expression profiles and DNA sequence motif information are integrated to infer underlying transcription factors' activities and the regulatory relationship from TFs to their target genes. The mNCA approach can facilitate the inference of regulatory networks in the situation that ChIP-chip data are not available. Since the initial connection information extracted from motif analysis is especially noisy, we further propose a knowledge-based stability analysis (kSA) approach to address the consistency between gene expression data and biological knowledge of motif information. We further propose target identification scheme with TFs prioritized by kSA. We have initially applied kSA to the application of yeast cell cycle study, and successfully revealed most of biologically important TFs associated with yeast cell cycle process. We derive the theoretical justification of stability analysis, revealing that if the perturbation is small enough the stability score is not affected by the perturbation level applied onto the knowledge.

Transcriptional regulatory network inference by regulatory component analysis (RCA)

For the regulatory network inference with incomplete protein-DNA interaction information, we propose a linear extraction scheme called regulatory component analysis (RCA), which can infer underlying regulatory components even with partial biological knowledge. The proposed scheme differs from the matrix decomposition optimization in NCA, and requires full knowledge of all regulatory components. Moreover, it can be applied even with par-

tial knowledge of one regulatory component. The RCA criterion is designed to maximize the coincidence of extracted component with knowledge, rather than fully follow the given knowledge that may be inconsistent to expression data. Thus, RCA is also less affected by false-positives (FPs) and false-negatives (FNs) within biological knowledge. The Rayleigh quotient function of RCA criterion enables efficient computation using generalized eigenvalue decomposition. Simulation-wise, to the best of our knowledge, statistical assumption-based methods (e.g., ICA and PCA) and knowledge guided methods (e.g., NCA and RCA) are fairly compared for the first time. This comparison is also performed with the realistic consideration that given biological knowledge could be incomplete and inconsistent to expression data. Furthermore, real biological expression data with ground truth collected from knowledge database are also designed to compare the performance of all the methods. Therefore, our comparison results would also serve as reference for other researchers in signal processing and bioinformatics.

Dys-regulated protein sub-network identification by Metropolis random walk on graph (MRWOG)

For integrative analysis in protein-protein interaction network, we proposed a novel scheme called Metropolis Random Walk On Graph (MRWOG) to identify the condition-specific sub-networks in a stochastic way. Instead of looking for single sub-network associated with maximum score, we sample multiple sub-networks through a designed random walk on interaction network. Then, we ensemble sampled sub-networks to form an average sub-network to assess the importance of each individual protein node in a global way, not only reflecting its individual association with clinical outcome but also indicating its topological role (hub, bridge) to connect other important proteins. Moreover, each protein node is associated with a sampling frequency score, which enables the statistical justification of each individual node and flexible scaling of sub-network results. Based on MRWOG approach, we further propose two strategies: one is bootstrapping used for assessing statistical confidence of detected sub-networks; the other is graphic division to separate a large sub-network to several smaller sub-networks for facilitating interpretations. MRWOG is easy to use with only two

parameters need to be adjusted, one is beta value for performing random walk and another is Quantile level for calculating truncated posteriori mean.

Related publications during Ph.D. study period:

Journal Publications:

1. **C. Wang**, J. Xuan, H. Li, Y. Wang, M. Zhan, E. P. Hoffman and R. Clarke, "Knowledge-guided gene ranking by coordinative component analysis," BMC Bioinformatics, 11:162, 2010 (related to section 2.5 in Chapter 2)

2. L. Chen, J. Xuan,**C. Wang**, I.-M. Shih, T.-L. Wang, Z. Zhang, R. Clarke, E. Hoffman and Y. Wang, "Biomarker identification by knowledge-driven multi-level ICA and motif analysis," Intl J. Data Mining and Bioinformatics, vol. 3, no. 4, pp. 365-381, 2009 (related to Chapter 2)

3. **C. Wang**, J. Xuan, L. Chen, P. Zhao, Y. Wang, R. Clarke, and E. P. Hoffman, "Motif-directed network component analysis for regulatory network inference, "BMC Bioinformatics, vol. 9, 2008 (related to section 2.3 in Chapter 2)

4. L. Chen, J. Xuan, **C. Wang**, I.-M. Shih, Y. Wang, Z. Zhang, E. Hoffman, and R. Clarke, "Knowledge-guided multi-scale independent component analysis for biomarker identification," BMC Bioinformatics, 9:416, 2008 (related to Chapter 2)

5. 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 (related to Chapter 2)

Conference Publications:

1. **C. Wang**, S. Ha, Y. Wang, J. Xuan and E. P. Hoffman, "Computational Analysis of Muscular Dystrophy Sub-types Using A Novel Integrative Scheme", in Proc. Intl

Conf. on Machine Learning and Applications (ICMLA 2010), Washington, DC, Dec. 2010. (related to Chapter 3)

2. **C. Wang**, J. Xuan, L. L. Chen, R. B. Riggins, E. P. Hoffman, and R. Clarke, "Reliability Analysis of Transcriptional Regulatory Networks," in Proc. Intl Conf. on Bioinformatics, Computational Biology, Genomics and Chemoinformatics, Orlando, FL, July 2008. (related to section 2.3 in Chapter 2)

3. J. Xiong, **C. Wang**, B. Zhang, Y. Wang, E. P. Hoffman, R. Clarke, and J. Xuan, "Inferring Condition-Specific miRNA-Gene Modules from miRNA and mRNA Profiling Data," in Proc. Intl Conf. on Bioinformatics, Computational Biology, Genomics and Chemoinformatics, Orlando, FL, July 2008. (related to Chapter 2)

4. **C. Wang**, J. Xuan, L. L. Chen, P. Zhao, Y. Wang, R. Clarke, and E. P. Hoffman, "Integrative Network Component Analysis for Regulatory Network Reconstruction," in Proc. Fourth Intl Symposium on Bioinformatics Research and Applications, Atlanta, GA, May 2008. (related to section 2.3 in Chapter 2)

5. L. Chen, **C. Wang**, I.-M. Shih, T.-L. Wang, Z. Zhang, Y. Wang, R. Clarke, E. Hoffman and J. Xuan, "Biomarker Identification by Knowledge-Driven Multi-Level ICA and Motif Analysis," in Proc. Sixth Intl Conf. on Machine Learning and Applications, Cincinnati, Ohio, Dec. 2007. (related to Chapter 2)

6. **C. Wang**, J. Xuan, T. Gong, R. Clarke, E. Hoffman, Y. Wang, "Stability Based Dimension Estimation of ICA with Application to Microarray Data Analysis," in Proc. The 2007 Intl Conference on Bioinformatics and Computational Biology, Las Vegas, Nevada, June 2007. (related to Chapter 2)

Chapter 2

Inference of transcriptional regulation network

Revealing of transcriptional regulation network is essential to understand the controlling of mRNA synthesis process in given biological conditions, which in turns affects translation of proteins that are needed for cellular system. To investigate the underlying mechanism using computational methods, many computational efforts have been made through different forms of modeling, using biochemistry PDE model (Kar, Baumann et al. 2009; Honkela, Girardot et al. 2010), Boolean network model (Shmulevich, Dougherty et al. 2002) and simple regression model (Conlon, Liu et al. 2003). In general, the more fine-resolution methods require the larger data-set and more detailed measurement. However, considering that number of available microarray samples is relatively small in focus biological study, simplified model such as linear regression or linear decomposition are very prevalent. In stead of aiming to predict specific regulations accurately, these models are developed to prioritize condition-specific TFs and TF-gene regulation relationships. Through prioritization of important molecular players, biologists can design detailed experiments to validate their functions and influence towards phenotypes. This chapter is divided into two parts: the first part is dedicated to address regulatory network inference when protein-DNA interaction in-

formation of all TFs is complete. The second part is focused on the inference task where only partial protein-DNA interaction information is available.

Table 2.1: Mathematical notations in Chapter 2.

Number of genes, microarray samples and underlying transcription factors	N, M, L
Expression concentration matrix	$\mathbf{E} \in \mathbb{R}^{+^{N\times M}}$
Expression activity matrix	$\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]^T \in \mathbb{R}^{N\times M}$
Expression pattern of the n-th gene	$\mathbf{x}_n \in \mathbb{R}^M$
Regulatory component matrix (RC matrix)	$\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_L] \in \mathbb{R}^{N\times L}$
the l-th regulatory component	$\mathbf{a}_l \in \mathbb{R}^N$
Transcription factor activity matrix (TFA matrix)	$\mathbf{S} = [\mathbf{s}_1, \cdots, \mathbf{s}_L]^T \in \mathbb{R}^{L\times M}$
the l-th TFA	$\mathbf{s}_l \in \mathbb{R}^M$
Score matrix of protein-DNA interactions	$\mathbf{K} \in \mathbb{R}^{N\times L}$
Connectivity pattern matrix	$\mathbf{B} \in (0,1)^{N\times L}$
Regulatory matrix set	$\mathbb{Z}_0$
	$\mathbb{Z}_0 \triangleq \{\mathbf{A} \in \mathbb{R}^{N\times L} \vert a_{nl} = 0 \text{ if } k_{nl} < \eta_l\}$
	or $\mathbb{Z}_0 \triangleq \{\mathbf{A} \in \mathbb{R}^{N\times L} \vert a_{nl} = 0 \text{ if } b_{nl} = 0\}$

2.1 Introduction of Transcriptional Regulation

With the advent of high-throughput microarray technology, researchers can simultaneously profile the dynamic activity of tens of thousands of genes, and generate hypotheses with various statistical and computational tools (Clarke, et al., 2008; Tusher, et al., 2001) to understand molecular mechanisms of phenotypes or disease sub-types. However, bioinformatics researchers have recently become aware that the molecular relationship between different types of genes cannot be ignored in computational analysis, not only for the identification of reliable biomarkers (Chuang, Lee et al. 2007) but also for the understanding of disease mechanisms (Lee, Chuang et al. 2008). Therefore, it is desirable to advocate a systems biology treatment of computational analysis by integrating gene expression data with different types of biological knowledge, such as protein-protein interaction networks (Chang, et al., 2008; Ideker, et al., 2002; Liao, et al., 2003; Wang, et al., 2008), function annotations (Pan, 2006; Zhou, et al., 2002), and pathway information (Lee, et al., 2008; Li and Li, 2008). These integrative approaches can reduce the impact of curse-of-dimensionality effectively by decreasing the number of estimated parameters, and facilitate the interpretation of computational results in order to generate biologically meaningful hypotheses. Nevertheless, a common pitfall is often shared by the abovementioned and many other integrative approaches; that is, given biological knowledge is assumed, with rare exceptions, to be consistent with gene expression data acquired under certain biological conditions. Such an assumption could yield misleading computational results that may generate incorrect biological hypotheses. In this section, we will restrict our discussions to infer transcriptional regulatory networks by integrating gene expression data with the binding knowledge of transcript factors to DNA sequences.

Transcriptional regulation is an essential mechanism for cells to respond fast-changing external conditions (Lee, et al., 2002; Luscombe, et al., 2004). Identification of transcriptional regulatory relationships can help us understand the activation of specific regulators, biological processes and pathways; particularly for cancer research, it is an important step to

reveal diverse molecular pathways dysregulated in cancers. Transcriptional regulatory relationships can be explicitly described as transcriptional regulatory networks (TRNs) (Liao, et al., 2003; Yu and Li, 2005), or functionally categorized as transcriptional regulatory modules (TRMs) (Li, et al., 2007; Segal, et al., 2003). Both TRNs and TRMs consist of transcription factors (TFs) as regulators and downstream genes as their target genes, transcription levels of which are modulated through TFs binding to their corresponding DNA promoter regions. Previously, many computational methods were proposed to identify transcriptional regulatory modules from either gene expression data (Pe'er, et al., 2001; Segal, et al., 2003) or TF-DNA interaction information (Sharan, et al., 2003; Zhou and Wong, 2004) but not both. Later, it became clear that these two information sources complement each other, based on which many integrative approaches were developed to better identify key TF regulators and their corresponding target genes (Bar-Joseph, et al., 2003; Bussemaker, et al., 2001; Chen, et al., 2007; Conlon, et al., 2003; Li and Zhan, 2008; Liao, et al., 2003; Nguyen and D'Haeseleer, 2006). However, some assumptions made in these approaches may not be valid in the reality of biological settings. For examples, some regression-based approaches (Bussemaker, et al., 2001; Conlon, et al., 2003; Nguyen and D'Haeseleer, 2006) assume that regulation relationship is approximately reflected by binding evidence, but this assumption ignores that TF binding does not necessarily suggest effective regulation. Some other approaches approximate the activity of a TF by its mRNA expression level (Wang, Xuan et al. 2008), which is also questionable as the protein level of a TF is known to only have weak correlation with its mRNA expression (Greenbaum, Colangelo et al. 2003); since post-translational modification also plays essential roles in the activation of TF regulators. Moreover, many of the regression schemes estimate the activity or influence of each TF individually (i.e., one by one), as a result possible collaborative regulation among different TFs would likely be overlooked.

Being aware of the above-mentioned limitations, network component analysis (NCA) was proposed to deduce the network and underlying transcription factor activity (TFA) (Liao, Boscolo et al. 2003). NCA is based on a biochemical model-based approach considering regulations of multi-TFs simultaneously: given the initial binary TF-DNA topological con-

nections, the regulation relationship is estimated according to gene expression data rather than being fixed according to initial binding evidence scores; TF activities are determined according to expression patterns of their targets. Several successful applications of NCA have been reported (Brynildsen and Liao, 2009; Rahib, et al., 2009; Yang, et al., 2005; Ye, et al., 2009), as well as a number of methodology improvements (Brynildsen, et al., 2006; Chang, et al., 2008; Dai, et al., 2009; Galbraith, et al., 2006; Sabatti and James, 2006; Tran, et al., 2005; Wang, et al., 2008). However, NCA, along with all other existing integrative approaches for TRN inference, still encounters one critical problem for many biological applications, that is, the given biological knowledge may be (more or less) inconsistent with the gene expression data. The inconsistent biological knowledge could lead computational methods to giving incorrect results and hence false discoveries. Many factors could contribute to such data-knowledge inconsistency, such as noises corrupting biological measurements, differences in biological conditions (e.g., in vivo vs. in vitro; cell line vs. tissue tumor), and inevitable false positives (or false negatives) in biological knowledge itself.

To delineate accurately condition-specific regulation, it is essential to prioritize TF regulators and their target genes with consistent knowledge support. For TF prioritization, conventional approaches either utilize prior knowledge of TF activities (Yang, Suen et al. 2005), or adopt statistical approaches to test the statistical significance of TF associated regulation (Bussemaker, et al., 2001; Conlon, et al., 2003; Tsai, et al., 2005). The former scheme requires prior information about specific biological conditions (e.g., stimulus, response, etc.) under study. For example, for yeast cell cycle studies, it is expected that cell cycle regulators should exhibit their activity with periodic cycles. However, prior information-based approaches will lead to biased results if a priori is inaccurate, and not all applications have complete prior information known beforehand especially for exploratory studies. The significance-based approaches are widely accepted as a norm to evaluate whether one TF (or its binding motif) can significantly contribute to the explanation of expression variations of its targets, or equivalently, whether the expression pattern of target genes can be well fitted by its regulator's activity (Brynildsen, et al., 2006; Galbraith, et al., 2006). These statistical approaches

require an explicit formulation of null hypothesis and generation of corresponding null distribution by permutation. However, null distribution sometimes cannot be easily generated by sample permutation, particularly when different microarray samples are not exchangeable (for example, in time-course microarray data). Moreover, when substantial amount of false knowledge occurs, the estimation accuracy of null distribution is also questionable.

Here, we proposed a novel approach based on stability analysis to address the inconsistency between data and knowledge. The approach is especially capable to reveal condition-specific TFs and downstream target genes for TRN inference. The basic idea of stability analysis is to introduce small perturbations on knowledge and interrogate the variation of resulting estimation when data are examined. Such variations reflect potential estimation deviations introduced by flawed knowledge.

The proposed scheme was extensively tested with simulation data, comparing with its counterparts. It has been shown that the stability analysis-based scheme is effective to prioritize condition relevant TFs and downstream targets when false connection number is relatively large and given connection knowledge is incomplete. We further applied the scheme to yeast cell cycle data to show its improved capability to highlight cell cycle related TFs and targets, without utilizing the prior knowledge of cycle pattern. Finally, the stability analysis is carried out to reveal different estrogen-related transcriptional regulatory networks in estrogen-induced and estrogen-deprived experiments. Such study might provide comparative pictures of regulatory relationship in breast cancer studies, and bring new insights to understand how regulatory networks are involved in estrogen signaling mechanisms.

2.2 Problem Formulation and Existing Methods

2.2.1 Unsupervised linear latent model analysis

Firstly, we briefly review the unsupervised linear latent model analysis for genomic signals. Given a high-dimensional genomic data matrix $\mathbf{X} \in \mathbb{R}^{N \times M}$, which can be seen as N realizations of random vector $\mathbf{x} \in \mathbb{R}^{M}$, the purpose of statistical latent variable algorithms such as PCA and ICA is to find a linear transformation $\mathbf{W} \in \mathbb{R}^{M \times L}$ to reduce the dimension of original genomic data, through which the transformed components of $\mathbf{Y} = \mathbf{XW} = [\mathbf{y}_1, \cdots, \mathbf{y}_L]$ are statistical uncorrelated and independent, respectively. When observed expression data is assumed to be the linear mixtures of underlying sources and components of sources are non-Gaussian distributed and independent, ICA can be used to perform blind separation, estimates of which correspond to underlying sources up to some scale and order ambiguities even without the exact distribution form of latent variable (Cardoso 1998; Lee, Girolami et al. 2000). Many ICA algorithms have been successfully applied to many biomedical problems where the source independence assumption holds (Jung, Makeig et al. 2000; Vigario, Sarela et al. 2000). All these linear models can be summarized as following matrix decomposition model:

$$\mathbf{X} = \mathbf{AS}, \tag{2.1}$$

where the observed data matrix $\mathbf{X}$ is decomposed as the product of two latent data matrices $\mathbf{A}$ and $\mathbf{S}$ with lower dimension.

We illustrate the common biological interpretation of latent variable model in Fig. 2.1. $\mathbf{X}$ is a genomic signal matrix of M measurements by N genomic instances, which could be genes (Liebermeister 2002), metabolisms (Scholz, Gatzek et al. 2004), or genome loci (Dawy, Sarkis et al. 2008). It is generally assumed that realizations of l-th latent component $[\mathbf{a}_1, \cdots, \mathbf{a}_L]$ reflect the genomic influences of underlying biological processes to all genomic instances. It is further assumed that (from energy efficiency point of view of cellular system) each biological process can only affect the activities of small portions of genomic instances, therefore a

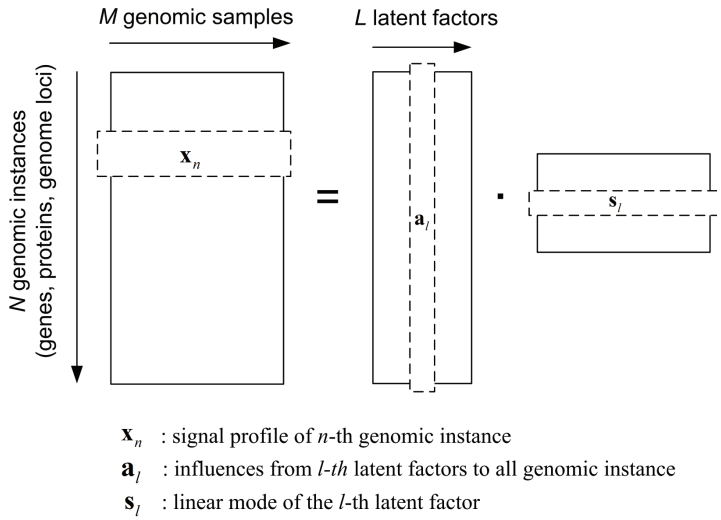


Figure 2.1: General linear latent variable model for genomic signals.

super-Gaussian distribution of each component $\mathbf{a}_l$ can be approximately assumed. This assumption is partially verified though previous studies involving the comparison between ICA and PCA, where non-Gaussianity assumption based ICA clearly outperforms PCA in revealing biological meaningful results (Liebermeister 2002; Lee and Batzoglou 2003).

Despite the initial successes of applying statistical linear latent variable methods for gene expression analysis, several limitations of totally "blind" approaches still exist: first, the underlying true dimension is hard to decide, and over-/under-estimation of source number will apparently lead to misinterpretation of expression data; second, expression level measurement is acquired through a sophisticated process where large amounts of errors and noises relying in data (Klebanov and Yakovlev 2007), pure data-driven approach suffers reproducibility problems (Kreil and MacKay 2003); third and most importantly, "blind" approaches can only produce expression analysis in relatively low resolution with gene groupings, and the specific biological interpretation is difficult to proceed with virtual signal sources based statistical assumptions. As a general trend for advanced bioinformatics and systems biology, it is usually required that the computational approach can lead to biologically testable hypothesis (Lander 2010).

The applications of statistical latent variable algorithms are mainly limited in exploratory analysis of genomic data, but the resulted computational results are too general to be interpreted in a specific biological context. Therefore, we focus on gene expression analysis for transcriptional regulatory network inference, where a clear generative model can be formulated with biological implications, discussed in following sections.

2.2.2 Log-linear model for transcription expression

Gene expression generally refers to an information conversion process from DNA sequence of one gene to its messenger RNA (mRNA), which will be further translated to corresponding protein(s). Therefore, mRNA molecular concentrations of genes are generally called as gene expression levels, or expression data for short. Expression data is acquired through a series of biochemistry-photo transformation, providing the parallel mRNA measurement of thousands of genes in a single microarray chip. Gene expression is one of the data types received the most intensive research attentions, not only because of its relatively cheap cost to acquire, but also it can reflect the genetic dynamics of cellular system (Stafford and Yidong 2007).

Having M microarray measurement with N genes, we denote the raw concentration data of mRNA as a matrix $\mathbf{E} \in \mathbb{R}^{+^{N\times M}}$, in which e_{nm} reflects the concentration of n-th gene in m-th microarray measurement. We denote normal concentration of n-th gene as $e_n^{(0)}$, which is usually generated in baseline condition as reference signal. It has been shown in (Liao, Boscolo et al. 2003) that transcription rate of one gene is decided by concentrations of several controlling proteins named transcription factors (TFs). Specifically, $V_{promoter,n}(t)$, the rate of mRNA synthesis (promoter activity) at time point t, and the rate of mRNA degradation $V_{degradation,n}(t)$ of the n-th gene are defined as follows according to Hill equation (Liao, et al., 2003; Ronen, et al., 2002):

$$V_{promoter,n}(t) = \frac{\lambda_n \prod_{l=1}^{L} \left(\frac{c_l(t)}{k_{nl}}\right)^{a_{nl}}}{1 + \lambda_n \prod_{l=1}^{L} \left(\frac{c_l(t)}{k_{nl}}\right)^{a_{nl}}}$$

and

$$V_{degradation,n}(t) = k_{degradation,n} e_n(t),$$

where k_{nl} and a_{nl} are kinetic parameters from the l-th TF to the n-th gene; λ_n and $k_{degradation,n}$ are synthesis and degradation parameter for the n-th gene. Assume that mRNA levels reach a quasi-steady state:

$$V_{promoter,n}(t) - V_{degradation,n}(t) = 0,$$

after some mathematical manipulations we have following relationship between the ratio level of mRNA and ratio levels of TFs:

$$\frac{e_n(t)}{e_n(0)} = \prod_{l=1}^{L} \left(\frac{c_l(t)}{c_l(0)} \right)^{a_{nl}} \frac{1 + \prod_{l=1}^{L} \left(\frac{c_l(0)}{k_{nl}} \right)^{a_{nl}}}{1 + \prod_{l=1}^{L} \left(\frac{c_l(t)}{k_{nl}} \right)^{a_{nl}}}.$$

It has been further assume in (Liao, et al., 2003) that $c_l(t)$ is around the neighborhood of $c_l(0)$ so that the term $\frac{1+\prod_{l=1}^{L} \left(\frac{c_l(0)}{k_{nl}} \right)^{a_{nl}}}{1+\prod_{l=1}^{L} \left(\frac{c_l(t)}{k_{nl}} \right)^{a_{nl}}} \approx 1$. Finally, we replace the time index using the discrete sample index m, denote baseline mRNA level of the n-th gene as $e_n^{(0)} = e_n(0)$, and denote baseline TF concentration of the l-th TF as $c_l^{(0)} = c_l(0)$, we come to a linear approximation of transcriptional regulation under equilibrium assumptions:

$$\frac{e_{nm}}{e_n^{(0)}} = \prod_{l=1}^{L} \left(\frac{c_{lm}}{c_l^{(0)}} \right)^{a_{nl}}, \tag{2.2}$$

in which, c_{lm} and $c_l^{(0)}$ are concentrations of the l-th TF under m-th microarray measurement and baseline condition, respectively; exponential item a_{nl} reflects how l-th TF regulates the transcription rate of n-th gene, $a_{nl} = 0$ as no regulation, $a_{nl} > 0$ as transcription promotion (or up-regulation), $a_{nl} < 0$ transcription suppression (or down-regulation). It needs to be emphasized that only expression concentration e_{nm} and $e_n^{(0)}$are directly observable, while c_{lm},$c_l^{(0)}$ and a_{nl} are all hidden variables.

By denoting

$$x_{nm} = \log \frac{e_{nm}}{e_n^{(0)}} \tag{2.3}$$

and

$$s_{lm} = \log \left(\frac{c_{lm}}{c_l^{(0)}} \right), \tag{2.4}$$

the Eq. (2.2) is expressed as

$$x_{nm} = \sum_{l=1}^{L} a_{nl} s_{lm} \tag{2.5}$$

or in a matrix multiplication form with an additive noise matrix $\mathbf{\Gamma} \in \mathbb{R}^{M \times N}$

$$\mathbf{X} = \mathbf{AS} + \mathbf{\Gamma}. \tag{2.6}$$

Eq. 2.6 can further be written in latent variable model with respect to gene index n:

$$\mathbf{x}_n = \sum_{l=1}^{L} a_{nl} \mathbf{s}_l + \gamma_n, \tag{2.7}$$

where $\mathbf{x}_n = [x_{n1}, \cdots, x_{nM}]$ and $\gamma_n = [\gamma_{n1}, \cdots, \gamma_{nM}]$ are gene expression profile and noise vectors of n-th gene; $\mathbf{s}_l = [s_{l1}, \cdots, s_{lM}]$ is the hidden activity vector of l-th TF. Eq. (2.6) is actually called log-linear model, considering the transformation from original Eq. (2.2) to (2.5) (Liao, Boscolo et al. 2003). It has also been observed that the log-ratio transformation of gene expression data approximately fit with Gaussian distribution (Liebermeister 2002). Different from general latent variable analysis, now everything has clear biological implication: latent factors of Eq. (2.6) correspond to the controlling proteins - transcription factors (TFs). We defined l-th row of matrix $\mathbf{S} = [\mathbf{s}_1, \cdots, \mathbf{s}_L]^T$ as l-th transcription factor activity (TFA), which reflects the hidden protein relative activity of l-th TF. The influence variable of l-th TF $\mathbf{a}_l$ is called l-th regulatory component (RC). Through this dissertation, $\mathbf{S}$ and $\mathbf{A}$ are referred as TFA matrix and regulatory component matrix (or RC matrix) for highlighting their biological implications.

2.2.3 Network component analysis (NCA)

For transcriptional regulatory network inference, some biological knowledge could facilitate the estimation of latent activities and controlling relationships. Recall that each regulatory component $\mathbf{a}_l$ in Eq. (2.7) corresponds to the controlling effect of certain TF to genes' transcription rates. It is known that one TF has to bind on DNA promoter region of certain gene in order to regulate the expression of this gene. Such physical binding relationship is usually measured through biological experiments (Wu, Smith et al. 2006) or predicted through computational sequence analysis (Ji and Wong 2006). Based on TF to gene binding evidences, we encoded regulation relationships from TFs to genes as a network connectivity pattern $\mathbf{B} \in (0,1)^{N \times L}$, which is a binary matrix with element $b_{nl} = 1$ indicating potential regulatory relationship from l-th TF to n-th gene.

Usually we called genes controlled by TFs as *target genes*. Assuming there is no feedback from target genes to TFs, the transcriptional regulatory network describing the relationship between TFs and target genes is a bipartite network, where the nodes of latent layer and of observed layer are TFs and downstream genes, respectively. Regulatory component matrix $\mathbf{S}$ describes weights of bipartite network edges. Therefore, estimation of hidden regulatory components is equivalent to inference of underlying regulatory network, illustrated as Fig. 2.2.

To solve Eq. (2.6) based on available biological knowledge $\mathbf{B}$, the original NCA algorithm is designed to estimate $\mathbf{A}$ and $\mathbf{S}$ by minimizing fitting errors (Liao, Boscolo et al. 2003):

$$(\hat{\mathbf{A}}, \hat{\mathbf{S}}) = \underset{(\mathbf{A},\mathbf{S})}{\arg\min} \, ||\mathbf{X} - \mathbf{AS}||_2^2, \tag{2.8}$$

$$\text{s.t. } \mathbf{A} \in \mathbb{Z}_0. \tag{2.9}$$

In (2.9), $\mathbb{Z}_0$ is a regulatory matrix set, deriving from biological knowledge of connectivity matrix:

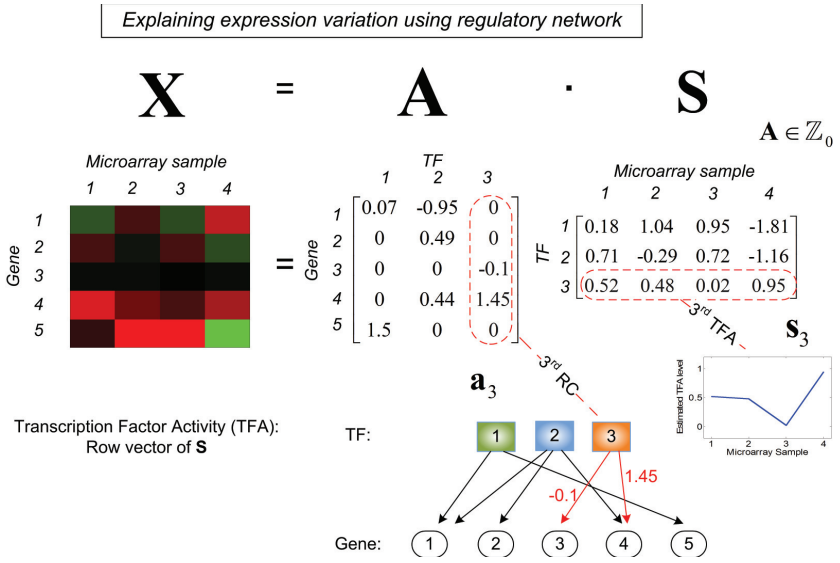


Figure 2.2: Illustrative figure of regulatory component (RC) and transcription factor activity (TFA) in NCA model.

$$\mathbb{Z}_0 \triangleq \{\mathbf{A} \in \mathbb{R}^{N\times L} | a_{nl} = 0 \text{ for } b_{nl} = 0\}. \tag{2.10}$$

Assuming the elements of noise matrix $\mathbf{\Gamma}$ is i.i.d Gaussian distributed, NCA criterion is equivalent to maximization likelihood with respect to noise distribution (Boscolo, Sabatti et al. 2005). It is interesting to notice that NCA criterion does not incorporate any statistical priori of $\mathbf{A}$ or $\mathbf{S}$. This is motivated by the discussions in (Liao, Boscolo et al. 2003) that statistical assumption may not fit to biological reality. Therefore, the NCA criterion is simply a least-squares with structure constraint on $\mathbf{A}$. From the perspective of source separation, we can regard $\mathbf{A}$ as mixing matrix and $\mathbf{S}$ as underlying source signals, or vice versa. Noticeably, $\mathbf{A}$ is more appropriate to be assumed as underlying sources than $\mathbf{S}$ for applying statistical latent variable methods. It is because non-Gaussianity assumption of each component $\mathbf{a}_l$ approximately holds, considering the fact that one TF can only regulate a small portion of genes (Liao, Boscolo et al. 2003; Boscolo, Sabatti et al. 2005).

2.2.4 Estimation ambiguities and identifiability conditions of N-CA

Although priori biological knowledge eliminates the ordering ambiguity of regulatory components, the scaling of underlying signals is still undetermined. Therefore, even with fulfillment of all identifiability conditions, estimated regulatory component could still differ from underlying true signals $\mathbf{A}$ up to some scaling ambiguity $\hat{\mathbf{A}} = \mathbf{AD}$, where $\mathbf{D}$ is arbitrary diagonal matrix with non-zero diagonal items. Such ambiguity is usually acceptable as it is wave-form preserved.

As NCA solution optimization involved biological knowledge $\mathbf{B}$, the structure characteristic of $\mathbf{B}$ is essential for NCA estimation. This is reflected from identifiability conditions of NCA. In the noiseless case, that is when $\mathbf{\Gamma} = \mathbf{0}$, the identifiability conditions for NCA are proved when the following four assumptions are met (Liao, Boscolo et al. 2003) (We adopt As. as the abbreviation for Assumption):

Identifiability conditions of NCA

(As. 1) The microarray sample number M should be greater or equal to TF number L.

(As. 2) Different TFAs $\mathbf{s}_{l(l=1,\cdots,L)}$ are linear independent.

(As. 3) For connectivity pattern matrix $\mathbf{B}$, if any TF is taken out, the modified connectivity pattern matrix $\tilde{\mathbf{B}}$ by removing the genes associated with this TF should have full row rank (rank $= L - 1$).

(As. 4) The network connectivity pattern $\mathbf{B}$ is perfectly known a priori.

Both As. 1 and As. 2 are almost universal presumptions for linear latent algorithm. As. 1 is generally needed to ensure the problem is not underdetermined. As. 2 is also similar to the presumptions for PCA/ICA models that mixing matrix $\mathbf{S}$ is non-singular.

As. 3 is equivalently saying that if one TF is determined, the rest $L - 1$ TFAs can still be determined uniquely. If the As. 3 is not meet, trimming of topological connections

is suggested to be performed. By explicitly exploiting property of As. 3, Chang and etc. (Chang, Ding et al. 2008) proposed an alternative algorithm fastNCA, which could be several tens times faster than original NCA algorithm. Since As. 3 is not always fulfilled for given connectivity, a condition check is usually carried out and connections violating As. 3 will be pruned (Liao, Boscolo et al. 2003). However, it is obvious that an effective condition check for As. 3 also relies on As. 4, assuming that given $\mathbf{B}$ reflects underlying true relationship $\mathbf{B}_0$. Therefore, it can be expected that the estimation accuracy of both NCA and fastNCA heavily depends on availability and quality of given biological knowledge.

2.2.5 Motif-directed network component analysis (mNCA)

Noticeably, most of current NCA applications are focused on simple cell system such as E. coli (Kao, Yang et al. 2004) and yeast (Liao, Boscolo et al. 2003; Yang, Suen et al. 2005). This is because complete biological connection data, such as high-throughput ChIP-chip data, are often not available for common species including rodent and human.

To solve this limitation, we propose a motif-directed NCA (mNCA) approach for regulatory network inference, which utilizes sequence motif information to construct initial connections and later integrates with gene expression data to estimate the activities and downstream targets of transcription factors. First, the upstream regions of the genes can be extracted from the database PromoSer (Halees, Leyfer et al. 2003). Second, Match^{TM} (Kel, Gossling et al. 2003) (or its improved version, P-Match (Chekmenev, Haid et al. 2005)) can be used to search the transcription factor binding sites (TFBSs) in each upstream region; this approach generates the scores of both "core similarity" and "matrix similarity" for each matched motif. Third, Match searches the TFBS for its position-weighted matrices (PWMs) that can be extracted from the TRANSFAC 11.1 Professional Database (Matys, Kel-Margoulis et al. 2006). Fourth, according to the PWMs, a motif score can be calculated for each TF-gene pair where the score is the maximum of the average scores of core similarity and matrix similarity. These motif scores provide the initial connection information for further mNCA

analysis as is detailed in the next section.

As TF binding motif is a relatively short sequence pattern, the topology obtained from motif information is very noisy and contain many false positives. Since the initial topology information is often unreliable for any specific TF-gene pair, we are going to address the consistency between biological knowledge derived from DNA motif information and expression data in the next section, through proposed stability analysis.

2.3 Knowledge-based Stability Analysis

2.3.1 Problems associated with biological knowledge

Since NCA is a network structure constraint approach to infer regulatory activities and relationships, the quality of $\mathbb{Z}_0$ and its consistency to data $\mathbf{X}$ will affect the accuracy of inferred networks. Previously, we simply assume that some binary matrix B is available to construct $\mathbb{Z}_0$. In reality, we know that $\mathbb{Z}_0$ is actually derived from some TF-DNA binding evidence such as DNA motif matching scores or ChIP-chip binding p-values, which have some dependence to regulation but cannot fully decide the occurrence of true regulation. To facilitate the following discussion, we use a matrix form to denote all binding evidence scores as a matrix $\mathbf{K} \in \mathbb{R}^{N\times L}$. Assume the higher a binding evidence score b_{nl} is, the more likely that the promoter region of the n-th gene can be bound by the corresponding l-th TF. To distinguish a likely regulatory relationship from an unlikely one, we can setup a cut-off threshold η_l for the l-th TF. If some TF-target pair between the l-th TF and the n-th gene has a binding evidence score below η_l, we determine this as an unlikely regulatory relationship. Therefore, we can redefine $\mathbb{Z}_0$ in Eq. (2.10) as

$$\mathbb{Z}_0 \triangleq \{\mathbf{A} \in \mathbb{R}^{N\times L} | a_{nl} = 0 \text{ if } k_{nl} < \eta_l\}. \tag{2.11}$$

Please see Fig. 2.3 for this basic idea of transforming binding score to $\mathbb{Z}_0$.

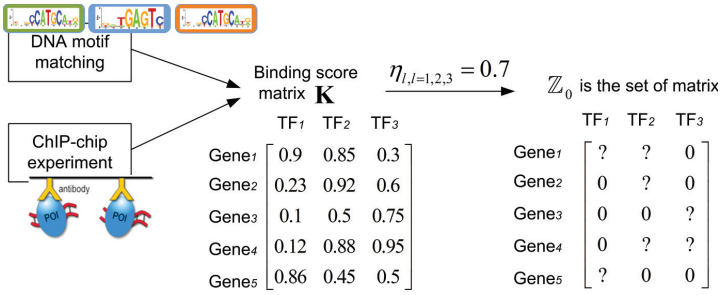


Figure 2.3: An illustration of constructing the regulatory matrix set $\mathbb{Z}_0$ from binding knowledge.

However, we understand that biological information sources could contain considerable amount of errors due to their incompleteness or lack of condition-specific knowledge. The binding evidence score can come from different information sources, such as ChIP-chip experiments (binding p-value), DNA motif sequence matching results (motif matching score) and literature surveys (text mining score). Even with reliable binding information, the occurrence of transcriptional regulation still cannot be fully decided, because transcription regulation is also affected by other factors, such as whether DNA is accessible at that time, and whether other necessary co-factors also bind onto the adjacent regions. Therefore, it is not sufficient to distinguish targets and non-targets purely based on binding scores. $\mathbb{Z}_0$ and its related NCA inference are heavily influenced by false positives (FPs) ($b_{nl} \geq \eta_l$ but the l-th TF does not regulate the n-th gene) and false negatives (FNs) ($b_{nl} < \eta_l$ but the l-th TF actually regulates the n-th gene) within $\mathbb{Z}_0$. As in many real applications the FPs and FNs of biological knowledge is known to be relatively high, the confidence of inference results could be very problematic. It has been shown that by directly applying NCA to yeast gene expression data with ChIP-chip data, the results could be as bad as only comparable to that with random network information (Brynildsen, Tran et al. 2006). Since the DNA binding score only provides incomplete evidence for TRN inference, we emphasize that it is essential to filter out less irrelevant and inconsistent knowledge for a further integrated analysis of gene expression data. We propose a novel stability analysis in the following section to perform

this task for TRN identification.

2.3.2 Basic idea of stability analysis

Stability analysis was originally proposed to perform model selection for various types of machine learning algorithms (Tilman, Volker et al. 2004). The basic idea of stability analysis can be described as follows: with multiple resampled versions of input data, the most consistent (stable) estimation will only occur when an appropriate model is chosen, which fits correctly to the underlying structure of the data. Specifically, stability has been shown to be tightly linked with the generalization capability of any supervised learning approach (Bousquet and Elisseeff 2002; Subramanian, Tamayo et al. 2005). Recently, stability has been also revealed as an effective criterion to perform feature or variable selection (Calle and Urrea, 2010; Kalousis, et al., 2007; Křížek, et al., 2007; Meinshausen and Bhlmann, 2010). The estimation consistency, or stability, reflects how relevant a feature or variable of interest is with respect to a specific machine learning task. In bioinformatics applications, stability analysis has also been used for gene ranking (Boulesteix and Slawski, 2009; Calle and Urrea, 2010) and classification label correction to deal with high levels of experimental noises and errors.

However, there is no existing stability analysis scheme to explicit evaluate the consistency between biological knowledge and gene expression data. Here, following a similar philosophy but with a different quest, we propose a new scheme, by adding small perturbations to given topological connections, to prioritize condition-specific TFs and their target genes. We will develop this so-called knowledge-based stability analysis scheme to identify condition-specific regulatory networks by integrating gene expression data and binding information. As we mentioned before, when network (topology) information is consistent with expression data, NCA can lead to correct TFA estimation hence regulatory relationships. But such consistency is not guaranteed, considering that topological information usually contains many false positives/negatives and expression data are often very noisy. As true TFA measure-

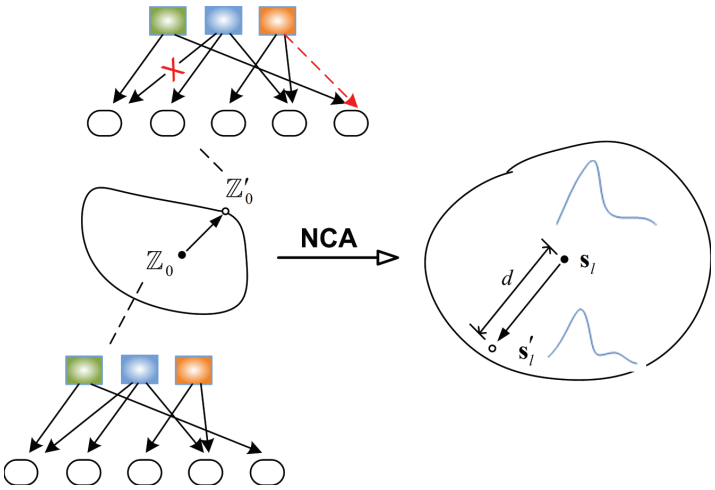


Figure 2.4: Illustration of basic ideas of stability analysis.

ments are unavailable and definite target genes are unknown in real experiments, we must clarify whether the estimated TFAs and regulatory relationships are reliable or arisen by chance. Only with purified TFs and their known targets, which are relevant to our study of interest, we can proceed to infer regulatory networks correctly via TFA estimation. For available topological information with errors, we propose to use a perturbation analysis to test the reliability of TFAs for regulatory network inference. To introduce proper perturbations of topological information, we can modify $\mathbb{Z}_0$ by deleting some existing edges or adding some non-existing edges; we can also add small amount noises in binding score $\mathbf{K}$ (if available) for perturbation analysis.

Taking an example shown in Fig. 2.4, we aim to study the regulatory role of the l-th TF through its TFA estimation $\mathbf{s}_l$ but the quality of biological knowledge, network information ($\mathbb{Z}_0$), is our concern. We can keep the expression unchanged and generated a perturbed $\mathbb{Z}'_0$ by intentionally altering a small amount of entries in given $\mathbb{Z}_0$. As a result, we would expect a deviation (denoted as d) between estimated $\mathbf{s}'_l$ (based on the perturbed $\mathbb{Z}'_0$) and original estimation $\mathbf{s}_l$(based on $\mathbb{Z}_0$). The rationale behind our stability analysis is that the TFA estimation of a condition-specific TF should be more robust to a small amount of perturba-

tion than any non-specific TF that generally lacks of support in data-knowledge consistency. With multiple different perturbations applied to $\mathbb{Z}_0$, we can define an evaluation metric of robustness as the average deviation of different estimations with respect to multiple small perturbations. By perturbing the initial network topology, we will perform a stability analysis on the variation of estimated TFAs and predicted TF-target relationships. A falsely estimated TFA, which is either caused by unspecific TF or inconsistent topological information, tends to be altered easily by small perturbations. On the contrary, an active TF (with a relatively good consistency between expression data and topology knowledge) will tend to keep its activity pattern stable throughout multiple perturbations. The prioritization of true target genes will depend on a stability analysis of prediction errors, as one gene could be regulated by multiple TFs and its expression pattern should be explained, at least in large part, by the TFAs of its regulators.

Specifically, we propose two strategies of stability analysis for TF and target gene prioritization, with both strategies focused on identification of condition-specific TFs and their target genes by intentionally altering the network topology information. After active TFs and target gene subsets are obtained, the TFAs are further refined (estimated) for a significance analysis of regulatory relationship. The overall workflow of the proposed stability scheme is shown in Fig. 2.5.

2.3.3 Knowledge perturbation and stability score

We apply multiple perturbations on biological knowledge (network topology) for stability analysis, each individual perturbation leading to a different estimated TFA. That is, we apply P independent perturbations on original $\mathbb{Z}_p = Perturb(\mathbb{Z}_0, \alpha, \beta)$ to obtain different versions of $\mathbb{Z}_p = Perturb(\mathbb{Z}_0, \alpha, \beta)$; we control the degree of perturbation to be very small (FP = FN = $\alpha \ll 1$ with respect to $\mathbb{Z}_p = Perturb(\mathbb{Z}_0, \alpha, \beta)$). Mathematically, we denote the perturbation function as follows:

$$\mathbb{Z}_p = \text{Perturb}_\alpha(\mathbb{Z}_0). \tag{2.12}$$

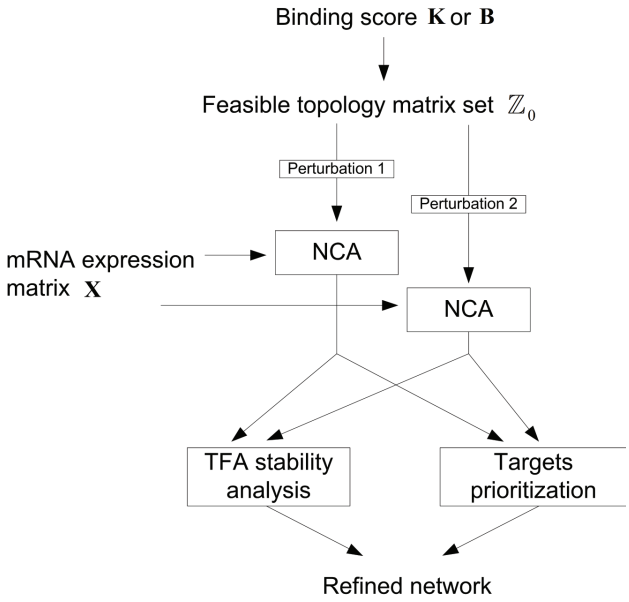


Figure 2.5: Workflow for proposed stability analysis schemes for transcription regulatory network inference.

The estimated TFA of the l-th TF (with respect to expression data $\mathbf{X}$) and the p-th perturbed regulatory matrix set $\mathbb{Z}_p$ can be represented, respectively, as

$$\hat{\mathbf{s}}_{l,p} = NCA_{l\text{-thTFA}}(\mathbf{X}, \mathbb{Z}_p) \tag{2.13}$$

and

$$\hat{\mathbf{a}}_{l,p} = NCA_{l\text{-thRC}}(\mathbf{X}, \mathbb{Z}_p). \tag{2.14}$$

We define a stability measure, namely instability score of TFA, (for the l-th TF) as follows:

$$\mathrm{IST}_l \doteq \frac{1}{P(1-P)} \sum_{p1=1}^{P} \sum_{p1=1, p1 \neq p2}^{P} d(\hat{\mathbf{s}}_{l,p1}, \hat{\mathbf{s}}_{l,p2}). \tag{2.15}$$

Similarly we also define a second instability score, called instability of RC (ISR), as

$$\mathrm{ISR}_l \doteq \frac{1}{P(1-P)} \sum_{p1=1}^{P} \sum_{p1=1, p1 \neq p2}^{P} d(\hat{\mathbf{a}}_{l,p1}, \hat{\mathbf{a}}_{l,p2}). \tag{2.16}$$

Table 2.2: Stability analysis algorithm for TFs.

$(\mathrm{IST}_l, \mathrm{ISR}_l) = \mathrm{Stab\_Analysis}(\mathbf{X}, \mathbb{Z}_0)$

Input: Expression matrix $\mathbf{X}$, regulatory matrix set $\mathbb{Z}_0$, perturbation level α

Output: Instability scores of TFA and RC for each TF: $(\mathrm{IST}_l, \mathrm{ISR}_l), l = 1, \cdots, L$

Algorithm flow:

For $p = 1$ to P

Add perturbations to the topology; $\mathbb{Z}_p = \mathrm{Perturb}_\alpha(\mathbb{Z}_0)$

Recalculate TFA and RC based on perturbed topology; $(\hat{\mathbf{A}}_p, \hat{\mathbf{S}}_p) = \mathrm{NCA}(\mathbf{X}, \mathbb{Z}_p)$

Store $(\hat{\mathbf{A}}_p, \hat{\mathbf{S}}_p)$;

Endfor

For $l = 1$ to L

Calculate In-stability Score of TFAs; $\mathrm{IST}_l = \frac{1}{P(1-P)} \sum_{p1=1}^{P} \sum_{p1=1, p1 \neq p2}^{P} d(\hat{\mathbf{s}}_{l,p1}, \hat{\mathbf{s}}_{l,p2})$

Calculate In-stability Score of RCs; $\mathrm{ISR}_l = \frac{1}{P(1-P)} \sum_{p1=1}^{P} \sum_{p1=1, p1 \neq p2}^{P} d(\hat{\mathbf{a}}_{l,p1}, \hat{\mathbf{a}}_{l,p2})$

Endfor

In both Eq. (2.15) and (2.16), the distance function $d_0(\mathbf{v}_1, \mathbf{v}_2)$ is defined in the way that intrinsic ambiguity of NCA solutions will not be taken into account:

$$d_0(\mathbf{v}_1, \mathbf{v}_2) = \min(\cos(\mathbf{v}_1, \mathbf{v}_2), \cos(\mathbf{v}_1, -\mathbf{v}_2)), \tag{2.17}$$

in which, $\cos(\cdot, \cdot)$ is the cosine distance. However, the range of such distance function is very narrow. When we calculate the averaged distance value, the final estimation could be heavily affected by the outliers. To further improve it, we propose a modified distance

$$d(\mathbf{v}_1, \mathbf{v}_2) = \log\left(\frac{d_0(\mathbf{v}_1, \mathbf{v}_2)}{1 - d_0(\mathbf{v}_1, \mathbf{v}_2)}\right). \tag{2.18}$$

The operator $\log\left(\frac{x}{1-x}\right)$ spans the distance range to be much larger and makes the resulted distribution more normal like. After the modification, the distance calculation is less affected by outliers than the un-modified distance function.

The algorithm of stability analysis for TFs is described as Table 2.2.

2.3.4 Target gene identification based on stably estimated TFA

Once having the TFA estimates of relevant TFs prioritized by kSA, we can identify target genes controlled by these TFs utilizing the relationship in Eq. (2.7), which describes that the individual expression profile $\mathbf{x}_n$ of the n-th gene can be represented as a summation of TFAs weighted by regulatory strength a_{nl} from the l-th TF to the n-th gene, plus a noise term γ_n.

A natural question is can we directly use a_{nl} to decide TF-gene controlling relationship? Although in general the larger value of a_{nl} the more likely the n-th gene is actually regulated by the l-th TF, different genes may not be directly comparable as they have distinct baseline activities. Moreover, we also wish to know statistical confidence for one particular regulatory relationship associate with value of each a_{nl}. Therefore, we use a multivariate regression scheme to further decide the statistical significance of TF-gene regulatory relationship. Basic idea of specifying the regulation relationship between TF and gene is as following: when we perform stability analysis for prioritizing condition-specific TFs, we can acquire stably estimated TFAs. For n-th gene, we are going to use the activities of its potential regulators ($\{\mathbf{s}_l | a_{nl} \neq 0\}$) to regress its expression profile $\mathbf{x}_n$ and calculate the resulted p-value of regression coefficient, which is the statistical significance level reflecting how well the gene expression can be fit by corresponding TFAs. Considering that many TF-target relationships will be tested, multiple hypothesis testing correction scheme described in (Storey and Tibshirani 2003) will be performed to calculate the false discovery rate (FDR) of TF-target relationship.

Based on the regression procedures described as above, we further propose to distinguish "foreground" genes (which are truly regulated by active TFs with consistent data-knowledge support) from "background" genes (the expression pattern of which cannot be reliably predicted by its TF regulators' activities). We simply define the relevance score of downstream

gene (RSDG) based on its average significance level based on TFA regression analysis:

$$\mathrm{RSDG}_n = \frac{-\sum_{l=1}^{L} \log(\text{p-value}(a_{nl}))}{\#(a_{nl} \neq 0)}, \tag{2.19}$$

where $\#(\cdot)$ is the operator to count the number of elements. The higher the score, the more likely this gene is a "foreground" gene truly regulated under this circumstance.

To summarize, for stability-based inference of transcriptional regulatory networks, we propose two consecutive steps: i) Identify condition-specific TF by stability analysis; ii) Identify condition-specific downstream target genes based on stable TFA estimates.

2.3.5 Under-determined case (more TFs than microarray samples)

Another issue has to be considered is one of NCA identifiability conditions, which requires larger microarray sample number than TF number. In the real application, we usually encounter the reverse situation, much larger TF number than the sample number. To avoid this limitation, we proposed to randomly sample a small number (¡L) of TFs each time, calculate and store their instability scores. After multiple times of above-mentioned random sampling, we will calculate averaged instability score of each TF across multiple random divisions. The reasoning behind this scheme is that if a TF actively participate under certain condition and regulate its targets, it should also be averagely stable in randomly sampled sub-networks. The algorithm for under-determined case is described in Table 2.3.

Table 2.3: Stability analysis algorithm for under-determined case.

$aISG_l$ = Stab_Analysis ($\mathbf{X}$, $\mathbb{Z}_0$)

Input: Expression matrix $\mathbf{X}$, regulatory matrix set $\mathbb{Z}_0$, perturbation level α
Number of random division R, sub-TF set number L_{sub} ($L_{sub} \le L$)

Output: Average Instability score $aISG_l$

Algorithm flow:

For r = 1 to R
  Select L_{sub} distinct TFs, their targets and corresponding $\mathbb{Z}_{sub}$ ($\mathbb{Z}_{sub} \subset \mathbb{Z}_0$)
  Record the index of selected TFs
  Stab_Analysis ($\mathbf{X}_{sub}$, $\mathbb{Z}_{sub}$)
  Record IST, ISR for each selected TF
Endfor

For l = 1 to L
  For l-th TF, according to the selections, calculate its averaged instability scores.
Endfor

2.4 Experiments for Stability Analysis

2.4.1 Simulation studies

To test the performance for regulatory network identification, we generated simulated regulatory networks consisting of TFs and target genes. The simulated expression data were generated based on Eq. (2.6). The performance of an algorithm was evaluated using 50 randomly generated networks across different signal-to-noise ratios (SNRs).

TF prioritization

To realistically simulate real microarray data studies, we purposely added some 'non-relevant' or 'less-relevant' TF connections in the simulation data. We simulated these two cases separately as follows: (a) 'non-relevant' TF case: gene expression data were generated as all genes were regulated by 15 TFs according to Eq. (2.6), and false connections of another 15

TFs, which have no impact on regulating expression of the all genes, were added. (b) 'less-relevant' TF case: gene expression data were generated as regulated by 30 TFs; within 30 TFs, 15 of them were associated with moderate false connections (FP/FN rate = 2%), while the connections of remaining 15 TFs were contaminated with considerable amount of FPs and FNs (20%). In both cases, we can separate the 30 TFs to be a positive set and a negative set. The positive set contains regulators having consistent topological information with expression data, and negative set contains regulators having inconsistent or less consistent expression-connection relationships. We compare kSA with several other typical methods to rank condition-specific TFs:

(a) Least-squares regression: for each TF, we directly use biological knowledge $\mathbf{b}_l$ to regress gene expression matrix $\mathbf{X}$. We rank TFs according to residue errors of using biological knowledge of each TF. The smaller the TF residual error, the more relevant this TF to biological condition that expression is measured.

(b) NCA regression: similar to least-squares regression scheme, except that we use estimated regulatory component $\mathbf{a}_l$ to regress gene expression.

(c) Averaged TFA: if we normalized each estimated regulatory component to be unit-standard deviation, we can have TFA average activity directly comparable. The higher the averaged absolute TFA level, the more relevant this TF to biological condition that expression is measured.

From performance comparison curves shown in Fig. 2.6, we can observe that least-square regression always lead to worst performance, it is understandable as there is no fine tune estimation about regulation strength a_{nl}. Averaged TFA is as good as both stability analysis schemes, and outperforms NCA regression approaches in distinguishing 'non-relevant' TFs, shown in Fig. 2.6(a). However, both stability analysis schemes clearly outperform all the other methods in distinguishing 'less-relevant' TFs shown in Fig. 2.6(b), demonstrating that stability analysis is more sensitive to prioritize condition-specific TFs.

Target gene prioritization

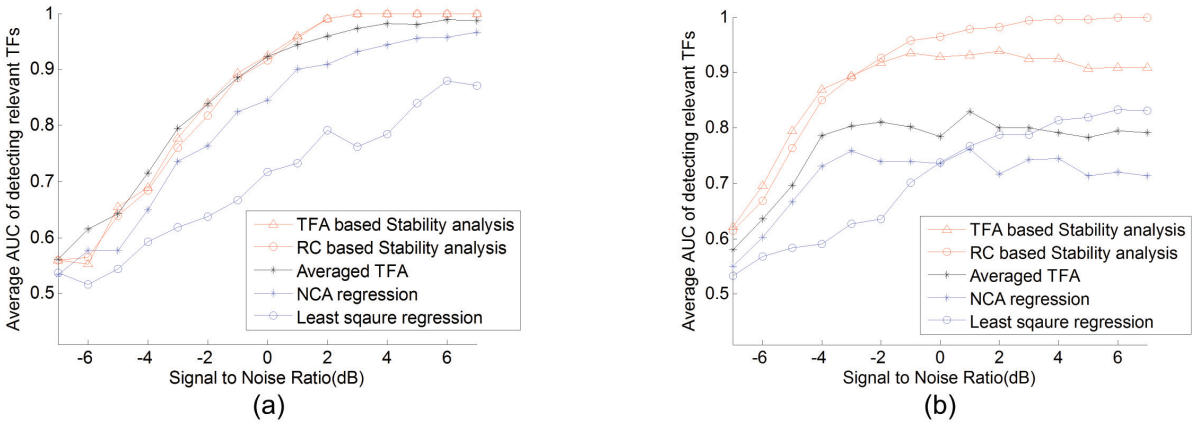


Figure 2.6: Comparison of prioritizing condition-specific TFs, in distinguishing relevant TFs from (a) 'non-relevant' TFs, and (b) 'less-relevant' TFs.

In order to evaluate the performance for target identification, we designed simulations as follows: an initial regulatory network was formed by 30 TFs ($L = 30$) and 500 downstream genes ($N = 500$), and 35 simulated microarray expression data samples ($M = 50$) were generated based on the original topology with varying SNR. For evaluating the prioritization capability of target genes, we keep half of downstream genes (250) having true topological information, while topological information of other half of genes was replaced with randomly generated connections, making topological knowledge of this set of genes inconsistent with expression data. We compare totally three different methods:

(a) Target gene ranking based on regression with estimated TFAs: we use relevance score of downstream gene (RSDG) defined in Eq. (2.19), which is based on average significance level of TFA regression analysis. The higher the RSDG score, the more relevant this downstream gene is.

(b) Target gene ranking based on averaged absolute regulation strength: for n-th gene, we calculate the averaged absolute value of all non-zero regulation strength $\{a_{nl}|a_{nl} \neq 0\}$ pinpointing to this gene. The higher the averaged absolute regulation strength, the more

relevant this downstream gene is.

(c) Target gene ranking based on NCA regression error: for n-th gene, we calculate the residual error regressed by NCA estimates: $\text{err}_{gene}(n) = ||\mathbf{x}_n - \sum_{l=1}^{L} \hat{a}_{nl}\hat{\mathbf{s}}_l||_2^2$. The lower the residual error, the more relevant this downstream gene is.

To test the robustness of each method, we purposely permute certain percentages of gene expression patterns to disrupt the ideal linear model assumptions. We vary the cases from no permutation to 10%, 20% and 50% permutation rate of expression patterns, the performances in terms of AUC are presented in Fig. 2.7(a), (b), (c) and (d), respectively. In every case, the proposed RSDG-based method outperforms other two methods to detect relevant downstream target genes. Specifically, when expression data is not permutated and well corresponds to underlying linear model, both regulation strength based and regression residual error based methods achieved very similar performance; when expression data is permuted slightly (10%), we can observe performance difference between regulation strength based and regression residual error based methods. This difference is enlarged with increasing permutation rate of expression data, suggesting that regression approach may over-fit to the expression data.

2.4.2 Yeast cell cycle experiment

The yeast cell cycle microarray experiment was performed using fluorescently labeled cDNA arrays, measuring the expression levels of 6,178 genes of wild-type S. cerevisiae cells. The cell cycle was synchronized by three independent methods: (1) α-pheromone (α-factor) was used to arrest the cells in G1 phase; (2) a temperature-sensitive mutation cdc15-2 was utilized to arrest cell in mitosis; (3) a temperature-sensitive mutation cdc28 was utilized to arrest cell in mitosis. We set the p-value threshold as 0.01 to integrate ChIP-chip data (Lee, Rinaldi et al. 2002) for initial network information. Although in (Lee, Rinaldi et al. 2002) the threshold was set as 0.001 (a stringent threshold) to better eliminating false-positives, here our motivation was to test the capability of the proposed stability-based

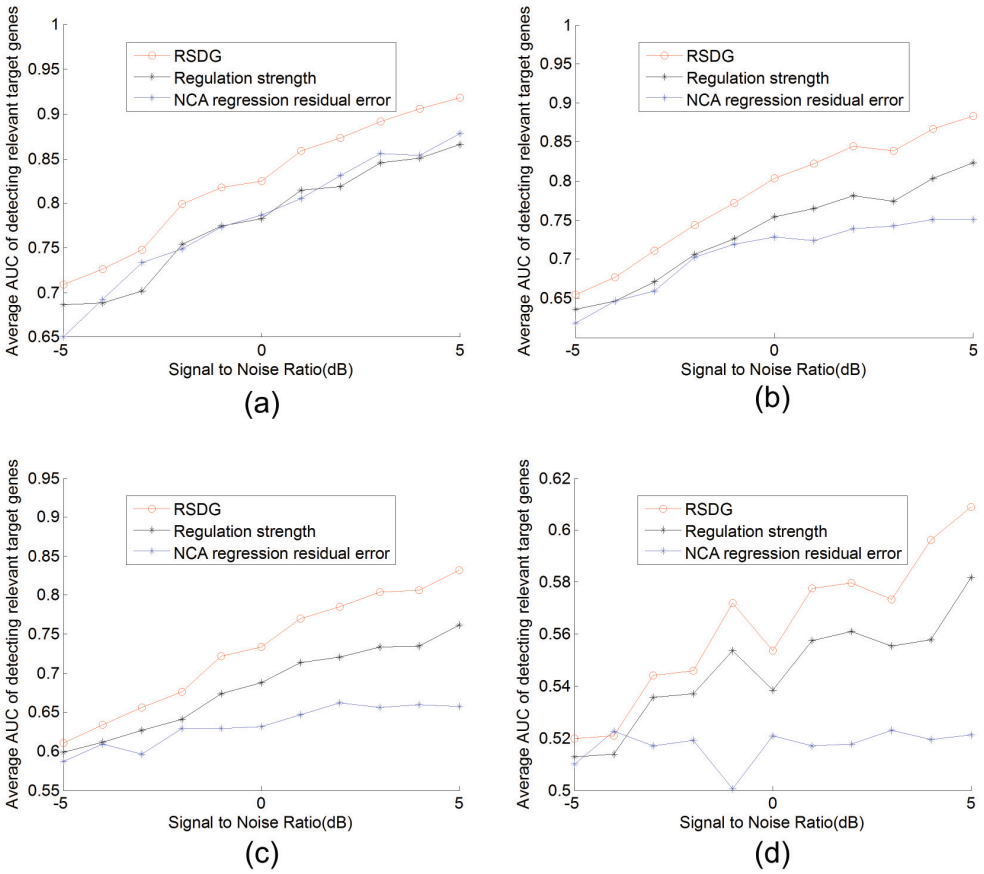


Figure 2.7: Comparisons of target prioritization when expression matrix (a) is not permuted, (b) permuted in 10%, (c) permuted in 20% and (c) permuted in 50%.

scheme for identifying cell cycle related genes even with false positives in knowledge. After preprocessing, the total number of genes is 4,419 and the number of TFs is 111. Within these 4,419 genes, there are 644 cell-cycle regulated genes as biologically validated in (Spellman, Sherlock et al. 1998).

After stability analysis, we extracted top 20 most stable TFs, 18 out of which are cell cycle-related regulators. Specifically, 14 of them are well established cell cycle regulators in different phases, as shown in Fig. 2.8. Based on these 20 TFs, we further identified their condition-specific target genes. With a FDR cut-off of 0.05, we obtained 164 cell cycle-related genes from top 300 stable downstream genes, with a significance p-value of 3.27e-13. The expression pattern of these cell cycle-related target genes is shown in Fig. 2.9. Notice that for the identification of both TF and target genes, unlike in (Spellman, Sherlock et al. 1998) we did not utilize any cell cycle pattern information (i.e., the cycle-like pattern), underlining the usefulness of the proposed stability analysis. It should be emphasized that if the prior pattern information (being biologically meaningful) is available, such as cycle pattern in yeast cell cycle experiment, it will be more informative to evaluate both expression pattern and stability of downstream targets. However, stability analysis is more generally applicable than pattern-based methods, especially for exploratory and discovery studies in which specific pattern knowledge is usually not known.

With the top 20 stable TFs, we show in Fig. 2.10 the estimated TFAs of three representative TFs: SWI4, MBP1 and HSF1. The cycle-like patterns of both SWI4 and MBP1 (TFAs) are well consistent with their cell cycle roles, in all three synchronization conditions. It suggests that these TFs are stably involved in the cell cycle process, and their TFAs thus can be reliably estimated even with a small percentage of topological connection errors. It is also interesting to notice that the estimated TFA of HSF1 only shows very small changes in the -factor synchronized condition, which is much smaller than the activities in CDC15 and CDC28 conditions. Actually, this phenomenon can be explained (at least partially) by the temperature related synchronizations of CDC15 and CDC28 conditions, as HSF1 is a heat shock stress responsive TF. Therefore, a caveat there should be raised that condition-

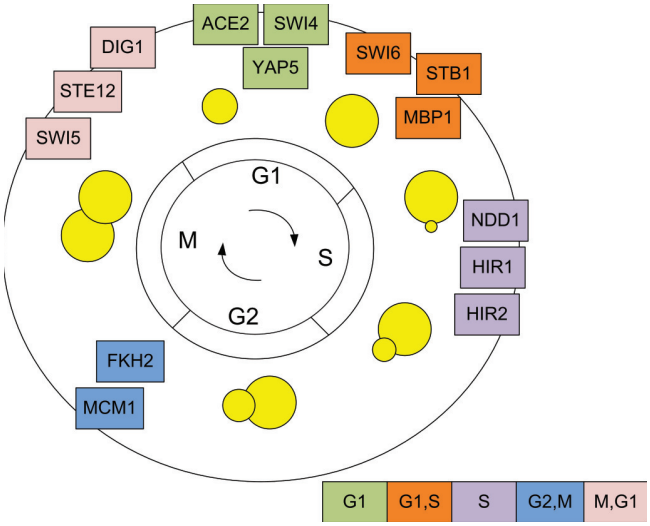


Figure 2.8: The distribution of top 20 TFs identified by stability analysis in different cell cycle phases.

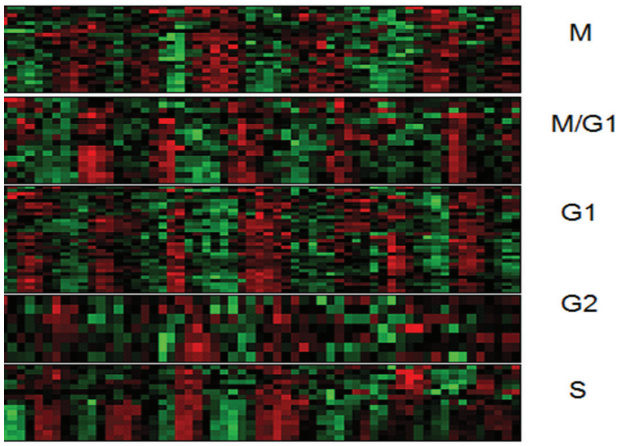


Figure 2.9: Heatmap of cell cycle related genes within top 300 stable targets.

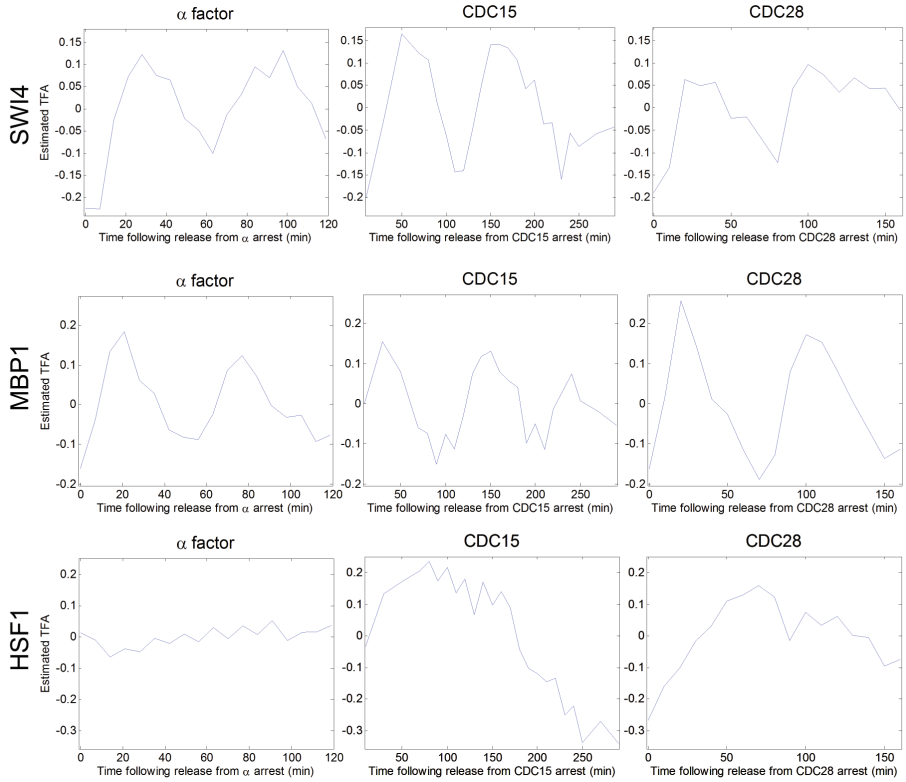


Figure 2.10: Estimated TFAs of SWI4, MBP1 and HSF1.

specificity does not directly suggest that a stable TF/target is relevant of the biological condition of interest, further biological context and interpretation need to be checked with.

With the confidently identified TFs, we further proceeded to evaluate how well each method can identify or highlight target genes. If regarding the 644 genes in (Spellman, Sherlock et al. 1998) having cell-cycle pattern as the true target genes (we should keep in mind that this ground-truth could be incomplete), we can rank target genes according to regression errors and stability scores, respectively. The precision-recall curves of three different methods (RSDG-based, NCA regulation strength-based, NCA regression-based and direct regression-based methods) are shown in Fig. 2.11. As we can see from the figure, RSDG based scheme shows a much improved performance (Area Under PRC curve (AUPC = 0.58)) in identifying condition-specific genes than both NCA regulation strength -based (AUPC =

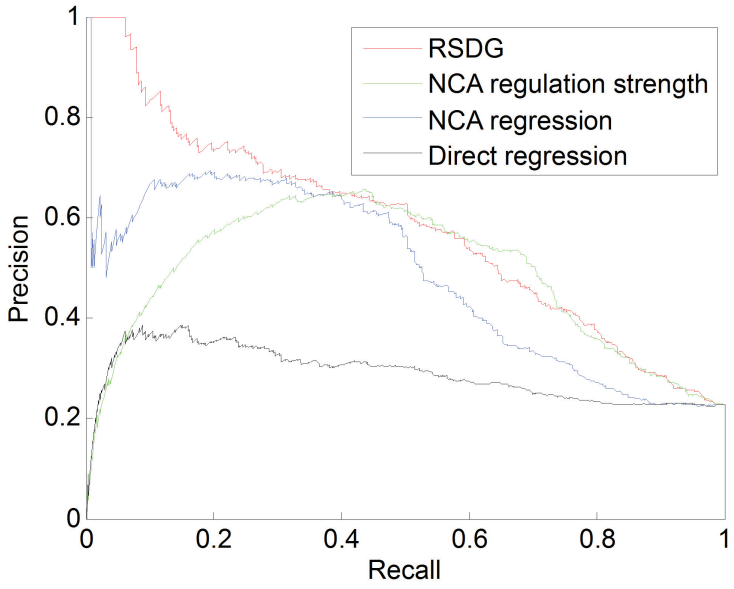


Figure 2.11: Precision-recall curves of different schemes to prioritize cell-cycle related genes.

0.48) NCA regression-based (AUPC = 0.49) and direct regression-based methods (AUPC = 0.29).

With condition-specific TFs and target genes identified by stability analysis, we can have a full picture of cell-cycle related transcriptional regulatory network, shown in Fig. 2.12.

2.4.3 Breast cancer cell line experiments

World widely breast cancer is the most common type of non-skin cancer in women population and one out of eight United States females could suffer from breast cancer in their lifetime. Some of the breast cancer cases, which are sensitive to hormones such as estrogen, can be treated effectively by blocking the effects of corresponding hormones. Unfortunately, one of the hurdles in the therapy is that substantial portions of breast cancer cases do not respond to anti-estrogen treatment, as the tumor could be inherently resistant to the drug or developing the drug resistance through time. Therefore, it is essential to establish under-

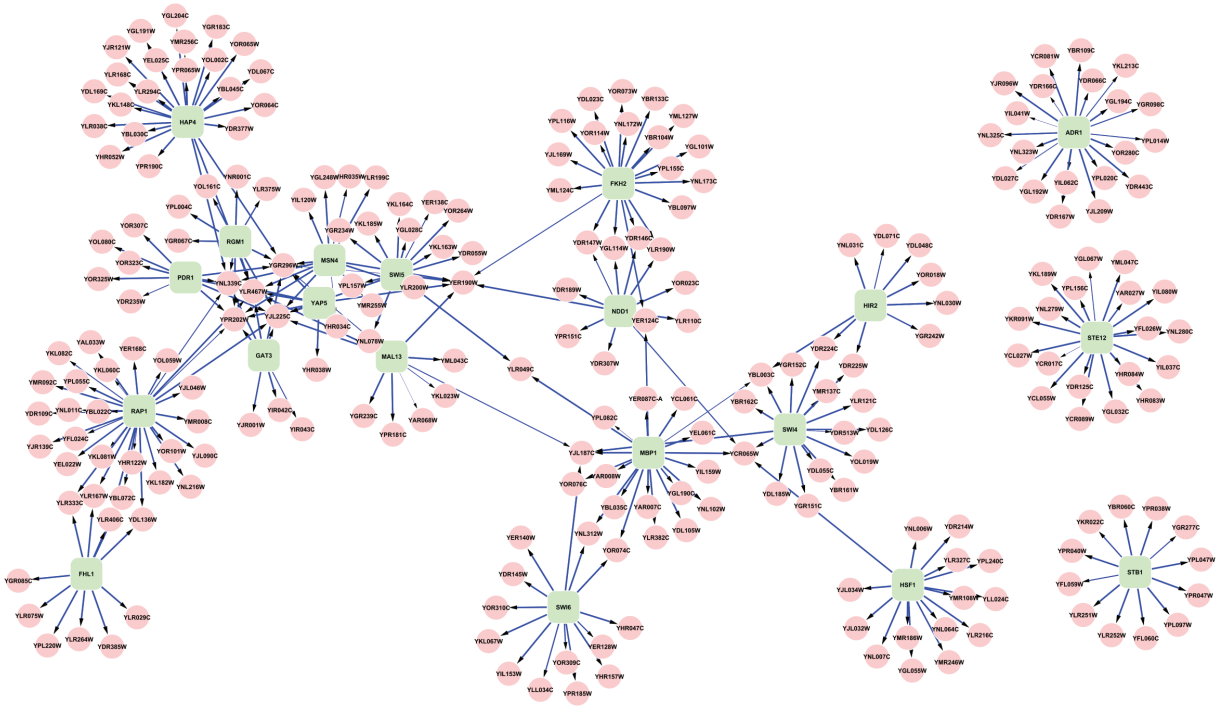


Figure 2.12: Yeast cell cycle regulatory network inference results based on proposed schemes.

standing of estrogen independent progression of breast cancer, identify molecular signature of drug resistance and develop novel treatment scheme.

Several known transcription factors play vital roles in the cancer progression to impact proliferation, differentiation and apoptosis these essential biological processes (Nebert 2002). In addition, TFs could also serve as therapy targets (Libermann and Zerbini 2006). To facilitate better understanding of transcriptional regulations in drug resistance/estrogen independent environment, in the following experiments we aimed to investigate the transcriptional regulatory networks in (a) E2 induction cell line study; (b) E2 deprivation cell line studies. We analyze three breast cancer gene expression data sets:

(1) E2 (17β-estradiol) induced dataset (E2 induced cell lines)

In corresponding experiment (Creighton, Cordero et al. 2006), three estrogen-dependent breast cancer cell lines (MCF-7, T47D and BT-474) were treated with E2(17β-estradiol) from 0 to 24 h, and then profiled for gene expression using Affymetrix GeneChip Arrays. This particular dataset provides us opportunity to investigate E2 dependent transcriptional regulation mechanisms in breast cancer cell lines.

(2) Estrogen deprived dataset (LCC cell lines)

With previously derived series of breast cancer variants (Brunner, et al., 1997) (Clarke, Brunner et al. 1989), which closely reflect clinical phenotypes of endocrine sensitive and resistant tumors, several E2 independent breast cell lines (MCF-7 stripped, M3, LCC1) were profiled. This profiling reflects the tumor progression in estrogen deprived case and therefore shed light on the understanding of regulation in such situations.

(3) Long term estrogen deprived dataset (LTED cell lines)

This dataset is measured based on a long-term estrogen-deprived (LTED) MCF7 cell model, which was made to study acquired resistance in postmenopausal women (Aguilar, Sole et al. 2010). Gene expression data were integrated into the time course of MCF7-LTED adaptation.

Computational results

Notice that here the number of motifs we evaluated is much larger than the available microarray sample number. Therefore, we adopted the scheme we proposed to tackle this underdetermined situations. Perturbation level α was set as 0.02, and number of stability running P and random partitions R were set as 50 and 200. The top motifs associated with most stable TFA estimations are displayed in Table 2.4. In the first E2-induced experiment, several E2F motifs (E2F-02, E2F-Q4-01, E2F-Q3-01) to be ranked in the top places, and it is consistent to the fact that proliferation was activated through E2F family TFs (Prall, Rogan et al. 1998). It is not surprised to see that estrogen receptor binding site (ER-Q6) only appear in this experiment as in the remaining experiments the activity of estrogen is purposely inhibited either through E2 deprivation or antiestorgen drug. As another evidence of estrogen receptor is activated, the co-activator of estrogen receptor CREB is also shown in the E2-induced list with multiple motifs (CREB-Q4, CREBATF-Q6, CREB-01). Interestingly, the binding motif ETS-1B of ETS1 (v-ets erythroblastosis virus E26 oncogene homolog 1) shows up in the stable lists of both E2 deprived case, and breast cancer study has shown that the over-expression of ETS1 is indicative of poor prognostics (Buggy, Maguire et al. 2004). In the E2-deprived experiment for LCC cell lines, in contrast several AP1 motifs (AP1-Q6, AP1-Q6-01, AP1-Q4-01, AP1-Q2-01) appear, and it is well accepted that AP1 is associated with drug resistance (Daschner, Ciolino et al. 1999). This confirms the observation in the previous E2-deprived cell line study. It is also interested to see that SP1 (Porter, et al., 1997), a TF could interact with ESR1, also activated in the corresponding patient group. Distinctly in the late case, MYB, a famous onco-gene related with breast cancer also (Drabsch, Hugo et al. 2007) has multiple stable motifs (MYB-Q6, MYB-Q5-01, MYB-Q3).

Some further discussions on motif analysis results

Motif analysis using kSA partially revealed complicated picture of estrogen-related regulation mechanism. As the signal receiver of estrogen ligand, estrogen receptor (ER) generally refers to a group of receptors that could be activated by E2 (Dahlman-Wright, et al., 2006). There

are two different forms of estrogen receptors ER-α and ER-β, which are encoded by two distinct genes: ESR1 and ESR2, respectively. The receptors could form ER-$\alpha\alpha$/ ER-$\beta\beta$ homodimers or ER-$\alpha\beta$ heterodimers.

In the classical mechanism of ER regulation (Hall, et al., 2001), E2 ligand activates ER complex by changing its conformation so that ER could bind to estrogen responsive elements (EREs) in the promoter regions of ER target genes. DNA-bound ER in turn recruits other co-factors such as CBP-p160 complex to regulate expression of the downstream target gene. In addition to E2 induced activation, it has also been shown that ER function can be initiated by growth factors, such as epidermal growth factor (EGF) and insulin-like growth factor-1 (IGF-1).

Whereas genomic actions of ER could be modulated by E2 or growth factors, ER also plays non-genomic roles without directly binding to ERE DNA sequence. This could occur in both E2-dependent and E2-independent conditions (Hall, et al., 2001; Kushner, et al., 2000). Taking the ER activation at AP-1 sites as an example (Kushner, et al., 2000), it is accepted that ER could activates gene transcription at AP-1 sites through the formation of protein complex with cJun/cFos, without utilizing ERE. The complicated activation mechanism have been summarized as two separate scenarios depending on the status of AF protein domain of ER: (a) ER-$\alpha$with estrogen or with tamoxifen could activate AP-1 though an AF mediated pathway; (b) if AF is absent, ER-α and ER-β could activate AP-1 in the presence of selective estrogen receptor modulators (SERMs).

Since motif analysis could only reveal active response elements on DNAs, the non-genomic regulation mechanism cannot be fully reflected unless more interaction measurements are acquired.

We further categorized the stable downstream targets into different gene families, according to growth factors, TFs, oncogenes and protein kinases (shown in Table 2.5, 2.6, 2.7 and 2.8), which play vitals roles to propagate the regulation in a cascading way. These genes are stably driven by the upstream TFs, which are identified by stability analysis, and they

Table 2.4: Top regulatory motifs ranked by proposed stability analysis scheme, for each of dataset analysis results.

Rank	**E2 induced cell lines**	**LCC cell lines**	**LTED cell lines**
1	CREB-Q4	T3R-Q6	E2F1-Q4
2	YY1-Q6	TAL1ALPHAE47-01	STAT1-03
3	CREBATF-Q6	NFKB-Q6	E2F-Q4-01
4	ATF4-Q2	ATF3-Q6	TAL1-Q6
5	OCT1-06	AP1-Q6	GATA1-03
6	CREBP1-Q2	TAL1BETAE47-01	NFY-Q6-01
7	CREB-01	NFKB-C	OCT1-B
8	ER-Q6	AP1-Q6-01	MAF-Q6-01
9	NFY-C	GATA1-03	P53-01
10	SRY-01	NFKB-Q6-01	E2F1-Q4-01
11	OCT1-Q6	ETS2-B	P53-02
12	E2F-Q3-01	ETS1-B	USF-Q6-01
13	STAT1-02	SRY-01	HNF1-Q6
14	E2F-02	AP1-Q2-01	SP1-Q4-01
15	ETS-Q6	STAT5A-03	YY1-Q6-02
16	E2F-03	STAT1-02	SP1-Q6
17	USF2-Q6	HFH3-01	BACH2-01
18	OCT1-03	HNF3ALPHA-Q6	HNF1-01
19	ETS2-B	AP1-Q4-01	SP1-Q6-01
20	ARNT-02	FOXJ2-01	LHX3-01

Table 2.5: Growth factors within top 300 downstream targets genes for each study.

E2 induced cell lines	**LCC cell lines**	**LTED cell lines**
ARTN	ARTN	AMH
CXCL10	CSF2	ARMET
CXCL12	CXCL10	ARTN
EPO	FGF9	CSPG5
FGF23	GAL	EPO
GAL	IL6ST	FGF9
IL19	NRTN	SEMA3C
IL6ST	TNFSF15	STC2
JAG2		
RABEP1		
SEMA3C		

Table 2.6: Oncogenes within top 300 downstream targets genes for each study.

E2 induced cell lines	**LCC cell lines**	**LTED cell lines**
CCNB1IP1	HMGA1	ARHGAP26
HMGA1	IL6ST	CBFA2T3
IL6ST	MITF	CCNB1IP1
MYB	MYB	RET
RABEP1	TFG	TPR
RARA		TRIM27
TFRC		WHSC1
TPR		
TRIM27		
WHSC1		

Table 2.7: Protein kinases within top 300 downstream targets genes for each study.

E2 induced cell lines	**LCC cell lines**	**LTED cell lines**
BUB1B	BUB1	CAMKK2
CAMKK2	CDK2	CHEK2
CDK2	CHEK1	CSNK2A1
CHEK1	IGF1R	GRK6
LIMK1	MST1R	HSPB8
PBK	SGK3	IGF1R
RAGE	SRPK1	PBK
ROCK2		RET
STK17A		SGK3
STK38L		TEX14

could in turn activate the secondary signaling transduction pathways and transcriptional regulations. As we are particularly focused on the transcriptional regulation mechanisms, we mainly inspected the TFs within the downstream targets (namely downstream TFs) and their potential relationships with upstream TFs. The collaboration of TFs is essential for the effective transcriptional regulation, as TFs need to form TF complex, bind on DNA promoter regions of target genes, and further recruit the RNA polymerase enzymes to perform transcription of specific genes. Take ESR1 as an example, the formation of TF complex can provide it an alternative way to bind on DNA indirectly, which could be explained by one of the drug resistance mechanisms (Bjornstrom and Sjoberg 2005).

We display the potential interactions among TFs in Fig. 2.13 by querying STRING web tool (Jensen, Kuhn et al. 2009), which can identifies interactions based on multiple sources of evidences. We can see that many TFs have multiple interactions, which may suggest collaborations among these TFs. Also, it is noticed that up-stream TFs tend to have higher interaction degree than down-stream TFs, which further confirm their pivotal roles in the

Table 2.8: TFs within top 300 downstream targets genes for each study.

E2 induced cell lines	**LCC cell lines**	**LTED cell lines**
ARID5A	ASCL1	CBFA2T3
ARIH2	ASH2L	E2F3
ATF5	ATF5	E2F5
CNOT4	ATF6	EED
CSDA	CDR2	EGR3
E2F1	CNOT4	EHF
E2F3	E2F3	GTF2A1
E2F5	EED	HIRA
EGR3	EGR3	HMGB1
FHL2	ESR2	HOXC5
HIRA	FHL2	SART3
HMGA1	GATA4	SMARCA1
HMGB2	HMGA1	SNAPC5
HOXC5	HMGB1	TEAD4
ILF2	KIAA0040	TP73
L3MBTL	KLF4	TRIM27
LARP1	LHX6	WHSC1
MAX	MITF	WT1
MSX1	MSX1	YBX1
MYB	MYB	YY1
NRF1	NRIP1	ZNF202
PAXIP1	RB1	ZNF592
RARA	SOX9	ZNF696
RNF24	TRMT1	
SART3	ZNF232	

regulations. Moreover, the TF interactions in E2 induced and E2 deprived conditions show quite different landscapes. In E2 induced case, ESR1 and CREB1 has known binding relationship, as the result of activation of ER signaling pathway. Several E2F family TFs show up in E2 induced case. All the E2F TFs are observed to associate with a stable upstream TF TFDP1, which has alias namely E2F-related transcription factor and can form heterodimerize with E2F proteins to enhance their DNA-binding activity and promote transcription of E2F target genes (Slansky and Farnham 1996). The activation of E2F has known association with increased proliferation activities, which again confirms transcriptional regulations in E2 induced condition promote the growth of cells. Specifically, the interaction between JUN (AP-1) and ESR2 has been shown associated with drug resistance mechanisms in multiple studies, and the cross-talk between AP-1 pathways and estrogen receptor could form the breast cancer alternative pathway (Jakacka, Ito et al. 2001) (Bjornstrom and Sjoberg 2005).

To further investigate the biological relevance of downstream targets in each condition, we performed functional enrichment analysis using David web tool (Huang da, Sherman et al. 2009). Specifically, it is interesting to notice that "cell cycle" (GO: 0007049) is one of significantly enriched biological processes in E2 induced case (FDR = 3.57e-4), while its enrichment significance greatly decreases in E2 deprived case (FDR = 3.34e-1). Furthermore, another biological process "negative regulation of programmed cell death" (GO: 0043069) is not enriched in E2 induced case (FDR = 9.9e-1) but in E2 deprived case (FDR = 1.1e-2). It implies that the cell growth is advanced differently in these two conditions. It is known that tumor cells can acquire resistance to therapy through the activation of anti-apoptotic genes (Igney and Krammer 2002), such as IGF1R, ESR2, ATF5. Insulin-like growth factor 1 receptor (IGF1R), a glycoprotein located on the cell membrane, is regulated by steroid hormones and growth factors and inhibited by anti-estrogen tamoxifen (Huynh, Tetenes et al. 1993).

According to current understanding of estrogen related regulation mechanism (Bjornstrom and Sjoberg 2005), we summarize the schematic plots of transcriptional regulations in different conditions as Fig. 2.14.

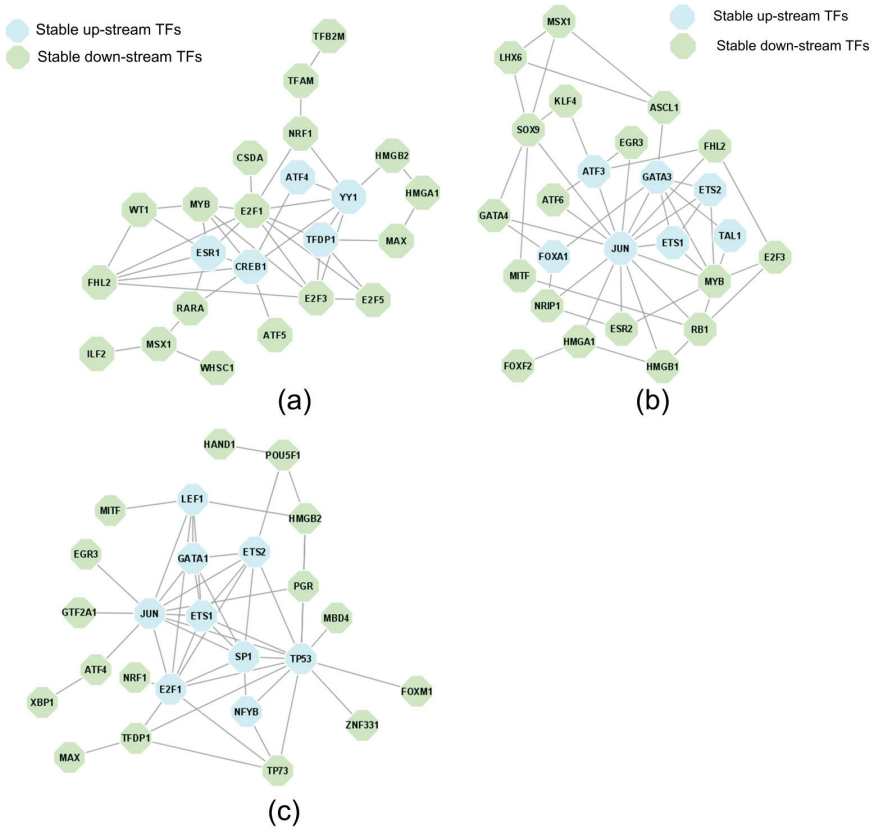


Figure 2.13: Protein interactions of stable TFs, identified by STRING web tool for (a) E2 induced cell lines, (b) E2 deprived LCC cell lines, and (c) E2 LTED cell lines.

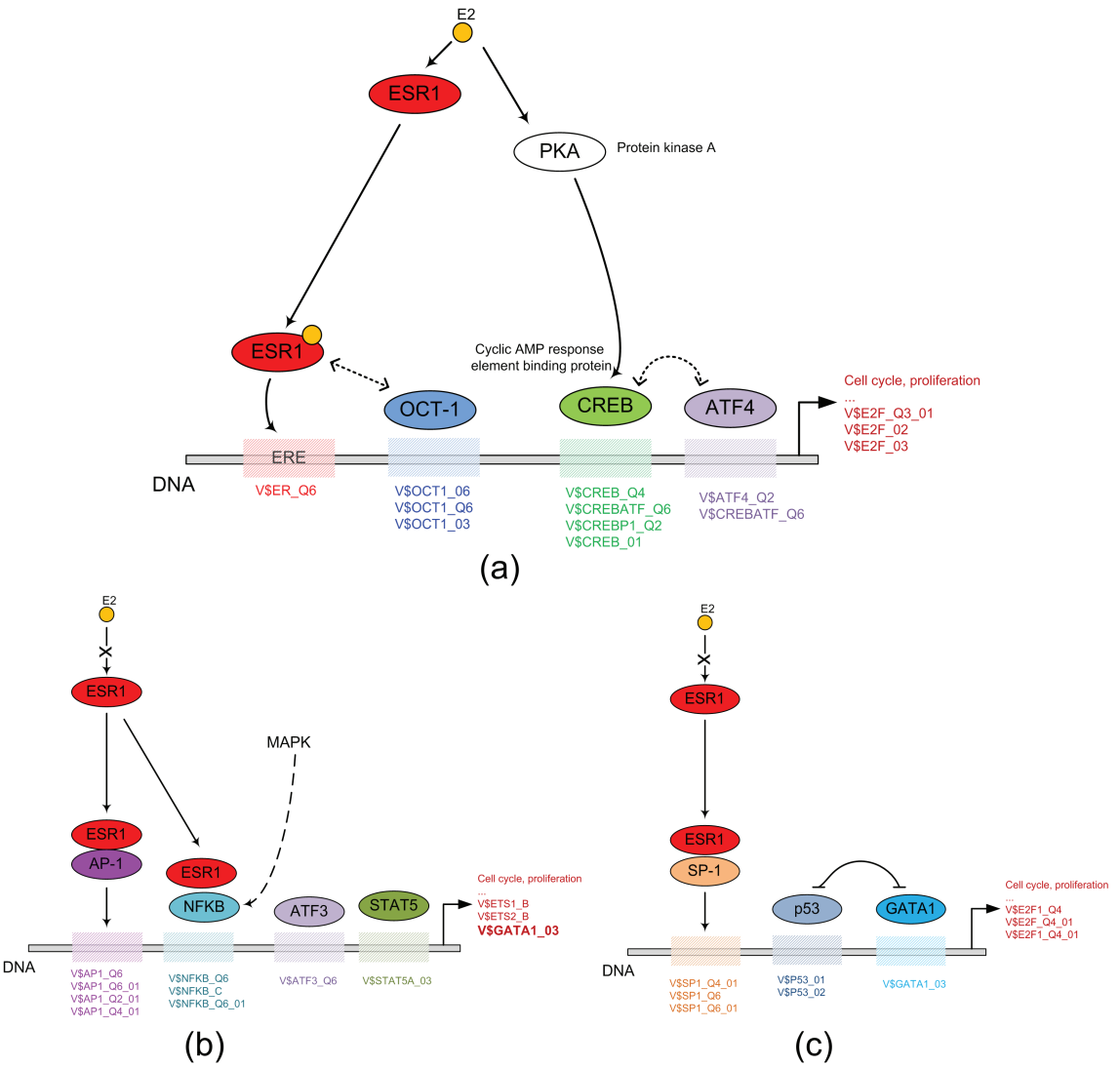


Figure 2.14: Schematic plots of underlying transcriptional regulations highlighted by mNCA and kSA, for (a) E2 induced cell lines, (b) E2 deprived LCC cell lines, and (c) E2 LTED cell lines.

2.4.4 Discussions on stability analysis

As an emerging trend of biomedical research, systems biology requires the advancement of integrated approaches fusing data and knowledge to describe complex properties and interactions within cellular systems, such as various forms of interaction networks, signaling transduction pathways and metabolic reaction pathways. Integrating genomic data with specific biological knowledge, such as interaction information, becomes a practical way to tackle these problems. Noticeably, an increasing amount of biological knowledge is being accumulated and categorized to facilitate the development of this prosperous field, through manual annotation (Camon, Magrane et al. 2004; Vastrik, D'Eustachio et al. 2007), high-throughput data acquisition (Carroll, et al., 2006; Lee, et al., 2002; Shoemaker and Panchenko, 2007), and computational prediction (Kummerfeld and Teichmann, 2006; Shoemaker and Panchenko, 2007). However, along with great abundance and diversity of biological knowledge, the quality of knowledge and its consistency with biological data become serious issues that any computational approach cannot avoid. Staring with unjustified biological knowledge, inferred biological interactions and activities could be less accurate and even misleading. Motivated by that, we proposed a novel computational scheme to filter out irrelevant and less consistent biological knowledge with a stability analysis procedure.

The proposed stability analysis scheme mainly focuses on assessing data-knowledge consistency for regulatory network inference. By purposely adding small perturbations to biological knowledge, we can distinguish reliable inference results from less confident ones, which can be caused by noises, errors and/or data-knowledge inconsistency. The scheme has been applied for prioritization of TFs and target genes. It has been shown that the proposed stability analysis outperforms several conventional approaches with superior robustness for regulatory network identification when given knowledge is incomplete. Note that it is easy to perform the proposed stability analysis and the procedure can be readily plugged into other knowledge integration approaches. Stability analysis of the consistency of data and knowledge could also serve as a testing and evaluation procedures for any knowledge-based integrated

approaches, considering that a practically useful approach should have a reasonable tolerance to certain amount of inaccurate knowledge.

It is worth noting that the proposed scheme not only is related to many previously stability analysis methods for model selection (Tilman, Volker et al. 2004), feature selection (Kalousis, Prados et al. 2007; Křížek, Kittler et al. 2007; Calle and Urrea 2010) and variable selection (Meinshausen and Bhlmann 2010), but also has strong connections with several other statistical or machine learning concepts:

i) Reproducibility: reproducibility is a fundamental requirement for designing any machine learning approaches, such as classifiers (Ruschhaupt, Huber et al. 2004). We expect a robust algorithm should be able to generate similar results even the input data are slightly perturbed or modified, due to noises or errors in measurements. From this perspective, our proposed method is aimed to address the problem that biological knowledge is often contaminated with noises and errors, and then to test the reproducibility of TRN inference by stability analysis.

ii) Variance: any statistical estimator gives the output based on random data samples as an estimate. Any estimate itself is also a random variable. The variance of estimate informs the potential variation range of obtained estimate values. The smaller the variance is, the more confidently we feel about this particular estimate (Delmar, Robin et al. 2005). The variance of estimate is affected by several factors: the function form of an estimator, the number of available data samples and the level of noise. Here, our proposed stability analysis can also be regarded as a way to assess the variance of estimate with respect to unreliable biological knowledge. Assuming that the knowledge-guided approach can provide unbiased estimation results with correct knowledge, stability analysis estimates the variance with respect to knowledge perturbation. Therefore, we will trust the estimated activity or regulatory relationship if associated with stable estimation (i.e., of less variance).

iii) Influence function: in the field of robust statistics (Huber 1981), influence function is used to measure how sensitive the output of estimator is with respect to an input outlier.

If the estimator is a derivable function with respect to the variable of interest, the value of influence function can be directly calculated. However, if the estimator is non-differentiable, very complex or mathematically intractable, stability analysis provides a Monte-Carlo way to calculate the empirical value of influence function. Moreover, influence function is a function defined on the domain of data, while stability analysis is a much more complex scheme defined for evaluating data-knowledge consistency.

The proposed stability analysis approach can be combined with other approaches such as (Chen, Xuan et al. 2010), to further decipher condition-specific TFs and targets. It can also be extended as a complementary technique for statistical significance analysis to assess the relevance of biological knowledge to experimental data. While for significance analysis the null hypothesis and corresponding distribution assumption under null hypothesis should be made, there is no need for the stability analysis-based approach to specify any particular assumption of distribution.

2.5 Regulatory Component Analysis (RCA)

Assumption 4 of the NCA identifiability conditions assumes the biological knowledge - connectivity pattern matrix $\mathbf{B}$ is (a) complete (including all TFs), and (b) accurate (consistent to expression data $\mathbf{X}$). However, in reality biological connection knowledge is often incomplete, especially for high species such as human, where only a few transcription factors can be known in advanced. Sometimes, transcriptional regulation network is even studied according to individual TF. Besides knowledge incompleteness, biological knowledge is also generally inconsistent with gene expression data. Such knowledge-data inconsistency is mainly stemmed from two situations: 1. part of given knowledge is also generated from other biological experiments, which may also introduce errors; 2. knowledge is very general and may not be specific to the biological conditions when expression data are acquired. As the result, biological knowledge usually contains considerable amount of false-positives and

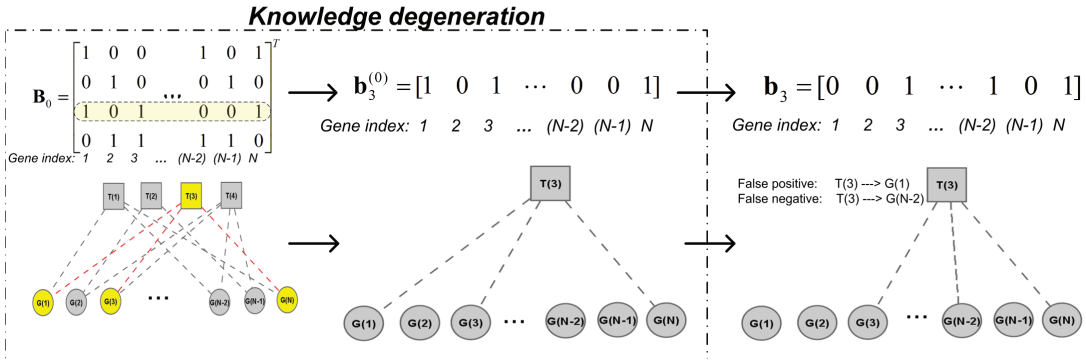


Figure 2.15: Illustration of biological knowledge degeneration. The left arrows indicate incompleteness of biological knowledge, and the arrows in the center false positives and false negatives could contaminate the final knowledge we obtained.

false-negatives, which should not be ignored for computational modeling.

We summarize incompleteness of biological knowledge and its inconsistency with expression data as knowledge degeneration, illustrated in Fig. 2.15. We denote $\mathbf{B}_0 = \left[\mathbf{b}_1^{(0)}, ..., \mathbf{b}_L^{(0)}\right]$, in which $\mathbf{b}_l^{(0)}$represent the true connectivity pattern for l-th TF. In Fig. 2.15, we make up the extreme case when only the knowledge of 3rd TF $\mathbf{b}_3$ is available, however, the given $\mathbf{b}_3$ is still different from true $\mathbf{b}_3^{(0)}$ because of false positives and false negatives relying in biological knowledge.

With the aware of degeneration of given biological knowledge, this section is dedicated to describe the motivation and criterion of proposed regulatory component analysis (RCA).

2.5.1 From matrix decomposition to linear extraction

Following As. 2 of NCA that different TFAs are linearly independent so that matrix $\mathbf{S}$ is invertible, a good regulatory components estimate $\hat{\mathbf{A}}$ can be regarded as a linear projection from expression matrix $\mathbf{X}$:

$$\hat{\mathbf{A}} = \mathbf{X}\mathbf{W}, \tag{2.20}$$

where the projection matrix $\mathbf{W}$ is also called de-mixing matrix in blind separation problem. Ideal $\mathbf{W}$ is the pseudo-inverse of mixing matrix up to a scaling ambiguity:

$$\mathbf{W} = \mathbf{S}^{\dagger}\mathbf{D}. \tag{2.21}$$

In Eq. (2.21), $\dagger$ is the notation for pseudo-inverse operator. While goals of PCA and ICA is to find projection matrix $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_L]$ so that resulting components that are statistically uncorrelated or independent, the purpose of NCA projection matrix is to find source matrix exactly following given connectivity knowledge and minimizing the fitting errors. Instead of matrix decomposition, PCA and ICA solutions can also be achieved in an extraction manner, by maximizing the variance and non-Gaussianity of estimated components, respectively. Extraction is usually implemented through a linear projection:

$$\mathbf{y} = \mathbf{X}\mathbf{w}, \tag{2.22}$$

where a good extraction filter $\mathbf{w} \in \mathbb{R}^M$ should correspond to one row of the de-mixing matrix in Eq. (2.21). When only certain sources are of interest, blind extraction appears to be a more efficient scheme than full blind separation. Typical blind extraction algorithms are designed to recover components of interests by maximizing certain desired characteristics of extracted components,

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} J(\mathbf{y}) = \arg\max_{\mathbf{w}} J(\mathbf{X}\mathbf{w}), \tag{2.23}$$

where function forms of $J(\cdot)$ are generally designed according to properties of underlying source signals, including non-Gaussianity, temporal continuity and etc (Cruces-Alvarez, Cichocki et al. 2004). Linear extraction scheme also avoids the dimension determination problem for latent components, which is difficult to carry on in real world application.

From previous discussions, we can see that an extraction scheme is very attractive for gene regulatory network inference, especially when only partial knowledge is available. However, an extraction scheme is not immediately clear for NCA scheme, as NCA fitting error minimization criterion required all regulatory components to be estimated in parallel. Motivated by limitations of NCA and inspired by extraction schemes originated from ICA, we propose a

linear extraction algorithm for regulatory network inference capable of incorporating partial biological knowledge, in next section.

2.5.2 Formulation of regulatory component analysis

Assume only one column of $\mathbf{B}$ is given, say l-th column $\mathbf{b}_l$, we proposed a scheme to extract corresponding regulatory component. First, according to $\mathbf{b}_l$ we divide column vectors of matrix $\mathbf{X}$ into two non-overlapped sets:

$$\mathbb{X}_+^{(l)} = \{\mathbf{x}_i | b_{il} = 1\} \tag{2.24}$$

and

$$\mathbb{X}_-^{(l)} = \{\mathbf{x}_j | b_{jl} = 0\}. \tag{2.25}$$

Number of members in $\mathbb{X}_+^{(l)}$ and $\mathbb{X}_-^{(l)}$ are denoted as N_+ and N_-, respectively ($N_+ + N_- = N$). Regulatory component analysis is designed to find a linear projection maximizing following cost function:

$$J_0(\mathbf{X}, \mathbf{b}_l, \mathbf{w}) = \frac{\frac{1}{N_+} \sum_{\mathbf{x}_i \in \mathbb{X}_+^{(l)}} (\mathbf{x}_i \mathbf{w})^2}{\frac{1}{N_-} \sum_{\mathbf{x}_j \in \mathbb{X}_-^{(l)}} (\mathbf{x}_j \mathbf{w})^2}. \tag{2.26}$$

The function value of $J_0(\mathbf{X}, \mathbf{b}_l, \mathbf{w})$ has intuitive explanation with data-knowledge consistency, reflecting how well the estimated regulatory component is supported by given biological knowledge. The larger function value of $J_0(.)$, the more consistent estimated component $\mathbf{y} = \mathbf{X}\mathbf{w}$ with given knowledge vector $\mathbf{b}_l$. In the noiseless case where $\Gamma = \mathbf{0}$ and perfect knowledge is given, $J_0 \to \infty$. With function value equals to 1, it suggests that estimated regulatory component is not consistent with biological knowledge, as the averaged controlling strength of potential target genes is the same with of non-target genes.

We further stack the members of each set to form two matrices $\mathbf{X}_+^{(l)}$ and $\mathbf{X}_-^{(l)}$, which corresponds to $\mathbb{X}_1^{(l)}$ and $\mathbb{X}_0^{(l)}$, respectively. The criterion function is rewritten as:

$$J_0(\mathbf{X}, \mathbf{b}_l, \mathbf{w}) = \frac{N_-}{N_+} \frac{\mathbf{w}^T \left(\mathbf{X}_+^{(l)}\right)^T \mathbf{X}_+^{(l)} \mathbf{w}}{\mathbf{w}^T \left(\mathbf{X}_-^{(l)}\right)^T \mathbf{X}_-^{(l)} \mathbf{w}}. \tag{2.27}$$

It has a Raleigh ratio form so that through some manipulations we can achieve following equation:

$$\mathbf{w}^T \left(\mathbf{X}_+^{(l)}\right)^T \mathbf{X}_+^{(l)} = \lambda \mathbf{w}^T \left(\mathbf{X}_-^{(l)}\right)^T \mathbf{X}_-^{(l)}, \tag{2.28}$$

which can be effectively solved using generalized eigenvalue decomposition between $\left(\mathbf{X}_+^{(l)}\right)^T \mathbf{X}_+^{(l)}$ and $\left(\mathbf{X}_-^{(l)}\right)^T \mathbf{X}_-^{(l)}$. Estimated extraction filter $\hat{\mathbf{w}}_{RCA}$ is the eigenvector associated with the maximum generalized eigenvalue of Eq. (2.28).

The proposed RCA criterion has several advantages over traditional NCA approaches (Liao, Boscolo et al. 2003; Boscolo, Sabatti et al. 2005; Chang, Ding et al. 2008):

1. Instead of requiring the complete priori knowledge of all TFs for pursuing a constrained least-square solution, RCA can incorporate incomplete knowledge to estimate individual regulatory component by maximizing on a knowledge-data consistency criterion.

2. Rather then strictly following given biological knowledge, RCA criterion function allows mismatch between estimated regulatory component and biological knowledge. This feature enables the detection of false-positives and false-negatives of biological knowledge, with the information from expression data.

3. Raleigh ratio function form of RCA criterion facilitates efficient optimization using generalized eigen-value decomposition. Moreover, it is convenient to incorporate other regularization items with form $J_r(\mathbf{X}, \mathbf{b}_l, \mathbf{w}) = \frac{\mathbf{w}^T F(\mathbf{X}, \mathbf{b}) \mathbf{w}}{\mathbf{w}^T \left(\mathbf{X}_-^{(l)}\right)^T \mathbf{X}_-^{(l)} \mathbf{w}}$ if extra priori knowledge is known. This is because the extended criterion function $J(\mathbf{w}) = J_0(\mathbf{w}) + \alpha J_r(\mathbf{w})$ can still be efficiently solved using generalized eigenvalue decomposition, where $\alpha$is some trade-off parameter. Notice that generalized eigenvalue decomposition has been widely used in various pattern

recognition applications (De Bie, Cristianini et al. 2005), as well as statistical criterion based blind separation problems (Parra and Sajda 2003). It suggests that the proposed RCA has the potentials to be extended with other priori property function terms, which is a topic under our further investigations.

2.5.3 Simulation studies

Simulation descriptions

Following the characteristics of true regulatory network, connectivity matrix is generated with sparse property. Transcription regulation could be involved with synergistic mechanism (one gene can be regulated through the collaboration of two or more TFs) so that regulatory components are dependent with each other. We generated dependent regulatory component with an average pair-wise correlation around 0.1. To evaluate the impact of biological knowledge to estimation, we consider two simulated scenarios:

(1) Perfect connectivity pattern is given ($\mathbf{B} = \mathbf{B}_0$).

(2) Imperfect connectivity pattern ($\mathbf{B} \neq \mathbf{B}_0$). In order to simulate the real situation where biological knowledge is incomplete and in-consistent, the given $\mathbf{B}$ input to algorithms is generated in two steps: first, only some row vectors of true $\mathbf{B}_0$ are given; second, the given partial $\mathbf{B}_0$ is corrupted with false positives (FPs) and false negatives (FNs).

In each scenario, we comprehensively test the estimation performance of multiple algorithms (PCA, fastICA, JADE, NCA, fastNCA and proposed RCA) under various signal-to-noise-ratio (SNR) conditions, where SNR is defined as follows according to Eq. (2.6):

$$\mathrm{SNR} = 10\log_{10}\frac{\mathrm{Power}_{\mathrm{signal}}}{\mathrm{Power}_{\mathrm{noise}}} = 10\log_{10}\frac{\sum_{n=1}^{N}\sum_{m=1}^{M}(x_{nm}-\gamma_{nm})^2}{\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_{nm}^2}. \tag{2.29}$$

As the regulatory component estimation problem is also equivalent to inference of transcrip-

tional regulatory network, we defined two performance evaluation functions for $\hat{\mathbf{a}}_l$ estimated by each algorithm:

Averaged pair-wise absolute correlation (APAC)

$$APAC = \frac{1}{L'} \sum_{l=1}^{L'} |corr\left(\hat{\mathbf{a}}_l, \mathbf{a}_l\right)| \tag{2.30}$$

and Averaged Area Under precision-recall Curve (AAUC)

$$AAUC = \frac{1}{L'} \sum_{l=1}^{L'} AUC\left(\hat{\mathbf{a}}_l, \mathbf{b}_l^{(0)}\right). \tag{2.31}$$

In Eq. (2.31), $b_l^{(0)}$ is the true biological knowledge of l-th TF, which is l-th row vector of true connectivity pattern matrix $\mathbf{B}^{(0)}$. $AUC(\cdot, \cdot)$ is a function calculate the value of area under precision-recall curve, which describes how well the estimated component can reveal the true target genes of corresponding TF. While APAC has clear implication for signal estimation accuracy, AAUC is more suitable for evaluating biological ground truth when quantitative regulatory component is usually not available.

Regulatory component estimations

(i) PCA and ICA: after regulatory components $\mathbf{y}_l, l = 1, \cdots, L$ are estimated, correspondence relationships need to be established with true components $\mathbf{a}_l$ for performance evaluation. Since NCA, fastNCA and RCA approaches implicitly incorporated with biological knowledge, correspondence is simple: $\widehat{\mathbf{a}}_l = \mathbf{y}_l$. But for PCA and ICA, ordering ambiguities still exist. Therefore, $\mathbf{y}_l$ is designed to correspond to $\widehat{\mathbf{a}}_{l'}$, knowledge vector $\mathbf{b}_{l'}$ of which has the highest similarity with $\mathbf{y}_l$. Two popular ICA algorithms were adopted in simulation studies: JADE (Cardoso and Souloumiac 1993; Cardoso 1999), which is based on algebra criterion to jointly diagonalize a set of higher-order statistics matrices, and fastICA (Hyvärinen 1999), which is based information theory derived criterion to maximize negative-entropy, or the distance with Gaussian distribution.

(ii) NCA and fastNCA: PCA, ICA and RCA allow $\hat{a}_{nl}$ to be arbitrary value even with no biological support $b_{ln} = 0$, while NCA and fastNCA explicitly require $\hat{a}_{nl} = 0, \forall b_{nl} = 0$. Since one of our purposes in simulations is to detect with false knowledge how well the underlying true regulatory component can still be recovered, we made a natural extension for NCA and fastNCA: assuming the non-singularity of mixing matrix $\mathbf{A}$(which is according to As. 2 in section 2.2.2), we use $\hat{\mathbf{A}} = \mathbf{X}\hat{\mathbf{S}}^{\dagger}$ as the estimates for regulatory components, in which $\hat{\mathbf{A}}$ is the estimate of TFA matrix from NCA or fastNCA algorithm. Through this simple transformation, $\hat{a}_{nl}$ could be of any value even for $b_{nl} = 0$. Now all methods can be fairly compared.

Simulation results

a) Biological knowledge is perfectly given ($\mathbf{B} = \mathbf{B}_0$)

To obtain a full spectrum of comparison, we test all the methods under SNR conditions from -1dB to 15dB. For each SNR condition, 50 times of experiments were carried out to calculate the average performance value: transcriptional regulatory network consists of 300 genes regulated by 15 TFs is randomly constructed; based on generated network simulated expression data with 35 samples are produced according to Eq. (2.6). ($M = 35, N = 300, L = 15$). From Fig. 2.16, we can observe that two performance evaluation display quite consistent pictures: in general, RCA and NCA show much better performance than JADE and fastICA these two ICA algorithms, while PCA remains the worst. It is understandable as the implicit utilization of knowledge give the advantages to NCA and RCA. However, it is interesting to notice that fastNCA shows similar performance with both NCA and RCA in high SNR region, but undergoes a dramatic degradation in low SNR region. This is because fastNCA is derived differently from least-squares solution of NCA; instead, it is based on a signal sub-space approach based on the As. 3 of NCA. As the result, the accurate estimation of sub-space is essential for its estimation accuracy. While in the high SNR conditions the sub-space estimation is generally reliable, its performance tends to degrade the performance in low SNR conditions. As a contrast, although matrix decomposition based NCA is more

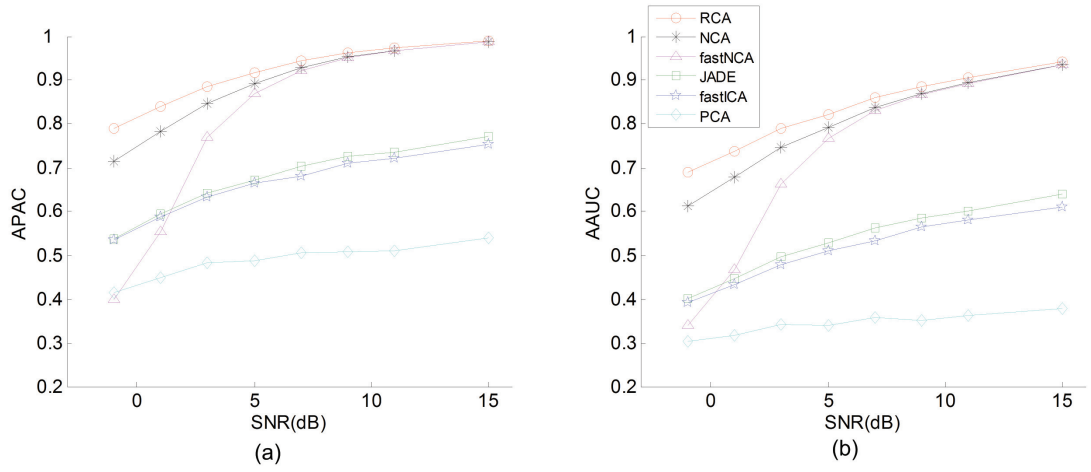


Figure 2.16: Curves of estimation performance for all the methods in scenario 1, where biological knowledge is perfectly given ($\mathbf{B} = \mathbf{B}_0$). (a) corresponds to the performance evaluation in averaged pair-wise absolute correlation (APAC) and (b) corresponds to the performance evaluation in Averaged Area-Under-precision-recall-Curve (APAC).

computationally costive than fastNCA, its performance is much more robust than fastNCA.

b) Biological knowledge is imperfectly given ($\mathbf{B} \neq \mathbf{B}_0$)

Keeping all the other simulation configuration parameters unchanged, we modify the quality of input biological knowledge B. This scenario is designed to evaluate the impact of imperfect biological knowledge to regulatory component estimation by only providing 10 TFs information out of underlying 15 TFs. Moreover, the given knowledge of these 10 TFs are contaminated with moderate FP and FN (FP rate = 1% and FN rate = 10%) to simulate the real biological study. Again, since estimation of regulatory component is equivalent to inference of regulatory network, two performance evaluations present consistent comparison orderings: RCA>NCA>(JADE and fastICA)>(fastNCA and PCA), shown in Fig. 2.17. It is brought to our attention that fastNCA fails miserably, sometimes the performance of which is even worse than PCA. This is because fastNCA heavily depends on As. 3 and As. 4, which are severely violated in this simulation case. Moreover, although least-squares

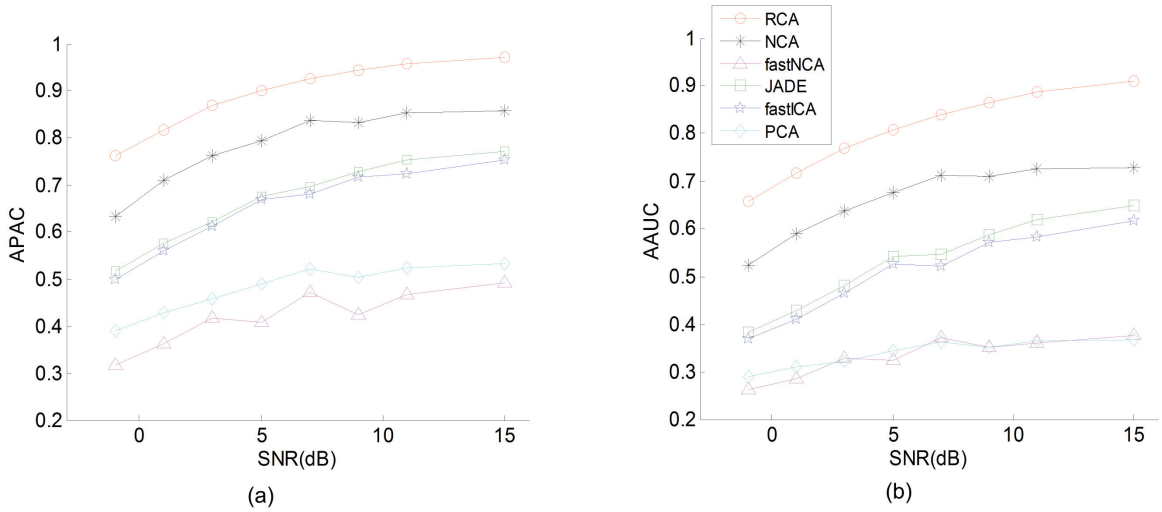


Figure 2.17: Estimation performance curves for all the methods in scenario 2, where biological knowledge is imperfectly given ($\mathbf{B} \neq \mathbf{B}_0$) . (a) corresponds to the performance evaluation in averaged pair-wise absolute correlation (APAC) and (b) corresponds to the performance evaluation in Averaged Area-Under-precision-recall-Curve (APAC).

based NCA remains a relatively robust performance, its performance apparently inferior to our proposed scheme RCA. In both simulations, two ICA algorithms consistently outperform PCA, this is due to the non-Gaussianity property used by ICA is well matched with sparse regulation relationship of regulatory components, even when independence assumption is violated.

We further present a few regulatory component estimation results and corresponding precision-recall curves, show in Fig. 2.18.

Different from the supervised learning utilizing the reference signal, unsupervised learning or blind signal processing is usually based on the statistical assumptions of underlying signals. Since no priori information is needed, unsupervised learning is especially suitable for exploratory data analysis where little biological knowledge is available. However, statistical assumptions such as statistical un-correlatedness and independence are difficult to justify

in the real biological studies (Liao, et al., 2003). NCA and fastNCA can be seen as semi-supervised learning method. This is because biological knowledge is only used to enforce some items of regulatory component matrix **A** without available knowledge support to be zero. Different from NCA and fastNCA, the proposed RCA scheme is a semi-supervised algorithm with relaxed constraint: the regulatory strength a_{nl} could be non-zero even there is no biological knowledge support ($b_{nl} = 0$). Such relaxation enables RCA to reveal false negative target gene and further improve the performance. Different from traditional classification where the label information is correctly given, biological knowledge for network inference is usually un-reliable and could be in-consistent to the biological study. Since the validity of statistical assumption and accuracy of biological knowledge cannot be guaranteed in practice, such comparison of "supervised" learning and "un-supervised" learning with inaccurate prior knowledge provides very meaningful reference for choosing appropriate method in practice.

2.5.4 Real biological experiments

In previous sections, simulation data verify the effectiveness and illustrate superior performance of proposed RCA algorithm. We are also willing to proceed to real biological data analysis. However, the revealing of real transcriptional regulation network for human being is still on-going and many related mechanisms remains unclear. Therefore, we purposely proceed to test all the algorithms on inferring transcriptional regulatory network for Escherichia coli, which is a simple bacterium and has been well studied as model system for various biological studies. We extracted biological knowledge of TFs from a knowledge database named RegulonDB (`http://regulondb.ccg.unam.mx`) with recently updated version 7.0 (Gama-Castro, Salgado et al. 2010). The RegulonDB database contains a collection of TF-target relationships that have been experimentally verified in Escherichia coli. Out of 169 TFs recorded in RegulonDB, we select 30 TFs with at least 15 experimental validated target genes to form initial connectivity pattern matrix, this selection criterion is based on the considerations for reliable precision-recall curve estimation and performance evaluation. The

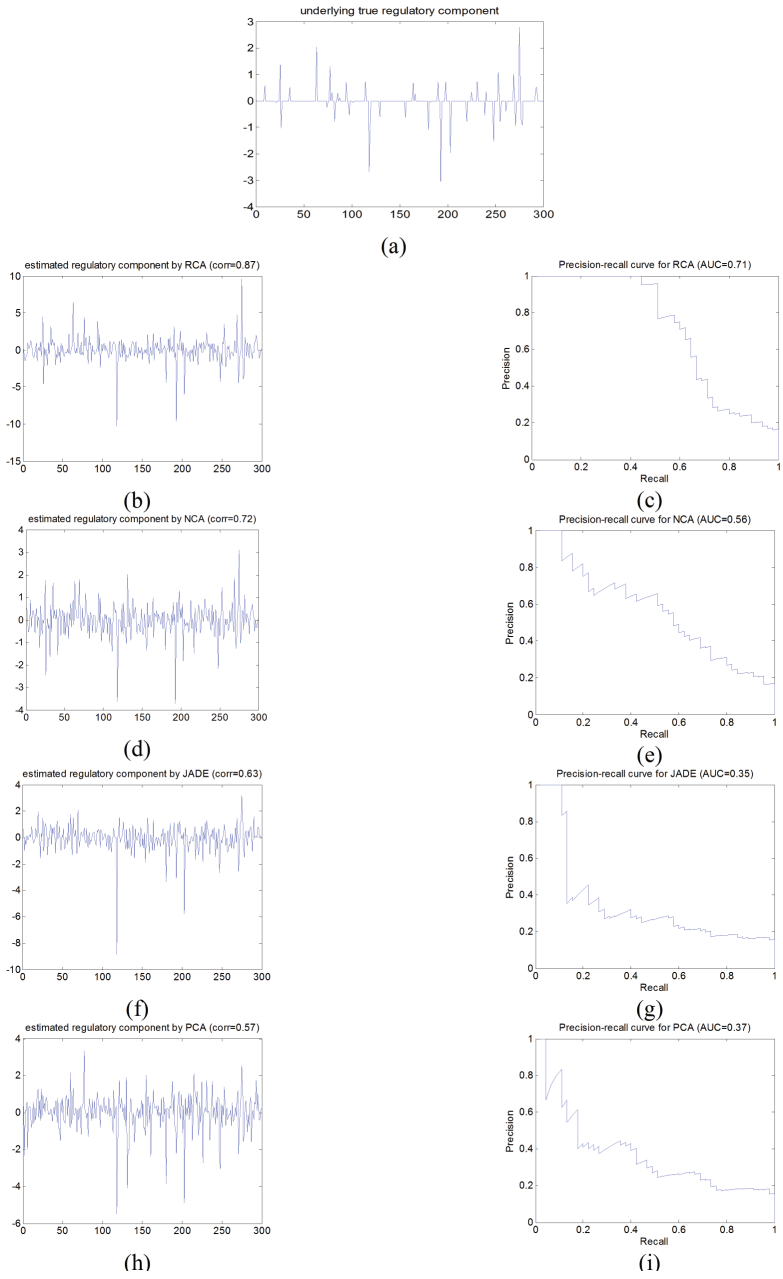


Figure 2.18: Estimated regulatory component profiles and associated precision-recalling curves for retrieving the genes truly affected by corresponding TFs. (a) is the underlying true regulatory component profile; (b), (d), (f) and (h) are estimated regulatory component profiles according to RCA, NCA, JADE and PCA, respectively. (c), (e), (g), and (i) are precision-recalling curves for retrieving the genes truly affected by corresponding TF, according to RCA, NCA, JADE and PCA, respectively.

targets genes of the 30 selected TFs were overlapped with a huge expression compendium (Faith, Hayete et al. 2007), which contains 445 Escherichia coli microarray samples under distinct biological conditions; after all above-mentioned procedures, a network connectivity pattern matrix with 1193 target genes and 30 TFs is obtained. Moderate amount of false positives and false negatives (FP =0.01, FN = 0.1) are added to connectivity pattern matrix. This is aiming to test how well the regulatory components can be estimated with incomplete and inconsistent knowledge.

As there is no quantitative ground truth for true regulatory component, we just used AAUC criterion to evaluate the performance. In addition, it has been observed that AAUC is highly correlated with APAC from our previous simulation studies. Each time we use 100 microarray samples randomly selected from total 445 microarray samples to estimate regulatory components for all the methods. 50 random randomly selections are done to calculate performance evaluation AAUC. Again, RCA significantly outperform all the other methods to retrieve the true target genes regulated by corresponding TFs, shown in Fig. 2.19. To further illustrate the retrieve performance of different methods, we present precision-recall curves for two TF ArgR and LexA as examples, shown in Fig. 2.20.

2.5.5 Discussions on RCA work

Linear latent variable models are widely used in biomedical applications for estimating underlying biological signals, which are corrupted by artifacts or undesired signals. Statistical assumptions such as un-correlatedness and independence are readily accepted in many of these applications such as ECG, EEG and MEG data analysis (Vigario, Sarela et al. 2000). However, when applying these statistical tools to analyze genomic signals with complicated underlying mechanisms, the results become very difficult to interpret. This is caused by many factors: 1. Underlying genomic activities could be dependent to ensure the robustness of biological system; 2. without clear biological context indication, the statistical approach may ignore or blur certain estimation.

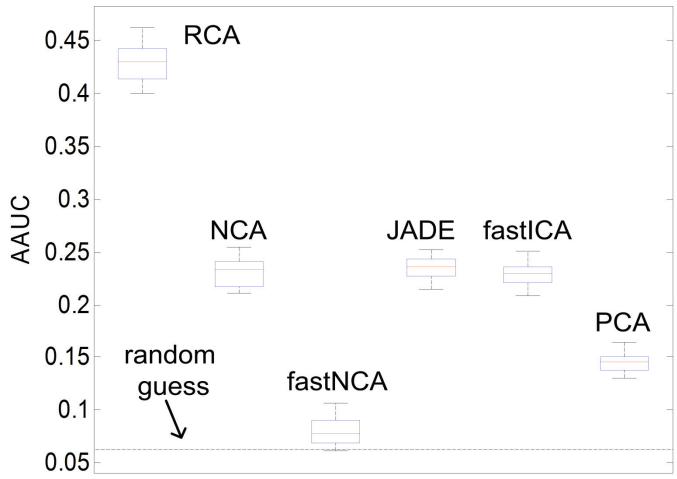


Figure 2.19: Boxplots for Averaged Area-Under precision-recall Curve (AAUC). Where the red-line of each boxplot corresponds to median of all AAUC values, and top and bottom of boxplot corresponds to 75% and 25% Quantile of all AAUC values.

Instead of enforcing strong statistical assumption, NCA incorporated biological knowledge into the solution process of linear latent model for gene expression application, leading to biologically interpretable sources, which we named as regulatory components through this paper. Noticeably, this linear model is also equivalent to a bipartite regulatory network describing the controlling relationships between TFs and genes. However, optimization of NCA is performed based on a biological knowledge constrained least-squares, making its estimation largely depending on available TF-gene binding knowledge, as well as the quality of given knowledge. Unfortunately, the real biological knowledge is generally incomplete and in-consistent to the expression data under study.

With aware of above-mentioned pitfalls in biological knowledge, we proposed a linear extraction based framework named RCA, which explicitly find the linear projection maximizing the coincidence with given partial biological knowledge. The linear extraction scheme also allows RCA to detect FPs and FNs of biological knowledge, which is inconsistent with gene expression data. The contributions of our works are multi-folded: first, transiting from

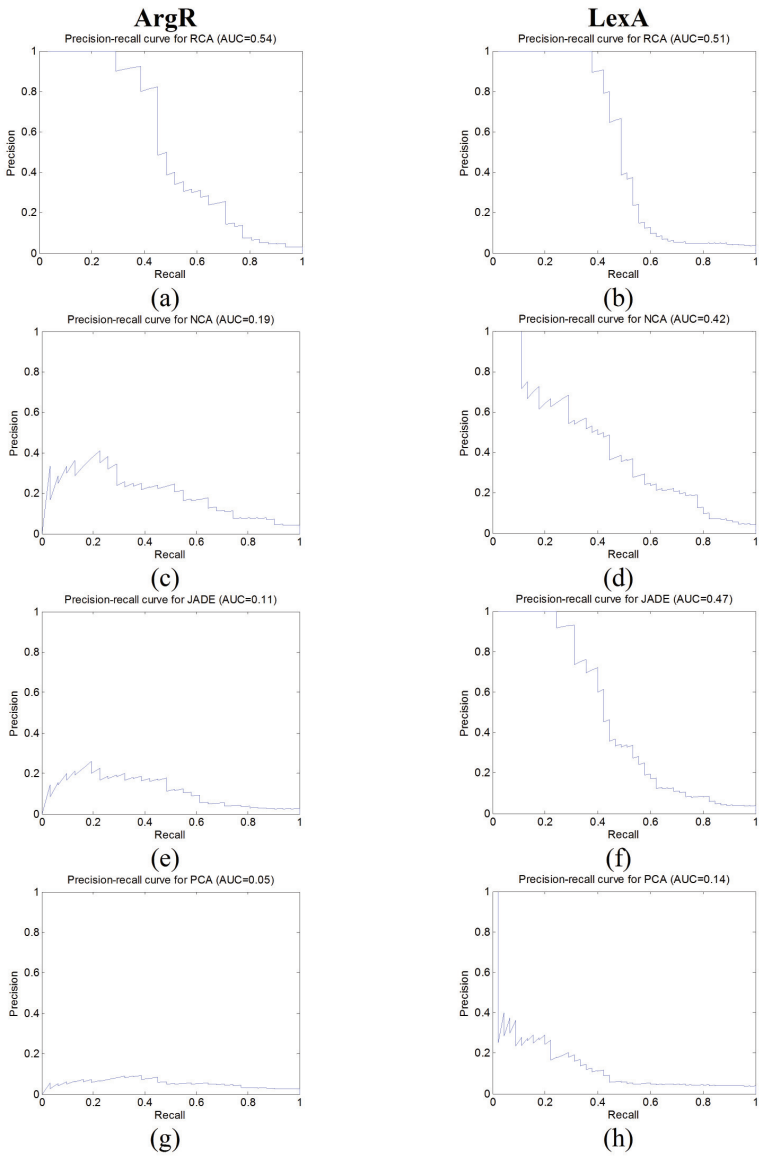


Figure 2.20: Precision-recalling curves for retrieving the genes truly affected by corresponding TFs. (a), (c), (e) and (g) are curves according to TF ArgR, by using RCA, NCA, JADE and PCA, respectively. (b), (d), (f) and (h) are curves according to TF LexA, by using RCA, NCA, JADE and PCA, respectively.

general linear latent model for genomic signals, we review the network inference problem's equivalence with linear latent variable model, which could sever as useful reference for signal processing researchers who are interested in genomic signal processing. Second, for the first time, we formulate a linear extraction scheme for transcriptional regulatory network inference problem, by utilizing incomplete but informative biological knowledge. The proposed scheme show significant performance improvement over traditional NCA methods in both simulations and real biological experiment in E. coli. Thirdly, through designed simulation studies, we showed that how to efficiently integrate biological knowledge is not a trivial problem, considering the given biological knowledge is usually incomplete and in-consistent to the data we have. An inappropriate incorporation of biological knowledge may even lead to worse performance than the methods without using biological knowledge.

Chapter 3

Identification of protein-protein interaction sub-networks

With increasingly accumulated protein interaction data, the identification of condition-specific protein sub-networks emerges as an attractive research problem, solutions of which can facilitate the understanding of molecular mechanisms, and provide reliable sub-network bio-markers for disease diagnosis/prognosis. Most of the existing algorithms mainly search for sub-networks enriched with differentially expressed genes, but overlook their potential interactions and topological importance. In addition, the identification of sub-network is usually solved through optimization schemes, and there is no condition-specific score associated with each gene/protein and each interaction. This makes prioritization of genes and interactions infeasible, and hinders the interpretation of network results. In this dissertation, we propose a novel scheme called Metropolis Random Walk On Graph (MRWOG) to identify the condition-specific sub-networks in a stochastic way. Instead of looking for single sub-network associated with maximum score, we sample multiple sub-networks through a designed random walk on interaction network. We then assemble the sampled sub-networks to form an aggregated sub-network to assess the importance of each individual protein node, which not only reflects its individual association with clinical outcome but also indicates its

topological role (hub, bridge) to connect other important proteins. Moreover, each protein node is associated with a sampling frequency score, which enables the statistical justification of each individual node and the flexible scaling of sub-network results.

3.1 Introduction

Protein, the workhorse of living cell, is involved in every biological process. As the evolution result of cellular systems, proteins usually collaborate with each other to perform biological functions (Hakes, Pinney et al. 2008). Such collaboration enhances the robustness of biological systems, avoiding the collapse of normal functionality when some genetic errors or unexpected environment changes occur (Maslov, Sneppen et al. 2004). The way that one protein physically associates with other protein is called protein-protein interaction (PPI). Thanks to the advanced microarray technology, physical interactions among proteins can now be measured using different experimental techniques, such as yeast two-hybrid (Y2H), tandem affinity purification (TAP) and mass spectroscopy (MS) (Shoemaker and Panchenko 2007). Through these technologies, we can access a rich information source called PPI network, which describes the potential interaction relationships among thousands of proteins in a network point of view.

With the PPI network, researchers have multiple ways to integrate this information according to different computational biology applications: in (Tornow and Mewes 2003), a clustering technique was proposed to dissect the whole PPI network into small groups of genes with functional coherence. PPI was also used to prioritize the cause genes of different diseases (Wu, Jiang et al. 2008). Chuang and his co-workers (Chuang, Lee et al. 2007) showed that PPI sub-networks are better biomarkers than individual proteins to predict metastasis status of breast cancer. In addition, since signal activation of proteins also requires physical interactions, PPI information is utilized to infer signaling transduction pathways, by using various mathematical methods ranging from integer programming (Zhao, Wang et al. 2008),

random coloring coding algorithm (Jung, Makeig et al. 2000), to network information flow approach (Yeger-Lotem, Riva et al. 2009). Besides all abovementioned applications, another very important focal point of utilizing PPI information is to identify condition-specific protein sub-networks with significant changes (Ideker, Ozier et al. 2002; Dittrich, Klau et al. 2008; Qiu, Zhang et al. 2010), which is the focus of this dissertation research. The sub-network analysis helps researchers break down the entire protein interaction network into small parts, and locate the abnormal local regions, which could reveal the disruption of cellular systems or the dys-regulated pathways of disease (Liu, Liberzon et al. 2007). Among the existing methods, Trey Ideker et al. (Ideker, Ozier et al. 2002) proposed to use a simulated annealing method to find sub-network with the largest conditional relevance. A prize-collecting tree-based cost function was designed in (Dittrich, Klau et al. 2008) and the optimization was obtained through mathematical programming techniques. With the edge scores derived from co-expression relationships, the problem was also solved using support vector regression with diffusion kernels (Qiu, Zhang et al. 2010).

Despite that the above-mentioned methods have been successfully applied to some biological studies, one important assumption commonly made in the approaches limits their further applications: that is, all the protein interactions are assumed to be equally reliable and the uncertainty is ignored. In reality, even protein interactions collected from biological experiments may also contain considerable amount of false positives and false negatives. Moreover, almost all the existing methods regard the sub-network identification problem as a '0 or 1' combinatorial search problem, and provide the search results with a bunch of proteins without any weights associated with. In specific biological studies, biologists are interested in not only sub-network constitution but also the relative importance of each protein in the sub-network, through which specific biological experiments can be designed to verify a few protein markers and their functionality. Last but not the least, even when all sub-network members are found, the edges between the members are not quantified or highlighted with their confidence levels. It is essential to assign each individual interaction with some continuous score so that biologists can prioritize different hypotheses regarding

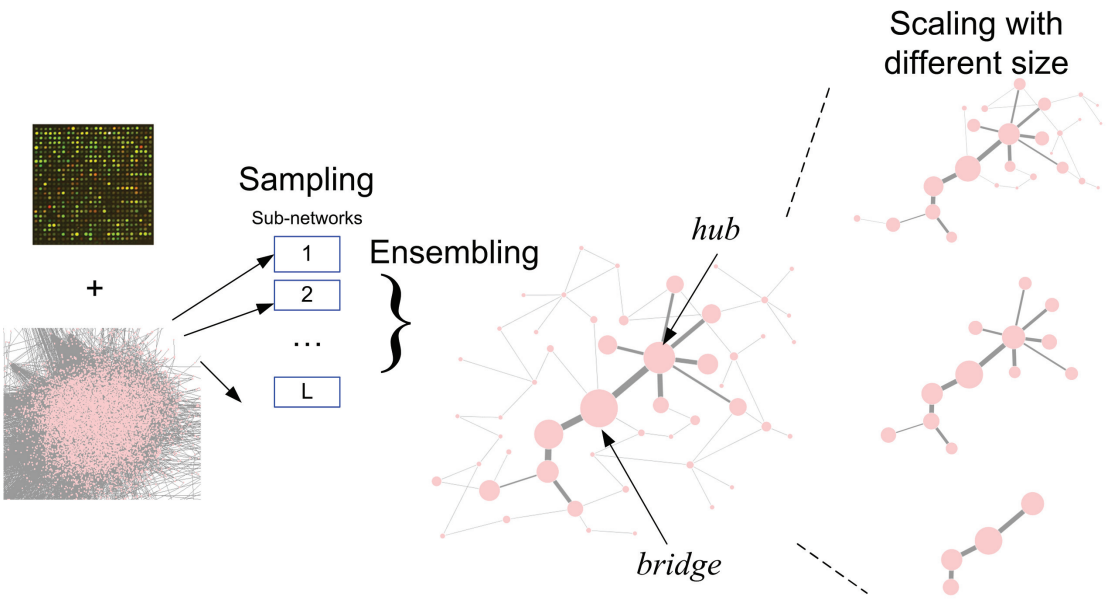


Figure 3.1: Graphic illustration of the proposed scheme. It starts from integrating gene expression and protein-protein interaction information, sampling multiple sub-networks, generating an ensemble of sub-networks to identify highly dys-regulated local regions with node and edge scores, and finally multiple scale results can be obtained according to the scores.

the signaling pathways of interest. In the following sections, we propose a novel method called Metropolis Random Walk On Graph (MRWOG) to overcome the above-mentioned limitations. A graphic illustration of the proposed scheme is shown in Fig. 3.1.

3.2 Existing Methods

3.2.1 PPI network and protein node score

Throughout this dissertation gene and protein is used exchangeable, as we assume that protein activity is approximately proportional to gene expression level. We use a graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ to represent given interaction network, such as PPI network. Vertex set $\mathbb{G} =$

$(\mathbb{V}, \mathbb{E})$ contains N proteins; edge set $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ comprises K physical interactions between proteins. In general, PPI is very sparse so that $K \ll N^2$. Given the expression measurement $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]^T \in \mathbb{R}^{N \times M}$ consisting mRNA levels of N genes across M microarray samples associated with some phenotype information vector $\mathbf{c} \in \mathbb{R}^M$, we can calculate the association score of each gene node through some function $\mathcal{A}(\cdot, \cdot)$:

$$z_n = \mathcal{A}(\mathbf{x}_n, \mathbf{c}), \tag{3.1}$$

in which, z_n reflects expression differentiation of n-th gene between two phenotypes (c_m= 0, 1) , or association between expression pattern of n-th gene and clinical trait ($c_m$is continuous).

Protein node score

Having expression data matrix $\mathbf{X} \in \mathbb{R}^{N \times M}$ and phenotype information of each microarray sample $\mathbf{c} \in \mathbb{R}^M$, statistical analysis is usually used to estimate an association score describing relevance of individual gene pattern to the given phenotype information. There are two cases: 1) Each element of $\mathbf{c}$ is of discrete value. For examples, to analyze gene knock experiments in the yeast (Ideker, Ozier et al. 2002), discrete genotype is used to construct phenotype information $\mathbf{c}$; in breast cancer metastasis study (Chuang, Lee et al. 2007), good or bad clinical outcome is also of discrete value. Mathematically, two biological phenotypes or clinical conditions can be defined as two index sets:

$$\mathbb{I}_1 = \{i | c_i = 0\} \text{ and } \mathbb{I}_2 = \{j | c_j = 1\} \tag{3.2}$$

so that p-value is usually calculated based on two-sample t-test:

$$p_n = \text{p-value}(n) = \text{t-test}(\{x_{n,i}\}_{i \in \mathbb{I}_1}, \{x_{n,j}\}_{j \in \mathbb{I}_2}). \tag{3.3}$$

2) Each element of $\mathbf{c}$ is of continuous value. For example, in (Dittrich, Klau et al. 2008), the phenotype information is continuous trait of clinical survival time. If given $\mathbf{c}$ is survival time vector of patients, p-value can also be obtained through cox-regression to assess how well the pattern of n-th gene can be used to determine survival time:

$$p_n = \text{p-value}(n) = \text{cox-regression}(\mathbf{x}_n, \mathbf{c}). \tag{3.4}$$

Having the estimated p_n from either case, a z-score transformation was proposed in (Ideker, Ozier et al. 2002):

$$z_n = \Phi^{-1}(1 - p_n), \tag{3.5}$$

where $\Phi(\cdot)$ is the CDF function of normal distribution. This z-score transformation was proposed to facilitate the statistical significance evaluation of identified sub-network, since the sub-network z-score of random case ($p_n \sim Unif(0, 1)$) follows normal distribution. Therefore, the association function $\mathcal{A}(\cdot)$ can be seen as the combination of z-score transformation and p-value calculation:

$$z_n = \mathcal{A}(\mathbf{x}_n, \mathbf{c}) = \Phi^{-1}\left(1 - p_n(\mathbf{x}_n, \mathbf{c})\right). \tag{3.6}$$

Remarks:

1. Noticeably, we can also combine multiple phenotype information $\{\mathbf{c}_1, \cdots, \mathbf{c}_O\}$ if available using order statistics (Ideker, Ozier et al. 2002; Dittrich, Klau et al. 2008). 2. A bias correction procedure is recommended in (Ideker, Ozier et al. 2002) to ensure sub-network scores with different sub-network sizes are directly comparable.

3.2.2 Optimization based approaches for identifying dys-regulated sub-networks

Protein interaction information is known to be very noisy (Bader, Chaudhuri et al. 2004), and inconsistent results may be obtained by different techniques (Blow 2009). Even if the interaction is measured through experiments, it still may not be specific to the certain biological condition we are interested. So it is meaningful to assign each interaction edge a score to reflect its condition specificity, rather than treating them equally. There are multiple ways to do so, for examples, in (Jung, Makeig et al. 2000) multiple evidences were

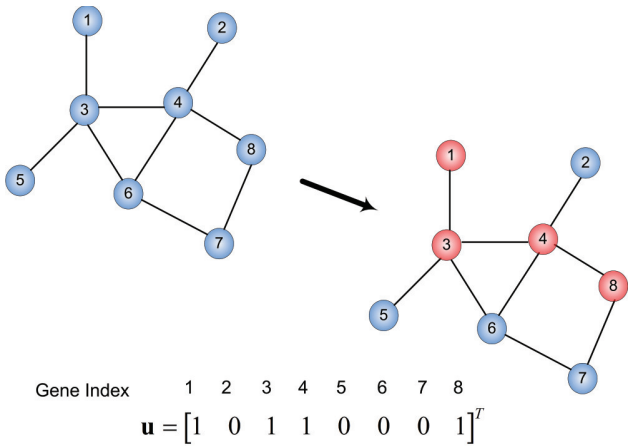


Figure 3.2: Illustration of representing a sub-network as a corresponding binary selection vector. A binary vector $\mathbf{u}$ is used to represent a sub-network, with 1 indicating the selection of gene node with corresponding index.

combined to calculate edge score, in (Zhao, Wang et al. 2008; Qiu, Zhang et al. 2010), gene-gene co-expression evidence was used to compute edge score. Once having edge score w_{ij} for each edge between protein i and$j$, we can define the sub-network score together with previously defined vertex score to reflect the condition-specific relevance/importance of given sub-network. We use a binary vector $\mathbf{u}$ of gene size to indicate the selection of corresponding vertices, depicted in Fig. 3.2. Similarly, another binary vector $\mathbf{y} \in (0,1)^K$ is used to indicate the selection of corresponding edges.

The sub-network score with vertex score only is defined as

$$\mathcal{S}(\mathbf{u};\mathbf{z}) = \sum_{n=1}^{N} z_n u_n / \sqrt{\sum_{n=1}^{N} u_n} \quad . \tag{3.7}$$

We can also incorporate edge information into the scoring:

$$\mathcal{S}_e(\mathbf{u},\mathbf{y};\mathbf{z},\mathbf{w}) = \left(\sum_{n=1}^{N} z_n u_n + \lambda \sum_{k=1}^{K} w_k y_k \right) / \sqrt{\sum_{i=1}^{N} u_i} \ , \tag{3.8}$$

in which, λ is the trade-off parameter controlling the balance between gene differentiation

and edge confidence.

Optimization based sub-network identification

Given the vertex based sub-network score defined in Eq. (3.7), the conventional algorithms purse the solution through optimization constrained by PPI topology:

$$\mathbf{u}_{opt} = \arg\max_{\mathbf{u}} \ \mathcal{S}(\mathbf{u}; \mathbf{z}), \tag{3.9}$$

$$\text{s.t. Selected nodes } (u_n = 1) \text{ are connected through } \mathbb{E}.$$

If edge weight information $\mathbf{w}$ is available, we can further incorporate it into the sub-network identification problem with most likely nodes and edges:

$$(\mathbf{u}_{opt}, \mathbf{y}_{opt}) = \arg\max_{(\mathbf{u},\mathbf{y})} \mathcal{S}(\mathbf{u}, \mathbf{y}; \mathbf{z}, \mathbf{w}), \tag{3.10}$$

$$\text{s.t. Selected nodes } (u_n = 1) \text{ and selected edges } (y_k = 1) \text{ are connected through } \mathbb{E}.$$

Since it is straightforward to add edge score into sub-network score function, using correlation or dependence values, we limit our discussion to be only node score associated for clarity. Noticeably, the sub-network identification problem in (3.9) was proved to be a NP-hard problem (Ideker, Ozier et al. 2002) so that only heuristic approaches can be applied to find sub-optimal solutions. Some algorithmic effort has been made to convert the original problem to be a prize collecting tree construction problem, and exact binary solution can be achieved using integer linear programming (Dittrich, Klau et al. 2008). However, the solutions provided by all the conventional methods are binary, not only insufficient to prioritize individual gene or interaction, but also difficult to adjust the sub-network size smoothly. More importantly, considering the inherent noises and errors in microarray expression data and protein interaction structure, as well as the complexity of cellular system, numerical optimal solution may not fully reflect the underlying biological mechanisms. For two distinct sub-optimal solutions with very similar scores, optimization based methods only prefer the solution with larger score, and overlook the other one, which may also convey important biological implications. Even heuristic average can be applied to ensemble multiple

sub-optimal solutions, it still lacks statistical justifications. Motivated by above-mentioned concerns, we formulate the sub-network identification problem differently as a probabilistic inference problem.

3.3 Stochastic Exploration of Interaction Network

3.3.1 Probabilistic formulation of sub-network identification problem

According to optimization according to network score function in (3.9), we define a non-negative score according to network constraint:

$$\mathcal{L}_0(\mathbf{u};\mathbf{z}) = \max(0, \mathcal{S}(\mathbf{z},\mathbf{u})\mathcal{C}_{\mathbb{E}}(\mathbf{u})) \tag{3.11}$$

and its shape could be adjusted according to some non-negative constant β

$$\mathcal{L}(\mathbf{u};\mathbf{z}) = \mathcal{L}_0(\mathbf{u};\mathbf{z})^{\beta}, \tag{3.12}$$

where $\mathcal{C}_{\mathbb{E}}(\mathbf{u})$ is connection check for vertex selection vector $\mathcal{C}_{\mathbb{E}}(\mathbf{u})$:

$$\mathcal{C}_{\mathbb{E}}(\mathbf{u}) = \begin{cases} 1, & \text{all the vertices associated with} = 1 \text{ connected through } \mathbb{E} \\ 0, & \text{else} \end{cases} \tag{3.13}$$

Following the definition in Eq. (3.12), sub-network $\mathcal{C}_{\mathbb{E}}(\mathbf{u})$ is associated with a likelihood score $\mathcal{L}(\mathbf{u};\mathbf{z})$ assessing how likely this sub-network contributes to phenotype/clinical difference. This likelihood function corresponds to some underlying conditional probability $\mathcal{L}(\mathbf{u};\mathbf{z}) \sim \Pr(\mathbf{z}|U=\mathbf{u})$.

Without loss of generality, we can define an energy function $Energy(\mathbf{u}) = \log \frac{1}{\mathcal{L}(\mathbf{u};\mathbf{z})}$ so that we will have higher probability in lower energy (higher $\mathcal{L}(\mathbf{u};\mathbf{z})$ value). This is similar to the probability definition in thermodynamics of physics, and $\mathbf{u}$ can be imagined as the position

(state) vector of a particle in high dimensional space. Due to thermodynamics, particles wander around in entire space randomly. The probability that particles are found in a position (state) with low potential energy is higher than the probability in high potential energy. Formally, it could be linked with the Boltzmann distribution, which is also known as Gibbs measure. Boltzmann distribution is a certain probability measure for the states of a system. For a system with N particles, the Boltzmann distribution is defined as:

$$\Pr(Energy_i) = \frac{N_i}{N} = \frac{g_i e^{-\frac{1}{k_B T} Energy_i}}{V(T)},$$

which reflects the portion of particles (N_i out of N) occupying a set of states i with energy $Energy_i$. k_B is Boltzmann constant, g_i is the degeneracy, T is temperature of the system and $V(T)$ is the partition function. Therefore, we can also write down the probability of proposed MRWOG scheme according to Boltzmann distribution form:

$$\Pr(U = \mathbf{u}) = \frac{\mathcal{L}(\mathbf{u}; \mathbf{z})}{V(\beta)} = \frac{\mathrm{e}^{-\beta \log \frac{1}{\mathcal{L}_0(\mathbf{u};\mathbf{z})}}}{V(\beta)} = \frac{\mathrm{e}^{-\frac{1}{T} \log \frac{1}{\mathcal{L}_0(\mathbf{u};\mathbf{z})}}}{V(T)},$$

where $\beta = \frac{1}{T}$ is the inverse of temperature T. Whereas the stochasticity of one thermodynamic system is mainly caused by the temperature and particle dynamics in the system, the uncertainty of network arises from the noisy measurements of expression data, dynamics of interaction network and inconsistency biological knowledge and data. Therefore, a sub-network with high network score may not ensure its aberrance, rather to suggest a higher likelihood that this sub-network undergoing certain degree of aberrance.

Importantly, if the energy function $Energy(.)$ can be written as a sum of parts, the Boltzmann distribution (Gibbs measure) has the Markov property so that a joint distribution/likelihood of random variables can be greatly simplified (Kindermann, 1980). Such energy function with Markov random field (MRF) property is preferred for its computational convenience: the likelihood function is decomposed as the product of different orders of cliques, so that a high-dimensional joint distribution can be approximated up to certain orders. The MRF approximation is usually motivated by enabling a simple optimization scheme for ML/MAP

like point estimation. If the sampling technique is deemed to used, the function form of potential energy function could be relaxed to be any form in order to accommodate higher-order and complex interactions. Notice that even though ML/MAP solution can be achieved through optimization of simplified/approximated MRF energy function, it is generally intractable to evaluate the probability and likelihood of underlying particles. This is because the computation of partition factor $V(T)$ requires the integration of high dimensional distribution.

From stochastic point of view, the typical optimization approach is just like a ML (maximum likelihood) or MAP (maximum a posterior) estimator. If multiple independent observations $\{\mathbf{z}_r\}_{r=1,\cdots,R}$ are available, we can combine all of them to obtain more accurate ML estimate:

$$\widehat{\mathbf{u}}_{\mathrm{ML}}(\mathbf{z}) = \arg\max_{\mathbf{u}} \quad \prod_{r=1}^{R} \Pr(\mathbf{z}_r|\mathbf{u}) = \arg\max_{\mathbf{u}} \quad \prod_{r=1}^{R} \mathcal{L}(\mathbf{u};\mathbf{z}_r). \tag{3.14}$$

If kernel function $g(\mathbf{u})$ of priori distribution $\Pr(U = \mathbf{u})$ is available, we can further use maximum a posterior (MAP) estimator:

$$\widehat{\mathbf{u}}_{\mathrm{MAP}}(\mathbf{z}) = \arg\max_{\mathbf{u}} \quad \mathcal{L}(\mathbf{u};\mathbf{z})g(\mathbf{u}). \tag{3.15}$$

Once the uncertainty within expression data and network topology is not negligible, which is usually the case in real biological application, we are more interested to an ensemble solution of multiple sub-optimal solutions accommodating the uncertainty. Therefore, a Bayesian mean estimator minimizing mean square error will be better of our interest than ML or MAP point estimates:

$$\mathbf{f}_{\mathbf{u}} = \arg\min_{\mathbf{f}} \mathrm{E}\left[(\mathbf{f} - U)^2\right]. \tag{3.16}$$

where $\mathbf{f}_{\mathbf{u}}$ is a vector of continuous values, indicating how frequently one node is selected into a condition-specific sub-network. The solution of (3.15) is simply a Bayesian mean estimate:

$$\mathbf{f}_{\mathbf{u},BM} = \hat{\mathbf{u}}_{BM} = \mathrm{E}[U|\mathbf{z}] = \sum_{\mathbf{u}\in\mathbb{U}} \mathbf{u} \Pr(U = \mathbf{u}|\mathbf{z}). \tag{3.17}$$

If the function shape $\Pr(\mathbf{z}|U=\mathbf{u})$ is simple and sharp implying not too much randomness, optimal solution through ML and MAP point estimations is good enough to reveal underlying true sub-network. Each element of the solution u_n could accurately reflect underlying activation status of the n-th gene. However, if uncertainty in the data cannot be ignored, a single "guess" from ML/MAP estimate is not informative to reflect the underlying network dynamics and inference uncertainty. In contrast, Bayesian mean estimate can reflect this uncertainty following probabilistic principal.

3.3.2 Metropolis random walk on graph (MRWOG)

Unfortunately, Bayesian mean in Eq. (3.17) is difficult to calculate analytically, since the posterior distribution $\Pr(U=\mathbf{u}|\mathbf{z})$ is difficult to compute according to Bayesian rule:

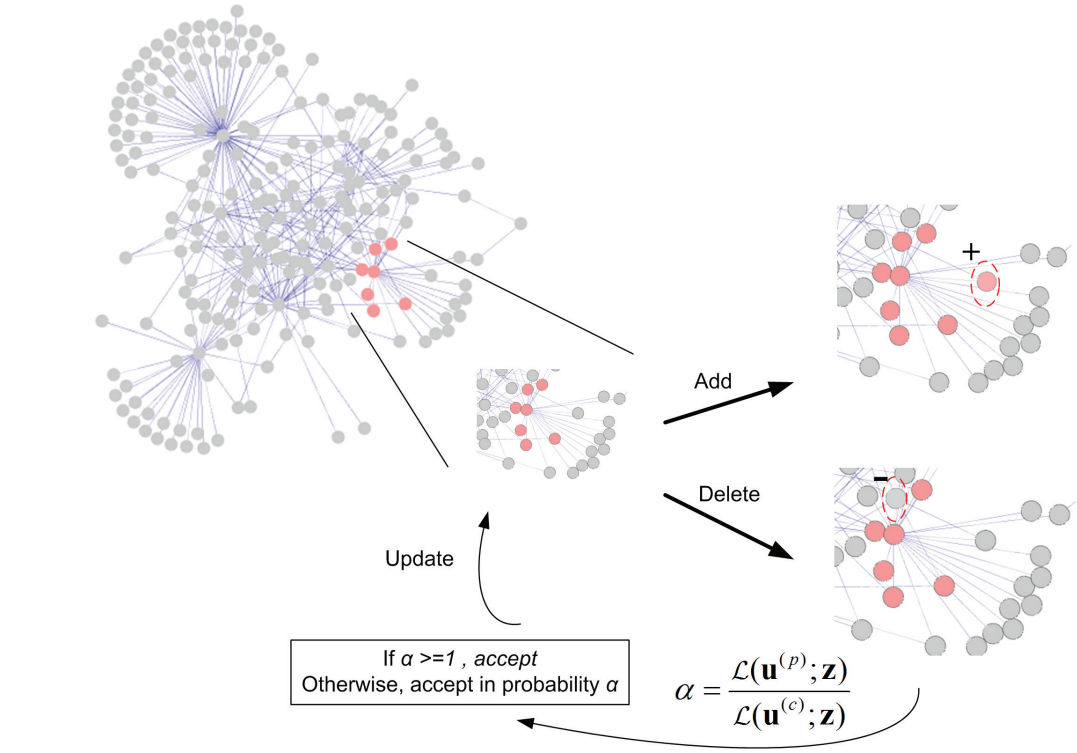
$$\Pr(U|\mathbf{z}) = \frac{\Pr(\mathbf{z}|U)\Pr(U)}{\sum_{U'\in\mathbb{U}} \Pr(\mathbf{z}|U')\Pr(U')}, \tag{3.18}$$

where $\Pr(U)$ is the priori distribution and $\mathbb{U}$ is its domain. To avoid this limitation, we proposed to use Monte-Carlo Markov Chain (MCMC) technique to accomplish the estimation in (3.18). Specifically, we used Metropolis sampling, a type of Monte-Carlo Markov Chain (MCMC) methods, to generate a series of samples that can be assumed as sampled from the posteriori distribution $\Pr(U|\mathbf{z})$. Here, we used Metropolis-Hastering sampling to obtain a sequence of random samples, by checking the following criterion value:

$$\alpha = \frac{\mathcal{L}(\mathbf{u}^{(p)};\mathbf{z})}{\mathcal{L}(\mathbf{u}^{(c)};\mathbf{z})} \frac{Q(\mathbf{u}^{(c)};\mathbf{u}^{(p)})}{Q(\mathbf{u}^{(p)};\mathbf{u}^{(c)})}, \tag{3.19}$$

in which, $Q(\mathbf{u}^{(p)};\mathbf{u}^{(c)})$ is a proposal function that is used to propose a new sample $\mathbf{u}^{(p)}$ based on current sample $\mathbf{u}^{(c)}$. Here, the function $Q(\cdot)$ is actually emulating a random walk as described in Fig. 3.3.

Since we assign the equal probability to add or delete one node from current sub-network,



Flip a coin,

If 'Head': select any unselected node j ($u_j = 0$), which is adjacent to at least one selected vertex ($u_i = 1$);

If 'Tail': deselect any selected leaf node j , which is already in the sub-network.

Figure 3.3: Illustration for proposal function $Q(\cdot)$ in proposed metropolis sampling scheme.

the sampling procedure is further simplified as Metropolis sampling:

$$\alpha = \frac{\mathcal{L}(\mathbf{u}^{(p)}; \mathbf{z})}{\mathcal{L}(\mathbf{u}^{(c)}; \mathbf{z})}. \tag{3.20}$$

If $\alpha \geq 1$, the newly proposed sample $\mathbf{u}^{(p)}$ will be accepted; if $\alpha < 1$, the newly proposed sample $\mathbf{u}^{(p)}$ will be accepted in probability α. The addition-deletion random walk starting from given initial selection $\mathbf{u}_{(0)}$ will eventually produce a Markov chain comprising of vertex selection samples:

$$MC_{\mathbf{u}} = \{\mathbf{u}_{(0)}, \mathbf{u}_{(1)}, \cdots, \mathbf{u}_{(l)}, \cdots, \mathbf{u}_{(L)}\}, \tag{3.21}$$

where each member $\mathbf{u}_{(l)}$ within this chain only depends on adjacent precedent member $\mathbf{u}_{(l-1)}$. Detail discussions, such as selection of acceptance rate and convergence check of Markov chain, are described as following sections.

Post-processing of resulting networks

Having sampled sub-networks according to proposed random walk scheme, we still have following questions regarding our ultimate goal - identifying dys-regulated sub-networks:
(a) Should we treat sampled sub-networks equally?
(b) How many underlying sub-networks are activated for given biological conditions?
The answer to question (a) is no, since we are more interested in sub-networks with relatively high scores and low-scored sub-networks are not of our interests. The answer to question (b) is less intuitively: if there are multiple underlying sub-networks and MRWOG could sample each true sub-networks enough times, we can use graph-cut to divide different sub-networks.

3.3.3 Metropolis sampling and MCMC (Monte-Carlo Markov Chain)

Markov chain is a discrete random process consisting of a series of successive random variables, and each variable is only dependent on the previous variable. Mathematically, a series of random variables $\{U_l\}_{l=1,\ldots,L}$ is Markov chain if the following equation holds:

$$\Pr(U_l|U_{l-1}, \cdots, U_1) = \Pr(U_l|U_{l-1}). \tag{3.22}$$

MCMC is a computational way to construct a Markov chain where the equilibrium distribution (which can also be called as steady-state distribution, limiting distribution, and stationary distribution in different literatures.) of samples within this chain is some desirable distribution. The convergence of Markov chain to the equilibrium distribution can be guaranteed, if (a) the chain is irreducible, and (b) every state is positive recurrent. In our sub-network identification application, each state is corresponding with a distinct sub-network, or equivalently a node selection vector $\mathbf{u}$. Let us assume there are H hidden statues and each state is associated with a different node selection vector, denoting as $\tilde{\mathbf{u}}(h)$ for h-th hidden state. Equilibrium distribution $\mathbf{\Pi} = [\pi_{ij}]_{i,j=1,\cdots,H}$ is satisfied with following condition: For any i and j,

$$\pi_{ij} = \sum_{i=1}^{H} \pi_i P_{ij}^r, \tag{3.23}$$

where P_{ij}^r is transition probability from i-th state to j-th state. Therefore, the Bayesian mean estimator is the ensemble of all the states weighted by equilibrium probability:

$$\hat{\mathbf{u}}_{BM} = \sum_{h=1}^{H} \tilde{\mathbf{u}}(h)\pi_h. \tag{3.24}$$

Gibbs sampler and Metropolis sampling are two most popular MCMC approaches to generate Markov chain, the samples of which follows some high-dimensional distribution $\Pr(U) = \Pr(U_1, \cdots, U_N)$. The difference between these two approach is that Gibbs sampler requires that every conditional distribution $\Pr(U_n|U_1, \cdots, U_{n-1}, U_{n+1}, \cdots, U_N)$ is known, while Metropolis only requires that the likelihood function $\mathcal{L}(\mathbf{u})$ of probability is known: $\mathcal{L}(\mathbf{u})$. In this work, we mainly adopt the Metropolis sampling scheme to generate Markov chain because $\mathcal{L}(\mathbf{u})$ is designed according to modified sub-network score function.

The random walk on graph is actually the transitions between different hidden states, where each state corresponds to one distinct sub-network. The basic principal is shown in Fig. 3.4 and Fig. 3.5.

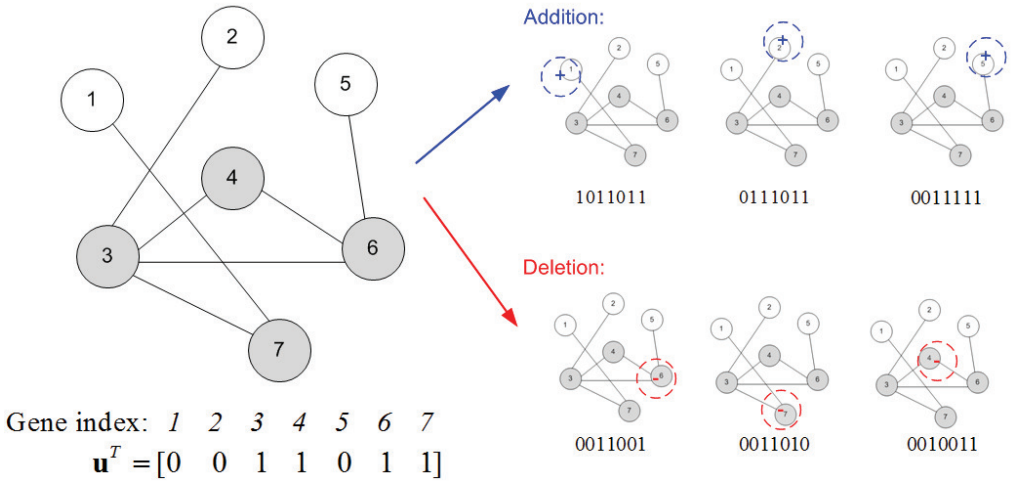


Figure 3.4: Illustration of basic principal of MRWOG. Starting from the sub-network on the left with four nodes being selected, there are six possibilities to propose new sub-networks, showing on the right.

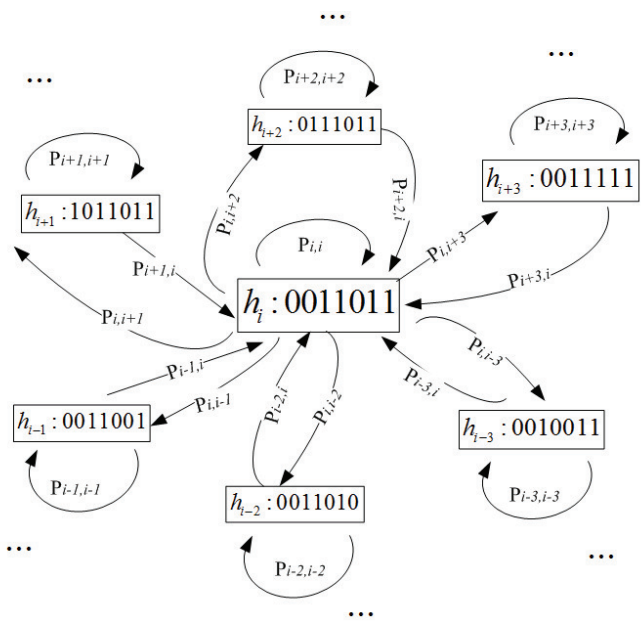


Figure 3.5: Illustration of hidden states of MRWOG, corresponding to the case shown in Fig. 3.4.

3.3.4 Priori distribution of metropolis sampling

Having the likelihood function

$$\mathcal{L}(\mathbf{u};\mathbf{z}) \propto \Pr(\mathbf{z}|U=\mathbf{u}), \tag{3.25}$$

we are interested to compute posteriori probability according to Bayesian rule

$$\Pr(U=\mathbf{u}|\mathbf{z}) = \frac{\Pr(\mathbf{z}|U=\mathbf{u})\Pr(U=\mathbf{u})}{\sum\limits_{U'\in\mathbb{U}} \Pr(\mathbf{z}|U')\Pr(U')}. \tag{3.26}$$

From (3.26) we have following relationship:

$$\Pr(U=\mathbf{u}|\mathbf{z}) \propto \Pr(\mathbf{z}|U=\mathbf{u})\Pr(U=\mathbf{u}) \propto \mathcal{L}(\mathbf{u};\mathbf{z})g(\mathbf{u}), \tag{3.27}$$

in which, $g(\mathbf{u})$ is the kernel function of the priori distribution: $\Pr(U=u) \propto g(\mathbf{u})$. If additional priori knowledge of sub-network is available, we can expect that the accuracy of algorithm can be further improved. In the simplest case, we just need to control the size of selected sub-network below preset size N_0:

$$g(\mathbf{u}) = \begin{cases} 1, & \sum u_n \leq N_0 \\ 0, & else \end{cases}, \tag{3.28}$$

which simply assumes every sub-network with size smaller than given threshold N_0 are equally possible. It is just like uniform priori. If sub-networks of certain size are interested in practice, we can also design specific priori distribution form. However, for the current study, we just use this simple priori for unbiased discovery purpose. If the priori distribution $g(\mathbf{u})$ is simple enough, it can also be incorporated into the design of proposal function $Q(\cdot)$. For example, instead of checking whether the size of sub-network is smaller than preset N_0, we can design the $Q(\cdot)$ to only propose the sub-network with size smaller than N_0. Another example of priori distribution can be about the density of the underlying sub-networks:

$$\mathcal{D}_{\mathbb{E}}(\mathbf{u}) = \frac{2N_{\mathbb{E}}(\mathbf{u})}{\sum u_n \left(\sum u_n - 1\right)}, \tag{3.29}$$

in which $N_{\mathbb{E}}(\mathbf{u})$ is the number of edges in selected sub-network according to $\mathbf{u}$, and $\sum u_n$ is the sub-network size. If the density of desired sub-network is known as priori, following form of $g(\mathbf{u})$ can be defined:

$$g(\mathbf{u}) = \exp\left(-\frac{\left(\mathcal{D}_{\mathbb{E}}(\mathbf{u}) - D_0\right)^2}{V}\right), \tag{3.30}$$

where V is a constant to control the degree of preference. Through this setup, we can extract the sub-networks of different priori interests.

Moreover, if molecular function and cellular component information from GO-term is of particular interests, we can also design priori likelihood function $g(\mathbf{u}, GO)$ to identify sub-network with specific context.

3.3.5 Convergence check

A Markov chain is deemed as "converged" if the samples produced after sufficient iterations are truly representative to the underlying stationary distribution. Once the convergence of MCMC can be determined, it is reasonable to terminate sampling process and compute statistics of interest using already generated samples.

Since the explicit form of stationary distribution is unknown, a direct check according to probability or likelihood is infeasible. Instead, the convergence diagnostics are mostly performed based on produced samples (Dodds and Vicini, 2004). We could have two ways to investigate the convergence of chain:

1. Split of a long chain:

By discarding the initial burn-in samples, we can split the remaining Markov chain $MC_{\mathbf{u}} = \{\mathbf{u}_{(1)}, \cdots, \mathbf{u}_{(L)}\}$ to be two halves: $MC_{\mathbf{u},left} = \{\mathbf{u}_{(1)}, \cdots, \mathbf{u}_{(\lfloor L/2 \rfloor)}\}$ and $MC_{\mathbf{u},right} = \{\mathbf{u}_{(\lfloor L/2 \rfloor+1)}, \cdots, \mathbf{u}_{(L)}\}$. If the chain is converged to equilibrium distribution, the same statistics derived from two halves:

$$\hat{\theta}(\mathbf{u} \in MC_{\mathbf{u},left})$$

and

$$\hat{\theta}(\mathbf{u} \in MC_{\mathbf{u},right})$$

should be very similar to each other. In practice, we can just compare the estimated posterior means by calculating Pearson or Spearman correlation value:

$$\rho = corr(\bar{\mathbf{u}}_{left}, \bar{\mathbf{u}}_{right}). \tag{3.31}$$

2. Comparison of multiple short chains:

An alternative way to check the convergence of MCMC is to check the closeness of multiple independent short chains, which are initialized randomly. If the Markov chains are converged, all of multiple chains will move away from initially assigned samples and generate samples following the same stationary distribution. Similarly with the scenario of single chain based check, we can calculate the statistics from each chain and comparing their closeness.

If cluster computing is available, the sampling of multiple chains can be employed in a parallel way to save computation time. After convergence is confirmed, the combination of produced samples from multiple chains will further enhance statistical power.

3.3.6 Acceptance rate and β value

Recalling the likelihood function $\mathcal{L}(\mathbf{u};\mathbf{z}) = \max(0, \mathcal{S}^{\beta}(\mathbf{u};\mathbf{z})\mathcal{C}_{\mathbb{E}}(\mathbf{u}))$, the value of $\beta(>0)$ does not affect the mode of underlying distribution $\Pr(\mathbf{z}|U=\mathbf{u})$, but does impact the way M-RWOG exploring the whole network. Specifically, β value affects the acceptance rate of metropolis sampling; a high acceptance rate makes the random walks wandering around the entire network while a low acceptance rate may make random walks trapped in local optimum. Acceptance rate is defined as:

$$\text{acpt}_L = \frac{1}{L}\sum_{l=1}^{L} \min(\alpha(l), 1), \tag{3.32}$$

where $\alpha(l)$ is the Metropolis sampling criterion value of l-th generated sub-network sample. A large acceptance rate tends to lead to a slow mixing process, that is, the generated Markov

chain needs a long time to converge to equilibrium distribution. A small acceptance rate will also make the chain converge slowly as a large portion of the proposed samples are rejected, and the chain tends to be stuck in a certain region, failing to explore the entire sample space.

In practice, we adjust the β value according to acceptance rate, since acceptance rate has clear implication for random walk within [0, 1].

3.3.7 Bootstrapping procedures

In reality, z_i, the relevance score of each gene, is derived from limited number of microarray measurements, the signal quality of which is known to be problematic, due to considerable amount of errors and noises from tissue sample preparation to hybridization. Therefore, it is very important to evaluate the statistical confidence of obtained computational results, with the awareness of limited sample effect and noise issues. To address statistical confidence of each node/edge selection, we propose to utilize bootstrapping technique to generate B bootstrap replicates $\mathbf{X}^{*1}, \cdots, \mathbf{X}^{*B}$ of original expression data $\mathbf{X}$, as well as associated phenotype information $\mathbf{c}^{*1}, \cdots, \mathbf{c}^{*B}$. For each $(\mathbf{X}^{*b}, \mathbf{c}^{*b})$ pair, MRWOG is applied to obtain corresponding Bayesian mean estimate $\mathbf{f}^{*b}_{\mathbf{u},BM} = MRWOG_{\mathbb{G}}(\mathbf{z}^{*b}) = MRWOG_{\mathbb{G}}(\mathbf{X}^{*b}, \mathbf{c}^{*b})$.

The confidence of each node selection can thus be computed from the bootstrap results as follows (Segal, Shapira et al. 2003):

$$\text{conf}(n\text{-th protein node}) = \frac{1}{B}\sum_{b=1}^{B} f_u^{*b}(n), \tag{3.33}$$

where $f_u^{*b}(n)$ is the n-th item of $\mathbf{f}^{*b}_{\mathbf{u},BM}$. Furthermore, we test the credibility of our confidence assessment by randomly permuting the phenotype information vector. Using random permutations, we can obtain an empirical distribution of the confidence score as the baseline. For a given confidence score $conf_0$, we can then calculate the false discovery rate (FDR) as follows:

$$\text{FDR}(\text{conf}_0) = \frac{\#\,\text{of expected false discoveries}}{\#\,\text{of true discoveries}} = \frac{\#\text{ of nodes with conf}_{b\text{aseline}} \geq \text{conf}_0}{\#\text{ of nodes with conf}_{observed} \geq \text{conf}_0}.$$

It should be noticed that this FDR is designed to assess the statistical significance of each individual node/edge.

3.3.8 Truncated mean

Let us assume the true sub-network is associated with a node selection vector $\mathbf{u}^{(True)}$. The probability $\Pr(U = \mathbf{u}|\mathbf{z})$ has a single mode (peak) in $U = \mathbf{u}^{(True)}$. It is expected that if the distribution of $\Pr(U = \mathbf{u}|\mathbf{z})$ is very wide and asymmetrical, Bayesian mean estimate $\mathbf{f}_{\mathbf{u},BM}$ approximated by sample average

$$\mathbf{f}_{\mathbf{u},BM} = \mathrm{E}_U[U|\mathbf{z}] = \sum_{U\in\mathbb{U}} U \Pr(U|\mathbf{z}) \approx \frac{1}{L}\sum_{l=1}^{L} \mathbf{u}_{(l)}, \tag{3.34}$$

which can be far away from $\mathbf{u}^{(True)}$. Two strategies can be adopted to address this problem:
(a) Increasing β value
By adjusting the value of β, we actually manipulate the shape of $\Pr(U|\mathbf{z})$. Ideally, if β is large enough, we can ensure the samples generated from MCMC is highly concentrated around $\mathbf{u}^{(True)}$. However, we also encounter the dilemma that larger β value could make the algorithm trapped in local optimum more easily, and consequently requires longer Markov chain to ensure the exploration of entire PPI network. Instead of greedily pursuing $\mathbf{u}^{(True)}$, we prefer to keep intermediate value of β and apply an alternative estimator.

(b) adopting truncated mean (TM) estimator
For single-peak distribution with heavy tails, truncated mean is robust to outliers. Here, we define truncated mean (TM) estimator as:

$$\mathbf{f}_{\mathbf{u},TM} \triangleq \frac{1}{\sum_{l=1}^{L} \mathbf{1}_q(\mathbf{u}_{(l)})} \sum_{l=1}^{L} \mathbf{u}_{(l)} \cdot \mathbf{1}_q(\mathbf{u}_{(l)}), \tag{3.35}$$

where

$$\mathbf{1}_q(\mathbf{u}_{(l)}) = \begin{cases} 1, & \mathcal{L}(\mathbf{u}_{(l)};\mathbf{z}) \geq Quantile(\{\mathcal{L}(\mathbf{u}_{(l)};\mathbf{z})\}_{l=1,\cdots,L}, q) \\ 0, & \text{otherwise} \end{cases} \tag{3.36}$$

is a 0-1 indicator function simply reflecting whether given sub-network sample $\mathbf{u}_{(l)}$ has associated function value larger than γ Quantile of function values of all the samples. $q \in [0, 1]$ serves as an adjustable parameter for user to decide how many sub-optimal solutions should be taken into consideration.

3.3.9 Further dissection of sub-network using graph-cut technique

Even with the sub-networks prioritized by MRWOG, occasionally we still observe that many genes tangling together with dense connections. To further clarify the sub-network results, we propose to dissect sub-network using graph-cut technique. Here, we only briefly describe graph cut for bipartition case as an example, as it is straight-forward extension to multi-partition situation. We denote the sub-network associated with given $\mathbf{u}$ as a graph $\mathbb{G}_{\mathbf{u}} = (\mathbb{V}_{\mathbf{u}}, \mathbb{E}_{\mathbf{u}})$. Denoting the two nodes connected through the k-th edge as v_{k1} and v_{k2}, we defined the similarity between these two nodes as the selection frequency of corresponding edge:

$$sim(v_{k1}, v_{k2}) \triangleq f_{\mathbf{y}}(k), \tag{3.37}$$

which reflects how likely these two nodes belongs to the same sub-network. By removing certain edges, the graph $\mathbb{G}_{\mathbf{u}}$ can be partitioned into two disjoint vertex sets $\mathbb{A}$ and $\mathbb{B}$, where $\mathbb{A} \cup \mathbb{B} = \mathbb{V}_{\mathbf{u}}$ and $\mathbb{A} \cap \mathbb{B} = \varnothing$. For applications such image segmentation the optimal partition is designed to be a minimum cut problem minimizing following cut function:

$$cut(\mathbb{A}, \mathbb{B}) = - \sum_{v_1 \in \mathbb{A}, v_2 \in \mathbb{B}} sim(v_1, v_2). \tag{3.38}$$

To avoid the partition bias preferring small sub-graphs, normalized cut (Ncut) was proposed to take total edge count connections into considerations:

$$Ncut(\mathbb{A}, \mathbb{B}) = \frac{cut(\mathbb{A}, \mathbb{B})}{assoc(\mathbb{A}, \mathbb{V})} + \frac{cut(\mathbb{A}, \mathbb{B})}{assoc(\mathbb{B}, \mathbb{V})}, \tag{3.39}$$

where $assoc(\mathbb{A}, \mathbb{V}) = - \sum_{v \in \mathbb{A}, t \in \mathbb{V}} sim(v, t)$ is the total connection from nodes in $\mathbb{A}$to all nodes in the graph. Many algorithms were proposed to achieve cut and later it has been reported that

spectral clustering methods can be readily used to solve graph cut problems (von Luxburg 2007). Review and in-depth discussions about spectral clustering can be seen in (von Luxburg 2007). Here, we adopt a classical graph-cut tool described in (Shi and Malik 2000), which was originally proposed to perform image segmentation using eigen-value decomposition technique.We utilized this tool of graph cut to do modular partition from Metropolis sampling results.

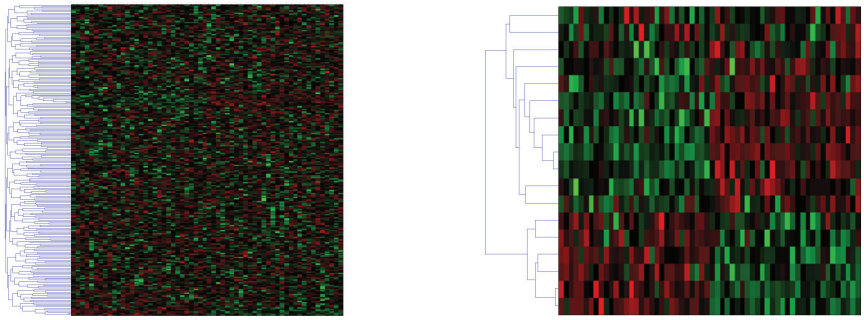
3.4 Simulation Studies of MRWOG

3.4.1 Simulation of gene expression data

To mimic real microarray data, we adopt the Gamma-Gamma (GG) model described in (Newton, Kendziorski et al. 2001) to generate simulation data. In GG model, observed gene expression x follows a Gamma distribution with shape parameter $\alpha_g > 0$ and scale parameter β_g, and the mean value of this distribution $\mu_g = \alpha_g \beta_g$. The probability density function of GG model is defined as

$$p(x|\alpha_g, \beta_g) = \frac{x^{\alpha_g - 1} \exp(-x/\beta_g)}{\beta_g^{\alpha_g} \Gamma(\alpha_g)}, \tag{3.40}$$

where the scale parameter β_g further follows a Gamma distribution with shape parameter α_0 and scale parameter β_0. Given these parameters, we can simulate the gene expression levels under two conditions with multiple replicates. Two types of expression patterns are generated according to whether this gene is "equally expressed" (EE) or "differentially expressed" (DE). In simulation, expression level of EE gene has same means under both conditions, while expression level of DE gene has different means. Examples of generated simulation expression are shown in Fig. 3.6.



(a) (b)

Figure 3.6: Examples of simulated expression data. (a) Heatmap of expression data containing both EE ("equally expressed") and DE ("differentially expressed") genes. (b) Heatmap of expression data only containing DE ("differentially expressed") genes.

3.4.2 Simulation network

To simulate different scenario in the real biological network, we constructed underlying ground truth sub-networks with a signaling structure with bridges connected with different modules, shown in Fig. 3.7. We purposely evaluate the performance of MRWOG using this simulation network and also compare with other related methods.

Starting with simulation network, we present in Fig. 3.8 influences of some related parameters to the algorithm. Fig. 3.8(a) displays the relationship between average acceptance rate and β: the larger value of β, the larger average acceptance rate. Fig. 3.8(b) presents the retrieving performance of MRWOG under various acceptance rates, where different Quantile level q are tested. It is observed that within a very wide range of acceptance rate between 0.3 to 0.7 with $q = 0.9$ almost equally good results are produced. Therefore, MRWOG is insensitive to selection of particular β value as long as the acceptance rate is in this range. Fig. 3.8(c) is the retrieving performance for bridges and hubs, where the performance is better than all the genes case. This is because MRWOG scheme is very effective to prioritize topologically important genes through the random walk scheme. Fig. 3.8(d) is the conver-

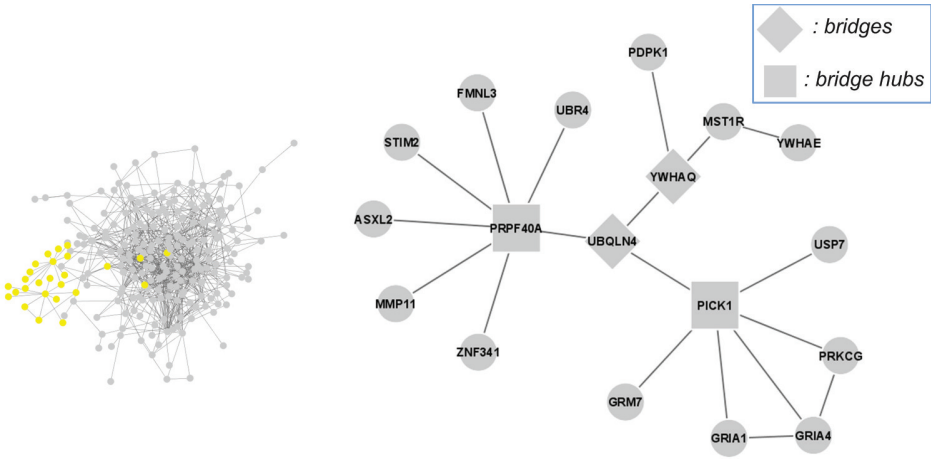


Figure 3.7: Illustration for the simulation network, where both hub and bridge nodes are considered. Left dense graph refers to the global network we used in simulation, and yellow nodes highlighted forming the underlying true sub-network, in a zooming view as right network.

gence check for Markov chains with varying q values, it is observed that generally higher q-value will converge slower in terms of spearman correlation value. This is understandable as fewer samples are used for higher q-value MRWOG models. Even though, we can see that the convergence is achieved consistently. After the spearman correlation with $q = 0$ as high as 0.7 (after around 9000 iterations), both the performance of all genes detection and hub/bridge detection will be very stable.

Comparison with other methods

In order to fully evaluate the performance of proposed MRWOG algorithm with or without edge information, we also perform comparison against two other methods, jactiveModule (Ideker, Ozier et al. 2002), and Heinz (Dittrich, Klau et al. 2008). As the sub-network cost functions are quite similar, we just briefly review these approaches as follows: jactiveModule solves the sub-network identification problem through simulated annealing search and simply return several sub-networks with maximum sub-network score. Its disadvantage is originated

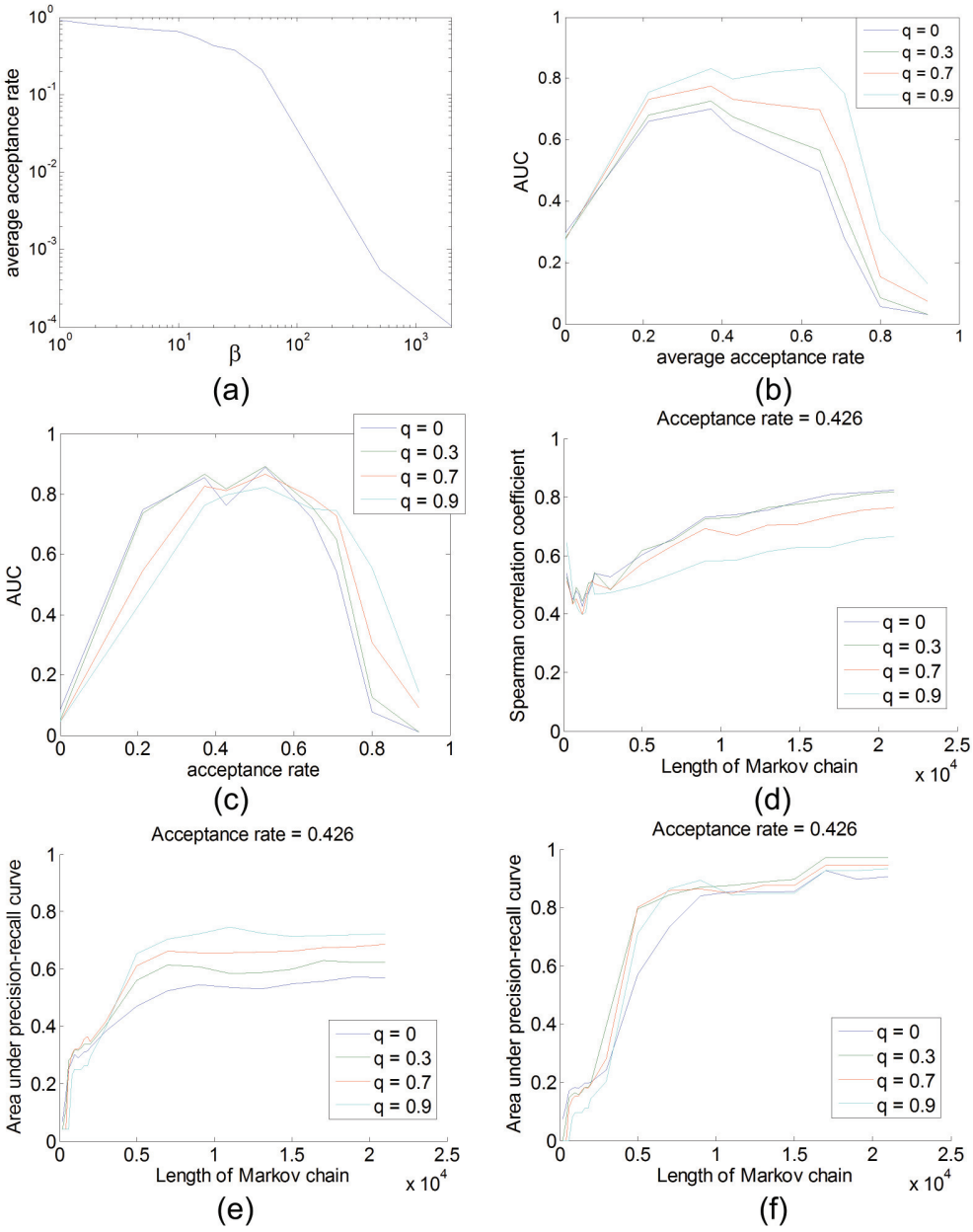


Figure 3.8: (a) β versus acceptance rate. (b) AUC for all genes under various q (Quantile) values. (c) AUC for hub genes under various q (Quantile) values. (d) Convergence check function according to varying length of Markov chain. (e) AUC for all the ground truth genes according to varying length of Markov chain. (f) AUC for all the ground truth hub genes according to varying length of Markov chain.

from its heuristic search procedure, it starts from given initial gene(s) and grows in the direction to increase the network score; Since every different running tends to lead to different sub-network results even with the same gene initialization, the computational results of jactiveModule is highly variable. Also, there are multiple heuristic parameters affecting the performance of jactiveModule as well: the maximum jumps allowing the method to explore, the temperature of simulated annealing and the stopping criterion (usually it is configured as without significant growing). Heinz actually reformulates the original sub-network identification problem; it converts the differentiation gene score to FDR (False Discovery Rate) and only considers the genes with FDR above certain threshold. Then it solves the sub-network identification problem using a prize-collecting tree. Because Heinz solves the problem in an exact way, the optimal solution can be achieved if appropriate FDR threshold is given. However, it can also only provide with single sub-network solution, without statistical confidence. Also, the selection of each gene vertex is in a binary way. In many situations, researchers are more interested to scale the sub-network inference results in a continuous way. Therefore, any sub-network associated properties such as sub-network size are also unclear to the users.

From the simulation results shown in Fig. 3.9, it is not surprised to observe that jactiveModule achieves the worst performance due to its heuristic searching strategy. Heinz achieves quite similar performance with MRWOG except for the cases that hub genes only have small or no differentiation between two simulated conditions. In contrast, MRWOG can incorporate the edge information to further improve its performance (with the assumption that genes with larger differentiation score are more likely to be connected through edges with higher confidence.) Good performance of MRWOG is understandable as the random walk on graph is guided towards the direction to identify most differentiable sub-networks. The random acceptance procedure also provides algorithm with opportunities to fully explore the whole solution space, without trapping into local optima. Moreover, the advantage of MRWOG is not limited to this as it can further provide the Bayesian mean estimate with associated statistics for each vertex and edge, which cannot be provided by conventional

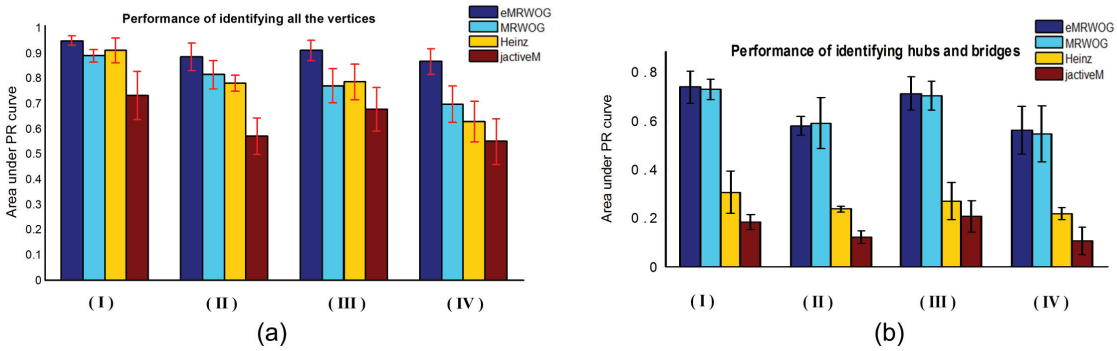


Figure 3.9: Performance comparisons of different algorithms for identification of (a) all the genes, (b) hub and bridge genes only. (I) and (III) correspond to simulations where ground truth sub-networks that are highly differentiable and moderately differentiable and all the vertices are equally differentiable. (II) and (IV) corresponds to simulations where ground truth sub-networks that are highly differentiable and moderately differentiable but hubs and bridges are not differentiable.

methods.

Performance comparison with point estimate

In real applications, it is always difficult to know the size of underlying true sub-network, and many algorithms are designed based on unrealistic assumption that sub-network size is known. The point estimate based on incorrect sub-network size leads to bias estimation where posteriori mean is much robust to such bias. We simulate the case that underlying true sub-network has 18 genes but the sampled sub-network size is 14, and such mismatch is expected to affect the performance of point estimate. Fig. 3.10 (a) and (b) show that ML point estimator generally performs worse than truncated mean based on MRWOG scheme. Presumably, if accurate priori (sub-network size) is given, ML should lead to very similar estimates with MRWOG. However, if improper priori given, point estimate could "over-fit" to the data, while sampling based scheme remain robust, if a reasonable q value is given (around 0.9). The performance evaluation we used here is F-measure, which considers both

precision and recall:

$$\text{F-measure} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{3.41}$$

Performance comparison with varying sub-network size

Furthermore, if we vary the size of generated sub-networks by MRWOG and Heinz, from 1 to 200. For each sub-network size case, we run multiple times to draw the performance interval based on F-measure, shown in Fig. 3.11. We can see that sampling based MRWOG is much robust than optimization based Heinz, as the 15%-85% Quantile interval of MRWOG is much narrower than interval of Heinz, and median performance of MRWOG is also clearly better than Heinz.

Prioritization of condition-specific nodes and edges

After MRWOG analysis, each gene node is assigned a selection frequency score we denoted as $f_{\mathbf{u}}(n)$, $n = 1, \cdots, N$, where

$$f_{\mathbf{u}}(n) = \mathrm{E}\left[U_n\right] \approx \frac{1}{L} \sum_{l=1}^{L} u_n(l). \tag{3.42}$$

Similarly, we also have edge selection frequency score $f_{\mathbf{y}}(k)$,$k = 1, \cdots, K$ assigned to each interaction edge, where

$$f_{\mathbf{y}}(k) = \mathrm{E}\left[Y_k\right] \approx \frac{1}{L} \sum_{l=1}^{L} y_k(l). \tag{3.43}$$

Using node score $f_{\mathbf{u}}(n)$ and edge score$f_{\mathbf{y}}(k)$, we can prioritize proteins and interactions specific to the biological conditions of our interests. We show the simulation results or node and edge prioritization as Fig. 3.12. For node prioritization, we used z-score based ranking as the baseline, and it can be observed from Fig. 3.12(a) that prioritization using both node scores and edge scores have much better than z-score based ranking, it is because the sampling consider the interrelationship between different genes. Similarly, Fig. 3.12 (b) shows that both edge score and node score based schemes perform much better than

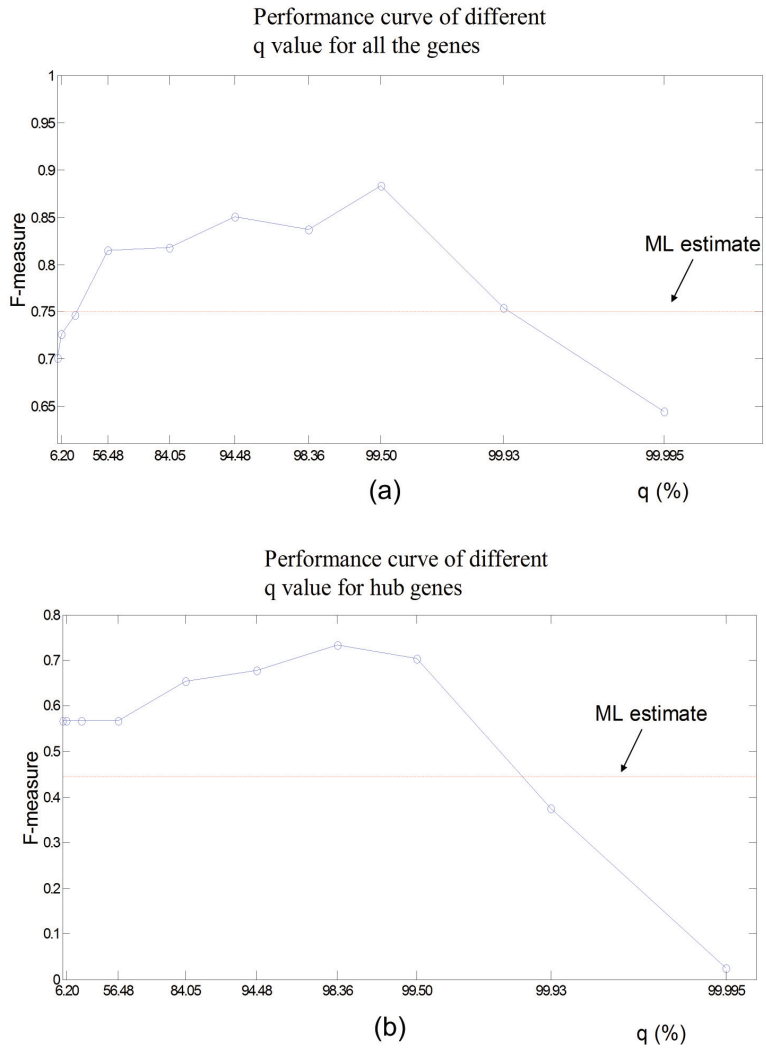


Figure 3.10: Performance curves of F-measure for different q values. (a) Performance curve for identifying all underlying genes. (b) Performance curve for identifying hub genes only. Red dashed lines indicate baseline performance according to using ML point estimates.

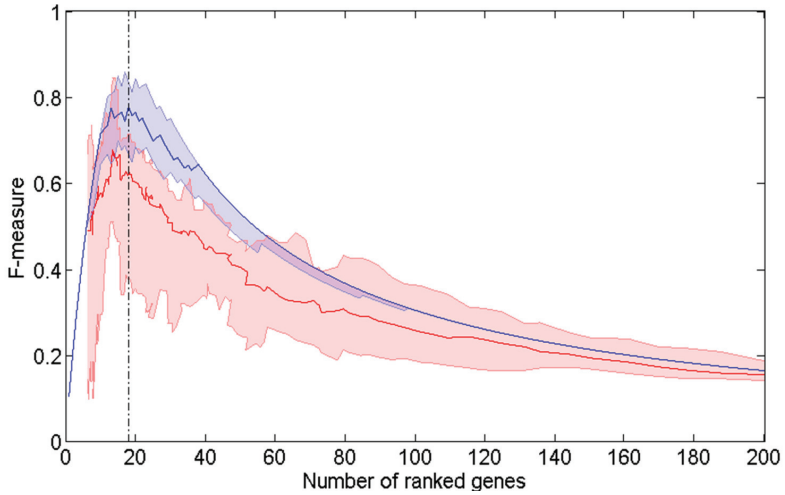


Figure 3.11: Performance comparison between MRWOG and Heinz. Blue shadow line is according to MRWOG, and red shadow line is according to Heinz. The range of the shadow line is 15% to 85% Quantile of F-measures after multiple runnings.

correlation approach, which use the correlation value of expression patterns of gene pairs to prioritize condition-specific edges. Interestingly, except for the region where q-value is very high (> 0.97), edge score perform much better than node score to prioritize condition-specific edges. This shows that although edge selection frequency is highly dependent to node selection frequency, it still emphasizes the specificity of edges that are not totally determined by the specificity of nodes.

Time complexity comparison

In general, the computation complexity of random walk approaches are difficult to assess analytically, considering that exploration efficiency is highly dependent on given graph and underlying stationary distribution. Empirically, we perform twenty runnings to compare the average running times of MRWOG, jactiveModule and Heinz in simulation studies. It turns out Heinz is the fastest algorithm with average computation time second; jactiveModule ranks the second with average time 1.35 seconds; while MRWOG is the most computation expensive in terms of running time (12.4 seconds). We can expect that the running time

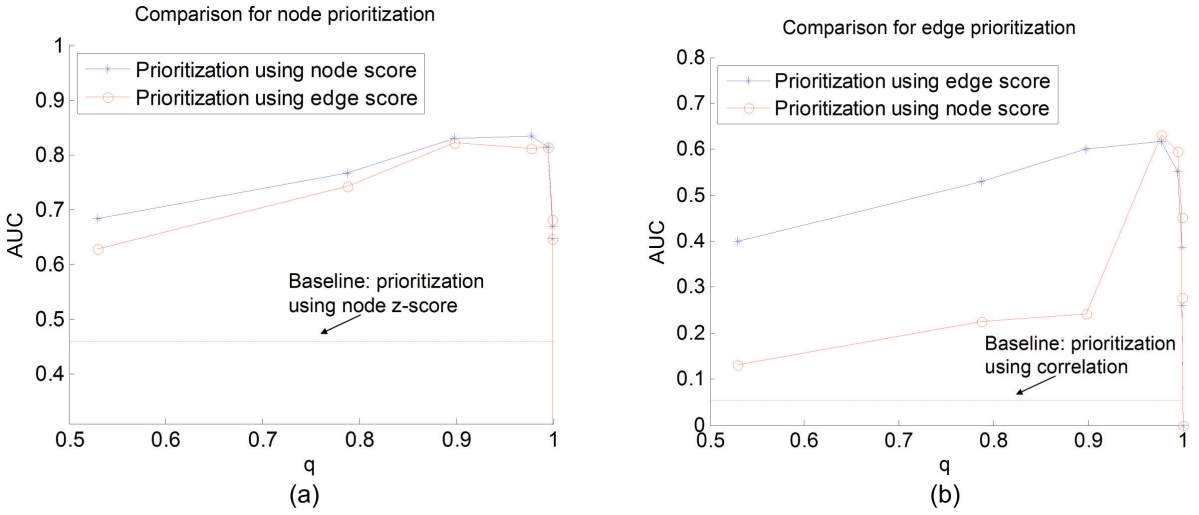


Figure 3.12: Prioritization of condition-specific (a) nodes and (b) edges using MRWOG.

difference would be even larger if a larger graph is tested, as MRWOG aims to thoroughly explore entire given network while the two other methods mainly focused on providing an ad-hoc single solution. However, since the sampling scheme could be easily distributed for parallel computation, the running time of proposed MRWOG can be greatly reduced.

3.5 Experiments on Real Biological Data

3.5.1 Experiments on yeast galactose-utilization pathway

In (Ideker, Thorsson et al. 2001), genetic perturbations on several key genes of galactose-utilization pathway were carried out to interrogate related cellular responses. Based on a focus physical interaction network incorporating both PPI and PDI interactions and differential scores derived from (Ideker, Ozier et al. 2002), MRWOG was run to prioritize genes involving with galactose utilization. Since MRWOG running results consists of node and edge selection frequency implying the importance of each gene and interaction edge, we can

easily visualize such prioritization effects by setting the size of nodes and width of edges proportional to selection frequency. Moreover, since MRWOG provides continuous selection frequency, we can setup different frequency threshold to look into the affected network in different scales, shown as in Fig. 3.13 (a), (b) and (c).

Overall, most edges prioritized by using high selection frequency belong to PDI categories as shown in Fig. 3.13(c), which suggests transcriptional regulation is heavily involved in galactose-utilization pathway. Moreover, a skeleton of most frequently visited genes were highlighted in this case, informing us the key components of the sub-networks. However, we can also investigate multi-scale results provided by MRWOG, whether they can lead us into different angles. Although large changes in down-stream transcriptional levels are expected and presumable, genes playing signaling roles with moderate changes are also of interested. In the result generated by setting low selection frequency threshold, we can observe a large part of PPI interactions also occurs in the top-left corner of Fig. 3.13 (a). By simply expanding one jump from genes with signaling transduction roles, we obtain a sub-network is mainly consisted of members of MAPK signaling pathway, especially in pheromone branches, shown in Fig. 3.14 (a) and (b). It is already known that MAPK pathway receives the environmental stresses, responds and regulates fundamental metabolisms (Gehart, Kumpf et al. 2010). While "high-resolution" result using high selection frequency threshold provides mostly changed sub-network, the "low-resolution" picture is also helpful to reveal some moderately alternated but closer to up-stream sub-network. Starting from "middle-resolution" result using intermediate selection frequency, we can already generate a rough graph-cut result with clear biological process divisions, shown in Fig. 3.15. This analysis can also been seen as a modularity analysis, as each module with much thicker edges imply their expression variations are more coordinated induced. In summary, MRWOG is a very useful tool to investigate dys-regulated gene networks in a scalable way and to facilitate the formation of novel hypotheses.

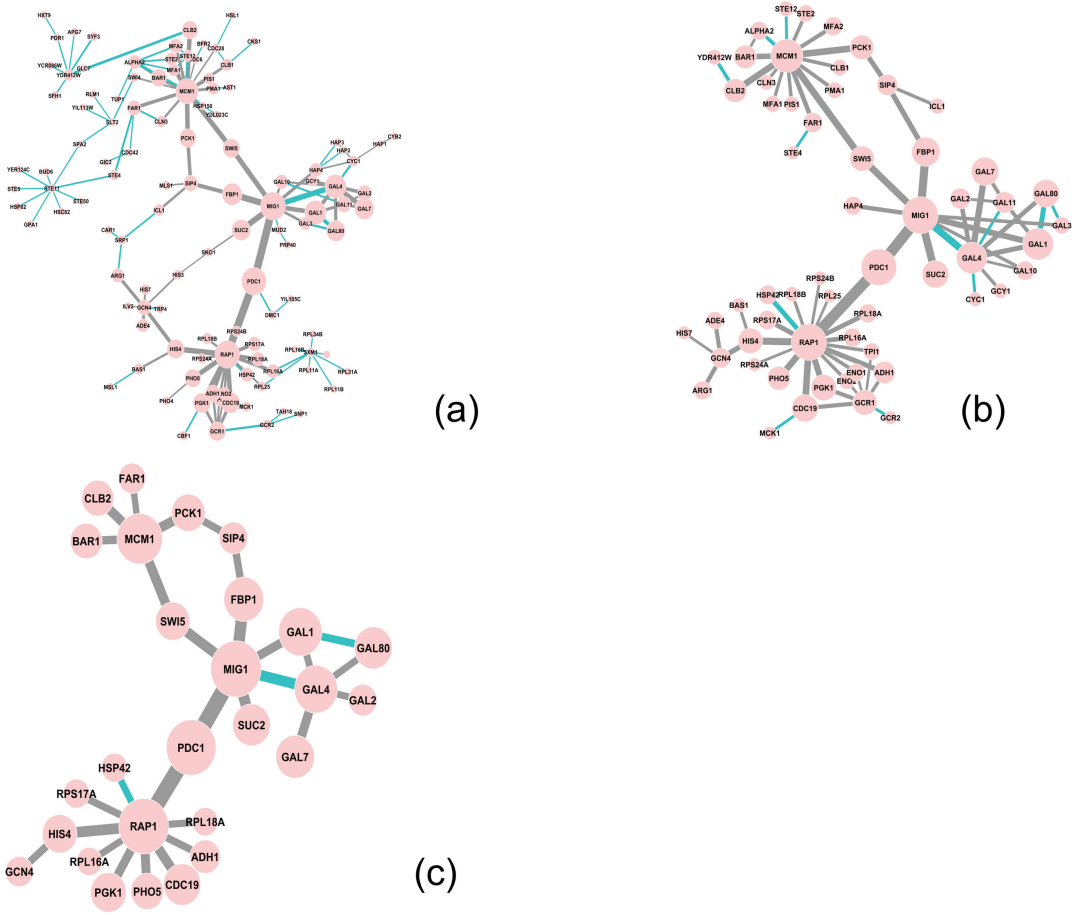


Figure 3.13: Visualization of MRWOG results for galactose experiments with various selection frequency threshold: (a) low selection frequency; (b) intermediate selection frequency; (c) high selection frequency.

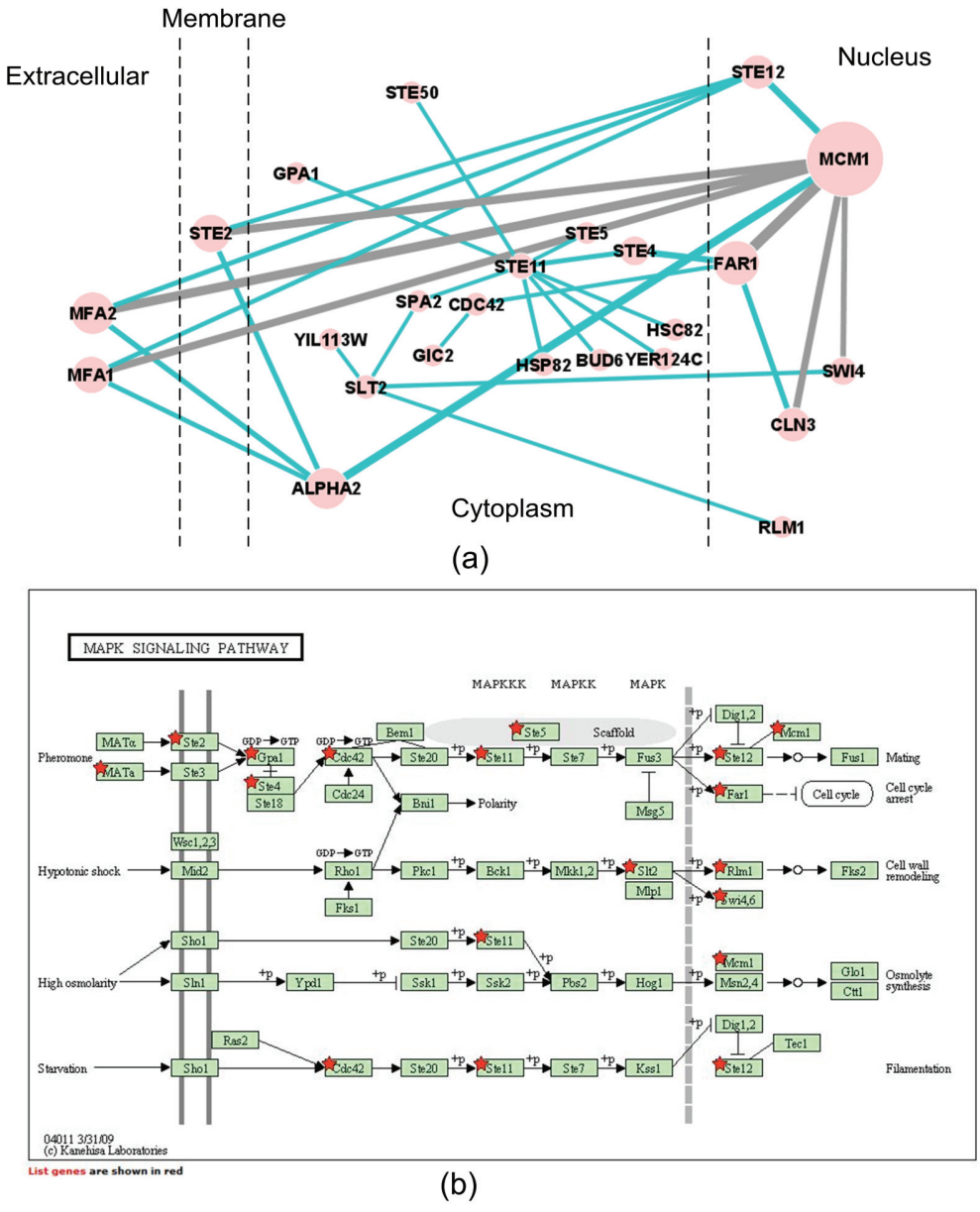


Figure 3.14: (a) The sub-network selected from one jump of genes with signaling transduction roles. (b) MAPK signaling pathway enrichment result where red star indicating genes belong to the selected sub-network.

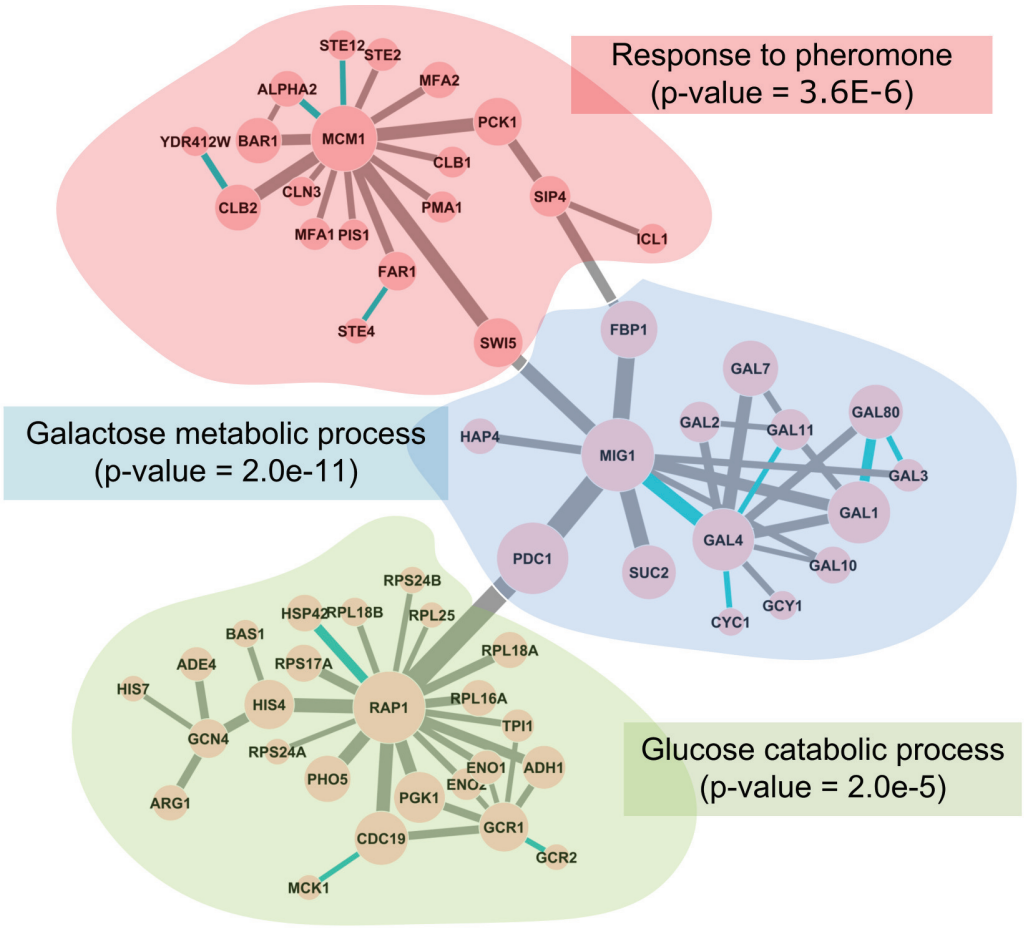


Figure 3.15: Graph-cut of MRWOG results for intermediate selection frequency. Each colored box describes the top enriched biological process with a Benjamin corrected p-value in corresponding division.

3.5.2 Experiments on breast cancer patient data

We further proceed to breast cancer studies, where the abnormal activities of protein sub-networks may also contribute to the understandings of breast cancer mechanisms. We divide patients into three different groups: early, late and none, according to their different clinical outcomes. While early and none groups imply the totally drug resistance and effective treatment, respectively, the late group corresponds to the progression that the treatment was initially effective but later turns to be resistance. Therefore, comparing to early vs. none (EvN) study, we are more interested to reveal the underlying sub-networks contributing to the differentiation of early vs. late study (EvsL), which could shed some lights on the signaling pathway studies of cancer metastasis and drug resistance.

We use two breast cancer data sets for experiment: Edinburg (in-house data set) and Loi data set (Loi, Haibe-Kains et al. 2007) with similar survival time division. Initial protein network is extracted from Protein-protein interaction network from HPRD (Keshava Prasad, Goel et al. 2009) with the selection criterion of two jumps from ESR1. We used Probe Logarithmic Intensity ERror (PLIER) algorithm with Quantile normalization to preprocess the original intensity data for gene expression measurements. We further convert gene expression data from probe set IDs to Entrez gene IDs. After mapping the PPI with available gene expression from both data sets, we obtain PPI network containing about 2,358 genes and 12,595 interactions.

By running MRWOG for 1,000,000 iterations, we obtain node selection frequency and edge selection frequency, which can help us prioritize condition-specific protein nodes and interaction edges. We visualize the final results in Fig. 3.16, where the selection frequency of nodes and edges are proportional to the node size and edge width, respectively. It is interesting to observe that although the resulted two sub-networks are quite different, several important players known to be related with breast cancer are shared by both sub-networks: ESR1, AR and EGFR. We demonstrate the overall overlap of individual genes using Venn diagram, shown in Fig. 3.17. Besides the high overlap rate in terms of individual genes, we also

find that both results are highly overlapped in terms of pathway enrichment using DAVID bioinformatics tools (Huang da, Sherman et al. 2009), shown as Table 3.2. It suggests that even difference between distinct data sets could be very large in terms of individual genes, the biological functionality and pathway activation could serve as the converging point to understand their underlying common mechanism.

We further lay out these overlapped 37 protein nodes on Edinburgh resulted sub-network and Loi resulted sub-network, shown in Fig. 3.18. Despite the commonality of protein nodes, we can observe a lot of discrepancies in terms of interaction edges, and as well as fold-change in different data sets. This observation informs us that the complexity and diversity of biological interactions.

Bootstrapping analysis

As we mentioned in previous sections, even having the calculation of node selection frequency to prioritize important proteins for each data set, we still need to cautious about the results as it is derived from noisy microarray data. One way to enhance our confidence is through statistical assessment such as bootstrapping. Using 100 random permutations, we obtained a baseline distribution of the confidence, shown as Fig. 3.19(b); the confidence calculated from the original (i.e., non-permuted) data set is also shown in Fig. 3.19(a) for a comparison. We used a gamma distribution function to fit the confidence distributions for both cases, and plotted the fitted confidence distributions in Fig. 3.19(c). For a given confidence score $conf_0$, we can then calculate the false discovery rate (FDR). Fig. 3.19(d) shows the number of nodes in relation to different FDR cutoff values; from the figure we can obtain the following FDR values for the top 50 nodes shown in Fig. 3.20: 50 nodes with FDR$\leq$1.9e-4; 35 nodes with FDR$\leq$1.7e-5; 26 nodes with FDR$\leq$2.5e-6; 15 nodes with FDR$\leq$3.6e-7. It should be noticed that this FDR is designed to assess the statistical significance of each individual node/edge. A network point of view is also important to assess multiple genes/interactions globally. The FDR of sub-networks can also be computed in a similar way.

Further graph-cut analysis

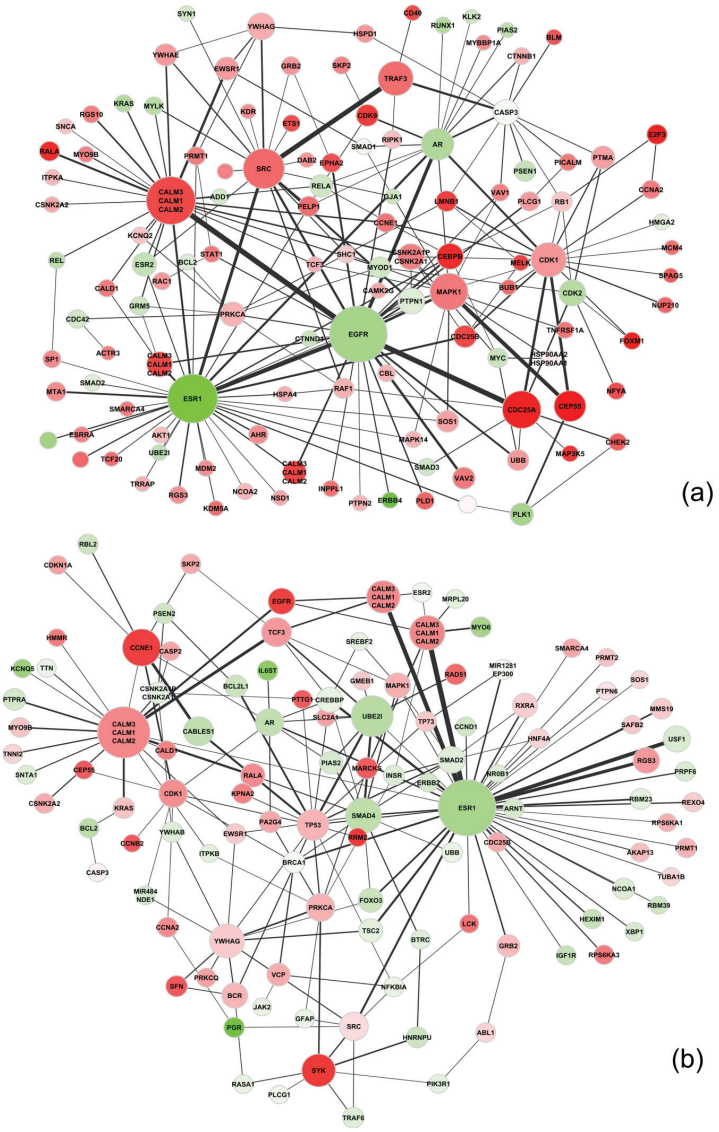


Figure 3.16: Visualization of MRWOG results on (a) Edinburgh data set and (b) Loi data set for early versus late survival analysis. The color of nodes indicates fold-change of gene expression of corresponding protein node. Red means over-expressed in 'early recurrent' patient group and green means over-expressed in 'late-recurrent' patient group. The node size and edge width is proportional to the node and edge selection frequency according to MRWOG.

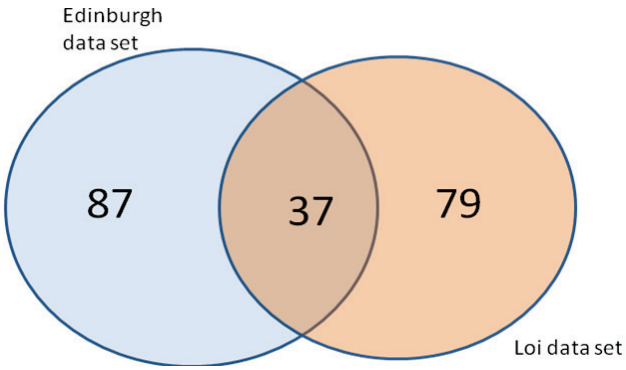


Figure 3.17: Venn diagram of overlapped protein nodes from two separate MRWOG analysis.

Using previously described graph-cut technique, which has close relationships with image segmentation (Tao, Jin et al. 2007) and emerging spectral clustering techniques (Archip, Rohling et al. 2005), we can divide MRWOG results into different modules, inter-connections of which are denser than connections among different modules. Using Edinburgh 'early recurrent' vs. 'late recurrent' study as an example, the original MRWOG resulting sub-network shown in Fig. 3.16 (a) can be divided into three different modules, shown in Fig. 3.21. The first module displayed in Fig. 3.21(a) has significant enrichment in 'Neurotrophin signaling pathway' (FDR = 0.02%), which has been shown having close relationship with MAPK pathway under estrogen induced condition (Singh, Setalo et al. 1999). Moreover, many evidences support that nerve growth factors and signaling stimulate the cell growth in breast cancer (Dolle, Adriaenssens et al. 2004). The second module is highly enriched in 'cell cycle pathway' (FDR = 3.8E-6%) and 'TGF-beta signaling pathway' (FDR=0.2%). The third module is dominated by 'ErbB signaling pathway' (FDR = 3.0E-4%), which has well known association with antiestrogen resistance mechanism (Shou, Massarweh et al. 2004) so that several clinical researches have been carried out to inhibit this pathway for achieving improved treatment results (Kurokawa and Arteaga 2001). From this initial study of using graph-cut technique, we illustrate that further detailed investigations can be performed based on 'fine-resolution' according to each divided local modules.

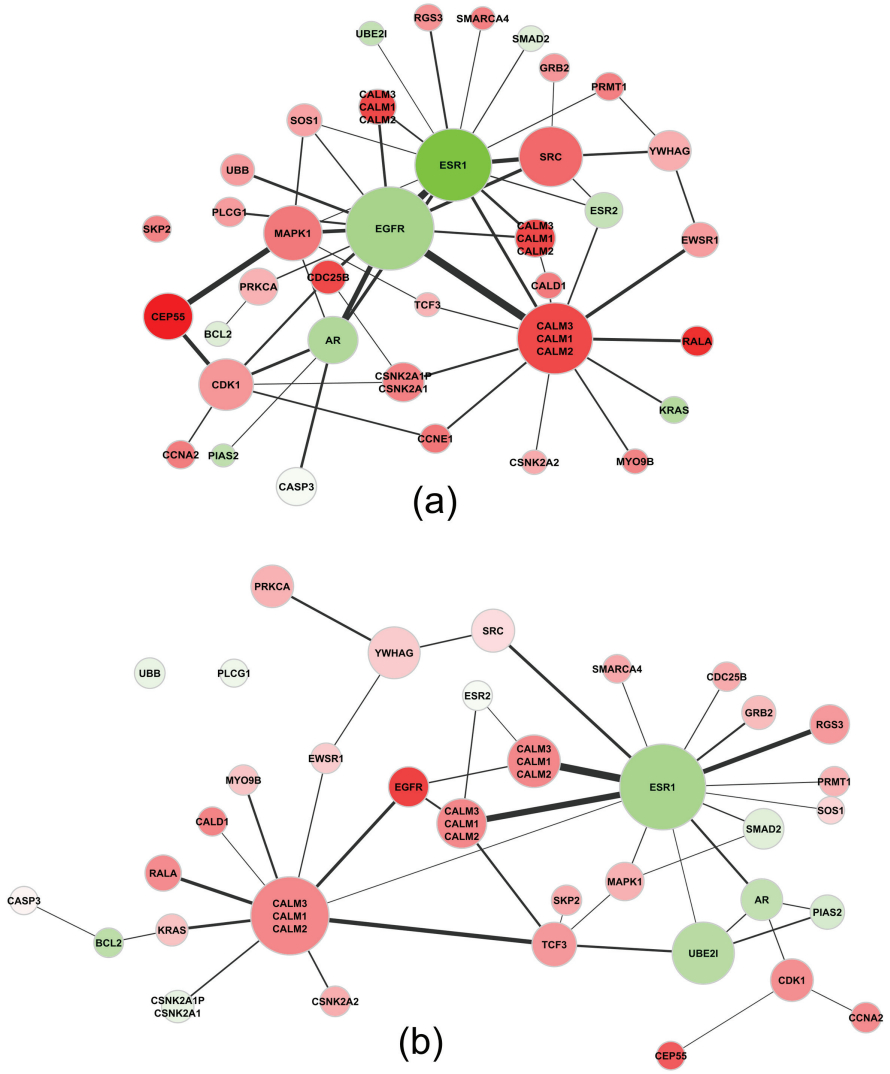


Figure 3.18: Visualization of overlapped protein nodes between two MRWOG results on Edinburgh and Loi data sets lay out on (a) Edinburg resulted sub-network and (b) Loi resulted sub-network. The color of nodes indicates fold-change of gene expression of corresponding protein node. Red means over-expressed in 'early recurrent' patient group and green means over-expressed in 'late-recurrent' patient group. The node size and edge width is proportional to the node and edge selection frequency according to MRWOG.

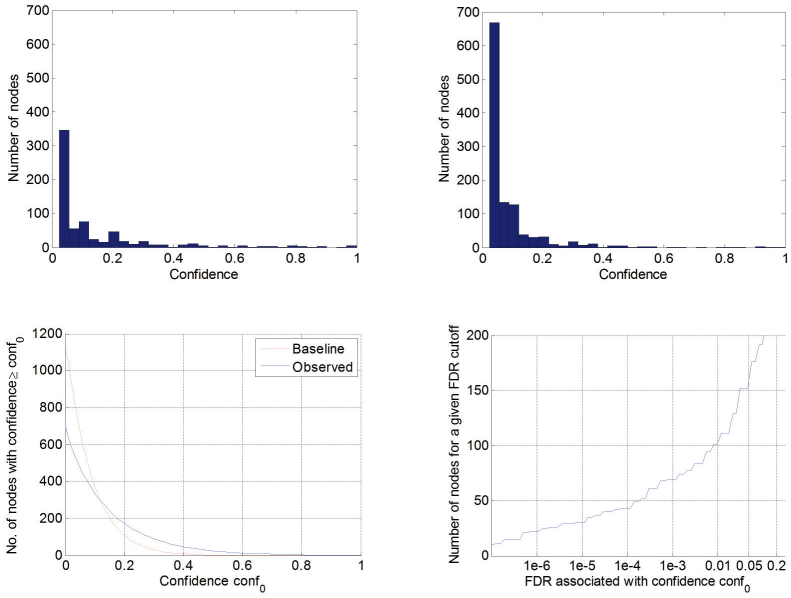


Figure 3.19: Credibility of confidence measured by permutations: (a) observed confidence; (b) baseline confidence obtained from permutations; (c) fitted confidence distributions; (d) number of nodes with respect to FDR cutoff value.

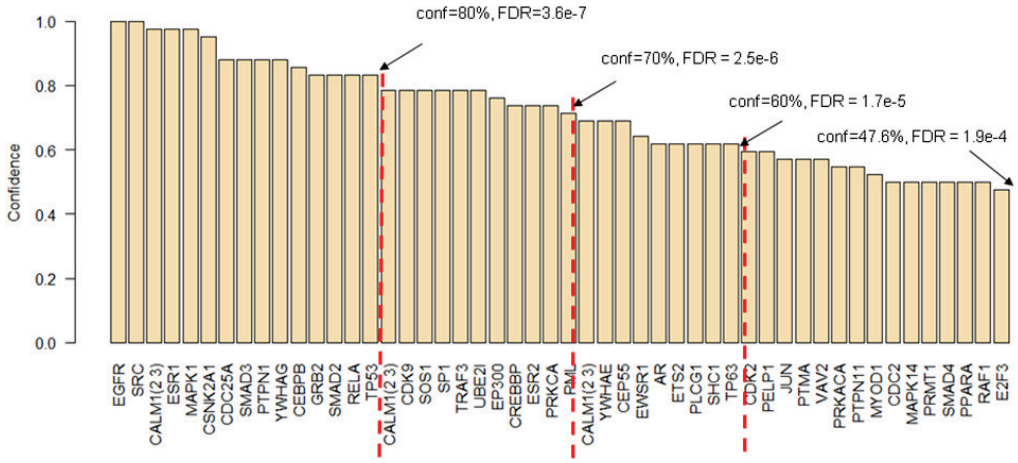


Figure 3.20: Confidence of the selected nodes as calculated by the bootstrap method. The list of the top 50 nodes with conf 0.476 is shown with their gene symbols.

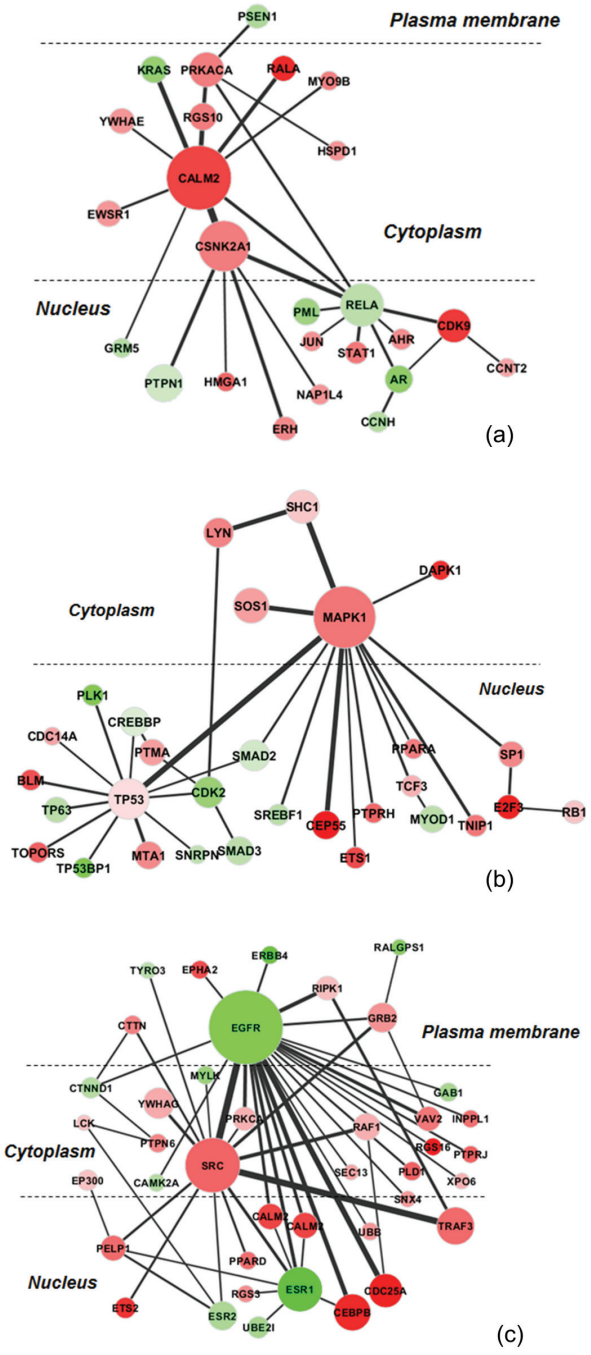


Figure 3.21: Graph-cut analysis for Edinburg 'early recurrent' vs. 'late recurrent' analysis, where (a), (b) and (c) are sub-networks divided from MRWOG result shown in Fig. 3.16(a).

3.6 Discussions on MRWOG

The identification of dys-regulated PPI sub-networks under certain biological condition or disease sub-type could provide important clues for further experimental investigations. In contrast with optimization based sub-network searching schemes, which only provide yes/no selection of protein nodes, our proposed MRWOG solves the sub-network identification problem in a stochastic manner, where association score of each sub-network is treated as a likelihood implying how possible this sub-network is dys-regulated. From this perspective, we formulate the objective of MRWOG as minimization of Bayesian mean square errors, taking into account all the sub-optimal sub-networks. With the Bayesian mean estimates, MRWOG is capable of assigning a probability score to each protein node and interaction edge. The score of one protein node indicates how often this node is selected in one sub-optimal sub-network. The score of one interaction edge suggests how often the nodes connecting with this edge is co-selected in one sub-optimal sub-network.

In addition, the probabilistic formulation makes it easy to incorporate additional information into sub-network discovery schemes, such as protein interaction confidence score, or sub-network modularity priori. The advantage of Bayesian mean estimate over original point estimate is that it accommodates inference uncertainties, and more over, it is not needed to provide any single solution, which could miss important information. Instead, the Bayesian mean doesn't require the complete coverage of the whole sample space.

Table 3.1: Mathematical notations in Chapter 3.

PPI network	$\mathbb{G} = (\mathbb{V}, \mathbb{E})$
Expression	$\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]^T \in \mathbb{R}^{N \times M}$
Phenotype label or disease outcome	$\mathbf{c}$
Protein node score	$\mathbf{z}$
Interaction edge score	$\mathbf{w}$
Node selection vector	$\mathbf{u}$
Edge selection vector	$\mathbf{y}$
Random vector of node selection	U
Random vector of edge selection	Y
Number of genes	N
Number of edges	K
Number of patient samples	M
Length of Markov chain	L
Number of bootstrapping	B
Number of hidden states	H
Node score function	$z_n = \mathcal{A}(\mathbf{x}_n, \mathbf{c})$
Sub-network score function	$\mathcal{S}(\mathbf{u}; \mathbf{z}) = \sum_{n=1}^{N} z_n u_n / \sqrt{\sum_{n=1}^{N} u_n}$
Sub-network score function with edge	$\mathcal{S}_e(\mathbf{u}, \mathbf{y}; \mathbf{z}, \mathbf{w}) = \left(\sum_{n=1}^{N} z_n u_n + \lambda \sum_{k=1}^{K} w_k y_k \right) / \sqrt{\sum_{i=1}^{N} u_i}$
Likelihood function	$\mathcal{L}(\mathbf{u}; \mathbf{z}) = \max(0, \mathcal{S}^{\beta}(\mathbf{u}; \mathbf{z}) \mathcal{C}_{\mathbb{E}}(\mathbf{u}))$
Probability mass function	$\Pr(\mathbf{z} \vert U = \mathbf{u}) \propto \mathcal{L}(\mathbf{u}; \mathbf{z})$
Bayesian mean estimator	$\mathbf{f}_{node} = \widehat{\mathbf{u}}_{\mathrm{BM}}(\mathbf{z})$
(node selection frequency)	$= \mathrm{E}_U[U \vert \mathbf{z}] = \sum_{U \in \mathbb{U}} U \Pr(U \vert \mathbf{z})$
Markov chain of node selection vector	$MC_{\mathbf{u}} = \{\mathbf{u}_{(0)}, \mathbf{u}_{(1)}, \cdots, \mathbf{u}_{(l)}, \cdots, \mathbf{u}_{(L)}\}$

Table 3.2: Pathways that enriched in MRWOG results on both data sets, with a FDR cut-off 5%.

KEGG pathway (ID: name)	**Loi** (# of genes (FDR%))	**Edinburgh** (# of genes (FDR%))
hsa05200:Pathways in cancer	33 (2.30E-16)	34 (3.50E-16)
hsa04110:Cell cycle	19 (9.90E-11)	19 (4.23E-10)
hsa04520:Adherens junction	13 (7.89E-7)	12 (2.60E-5)
hsa04115:p53 signaling pathway	11 (4.15E-5)	6 (4.57)
hsa04012:ErbB signaling pathway	12 (4.16E-5)	15 (4.95E-8)
hsa04010:MAPK signaling pathway	13 (0.49)	17 (3.55E-3)
hsa04910:Insulin signaling pathway	8 (3.7)	10 (0.28)

Chapter 4

Conclusion and future work

In this dissertation, we propose an integrative network analysis framework for genomic pathway inference from two different levels: transcriptional regulatory network (TRN) inference and protein-protein interaction (PPI) sub-network identification. Specifically, we develop several computational approaches to accomplish condition-specific network analysis tasks by integrating mRNA expression data, protein-DNA interaction and protein-protein interaction information.

In this Chapter, we first summarize original contributions of this dissertation, and then discuss some limitations of current approaches. Finally, we conclude this dissertation with several potential research questions for extending current framework.

4.1 Summary of Contributions

Nowadays, high-throughput screening of genomic signals become a routine experiment for biologists to investigate the abnormal changes in different conditions. However, revealing of underlying mechanistic system remains a non-trivial task, considering the complexity of cellular system and inherent noises in measurements. To study genomic data in a biological

context, we present an integrative framework to dissect molecular pathways into two types of network analysis: protein-DNA network analysis and protein-protein interaction network analysis. The major contributions of this dissertation are summarized as follows.

4.1.1 Condition-specific transcription regulatory network inference with biological knowledge of all TFs

Since transcriptional regulation is the major driving force for gene expression, the inference of transcription regulatory network could facilitate to understand the controlling mechanisms in normal cellular system (Yang, Suen et al. 2005), as well as programmed dynamics in cancer cells (Creighton, Cordero et al. 2006). Since the inference task purely based gene expression is very challenging, many computational methods were proposed to integrate expression with protein-DNA interaction information to combat noise in expression data. However, two major issues remain unsolved: (1) biological knowledge of protein-DNA interactions is not always available, especially for species other than yeast and E. coli. ; (2) Integrative analysis based on inconsistent biological knowledge and data could produce misleading results and confound our understandings.

With the awareness of these problems, we have proposed several strategies. For the problem of transcriptional regulatory network inference where all potentially active TFs are known, we have proposed a computational approach, namely motif-directed Network Component Analysis (mNCA), to integrate gene expression profiles and DNA sequence motif information so that we can infer the underlying transcription factor activities (TFAs) and the regulatory relationship from TFs to their targets. mNCA is designed to facilitate the regulatory network inference problem in the situation that ChIP-chip data is not available. Since the initial connection information extracted from DNA motif is very noisy and contains considerable amounts of false positives, the estimated TFA cannot reflect the relevance of one TF to certain biological condition. To address the relevance of TF by using noisy biological knowledge, we further proposed a novel computational scheme called knowledge-based stability analysis

to test the consistency between mRNA expression and biological knowledge. In the proposed stability analysis, multiple random perturbations are first applied onto network connections, and then a stability score is calculated as the average distance of different TFA estimations after perturbations. The rationale behind stability analysis is that the TFA estimation of an active TF is less likely to be affected by the random perturbation on knowledge than other non-active TFs. Besides mNCA and kSA, we also propose a computational strategy to handle the under-determined case, where number of microarray samples smaller than number of TFs. This strategy greatly extends the applicability of mNCA to explore larger number of TFs.

The proposed scheme showed robust and superior performance over conventional approaches. We applied stability analysis to yeast cell cycle experiment and further to a series of anti-estrogen breast cancer studies. In both experiments not only biologically relevant regulators are highlighted, the condition-specific transcriptional regulatory networks are also constructed, which could provide further insights into the corresponding cellular mechanisms.

4.1.2 Transcription regulatory network inference with biological knowledge of single TF

Since NCA is a matrix decomposition method based on biological knowledge constraints, which is designed for regulatory network inference under assumption that all potentially active TFs are known beforehand, it is not suitable for the case where only a partial TF list is known. We further propose a single TF knowledge guided approach, regulatory component analysis (RCA), which explicitly find the linear projection maximizing the coincidence with given partial biological knowledge. The linear extraction scheme also allows RCA to detect FPs and FNs of biological knowledge, which is inconsistent with gene expression data. RCA can be regarded as a complementary work to matrix decomposition based methods such as NCA to infer transcriptional regulatory networks where the knowledge of TFs is incomplete.

The contributions of RCA works are two-folded: first, we formulate a linear extraction scheme for transcriptional regulatory network inference problem, by utilizing incomplete but informative biological knowledge. The proposed scheme show significant performance improvement over traditional NCA methods in both simulations and real biological experiment in E. coli. Second, through designed simulation studies, we showed that how to efficiently integrate biological knowledge is not a trivial problem, considering the given biological knowledge is usually incomplete and in-consistent to the data we have. An inappropriate incorporation of biological knowledge may even lead to worse performance than the methods without using biological knowledge.

4.1.3 Protein-protein interaction sub-network identification

With increasingly accumulated protein interaction data, the identification of condition-specific protein sub-networks emerges as an attractive research problem, solutions of which can facilitate understandings of molecular mechanisms, and provide reliable sub-network bio-markers for disease diagnosis/prognosis. Most of the existing algorithms mainly search for sub-networks enriched with differentially expressed genes, but overlook their potential interactions and topological importance. In addition, the identification of sub-network is usually solved through optimization schemes, and there is no condition-specific score associated with each gene and each interaction. This makes prioritization of genes and interactions infeasible, and hinders the interpretation of network results.

With increasingly accumulated protein interaction data, the identification of condition-specific protein sub-networks emerges as an attractive research problem, solutions of which can facilitate understandings of molecular mechanisms, and provide reliable sub-network bio-markers for disease diagnosis/prognosis. Most of the existing algorithms mainly search for sub-networks enriched with differentially expressed genes, but overlook their potential interactions and topological importance. In addition, the identification of sub-network is usually solved through optimization schemes, and there is no condition-specific score associ-

ated with each gene and each interaction. This makes prioritization of genes and interactions infeasible, and hinders the interpretation of network results.

4.2 Future Extensions

There are many potential extensions can be done starting from current framework. We describe some major directions in this section.

Generalization of knowledge-based stability analysis

In any integrative analysis scenarios, data-knowledge consistency or consistency between different data types is an issue could mislead computational analysis. Although we demonstrate the effectiveness of stability analysis in transcriptional regulatory network inference application, through extensive simulation studies and experiments on real biological data, it remains unclear how well this scheme can be extended to other integrative analysis applications, for example, protein-protein interaction sub-network identification. It certainly requires more theoretical work to gain the insights about how to generalize stability analysis to other knowledge integration applications. This part of future work is not isolated as ideas could be borrowed from many closely related research fields, such as perturbation analysis and robust estimators.

Statistical extension of regulatory component analysis

In current formation of regulatory component analysis, we mainly focus on the coordination between biological knowledge and expression data. Noticeably, pure statistical approaches such as PCA and ICA can also reveal certain portion of target genes, without using any biological priori. This raises an interesting research question that how to develop the method utilizing both biological knowledge and statistical properties of underlying signals, for example, sparse property of regulatory component. Another following question is how to find such a balance point to fully exploit the information from both sides. Fortunately, the

current formation of RCA allows flexible incorporation of extra statistical criterions, which have been extensively studied in pattern recognition area (De Bie, Cristianini et al. 2005) and signal processing field (Parra and Sajda 2003).

Incorporation of domain-knowledge in MRWOG

Currently, we only make very mild assumption about priori of MRWOG: every protein node and interaction edge is equally possible to be activated. Such uniform priori is appropriate to be used for unbiased exploration at initial stage. However, for further focused study it is essential to have domain-knowledge from the biological problems we studied, such as cellular location, regulation/activation direction of the interaction edge, implication of each interaction (inhibition/activation/collaboration) and etc. The power of Bayesian principal is relying on the incorporation of both data likelihood and prior belief without seeing the data. To move forward to infer a biological context specific pathway, the incorporation of domain-knowledge and formulation of informative priori distribution is very essential. In this dissertation, we lay out foundation for further study on this topic.

Directional inference of protein-protein interactions

The current modeling of PPI using MRWOG is a simplification of undirected graph. As we know that PPI involves both undirected interactions such as protein complex formation, as well as directed interactions such as post-translation modification. There are several potential ways to incorporate or infer directions of PPI: first, based on current MRWOG scheme, we can enforce the random walk to be directional if additional direction information is available, such as post-translation modification information collected in various of PPI database (Mathivanan, et al., 2006) and cellular location information derived from gene-ontology database (Ashburner, et al., 2000); second, given the upstream proteins (ligands and receptors) and downstream TFs of signaling pathway, several methods treat inference signaling transduction from upstream to downstream as a path finding problem (Scott, et al., 2006; Yeger-Lotem, et al., 2009; Zhao, et al., 2008), and directions of intermediate interactions can be assigned according to optimal path finding.

Utilization of hyper-graph cut for protein sub-network division

In MRWOG scheme, one of potential research directions is to investigate the application of advanced graph-cut techniques for sub-network dissection. Generalized from traditional graph, a hyper-graph contains hyper-edges that can connect arbitrary number of nodes (Klamt, et al., 2009). The graph-cut scheme has also been extended to accommodate hyper-graph scenario. The concept of hyper-graph is especially suitable for describing formation of protein complex, since each protein complex contains multiple protein members and each protein could also involve into multiple protein complexes. It is worthy to notice that hyper-graph has not only been utilized in some of exiting bioinformatics research works for integrating multiple data-type (Tian, et al., 2009) or analyzing metabolic pathway (Mithani, et al., 2009), but also been proposed to interrogate interaction network. It is certainly a promising research direction to extend MRWOG beyond pair-wise interaction network to hyper-graph network (Ladroue, et al., 2009).

4.3 Conclusion

Biotechnology is moving in an unexpected fast pace. While the initial draft of first human genome sequence finished in 2001 costs around $3 billion in total, it only took $22 million in 2006 to acquire the complete sequence of a nonhuman primate (Service 2006). In 2010, it was announced by multiple biotech companies that a human genome can be sequenced in a day for less than $6,000 (Venter 2010). With much cost-effective sequence techniques, gene expression, methylation and many other genomic signals can be measured in a much more accurate way (Metzker 2009). However, how to interpret huge amount of data and advance our understandings of biological systems based on these data still remained to be challenges, considering large gene population versus small sample number, inherent measurement noises, heterogeneity in tissue preparation, and individual genomic difference.

Since it is known that molecules interact within each other cellular system, a network analysis

appears to be useful to delineate the data and facilitate biological interpretations. In this dissertation, we present an integrative analysis framework for dissecting molecular pathway into two parts: protein-DNA interaction network and protein-protein interaction network. For different network analysis scenario, we proposed several integrative approaches: (1) motif-guided network component analysis (mNCA) and knowledge-based stability analysis (kSA) for identifying condition-specific TFs and target genes, (2) regulatory component analysis (RCA) for identifying target genes with partial protein-DNA knowledge, and (3) Metropolis random walk on graph (MRWOG) for identifying dys-regulated protein sub-networks and prioritizing condition-specific proteins and interactions. Through extensive simulation and experiments on real microarray data, we have demonstrated that (1) condition-specific TFs can be highlighted with proposed stability analysis and a condition-specific regulatory network can be constructed based on that, (2) gene targets can be obtained even with partial biological knowledge, and (3) the given interaction network knowledge can be re-evaluated a specialized MCMC algorithm following Bayesian principal. All of above algorithms and applications involve the careful design with respect to integrative analysis, especially with the emphasis on the consistency between multiple data types and information sources. It can be foreseen that with more data, integrative analysis will become more and more important, and eventually lead to the complete understandings of pathway.

Bibliography

Aguilar, H., X. Sole, et al. (2010). "Biological reprogramming in acquired resistance to endocrine therapy of breast cancer." Oncogene 29(45): 6071-83.

Allison, D. B., X. Cui, et al. (2006). "Microarray data analysis: from disarray to consolidation and consensus." Nat Rev Genet 7(1): 55-65.

Alter, O. and G. H. Golub (2004). "Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription." Proc Natl Acad Sci U S A 101(47): 16577-82.

Archip, N., R. Rohling, et al. (2005). "Ultrasound image segmentation using spectral clustering." Ultrasound Med Biol 31(11): 1485-97.

Ashburner, M., et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nat Genet, 25, 25-29.

Bader, J. S., A. Chaudhuri, et al. (2004). "Gaining confidence in high-throughput protein interaction networks." Nat Biotechnol 22(1): 78-85.

Bjornstrom, L. and M. Sjoberg (2005). "Mechanisms of estrogen receptor signaling: convergence of genomic and nongenomic actions on target genes." Mol Endocrinol 19(4): 833-42.

Blow, N. (2009). "Systems biology: Untangling the protein web." Nature 460(7253): 415-8.

Boscolo, R., C. Sabatti, et al. (2005). "A generalized framework for network component analysis." IEEE/ACM Trans Comput Biol Bioinform 2(4): 289-301.

Bousquet, O. and A. Elisseeff (2002). "Stability and Generalization." Journal of Machine Learning Research 2: 499-526.

Brynildsen, M. P. and J. C. Liao (2009). "An integrated network approach identifies the isobutanol response network of Escherichia coli." Mol Syst Biol 5: 277.

Brynildsen, M. P., L. M. Tran, et al. (2006). "A Gibbs sampler for the identification of gene expression and network connectivity consistency." Bioinformatics 22(24): 3040-6.

Buggy, Y., T. M. Maguire, et al. (2004). "Overexpression of the Ets-1 transcription factor in human breast cancer." Br J Cancer 91(7): 1308-15.

Calle, M. L. and V. Urrea (2010). "Letter to the Editor: Stability of Random Forest importance measures." Brief Bioinform.

Camon, E., M. Magrane, et al. (2004). "The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology." Nucleic Acids Res 32(Database issue): D262-6.

Carbon, S., A. Ireland, et al. (2009). "AmiGO: online access to ontology and annotation data." Bioinformatics 25(2): 288-9.

Cardoso, J. F. (1998). "Blind signal separation: statistical principles." Proceedings of the IEEE 86(10): 2009-2025.

Cardoso, J. F. (1999). "High-order contrasts for independent component analysis." Neural Computation 11: 157-192.

Cardoso, J. F. and A. Souloumiac (1993). "Blind beamforming for non-Gaussian signals." Radar and Signal Processing, IEE Proceedings F 140(6): 362-370.

Carro, M. S., W. K. Lim, et al. (2010). "The transcriptional network for mesenchymal transformation of brain tumours." Nature 463(7279): 318-25.

Carroll, J. S., C. A. Meyer, et al. (2006). "Genome-wide analysis of estrogen receptor binding sites." Nat Genet 38(11): 1289-97.

Chang, C., Z. Ding, et al. (2008). "Fast network component analysis (FastNCA) for gene regulatory network reconstruction from microarray data." Bioinformatics 24(11): 1349-58.

Chekmenev, D. S., C. Haid, et al. (2005). "P-Match: transcription factor binding site search by combining patterns and weight matrices." Nucleic Acids Res 33(Web Server issue): W432-7.

Chen, L., J. Xuan, et al. (2010). "Multilevel support vector regression analysis to identify condition-specific regulatory networks." Bioinformatics 26(11): 1416-22.

Chuang, H. Y., E. Lee, et al. (2007). "Network-based classification of breast cancer metastasis." Mol Syst Biol 3: 140.

Clarke, R., N. Brunner, et al. (1989). "Progression of human breast cancer cells from hormone-dependent to hormone-independent growth both in vitro and in vivo." Proc Natl Acad Sci U S A 86(10): 3649-53.

Clarke, R., H. W. Ressom, et al. (2008). "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data." Nat Rev Cancer 8(1): 37-49.

Conlon, E. M., X. S. Liu, et al. (2003). "Integrating regulatory motif discovery and genome-wide expression analysis." Proc Natl Acad Sci U S A 100(6): 3339-44.

Creighton, C. J., K. E. Cordero, et al. (2006). "Genes regulated by estrogen in breast tumor cells in vitro are similarly regulated in vivo in tumor xenografts and human breast tumors." Genome Biol 7(4): R28.

Cruces-Alvarez, S. A., A. Cichocki, et al. (2004). "From blind signal extraction to blind instantaneous signal separation: criteria, algorithms, and stability." Neural Networks, IEEE Transactions on 15(4): 859-873.

Dahlman-Wright, K., et al. (2006) International Union of Pharmacology. LXIV. Estrogen receptors, Pharmacological reviews, 58, 773-781.

Daschner, P. J., H. P. Ciolino, et al. (1999). "Increased AP-1 activity in drug resistant

human breast cancer MCF-7 cells." Breast Cancer Res Treat 53(3): 229-40.

Dawy, Z., M. Sarkis, et al. (2008). "Fine-Scale Genetic Mapping Using Independent Component Analysis." Computational Biology and Bioinformatics, IEEE/ACM Transactions on 5(3): 448-460.

De Bie, T., N. Cristianini, et al. (2005). Eigenproblems in Pattern Recognition. Handbook of Geometric Computing : Applications in Pattern Recognition, Computer Vision, Neuralcomputing, and Robotics, Springer-Verlag: 129-170.

Delmar, P., S. Robin, et al. (2005). "VarMixt: efficient variance modelling for the differential analysis of replicated gene expression data." Bioinformatics 21(4): 502-8.

Dittrich, M. T., G. W. Klau, et al. (2008). "Identifying functional modules in protein-protein interaction networks: an integrated exact approach." Bioinformatics 24(13): i223-31.

Dodds, M. G. and P. Vicini (2004). "Assessing convergence of Markov chain Monte Carlo simulations in hierarchical Bayesian models for population pharmacokinetics." Ann Biomed Eng 32(9): 1300-13.

Dolle, L., E. Adriaenssens, et al. (2004). "Nerve growth factor receptors and signaling in breast cancer." Curr Cancer Drug Targets 4(6): 463-70.

Drabsch, Y., H. Hugo, et al. (2007). "Mechanism of and requirement for estrogen-regulated MYB expression in estrogen-receptor-positive breast cancer cells." Proc Natl Acad Sci U S A 104(34): 13762-7.

Ein-Dor, L., I. Kela, et al. (2005). "Outcome signature genes in breast cancer: is there a unique set?" Bioinformatics 21(2): 171-8.

Faith, J. J., B. Hayete, et al. (2007). "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles." PLoS Biol 5(1): e8.

Gama-Castro, S., H. Salgado, et al. (2010). "RegulonDB version 7.0: transcriptional reg-

ulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units)." Nucleic Acids Res 39(Database issue): D98-105.

Gehart, H., S. Kumpf, et al. (2010). "MAPK signalling in cellular metabolism: stress or wellness?" EMBO Rep 11(11): 834-40.

Greenbaum, D., C. Colangelo, et al. (2003). "Comparing protein abundance and mRNA expression levels on a genomic scale." Genome Biol 4(9): 117.

Hakes, L., J. W. Pinney, et al. (2008). "Protein-protein interaction networks and biology–what's the connection?" Nat Biotechnol 26(1): 69-72.

Halees, A. S., D. Leyfer, et al. (2003). "PromoSer: A large-scale mammalian promoter and transcription start site identification service." Nucleic Acids Res 31(13): 3554-9.

Harbison, C. T., D. B. Gordon, et al. (2004). "Transcriptional regulatory code of a eukaryotic genome." Nature 431(7004): 99-104.

Hall, J.M., Couse, J.F. and Korach, K.S. (2001) The multifaceted mechanisms of estradiol and estrogen receptor signaling, The Journal of biological chemistry, 276, 36869-36872.

Hoheisel, J. D. (2006). "Microarray technology: beyond transcript profiling and genotype analysis." Nat Rev Genet 7(3): 200-10.

Honkela, A., C. Girardot, et al. (2010). "Model-based method for transcription factor target identification with limited data." Proc Natl Acad Sci U S A 107(17): 7793-8.

Huang da, W., B. T. Sherman, et al. (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nat Protoc 4(1): 44-57.

Huber, P. (1981). Robust Statistics, Wiley-Interscience.

Huynh, H. T., E. Tetenes, et al. (1993). "In vivo inhibition of insulin-like growth factor I gene expression by tamoxifen." Cancer Res 53(8): 1727-30. Hyvärinen, A. (1999). "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis." IEEE Transactions on Neural Networks 10(3): 626-634.

Ideker, T., J. Dutkowski, et al. (2011). "Boosting Signal-to-Noise in Complex Biology: Prior Knowledge Is Power." Cell 144(6): 860-863.

Ideker, T., O. Ozier, et al. (2002). "Discovering regulatory and signalling circuits in molecular interaction networks." Bioinformatics 18 Suppl 1: S233-40.

Ideker, T., V. Thorsson, et al. (2001). "Integrated genomic and proteomic analyses of a systematically perturbed metabolic network." Science 292(5518): 929-34.

Igney, F. H. and P. H. Krammer (2002). "Death and anti-death: tumour resistance to apoptosis." Nat Rev Cancer 2(4): 277-88.

Jakacka, M., M. Ito, et al. (2001). "Estrogen receptor binding to DNA is not required for its activity through the nonclassical AP1 pathway." J Biol Chem 276(17): 13615-21.

Jensen, L. J., M. Kuhn, et al. (2009). "STRING 8–a global view on proteins and their functional interactions in 630 organisms." Nucleic Acids Res 37(Database issue): D412-6.

Ji, H. and W. H. Wong (2006). "Computational biology: toward deciphering gene regulatory information in mammalian genomes." Biometrics 62(3): 645-63.

Jung, T.-P., S. Makeig, et al. (2000). "Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects." Clinical Neurophysiology 111(10): 1745-1758.

Kalousis, A., J. Prados, et al. (2007). "Stability of feature selection algorithms: a study on high-dimensional spaces." Knowledge and Information Systems 12(1): 95-116.

Kao, K. C., Y. L. Yang, et al. (2004). "Transcriptome-based determination of multiple transcription regulator activities in Escherichia coli by using network component analysis." Proc Natl Acad Sci U S A 101(2): 641-6.

Kar, S., W. T. Baumann, et al. (2009). "Exploring the roles of noise in the eukaryotic cell cycle." Proc Natl Acad Sci U S A 106(16): 6471-6.

Kel, A. E., E. Gossling, et al. (2003). "MATCH: A tool for searching transcription factor

binding sites in DNA sequences." Nucleic Acids Res 31(13): 3576-9.

Keshava Prasad, T. S., R. Goel, et al. (2009). "Human Protein Reference Database–2009 update." Nucleic Acids Res 37(Database issue): D767-72.

Khurana, T. S. and K. E. Davies (2003). "Pharmacological strategies for muscular dystrophy." Nat Rev Drug Discov 2(5): 379-90.

Kindermann, R. (1980) Markov Random Fields and Their Applications (Contemporary Mathematics ; V. 1). Amer Mathematical Society.

Klamt, S., Haus, U.U. and Theis, F. (2009) Hypergraphs and cellular networks, PLoS Comput Biol, 5, e1000385.

Klebanov, L. and A. Yakovlev (2007). "How high is the level of technical noise in microarray data?" Biol Direct 2: 9.

Kreil, D. P. and D. J. MacKay (2003). "Reproducibility assessment of independent component analysis of expression ratios from DNA microarrays." Comp Funct Genomics 4(3): 300-17.

Křížek, P., J. Kittler, et al. (2007). Improving Stability of Feature Selection Methods. Computer Analysis of Images and Patterns: 929-936.

Kurokawa, H. and C. L. Arteaga (2001). "Inhibition of erbB receptor (HER) tyrosine kinases as a strategy to abrogate antiestrogen resistance in human breast cancer."Clin Cancer Res 7(12 Suppl): 4436s-4442s; discussion 4411s-4412s.

Kushner, P.J., et al. (2000) Estrogen receptor pathways to AP-1, The Journal of steroid biochemistry and molecular biology, 74, 311-317.

Ladroue, C., et al. (2009) Beyond element-wise interactions: identifying complex interactions in biological processes, PloS one, 4, e6899.

Lander, A. D. (2010). "The edges of understanding." BMC Biol 8: 40.

Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature 409(6822): 860-921.

Lee, E., H. Y. Chuang, et al. (2008). "Inferring pathway activity toward precise disease classification." PLoS Comput Biol 4(11): e1000217.

Lee, S. I. and S. Batzoglou (2003). "Application of independent component analysis to microarrays." Genome Biol 4(11): R76.

Lee, T.-W., M. Girolami, et al. (2000). "A unifying information-theoretic framework for independent component analysis." Computers & Mathematics with Applications 39(11): 1-21.

Lee, T. I., N. J. Rinaldi, et al. (2002). "Transcriptional regulatory networks in Saccharomyces cerevisiae." Science 298(5594): 799-804.

Liao, J. C., R. Boscolo, et al. (2003). "Network component analysis: reconstruction of regulatory signals in biological systems." Proc Natl Acad Sci U S A 100(26): 15522-7.

Libermann, T. A. and L. F. Zerbini (2006). "Targeting transcription factors for cancer gene therapy." Curr Gene Ther 6(1): 17-33.

Liebermeister, W. (2002). "Linear modes of gene expression determined by independent component analysis." Bioinformatics 18(1): 51-60.

Liu, M., A. Liberzon, et al. (2007). "Network-based analysis of affected biological processes in type 2 diabetes models." PLoS Genet 3(6): e96.

Loi, S., B. Haibe-Kains, et al. (2007). "Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade." J Clin Oncol 25(10): 1239-46.

Mithani, A., Preston, G.M. and Hein, J. (2009) Rahnuma: hypergraph-based tool for metabolic pathway prediction and network comparison, Bioinformatics, 25, 1831-1832. Ronen, M., et al. (2002) Assigning numbers to the arrows: parameterizing a gene regulation

network by using accurate expression kinetics, Proc Natl Acad Sci U S A, 99, 10555-10560.

Manoli, T., N. Gretz, et al. (2006). "Group testing for pathway analysis improves comparability of different microarray datasets." Bioinformatics 22(20): 2500-6.

Maslov, S., K. Sneppen, et al. (2004). "Upstream plasticity and downstream robustness in evolution of molecular networks." BMC Evol Biol 4: 9.

Mathivanan, S., et al. (2006) An evaluation of human protein-protein interaction data in the public domain, BMC Bioinformatics, 7 Suppl 5, S19.

Matys, V., O. V. Kel-Margoulis, et al. (2006). "TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes." Nucleic Acids Res 34(Database issue): D108-10.

Meinshausen, N. and P. Bhlmann (2010). "Stability selection." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72(4): 417-473.

Metzker, M. L. (2009). "Sequencing technologies - the next generation." Nat Rev Genet 11(1): 31-46.

Nebert, D. W. (2002). "Transcription factors and cancer: an overview." Toxicology 181-182: 131-41.

Newton, M. A., C. M. Kendziorski, et al. (2001). "On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data." J Comput Biol 8(1): 37-52.

Nguyen, D. H. and P. D'Haeseleer (2006). "Deciphering principles of transcription regulation in eukaryotic genomes." Mol Syst Biol 2: 2006 0012.

Parra, L. and P. Sajda (2003). "Blind source separation via generalized eigenvalue decomposition." J. Mach. Learn. Res. 4: 1261-1269.

Prall, O. W., E. M. Rogan, et al. (1998). "Estrogen regulation of cell cycle progression in breast cancer cells." J Steroid Biochem Mol Biol 65(1-6): 169-74.

Qiu, Y. Q., S. Zhang, et al. (2010). "Detecting disease associated modules and prioritizing active genes based on high throughput data." BMC Bioinformatics 11: 26.

Quackenbush, J. (2002). "Microarray data normalization and transformation." Nat Genet 32 Suppl: 496-501.

Ronen, M., et al. (2002) Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics, Proc Natl Acad Sci U S A, 99, 10555-10560.

Ruschhaupt, M., W. Huber, et al. (2004). "A compendium to ensure computational reproducibility in high-dimensional classification tasks." Stat Appl Genet Mol Biol 3: Article37.

Scott, J., et al. (2006) Efficient algorithms for detecting signaling pathways in protein interaction networks, Journal of computational biology : a journal of computational molecular cell biology, 13, 133-144.

Scholz, M., S. Gatzek, et al. (2004). "Metabolite fingerprinting: detecting biological features by independent component analysis." Bioinformatics 20(15): 2447-54.

Segal, E., M. Shapira, et al. (2003). "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data." Nat Genet 34(2): 166-76.

Sharan, R., I. Ulitsky, et al. (2007). "Network-based prediction of protein function." Mol Syst Biol 3: 88.

Shi, J. and J. Malik (2000). "Normalized Cuts and Image Segmentation." Pattern Analysis and Machine Intelligence, IEEE Transactions on 22(8): 888-905.

Shmulevich, I., E. R. Dougherty, et al. (2002). "Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks." Bioinformatics 18(2): 261-74.

Shoemaker, B. A. and A. R. Panchenko (2007). "Deciphering protein-protein interactions. Part I. Experimental techniques and databases." PLoS Comput Biol 3(3): e42.

Shoemaker, B. A. and A. R. Panchenko (2007). "Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners." PLoS

Comput Biol 3(4): e43.

Shou, J., S. Massarweh, et al. (2004). "Mechanisms of tamoxifen resistance: increased estrogen receptor-HER2/neu cross-talk in ER/HER2-positive breast cancer." J Natl Cancer Inst 96(12): 926-35.

Singh, M., G. Setalo, Jr., et al. (1999). "Estrogen-induced activation of mitogen-activated protein kinase in cerebral cortical explants: convergence of estrogen and neurotrophin signaling pathways." J Neurosci 19(4): 1179-88.

Slansky, J. E. and P. J. Farnham (1996). "Introduction to the E2F family: protein structure and gene regulation." Curr Top Microbiol Immunol 208: 1-30.

Smoot, M. E., K. Ono, et al. (2011). "Cytoscape 2.8: new features for data integration and network visualization." Bioinformatics 27(3): 431-2.

Sobhanifar, S. (2003). "The yeast two-hybrid assay: an exercise in experimental eloquence." The science creative quarterly(2).

Spellman, P. T., G. Sherlock, et al. (1998). "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization." Mol Biol Cell 9(12): 3273-97.

Stafford, P. and C. Yidong (2007). "Expression technology - A review of the performance and interpretation of expression microarrays." Signal Processing Magazine, IEEE 24(1): 18-26.

Storey, J. D. and R. Tibshirani (2003). "Statistical significance for genomewide studies." Proc Natl Acad Sci U S A 100(16): 9440-5.

Storey, J. D., W. Xiao, et al. (2005). "Significance analysis of time course microarray experiments." Proc Natl Acad Sci U S A 102(36): 12837-42.

Subramanian, A., P. Tamayo, et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proc Natl Acad Sci U S A 102(43): 15545-50.

Tao, W., H. Jin, et al. (2007). "Color image segmentation based on mean shift and normalized cuts." IEEE Trans Syst Man Cybern B Cybern 37(5): 1382-9.

Taylor, I. W., R. Linding, et al. (2009). "Dynamic modularity in protein interaction networks predicts breast cancer outcome." Nat Biotechnol 27(2): 199-204.

TCGA-Research-Network (2008). "Comprehensive genomic characterization defines human glioblastoma genes and core pathways." Nature 455(7216): 1061-8.

Tian, Z., Hwang, T. and Kuang, R. (2009) A hypergraph-based learning algorithm for classifying gene expression and arrayCGH data with prior knowledge, Bioinformatics, 25, 2831-2838.

Tilman, L., R. Volker, et al. (2004). "Stability-based validation of clustering solutions." Neural Comput. 16(6): 1299-1323.

Tornow, S. and H. W. Mewes (2003). "Functional modules by relating protein interaction networks and gene expression." Nucleic Acids Res 31(21): 6283-9.

Tusher, V. G., R. Tibshirani, et al. (2001). "Significance analysis of microarrays applied to the ionizing radiation response." Proc Natl Acad Sci U S A 98(9): 5116-21.

Ulitsky, I. and R. Shamir (2007). "Identification of functional modules using network topology and high-throughput data." BMC Syst Biol 1: 8.

Vastrik, I., P. D'Eustachio, et al. (2007). "Reactome: a knowledge base of biologic pathways and processes." Genome Biol 8(3): R39.

Venter, J. C. (2010). "Multiple personal genomes await." Nature 464(7289): 676-7.

Vigario, R., J. Sarela, et al. (2000). "Independent component approach to the analysis of EEG and MEG recordings." IEEE Trans Biomed Eng 47(5): 589-93.

von Luxburg, U. (2007). "A tutorial on spectral clustering." Statistics and Computing 17(4): 395-416.

Wang, C., J. Xuan, et al. (2008). "Motif-directed network component analysis for regulatory network inference." BMC Bioinformatics 9, Suppl (S1):S21.

Wu, J., L. T. Smith, et al. (2006). "ChIP-chip comes of age for genome-wide functional analysis." Cancer Res 66(14): 6899-902.

Wu, X., R. Jiang, et al. (2008). "Network-based global inference of human disease genes." Mol Syst Biol 4: 189.

Yang, Y. L., J. Suen, et al. (2005). "Inferring yeast cell cycle regulators and interactions using transcription factor activities." BMC Genomics 6(1): 90.

Ye, C., S. J. Galbraith, et al. (2009). "Using network component analysis to dissect regulatory networks mediated by transcription factors in yeast." PLoS Comput Biol 5(3): e1000311.

Yeger-Lotem, E., L. Riva, et al. (2009). "Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity." Nat Genet 41(3): 316-23.

Zhao, X. M., R. S. Wang, et al. (2008). "Uncovering signal transduction networks from high-throughput data by integer linear programming." Nucleic Acids Res 36(9): e48.