Novel Monte Carlo Approaches to Identify Aberrant Pathways in Cancer

Jinghua Gu

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 and Computer Engineering

Jason J. Xuan
Yue Wang
Luiz A. DaSilva
A. Lynn Abbott
Chang-Tien Lu

Aug. 2nd, 2013
Arlington, Virginia

Keywords: Microarray Data Analysis, Transcriptional Regulatory Network
Protein-Protein-Interaction Network, Aberrant Signaling Pathway
Markov Chain Monte Carlo, Gibbs Sampling

Novel Monte Carlo Approaches to Identify Aberrant Pathways in Cancer

Jinghua Gu

ABSTRACT

Recent breakthroughs in high-throughput biotechnology have promoted the integration of multi-platform data to investigate signal transduction pathways within a cell. In order to model complicated dynamics and heterogeneity of biological pathways, sophisticated computational models are needed to address unique properties of both the biological hypothesis and the data. In this dissertation work, we have proposed and developed methods using Markov Chain Monte Carlo (MCMC) techniques to solve complex modeling problems in human cancer research by integrating multi-platform data. We focus on two research topics: 1) identification of transcriptional regulatory networks and 2) uncovering of aberrant intracellular signal transduction pathways.

We propose a robust method, called GibbsOS, to identify condition specific gene regulatory patterns between transcription factors and their target genes. A Gibbs sampler is employed to sample target genes from the marginal function of outlier sum of regression t statistic. Numerical simulation has demonstrated significant performance improvement of GibbsOS over existing methods against noise and false positive connections in binding data. We have applied GibbsOS to breast cancer cell line datasets and identified condition specific regulatory rewiring in human breast cancer.

We also propose a novel method, namely Gibbs sampler to Infer Signal Transduction (GIST), to detect aberrant pathways that are highly associated with biological phenotypes or clinical information. By converting predefined potential functions into a Gibbs distribution, GIST estimates edge directions by learning the distribution of linear signaling pathway structures. Through the sampling process, the algorithm is able to infer signal transduction directions which are jointly determined by both gene expression and network topology. We demonstrate the advantage of the proposed algorithms on simulation data with respect to different settings of noise level in gene expression and false-positive connections in protein-protein interaction (PPI) network.

Another major contribution of the dissertation work is that we have improved traditional perspective towards understanding aberrant signal transductions by further investigating structural linkage of signaling pathways. We develop a method called Structural Organization to Uncover pathway Landscape (SOUL), which emphasizes on modularized pathways structures from reconstructed pathway landscape. GIST and SOUL provide a very unique angle to computationally model alternative pathways and pathway crosstalk. The proposed new methods can bring insight to drug discovery research by targeting nodal proteins that oversee multiple signaling pathways, rather than treating individual pathways separately. A complete pathway identification protocol, namely Infer Modularization of PAthway CrossTalk (IMPACT), is developed to bridge downstream regulatory networks with upstream signaling cascades. We have applied IMPACT to breast cancer treated patient datasets to investigate how estrogen receptor (ER) signaling pathways are related to drug resistance. The identified pathway proteins from patient datasets are well supported by breast cancer cell line models. We hypothesize from computational results that HSP90AA1 protein is an important nodal protein that oversees multiple signaling pathways to drive drug resistance. Cell viability analysis has supported our hypothesis by showing a significant decrease in viability of endocrine resistant cells compared with non-resistant cells when 17-AAG (a drug that inhibits HSP90AA1) is applied.

We believe that this dissertation work not only offers novel computational tools towards understanding complicated biological problems, but more importantly, it provides a valuable paradigm where systems biology connects data with hypotheses using computational modeling. Initial success of using microarray datasets to study endocrine resistance in breast cancer has shed light on translating results from high throughput datasets to biological discoveries in complicated human disease studies. As the next generation biotechnology becomes more cost-effective, the power of the proposed methods to untangle complicated aberrant signaling rewiring and pathway crosstalk will be finally unleashed.

Acknowledgments

I owe a debt of gratitude to the people who have helped me with the completion of the dissertation work.

I am most grateful to my advisor and mentor, Dr. Jason Xuan, for his constant guidance and encouragement in my research and life. Dr. Xuan has set up a stage for me to grow and thrive ever since I joined his team. Not only did he bring me exiting research topics, but also educate me the right mindset to face success and failure during my journey of scientific exploration. His unique perspective in identifying signaling networks related to endocrine resistance inspired my idea of aberrant pathway identification, which finally becomes the backbone of the dissertation work. The major contributions of the dissertation work, which is to develop useful computational tools that facilitate biological discovery in cancer research, cannot be made without the help and effort from Dr. Xuan.

I would like to thank Dr. Joseph Wang for his great insight and rigorous examination of the dissertation work. Dr. Wang has encouraged me to integrate prior knowledge for pathway identification. He has raised many valuable comments to help improve the quality of dissertation work. As the director of CBIL, Dr. Wang has provided us with many opportunities to communicate with top researchers from all over the world.

I am also grateful to Dr. Chang-Tien Lu, Dr. Luiz DaSilva and Dr. Lynn Abbott, who have served as my committee members for Ph.D dissertation research. They have provided me with very useful comments during my preliminary exam to improve the completeness and presentation of the research work.

I also want to express my gratitude to our collaborators at Georgetown University and Johns Hopkins University, including Dr. Robert Clarke, Dr. Leena Hilakivi-Clarke, Dr. Ie-Ming Shih, Dr. Tian-Li Wang, Dr Rebecca Riggins, Dr. Ayesha Shajahan-Haq, and Diane Demas. They have always been our best ally and leader in the war against cancer. Their expertise in biology, hardworking and devotion, open-mindedness and belief in teamwork have promoted the successful marriage between biologists and engineers. I sincerely thank our collaborators to devote their time, energy and resources

to biological validation experiments, which has been a strong support of the effectiveness of our computational methods.

I also want to thank people in CBIL, including all faculties, current students and alumni, for building such a great environment to conduct my dissertation work. I especially benefited from discussions with Dr. Guoqiang Yu, Dr. Li Chen, Dr. Bai Zhang and Dr. Chen Wang for their generous inputs and inspiring suggestions. It has been a wonderful experience to be part of the big family, which makes unique and memorable notes in the symphony of life.

I particularly want to thank my family, including my parents and my parents in law, for their support and trust over the past four years. Even if they are far away on the other side of the planet, their encouragement and consideration has always made my workspace a little warmer and cozier.

Finally, I gave my last but deepest thankfulness to my wife, Ye Chen, for every minute of hers to accompany me in my lab on weekends, for all her dream vacations that I have ruined, for all her faith in my research work that she even does not understand, for every dinner, for all her love, and for everything else.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AP Average Precision

AUC Area under the ROC Curve

CDF Cumulative Density Function

cDNA complementary DeoxyriboNucleic Acid

ChIP Chromatin ImmunoPrecipitation

CN condition number

COGRIM Clustering of Genes into Regulons using Integrated Modeling

CRMs cis-Regulatory Modules

DAVID Database for Annotation, Visualization and Integrated Discovery

DIP Database of Interacting Protein

DMFS Distant-Metastasis-Free-Survival

DNA DeoxyriboNucleic Acid

E2 Estradiol

EO Edge Orientation

ER+ Estrogen Receptor positive

FAS Faslodex

FastNCA Fast Network Component Analysis

FDR False Discovery Rate

GESA Gene Set Enrichment Analysis

GG Gamma-Gamma

GibbsOS Gibbs sampler on Outlier Sum statistic

GIST Gibbs sampler to Infer Signal Transduction

GO Gene Ontology

GRAM Genetic Regulatory Modules

HPRD Human Protein Reference Database

HSP Heat Shock Protein

ILP Integer Linear Programming

IMPACT Infer Modularization of Pathway CrossTalk

IPA Ingenuity Pathway Analysis

KEGG Kyoto Encyclopedia of Genes and Genomes

LARS Least Angle Regression

LTED Long-Term Estrogen-Deprived

MCMC Markov Chain Monte Carlo

mRNA messenger RiboNucleic Acid

NCA Network Component Analysis

OS outlier sum statistic

OSRT outlier sum of regression t-statistic

PARADIGM PAthway Recognition Algorithm using Data Integration on Genomic Models

PCA Principal Component Analysis

PDF Probability Density Function

PFC Percentage of false connections

PFP Proportion of False Positives

Plier Probe Logarithmic Intensity ERror

PPI Protein-Protein Interaction

ROC Receiver Operating Characteristics

SAM Significance Analysis of Microarray

siRNA small interfering RNA (siRNA)

SNR Signal to Noise Ratio

SNR Signal-to-Noise Ratio

SOM Self-Organizing Maps

SOUL Structural Organization to Uncover pathway Landscape

STRING Search Tool for the Retrieval of Interacting Genes/Proteins

SVM Support Vector Machine

SVR Support Vector Regression

SynTReN Synthetic Transcriptional Regulatory Networks

TAM Tamoxifen

TF Transcription Factor

TFA Transcription Factor Activity

TFBS Transcription Factor Binding Site

TRN Transcriptional Regulatory Network

1. Background and Introduction

1.1. Motivations

Systems biology [1, 2] is an inter-disciplinary field of life science that focuses on obtaining, integrating and analyzing large-scale and complex biological data from different platforms, including genomics, epigenetics, transcriptomics, proteomics, etc. The accomplishment in profiling large datasets have provided researchers with broad scope and great depth to understand the mechanism of biological systems, such as cancer cells, from a system point of view [3, 4]. To extract biologically meaningful information from gigabits or even terabits of data in hundreds of samples, computational scientists are devoted to building mathematical models, simulating biological systems and analyzing huge genomic datasets [5]. Due to overwhelming complexity of human disease, there arises unprecedented demand of effective and robust computational tools to integrate information from multi-platform datasets.

In the field of systems biology, one of the most attractive yet most challenging application is cancer [6]: to understand the mechanisms of different cancer types and to develop new therapies to treat cancer patients. Despite traditional chemotherapy and radiation treatment, or even the emerging trend of personalized medicine [7], our weapons against cancer are far from sufficient. Recently, tremendous efforts have been made by The Cancer Genome Atlas [8, 9] project to profile and sequence a large number of tumor samples in various cancer types, including lung cancer, breast cancer, ovarian cancer, prostate cancer, glioblastoma multiform, etc. In addition to rapid growth in sample size, multi-platform data have presented comprehensive sketch of the cancer systems in both genomic and non-genomic levels. This offers us great opportunity to study the molecular mechanism of cancer by integrating information from different components inside cancer cells.

Accompanying the growing trend to apply high-throughput biotechnology in cancer studies, a tremendous effort in data collection, annotation and analysis has been

conducted. Numerous useful tools and databases are available through the Internet [10-15]. However, these open resources usually provide generic information of gene/protein functions and/or their interactions, which is usually not specific to particular cell types or environmental conditions. Direct data mining based on databases/literature may hardly lead to novel discoveries because we know that cancer cells are heterogeneous and condition-specific. Therefore we propose to integrate high-throughput, multi-platform datasets with known biological knowledge that comes from existing databases or literature. We term this paradigm as 'condition-specific modeling' of cancer system, through which we seek to refine our understanding towards how cancer originates, progresses and evolves at the molecular level.

1.2. Background and data sources

High throughput datasets used in cancer research are acquired at different levels including DNA level, mRNA level and protein level. Recently, new data types such as small molecule RNA (miRNA) have drawn increasing attention [16, 17]. At DNA level, accomplishment of human genome sequence in 2001 is a significant step towards the exploration of human biology at molecular level. By sequencing human genome, we are able to identify diseases associated single-nucleotide polymorphisms (SNPs) [18], mutations or copy number variations [19-21]. We can also identify epigenetic changes (DNA methylation) [22, 23] in the promoter regions of tumor suppressor genes. At RNA level, profiling gene expression using microarray has been in dominant place in the last ten years [24, 25]. A great number of studies have been focused on the association between gene expression changes and phenotype information, such as survival information or clinical subtypes [26]. Typically when time-course data are available, we can use differential equations to model the change of mRNA concentrations that reveals how biological components are wired together in human cells [27, 28]. As the next-generation sequencing technique becomes more cost-effective [28, 29], we are capable of studying the entire transcriptome in a much higher resolution. Deep sequencing techniques enable us to study gene regulation at exon or transcript level where the discovery of novel splice junctions and gene fusion also becomes possible [30, 31]. For

protein data, which often referred to as 'proteomics' [32], are designed for a large-scale study of proteins' functionality and structure. The emerging of high-throughput proteomics data fills up the gap between transcription and translation, which is an important complementary component to gene expression data that provides us with means to study post-translational modifications such as phosphorylation and ubiquitination.

In addition to the above-mentioned data at DNA, mRNA and protein levels, we also use another type of data that investigate interactions between different levels of molecules, such as protein-DNA interaction and protein-protein interaction, etc. ChIP-on-chip experiments (Chromatin Immunoprecipitation on microarray chip) [33] and ChIP-Seq (ChIP sequencing) [34] are designed to identify the interactions between proteins (e.g., transcription factors) and DNA in vivo. Due to that high-throughput ChIP experiments are so far not completely available for Homo sapiens, we have to resort to information from the sequence level, namely Cis-regulatory module (CRM). CRM is a short DNA region (usually within thousands of bases) that a transcription factor will bind on through some unique Cis-regulatory element (transcription factor binding site) to regulate the expressions of nearby genes [35, 36].

Protein-protein interaction (PPI) data measure the possibility of two proteins that bind together to fulfill certain biological functions, which can be measured using techniques such as yeast two-hybrid screening [37]. The latest version of PPI data for human species can be downloaded at databases such as Human Protein Reference Database (HPRD) or the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database [38, 39].

Moreover, there are a number of databases and tools available online, which collect knowledge from literature or biological experiments. They are very useful resources to interpret computational results, which may be further translated into new biological discoveries. For functional annotation, Saccharomices Genome Database [40] has provided comprehensive information on budding yeast. Database for Annotation, Visualization and Integrated Discovery (DAVID) [41, 42] offers comprehensive functional annotation for human molecules and their interactions. The Kyoto Encyclopedia of Genes and Genomes (KEGG) [43] database embraces a great number of

canonical pathways, which is a good tool for initial validation of computationally recovered signal transductions from data. Figure 1.1 shows three types of interaction data: protein-DNA interaction, protein-protein interaction and pathway interaction.

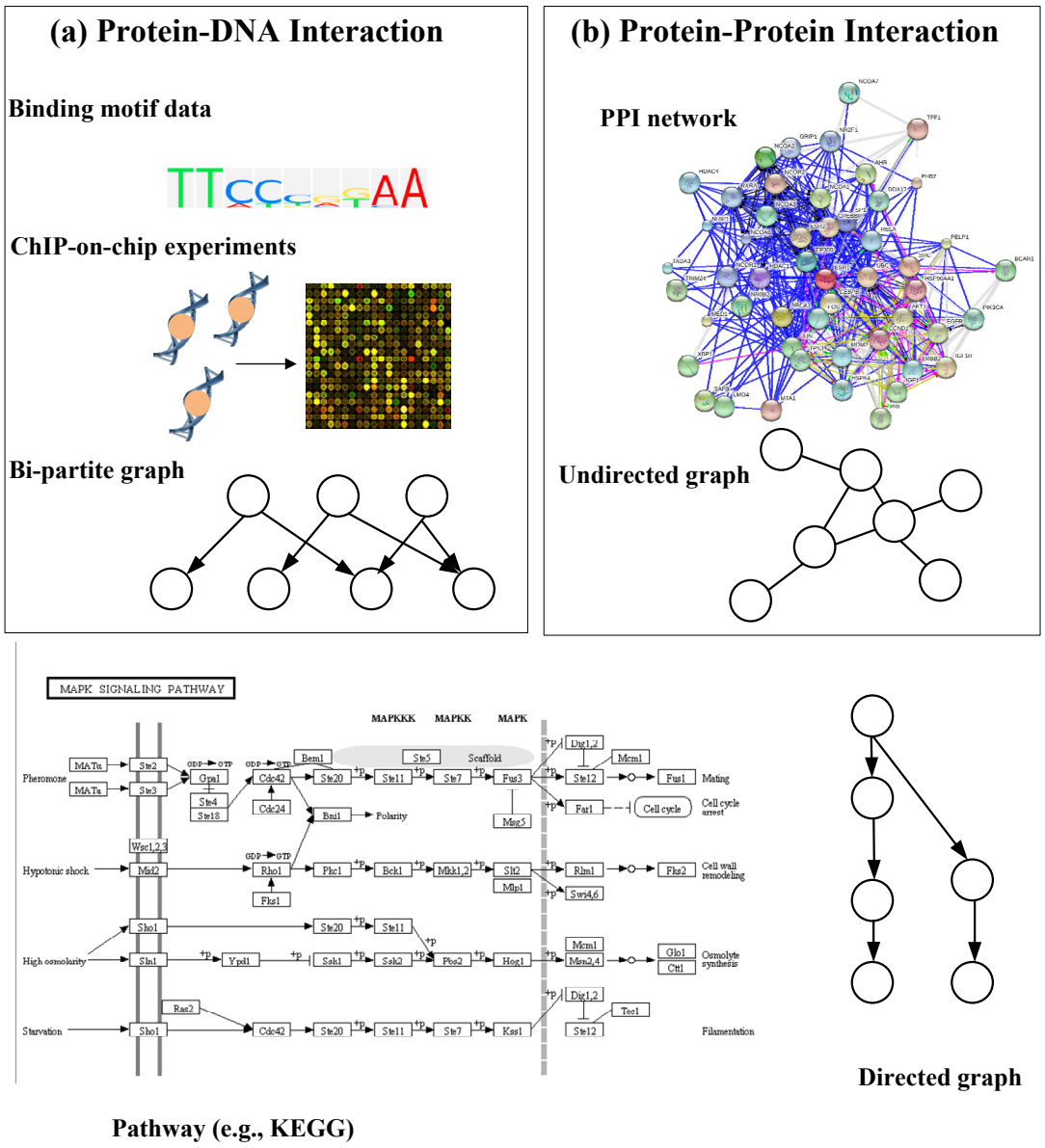


Figure 1.1. Biological interactions and their graphical representations. (a) Protein-DNA interaction from ChIP-on-chip experiments or binding motif information can be modeled as bipartite network. (b) Protein-protein interaction network is equivalent to un-directed graph. (c) Pathway interaction, e.g., canonical yeast MAPK pathways from KEGG database, can be represented as directed graph. (microarray image is from: http://bio.cse.ohio-state.edu/MicroarrayDesigner/ PPI figure is from: http://string-db.org/newstring_cgi/show_network_section.pl. Pathway figure is from: http://www.genome.jp/kegg-bin/show_pathway?map04010).

Each of the interaction data type in Figure 1.1 can be easily translated into mathematical language using graph theory. Protein-DNA interactions, such as interaction from binding motif data or ChIP-on-chip data, can be modeled as bipartite graphs that consist of transcription factors and their target genes [44]. PPI networks differ from pathway networks in that the former have no directional information while the later are directed graphs [45, 46].

In this study, we focus on integrating microarray gene expression, protein-DNA interaction, protein-protein interaction, and other biological knowledge to understand aberrant signal transductions in human cancer systems. The main objective of the dissertation work is through integrative modeling of information from different platforms, to develop novel and useful computational tools to generate new biological hypotheses.

1.3. A paradigm of integrative analysis: bridging components in cells from multi-platform genomic data

Systems biology emphasizes studying the dynamics and interactions within a cell from multiple experimental sources, which range from genomic sequencing, transcriptional regulation to protein function and structure, and post-translational modifications, etc. All components play cooperative roles to maintain regular functionality of a normal cell. When a perturbation occurs, such as an extracellular stimulus is introduced to the system, a sequence of reactions will be triggered in different biological counterparts. For example, binding of endocrine treatment drugs (such as Tamoxifen and Faslodex) to the estrogen receptor (ESR1) in the membrane will inhibit its activity, which further impacts its function as a transcription factor when it translocates into the nucleus to bind to estrogen response elements.

However, stories become more complicated when coming to human cancer studies due to the followings: 1) Traditional differential gene-based analysis methods have limited power in revealing true biological insights about the cancer genome because most differential genes tend to be downstream (the ‘end product’) of aberrant signal transduction. 2) Some important genes in biological networks are termed as ‘hub genes’,

which interact with a group of other genes or bridge different functional modules. They may influence the whole system through mechanisms that cannot be well captured by gene expression data, but rather through post-translational modifications such as phosphorylation. 3) Heterogeneity and complexity of the cancer genome determine the diversity and dynamics in cancer signal transduction. There have been increasing awareness that most cancer diseases are not driven by single or a few signaling pathways, but rather through the crosstalk of multiple signal transduction components. This calls for the effort of reconstructing the 'blue print' of cancer interactome through integrating high throughput data with biological knowledge. With this purpose in mind, we propose a framework, namely Infer Modularization of PAthway CrosTalk (IMPACT), for integrative analysis to bridge cellular components using data from multiple platforms (Figure 1.2).

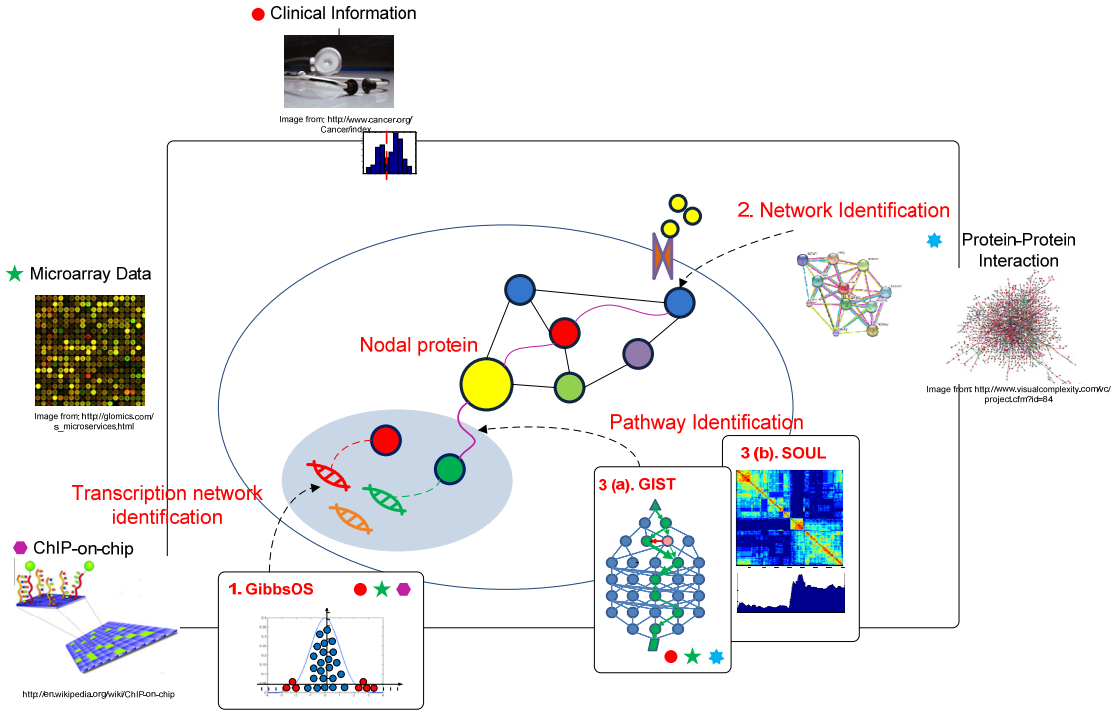


Figure 1.2. IMPACT framework for integrative analysis of multi-platform genomic data, which consists of four main methods: 1. GibbsOS for regulatory network identification; 2. PPI subnetwork identification; 3. GIST for identifying aberrant signal transduction and 4. SOUL for pathway modularization and crosstalk.

In Figure1.2, we study the complex biological system using multiple data resources, including microarray gene expression data, protein-protein interaction data, protein-DNA data (ChIP-on-chip data or motif binding data) and clinical information (e.g., distant

metastasis free survival). We concentrate on investigating how a cell responds to perturbations of the system including external stimulus (e.g., estrogen deprivation of breast cancer tumor). We seek to use gene expression to measure whether a target gene is hereby activated by a transcription factor or whether two proteins are possibly interacting with each other as a response.

Instead of perceiving complicated dynamics and interactions within a cell from a gene-based angle, we study the molecules in a system perspective, such as networks, modules or pathways [47]. There are three types of networks within our research scope: transcriptional regulatory networks (TRNs), protein-protein interaction (PPI) networks and pathway networks. Transcriptional regulatory networks include interactions among transcription factors and their target genes. To investigate the rewiring of transcriptional regulation, we have developed a method namely 'GibbsOS' to identify activated transcription factors and their down-stream targets. Protein-DNA interaction data from either ChIP-on-chip experiments or motif binding matrix are used as prior knowledge to build the initial network, which are further trimmed by gene expression to elicit condition-specific interactions. PPI network, as the name implies, represents the interaction among proteins. PPI serves as the bridge to connect proteins from different cellular compartments. Much work has been done to assign functional/biological interpretations to different sub-components of the entire PPI, which known as subnetworks. For PPI network identification, Wang and Chen [48, 49] have proposed novel methods to study protein-protein interactions at either global or local level.

A fundamental question that remains unaddressed by either TRN or PPI subnetwork identification is the causality among network components. In other words, to understand how signals are transduced from one component to another in the form of 'pathways', holds the key to crack the complicated biological riddle. The term 'pathway network' emphasizes the tangling and crosstalk among multiple singling pathways in contrast to traditional understanding that single canonical pathway determines the disease phenotype.

In this dissertation research, we propose a novel pathway identification method called Gibbs sampler to Infer Signal Transduction (GIST) to connect components in different parts of a cell by integrating gene expression data and protein-protein interaction data. Based on the topology of PPI network or additional information from cellular

locations, we aim to infer the direction of condition-specific protein-protein interactions. After the GIST algorithm has prioritized thousands of signaling pathways through data integration, we further investigate the interaction of pathways using method called Structural Organization to Uncover pathway Landscape (SOUL), which identifies critical proteins (known as 'nodal proteins') as crossroad among multiple pathways.

The entire IMPACT protocol utilizes the abovementioned four computational methods to identify aberrant signal transduction. IMPACT connects stimuli from extracellular space to proteome in different cellular compartments, and finally anchors at the nucleic transcriptome that implement disease phenotypes. Unlike traditional pathway identification methods, we are not merely looking at one or several disease related pathways but rather investigate the interaction and crosstalk among multiple functional pathway modules from a global angle.

1.4. Objective and statement of the problem

In this dissertation, we develop computational methods to model signal transduction in cancer systems by integrating multi-platform data. Our major focuses of this dissertation research are: 1) Develop statistical methods for robust identification of condition-specific transcriptional regulatory networks to study the rewiring between transcription factors and their down-stream targets. 2) Develop computational methods to infer signal transduction and pathway crosstalk in response to carcinogens or drug treatment. Due to the complexity of the problem (as discussed in Section 1.3) and inevitable noise contamination from biological experiments, it is very difficult to establish accurate mathematical models for cancer systems. Therefore we propose to use powerful statistical methods, i.e., Markov Chain Monte Carlo (MCMC) methods [50-52], to draw a comprehensive picture of intracellular dynamics and interactions by integrating data from multiple sources.

1.4.1. Robust identification of transcriptional regulatory networks using Gibbs sampler on outlier sum statistic

Transcription regulation has been an important topic in systems biology. Central dogma of molecular biology implies that by measuring the gene expression profile should

we get a general picture of proteins' activities, which are products of translation. With the advance of high–throughput microarray experiments and the increasing ability to investigate the interactions between proteins and DNA using ChIP-on-chip techniques, we can recover activated transcriptional regulatory networks within a biological system. Recent break-through in next-generation sequencing techniques and ChIP-Seq provide researchers with more reliable, higher resolution genomic data.

In the past a few years, many algorithms have been developed to model regulatory networks [53, 54]. Despite the notable progress achieved in revealing yeast's regulatory networks, the problem to study gene regulation for human cancer diseases remains challenging: 1) unlike the yeast case, ChIP-on-chip data are so far not completely available for human. We have to use information about the sequence of promoter regions as prior knowledge for possible protein-DNA interactions. 2) The protein-DNA binding data contain a lot of false information [55]. They are generic information which may not truly reflect the potential of binding under specific conditions. 3) In human cancer studies, data are usually contaminated by noise from biological experiments. We don't want to accept noise as 'knowledge'[56]. 4) Most biological processes are nonlinear so that we need to be mindful about the scalability and robustness of applying any linear methods to real biological problems [57]. 5) Transcription factors can work co-operatively, which requires joint estimate of individuals' activities. To tackle the above challenges, we propose a novel method namely Gibbs sampler on outlier sum statistic (GibbsOS) to identify condition-specific regulatory modules for human cancer research.

1.4.2. Investigating aberrant pathway modularization and crosstalk through data integration by IMPACT

In Figure 1.2, we have introduced a general framework of the proposed IMPACT protocol to identify aberrant signal transduction and pathway crosstalk by integrating multi-platform data from cancer genome. IMPACT is a pathway identification protocol that builds upon prerequisite methodologies for regulatory network identification and PPI subnetwork identification. The basic idea of IMPACT is to further uncover directed signal transduction flows among given source proteins (e.g., receptors in the membrane) and target proteins (e.g., transcription factors in the nucleus) based on undirected PPI

subnetworks. By integrating multi-platform data with knowledge, IMPACT provides a usual mean to evaluate signal transduction directions as well as the interactions among multiple pathway modules. The IMPACT protocol has two major components: Gibbs sampler to Infer Signal Transduction (GIST) and Structural Organization to Uncover pathway Landscape (SOUL). GIST is a sampling based method to identify directed linear signaling pathways from PPI and gene expression data. The identified pathways by GIST consist of protein-protein interactions that are highly associated with phenotype. On the other hand, SOUL further investigates structural interactions among elicited pathways from GIST by exploring pathway landscape. Proteins of topological importance to connect different pathway modules are proposed as nodal proteins, which may be potential targets for cancer therapy.

1.4.3. GIST: detecting aberrant signal transduction pathways from high-throughput data

Understanding how intracellular components respond to external stimuli is very important for cancer research, such as to study the mechanism of endocrine resistance in breast cancer treatment [58-60]. Large-scale analysis of protein-protein interaction (PPI) network sets a good base for identifying pathways among specific proteins like membrane proteins and transcription factors. Unlike the case in yeast studies that the protein-protein interaction data have quite confident scores derived from yeast two-hybrid screening experiments [61], PPI data for human contain a lot of false positives and false negatives connections. In order to screen those interactions on human cancer data, it is very important that we perceive the problem from a condition-specific point of view. We need to re-weight the nodes or edges in PPI network and establish new computational models that take into consideration gene expression change, phenotype information and PPI topology. To distinguish from network identification, pathway identification is more focused on recovering possible signal transduction directions between the start and end nodes. More specifically, it is of significant importance yet very challenging to infer the directions of protein-protein interactions from undirected PPI network in the context of human cancer research. To fulfill the above purpose, we propose a novel algorithm called Gibbs sampler to Infer Signal Transduction (GIST), to detect aberrant pathways that are highly associated with biological phenotypes or clinical information.

1.4.4. SOUL: exploring pathway landscape to study multiple pathway crosstalk and pathway modularization

Latest researches on cancer pathways have revealed increasing evidence that human cancer is complicated disease that cannot be explained by individual pathways, but rather the interaction among multiple pathways [62]. One major challenge about identifying multiple signal transduction pathways from PPI data is that human PPI contains more than 9,000 nodes and 60,000 interactions, which yields a tremendous state space even for pathways of moderate length (<9). Recent advance in drug targeting study has proposed the concept of 'nodal proteins', which are the crossroad of multiple signaling pathways. Successful examples of nodal proteins include heat shock proteins and survivins, which supervise both cell division and apoptosis pathways [62, 63]. Knockdown of these topologically important proteins will affect the activity of multiple functional pathways. Therefore, developing computational methods to identify potential nodal proteins may hold the key to pathway-targeted therapy in cancer. We develop a method, namely Structure Organization to Uncover pathway Landscape (SOUL), to further analyze sampled aberrant signaling pathways from GIST algorithm. By clustering pathways according to their structural profile, we group structural similar pathways with large potentials into pathway modules. Proteins that work as inter-modular hubs are computationally identified as nodal proteins due to their potential importance in signal transduction among multiple pathway components.

1.5. Summary of contributions

In this dissertation, we focus on developing computational methods to integrate multi-platform data in order to investigate important biological problems in human cancer studies. We summarize the potential contributions as follows:

(1) We propose a computational method to robustly identify gene regulatory networks using outlier sum statistics. For each transcription factor, we use a few candidate genes from biological knowledge and further evaluate their role in transcriptional regulation using expression data. The proposed algorithm aims to filter out false-positive interactions between the transcription factors and their targets from prior

knowledge so that we will be able to identify the rewiring of regulatory networks in a condition-specific manner.

(2) We develop the GIST algorithm to identify intracellular signal transduction pathways based on PPI network and gene expression data. A Gibbs strategy is employed so that we can examine the possible signal transduction flow at different scales. We can estimate the joint distribution of the pathway configurations by sampling from the conditional distribution, which can reduce the computational cost compared to exhaustive search. The GIST algorithm also provides a feasible way to infer signal transduction directions from undirected PPI network.

(3) A novel computational method, namely SOUL, is proposed, which focuses on the interaction among multiple signaling pathway modules instead of studying individual pathways independently. The development of SOUL has been a timing effort, which echoes the latest advance to study pathway crosstalk in signaling pathway networks for drug target research. SOUL dissects and regularizes sampled pathways from GIST algorithm to elicit topologically important nodal proteins as potential therapeutic target.

(4) We propose to use Markov Chain Monte Carlo (MCMC) methods to solve difficult modeling problems in cancer research. MCMC offers an effective means to solve complicated models when analytical solution is not available. In this dissertation, we have developed sophisticated computational models to address the complexity of biological problems and imperfection of biological data. The computational models defined over categorical random variables (e.g., genes or proteins), are constrained by networks derived from either protein-DNA interaction data or protein-protein interaction data. Therefore the models that we have proposed are not convex in general, which means that we do not have effective searching algorithms to obtain the optimal solutions. Moreover, we do not regard our problems as classic optimization problems, but rather distribution learning problems where global optimum and sub-optimum solutions are both very important. The MCMC methods are good alternatives for distribution learning, which provides a more robust and confident solution to the model. In our dissertation research, we use MCMC methods to solve two fundamental problems regarding regulatory network identification and pathway identification. In regulatory network identification, we use a Gibbs sampler to extensively evaluate the goodness of fitting

among the expression of seed genes and that of the candidate genes, which makes it possible to estimate the marginal distribution of outlier sum statistic based on samples from conditional distributions. In pathway identification, we use a Gibbs sampling strategy to sample pathway states with highest potentials from huge pathway state space, which are later used to estimate edge directions. Those suboptimal pathways with competing scores are referred to as 'alternative pathways'. The sampled pathways from the entire pathway space offers a snapshot of signaling rewiring in the global pathway landscape, which are further clustered for investigating pathway modularization and crosstalk by the proposed SOUL method.

1.6. List of relevant publications

Peer-reviewed Journal Publications

[1] **J. Gu**, J. Xuan, R. B. Riggins, L. Chen, Y. Wang and R. Clark, 'Robust identification of transcriptional regulatory networks using a Gibbs sampler on outlier sum statistic', ***Bioinformatics***, 28(15):1990-1997, 2012.

[2] IeM. Shih, L. Chen, C. Wang, **J. Gu**, B. Davidson, L. Cope, R. J. Kurman, J. Xuan, TL. Wang, 'Distinct DNA methylation profiles in ovarian serous neoplasms and their implications in ovarian carcinogenesis', ***Am J Obstet Gynecol***, 203(6):584.e1-22, 2010.

Manuscripts Submitted/In Preparation

[3] **J. Gu**, J. Xuan, A. N. Shajahan-Haq, D. M. Demas, Y. Wang and R. Clarke, 'IMPACT: Unraveling intracellular signal transduction and pathway crosstalk by exploring pathway landscape ', submitted to ***Genome Research***.

[4] **J. Gu**, J. Xuan, X.Shi, A. N. Shajahan, L. H. Clarke and R. Clarke, 'Revealing co-operative regulatory networks by probabilistic network component analysis', to be submitted to ***Bioinformatics***.

[5] **J. Gu** *et al.* 'IsoDiff: a novel Bayesian approach to identify differentially expressed isoforms from RNA-Seq data', in preparation.

Conference Publications

[1] **J. Gu**, J. Xuan, X. Wang, A. N. Shajahan, L. Hilakivi-Clarke, and R. Clarke, "Reconstructing transcriptional regulatory networks by probabilistic network component analysis," to appear in ***Proc. The fourth ACM International Conference on Bioinformatics, Computational Biology, and Biomedical Informatics (ACM BCB 2013)***, Washington, DC, Sept. 2013.

[2] X. Shi, **J. Gu**, X. Chen, A. N. Shajahan, L. Hilakivi-Clarke, R. Clarke, and J. Xuan, 'mAPC-GibbsOS: An integrated approach for robust identification of gene regulatory networks,' to appear in ***Proc. the International Conference on Intelligent Biology and Medicine (ICIBM 2013)***, Nashville, TN, Aug. 2013.

[3] X. Wang, **J. Gu**, L. Chen, A. N. Shajahan, R. Clarke, and J. Xuan, 'Sampling-based subnetwork identification from microarray data and protein-protein interaction network,' in ***Proc. of the 11th International Conf. on Machine Learning and Applications***, Boca Raton, Florida, Dec. 2012.

[4] L. Chen, J. Xuan, **J. Gu**, Y. Wang, Z. Zhang, TL. Wang and IeM. Shi, 'Integrative network analysis to identify aberrant pathway networks in ovarian cancer', in ***Proc. of Pacific Symposium on Biocomputing2012***, Big Island, Hawaii, Jan. 2012.

[5] **J. Gu**, J. Xuan, C. Wang, L. Chen, TL. Wang, IeM. Shih, ' Detecting aberrant signal transduction pathways from high-throughput data using GIST algorithm', in ***Proc. of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology***, San Diego, CA, May 2012.

[6] **J. Gu**, C. Wang, I. Shi, T. Wang, Y. Wang, R. Clark and J. Xuan, 'GIST: A Gibbs sampler to identify intracellular signal transduction pathways', in ***Proc. of the 33rd IEEE Engineering -in-Medicine-and-Biological-Society*** , Boston, MA, Aug. 2011.

[7] **J. Gu**, J. Xuan, Y. Wang, R. B. Riggins, R. Clarke, 'Identification of transcriptional regulatory networks by learning the marginal function of outlier sum statistic', in ***Proc. of the 9th International Conf. on Machine Learning and Applications***, Washington, DC, Dec. 2010.

1.7. Organization of the dissertation

The goal of this dissertation is to develop effective methods to identify aberrant pathways in cancer through data integration. As is described in Section 1.3, the major task of data integration at system level consists of three parts: regulatory network identification, subnetwork identification and pathway identification. In this work, we will focus on the first (regulatory network) and the third part (pathway) while previous researchers have addressed the second part (subnetwork) by developing several subnetwork identification methods based on integration of PPI data and microarray gene expression [37, 38]. The proposal is organized as follows:

In Chapter 2, we address the problem of regulatory network identification using a Gibbs sampler on outlier sum statistic (GibbsOS). We first summarize the potential challenges of the work such as noise in gene expression data and false-positives in protein-DNA data for human cancer studies. We then introduce our regulatory network identification method based on regression t-statistic. To evaluate whether target genes from a candidate pool are likely to have regulatory relationship with a given transcription factor, we introduce the outlier sum statistic based on which a Gibbs sampler is designed. We compare our method with competing methods on simulation data and yeast cell cycle data to demonstrate its efficacy. Synthetic expression datasets based on log-linear model and non-linear kinetic model are generated to test the robustness of GibbsOS under different signal-to-noise ratios (SNR). We have also simulated initial binding data with different false-positive connections to evaluate the proposed algorithm against structural error. Finally, we apply the GibbsOS algorithm on two breast cancer cell line datasets to identify several condition-specific regulatory networks related to different estrogen conditions.

In Chapter 3, we discuss a novel computational protocol, namely IMPACT, to identify intracellular signal transduction pathways and their crosstalk. The IMPACT protocol contains two main pathway identification methods: GIST and SOUL. First, we address our concern on existing methods of pathway identification. Next, we introduce in details about the GIST algorithm: 1) propose three potential functions of a pathway configuration by integrating gene expression data, clinical information and prior

knowledge about cellular locations. 2) Formulate a Gibbs distribution based on the potential functions and use a sampling technique to estimate the joint distribution of the pathway configurations. The joint distribution is the posterior distribution of the pathway configurations from the observed gene expression and knowledge. 3) Estimate edge directions of the PPI network from sampled pathways. To show the effectiveness of the proposed algorithm, we compare GIST with several existing methods using synthetic datasets. We evaluate the performance of GIST to identify pathway nodes and directed edges with competing methods under diverse experimental settings with different levels of expression noise and structural error in initial PPI.

After we have introduced the GIST algorithm, we further talk about the SOUL method that pools sampled pathways from GIST output to reconstruct pathway landscape. We introduce detailed procedures of the SOUL method including: calculate structural profile of sampled pathways, generate structural heatmap, reconstruct pathway landscape and identify nodal proteins.

Finally, we will apply the IMPACT pathway identification protocol to two breast cancer treated patient datasets from Loi and Symmans [64, 65]. We use cell line models from Georgetown University to validate computational results obtained from the patient datasets. We hypothesize that nodal proteins HSP90AA1 and a few other signaling hubs are related to breast cancer drug resistance or tumor progression, which are further validated using cell viability analysis.

In Chapter 4 we summarize the contribution of the dissertation research, discuss several important aspects for improvement and draw conclusion about the dissertation.

2. Robust Identification of Transcriptional Regulatory Networks Using a Gibbs Sampler on Outlier Sum Statistic

2.1. Introduction

Identification of transcriptional regulatory networks is essential to understanding the functional roles of molecules and the mechanisms of complicated biological processes. In cancer research, those identified regulatory networks, also referred to as regulatory modules, can give insight into important biological pathways that drive the development of disease, based on which novel drugs or therapies may be developed. In addition to high-throughput microarray techniques that profile the entire transcriptome of cells, more specific information about regulation can be obtained from binding motif data from DNA motif analysis or ChIP-on-chip experiments. Binding motif is also called transcription factor binding site, a segment of DNA regulatory sequence in the promoter region, to which a specific regulatory protein binds preferentially. ChIP-on-chip experiments can provide whole-genome analysis of binding site locations of all regulatory proteins. Both binding motif data and ChIP-on-chip data provide an initial structure of the unknown network topology. The availability of the above techniques has promoted extensive researches on regulatory networks in the last decade and some of which are quite successful in applications such as yeast studies.

Early exploration in gene modules depends heavily on statistical methods such as principal component analysis [66] and independent component analysis [67]. These statistical methods utilize information solely from gene expression data and have strong assumptions that the components are mutually orthogonal or statistically independent, which can hardly be satisfied in real biological systems. Other statistical methods include dynamic Bayesian networks [54] and probabilistic Boolean networks [68], but success of these methods is often limited by the relatively small number of tumor samples available for gene expression profiling in cancer studies. On the recent integrated use of gene

expression data and binding data such as ChIP-on-chip data, Liao *et al*. proposed a computational method called network component analysis (NCA), which is based on a log-linear model of the relationship between mRNA abundance and transcription factor activity [69]. Subjected to some mild identifiability conditions referred to as NCA criteria, the NCA method can be used to effectively estimate hidden transcription factor activities. Chang *et al*. later developed a fast version of the NCA method with a closed-form solution by using matrix factorization techniques [70].

Despite the above-mentioned promising progress made for gene regulatory module identification, two major concerns need to be further addressed. First, both microarray gene expression data and binding data are often corrupted by noises and the situation is even worse in real microarray data acquired from cancer patients. Noises in gene expression data will weaken the regulatory pattern, while noises in binding data will introduce false positive and false negative connections between transcript factors and their target genes. Second, the discrepancy between binding affinity and true regulation cannot be ignored. Recent studies show that transcription factors regulate their target genes in a condition-specific manner, i.e., the regulation is not only determined by the potential of binding strength but also depends on the environmental condition. Notably, Brynildsen *et al*. claimed that using random network information, instead of that from ChIP-on-chip data, can approximately yield the same amount of variation explaining changes in gene expression; this claim in certain extent echoes our concern that noises in binding data may have strong impact on transcript factor activity estimation [71]. In the study, they also showed that by selecting a subset of genes with high confidence only, a significant increase in the fitting performance can be achieved and more variations in gene expression data can be explained. However, Brynildsen *et al*. used 'condition number' as a metric to measure the consistency between regulators and their target genes, which is not effective when signal-to-noise ratio of the expression data is low, and hence limits its further applications to cancer studies .

As such, more robust methods are needed in order to correctly infer condition-specific gene regulatory modules from noisy data sources. In this paper, we propose a new Gibbs sampling-based method for target gene identification and regulatory network reconstruction. Our method is capable of reducing false positives contained in binding

data by iteratively selecting the highly confident genes from an initial pool. In particular, a novel statistic for testing the confidence of target genes, namely outlier sum of regression t-statistic, is specifically designed to pin-down confident target genes; based on this statistic a Gibbs sampling strategy is utilized to sample target genes in a high probability as governed by the underlying distribution. This Gibbs sampler on outlier sum statistic is an effective and robust approach to gene regulatory module identification because of the following features: (i) the regression t-statistic is a effective metric to statistically measure the dependency of genes and works particularly well when signal-to-noise ratio is low; (ii) the outlier sum of regression t-statistic reflects the aggregated evidence of the regulation between one transcription factor and its target genes; (iii) the Gibbs sampling scheme allows us to study the marginal confidence of the target genes, which may be regulated by multiple transcription factors; (4) in addition to target gene identification, a procedure of significance analysis is also incorporated into our scheme to test the statistical significance of transcription factors being studied.

Experimentally, we have tested the outlier sum-based Gibbs sampler on both simulation and yeast cell cycle data to demonstrate its effectiveness and efficacy. We compared the performance of our method with competing methods under experimental settings. The results showed that our method exhibited robust performance in reconstructing regulatory networks. Finally, breast cancer case study has demonstrated the feasibility of applying GibbsOS to reveal condition-specific regulatory rewiring.

2.2. Methods

Our Gibbs sampling method based on outlier sum statistics aims to reliably identify regulatory networks consisting of transcription factors and their target genes. Figure 2.1 shows the workflow of our proposed sampling-based framework. The Gibbs method starts from the initial topology (A) of the transcriptional regulatory network by selecting a pool of candidate target genes for each transcription factor. Regression analysis among the candidate gene expression X is performed and the regression t-statistic is then aggregated into an outlier sum. A Gibbs strategy is adopted so that we can sample genes according to their marginal outlier sum of regression t-statistic (OSRT). After the

sampled distribution is converged, both transcription factors and their target genes can be ranked for regulatory module reconstruction. Moreover, a statistical procedure based on outlier sum statistic is designed to test the significance of transcription factors, which can help us prioritize some important regulatory modules for real biological studies.

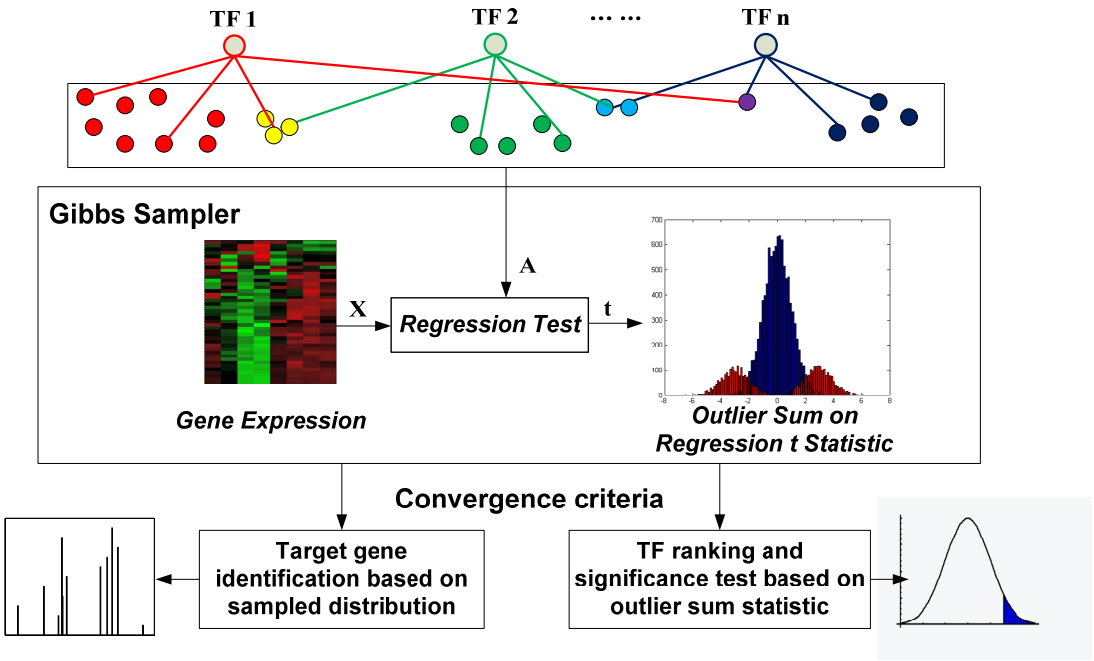


Figure 2.1. A general workflow of the Gibbs sampler on outlier sum statistic.

2.2.1. A log-linear model for regulatory network identification

The expression levels of genes in live cells are controlled by many activated regulatory proteins in the form of transcription factor activities (TFAs) as binding to the promoters of the genes. Liao *et al*. have modeled the relationship between gene expression level and transcription factor activity as a log-linear model which can be written in the following compact form [69] :

$$\mathbf{E} = \mathbf{AP}, \tag{2.1}$$

where $\mathbf{E}$ is an $N \times M$ gene expression matrix and $\mathbf{A}$ is an $N \times L$ connectivity matrix representing the regulation strength of a regulatory network. Here $\mathbf{P}$ (with a size of $L \times M$) is a matrix representing the unknown TFAs.

The log-linear model of transcription regulation depicts a bipartite network constituted by transcription factors and their target genes. An initial skeleton of the

network can be constructed from binding data like ChIP-on-chip data and motif information. However, due to the intrinsically noisy attribute of binding data and the gap between binding ability and true regulation, this initial skeleton or 'snap-shot' of the network contains a lot of false connections. In terms of one specific microarray data set acquired under certain biological or environmental condition, true target genes are referred to as 'foreground genes' in this paper; on the other hand, the term, 'background genes', refers to those genes that are irrelevant to true regulation but are included in the initial regulatory network (due to experimental noise or technological deficiency). It is reasonable to assume that the TFA estimation based on foreground genes is more reliable than that from the entire population of genes or a group of initial target genes contaminated by background genes. Hence identification of foreground genes becomes a crucial step for reliable reconstruction of gene regulatory modules. With a group of highly confident genes with strong regulatory evidence from their expression and binding data, we can assess the significance of any given regulatory module in a condition-specific manner, while gaining an improved estimation of the activity of transcript factors regulating the network.

2.2.2. Target gene identification based on regression model and outlier sum statistic

Due to the existence of background genes from binding data, the application of Equation (2.1) on a group of genes defined by initial binding data will often lead to a biased over-fitting of the data. As matter of fact, the log-linear model is applicable to only a group of foreground genes whose expression data carry information about the activity of their regulators. Suppose that we have L true target genes $[\theta_1, \theta_2, ..., \theta_L]$, each corresponding to at least one unique transcription factor. These L target genes are also referred to as 'seed genes' or 'seeds' for the L transcription factors. We can write the expression of any candidate gene as linear regression form in terms of the seed gene expressions as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}, \tag{2.2}$$

where $\mathbf{y}$ denotes expression of the candidate gene, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_L]$ is the expression of the L seed genes. $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_L]^{\mathrm{T}}$ is the scalar coefficient vector. Under some mild

assumptions, it can be derived that $\beta_i = 0$ indicates that gene *y* is not regulated by TF *i*. Hence, we can conduct a significance test on β_i, to determine whether a gene is regulated by TF *i* or not. The least square estimate of $\boldsymbol{\beta}$ [72] is given by $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{C}\mathbf{X}^{\mathrm{T}}\mathbf{y}$. Note that we also denote $\mathbf{C} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$ in the above equation. The hypotheses for testing the significance of any individual coefficient β_i are as follows:

$$H_0(\text{null hypothesis}): \beta_i = 0;\ H_1(\text{alternative hypotheisis}): \beta_i \neq 0.$$

A suitable test statistic for the above hypotheses is given by

$$t = \frac{\hat{\beta}_i}{S_b} = \frac{\hat{\beta}_i}{\sqrt{\frac{SSR}{M-L-1}\cdot C_{ii}}}, \tag{2.3}$$

where $SSR = \mathbf{y}^{\mathrm{T}}\mathbf{y} - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ (sum of squared residuals) and $M-L-1$ is the degree of freedom. C_{ii} is the i^{th} diagonal element of $\mathbf{C} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$. Under the null hypothesis the test statistic *t* follows Student's-t distribution with $M-L-1$ degrees of freedom (dofs). The null hypothesis is rejected if $|t| > t_{\alpha/2,M-L-1}$, where $t_{\alpha/2,M-L-1}$ denotes the statistic with a $100\times(1-\alpha)\%$ confidence interval. This means that if θ_i is a true target gene of TF *i*, the calculated test statistic *t* should fall into the rejection region (i.e., $|t| > t_{\alpha/2,M-L-1}$). On the contrary, if θ_i is a background gene (whose expression pattern is irrelevant to TFA *i*), the test statistic will remain in the acceptance region of the null hypothesis (i.e., $|t| \le t_{\alpha/2,M-L-1}$). Figure 2.2 illustrates how the test statistic *t* contributes to the identification of target genes for TF *i*.

The above discussion suggests how the quality of the seeds can be examined. Now suppose that we have a group of candidate genes for TF *i* denoted as set $\Theta_i = \{\theta_{i1},...\theta_{iK_i}\}$ (K_i possible target genes for TF *i* as reported by either ChIP-on-chip data or motif information data), among which several of the genes are true targets denoted as set Θ_{iF}; the rest of the genes in Θ_i are background genes, which have no regulation relationship with TF *i*, denoted as set Θ_{iB}; thus, $\Theta_i = \Theta_{iF} \cup \Theta_{iB}$. To start our regression analysis

procedure, we select a random gene $\theta_{ik} \in \Theta_i$ $(1 \le k \le K_i)$ as seed gene θ_i and carry out regression analysis between each single gene in Θ_i (except θ_{ik} itself) and the L seed genes $[\theta_1,...,\theta_{i-1},\theta_i=\theta_{ik},\theta_{i+1},...,\theta_L]$. Using the remaining candidate genes for TF i, we will obtain a total number of K_i-1 test statistics, t_{kj}, j=1…K_i, $j \neq k$. If $\theta_{ik} \in \Theta_{iF}$, because genes in Θ_{iF} are foreground genes, t_{kj} will be significantly larger than 0. This means that t_{kj} corresponding to foreground genes are outliers (represented by red dots in Figure 2.2) to the null distribution, which is a Student's-t distribution with $M-L-1$ degrees of freedom (denoted as $t(\nu = M-L-1)$ and shown as the blue curve (probability density function) in Figure 2.2). On the other hand, if $\theta_{ik} \in \Theta_{iB}$, all t_{kj} should follow the null distribution very well because a background gene should not have a strong regression relationship with either true target genes of TF i or other background genes.

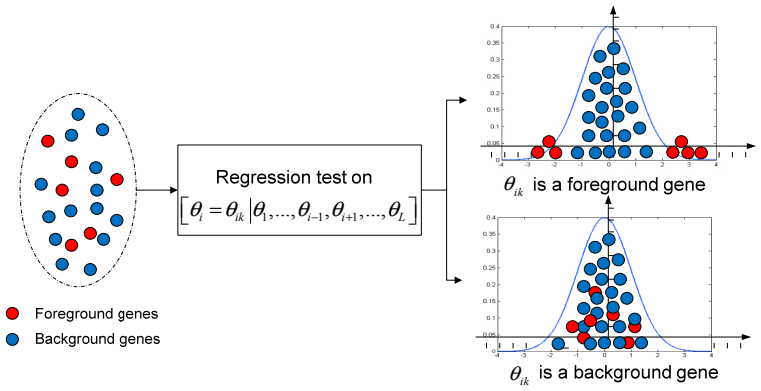


Figure 2.2.Target gene identification based on regression analysis.

In order to determine whether a seed gene for TF i is well selected (which means θ_{ik} is a true target gene) we propose to use the outlier sum statistic [73] of θ_{ik} as follows:

$$OS_{ik} = \sum_{j \neq k} t_{kj} \cdot I\left(t_{kj} - t_{\alpha/2,M-L-1}\right), \tag{2.4}$$

where $I(x) = \begin{cases} 1, & x \ge 0 \\ 0 & x < 0 \end{cases}$ is an indicator function. Equation (2.4) shows that for any selected seed θ_{ik} of TF i, we will sum up all t_{kj} that are 'outliers' to the null distribution using a threshold $t_{\alpha/2,M-L-1}$. If the selected θ_{ik} is a foreground gene, the corresponding outlier sum statistic should be significantly larger than 0. Contrarily, the outlier sum (i.e.,

OS_{ik}) will be close to 0 if θ_{ik} is a background gene. The outlier sum statistic represents a conditional test of the quality of a selected seed for TF *i*, given *L*-1 seed genes for other transcription factors, which will be further discussed in the next subsection.

2.2.3. A Gibbs sampler on the outlier sum statistic

The outlier sum statistic defined in the previous section is specifically designed to test the significance of a seed θ_i by evaluating its suitability to be a true target gene of TF *i* based on other seeds. Hence we can view the outlier sum statistic as a conditional function as follows:

$$OS\big|_{\theta_i} = f_{os}\left(\theta_i \,\middle|\, \theta_1,...,\theta_{i-1},\theta_{i+1},...,\theta_L\right), \tag{2.5}$$

where $f_{os}(\cdot)$ denotes the test function of outlier sum. From Equation (2.5) we can see that the power of the outlier sum statistic is influenced by the quality of the other seeds $\theta_j\,(j=1,...,L, j\neq i)$. If the conditional seeds for other transcription factors are contaminated by background genes, the performance of test statistic *t* from regression analysis may degrade, consequently weakening the power of the outlier sum statistic. Theoretically, the optimal solution to this identification problem can be obtained by maximizing the likelihood function based on the joint distribution of $OS\big|_{\theta_i}\,(i=1,...,L)$, i.e., $f_{os}\left(\theta_1,...,\theta_i,...,\theta_L\right)$. The optimality is underlined by the fact that we aim to simultaneously find a joint set of seed genes (i.e., $\theta_i(i=1,...,L)$) for *L* TFs, with which the likelihood function of $f_{os}\left(\theta_1,...,\theta_i,...,\theta_L\right)$ is maximized. However, as we all know, it is not a trivial task (often infeasible) to estimate the joint distribution in a high dimension space. In this paper, we opt to find a suboptimal solution to this problem by focusing on identification of target genes for one TF at a time, rather than on a joint set of target genes for all the TFs. Specifically, this suboptimal solution can be obtained based on the marginal distribution of seed genes for each TF, which can be estimated from its conditional distributions by a sampling scheme described later. In other words, the suboptimal solution to identification of target genes for TF *i* should be based on the

marginal analysis (in contrast to the conditional analysis) of the target genes of TF *i*, which does not depend on the selection of other seed genes.

In order to translate the conditional test statistic (i.e., Equation (2.5)) into a marginal one, we will employ a Gibbs sampling strategy that can be used to effectively estimate the marginal distribution from conditional distributions [74]. Now let us define a conditional probability density function as follows:

$$p\left(\theta_i \middle| \theta_1,...,\theta_{i-1},\theta_{i+1},...,\theta_L\right) = \frac{1}{K}\cdot f_{os}\left(\theta_i \middle| \theta_1,...,\theta_{i-1},\theta_{i+1},...,\theta_L\right). \qquad (2.6)$$

The probability distribution function p differs from f_{os} only in an unknown normalizing constant K. Now we have assigned each $\theta_i \in \Theta_i$ with a probability density conditioned on other seeds $\theta_j\,(j=1...K, j\neq i)$ and obtained the conditional distribution of θ_i. Based on the conditional probability density p, we can estimate the marginal distribution $p(\theta_i)$ by using a Gibbs sampling strategy as follows:

$$\begin{cases} \theta_1^{(t+1)} \sim p(\theta_1 \middle| \theta_2^{(t)},\theta_3^{(t)},...,\theta_L^{(t)}) \\ \theta_2^{(t+1)} \sim p(\theta_2 \middle| \theta_1^{(t+1)},\theta_3^{(t)},...,\theta_L^{(t)}) \\ \quad ..... \\ \theta_i^{(t+1)} \sim p(\theta_i \middle| \theta_1^{(t+1)},...,\theta_{i-1}^{(t+1)},\theta_{i+1}^{(t)},...,\theta_L^{(t)}) \\ \quad ..... \\ \theta_L^{(t+1)} \sim p(\theta_L \middle| \theta_1^{(t+1)},\theta_1^{(t+1)},...,\theta_{L-1}^{(t+1)}) \end{cases}, \qquad (2.7)$$

where t denotes the t^{th} step in the sampling iteration process. At iteration t, we will sequentially sample one seed gene for each TF based on the corresponding conditional probability density function. After we have sampled one gene for TF *i*, the seed for this transcription factor will be updated by the newest sample. Through a sufficient number of sampling iterations, we will have a group of samples that are drawn from the marginal distribution of seed genes for each TF. From Equation (2.6) we know that during the sampling procedure, we are also estimating the marginal function of the outlier sum statistic because it equals to the probability density function except a normalizing constant. This means that genes with larger marginal probability density have consistently larger marginal outlier sum. Hence we can approximate the empirical probability density function of $p(\theta_i)$ by calculating the frequency of the samples. The

candidate genes for θ_i with large frequency count are more likely the true target genes for TF *i*. Note that the proposed Gibbs sampling framework is very efficient since its convergence can be theoretically guaranteed.

Besides for target gene identification, one advantage of the proposed Gibbs sampling method is that it can facilitate to identify true (active) transcription factors. Outlier sum statistic is an index showing the possibility that one gene is a true target gene for given TF. If one TF is a background (inactive) TF that presumably does not regulate any target genes, the test statistic *t* will hardly be significant between any two of its candidate genes. As a result, the outlier sum of all the candidate genes should be close to zero. On the other hand, if one TF is a true regulator, the true target genes of this TF should support each other and yields large outlier sum. Hence the outlier sum based sampling method can prioritize both transcription factors and their corresponding target genes to reconstruct regulatory networks from gene expression and binding data.

2.2.4. Significance test of identified transcriptional regulatory modules

Our Gibbs sampling method is designed to extract a confident set of target genes through a probabilistic manner for condition-specific regulatory module identification. However when working on the real microarray data such as breast cancer data, we need strong statistical evidence to support and justify the identified regulatory modules before we can proceed to conduct costly biological experiments for validation. In order to address the concern that the results obtained from certain data set is not by chance, we propose a workflow to test the statistical significance of any given regulatory module based on the outlier sum distribution associated with the corresponding transcription factor. We summarize the procedure for testing the significance of the regulatory module of TF *i* as follows:

(1) Generate a null distribution of the regression t-statistic from the entire gene set being studied, denote as N_1.

(2) Suppose TF *i* has *n* candidate target genes in the initial pool. Generate the null distribution of the outlier sum statistic of size *n* based on N_1, denote as $N_{2,n}$.

(3) Test whether the sampled distribution for TF i has the same mean as the null distribution $N_{2,n}$. The statistical significance (i.e., p-value) is conventionally denoted as p_i. By definition, p_i is the probability that the mean of sampled outlier sum distribution is larger than that obtained from random selection of n genes for TF i in the population.

Through this significance test procedure, we are able to highlight the modules with strong evidence supporting regulation events under certain condition by both microarray data and binding information.

When applying our Gibbs sampling method to any real biological data, we need to be fully aware that the null distribution of regression t statistic is no longer centered at zero. This is because that a considerable number of genes may be correlated due to their roles in many common biological functions. Hence it is important to build the proper null distribution for a specific data set. We now explain the significance test on transcriptional regulatory modules for breast cancer cell line data in estrogen induced condition (early up-regulated group) with more details.

The regression t statistic is impacted by the number of independent variables (possible seed genes) that are involved in the regression analysis. Each time we randomly select one gene from the population and calculate the regression t statistic using l genes that are also randomly selected, where l follows the same distribution as the number of possible regulating TFs for one target gene obtained from ChIP-on-chip data or binding motif data (Figure 2.3).

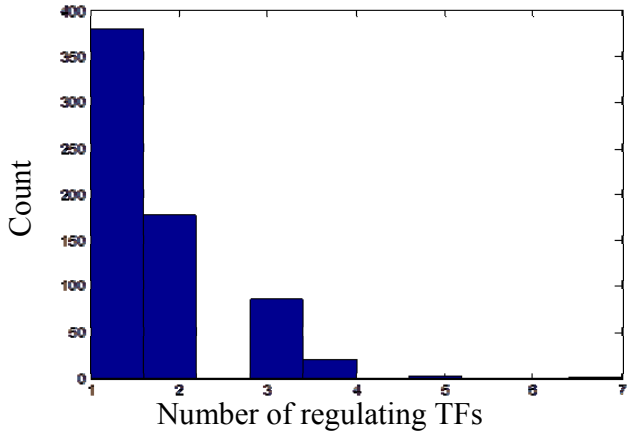


Figure 2.3. Distribution of the number of regulating TFs for one gene.

We conducted the above regression test for 10,000 times by randomly drawing genes from the population and obtain the distribution of the regression t statsitic N_1 as shown in Figure 2.4:

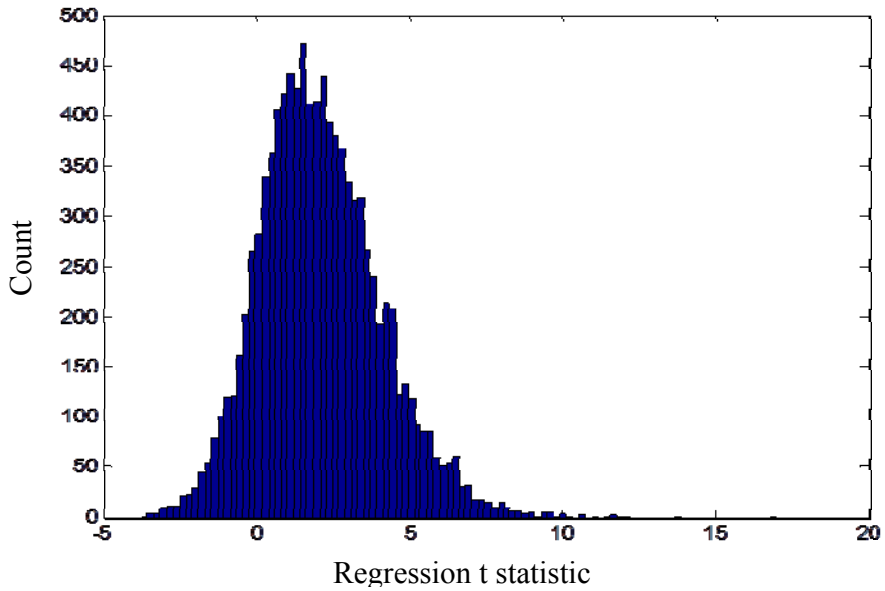


Figure 2.4. Null distribution of regression t statistic.

The above distribution has a mean value around 2, showing a high correlation of expression in the gene population. This is reasonable because the genes being studied all exhibit early up-regulation patterns. Having obtained the null distribution of regression t statistic, we can further estimate the null distribution of the outlier sum statistic. Note that the outlier sum statistic is defined as the sum of absolute value of regression t statistic that is above certain threshold (e.g., 1.96, corresponding to two-sided significance level of 0.05 in normal distribution), which is highly correlated with the number of candidate genes. Hence for TFs with different candidate pool sizes, ranging from 1 to 100, we generated 100 null distributions of the outlier sum statistic, denoted as $N_{2,n}$ where n=1,2,3,…100. We show the null distribution of the outlier sum statistic regarding the number of candidate genes in Figure 2.5.

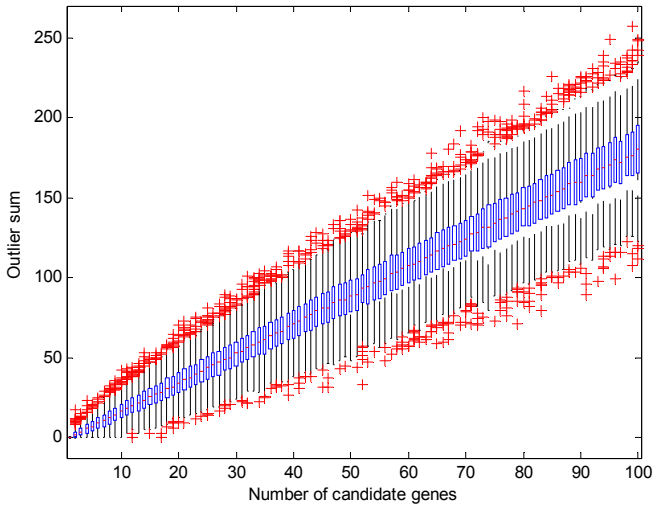
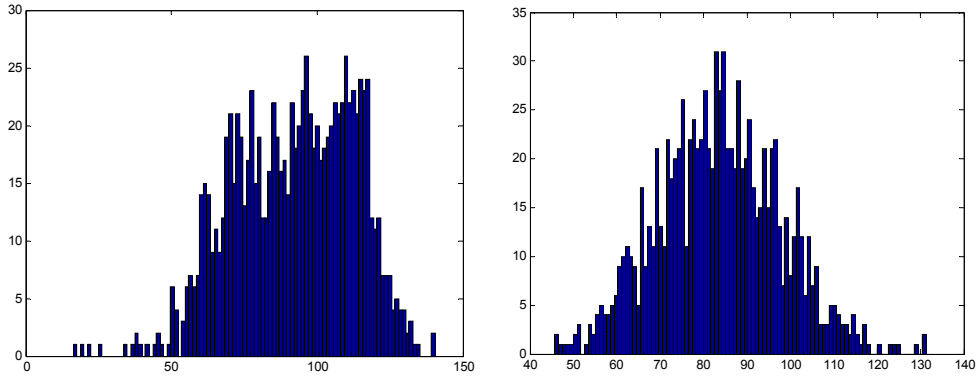


Figure 2.5. Box-plot of null distributions of outlier sum statistic with regard to different candidate gene number.

Having obtained the null distribution of the outlier sum statistic, we can test whether the activity of a given transcriptional regulatory module is significant or not. Let's take estrogen receptor (ESR1) as an example (Figure 2.6). Using 10 percentile cutoff of binding motif score, we obtain 48 possible targets for ER binding. After running the Gibbs sampler for 1,000 iterations, we obtained the distribution of outlier sum statistic associated with ER. We have also pre-calculated the null distribution of an arbitrary TF with 48 candidate genes, which is denoted as $N_{2,48}$.



(a) outlier sum distribution of ER (b) null distribution of outlier sum,$N_{2,48}$

Figure 2.6. Significant test for transcriptional regulatory module associated with motif ER. Note: x-axis – 'outlier sum'; y-axis – 'count'.

We tested whether the mean of the sampled outlier sum is larger than that of the null distribution $N_{2,n}$ using two sample student's t-test. For the ER motif, we obtained a p-

value of e-22, indicating very strong evidence that the target genes proposed by the binding motif data are more capable of explaining the uncertainties of each other than randomly pooled candidate genes.

2.2.5. Bridging network component analysis with regression model

The expression levels of genes in live cells are controlled by many activated regulatory proteins in the form of transcription factor activities (TFAs) as binding to the promoters of the genes. Liao *et al*. have modeled the relationship between gene expression level and transcription factor activity as a log-linear model defined by [69] :

$$\frac{E_i(t)}{E_i(0)} = \prod_{j=1}^{L} \left(\frac{TFA_j(t)}{TFA_j(0)} \right)^{CS_{ij}}, \tag{2.8}$$

where $E_i(t)$ is the expression level of gene i in the t^{th} sample, and $TFA_j(t)$ is the activity of transcription factor j in the t^{th} sample. CS_{ij} is the regulation strength of transcription factor j on gene i. If we take the logarithm of Equation (2.8), the regulatory model becomes linear and we can rewrite it in a matrix form. Suppose we have a total number of N genes with M samples regulated by L transcription factors, let's recall the linear equation introduced in Equation (2.1):

$$\mathbf{E} = \mathbf{AP},$$

where E is an $N \times M$ gene expression matrix and A is an $N \times L$ connectivity matrix representing the regulation strength of a regulatory network. Here P (with a size of $L \times M$) is a matrix representing the unknown TFAs.

Suppose that we have L true target genes $[\theta_1, \theta_2, ..., \theta_L]$, each corresponding to at least one unique transcription factor. These L target genes are also referred to as 'seed genes' or 'seeds' for the L transcription factors. Because we assume that all seed genes are foreground genes so that the log-linear relationship between TFA and gene expression profile should be satisfied for the L seeds. Thus we have:

$$\mathbf{E}_\theta = \mathbf{A}_\theta \mathbf{P}, \tag{2.9}$$

where E_θ is an $L \times M$ gene expression matrix and A_θ is an $L \times L$ connectivity matrix.

Now suppose that we have another candidate gene *y*, gene *y* is a foreground gene and TF *i* is one of its regulators. The expression of gene y satisfies

$$\mathbf{e}_y = \mathbf{aP}, \tag{2.10}$$

where $\mathbf{e}_y$ is a $1\times M$ gene expression vector and $\mathbf{a} = [a_1, a_2, ..., a_L]$ is a $1\times L$ connectivity vector. Here the physical meaning of **a** is the regulation (or binding) strength of *L* transcription factors over gene *y*. If $a_i = 0$, it means that gene *y* is not regulated by TF *i*. Because we have assumed that gene y is regulated by TF i, the corresponding a_i should be significantly different from 0. By taking advantage of the defined regulation relationship as Equation (2.9) we have

$$\mathbf{e}_y = \left(\mathbf{a}\cdot\mathbf{A}_\theta^{-1}\right)\cdot\mathbf{E}_\theta = \sum_{i=1}^{L}\beta_i\cdot\mathbf{e}_{\theta_i}, \tag{2.11}$$

provided that $\mathbf{A}_\theta$ is invertible. Here $\mathbf{e}_{\theta_i}$ is a $1\times M$ expression vector of seed θ_i and β_i is a scalar coefficient. As long as all the seed genes are regulated by at least one unique transcription factor, $\mathbf{A}_\theta$ is assured to be of full rank. Equation (2.11) shows that the expression of a foreground gene can be represented as the linear combination of the expressions of *L* seed genes. Without loss of generality, $\beta_i = 0$ indicates that gene *y* is not regulated by TF *i*. Hence, we can conduct a significance test on β_i, i.e., to test if $\beta_i = 0$, in order to determine whether a gene is regulated by TF *i* or not.

For Equation (2.11) to hold, we need to assume that $\mathbf{A}_\theta^{-1}$ is invertible. Justification of this assumption has been discussed in the paper by Brynildsen *et al*., 2006. The Gibbs sampler, as Brynildsen *et al*. claim, is designed to find the seed genes when $\mathbf{A_\theta}$ is invertible. However, they also realize that in many datasets such an invertible $\mathbf{A_\theta}$ may not exist, or in other words, there is no invertible matrix $\mathbf{A_\theta}$ that is formed by only foreground genes ('consistent genes' in Brynildsen's terminology)**.** Instead, some background genes ('error genes' in Brynildsen's terminology) will be populated during the sampling process. In this case, the Gibbs sampler will identify the genes that maximize the number of consistent genes [71].

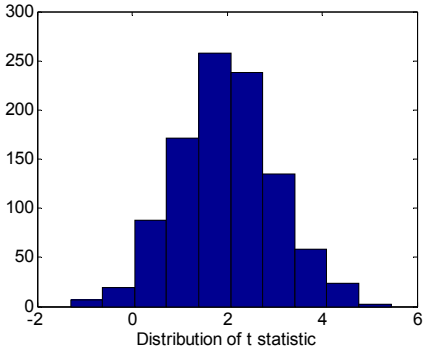
We consider Equation (2.11) as a conceptual explanation of the relationship between a potential target gene's expression and that of the selected seed genes. Actually, matrix inversion is not performed in our implementation. The ability of the GibbsOS method to identify true target genes does not totally rely on the 'invertible' assumption, which means that it is quite robust to underlying network topology. The robustness is inherited from the regression analysis that we proposed to evaluate the quality of the seed gene. Below we use an example to show the robustness of regression analysis against ill-conditioned (rank deficient) network topology.

Case 1: We have a small regulatory network with L=3 transcription factors. Assume that the current three seed genes gene 1, gene 2 and gene 3 have binding strength matrix **A** defined as:

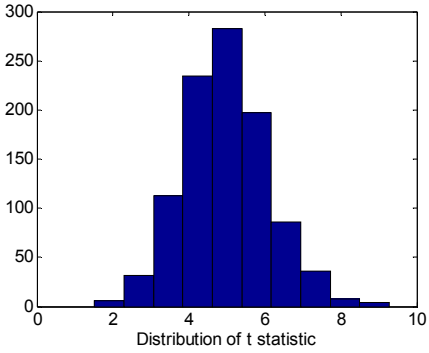
$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{matrix} \text{gene 1} \\ \text{gene 2} \\ \text{gene 3} \end{matrix},$$

where the i^{th} column denotes the i^{th} transcription factor. In this setting, matrix **A** is certainly not invertible. Now we want to evaluate whether the regression based method can still identify consistent target genes for the same transcription factor. We synthesize another foreground candidate gene, gene 4, for transcription factor 1 with binding matrix [1 0 0]. We regress the expression of gene 4 using gene 1's (current seed gene for TF 1) expression based on other seed genes (gene 2 and gene 3), and we calculate the corresponding regression t-statistic. If the t-statistic is larger than 1.96 (0.05 two-tailed significance level; 95% confidence level for normal distribution), it means that the current seed gene for TF 1 can be well supported by candidate gene 4.

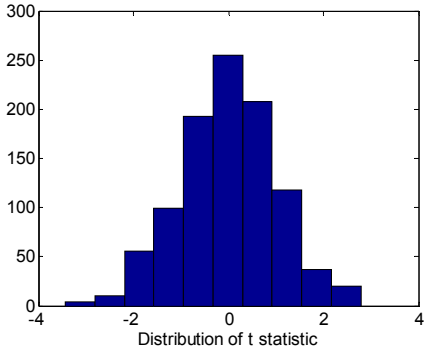
We use Matlab randn() function to generate a 3×M transcription factor activity matrix, where M=100 is the sample size. The gene expression is generated using linear equation E=AP+N, where N follows standard Gaussian distribution. We ran 1,000 random trials and the distribution of the t statistic for seed gene 1 is shown in Figure 2.7 (a). We then calculate the mean of t distribution in 1,000 trials, which is 2.0009.



(a) A is rank deficient



(b) A is full rank



(c) A is full rank but seed gene is a background gene

Figure 2.7. Impact of topology A on regression t-statistic. Note: x-axis – 't statistic'; y-axis – 'count'.

Case 2: We also tested the scenario when A is invertible. We performed a regression analysis of the gene expression of gene 4 using a different seed gene set: gene 1, gene 2 and gene 5 where gene 5 has the binding pattern of [0 0 1]. In this case, the binding matrix of seed genes,

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{matrix} \text{gene 1} \\ \text{gene 2} \\ \text{gene 5} \end{matrix},$$

is invertible. The distribution of the corresponding regression t-statistic is shown in Figure 2.7 (b) with a mean around 5, which is much larger than 1.96 (0.05 two-tailed significance level; 95% confidence level).

Case 3: For comparison, we also introduced a background gene, gene 6, with binding pattern [1 0 0]. We tested the significance of regression of gene 6 on seed gene 1 based on gene 2 and gene 5. The distribution of regression t-statistic is shown in Figure 2.7 (c) with 0 mean, which is not significant at all.

From the above study, we can summarize the summation results as follows:

(1) If A is invertible and the genes are all foreground genes, the regression analysis will generate a t-statistic larger than 1.96 in a probability of 0.998 (Figure 2.7 (b), Case 2).

(2) If A is not invertible but the genes are all foreground genes, the regression analysis will generate a t-statistic larger than 1.96 in a probability of 0.512 (Figure 2.7 (a), Case 1).

(3)If A is invertible but the candidate gene is a background gene, the regression analysis will generate a t-statistic larger than 1.96 in a probability of 0.021 (Figure 2.7 (c), Case 3).

Thus we can see that even when the 'invertible' assumption of matrix A cannot be satisfied, the regression t-statistic remains still capable of, in certain degree (with a 50% success rate in our simulation setting), detecting the consistency between the foreground genes of the same transcription factor, which suggests the robustness of the proposed method for reconstructing complex biological networks.

2.3 Results

2.3.1. Performance of regression test versus condition number against noise

As is shown in our simulation study, the outlier sum based Gibbs sampler significantly outperforms the condition number based sampling method. Especially when signal-to-noise ratio degrades to about 0 dB, the condition number based method almost performs like random guess while the outlier sum based method can maintain reasonable detection power. The robustness of our Gibbs sampler is inherited from the regression t statistic. Both regression t statistic and the condition number are measuring whether the expression of one candidate gene can be explained by linear combinations of a given set of seed genes. Assume that the candidate gene y has expression $\mathbf{e}_y$ and we also have L seed genes $[\theta_1, \theta_2, ..., \theta_L]$ whose expressions correspond to the rows of matrix $\mathbf{E}_\Theta$. To test whether the expression of gene y is a linear combination of expressions of the seed genes using the condition number method, we need to merge the expression of gene y and the

seed genes to get a new matrix called $\mathbf{E}$ where $\mathbf{E} = \begin{bmatrix} \mathbf{e}_y \\ \mathbf{E}_\Theta \end{bmatrix}$. The condition number is defined as

$$\kappa(\mathbf{E}) = \left| \frac{\lambda_{\max}(\mathbf{E})}{\lambda_{\min}(\mathbf{E})} \right|, \tag{2.12}$$

where $\lambda_{\max}(\mathbf{E})$ and $\lambda_{\min}(\mathbf{E})$ are the largest and smallest eigenvalues of matrix $\mathbf{E}$. If the condition number is large (infinity in ideal case), the matrix $\mathbf{E}$ is rank deficient indicating the uncertainties in gene *y* can be well explained by the expression of the seed genes. Otherwise, a full rank matrix $\mathbf{E}$ implies the expression of gene *y* is linearly independent of the expression of the seed genes.

We can also test the linear independency between the candidate gene y and seed genes from a regression point of view. Assume that the matrix $\mathbf{E}_\Theta$ has the form

$$\mathbf{E}_\Theta = \begin{bmatrix} \mathbf{e}_{\theta_1} \\ \mathbf{e}_{\theta_2} \\ \cdots \\ \mathbf{e}_{\theta_L} \end{bmatrix}. \tag{2.13}$$

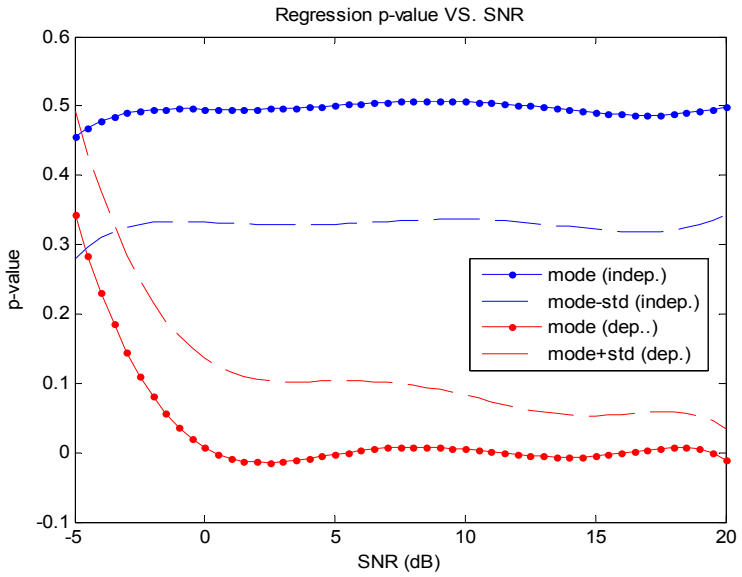
Recall the linear regression representation of the problem from Equation (2.11) as follows:

$$\mathbf{e}_y = \left(\mathbf{a} \cdot \mathbf{A}_\theta^{-1}\right) \cdot \mathbf{E}_\theta = \sum_{i=1}^{L} \beta_i \cdot \mathbf{e}_{\theta_i}.$$

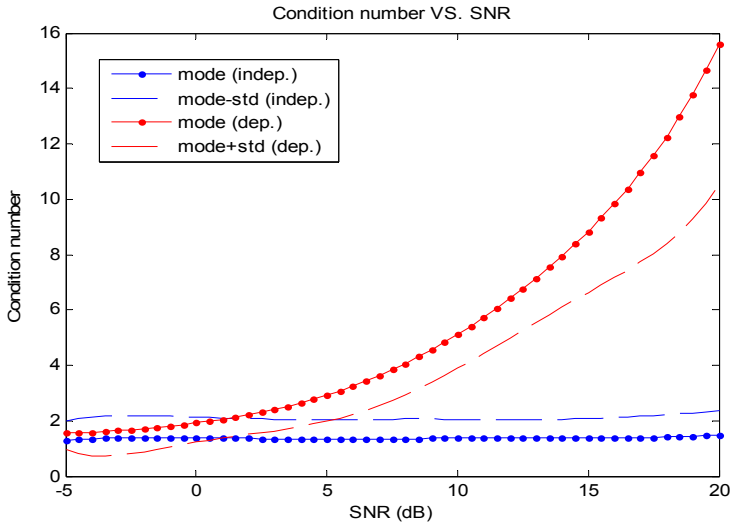
If the uncertainty carried by gene y can be explained by the seed gene θ_i, the regression t statistic associated with coefficient β_i should be significantly different from zero, yielding very significant regression p-values.

We now use simulation data to compare the performance of the two measurements in testing the linear independency of a group of genes. We generate the expression matrix $\mathbf{E}$ using linear equation $\mathbf{E} = \mathbf{AP}$ and add different levels of noise to the expression. As a control study, we also generated linear independent gene expression $\mathbf{E}$. Figure 2.8 shows

how the performance of the two measurements changes as signal-to-noise ratio increases from -5dB to 20dB in both independent and dependent cases.



(a) Testing robustness of regression t statistic in different noise conditions



(b) Testing robustness of condition number in different noise conditions

Figure 2.8. Performance of regression t statistic and condition number in testing linear independency of gene expression matrix **E.**

The curve in blue color in Figure 2.8 represents the linear independent case, and the curve in red color corresponds to the linear dependent case. The solid curve is the mode of the distribution and the dashed curve is one standard deviation away from the mode.

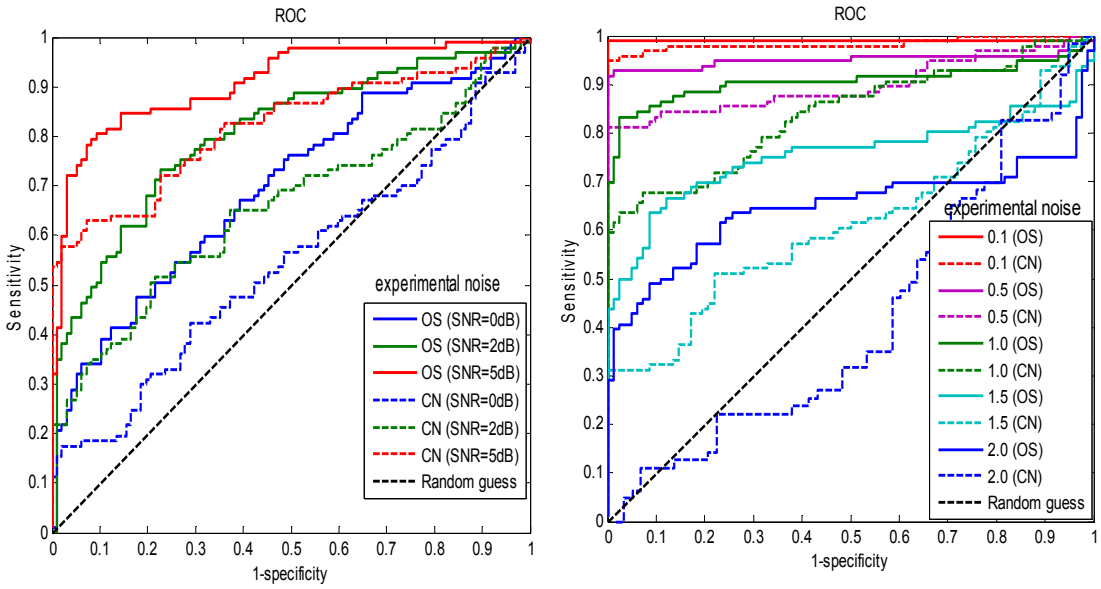
Figure 2.8 (b) shows that when SNR is high (>5 dB), the distribution of condition number in both linear independent and linear dependent cases are clearly separated, which guarantees its efficiency in distinguishing the two cases. As the SNR further degrades, the distributions of the condition number for the two cases start to overlap, hindering its detection of linear dependency of gene expression profiles. This explains why as SNR goes below 0 dB, the condition number based sampler totally behaves as random guess.

In contrast, the distributions of the regression p-values in linear independent and linear dependent cases are well separated from each other even when SNR decreases to 0 dB (Figure 2.8 (a)), demonstrating the robustness of the regression based method against the quality degradation in expression data. The robustness of the regression t statistic in testing linear dependency among gene expressions finally leads to the performance improvement of the outlier sum based Gibbs sampler for target gene identification.

2.3.2. Performance comparison between Gibbs samplers based on outlier sum statistic and condition number against experimental noise

We first tested our sampling method on simulation data to assess its performance for target gene identification. We used Matlab functions to synthesize an initial network consisting approximately 200 target genes each with 30 experiments and 20 transcription factors with linearly independent TFAs. Among all the target genes, half of them are foreground genes whose expression are generated according to Equation (1) to guarantee the linearity, while for the other half, namely background genes, we randomly generate their gene expression profiles for this experiment. We compared the performance of our proposed sampling method for target gene identification with an existing Gibbs sampling method that is based on condition number evaluation [71]. As demonstrated in the paper, the competing method outperforms traditional model-fitting methods (e.g. Ordinary Least Squares; Huber, Cauchy and Fair weighted robust regressions) by reducing the over-fitting to background genes. We evaluated the algorithm performance based on several simulated data sets with different signal-to-noise ratios (SNRs). Figure 2.9 (a) shows the receiver operating characteristic (ROC) curves for target gene identification using the two Gibbs samplers based on outlier sum statistic (OS) and condition number (CN), respectively.

From Figure 2.9 (a) we can see that for different signal-to-noise ratios (5dB, 2 dB and 0dB), our OS-based sampling method consistently outperforms the CN-based sampling method with a 0.1 increase in area-under-curve (AUC). When signal-to-noise ratio decreases to 0 dB, the CN-based sampling method performs slightly better than random guess (AUC = 0.5), while the OS-based sampler still maintains a reasonable performance with the AUC of 0.695.



(a) (b)

OS: outlier sum
CN: condition number

Figure 2.9. ROC comparison of Gibbs sampling method based on outlier sum statistic and condition number using linear simulation data and SynTReN. (a) Experiment on linear simulation. (b) Experiment on SynTReN data.

For a more realistic simulation of transcription regulatory net-works, we use the SynTReN software [75] to generate synthetic networks and gene expression profiles. SynTReN extracts sub-networks from known yeast networks and utilizes Michaelis-Menten and Hill kinetics to model the interaction kinetics. The main difference between the expression data generated by SynTReN and those by Equation (2.1) is that SynTReN models the expression of target genes as a nonlinear function of the expression of their regulators, which is a more realistic representation of biological regulation relationships than Equation (2.1). The network topology and expression data generated by SynTReN are better mimic of the real transcriptional regulatory networks (TRNs) than simple linear combinations. We generated a network consisting of 100 foreground genes and 100

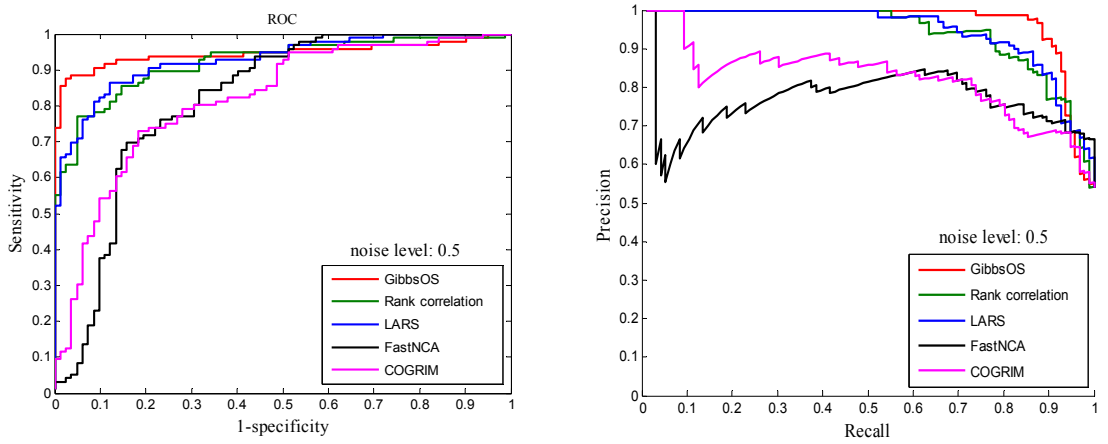
background genes regulated by 24 TFs in total. We also set different levels of experimental noise in order to test the robustness of our proposed method against the quality degradation of gene expression data. Figure 2.9 (b) shows the ROC curves in terms of target gene identification using our OS-based sampling method and CN-based sampling method. As the noise level increases, both methods suffer from performance degradation showing a corresponding decrease in the AUCs. How-ever under the same noise condition, our OS-based sampling method consistently performs better than the Gibbs sampler based on condition number with an average AUC increase of 0.1.

2.3.3. Performance comparison of multiple regulatory network identification methods under different noise conditions

We compared the performance of the GibbsOS method with four additional existing methods, including rank correlation [76], least angle regression (LARS) [77], FastNCA [70] and COGRIM [78], on SynTReN (Van den Bulcke, Van Leemput et al. 2006) simulated networks. Rank correlation test was studied as being more robust than other similarity-based methods when systematic noise level increases. Least angle regression (LARS) is a technique to solve linear regression problems with sparsity constraint, which particularly fits well with the context of regulatory network reconstruction. FastNCA is an efficient method for Network Component Analysis (NCA) by using eigenspace analysis, which is quite robust to expression noise and different network topologies. COCRIM used a Bayesian hierarchical model that integrated gene expression and ChIP-on-chip data to cluster genes into regulatory networks. In addition to using receiver operating characteristic (ROC) curve to measure the sensitivity versus specificity, we also provided the precision-recall (PR) curve [79] to show the precision of the algorithms in retrieving the true target genes. Figures 2.10-12 show the ROC curves and PR curves for target gene identification on synthetic data generated by SynTReN with different noise settings (noise level = 0.5, 1.0 and 1.5).

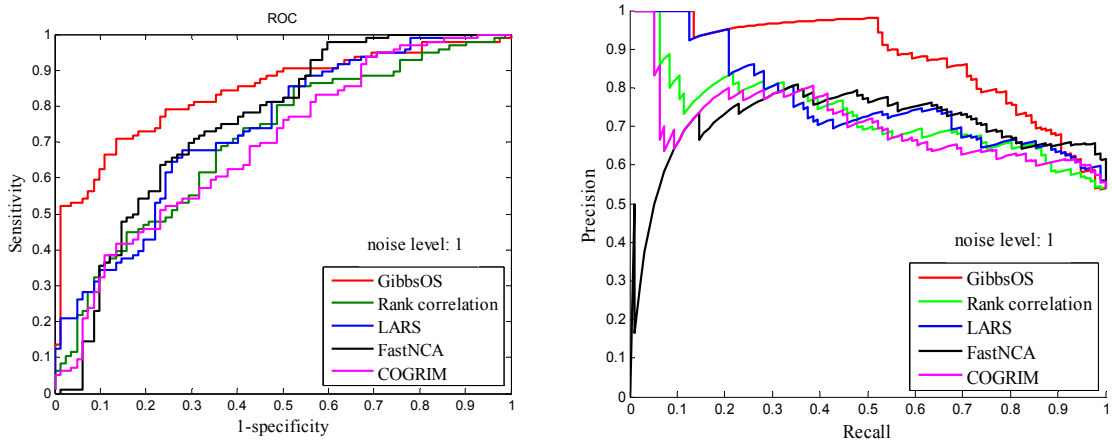
Table 2.1 also gives the area-under-the-curve (AUC) of each ROC curve. We also included average precision (AP) [79] in the table to indicate the proportion of false connections in total identified connections. The larger an AP value is, the fewer false connections are in the total identified connections. From Figures 2.10-12 and Table 2.1,

we observed that GibbsOS consistently outperformed other existing methods on synthetic yeast networks generated by SynTReN. More specifically, when noise level increased to 1.0, GibbsOS achieved the largest improvement compared to other methods by an increase in AUC of at least 0.08 (compared to FastNCA, which ranked second in terms of AUC) and an increase in AP of at least 0.11 (compared to LARS, which ranked second in terms of AP).



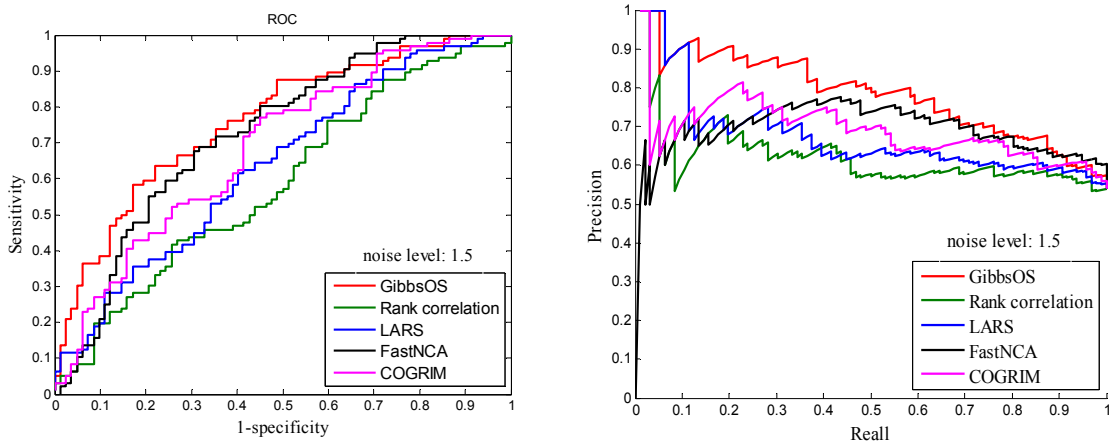
(a) (b)

Figure 2.10. Receiver operating characteristic (ROC) curve (a) and Precision-recall (PR) curve (b). Experimental noise level for SynTReN is set to 0.5. The simulated network consists of approximately 100 foreground genes and 100 background genes. The number of total active TFs is 10. The gene expression data are generated by SynTReN using a nonlinear model.



(a) (b)

Figure 2.11. Receiver operating characteristic (ROC) curve (a) and Precision-recall (PR) curve (b). Experimental noise level for SynTReN is set to 1. The simulated network is the same as the one described in Figure 2.10.



(a) (b)

Figure 2.12. Receiver operating characteristic (ROC) curve (a) and Precision-recall (PR) curve (b). Experimental noise level for SynTReN is set to 1.5. The simulated network is the same as the one described in Figure 2.10 and 2.11.

Table 2.1. Area-under-the-curve (AUC) of ROC and average precision (AP).

	Noise level = 0.5		Noise level = 1.0		Noise level = 1.5	
	AUC	AP	AUC	AP	AUC	AP
GibbsOS	0.9463	0.9687	0.8397	0.8776	0.7625	0.7863
Rank correlation	0.9212	0.9434	0.7048	0.7325	0.5877	0.6243
LARS	0.9289	0.9478	0.7362	0.7674	0.6388	0.6767
FastNCA	0.8186	0.7717	0.7532	0.7060	0.7271	0.6964
COGRIM	0.8181	0.8264	0.6921	0.7100	0.6874	0.6959

2.3.4. Performance comparison between Gibbs samplers based on outlier sum statistic and condition number against false positive connections

As mentioned in the introduction section, protein-DNA interaction data have imperfections with certain false positive and/or false negative rate. Here we particularly focus on addressing the problem of filtering out false positive connections as obtained from ChIP-on-chip or binding motif data. Specifically, we aim to separate foreground genes from background genes by integrating gene expression data and protein-DNA interaction data. A high false positive rate in ChIP-on-chip data means that in the initial candidate pool, a large number of candidate genes are background genes. We systematically studied the influence on the performance as the density and false positive rate in initial network topology varied in a certain range. We compared the performance of GibbsOS method with that of the condition number-based method for regulatory network identification.

Figure 2.13 shows the ROC comparison result on simulated networks with different false positive rates. Note that a parameter, nf:nb, is used to control the proportion of

foreground genes contained in the initial candidate pool. To quantify the false positives, we also define the term 'proportion of false positives' (PFPs) as follows:

$$\text{PFP}=\frac{\text{number of background genes}}{\text{number of background genes + number of foreground genes}},$$

which reflects how many background genes are introduced in the simulated network.

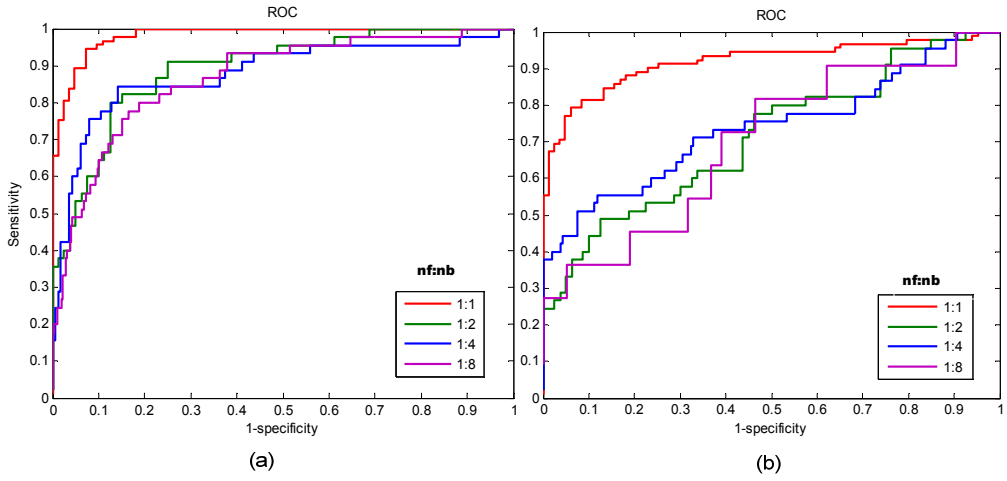


Figure 2.13. Performance comparison between the proposed GibbsOS method (a) and the condition number-based method (b), when false positive rate in the initial topology increases. 'nf:nb' is a parameter used to control the ratio of foreground genes and background genes. 'nf:nb = 1:1' means that the number of foreground genes is equal to that of background genes.

Table 2.2. Algorithm comparison on networks with different size and proportion of false positives (PFRs) in initial topology.

nf:nb	Actual proportion of false positives in initial topology	Number of foreground genes	Number of background genes	Total size of network	AUC (GibbsOS)	AUC (condition number)
1:1	0.4716	93	83	176	0.98238	0.91894
1:2	0.64	45	80	125	0.89222	0.71194
1:4	0.7837	45	163	208	0.87894	0.73968
1:8	0.8767	45	320	365	0.86883	0.69952

From Table 2.2 we can see that our proposed GibbsOS method is quite robust against the increased number of background genes in initial topology by sustaining a reasonable AUC of 0.86883 when the actual proportion of false positives (PFPs) in the network reaches 87% (nf:nb≈1:8). Comparatively, we also see that GibbsOS outperforms the condition number-based method in our simulation setting.

2.3.5. The impact of network topology on target gene identification

Besides the proportion of false positives, network topology will affect the performance of the proposed method for regulatory network identification through other factors such as: network size (cardinality), network degree (connection density) and number of transcription factors. To comprehensively evaluate the robustness of our GibbsOS method, we studied the ROC/AUC performance under three typical network configurations with different network sizes and degrees (Figure. 2.14).

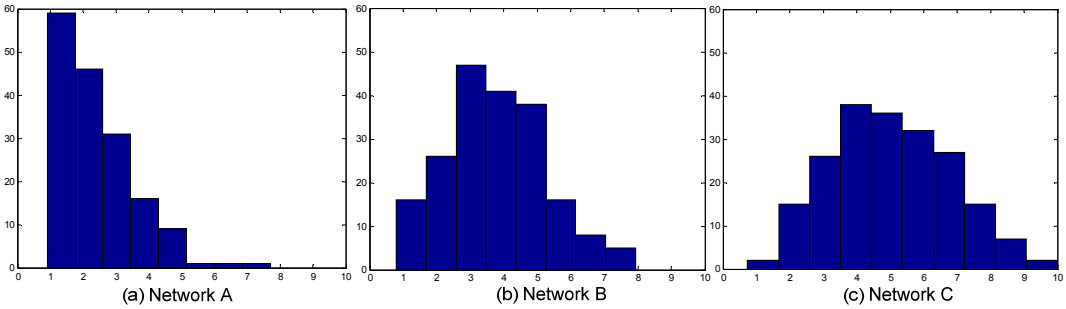


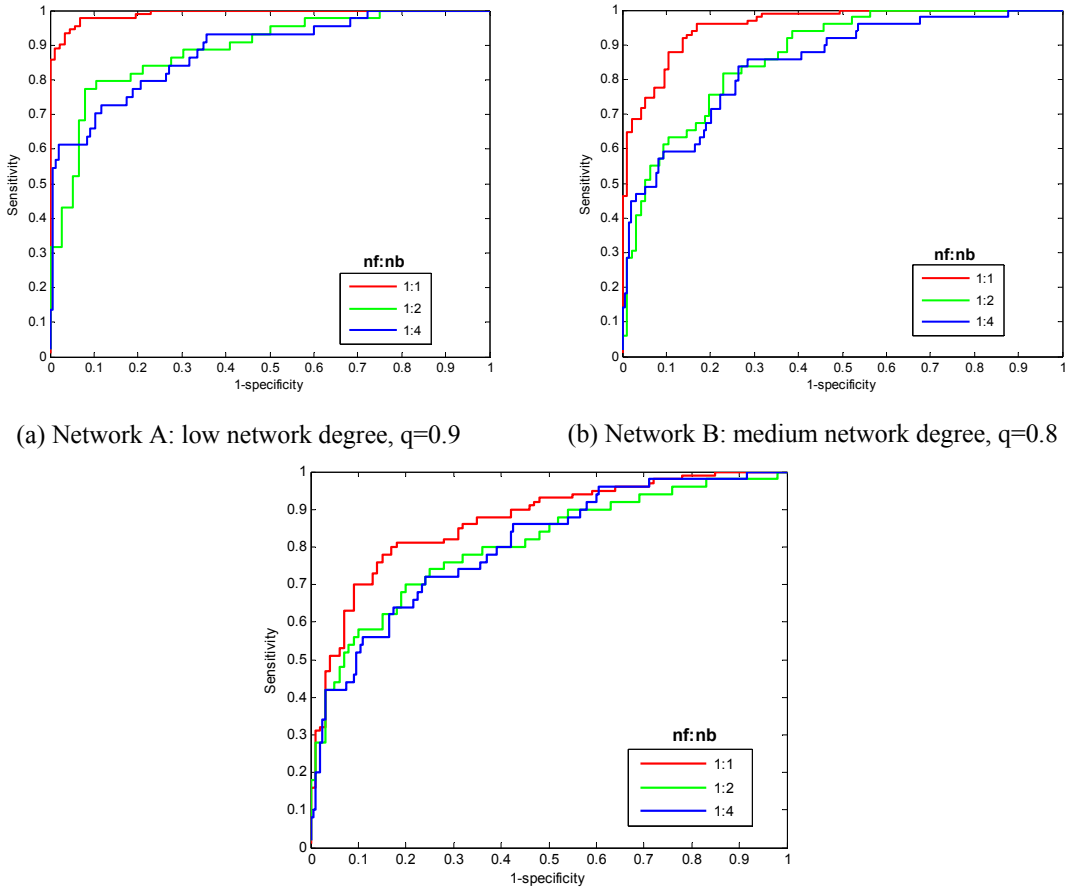
Figure 2.14. Degree distribution of simulated networks: (a) sparse network (q=0.9); (b) dense network (q=0.8); (c) much denser network (q=0.7). Note: x-axis – 'degree of network'; y-axis – 'count'.

The majority of genes in Network A have degree 1 or 2, which means that the overall network topology is quite sparse. The average degree in Network B increases to above 3 and Network C is densely connected with an average degree of about 5. Network A and C are similar to the example networks in [80], and network B is similar to the example in [81]. We have carried out a quite comprehensive study to evaluate the robustness of GibbsOS on these synthetic networks that differ in size, average degree, PFP and number of transcription factors.

Study 1: Investigating the influence of false positives and degree of the network to our proposed GibbsOS method

The degree of a target gene is defined as the number of possible transcription factors that regulate its gene expression at the same time. The network degree is an overall measure about how densely the target genes in a network are connected to the transcription factors. In our simulation, we use rand() function in Matlab to generate the connectivity matrix **A**, whose elements are uniformly distributed random numbers within 0 and 1.Those entries in **A** whose values are larger than threshold q ($0<q<1$) will be set as

1, which means a possible connection. When q is large, the network will become sparse and the average degree of the target genes will be small; on the other hand when q becomes smaller, the network will correspondingly become denser and the degree of the genes will increase. The degree distribution in Figure 2.14 is generated using q=0.9, 0.8 and 0.7 respectively, with approximately 100 foreground genes, 100 background genes and 20 transcription factors. For simplicity, we fix the sample size (30), SNR (2dB), number of transcription factors (20) in the simulation study. Figure 2.15 shows the receiver operating characteristic (ROC) curves with regard to different networks with different degree and proportion of false positives.



(a) Network A: low network degree, q=0.9

(b) Network B: medium network degree, q=0.8

(c) Network C: high network degree, q=0.7

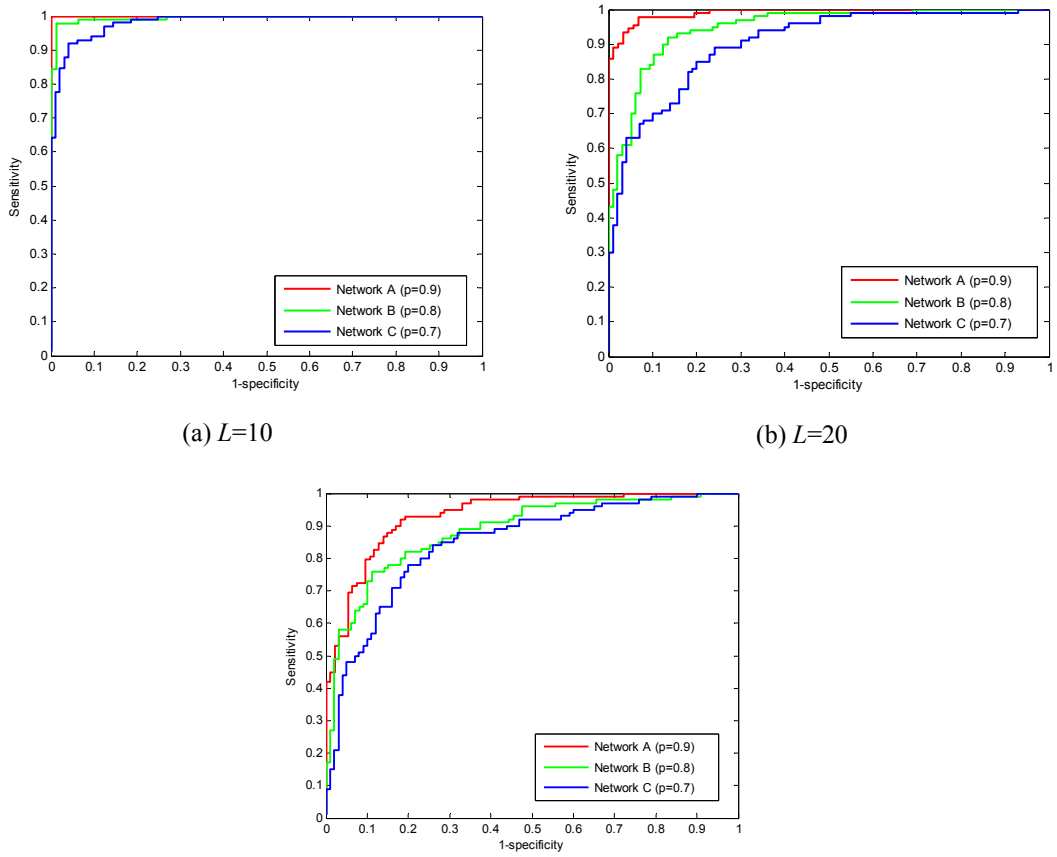
Figure 2.15. ROC performance of GibbsOS method in simulated networks with different average degree and proportion of false positives (PFPs).

Similarly as in Figure 2.13, nf:nb is a simulation parameter that controls the ratio between the number of foreground genes and that of background genes. Due to the randomness introduced in simulation, the actual ratio might be slightly deviated from the pre-set value. We also summarize the area-under-curve (AUC) values of ROC study in Table 2.3. From both Figure 2.15 and Table 2.3 we clearly see that the proposed method is quite robust when applied to networks with different average degrees and proportions of false positives. Despite a minor performance degradation when the average network degree is very high (Network C) or false positive rate becomes quite large (nf:nb=1:2 or 1:4), the general ROC curves sustain quite well with a minimum AUC of 0.8. This observation is very consistent with our intuition that the complexity of the network (network degree) and high false positive rate will make the regulatory network identification more challenging.

Table 2.3. AUC performance of the GibbsOS method with regard to network topologies of different average degrees and proportion of false positives (PFPs).

Network	nf:nb	Actual proportion of false positives in initial topology	Number of foreground genes	Number of background genes	Total size of network	AUC
Network A (p=0.9)	1:1	0.4888	91	87	178	0.9909
	1:2	0.6333	44	76	120	0.8888
	1:4	0.7789	44	155	199	0.8834
Network B (p=0.8)	1:1	0.4897	99	95	194	0.9559
	1:2	0.6621	49	96	145	0.8686
	1:4	0.7984	49	194	243	0.8461
Network C (p=0.7)	1:1	0.5	100	100	200	0.8682
	1:2	0.66667	50	100	150	0.8054
	1:4	0.80	50	200	250	0.8043

Study 2: investigating the influence of number of transcription factors



(a) L=10

(b) L=20

(c) L=30

Figure 2.16. ROC performance of the GibbsOS method regarding networks with different network scale and density. q is the parameter that controls the density of the network and L is the number of transcription factors that controls the scale of the network. The sample size is 30 and SNR =2dB.

Another factor that determines the complexity of a network is the number of transcription factors. In fact, the actual average degree of a network is jointly determined by threshold q and number of transcription factors. Previously we have carried out extensive simulations with the number of transcription factors L=20. In this section, we intent to further test how the performance will change if L becomes larger or smaller, for a fixed number of samples (M=30), SNR (2dB), around 100 foreground genes and 100 background genes. We tested three networks (q=0.9, 0.8 and 0.7) that are similar as the ones shown in Figure 2.14. Meanwhile we set the number of transcription factors L to 10,

20 and 30, to test whether the proposed method will remain robust against a variety of network configurations in simulation. The results are shown in Figure 2.16 and Table 2.4.

Table 2.4. AUC performance of the GibbsOS method with regard to network topologies of different scale and density.

Network	*L*	Actual proportion of false positives in initial topology	Number of foreground genes	Number of background genes	Total size of network	AUC
Network A (q=0.9)	10	0.3564	65	36	101	1
	20	0.4888	91	87	178	0.99459
	30	0.4896	98	94	192	0.98327
Network B (q=0.8)	10	0.4643	90	78	168	0.99091
	20	0.4924	100	97	197	0.94773
	30	0.4975	100	99	199	0.9031
Network C (q=0.7)	10	0.5	98	97	195	0.93335
	20	0.5	100	100	200	0.88505
	30	0.5	100	100	100	0.8476

From Figure 2.16 and Table 2.4 we can see that the proposed GibbsOS method is quite robust for all the network settings studied. Even in the hardest case scenario (Network C and L=30), the proposed method still can achieve a very good performance with around 0.85 AUC.

2.3.6. Yeast Cell Cycle Study

We also applied our Gibbs sampling method to Spellman's yeast cell cycle data for algorithm validation [82]. This microarray data consist of a total of 77 samples from different experiments including synchronized samples (alpha factor, elutriation or Cdc15), Cln3 and Clb2 experiments. The initial binding in-formation for network connectivity comes from the ChIP-on-chip data [83]. For this comparison experiment, we first selected a small set of TFs, 13 TFs in particular, which have strong transcriptional activities in cell cycles as reported by literature [84, 85]. We then used Spellman's 800 genes as an initial pool for the study and the p-value cutoff for the ChIP-on-chip data is 0.01. As in

the previous simulation study, we also compared our outlier sum based Gibbs sampling method with the condition number-based one. The goal of this study is to verify the efficiency of our proposed method and demonstrate its advantage over the condition number-based sampling method for cell cycle-related gene identification.

We ran both Gibbs samplers on the yeast cell cycle data and ranked the target genes according to their sampling frequencies. To assess the biological significance of the results, we use the Gene Ontology (GO) term finder from Saccharomyces Genome Database for functional enrichment analysis. For the top 100 ranked genes, our outlier sum based sampler is more successful in discovering significant clusters associated with mitosis, cell cycle regulation and processes than the condition number based sampler (Table 2.5 and Table 2.6).

Table 2.5. Top enriched yeast functional clusters identified by the GibbsOS method.

Gene Ontology term	Cluster frequency	Background frequency	p-value	FDR
Cell cycle	29 out of 100	526 out of 7167	1.98e-08	0.00%
Mitotic cell cycle	21 out of 100	282 out of 7167	6.80e-08	0.00%
Regulation of cell cycle	17 out of 100	180 out of 7167	1.18e-07	0.00%
Cell cycle process	26 out of 100	486 out of 7167	4.90e-07	0.00%
Regulation of cellular process	37 out of 100	949 out of 7167	5.05e-07	0.00%
Biological regulation	42 out of 100	1217 out of 7167	9.84e-07	0.00%

Table 2.6. Enriched functional clusters identified by the condition number-based sampling method.

Gene Ontology term	Cluster frequency	Background frequency	p-value	FDR
Mitotic cell cycle	15 out of 100	282 out of 7167	0.00247	0.00%
Cell cycle phase	17 out of 100	372 out of 7167	0.00435	0.01%

Moreover, we used cell cycle phase information associated with the given transcription factors to evaluate the quality of the extracted target genes. The cell cycle includes four stages referred to as phases: G1 phase, S phase, G2 phase and M phase. The time of peak expression of the target genes should be correlated with the cell cycle phase of the corresponding transcription factors. We selected a few transcription factors representing each cell cycle phase and plotted the expression pattern of their target genes that are ranked among top 100 in the population.

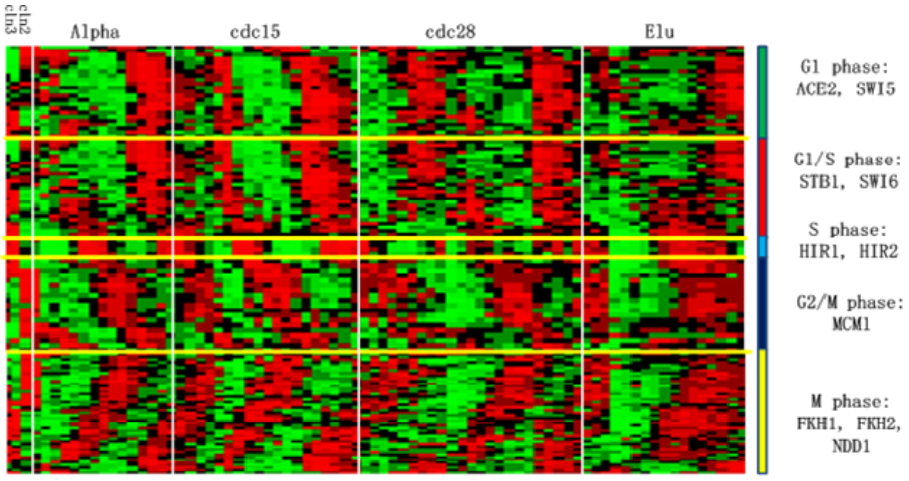


Figure 2.17. Target genes' expression pattern for transcription factors associated with different cell cycle phase.

In Figure 2.17, each row represents a gene and each column corresponds to a time point. All samples in each row have been standardized (mean subtracted and then scaled by its standard deviation). Red color means that corresponding samples are relatively over-expressed while green color denotes samples that are under-expressed. We separated the transcription factors into several groups according to their major cell cycle phases (G1, G1/S, S, G2/M and M) and extracted the top ranked target genes for each TF using the GibbsOS method. After clustering the target genes within each cell cycle phase

group, we can see a clear pattern showing the correlation between peak expression of the target genes and the cell cycle phase of the transcription factors. The genes whose regulator belongs to later phases will also have a corresponding delay in the peak expression. More interestingly, we observe that a substantial number of genes regulated by M phase transcription factors have returned to G0, S and G1 phases, which is consistent with the fact that the daughter cells will start a new cycle after cell division.

2.3.7. Breast cancer cell line study

We applied our Gibbs sampler to two breast cancer cell line data sets to identify regulatory modules associated with two different estrogen-associated conditions: estrogen-induced [86] and long-term estrogen-deprived (LTED) [87]. The estrogen-induced data set consists of three estrogen- dependent breast cancer cell lines (MCF-7, T47D and BT-474) treated with 17β-estradiol (E2 or estrogen) *in vitro* over a time course ranging from 0 to 24 hours. The original paper reported four E2-induced clusters with clear over-expression patterns at different starting or ending points. The long-term estrogen deprived (LTED) MCF-7 dataset contains eight time points ranging from day 0, when the cells were first cultured in media depleted of estrogen, up to six months when the cells are fully estrogen independent. We used Hierarchical Clustering Explorer (http://www.cs.umd.edu/hcil/hce/) to identify two clusters that are over-expressed or under-expressed after day 90, which is a critical time point when these (and other) LTED cell models have recovered from the initial effects of estrogen deprivation and begin to proliferate again [87-89].

To understand the role that estrogen signaling plays in transcriptional regulation that leads to estrogen independence, and ultimately to breast cancer progression, we selected 26 estrogen receptor (official gene symbol: ESR1) related transcription factors. These transcription factors are known to be directly or indirectly associated with ER signaling and participate in important cellular functions such as apoptosis and cell cycle regulation. Because comprehensive ChIP-on-chip data are not fully available for the human transcriptome, we used binding motif information for these 26 transcription factors to construct an initial pool of target genes. A threshold (e.g., 10 percentile of binding motif scores) was used and genes with binding motif scores larger than this threshold were

regarded as candidate targets. In order to study gene regulatory modules in a condition-specific manner, we then divided the genes identified in each condition into two groups, namely 'early up-regulated' and 'late up-regulated'. We applied our Gibbs sampler to both groups of genes under different E2 conditions and identified significant transcriptional regulatory modules associated with each condition.

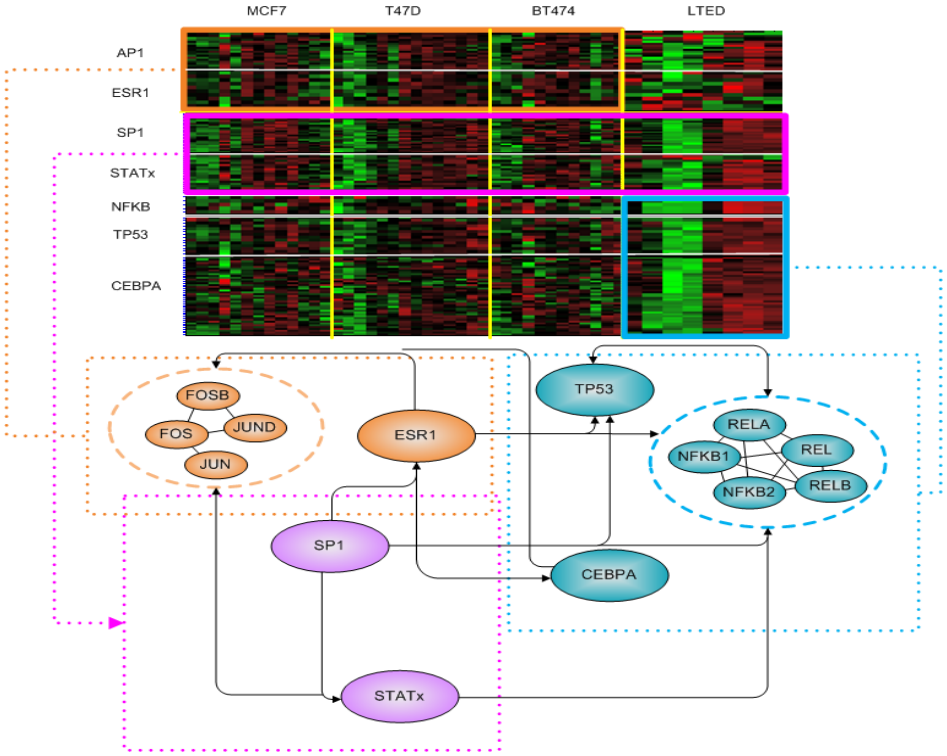


Figure 2.18. Identified transcriptional modules and their interactions associated with different estrogen conditions. Note that the interactions between different transcription factors were obtained from Ingenuity Pathway Analysis (IPA) (www.ingenuity.com). AP1 and ESR1 are significantly enriched in early up-regulated estrogen induced group while NFκB, CEBPA and TP53 have significant regulatory activities with late up-regulated long-term deprived group. Meanwhile there is strong evidence showing that SP1 and STAT family proteins are important mediators for transcriptional regulation under both groups.

In summary, we observed that ESR1 has strong regulatory activity in the estrogen induced condition but is no longer a significant transcriptional modulator in LTED cells, which is consistent with global transcriptional reprogramming and decreased reliance on pro-proliferative ESR1 signaling during the acquisition of estrogen independence. V$AP1_Q2_01, V$AP1_Q4_01 and V$CREB_Q2 were all highly enriched in the early E2 up-regulated condition, while V$NFKB_Q6_01 had strong regulatory activity under

LTED conditions. The latter observation echoes recent studies in which up-regulation of NFκB contributes to endocrine resistance in ER-positive breast cancer cells [90-92].

When we compared the early up-regulated group under estrogen induced conditions (indicative of the acute response to E2) to the late up-regulated group under LTED conditions (indicative of E2 independence), we uncovered several underlying regulatory networks (Figure 2.18). AP1 and ESR1 were uniquely activated under estrogen-induced conditions while expression of NFκB, TP53 and CEBPA target genes were associated with the LTED condition. STAT family transcription factors and SP1 showed strong regulation under both conditions. We further obtained network information of the transcription factors from Ingenuity Pathway Analysis software. From Figure 2.18 we can see that ESR1 is extensively connected to other transcription factors in a variety of ways, indicating its crucial role in mediating transactional regulation under E2 conditions. ESR1 can either interact with SP1 and TP53 through protein-DNA interactions and protein-protein interactions, or promote/inhibit activities of NFκB, CEBPA and C-fos, which is the component of AP1 protein complex. Although STAT family proteins do not directly interact with ESR1, their association with the estrogen receptor may be indirectly obtained from interactions with SP1, NFκB and AP1. Hence we conclude that ESR1 remains highly active under estrogen induced conditions while its canonical transcriptional function is weakened in LTED conditions. Concomitantly, alternate transcriptional pathways are activated through either direct or indirect crosstalk with estrogen receptor.

Interestingly, the TRANSFAC motif V$STAT_01 was activated in all four groups, suggesting that this family of transcription factors is particularly important in breast cancer progression and estrogen independence. This is a general STAT-binding motif that does not preferentially recruit any particular family member. However, under LTED conditions the family member STAT5B is the most significantly up-regulated, suggesting that this member of the STAT family is responsible for driving the network in this setting. This is consistent with the known activities of STAT5B in breast cancer; a naturally-occurring mutant of STAT5B functions as a dominant-inhibitory molecule that blocks ESR1 transcriptional activity [93], STAT5B is an essential target of BCAR1/c-Src signaling in Tamoxifen-resistant breast cancer [94], and a constitutively active variant of

STAT5B can independently induce Tamoxifen resistance in ER+ breast cancer cell lines [95].

Given the apparent importance of STAT5B signaling in this context, we more closely examined the V$STAT_01 target genes that our method identified. GCSH [96], CYP1B1 [97, 98], CISH [99], and PHB [100] have all previously been shown to be either associated with or directly regulated by estrogen/ESR1 signaling, yet all four of these genes are strongly up-regulated under late LTED conditions. Our findings therefore suggest that STAT signaling via STAT5B may functionally replace ESR1 activation of key genes that support the pro-proliferative and/or pro-survival phenotype of LTED cells, ultimately contributing to the acquisition of estrogen independence and breast cancer progression.

2.3.8 Convergence of GibbsOS

The general convergence of GibbsOS can be theoretically guaranteed (Appendix A2). However, in real biological applications, we do not have effective means to pre-determine the number of iterations needed before convergence. Therefore, we suggest to perform convergence diagnose after running the sampling chain for thousands of iterations when working on biological datasets. We have followed Cauchy's convergence criterion to evaluate the difference between two sampled cumulative distribution functions (CDFs) of target genes. The Gibbs sampler is believed to achieve reasonable convergence if inclusion of new samples does not significantly change the estimated CDF of target genes. In our breast cancer cell line study, good convergence has been achieved after typically 1000 iterations. Please see detailed definition of CDF function of target gene distribution and convergence check for breast cancer cell line study in Appendix A1 and A2.

2.4. Discussion

Identification of transcriptional regulatory modules plays a key role in understanding the mechanism of cancer progression. However, intrinsic defects of microarray data and ChIP-on-chip/binding-motif that caused by noises and the discrepancy between gene

profiles and prior binding knowledge make it challenging to correctly recover hidden regulation patterns. The condition number-based sampling method proposed by Brynildsen *et al.* is endeavored to filter out the 'background genes' that exhibit no regulation patterns but are falsely introduced by ChIP-on-chip / binding data. As we know that the condition number is defined as the ratio of the largest eigenvalue to the smallest eigenvalue, which may be affected by noise because the smallest eigenvalue typically represents the noise power while the larger eigenvalues come from the signal. Hence as signal-to-noise ratio decreases, the algorithm suffers from increasingly large performance degradation. On the other hand, the condition number of a gene profile matrix is an index showing the linear independency of gene expression levels regardless of which specific genes having linear dependency. Thus the evaluation of the seed gene for a given TF may be impaired by highly correlated seeds for other TFs.

For more robust identification of gene regulatory modules including both target genes and transcription factors, we propose a Gibbs sampler based on outlier sum of regression t-statistic. The robustness of the new statistic against experimental noises in microarray data is inherited from that the regression t-statistic maintains a reasonable statistic power when the data are relatively noisy.

In real biological applications such as breast cancer cell line study, we are fully aware of the fact that a considerable number of genes may be co-expressed through some unknown mechanism while only a small portion of them should be the true drivers of cancer progression under certain condition. By using a significance test on outlier sum, we are able to identify the TFs that have the maximum consistency between its binding knowledge and the target genes' expression pattern. We generate different null distributions associated with different target gene pool sizes to prevent the algorithm from being biased to TFs with more candidate genes.

2.5. Conclusion

We proposed a Gibbs sampler to identify condition-specific transcriptional regulatory modules based on the outlier sum of regression t-statistics. By utilizing a Gibbs strategy we are able to estimate the marginal distribution of the outlier sum statistic

for each transcription factor and evaluate the significance of the corresponding regulatory module. Our proposed method is aimed to extract the true target genes through the integration of both microarray gene expression data and ChIP-on-chip/binding motif data.

We used simulation data and model organism (e.g., yeast) to validate the efficacy of the GibbsOS method. For simulation study, we particularly compared the performance of the GibbsOS method to existing methods under two scenarios: 1) different noise levels in gene expression data and 2) different false positive connection levels in binding data. Computational results demonstrated superior performance of our proposed method regarding expression noise and structural error in binding data. We also applied GibbsOS to yeast cell cycle dataset to identify cell cycle phase specific regulatory components. We used Gene Ontology (GO) analysis from Saccharomyces Genome Database (http://www.yeastgenome.org/) to compare the computationally identified results from GibbsOS with those from the competing method (condition number based Gibbs sampler). Functional annotation analysis showed that GibbsOS successfully retrieved more target genes related to cell cycle functions (mitotic cell cycle, cell cycle regulation, etc).

Finally, we applied GibbsOS to two breast cancer cell line datasets and uncovered condition-specific regulatory rewiring in human breast cancer. Particularly, transcription factors such as NFκB and STAT family proteins are well supported by literatures that these proteins are potential contributors to endocrine resistance in ER+ breast cancer.

3. Unraveling intracellular signal transduction and pathway crosstalk by exploring pathway landscape

3.1. Introduction

Recent advances in drug discovery have posed a new direction towards designing effective anti-cancer therapies [101]. Unlike traditional targeted cancer therapy that intends to eliminate tumor cells by shutting-down single molecular pathway, a 'global' pursuit in targeting proteins that crosslink multiple cancer-related signaling pathways has drawn increasing attention [62]. Several hallmark proteins that have been identified, such as HSP90AA1 and BIRC5 (survivin), are good paradigms of crossroads intersecting multiple signaling pathways that are essential in cell proliferation, survival and resistance to growth inhibition, etc [62]. It is noteworthy that the concept of compartmentalized cancer drugs (that interfere with multiple molecular targets in different subcellular compartments) receives growing appreciation in its ability to provide tumor elimination without damaging normal cells [101-103]. Pioneer work in combinatory drug design calls for the development of novel systems biology tools to identify multiple cancer-specific pathways or signaling pathway networks, in order to counteract the complexity and heterogeneity in cancer therapy with a global strategy. Given apparent importance of understanding intracellular signal transductions that are associated with disease or environmental stress, efforts have been made in developing computational and integrative methods to combine multi-platform genomic data with biological knowledge for pathway identification. Manually curated pathway databases, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [104], have been widely incorporated into statistical methods to analyze high-throughput genomic data. Subramanian et al. developed the well-known Gene Set Enrichment Analysis (GESA) method [105] to test the statistical significance of the association between a group of genes (e.g., genes from canonical pathways) and phenotypic information. More sophisticated methods were later developed,

taking into account pathway structural information, to put a special emphasis on genes with topological importance (e.g., hub genes with many interactions) [106, 107]. Vaske et al. developed an integrative method, namely PAthway Recognition Algorithm using Data Integration on Genomic Models (PARADIGM) [108], to infer disease-related pathways; PARADIGM was built upon a probabilistic graph model to integrate gene expression data and copy number variation data for pathway identification. These methods are useful computational tools that can help capture disease-specific activities of canonical pathways; nevertheless, their ability to discover novel pathway interactions remains debatable.

Growing knowledge about genome-wide protein-protein interactions (PPIs) [38, 109] has offered a preferable source of information for signaling pathway identification, typically formulated as a mathematical problem to reconstruct signaling paths between given source proteins and target proteins [110]. One major challenge, yet to be highly emphasized, is to infer pathway edge directions based on un-directed PPI network. This problem becomes even more critical when mining aberrant signaling pathways in human cancer studies since the complete human PPI network typically contains about 9,000 proteins with more than 33,000 interactions [111]. Some proteins (or hub proteins) that connect multiple functional modules might interact with up to 200 partners (e.g., SRC). Therefore, it is an essential step to infer edge directions in PPI for the identification of signaling pathways by untangling the densely connected human PPI network towards systematic interpretation of human cancer data.

Early explorations to combine PPI data with gene expression data for pathway identification can be traced back to Netsearch [112], which performs an exhaustive search for paths in PPI network up to certain length and ranks them according to expression clusters. Scott et al. developed a random color coding scheme [113] to search for pathway networks; this algorithm was further accelerated by Hüffner et al. to reduce its runtime significantly [114]. Zhao et al. modeled the problem of finding pathway networks between given source and target proteins as an integer linear programming (ILP) problem, which can be relaxed and solved by standard linear programming techniques [110]. However, the result from the ILP approach is quite sensitive to the trade-off parameter in the algorithm that controls the network size and score. Moreover, the

authors did not address the problem of determining edge directions for pathway identification. Similarly, a minimum-cost flow optimization formulation was proposed by Yeger-Lotem et al. in ResponseNet to explore cellular response pathways [115, 116]. A flexible framework was proposed by Yosef et al. to re-weight two objective functions so that locally shortest paths between one source protein (anchor point) and several target proteins (terminals) can be reconstructed under global optimality constraint [117]. Bowtie builder is a computational method that learns the bow-tie structured pathway networks using a greedy graph search algorithm [118]. PathFinder tries to examine false positive and false negative connections in PPI by integrating multiple data resources including gene expression data and subcellular information, followed by a similarity filtering strategy of pathway segments according to association rules [119]. Although it is very encouraging from the preliminary success achieved by these methods to identify intracellular signaling transduction (e.g., MAP Kinase pathways in Saccharomyces cerevisiae), none of them explicitly address how pathway edge directions can be inferred from un-directed PPI network. Moreover, many of the existing methods focus on finding single optimal pathway or pathway network, which might only provide an oversimplified view of the diversity of signaling wiring in human cancer studies.

Among limited methods [120, 121] that specifically tackle the problem of inferring edge directions for pathway identification, most of them either reply on protein domain interactions or are not scalable to entire proteome. Notably, Gitter et al. proposed to use maximum edge orientation on PPI network to determine the most likely signaling directions that fulfill global optimality [122, 123]. By using a random orientation algorithm with local search, edge orientation (EO) algorithm is able to fine-tune signaling directions between protein pairs in the initial PPI network to improve pathway identification. EO is the first method to directly infer edge directions by integrating gene expression data with PPI network, achieving an encouraging success in predicting the osmotic stress response networks [123]. However, the EO approach relies on an assumption that most biological pathways are short (length<5) to accommodate exhaustive enumeration of all possible pathways. EO deterministically estimates the best direction for each individual edge to achieve global optimality, without providing a confidence level for inferred edge directions. Neither does EO take into account

important biological knowledge such as subcellular information, which often yields signal transduction directions that are difficult for biological interpretation. Finally, EO prioritizes a ranked list of linear pathways without linking individual pathways at structural or functional level. As a result, EO tends to favor most differential pathways while downplays less differentially expressed signaling components.

To reconstruct aberrant pathway modules and decipher inter-module pathway crosstalk in cancer studies, we develop a new approach, namely Inferring Modularization of PAthway CrossTalk (IMPACT), which integrates gene expression data with biological knowledge based on a distribution learning framework. A potential function, which is later converted to a target distribution for sampling, is formulated to measure the overall aberrancy of individual pathways. The proposed method is able to estimate edge directions by aggregating pathway samples from target distribution. IMPACT further clusters linear pathways with structural similarities into pathway modules that are bridged through nodal proteins. From a computational viewpoint, we hypothesize that these nodal proteins of inter-module pathway crosstalk are potential candidates for drug target discovery.

3.2. Overview of the pathway identification protocol

We developed a complete framework (IMPACT) to identify signaling pathway modules that link intracellular signal transduction with nucleic transcriptional regulation. IMPACT integrates gene expression data, protein-protein interaction data and subcellular location information to infer directed signaling pathway modules that are associated with clinical outcome. The core functionalities of IMPACT are built upon two novel pathway identification methods, (1) Gibbs sampler to Infer Signal Transduction (GIST) and (2) Structural Organization to Uncover pathway Landscape (SOUL) (Figure 3.1 (a)).

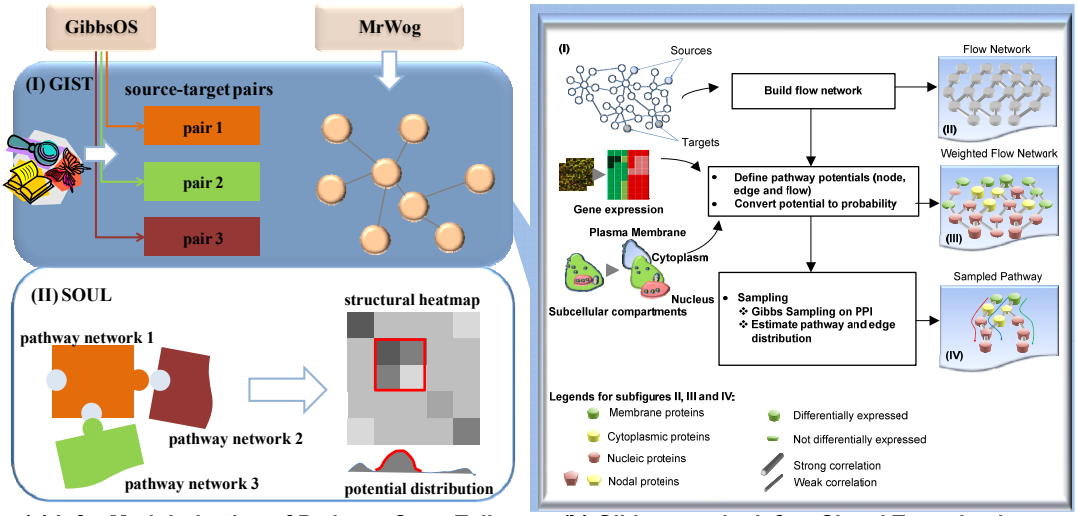


Figure 3.1. Flow chart of IMPACT method. (a) Block diagram of IMPACT method. Two prerequisite computational methods (GibbsOS and MrWog) are suggested before running GIST for pathway identification. GibbsOS, together with literature search and database mining, provide putative source(s)-target(s) pairs for biological hypotheses. MrWog is used to pre-screen PPI network to extract biologically relevant subnetworks for sampling (a.I). After running GIST algorithm, samples are pooled to reconstruct pathway landscape (a.II). (b) Detailed block diagram for GIST algorithm. Multi-platform data are integrated in GIST framework (b.I). After building flow network, nodes and edges are re-weighted according to potential functions (b.II and b.III). Gibbs sampler is employed to draw pathway samples and aggregate sampled pathways to estimate edge distributions (b.IV).

3.2.1. Gibbs sampler to Infer Signal Transduction (GIST)

A novel framework based on Gibbs sampling is proposed to integrate multi-platform data to study aberrant signal transduction and alternative pathways (Figure 3. 1 (a) (I)). Figure 3.1 (b) (I) gives a detailed block diagram of the GIST algorithm, which consists of three major components (Figure 3. 1 (b), subfigure II, III and IV). Firstly, given some predefined source proteins and target proteins, a flow network (Figure 3.1 (b) (II)) of given length is constructed between the sources and targets. Three potential functions, namely node potential, edge potential and flow potential, are defined for individual pathways to re-weight the flow network (Figure 3.1 (b) (III)). Further, a probabilistic representation is formulated to convert pathway potential into probability distribution. We apply a Gibbs sampling method to draw pathway samples according to the conditional distribution, which approximate the joint target distribution when the chain is long enough. The algorithm finally highlights interconnected linear pathways with largest

potentials (Figure 3.1 (b) (IV)) where edge directions are inferred by aggregating pathway samples. Sampled pathways from GIST algorithm will be further analyzed by computational algorithm (SOUL) to identify nodal proteins that 'oversee' several pathway modules hence to identify potential therapeutic drug targets.

3.2.2. Structural Organization to Uncover pathway Landscape (SOUL)

After running the GIST algorithm, users may obtain lists (ordered according to potentials) of sampled pathways, initiated by specific source and target proteins. We provide an algorithm namely Structural Organization to Uncover pathway Landscape (SOUL), which further analyzes pooled samples from GIST output (Figure 3. 1 (a) (II)). We first compute a structural profile whose elements are structural similarity scores of two individual pathways. A hierarchical clustering is performed to group pathways with similar structures into a 'structural heatmap'. The heatmap and pathway potential distribution jointly define the 'pathway landscape', where some pathway clusters with higher potentials will be identified as aberrant 'pathway modules'. The identified pathway modules may be cross-linked with each other through one or more nodal proteins, which are proteins that 'supervise' pathway crosstalk among multiple modules.

3.2.3. Inferring Modularization of PAthway CrossTalk (IMPACT)

Complication to apply existing PPI based pathway identification methods onto clinical patient datasets arises in several ways as follows. 1) We need to specify source and target proteins to initiate pathway identification. Gitter et al. have proposed using hidden Markov model to infer dynamic regulatory events based on time series data. However, time specific expression or binding data are hardly available for patient datasets, which calls for computational methods that are capable of modeling regulatory mechanism in steady-state. 2) Human PPI contains thousands of nodes and tens of thousands of interactions. As a result, the number of possible pathways even for moderate length (L=8) is numerically too large. 3) Most existing methods either return a single 'best' pathway or a ranked list of all pathways, without further studying their interconnection and structural linkage. We argue that it is particularly important to investigate whether the identified pathways form functional modules and how these

modules are cross-linked. This is practically meaningful when studying complex human diseases such as cancer, where increasing evidence has supported the role of alternative pathways and pathway crosstalk [62].

We provide a complete pathway identification protocol, namely Infer Modularization of PAthway CrossTalk (IMPACT), to address the above mentioned challenges. Fig.1a gives the block diagram of the entire analysis protocol, which includes two previously developed methods (GibbsOS [124] and MrWog [48]), and two aforementioned pathway identification methods (GIST and SOUL). GibbsOS is a method for regulatory network reconstruction, which identifies significant transcription factors from expression data and binding data (e.g., ChIP-on-chip and binding motif data). Significant transcription factors identified from GibbsOS will be used as target proteins for GIST algorithm. On the other hand, we use literature search and database mining to propose source proteins.

MrWog is a subnetwork identification method that draws networks from PPI using Metropolis-Hastings sampling scheme [125]. As an output, proteins with higher sampling frequencies will be selected as the base for GIST to build flow network. We regard MrWog as an important preprocessing step for GIST by retrieving proteins with larger differential power and topological importance. MrWog provides a 'shrunk' subnetwork from PPI data that significantly reduces the sampling space for GIST algorithm, yet preserving most biologically relevant information through probabilistic integration of gene expression data and PPI data.

Based on source-target proteins and identified subnetworks from prerequisite methods, IMPACT calls GIST to draw pathway samples, which are later utilized by SOUL to reconstruct pathway landscape. Unlike traditional methods that either prioritize linear pathways or extract pathways with certain structures (e.g., tree structure), IMPACT suggests highlighting structural correlated pathway modules to investigate their inter-module crosstalk. By prioritizing proteins that 'oversee' crosstalk among multiple aberrant pathway modules, IMPACT offers an alternative mean to investigate potential drug targets for human cancer study.

3.3. Gibbs sampler to Infer Signal Transduction

3.3.1 Defining pathway potentials

Let $\boldsymbol{\theta}_{1\times L}=\{\theta_1,\theta_2,...,\theta_L\}$ denote a linear pathway of length L, where θ_i is a categorical variable that represents a specific protein at the i^{th} position of the pathway. θ_1 is the source protein and θ_L is the target protein. Let Ω_i denote the domain of θ_i and we have $\Omega_1\bigcap\Omega_2\bigcap...\bigcap\Omega_L\subseteq\Omega$, where Ω denotes all proteins in the PPI. We also have gene expression data $\mathbf{X}_{n\times m}$, where n denotes number of genes and m denotes number of samples. We adopt the terminology of 'potential function', which originated from physics and was later widely used in image processing, to measure abnormality of pathways. Three potential functions have been defined, which are termed as node potential, edge potential and flow potential.

Node potential

We use node potential to investigate how proteins within a pathway are correlated to phenotypic data. Either categorically or continuously valued phenotype information can be incorporated to formulate node potential. For categorical data, let us denote $\mathbf{c}=[c_1,c_2,...,c_M]^{\mathrm{T}}$ as the group label for M samples under two clinical conditions, e.g., $c_m=1$ means patients with high disease risk while $c_m=0$ means vice versa. We use two-sample t-test to evaluate the differential expression of i^{th} protein in the pathway, and the test statistic and p-value are given by:

$$t_i^{(1)}=\frac{\bar{\mathbf{x}}_{\theta_i,1}-\bar{\mathbf{x}}_{\theta_i,0}}{s\cdot\sqrt{\frac{1}{m_1}+\frac{1}{m_0}}}, \tag{3.1}$$

$$p_i^{(1)}=2\times\left(1-\int_{-\infty}^{\left|t_i^{(1)}\right|}f_{\text{student's }t}(t,\nu)dt\right) \tag{3.2}$$

where $\bar{\mathbf{x}}_{\theta_i,1}$ and $\bar{\mathbf{x}}_{\theta_i,0}$ are the averaged expressions of i^{th} protein respectively under two conditions. m_1 and m_0 are the numbers of samples and s is the pooled sampled standard

deviation. $\nu = m_1 + m_0 - 2$ is the degree of freedom for student's t distribution. The potential of protein θ_i is therefore calculated as:

$$\mathrm{V}_1\left(\theta_i;\mathbf{X}\right) = \Phi^{-1}\left(1 - p_i^{(1)}\right), \tag{3.3}$$

where Φ^{-1} is the normal inverse cumulative distribution function. For continuously valued phenotypic information c, we can use Cox proportional hazards model [126] to calculate the corresponding p-value.

Edge potential

We use the Pearson's correlation test to measure co-expression of two adjacent proteins θ_i and θ_{i+1} in the pathway. The test statistic $t_{i,i+1}^{(2)}$, correlation coefficient r and corresponding significance value $p_{i,i+1}^{(2)}$ are given by:

$$t_{i,i+1}^{(2)} = r\sqrt{\frac{m-2}{1-r^2}}, \tag{3.4}$$

$$r = \frac{1}{m-1}\sum_{1}^{m}\left(\frac{\mathbf{x}_{\theta_i} - \overline{\mathbf{x}}_{\theta_i}}{s_{\mathbf{x}_{\theta_i}}}\right)\left(\frac{\mathbf{x}_{\theta_{i+1}} - \overline{\mathbf{x}}_{\theta_{i+1}}}{s_{\mathbf{x}_{\theta_{i+1}}}}\right) \tag{3.5}$$

$$p_{i,i+1}^{(2)} = 2\times\left(1 - \int_{-\infty}^{\left|t_{i,j}^{(2)}\right|} f_{\text{student's } t}\left(t,\nu\right)dt\right) \tag{3.6}$$

where $\mathbf{x}_{\theta_i}$ and $\mathbf{x}_{\theta_{i+1}}$ are gene expression vectors of i^{th} protein and j^{th} protein in the pathway, each with m samples. $\overline{\mathbf{x}}_{\theta_i}$ is the sample mean and $s_{\mathbf{x}_{\theta_i}}$ is the standard deviation. $\nu = m - 2$ is the degree of freedom. The potential of the edge between protein θ_i and protein θ_{i+1} is given by:

$$\mathrm{V}_2\left(\theta_i,\theta_{i+1};\mathbf{X}\right) = \Phi^{-1}\left(1 - p_{i,i+1}^{(2)}\right) \tag{3.7}$$

Flow potential

Neither node nor edge potential directly addresses causality between pathway components. To infer signaling transduction between source proteins and target proteins, additional information (e.g., subcellular locations) can be incorporated into the model.

We assume that biological valid pathways most likely start from extracellular space or membrane, go through cytoplasm and finally enter into nucleus. A flexible framework that allows users to incorporate their domain knowledge of how biological signal flows across different subcellular compartments is regarded useful. An entire pathway $\boldsymbol{\theta}$ of length L is a clique $C_L = \{\theta_k | \theta_k \in \boldsymbol{\theta}\}$. By encoding subcellular locations into numeric values (Extracellular space: 1, membrane: 2, cytoplasm: 3, nucleus 4), we define the flow potential as follows:

$$\mathrm{V}_3(\boldsymbol{\theta}) = \sum_{i=1}^{L-1} \mathrm{I}(l_{i+1} - l_i) + \sum_{k=1}^{4} \psi_k \sum_{i}^{L-1} (l_i = k), \tag{3.8}$$

where l_i is the encoded subcellular location of i^{th} protein in pathway $\boldsymbol{\theta}$. ψ_k is the parameter that weights the importance of each subcellular compartment. The first term of Equation (3.8) measures how many directed edges are consistent to subcellular location information. The second term is designed for users to flexibly weight each subcellular layer. For example, if we want to study pathways in the cytoplasm, we can set ψ_3 to a positive value while keep all other $\psi_k, k \neq 3$ to zero.

3.3.2 Pathway energy function, Gibbs distribution and Gibbs sampling

Having defined three potential functions, we combine them to get the energy function of pathway $\boldsymbol{\theta}$ as:

$$\mathrm{U}(\boldsymbol{\theta}; \mathbf{X}) = \frac{\sum_{i=1}^{L} \mathrm{V}_1(\theta_i; \mathbf{X}) + \lambda \sum_{i=1}^{L-1} \mathrm{V}_2(\theta_i, \theta_{i+1}; \mathbf{X})}{\sqrt{1+\lambda^2}} + \mathrm{V}_3(\boldsymbol{\theta}), \tag{3.9}$$

where parameter λ controls the trade-off between node potential and edge potential. The first two potentials are driven by the data, which we refer to as 'likelihood potentials', while the flow potential comes from knowledge and is therefore referred to as 'prior potential'. From optimization point of view, Equation (3.9) defines the cost function by maximizing which we should get an optimal pathway. However, due to many confounding factors (such as biological noise and experimental noise in the data, sample heterogeneity, false-positives and false-negatives in PPI, uncertainty of subcellular information induced by protein translocation, etc.), finding the 'best' single or a few

pathways may not comprehensively capture signal transduction in the application of human cancer study. Instead, we prefer to translate the optimization task into a distribution learning problem. We define an intermediate variable $\mathrm{S}(\boldsymbol{\theta}) = -\mathrm{U}(\boldsymbol{\theta})$, and convert $\mathrm{S}(\boldsymbol{\theta})$ into probabilistic form using Gibbs distribution [127] as follows:

$$\begin{aligned} P(\boldsymbol{\theta};\mathbf{X}) &= \frac{1}{Z}\cdot e^{\frac{-\mathrm{S}(\boldsymbol{\theta};\mathbf{X})}{T}} = \frac{1}{Z}\cdot e^{\frac{\mathrm{U}(\boldsymbol{\theta};\mathbf{X})}{T}} \\ &= \frac{1}{Z}\cdot \exp\left(\frac{\sum_{i=1}^{L}\mathrm{V}_1(\theta_i;\mathbf{X}) + \lambda\sum_{i=1}^{L-1}\mathrm{V}_2(\theta_i,\theta_{i+1};\mathbf{X})}{T\sqrt{1+\lambda^2}} + \frac{\mathrm{V}_3(\boldsymbol{\theta})}{T} \right) \end{aligned} \tag{3.10}$$

where $Z = \sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} e^{\frac{1}{T}\mathrm{U}(\boldsymbol{\theta};\mathbf{X})}$ is called the partition function and $\boldsymbol{\Theta}$ is the set of all possible paths. T is referred to as 'temperature' that controls the shape of the distribution. To efficiently sample pathways from $P(\boldsymbol{\theta};\mathbf{X})$, we employ a Gibbs strategy by drawing samples from the conditional distribution. The sampling process can be formulated as:

$$\theta_i^{(t+1)} \sim P(\theta_i \,|\, \theta_1^{(t+1)},\ldots,\theta_{i-1}^{(t+1)},\theta_{i+1}^{(t)},\ldots\theta_L^{(t)};\mathbf{X}) \tag{3.11}$$

where t is the t^{th} sampling iteration. In each iteration, we probabilistically accept a new sample for θ_i based on the conditional distribution defined by Equation (3.10) while fix the other proteins θ_{-i} in the pathway. The samples drawn from the conditional distribution are theoretically good approximation of samples from the joint posteriors distribution.

3.3.3 Flow network and sampling on PPI

GIST does not apply Gibbs sampling directly on PPI network, but rather on a regularized structure that we refer to as 'flow network'. Given the source protein(s) and target protein(s) of interest, we first need to build a directed pathway flow network of L layers from the original PPI as shown in Figure 3.2. First, we start from the source nodes (nodes in the first layer) and search its neighbors in the PPI network. The direct neighbors of the source node are included into the second layer, based on which we successively define the third layer, forth layer, etc. This is called the 'forward search' of the PPI network and the target node will present at the L^{th} layer (the target node can also show up

in the upper layers if there is a path between source and target that has length smaller than L). Similarly, we also perform 'backward search' of the PPI network which starts from the target node and rebuild the L layers in the reverse direction. Once we obtain the 'forward' and 'backward' network, we can merge them use 'AND' logic: for each layer we only keep the nodes that are present at both 'forward' and 'backward' networks to get the final flow network.

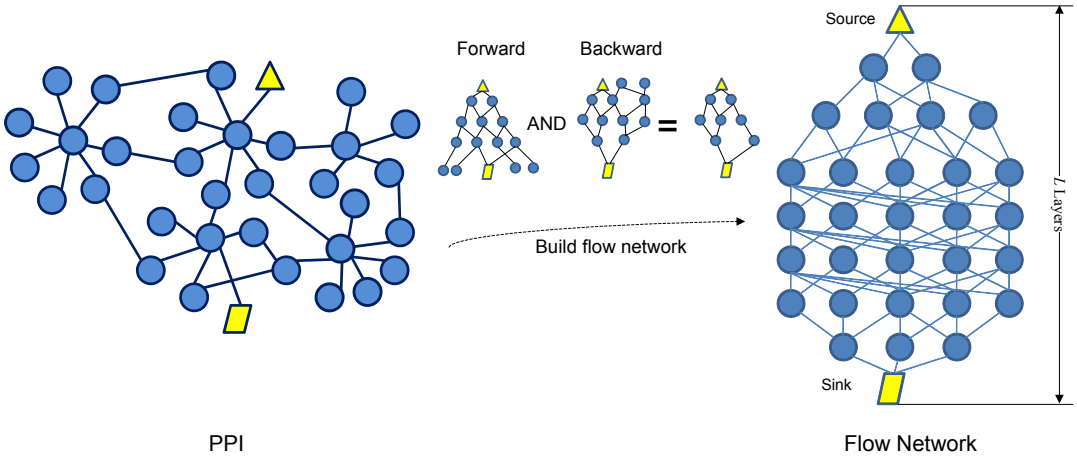


Figure 3.2. An illustration of constructing flow network from protein-protein interaction data.

Figure 3.3 shows how we sample pathway configurations from the flow network. First of all, we will obtain a flow network of L layers, between the sources and targets (Figure 3.3 (a)). One protein may appear at different layers in the flow network. A flow network defines a regularized and directed structure between given source proteins and target proteins but assumes uniform weight on each individual paths. We re-weight the flow network by assigning potentials to individual nodes and edges (Figure 3.3 (b)). The size (thickness) of the nodes (edges) indicates the significance of protein-phenotype association (protein-protein interaction). Different colors represent different subcellular locations. Starting from any arbitrary initial pathway, GIST iteratively replaces nodes in the i^{th} ($1 \leqslant i \leqslant L$) layer of current pathway by probabilistically sampling one candidate gene from i^{th} layer. Figure 3.3 gives an example on how pathways are being sampled on the weighted flow network.

However, the above network-constrained Gibbs sampler may be stuck into local distribution especially when there are multiple parallel (non-intersecting) pathways

between given sources and targets. In theory, this is due to that the Markov chain defined by the sampling process is reducible (i.e., not all states can be reached from any other state). To resolve this issue, we introduce the concept of 'pseudo-edge', to communicate any two states in the pathway state space. More specifically, for any two proteins in the adjacent layers of the weighted flow network, even though they are not connected in the original PPI network, we will add an edge with very small weight δ (usually zero or negative) to connect them. This simple modification will make the sampling process on the flow network an ergodic Markov chain, which will converge to unique stationary distribution (see Section 3.3.4).

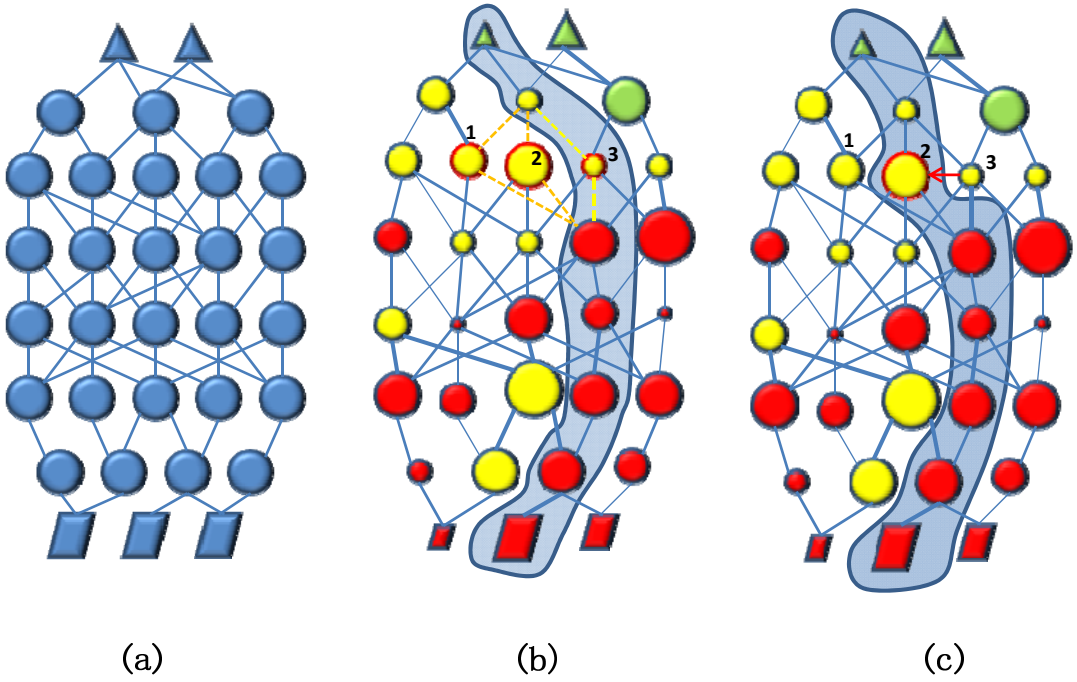


Figure 3.3. Sampling on the flow network. (a) Flow network is constructed among given source(s) and target(s). (b) Re-weight nodes and edges based on potential functions. Different node color represents different subcellular compartments (green: membrane; yellow: cytoplasm; red: nucleus). The current pathway is highlighted in the shaded area. A Gibbs sampler will update one gene θ_i at a time, and iteratively updates all the other pathway members. Suppose we want to update the third protein θ_3 in the pathway. Based on the flow network, three proteins in the third layer (marked 1, 2 and 3) are potential candidates that connect the existing proteins to the second and fourth layer. (c) We calculate the pathway probabilities according to Equation (3.10) for all three corresponding pathways and probabilistically accept one of the proteins to update the previous one (protein 2 is selected to update θ_3). We also correspondingly update the edges of the new protein. This procedure will be sequentially applied to the fourth, fifth until the L^{th} layer, and will be repeated for many iterations until convergence.

3.3.4. Modified Gibbs sampler and Markov properties

In order that through enough sampling iterations, the estimated distribution will be a stationary distribution that is irrelevant to its initial states, the proposed Gibbs sampler should have some basic properties such as irreducibility and reversibility. A Markov chain is said to be 'irreducible' if it is possible to get to any state from any other state, which can be concisely formulated as:

$$\forall \boldsymbol{\theta}_i, \boldsymbol{\theta}_j \in \boldsymbol{\Theta},\ \exists\, n,\ \text{s.t.}\ P\left(\boldsymbol{\theta}^{(n)} = j \middle| \boldsymbol{\theta}^{(0)} = i\right) = p_{ij}^{(n)} > 0\,. \tag{3.12}$$

Unfortunately, this property cannot always be satisfied when we are sampling pathways from the PPI network using the framework described by Figure 3.3, which means that the sampler may be stuck at local distribution. This is because that in our Gibbs sampling process, we only allow changing one layer in the current path at a time. One *ad hoc* solution to improve the algorithm is to increase the number of layers being sampled at the same time [128] and in the extreme case when the step-size equals to L, the algorithm simply becomes a exhaustive search method. Hereby we propose a simple modification of the previous Gibbs sampler by introducing a small baseline sampling frequency δ to states (pathway configurations) that are not allowed by the initial flow network, so that any two states in $\boldsymbol{\Theta}$ communicates with each other. Edges connecting nodes in two adjacent layers that are not connected in the original PPI are referred to as 'pseudo-edges'. Figure 3.4 shows a toy model of how we make the Gibbs sampler irreducible:

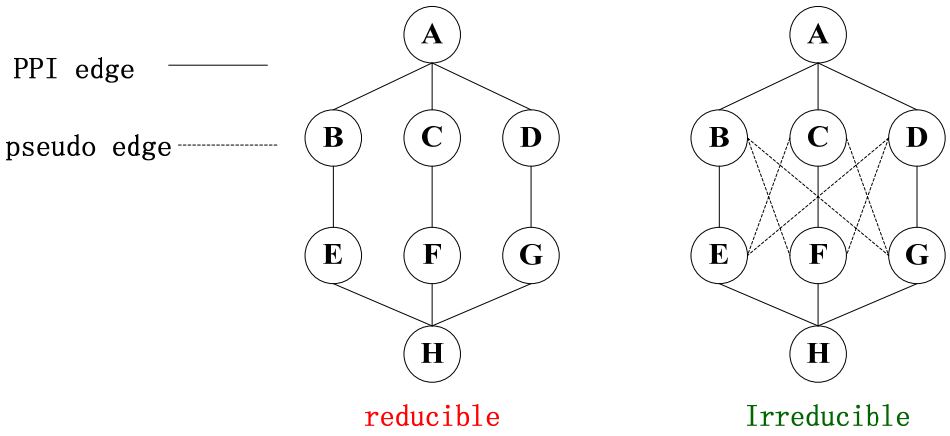


Figure 3.4. An example of irreducibility for pathway sampling.

The sampling process defined on Figure 3.4 (a) is reducible because if the algorithm starts from an initial state (1,2,5,8,11) , it will always be stuck at the same state. So we have

$$p\left(\boldsymbol{\theta}=[1,2,5,8,11]^{\mathrm{T}} \rightarrow \boldsymbol{\theta}=[1,3,6,9,11]^{\mathrm{T}}\right)=0$$

$$\text{and } p\left(\boldsymbol{\theta}=[1,2,5,8,11]^{\mathrm{T}} \rightarrow \boldsymbol{\theta}=[1,4,7,10,11]^{\mathrm{T}}\right)=0 .$$

In Figure 3.4 (b), the nodes in any two adjacent layers are mutually connected regardless whether they are truly connected in the original PPI network or not. For those 'true connections' that derived from the original PPI, we use the PPI score as the edge score (in this toy example, all true PPI connection scores are set to 1). For those pseudo-edges that do not exist in the original PPI network, we set their weight to 0. Hence any two states defined on the flow network are communicating with each other. For example, suppose the current state is (1,2,5,8,11) and it can reach the new state (1,3,6,9,11) by either:

(1,2,5,8,11)→ (1,3,5,8,11)→ (1,3,6,8,11)→ (1,3,6,9,11)

or (1,2,5,8,11)→ (1,2,6,8,11)→ (1,2,6,9,11)→ (1,3,6,9,11) or… etc.

Before we go to further discussion, we want to define the 'relative distance' between two pathway state vectors as:

$$d\left(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j\right)=\sum_{l=1}^{L} \mathrm{I}\left(\theta_{i,l} \neq \theta_{j,l}\right), \tag{3.13}$$

where $\mathrm{I}(x)=\begin{cases}1, & x \text{ is true} \\ 0, & x \text{ is false}\end{cases}$. The 'relative distance' is simply the total number of different elements in the same entry. Because nodes from any two adjacent layers are connected in the flow network, we know that for the Gibbs sampler that updates one variable in the state vector, the step-one transition probability should satisfy that for any $\boldsymbol{\theta}_i, \boldsymbol{\theta}_j$:

$$p^{(1)}\left(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j\right)=p_{i,j}^{(1)}>0 \text{ for } d\left(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j\right)=1$$

$$\text{and } p^{(1)}\left(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j\right)=p_{i,j}^{(1)}=0 \text{ for } d\left(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j\right)>1. \tag{3.13}$$

Similarly, we can also define the '*l*-relative distance' as:

$$d^l\left(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j\right)=\mathrm{I}\left(\theta_{i,l}\neq\theta_{j,l}\right),\quad 1\le l\le L\,, \tag{3.14}$$

and we have $d\left(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j\right)=\sum_{l=1}^{L}d^l\left(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j\right)$.

The Gibbs sampling process on the modified flow network has defined an irreducible Markov chain that draw samples (states) from an enlarged state space $\boldsymbol{\Theta}'$ and $\boldsymbol{\Theta}\subset\boldsymbol{\Theta}'$, where $\boldsymbol{\Theta}$ is the state space of all 'valid pathways' defined by the original flow network (see Equation (3.10)). We have the following theorems regarding the modified Gibbs sampler:

Theorem 1: Introducing pseudo-edges does not change the probability of any valid pathway $\boldsymbol{\theta}\in\boldsymbol{\Theta}$.

Proof: The proof of theorem 1 is quite straightforward. Let's denote the probability of pathway $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ under the modified sampling process as $p'(\boldsymbol{\theta})$, which can be written as:

$$p'(\boldsymbol{\theta})=\frac{f'(\boldsymbol{\theta})}{\sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}f'(\boldsymbol{\theta})}=\frac{\dfrac{e^{-\frac{1}{T}\mathrm{S}(\boldsymbol{\theta})}}{\sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}'}e^{-\frac{1}{T}\mathrm{S}(\boldsymbol{\theta})}}}{\sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\dfrac{e^{-\frac{1}{T}\mathrm{S}(\boldsymbol{\theta})}}{\sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}'}e^{-\frac{1}{T}\mathrm{S}(\boldsymbol{\theta})}}}=\frac{e^{e^{-\frac{1}{T}\mathrm{S}(\boldsymbol{\theta})}}}{\sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}e^{-\frac{1}{T}\mathrm{S}(\boldsymbol{\theta})}}=p(\boldsymbol{\theta})\,. \tag{3.15}$$

For Equation 3.15 to hold, the sampling chain should be long enough so that the relative frequency can approximate the true probability well.

Theorem 2: The proposed Gibbs sampler defines a reversible Markov chain.

Proof: A Markov process is reversible when the 'detailed balance' condition is satisfied:

$$\pi_i p_{i,j}=\pi_j p_{j,i}\,, \tag{3.16}$$

where π_i is the equilibrium probability of state $\boldsymbol{\theta}_i$ and $p_{i,j}$ is the transition probability from $\boldsymbol{\theta}_i$ to $\boldsymbol{\theta}_j$. For any two states $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$, we have:

$$p_{i,j}=\begin{cases}\dfrac{1}{L}\cdot\dfrac{e^{-\frac{1}{T}\mathrm{S}_{\boldsymbol{\theta}_j}}}{\sum_{\boldsymbol{\theta}_k\in\boldsymbol{\Theta}_l}e^{-\frac{1}{T}\mathrm{S}_{\boldsymbol{\theta}_k}}}, & d^l\left(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j\right)=d\left(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j\right)\le 1\\ 0, & \text{otherwise}\end{cases}, \tag{3.17}$$

where $\boldsymbol{\Theta}_l=\left\{\boldsymbol{\theta}_k\middle|d^l\left(\boldsymbol{\theta}_i,\boldsymbol{\theta}_k\right)=d\left(\boldsymbol{\theta}_i,\boldsymbol{\theta}_k\right)\le 1\right\}$. Hence for $d^l\left(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j\right)=d\left(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j\right)\le 1$, we have:

$$\begin{aligned}\pi_i p_{i,j}&=\frac{1}{Z}\times e^{-\frac{1}{T}\mathrm{S}_{\boldsymbol{\theta}_i}}\cdot\frac{1}{L}\cdot\frac{e^{-\frac{1}{T}\mathrm{S}_{\boldsymbol{\theta}_j}}}{\sum_{\boldsymbol{\theta}_k\in\boldsymbol{\Theta}_l}e^{-\frac{1}{T}\mathrm{S}_{\boldsymbol{\theta}_k}}}\\ &=\frac{1}{Z}\times e^{-\frac{1}{T}\mathrm{S}_{\boldsymbol{\theta}_j}}\cdot\frac{1}{L}\cdot\frac{e^{-\frac{1}{T}\mathrm{S}_{\boldsymbol{\theta}_i}}}{\sum_{\boldsymbol{\theta}_k\in\boldsymbol{\Theta}_l}e^{-\frac{1}{T}\mathrm{S}_{\boldsymbol{\theta}_k}}}.\\ &=\pi_j p_{j,i}\end{aligned} \tag{3.18}$$

For cases that $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ have a relative distance larger than 1, the detailed balance is always satisfied because the transition probability is zero.

3.3.5. Parameter learning using Bayesian inference

Parameter λ in Equation (3.9) determines the trade-off between node potential and edge potential. However, the maximum likelihood estimate of λ might not exist, or in other words has infinite solution (e.g., $\lambda=\pm\infty$). Here we offer a remedy to explore λ in a given domain $\left[\lambda_{\text{inf}},\lambda_{\text{sup}}\right]$ by sampling it jointly with pathway states $\boldsymbol{\theta}$. Given a newly sampled pathway $\boldsymbol{\theta}^t$, we sample λ^t from the posterior distribution:

$$P\left(\lambda\middle|\boldsymbol{\theta}^t,\mathbf{X}\right)\propto P\left(\boldsymbol{\theta}^t\middle|\lambda,\mathbf{X}\right)P(\lambda). \tag{3.19}$$

We use Metropolis-Hastings method to sample λ. A uniform proposal function defined over $\left[\lambda_{\text{inf}},\lambda_{\text{sup}}\right]$ is used to propose samples. If we use the prior distribution of λ to propose new λ, we compute the following ratio

$$\begin{aligned}\alpha &= \frac{p\left(\lambda'|\boldsymbol{\theta},\mathbf{X}\right)\cdot q\left(\lambda^{(t)}\right)}{p\left(\lambda^{(t)}|\boldsymbol{\theta},\mathbf{X}\right)\cdot q\left(\lambda'\right)} \\ &= \frac{p\left(\boldsymbol{\theta}|\lambda',\mathbf{X}\right)\cdot q\left(\lambda'\right)\cdot q\left(\lambda^{(t)}\right)}{p\left(\boldsymbol{\theta}\middle|\lambda^{(t)},\mathbf{X}\right)\cdot q\left(\lambda^{(t)}\right)\cdot q\left(\lambda'\right)}, \\ &= \frac{p\left(\boldsymbol{\theta}|\lambda',\mathbf{X}\right)}{p\left(\boldsymbol{\theta}\middle|\lambda^{(t)},\mathbf{X}\right)}\end{aligned} \tag{3.20}$$

where $\lambda^{(t)}$ is the current value of parameter λ. By generating a random number η, where $0 \le \eta \le 1$ and we can update λ according to the following procedure:

$$\lambda^{(t+1)} = \begin{cases} \lambda', & \alpha \ge \eta \\ \lambda^{(t)} & \alpha < \eta \end{cases}. \tag{3.21}$$

If we substitute the explicit form of $p\left(\boldsymbol{\theta}|\lambda',\mathbf{X}\right)$ into Equation (3.20), we have:

$$\alpha = \frac{p\left(\boldsymbol{\theta}|\lambda',\mathbf{X}\right)}{p\left(\boldsymbol{\theta}\middle|\lambda^{(t)},\mathbf{X}\right)} = \frac{\dfrac{e^{\frac{\mathrm{U}_{\lambda'}(\boldsymbol{\theta};\mathbf{X})}{T}}}{\sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_l} e^{\frac{1}{T}\mathrm{U}_{\lambda'}(\boldsymbol{\theta};\mathbf{X})}}}{\dfrac{e^{\frac{\mathrm{U}_{\lambda^{(t)}}(\boldsymbol{\theta};\mathbf{X})}{T}}}{\sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_l} e^{\frac{1}{T}\mathrm{U}_{\lambda^{(t)}}(\boldsymbol{\theta};\mathbf{X})}}}, \tag{3.22}$$

where $\boldsymbol{\Theta}_l = \left\{\boldsymbol{\theta}_k \middle| d^l\left(\boldsymbol{\theta}_i,\boldsymbol{\theta}_k\right) = d\left(\boldsymbol{\theta}_i,\boldsymbol{\theta}_k\right) \le 1\right\}$.

3.3.6. Interpreting the posterior distribution

One major effort in our research is to combine multi-platform data (gene expression data, PPI network and subcellular location) to infer signaling directions. In our terminology, PPI network and subcellular information are referred to as ‘knowledge’, while gene expression profiles are referred to as ‘data’. Bayes' rule, which has been long

adopted to aggregate belief from both knowledge and observations (data), especially fits the scenario to interpret our pathway probability model. For any pathway configuration $\boldsymbol{\theta}$, we define the prior probability as:

$$p(\boldsymbol{\theta}) = \begin{cases} \dfrac{e^{\frac{V_3(\boldsymbol{\theta})}{T}}}{\sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} e^{\frac{V_3(\boldsymbol{\theta})}{T}}} = \dfrac{1}{Z_1} e^{\frac{V_3(\boldsymbol{\theta})}{T}}, & \boldsymbol{\theta}\in\boldsymbol{\Theta} \\ 0, & \boldsymbol{\theta}\notin\boldsymbol{\Theta} \end{cases}, \tag{3.23}$$

where $V_3(\boldsymbol{\theta})$ is the flow potential which is given by Equation (3.8). $\boldsymbol{\Theta}$ is the entire set of pathway configurations (states) defined by the flow network. In other words, only pathways between given sources and targets in the flow network will be assigned a non-zero probability; otherwise the prior probability of $\boldsymbol{\theta}$ is always zero. The prior probability $p(\boldsymbol{\theta})$ encodes PPI topology, sources and targets, and subcellular locations to prioritize pathway structures, which is totally independent to observed gene expression data.

On the other hand, we propose the following probability formulation as the likelihood function based on node and edge potentials:

$$p(\mathbf{X}|\boldsymbol{\theta}) = \frac{1}{Z_2} \cdot \exp\left(\frac{\sum_{i=1}^{L} V_1(\theta_i;\mathbf{X}) + \lambda \sum_{i=1}^{L-1} V_2(\theta_i,\theta_{i+1};\mathbf{X})}{T\sqrt{1+\lambda^2}} \right), \tag{3.24}$$

where $Z_2 = \sum_{all\ \boldsymbol{\theta}} \exp\left(\frac{\sum_{i=1}^{L} V_1(\theta_i;\mathbf{X}) + \lambda \sum_{i=1}^{L-1} V_2(\theta_i,\theta_{i+1};\mathbf{X})}{T\sqrt{1+\lambda^2}} \right)$. For any pathway configuration $\boldsymbol{\theta}$, no matter it is in the flow network or not, we can assign a likelihood probability based on its association with phenotype and correlation between node expressions. Pathways with large node potentials and edge potentials are preferred by Equation (3.24). By

multiplying Equation (3.23) and Equation (3.24), we get the 'posterior probability' of pathway $\boldsymbol{\theta}$ as:

$$p(\boldsymbol{\theta}|\mathbf{X}) = \begin{cases} \frac{1}{Z_1 Z_2} \exp\left(\frac{\sum_{i=1}^{L} \mathrm{V}_1(\theta_i;\mathbf{X}) + \lambda \sum_{i=1}^{L-1} \mathrm{V}_2(\theta_i, \theta_{i+1};\mathbf{X})}{T\sqrt{1+\lambda^2}} + \frac{\mathrm{V}_3(\boldsymbol{\theta})}{T} \right), & \boldsymbol{\theta} \in \boldsymbol{\Theta} \\ 0, & \boldsymbol{\theta} \notin \boldsymbol{\Theta} \end{cases}, \quad (3.25)$$

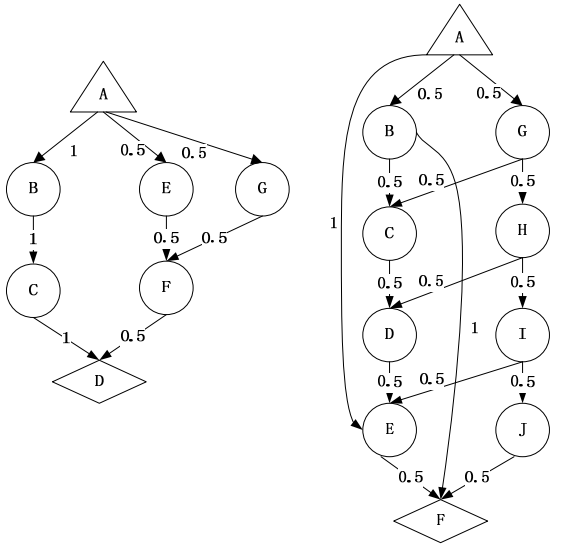
$$\propto \begin{cases} \frac{1}{Z} \exp\left(\frac{\sum_{i=1}^{L} \mathrm{V}_1(\theta_i;\mathbf{X}) + \lambda \sum_{i=1}^{L-1} \mathrm{V}_2(\theta_i, \theta_{i+1};\mathbf{X})}{T\sqrt{1+\lambda^2}} + \frac{\mathrm{V}_3(\boldsymbol{\theta})}{T} \right), & \boldsymbol{\theta} \in \boldsymbol{\Theta} \\ 0, & \boldsymbol{\theta} \notin \boldsymbol{\Theta} \end{cases}$$

where $Z = \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} e^{\frac{1}{T}\mathrm{U}(\boldsymbol{\theta};\mathbf{X})}$. Equation (3.25) shows that the Gibbs distribution introduced in Equation (3.10) of the main text is proportional to the posterior distribution by an unknown scaling factor.

3.3.7 Infer edge direction

It is well justified to regard PPI scores as additive scores for pathway inference [113]. However, other than defining the pathway identification problem as searching for a simple path or multiple intersected paths in a deterministic flavor by maximizing the summation of edge scores, we propose to perceive the complicated interactions and crosstalk between intracellular signaling pathways from a probabilistic point of view. In simple scenarios where we only consider edge scores (node score is not considered), the pathway sampling problem can be defined as follows: given proper starting node of the path (e.g. membrane proteins) that we refer to as ‘source’, and ending node of the path (e.g. transcription factors) that we refer to as ‘target’ (or 'sink'), we sample a pathway of given length L (number of nodes) according to certain function (e.g., summation or Gibbs distribution) of all its edge scores. Figure 3.5 (a) depicts a simple network of seven nodes

where node A is the source and node D is the sink. The edge scores and the true edge directions are also demonstrated in the figure.



(a) (b)

Figure 3.5. Illustration of sampling pathways from PPI network.

Assume that we are interested in studying pathways of length four between source A and sink D. There are in total three possible paths of length four which are: ABCD, AEFD and AGFD. We can calculate the pathway score S for each path as:

$$S(ABCD) = 1+1+1 = 3;$$
$$S(AEFD) = 0.5+0.5+0.5 = 1.5;$$
$$S(AGFD) = 0.5+0.5+0.5 = 1.5.$$

If we sample the three paths according to their pathway score S, the probabilities that each path is being sampled are:

$$P(ABCD) = 3/(3+1.5+1.5) = \frac{1}{2};$$
$$P(AEFD) = 1.5/(3+1.5+1.5) = \frac{1}{4};$$
$$P(AGFD) = 1.5/(3+1.5+1.5) = \frac{1}{4}.$$

Since the events that each path being sampled are mutually exclusive, we can sum up the individual probabilities to obtain the marginal probability of each edge that being sampled. Hence, we have

$$P(AB)=P(BC)=P(CD)=\frac{1}{2},\ P(AE)=P(EF)=\frac{1}{4} \text{ and } P(FD)=\frac{1}{2}.$$

Though edge FD does not belong to the best path with maximum pathway score, it has the same edge probability as those edge members in the best path due to its unique topological importance. It is easy to observe that through the sampling process, we will highlight edges that either belong to paths with large overall pathway scores or have strong evidence to participate in crosstalk between different pathways. Until now, we have explained the sampling model for pathway identification and we will further show how the edge directions can be inferred using the example in Figure 3.5 (b).

In Figure 3.5 (b) the underlying true edge directions are illustrated using arrows which in reality are unknown to us. There are altogether six paths of length six between source A and sink F. The path AEDCBF has the largest pathway score of 3.5 and all other paths have score 2.5. Note that in this particular demo, three edges within the best path are in the wrong direction. If we evaluate the quality of pathway solely based on edge scores, it is impossible for us to infer the correct edge directions. However, through the sampling process as we have previously defined, we are able to detect the correct signal flow along a single path and hence infer the edge directions. Similar as we did for the example in Figure 3.5 (a), we calculate the pathway scores for all six possible paths and further calculate the individual probability of each pathway as:

$$\mathrm{P(AEDCBF)} = \frac{7}{32},$$

$$\mathrm{P(ABCDEF)} = \mathrm{P(AGCDEF)} = \mathrm{P(AGHDEF)} = \mathrm{P(AGHIEF)} = \mathrm{P(AGHIJF)} = \frac{5}{32}.$$

We can aggregate the pathway probabilities into edge probabilities and get marginal distribution of the edges as:

$$\mathrm{P(AE)=P(ED)=P(DC)=P(CB)=P(BF)} = \frac{7}{32},$$

$$\mathrm{P(AB)=P(BC)} = \frac{5}{32},\ \mathrm{P(CD)} = \frac{10}{32},\ \mathrm{P(DE)} = \frac{15}{32} \text{ and } \mathrm{P(EF)} = \frac{20}{32}.$$

From the above calculation, we can see that the single path with largest score have three edges (ED, DC, CB) with wrong directions. However, the proposed sampling method can detect this 'abnormality' of edge direction, yielding that the edges with correct directions have larger chance to be sampled, e.g. P(CD)>P(DC) and P(DE)>P(ED). The ability to infer edge directions is coming from the usage of network topology. It is the cross-talk between different paths by sharing common edges provides us with extra information about the edge direction.

Figure 3.5 (b) gives a simple example with only a dozen of connections to show how pathway directions can be inferred from un-directed PPI connections. When working on real biological data, one may deal with thousands of nodes and tens of thousands of edges, where Gibb sampler may be preferred over exhaustive search to approximate pathway distributions. After the Gibbs sampler converges, GIST pools the samples together to estimate edge directions. We introduce variable $e_{i,j}$, where $e_{i,j}=1$ denotes a directed edge from protein $\omega_i \in \Omega$ to protein $\omega_j \in \Omega$. The underlying true probability of $e_{i,j}$ is determined by the probability distribution of pathway $\boldsymbol{\theta}$, which can be theoretically defined by:

$$p^*_{i,j} = P(e_{i,j}=1) = \sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} P\left(e_{i,j}=1\middle|\boldsymbol{\theta}\right) P(\boldsymbol{\theta}), \tag{3.26}$$

where $P\left(e_{i,j}=1\middle|\boldsymbol{\theta}\right)=1$ if $e_{i,j}$ corresponds to a connected edge in pathway $\boldsymbol{\theta}$; otherwise 0. Equation (3.26) models the probability of each directed edge as a Bernoulli random variable with a success rate $p^*_{i,j}$. A truncated version of $p^*_{i,j}$ is also made available to favor practice when users want to focus on top ranked pathway with large potentials (Section 3.3.8). Now that we have calculated the probability of a directed edge ($e_{i,j}$), we define the confidence of edge direction as:

$$q\left(e_{i,j}\right) = \frac{p^*_{i,j}}{p^*_{i,j}+p^*_{j,i}}, \text{ s.t. } \max\left(p^*_{i,j}, p^*_{j,i}\right) \neq 0, \tag{3.27}$$

where $q\left(e_{i,j}\right)\in[0,1]$. If $q\left(e_{i,j}\right)$ is close to 1, it means the signal goes from protein ω_i to ω_j almost surely, while $q\left(e_{i,j}\right)=q\left(e_{j,i}\right)=0.5$ means lack of confidence to determine the direction of signal flow. Therefore, GIST not only assesses the importance of any individual edge, but also provides the confidence measure for edge directions.

3.3.8. Truncation of pathways samples to estimate edge probability

When number of proteins in Ω is large, the sample space $\boldsymbol{\Theta}$ (e.g., for L>8) becomes huge. In this case, it is computationally infeasible to enumerate all possible pathways to calculate $p^*_{i,j}$. Practically, computational biologists are more interested in

top ranked hundreds of samples (from totally tens of thousands of samples) with highest differential power and edge correlations. Therefore, we offer an un-normalized edge probability score based on truncated samples as follows:

$$p_{i,j}^{K} = \sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^{K}} \left(V_1(\boldsymbol{\theta}) + V_2(\boldsymbol{\theta})\right) \cdot P\left(e_{i,j}=1 \middle| \boldsymbol{\theta}\right) \sim \frac{\sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^{K}} \left(V_1(\boldsymbol{\theta}) + V_2(\boldsymbol{\theta})\right) \cdot P\left(e_{i,j}=1 \middle| \boldsymbol{\theta}\right)}{\sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^{K}} \left(V_1(\boldsymbol{\theta}) + V_2(\boldsymbol{\theta})\right)} = P^{K}(e_{i,j}=1), \quad (3.28)$$

where $\boldsymbol{\Theta}^{K}$ denotes top K truncated pathway samples. In Equation (3.28) we no longer take exponential form of potential functions to make $p_{i,j}^{K}$ follow Gibbs distribution for practical purpose to emphasize sub-optimum pathways (we only use Gibbs distribution in sampling process to sample pathways with large potentials from huge sample space). Technically $p_{i,j}^{K}$ is not a probability but rather a score function that equals the true edge probability by a linear transformation. In Equation (3.28) we purposely downplay prior knowledge by only using node and edge potentials, because samples in $\boldsymbol{\Theta}^{K}$ have already been well constrained by prior knowledge (e.g., cellular location) through Gibbs sampling. Consequently, the corresponding confidence of edge direction is given by:

$$q^{K}\left(e_{i,j}\right) = \frac{p_{i,j}^{K}}{p_{i,j}^{K} + p_{j,i}^{K}} = \frac{P^{K}(e_{i,j}=1)}{P^{K}(e_{i,j}=1) + P^{K}(e_{j,i}=1)}, \text{ s.t. } \max\left(p_{i,j}^{K}, p_{j,i}^{K}\right) \neq 0 . \quad (3.29)$$

3.4. Structural Organization to Uncover pathway Landscape (SOUL)

One major component of IMPACT is SOUL (Structural Organization to Uncover pathway Landscape), which is a post-processing algorithm that further investigates

sampled pathways to hypothesize aberrant pathway modules and their inter-connections. Instead of ranking linear pathways according to their scores, SOUL aims to prioritize pathways as topologically correlated clusters, namely pathway modules. It is especially interesting to examine 1) what are the common functions of pathways within the same module? 2) What are the possible diversities (i.e. alternative pathways) to realize the functionality in the same pathway module? 3) How different pathway modules are inter-connected and crosstalk with each other? To our best knowledge, none of the existing pathway identification methods have formally addressed the above questions, which are nevertheless very important for the understanding of complex disease (e.g., cancer) at system level. Li *et al*. [129, 130] clustered canonical pathways based on shortest distance profile and identified global pathway configurations. We propose novel adaptation to Li's strategy by reconstructing 'pathway landscape' from structural clusters and potential distributions of pathway samples estimated by GIST algorithm.

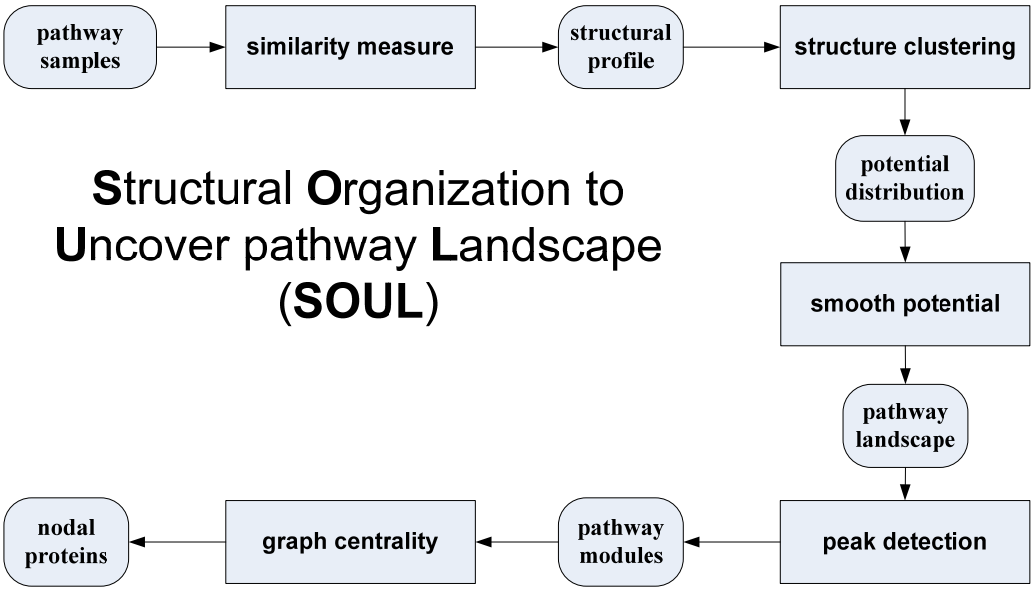


Figure 3.6. Block diagram of the SOUL algorithm.

Figure 3.6 gives a block diagram of the SOUL algorithm, which consists of the following steps:

1. Pool pathways samples together. Calculate the structural profile $\mathbf{d} = \{d_{ij}\}$ between any two pathway samples $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$. d_{ij} is the number of overlapped proteins.

2. Perform hierarchical clustering based on structural profile $\mathbf{d}$ and obtain cluster label c. Re-organize structural profile $\mathbf{d}$ (both row-wise and column-wise) according to label c and obtain an organized structural profile $\mathbf{d}_c$. We visualize $\mathbf{d}_c$ to generate the structural heatmap. Re-order the pathway potential to get the potential distribution.

3. The potential distribution is in the form of discrete peaks. Smooth the potential (e.g., using moving average filter) to obtain a continuous response as the new distribution. This continuous distribution is referred to as 'pathway landscape' in this paper.

4. Select distinct local peaks from pathway landscape. The pathways associated with these peaks are the identified pathway modules.

5. For the identified pathway modules, use betweenness centrality [131, 132] to prioritize nodal proteins. We hypothesize that the proteins with highest centrality scores are potential candidates of nodal proteins due to their topological importance in the pathway network.

3.5. Simulation Results

3.5.1. Simulation design

We evaluated the performance of GIST for pathway identification on synthetic datasets. Based on alternative pathway models from Gong *et al*. [133], we summarized two basic pathway structures: type I structure and type II structure. Type I structure refers to wiring of alternative pathways between single source protein and single target protein, while type II structure is established among multiple sources and targets. PPI data from Human Protein Reference Database [38, 111] and canonical pathways from KEGG were downloaded to build simulation network and ground truth pathways. We compared GIST method with existing methods in terms of identifying or retrieving both pathway nodes and edges. Different levels of noise in expression data and false positive connections in PPI were simulated to comprehensively evaluate the robustness of competing methods. Precision-Recall curve and average precision (AP) were used as performance measures in this study [79].

3.5.2. Alternative pathway structures and pathway crosstalk

On the structure of alternative pathways, Gong *et al*. have proposed five basic types of alternative pathways including: A (divergent), V (convergent), O(single/multiple), M (multiple/multiple) and N (nested) [133]. We have further summarized two types of fundamental alternative pathway structures as: 1) alternative pathways between single source and single target (type I); 2) alternative pathways between multiple source(s) and multiple target(s) (type II). Figure 3.7 gives a schematic diagram of the two pathway structures. Type I pathway structure can be regarded as a special case of type N (nested) pathway between single source protein and single target protein, which embraces type O (single/multiple) pathways as sub-components. Type II pathway is a more general case of type N (nested) structure between multiple source proteins and multiple target proteins, which is build upon a mixture of type A (divergent), type V (convergent) and type M (multiple/multiple) pathway components.

Both type I and type II pathway structures are designed to study alternative signal transduction, while the latter one is also used to model crosstalk among multiple pathways. For example, Saito studied crosstalk among three MAPK pathways (pheromone response, filamentous growth response, and osmostress adaptation), each started from partially overlapped membrane anchor/sensor proteins (Cdc42, Msb2,Sho1 and Opy2, etc.) and sent signal to transcription factors (Ste12, Tec1, Sko1, Hot1, Msn2/4 and Smp1), to implement different biological functions [134].

To generate simulation data involving two types of pathway structures, a subnetwork centered at some putative hub proteins (tumor protein 53) from human PPI network was selected as the base topology. Proteins that are involved in canonical pathways (MAPK, ERBB, JAK/STAT, et al.) were also extracted from knowledge database [104, 135] as the candidate pool for building ground truth pathways. We also collected subcellular location information for human proteome and assumed that a valid path should start form extracellular space/membrane and end in nucleus. A mixture model was employed to synthesize the edge z-scores plus different levels of noise so that we can simulate real biological scenarios with experimental/biological noise. Exhaustive search was conducted to obtain the ground truth distributions of both nodes and edges.

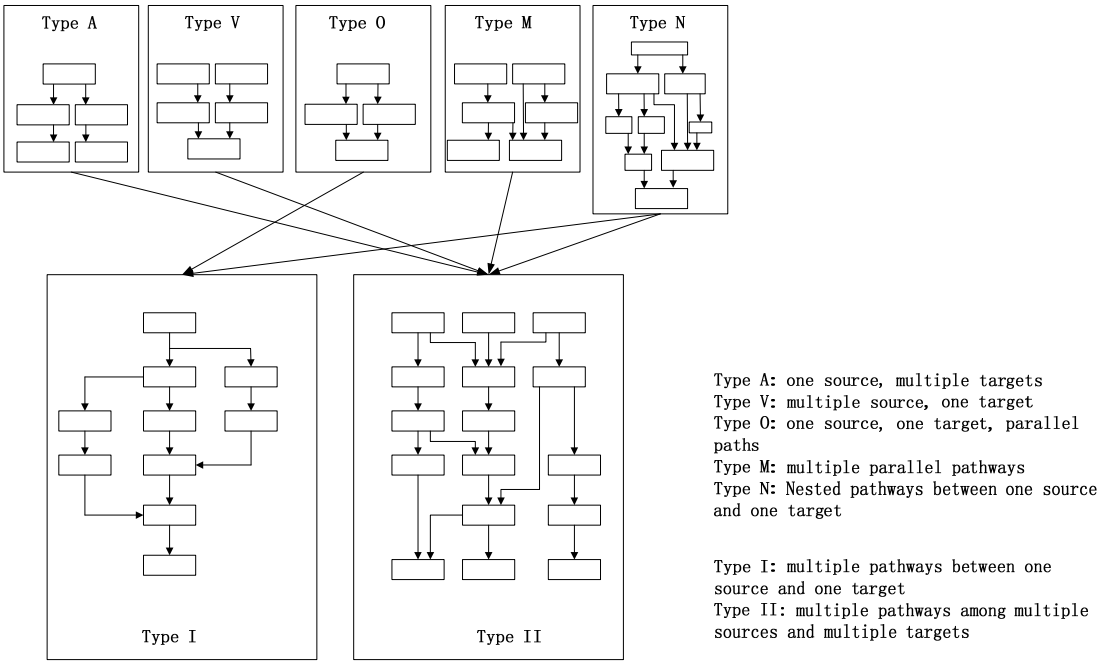


Figure 3.7. Type 1 and type 2 alternative pathway structures.

3.5.3. Performance comparison under different noise settings

We compared the proposed GIST algorithm with three existing methods that are random color coding [113], edge orientation [122] and integer linear programming (ILP) [110]. We used the improved version of random color coding called Faspad [114] to identify linear pathways with maximum product of edge probabilities. For edge orientation method, we used random orientation followed by a local search, which was demonstrated as more effective and scalable compared to MIN-SAT and MAX-CSP algorithms. The limitation of the ILP method is that it only identifies one single pathway (network) with best score. The result of the algorithm also highly depends on the size parameter λ. We adjusted λ smoothly in certain range and aggregated the identified results for comparison. We only compared GIST with ILP in node identification because ILP cannot infer edge directions.

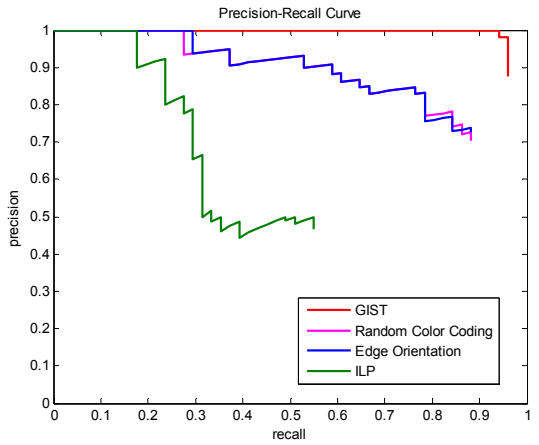
Our goal is not only to recover the 'best' linear pathway in the PPI network, but also to identify alternative pathways (suboptimal solutions) between given sources and targets. For type I structure, the initial network contained 261 proteins and 998 edges while for

type II case, the initial network contained 266 proteins and 1,026 edges. Table 3.1 summarizes the average precision of four algorithms on two simulation studies.

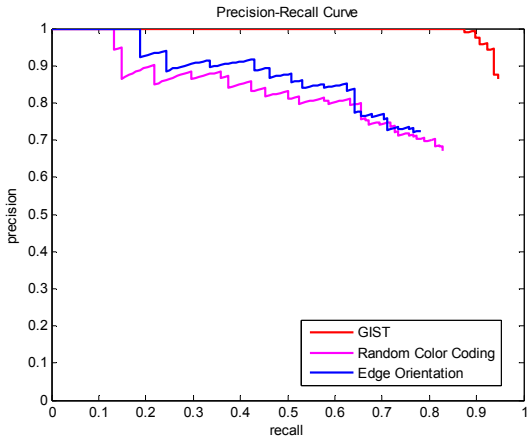
From Table 3.1, we can see that the GIST algorithm had the most desirable performance in identifying relevant pathway nodes and edges for both type I and type II structures. This was mainly due to incorporation of subcellular location to filter out pathways with inconsistent flow. When the level of noise was set to 0.2, GIST achieved about 0.16 (0.15) increase of average precision in node identification for type I (II) pathway, and an even larger improvement of 0.24 (0.17) in average precision for edge identification. As the level of noise increased, all methods suffered from performance degradation while GIST was still able to maintain better detection power. From Table 3.1, Figure 3.8 and Figure 3.9, we observed that random color coding and edge orientation had comparative performance in most cases. The ILP method, though successfully identified top-ranked nodes and edges with high precision, was not able to capture suboptimal pathways as the other three methods did.

Table 3.1. Average precision for pathway identification under different noise settings.

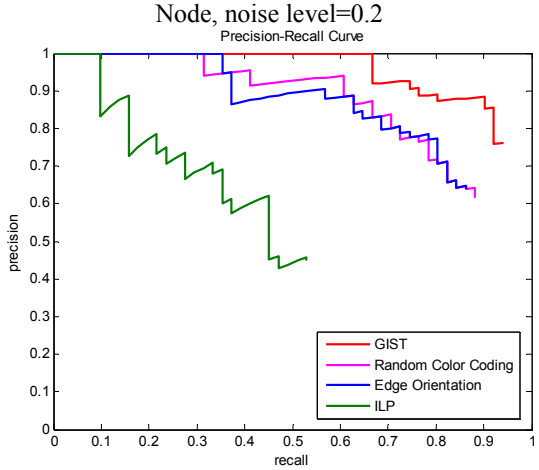
Average precision	Noise Level	Structure 1		Structure 2	
		node	edge	node	edge
GIST	0.2	**0.9604**	**0.9424**	**0.9002**	**0.8359**
Random Color Coding		0.8062	0.7013	0.8509	0.6686
Edge Orientation		0.8066	0.6951	0.7772	0.4728
ILP		0.4059	N/A	0.5860	N/A
GIST	0.5	**0.9098**	**0.8353**	**0.8058**	**0.6902**
Random Color Coding		0.8010	0.5916	0.7684	0.5678
Edge Orientation		0.7797	0.5689	0.7375	0.4412
ILP		0.3873	N/A	0.5704	N/A
GIST	0.8	**0.8425**	**0.6985**	**0.8004**	**0.4657**
Random Color Coding		0.7551	0.4805	0.7146	0.4167
Edge Orientation		0.7163	0.4715	0.6774	0.3509
ILP		0.3638	N/A	0.4853	N/A



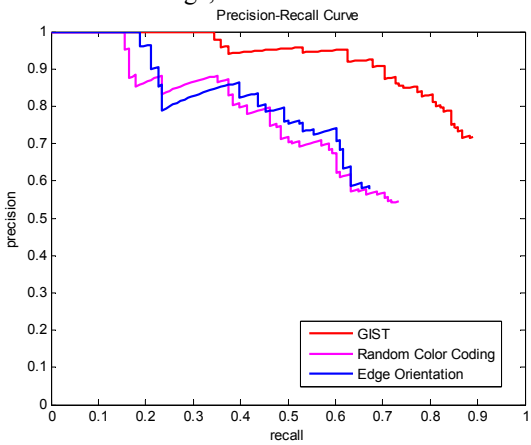
Node, noise level=0.2



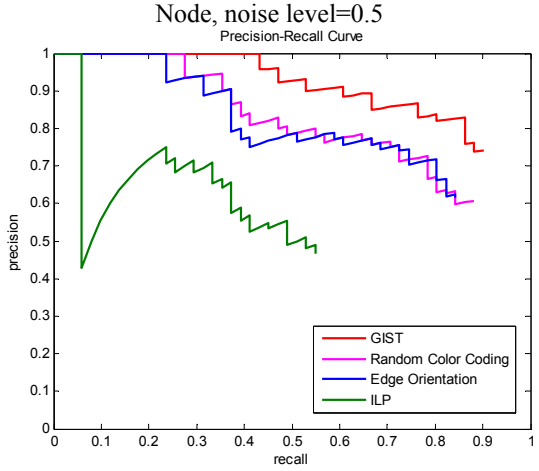
Edge, noise level=0.2



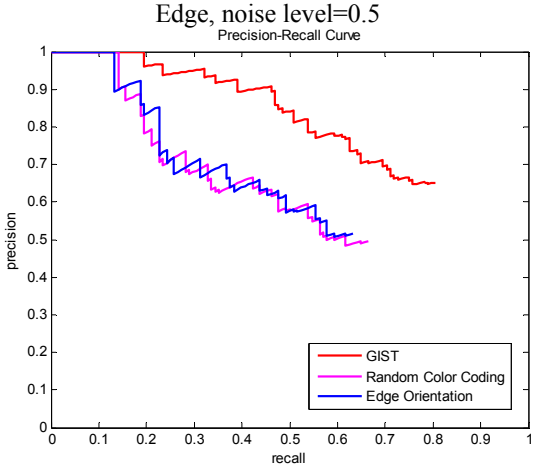
Node, noise level=0.5



Edge, noise level=0.5

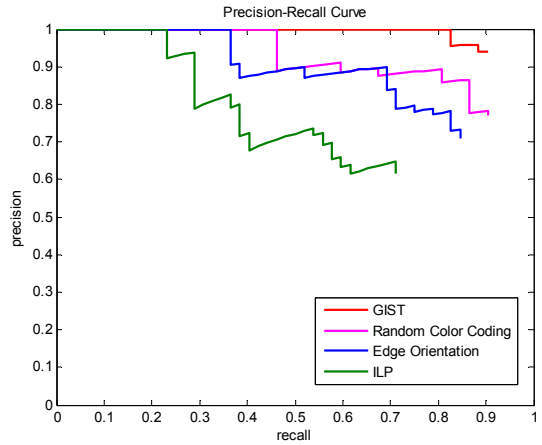


Node, noise level=0.8

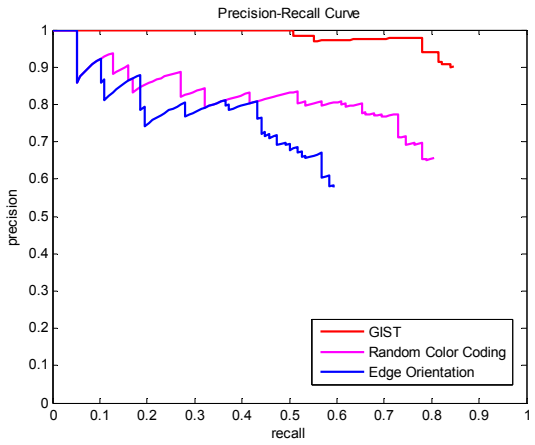


Edge, noise level=0.8

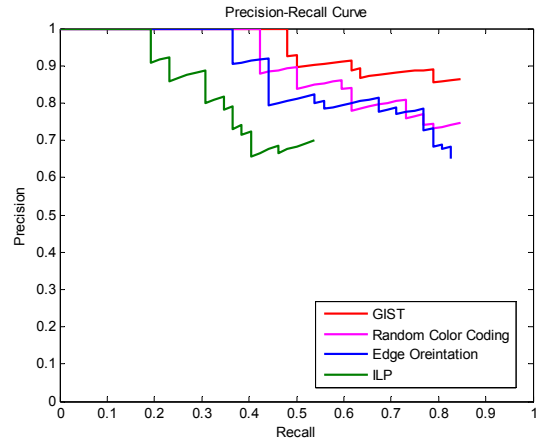
Figure 3.8. Precision-Recall curve for node/edge identification on type I pathway structure under different noise level.



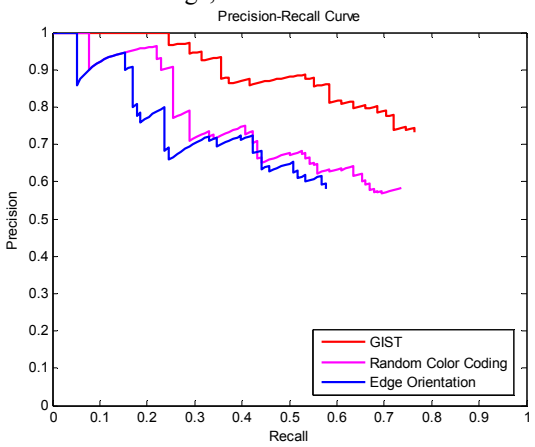
Node, noise level=0.2



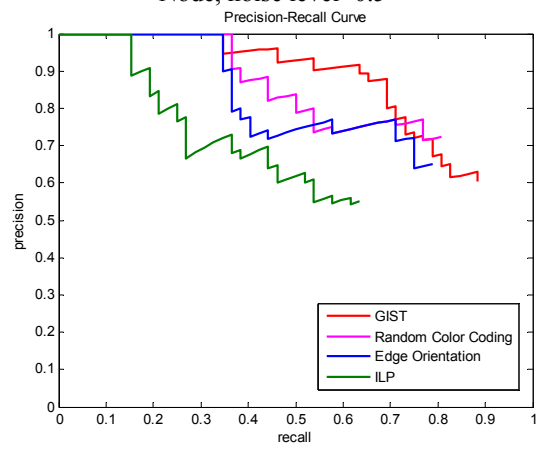
Edge, noise level=0.2



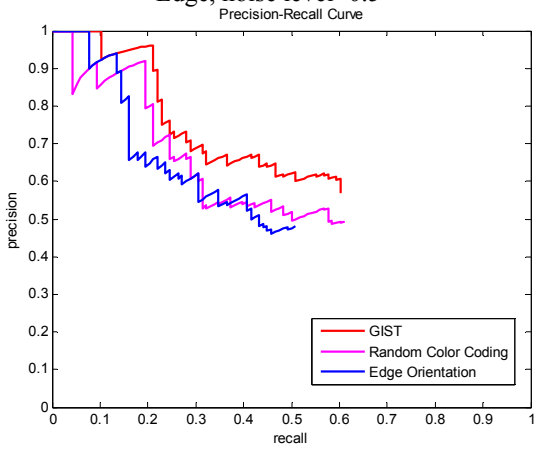
Node, noise level=0.5



Edge, noise level=0.5



Edge, noise level=0.8



Edge, noise level=0.8

Figure 3.9. Precision-Recall curve for node/edge identification on type II pathway structure under different noise level.

3.5.4. Evaluating robustness against false positives in PPI

In addition to comparing node and edge identification under different noise settings, we also studied how false positive connections in PPI network affected pathway identification. We introduced a matrix called Percentage of False Connections (PFC) as:

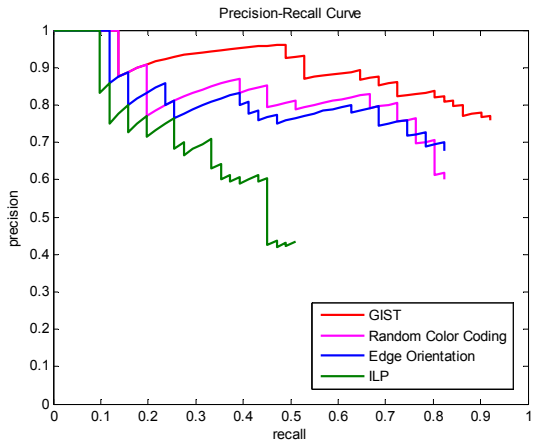
$$\text{PFC} = \frac{\text{number of added false connections}}{\text{number of connections in original PPI}} \times 100\% ,$$

to measure how many false positive connections (edges) were added into the original PPI network. In our simulation, we specifically studied the influence of false positives by generating PPI networks with 10%, 25% and 50% of false connections under moderate noise level (0.5). We used average precision to measure the performance of the competing methods in retrieving ground truth nodes and edges. Table 3.2 summarizes the average precision of four methods (GIST, random color coding, edge orientation and integer linear programming) for both type I and II pathway structures. We also plotted precision-recall curves in Figure. 3.10 and Figure 3.11.

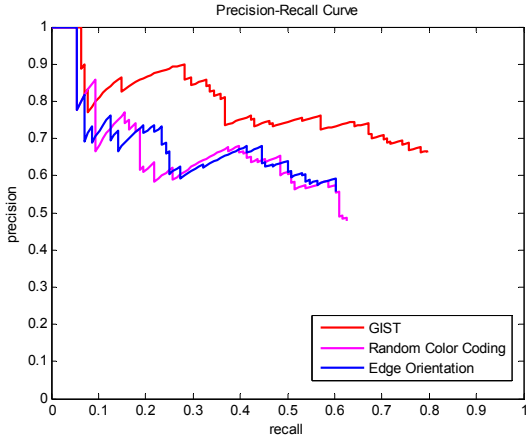
From Table 3.2, Figure 3.10 and Figure 3.11, we can clearly see that the performance of all four methods degrades as more false positive connections are introduced to PPI network. However, the proposed GIST algorithm always maintains comparable or better performance in all cases. One interesting observation is that despite relatively minor advantages in identifying nodes in the ground truth pathway, GIST achieved remarkably better performance in directed edge identification with an at least 0.1 increase in average precision (e.g., Table 3.2, type II structure). This suggests a potential gap between correct identification of nodes and directed edges, whereas GIST algorithm has distinguished itself by offering a high power to infer edge directions. From synthetic data analysis, we conclude that GIST is more preferable in identifying alternative directed edges from noisy expression data and imperfect PPI data, showing its efficacy to infer complex signaling pathways.

Table 3.2. Average precision for pathway identification under different proportion of false positive connections (PFC) in PPI network.

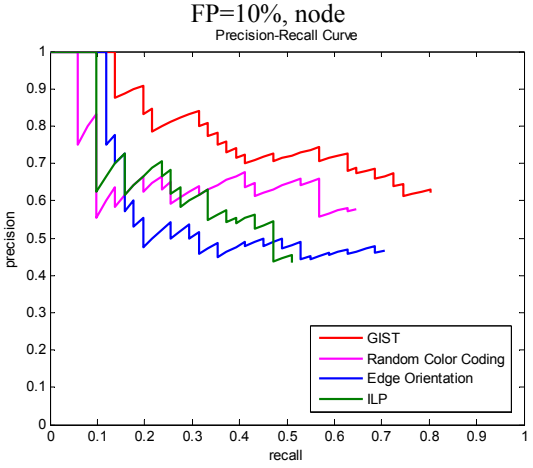
Average precision	PFC	Structure 1		Structure 2	
		node	edge	node	Edge
GIST	10%	**0.8348**	**0.6354**	**0.7683**	**0.6601**
Random Color Coding		0.6950	0.4259	0.7157	0.5170
Edge Orientation		0.6753	0.4170	0.7373	0.4900
ILP		0.3719	N/A	0.5165	N/A
GIST	25%	**0.6343**	**0.3092**	**0.7695**	**0.5424**
Random Color Coding		0.4385	0.1321	0.7613	0.4350
Edge Orientation		0.4180	0.1847	0.7013	0.3818
ILP		0.3482	N/A	0.4723	N/A
GIST	50%	**0.4938**	**0.1830**	**0.6823**	**0.5245**
Random Color Coding		0.4238	0.1157	0.6635	0.4101
Edge Orientation		0.3813	0.1360	0.5873	0.3254
ILP		0.3273	N/A	0.4726	N/A



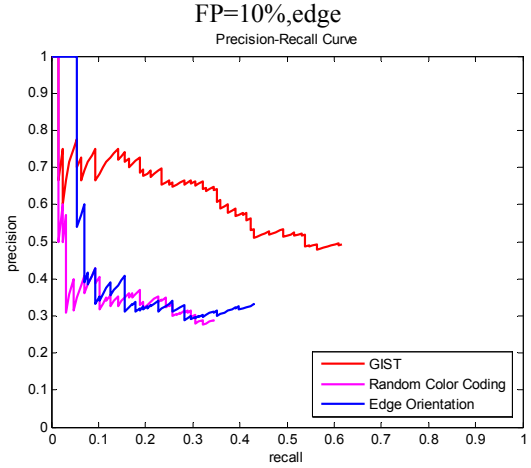
FP=10%, node



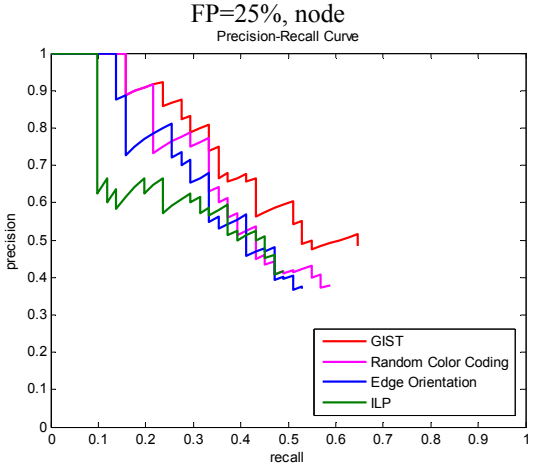
FP=10%,edge



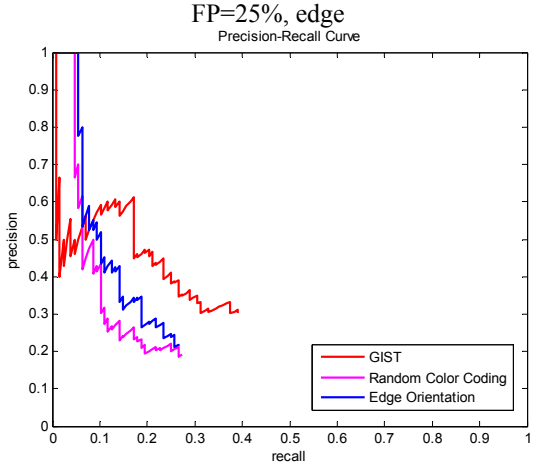
FP=25%, node



FP=25%, edge

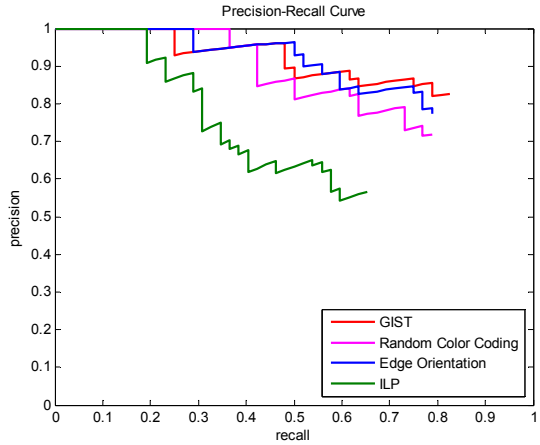


FP=50%, node

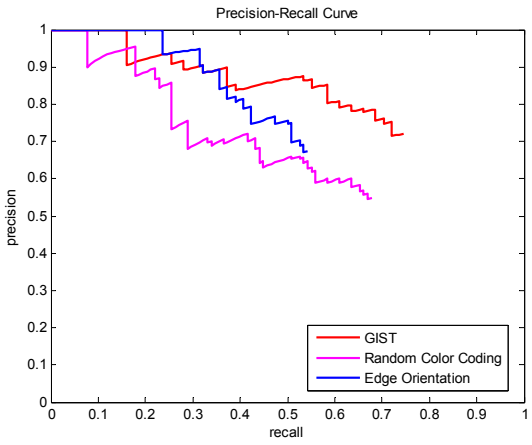


FP=50%, edge

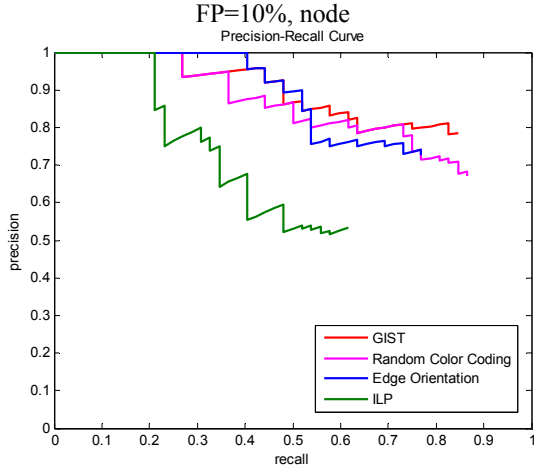
Figure 3.10. Precision-Recall curve for node/edge identification on type I pathway structure under different false-positive rate in PPI.



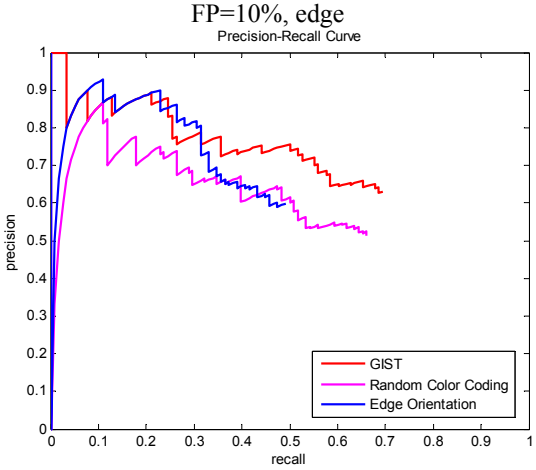
FP=10%, node



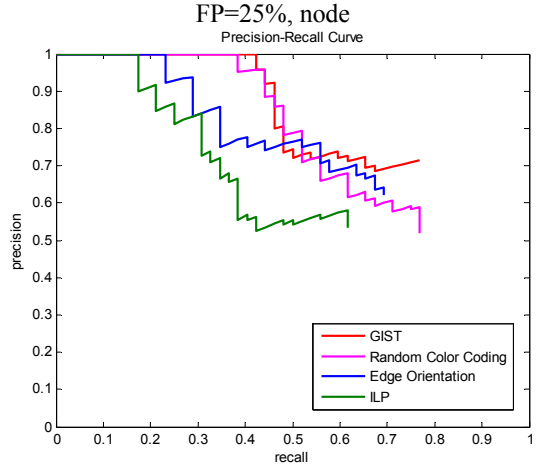
FP=10%, edge



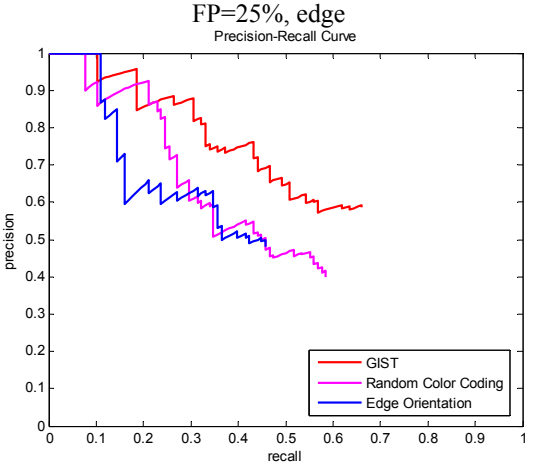
FP=25%, node



FP=25%, edge



FP=50%, node



FP=50%, edge

Figure 3.11. Precision-Recall curve for node/edge identification on type II pathway structure under different false-positive rate in PPI.

3.6. Biological case studies

3.6.1. GIST case study on yeast PPI data

To evaluate the efficacy of the algorithm, we first test GIST on the yeast protein-protein interaction data to identify several MAPK signaling pathways. The PPI network was obtained from the Database of Interacting Protein (DIP) [136-138]which contained 4389 proteins with 14,319 interactions. The calculation of the edge confidence scores is described in [139], which is of high precision in the yeast study. We compared the GIST algorithm with some existing methods such as Integer Linear Programming (ILP) and Random Color Coding on two MAPK pathways, which are pheromone response pathway and filamentous growth pathway.

Figure 3.12 (a) shows the numerical results of the identified pathway of length nine between Ste3 and Ste12. Based on the sampled edge distribution, we selected the edges with the top edge probabilities and recovered the signaling pathway between the given sink and source. We compared our result with the knowledge from KEGG database and the GIST algorithm successfully recovered 16 out of 19 components in pheromone response pathway. The Venn diagram in Figure 3.12 (a) shows that GIST detected all 12 proteins in the KEGG pathway, which were also detected by the other two competing methods. Meanwhile GIST was also able to identify proteins that were supported by the KEGG database but was not reported by either ILP (Gpa1, Dig1 and Dig2) or random color coding (Ste20).

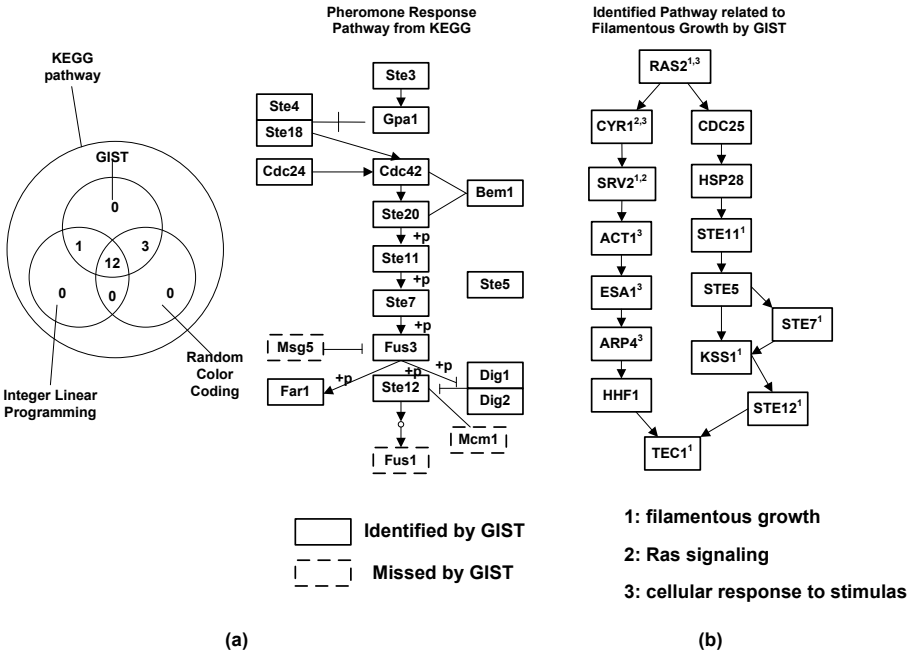


Figure 3.12. Finding MAPK signaling pathways on yeast protein-protein network.

Figure 3.12 (b) shows the identified paths between Ras2 and Tec1. The GIST algorithm successfully detected 6 out of 9 proteins in the canonical filamentation growth pathway and in addition found an alternative pathway including Cyr1, Srv2, Act1, Esa1, Arp4 and Hhf1.By using knowledge from GO Term Finder from Saccharomyces Genome Database and KEGG database, we found that most of the proteins identified by GIST are enriched in functional processes such as filamentous growth, Ras signaling and cellular response to stimulus. This verifies the merit of the GIST algorithm to perceive the pathway interactions from a global level, taking consideration the potential crosstalk between multiple pathways.

3.6.2. Convergence of Gibbs sampler: a case study on yeast expression dataset

One important problem regarding Gibbs sampling is the convergence of the Markov chain. Hereby we used gene expression data from yeast as a case study to evaluate the convergence of the proposed GIST algorithm. Previous example to identify MAPK signaling pathways utilizes yeast PPI confidence scores, which are quite precise in the yeast case. Nevertheless, we do not have the luxury to derive reliable edge scores in many other organisms, especially for Homo sapiens. With recent availability of high-throughput microarray expression data, we are able to estimate the edge scores from the

expression profiles. One feasible scheme of calculating the edge score is to use the correlation coefficients based on the gene expression [110]. We collected phosphate metabolism data which measure expression level of 6546 Cy3- or Cy5-labeled cDNA probes for eight yeast samples under the low and high-Pi conditions [140]. We assign the source node as Ras2 and the target as Tec1. We ran 200 short Gibbs chains each with 100 iterations.

Figure 3.13 (a) shows the results of the identified paths between the Ras2 and Tec1. The node size represents the node probability of the protein and the edge width is proportional to the edge probability. The identified pathway components are highly enriched in functional processes such as cellular response to stimulus, ribosome biogenesis and signal transductions. More specifically, GIST successfully detects six proteins (STE4, STE5, KSS1, STE12, RAS2 and TEC1) related to filamentous growth of unicellular organism. Moreover, the cAMP metabolic and biosynthesis process pathways (IRA2, RAS1 and RAS2) are also enriched. It has been studied that both MAP kinase and cAMP can affect filamentation process of yeast cells, as mutations of either pathway will affect a cell surface protein, which again supports the crosstalk between two pathways [141].

To test the convergence of the GIST algorithm, we have plotted the convergence curve of the node probability as is shown in Figure 3.13 (b). Because the ground truth is not available to us, we have to compare the estimated distribution of the top ranked genes at each iteration to that of the final distribution (after 100 iterations). The Cumulative Distribution Function (CDF) is calculated and we design the following similarity score as:

$$\text{Similarity score} = \left| F_1(x) - F_2(x) \right|^2, \tag{3.30}$$

where $F_1(x)$ and $F_2(x)$ are the CDF function of two distributions. From Figure 3.13(b) we can see that the GIST algorithm has good convergence behavior and after about 50 iterations, the sampler has reached the stationary point and is drawing samples from a stable distribution. We also plot the estimated node probability of all the genes in Figure 3.13 (c). The estimated distribution is very ‘spiny’ and those tall ‘spikes’ correspond to genes that are most likely to be involved in a particular pathway process between the given source and target.

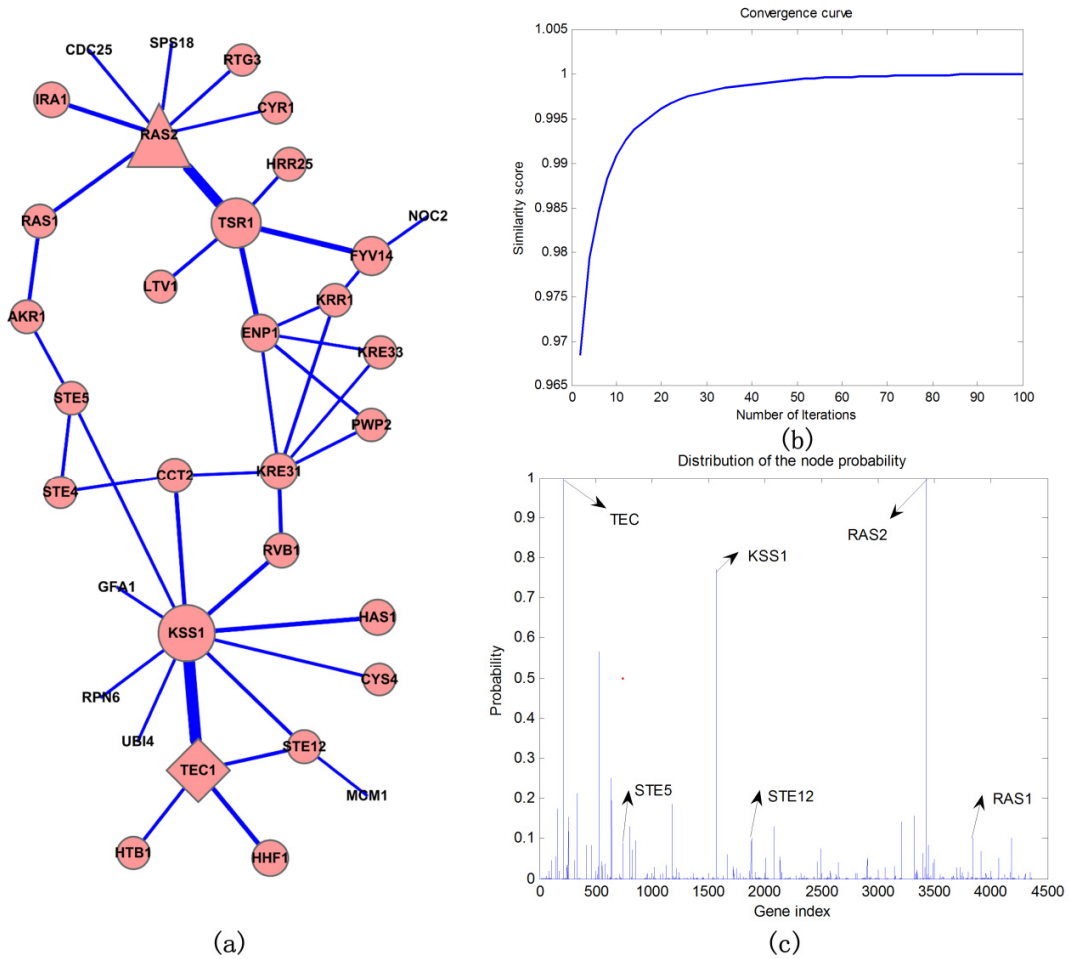


Figure 3.13. Applying GIST on phosphate metabolic expression dataset to study filamentous growth signaling.

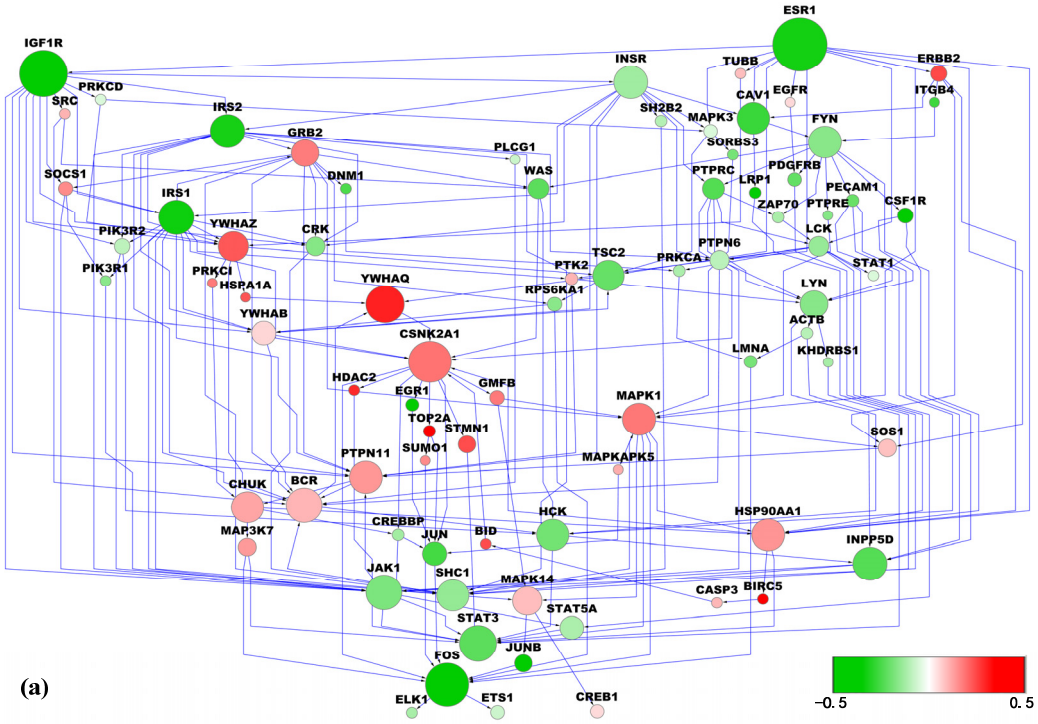
3.6.3. Breast cancer study using IMPACT

Aberrant signal transduction in Tamoxifen treated patient data from Loi et al.

We first applied the proposed pathway identification method IMPACT on a gene expression dataset from Jules Broad Institute to identify Tamoxifen resistance-related pathways in breast cancer [64]. The microarray dataset (termed Loi data here) was profiled on Affymetrix U133A platform and we used PLIER algorithm from Affymetrix Expression Console to normalize all samples in each dataset (http://www.affymetrix.com). We observed quite severe institutional batch effect on the Loi dataset and used ComBat algorithm from GenePattern for correction [142]. A 5-year cutoff on distant-metastasis-free-survival (DMFS) was used to divide samples into 'early

recurrence' group (DMFS ≤ 5 years)) and 'late recurrence' group (DMFS > 5 years), which finally yielded 88 'early recurrence' samples and 92 'late recurrence' sample in the Loi dataset.

We were particularly interested in estrogen receptor (ER or ESR1) related signaling events and their connections to downstream functional processes (e.g., cell cycle and apoptosis) in the context of drug resistance. We designed three functional case studies to discover: 1) ER signaling pathway, 2) cell cycle pathway and 3) apoptosis pathways, where the later two case studies would give us further insights about how message from ER signaling pathways would lead to cancer cell growth because of cell proliferation and apoptosis.



(a)

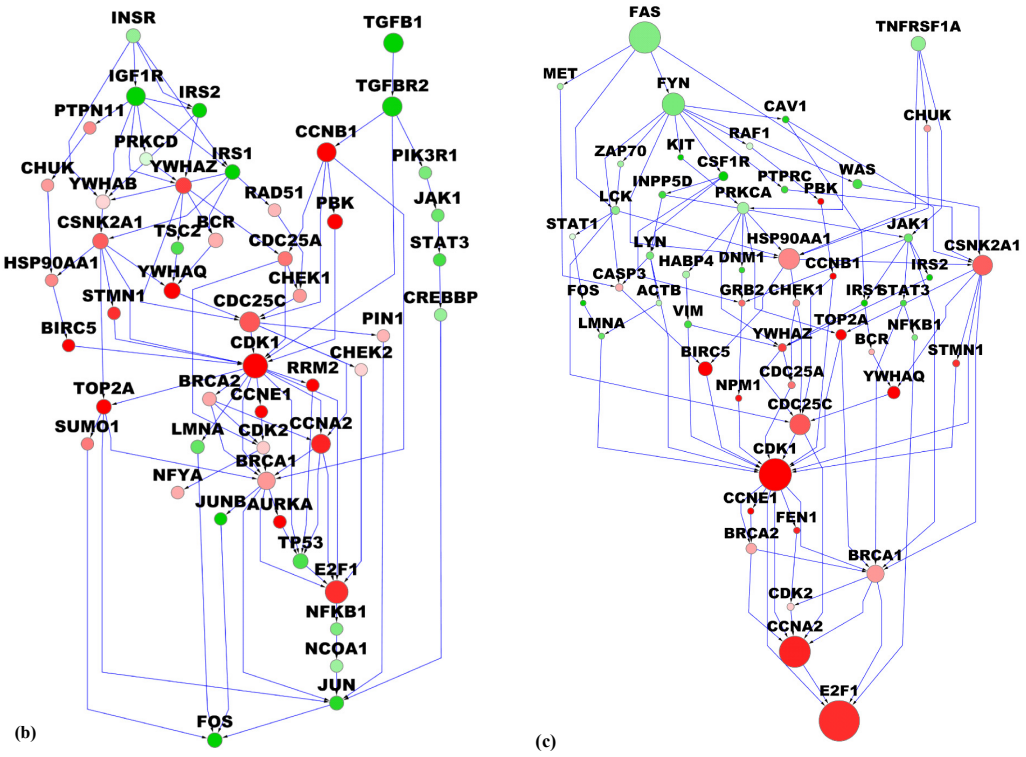


Figure 3.14. Identified pathway networks from Loi's dataset. (a) ER signaling pathway network. (b) Cell cycle pathway network. (c) Apoptosis pathway network. Node color represents log2 fold-change of gene expression between early and late group of patients (red: over-expressed in early group; green over-expressed in late group).

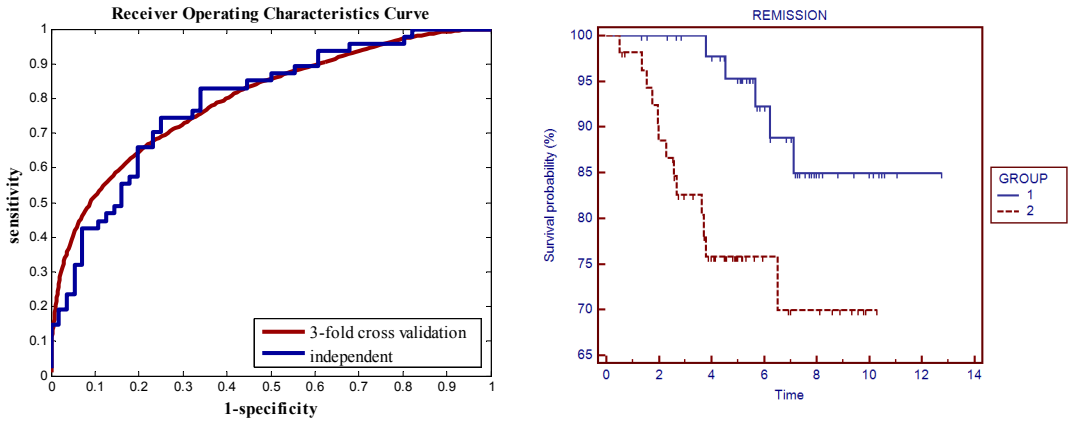


Figure 3.15. Prediction analysis of ER signaling pathways identified from Loi data. (a) 3-fold cross validation on ER signaling pathway network identified from Loi was performed, which gave an AUC of 0.8. Independent test of Loi ER signaling pathways on Symmans data had an AUC of 0.79. (b) Kaplan Meier plot [143]of Symmnas data predicted by pathway network from Loi data. Group 1 is the 'late recurrence' group and group 2 is the 'early recurrence' group. The hazard ratio is 3.26.

We first applied the GIST algorithm to reconstruct signaling pathways related to estrogen receptor signaling, which connected source protein (ESR1) and identified transcription factors from GibbsOS. Figure 3.14 (a) is the inferred estrogen signaling pathway network between ESR1 and nuclear transcription factors, which is assembled by collapsing top 200 linear pathway samples from the GIST algorithm. The size of a node is proportional to the probability of the protein that being sampled. Among totally 79 proteins in the pathway network, seven of them are transcription factors (JUN, FOS, STAT1, STAT3, STAT5A, ELK1 and ETS1) with significant regulatory activities that are identified by the GibbsOS method. We tested the prediction performance of the 79-gene pathway network on an independent dataset from MD Anderson Cancer Center [65], which yielded an area-under-the-curve (AUC) of 0.79 for receiver-operating-characteristic (ROC) curve [144] and a hazard ratio [145] of 3.26 (Figure 3.15).

Figure 3.14 (a) has revealed complicated wiring of alternative pathways that are interconnected through frequently sampled cytoplasmic proteins such as IRS1/2, JAK1, YWHAZ, CSNK2A1, MAPK1 and HSP90AA1. We hypothesize that these proteins may play critical roles in exchanging message among different functional modules, which in other words are responsible for activating or suppressing alternative pathways or initiating pathway crosstalk. We uploaded the 79 proteins to the Database for Annotation, Visualization and Integrated Discovery (DAVID) [41, 42] for functional annotation and pathway enrichment analysis. The identified pathway network are highly enriched in insulin signaling pathway (Benjamin p-value=2.4E-10), ErbB signaling pathway (Benjamin p-value=4.0E-13), MAPK signaling pathway (Benjamin p-value=5.1E-8), and JAK-STAT signaling pathway (Benjamin p-value=2.0E-5). Pathway enrichment analysis has depicted a diverse picture where multiple signaling pathways contribute to recurrence of breast cancer, which supports the increasing acknowledgement of alternative pathways and pathway crosstalk in cancer.

We also applied the GIST algorithm to identify cell cycle and apoptosis related pathways. We selected a few breast cancer cell cycle related membrane proteins or growth factors (EGFR, TGFB1, IGF1R, INSR, FGFR1) from literature as source proteins for this cell cycle study, while canonical death receptors (IL1R1, FAS and TNFRSF1A) were selected as source proteins for the apoptosis study. GibbsOS was applied on cell

cycle and apoptosis related motifs to identify significant transcription factors as target proteins. Figure 3.14 (b) and (c) show the identified signaling networks for cell cycle and apoptosis from Loi data. Our analysis for both cell cycle and apoptosis pathways in early recurrence tumors show increase in proteins associated with signal transduction (YWHAQ, YWHAZ, PTPN11), chaperone proteins (HSP90AA1, HSPA1A) and proteins involved in cytoskeletal rearrangement (STMN1, SRC). Increase in demand in cell proliferation in early recurrence tumors may be supported by their cumulative functions. In late recurrence tumors, however, there is an increase in proteins associated with ER signaling (ESR1, IGF1R, IRS1, IRS2, CAV1), tyrosine kinases (FYN, LYN, HCK), phosphatases (PTPRC, INPP5D) and JAK-STAT pathway (JAK1, STAT5A, STAT3). Expression of ER signaling pathway proteins in tumors that recur late may explain why these tumors initially respond to antiestrogens but later use the signaling revealed in this pathway to increase cell proliferation, hence, to recur.

Pathway modularization and crosstalk identified from Loi data

To address how ER signaling pathways interact with cell cycle and apoptosis pathways, we further applied SOUL to investigate significant pathway modules and their inter-module crosslink. Pathway samples obtained from three GIST case studies were pooled to form a 'pathway landscape' centered on estrogen receptor (Figure 3.16). We first applied hierarchical clustering to group pathways according to their structural similarity, yielding a re-organized pathway topological pattern referred to as 'pathway structural heatmap' (Figure 3.16 (a)). Structural heatmap presented two major pathway components with different functional emphasis ('ER signaling' versus 'cell cycle and apoptosis'). Despite the clear boundary between the two major pathway components, we also observed highlighted spots in the off-diagonal regions, suggesting potential connections between ER signaling pathways and its downstream machineries (i.e., cell cycle and apoptosis). We re-shuffled the potentials of the individual pathways according to the structural clusters in Figure 3.16 (a), which formed multimodal distribution (Figure. 3.16 (b)). We referred to the re-organized potential distribution as the 'pathway landscape'. Multiple local peaks in the pathway landscape indicate modularization of linear pathways. We hypothesize that functionally related or complementary linear

pathways are coupled as pathway modules, whereas different pathway modules crosstalk through signaling hubs (nodal proteins). We identified four pathway modules with local peak potentials in Figure 3.16 (b), two from ER signaling pathways and the other two from cell cycle and apoptosis pathways (Appendix B.1). Though starting from common source protein ESR1, pathway module 1 and module 2 have quite distinct local configurations. Proteins in module 1 are more enriched in response to hormone, which contains several members from the canonical MAPK signaling pathway (HSPA1A, JUN, RPS6KA1, STMN1 and FOS) and insulin signaling pathway (IRS1, IRS2 and TSC2). On the other hand, pathways in module 2 are converged to JAK-STAT pathway (JAK1, STAT3 and STAT5A). Module 3 is highly enriched in cell cycle related proteins including: E2F1, TGFB1, CDC25C/A, CHEK1, TP53, CCNA2 and CCNB1. Module 4 contains proteins related to both apoptosis/cell death (FAS, TNFRSF1A, LCK, BRCA1, E2F1 and CHUK) and cell cycle (CCNA2,CDC25C,E2F1), suggesting potential coupling of the two biological processes [146].

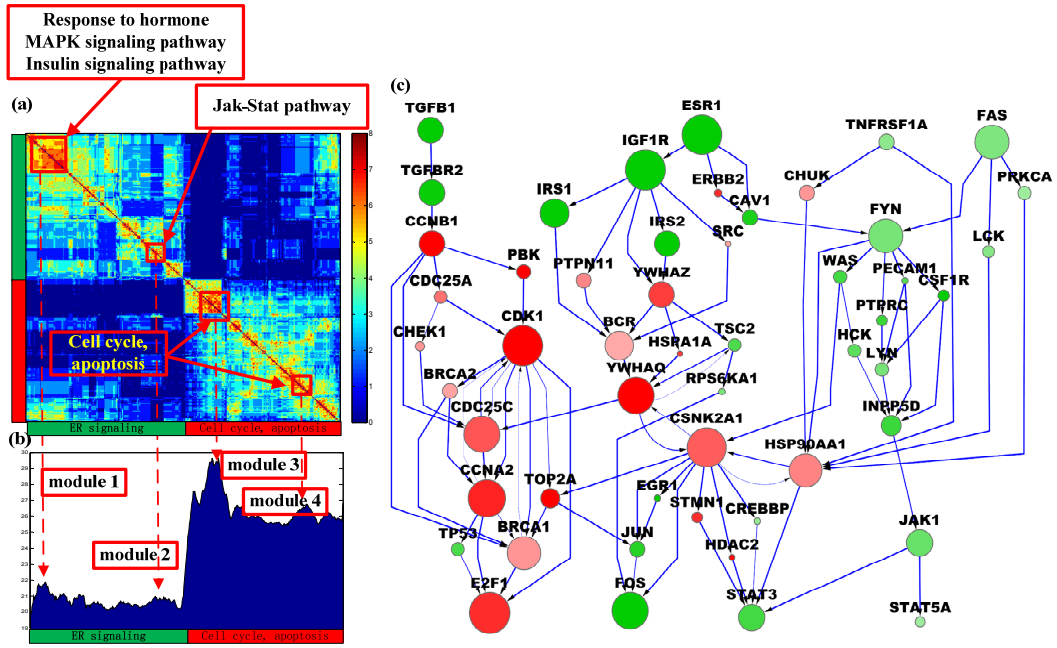


Figure 3.16. Explore pathway landscape to investigate pathway modularization and crosstalk. (a) Structural heatmap of sampled pathways from GIST algorithm. Structural similarity analysis reveals functional diversity in sampled pathways. (b) Potential distribution of pathway samples. Pathway clusters that correspond to locally enriched potentials are identified as pathway modules. Four pathway modules have been identified. (c) Layout of pathway network from four identified pathway module. The color scheme is the same as Figure 3.14.

Figure 3.16 (c) shows a merged pathway network from the four identified modules, which consists of several parallel pathways bridged through nodal proteins. On the left part of the network we have cell cycle pathways starting from TGFB pathway and converging to transcription factor E2F1. On the middle and right part of the network, signals have been initiated by various signaling membrane molecules, through several highlighted cytoplasm proteins such as YWHAQ, HSP90AA1 and CSNK2A1, and finally merge into transcription factors related to cell cycle (FOS and JUN) or JAK-STAT pathway. We hypothesize that proteins YWHAQ, CSNK2A1 and HSP90AA1 are potential nodal proteins that oversee several pathways of diverse functionalities. Their up-regulation in 'early-recurrent' group of patients (node color red) also suggests their potential role in pathway crosstalk associated with drug resistance in breast cancer.

Figure 3.16 (c) has also revealed several putative candidates and interactions that may relate to breast cancer progression and antiestrogen resistance. CDK1 (cyclin-dependent kinase 1; CDC2) is a serine/threonine kinase that is an essential modulator in the initiation and transition process through mitosis of the cell cycle primarily through its interaction with CCNB1 (cyclin B1). Deregulation or aberrant expression of CDK1 has been reported in many tumor types that makes it a suitable therapeutic target [147]. CDK1/CCNB1 helps to protect mitotic cells against extrinsic death stimuli [148]. Thus, increased expression of CDK1 in breast tumors that recur early may contribute to Tamoxifen resistance by shielding these breast tumor cells from antiestrogen-mediated cell death.

Pathway modularization and crosstalk identified from Symmans data

We also applied the proposed pathway identification protocol to a second Tamoxifen treated dataset from MD Anderson Cancer Center [65] (Appendix B.2). The dataset (termed Symmans data here) contains 47 'early recurrence' samples and 56 'late recurrence' samples using a 5-year survival cutoff. The dataset was profiled on Affymetrix U133A platform and we used PLIER algorithm for sample normalization (http://www.affymetrix.com). Similar as we did for the Loi dataset, we performed three case studies to investigate pathways related to ER signaling, cell cycle and apoptosis.

We compared the results of three case studies to evaluate the consistency between inferred pathways from Loi data and Symmans data using Venn diagram analysis (Figure 3.17). For ER signaling pathways, 29 out of 40 proteins identified from Symmans data were overlapped with the proteins from Loi data (p-value<1.48E-36). We merged cell cycle and apoptosis pathways together for a 4-way Venn diagram analysis because many proteins play dual roles in both cell proliferation and apoptosis. 26 out of totally 49 cell cycle proteins from Loi data were confirmed in the Symmans dataset (p-value<1.419E-29), while 28 out of 52 apoptosis proteins from Loi data overlapped with the proteins from Symmans data (p-value<3.3E-34).

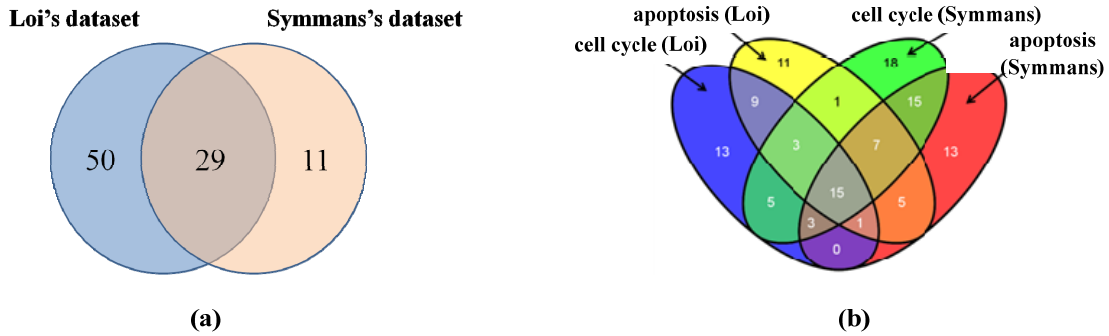


Figure 3.17. Venn diagram analysis for pathway results from Loi data and Symmans data: (a) ER signaling pathways; (b) Cell cycle and apoptosis pathways.

Pathway validation using breast cancer cell line data

We also used breast cancer cell culture to validate aberrant pathway modules identified from patient datasets. Four MCF7 derived cell models were included in the study: MCF7-STR, MCF7RR-STR, LCC1 and LCC2 cell lines [149-152]. MCF7RR-STR and LCC2 were Tamoxifen resistant cells while MCF7-STR and LCC1 were drug sensitive cells. The four MCF7-derived cell lines formed two pairs for Tamoxifen resistance study: MCF-STR v.s. MCFRR-STR (MCF7-STR for short) and LCC1 v.s. LCC2 (LCC for short), which were used as validation datasets to evaluate the identified aberrant pathways from patient datasets.

We first applied Gene Set Enrichment Analysis (GSEA) to investigate the concordance between patient datasets and cell line models [105], which shows that cell cycle and apoptosis pathways were significantly enriched in the resistant groups (i.e., MCFRR-STR and LCC2). We confirmed from Figure 3.14 (b) and (c) that most proteins

in cell cycle and apoptosis pathways had elevated expression in the 'early recurrence' group, which was in agreement with the cell line profiles. For ER signaling pathways, a subset of genes and their expression levels were also confirmed by both cell line studies (e.g., IRS1/2, FOS and JUN).

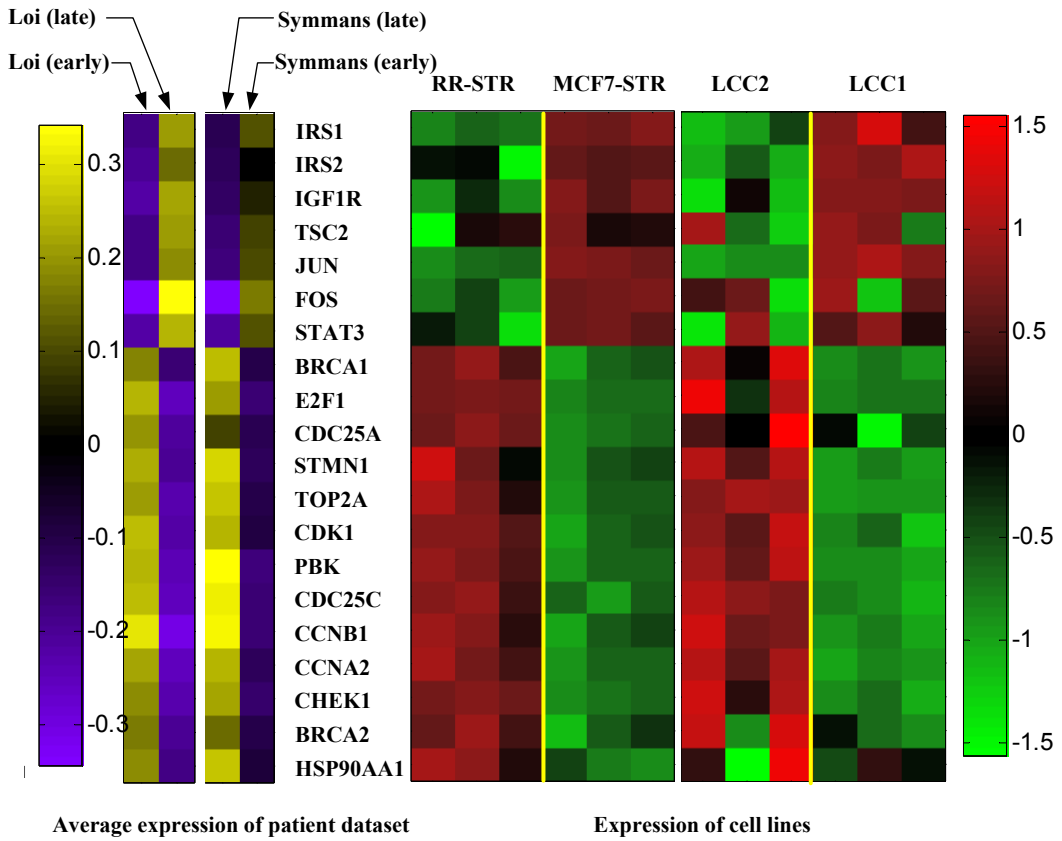


Figure 3.18. Cell line validation for identified pathway nodes from patient datasets. The left panel shows the average log2 expression of selected pathway proteins from Figure 3.16 (c). The right panel gives the log2 expression of two cell line studies: (MCF7-STRP v.s. MCF7RR-STRP and LCC1 v.s. LCC2). Seven proteins (IRS1, IRS2, IGF1R, TSC2, JUN, FOS and STAT3) are consistently over-expressed in non-resistant groups from both patient and cell line datasets. The rest of the proteins, which mainly relate to cell cycle and apoptosis, are over-expressed in resistant groups.

Figure 3.18 shows 20 out of totally 52 genes from the four pathway modules (Figure 3.16) with most consistent expression patterns between patient data and cell line data. The left panel shows the average log2 expression in each patient group ('early recurrence' vs. 'late recurrence'). The right panel shows the log2 expression of two cell line studies. Nucleic proteins from cell cycle/apoptosis modules (BRCA1, BRCA2, CCNA2, E2F1, CDC25A, CDC25C, TOP2A, CDC2, CHUK) show dominant up-regulation in both 'early recurrence' group of patient data and resistant cell line data, while transcription factors

JUN, FOS and STAT3 are down-regulated. In the upper cellular level we found that STMN1, PBK, CCNB1 and HSP90AA1 were over-expressed in early recurrence/resistant groups, while IRS1, IRS2 IGF1R and TSC2 were over-expressed in late recurrence/non-resistant groups. The concordance between patient and cell line data supports that the identified pathway modules may be actively involved in the recurrence of breast cancer and drug resistance of Tamoxifen treatment.

HSP90AA1 (HSP90) and endocrine resistance

Given its apparent topological importance in bridging pathways with diverse functions and consistent up-regulation in the resistant groups from multiple datasets, we hypothesize that HSP90AA1 could be a potential nodal protein that oversees several Tamoxifen resistance pathways. Heat shock proteins (HSPs) are chaperone proteins involved in stress response, which correctly fold proteins in the active form [153]. Our pathway network analysis of recurrent tumors shows an association of increased expression of HSP90 with TAM resistance. Previously, it has been demonstrated that HSP90 inhibitor 17-(dimethylaminoethylamino)-17-demethoxygeldanamycin (17-DMAG) can inhibit the growth of aromatase inhibitor-resistant breast tumors by inducing apoptosis and G(2) cell cycle arrest, and by degradation of HSP90 client proteins AKT and HER2 [154]. Our antiestrogen resistant LCC2 cells (resistant to Tamoxifen) showed significantly increased response (decrease in relative cell proliferation) to HSP90 inhibitor 17-N-Allylamino-17-Demethoxygeldanamycin (17-AAG) compared to antiestrogen at 0.1 μM compared to sensitive LCC1 cells (Figure 3.19 (a)). Notably, ER+ antiestrogen breast cancer cell line, LCC9 (cross-resistant to both Tamoxifen and Faslodex) showed even stronger inhibition of cell proliferation at different dosage of 17-AAG (Figure 3.19 (b)). Furthermore, knockdown of HSP90AA1 with siRNA showed significant inhibition of cell proliferation in LCC2 cells under control conditions (vehicle alone) and sensitization of LCC9 cells to Faslodex but not Tamoxifen (Figure 3.19 (c)). LCC9 cells also showed significant inhibition in cell growth and re-sensitization to both Faslodex and Tamoxifen with HSP90AA1 knockdown (Figure 3.19 (d)). Figure 3.19 (e-f) shows knockdown of HSP90AA1 protein under the different treatment conditions. These

data suggests that activation of HSP90AA1-regulated pathways partly contribute to antiestrogen resistance in breast cancer.

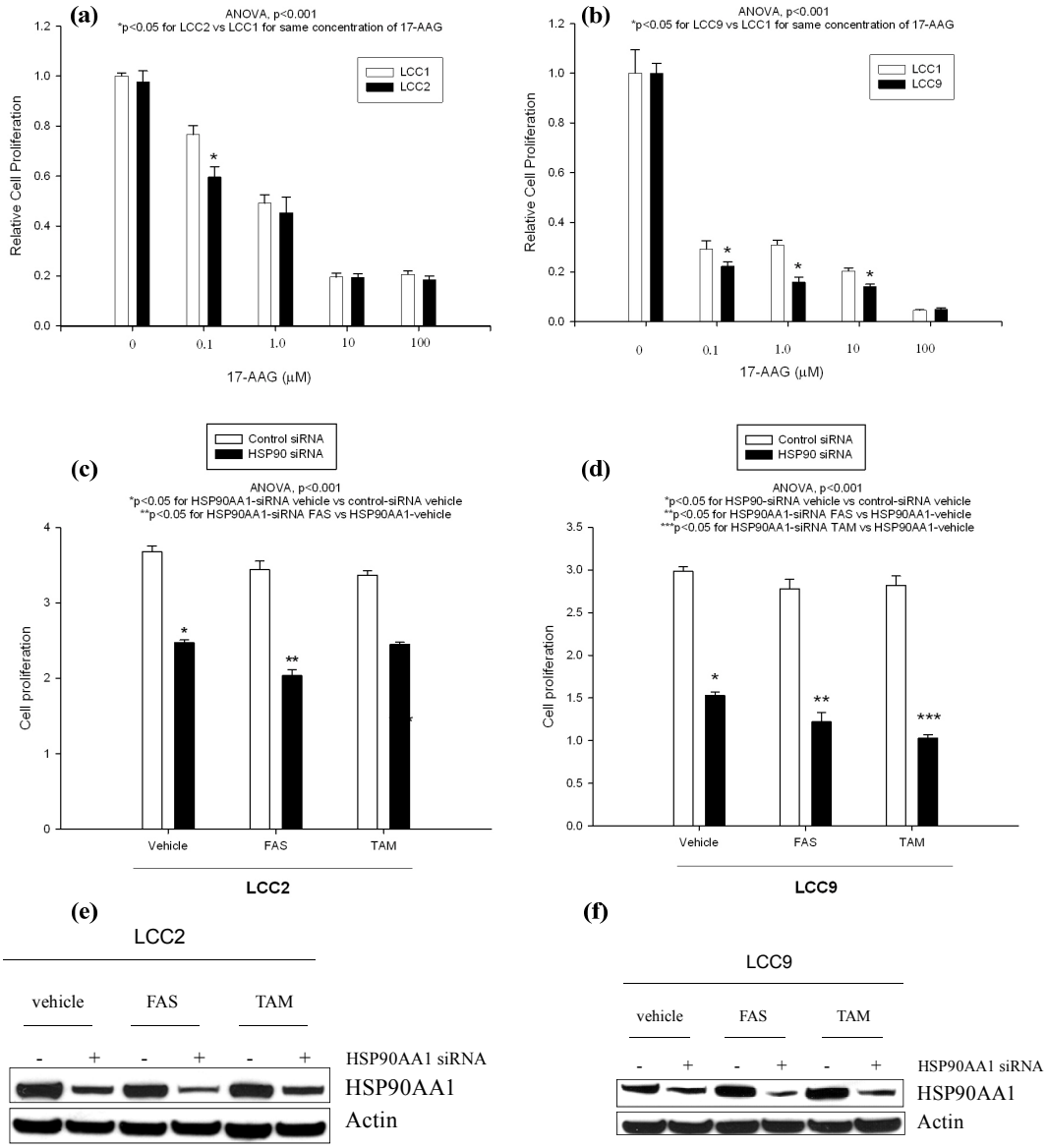


Figure 3.19. HSP90AA1 is a potential drug target for antiestrogen resistance. (a) Cell proliferation is significantly reduced in LCC2 cells compared to LCC1 cells in treatment of 17-AAG at 0.1 μM. (b) Significant inhibition of cell proliferation of resistant LCC9 cells compared to LCC1 cells in treatment of 17-AAG at 0.1, 1.0, and 10μM. (c) Knockdown of HSP90AA1 with siRNA showed significant inhibition of cell proliferation in Faslodex sensitized LCC2 cells compared to control (vehicle) but not Tamoxifen. (d) Knockdown of HSP90AA1 with siRNA showed significant inhibition of cell proliferation in both Tamoxifen and Faslodex sensitized LCC9 cells compared to control (vehicle). (e-f) Western blot showing protein concentration after knockdown of HSP90AA1 under different conditions.

3.7. Discussion

One fundamental limitation of applying existing pathway identification methods to cancer study is the lack of effort to dissect complicated signaling rewiring in a systematic manner: most methods will return a large prioritized pathway list without further stratification of the pathway interactions according to their functions or structures. Increasing evidences have been revealed by recent researches that cancer system is driven by complicated interactions of multiple signaling pathways, which promotes the idea to target the crossroads of multiple pathways for therapeutic study. Therefore, we propose to investigate aberrant signal transduction in functional modules constrained by protein-protein interactions, where proteins that inter-connect multiple modules may work as 'supervisor' for pathway crosstalk.

The proposed methodology has several notable advantages over existing methods, which can be summarized as follows. Firstly, GIST can easily incorporate biological knowledge such as cellular location information, which makes the identified pathways more biologically interpretable. It can also allow users to incorporate domain knowledge such as subcellular compartment by emphasizing signal transduction in either nucleus or its up-stream (cytoplasm or membrane). Without utilizing prior knowledge, computational methods purely based on data tend to over-fit each individual dataset whereas the results are often hard for interpretation or generalization across multiple datasets. Secondly, most existing methods either do not address the identification of edge direction or infer edge direction in a deterministic manner (e.g., the edge orientation algorithm). In the context of recovering directed pathways from complex human PPI, it is of unique significance to evaluate the confidence level of inference before forming biological hypotheses from computational results. We hereby use GIST to assign each edge with a posterior probability of its direction, showing the degree of confidence of inferred direction by integrating data and knowledge. Thirdly, we present the final pathways in forms of structural related modules using SOUL instead of ranking thousands of pathways. It has been shown that pathways of different functionalities have different local potential profiles (e.g., cell cycle related proteins have significantly higher potential than cytoplasmic signaling proteins, Figure 3.16 (b)). Ranking pathways

without considering the relationship among individual pathways only highlights the most differentially expressed pathways yet downplays important signaling components. By exploring the pathway landscape and clustering pathways according to their structural similarities, we can further categorize complicated pathway interactions into within-module interactions and between-module interactions. Finally, we propose for the first time, a computational interpretation of nodal proteins as hub proteins that link multiple pathway modules to oversee signaling crosstalk. This echoes the increasing acknowledgement of hypotheses that human cancer is such complicated disease driven by multiple pathways rather than one single pathway.

3.8. Conclusion

We have developed a computational method, i.e., IMPACT, to explore intracellular signal transduction and pathway crosstalk by integrating multi-platform data. We have proposed a unique perspective to investigate structure linkage of pathways from distribution point of view, where locally coherent pathway modules are identified instead of ranking individual pathways. Moreover, we have fully addressed the problem and complexity in identifying aberrant pathways in human cancer studies by developing a complete protocol that integrates gene expression data, protein-DNA interaction data, protein-protein interaction data and knowledge from literature. IMPACT implements a two-step pathway identification workflow, which first estimates aberrant linear signal transduction between given source and target proteins followed by a structural re-organization of sampled pathway states. Based on identified pathway modules, topologically important hub proteins that inter-connect multiple pathway modules are identified as potential nodal proteins to oversee pathway crosstalk.

We used synthetic datasets and datasets acquired from yeast to validate the performance of the proposed pathway identification method. We compared the ability of the GIST algorithm to identify pathway nodes and directed edges using simulated data in presence of additive noise in gene expression and false positive connections in protein-protein interaction data. Numerical results showed significant advantage of GIST to identify directed edges and alternative pathways compared to existing methods.

We also applied GIST on yeast protein-protein interaction data to identify yeast MAPK signaling pathways. Canonical pathways from Kyoto Encyclopedia of Genes and Genomes (KEGG) are used as golden standard to evaluate identified proteins from multiple pathway identification methods. We validated that GIST not only achieved comparable or better performance in identifying proteins that are involved in canonical MAPK pathways, but was also capable of revealing novel alternative pathways through data integration.

Finally, we applied the complete pathway identification protocol to breast cancer patient datasets to identify nodal proteins that supervise multiple aberrant signaling pathways. Breast cancer cell line models were used to validate the identified signaling pathway networks from the patient datasets. The proposed nodal protein, HSP90AA1, was further validated by wet-lab experiments to be associated with endocrine resistance in breast cancer cells.

4. Contribution, Future Work and Conclusion

4.1. Summary of original contribution

In this proposal, we aim to develop computational methods to integrate gene expression data, protein-protein interaction data and protein-DNA interaction data for the identification of signal transduction pathways in cancer. The original contributions of this dissertation research can be summarized as follows:

4.1.1. Robust identification of transcriptional regulatory networks using Gibbs sampler on outlier sum statistic

A novel method, namely Gibbs sampler on outlier sum statistic (GibbsOS), is developed to identify regulatory networks from the combinatory use of microarray data and protein-DNA interaction data. Based on the assumption that the target genes of the same transcription factor should be correlated with the hidden transcription factor activity, we propose a novel test statistic called outlier sum of regression t-statistic to evaluate the consistency of candidate target genes. The proposed method is capable of investigating complicated cases that one target gene is being regulated by multiple transcription factors. It is designed to identify true target genes whose expression changes are more likely to be correlated with its regulator's activity.

We have tested our GibbsOS method on two simulation datasets according to different noise levels. The result shows that our method consistently outperforms the competing method under different signal-to-noise ratios that demonstrates its robustness against additive noise. On the other hand, we have also investigated the impact of false-positive connections under log-linear simulation data. By controlling the structure error (also known as 'false positive connections' or 'background genes') contained in the initial binding matrix, we have evaluated the robustness of GibbsOS against competing methods with regard to structural perturbations in the binding data. A validation study on yeast data also shows the efficacy of GibbsOS for target gene identification, which is supported by functional annotation analysis that target genes identified by GibbsOS is more

significantly enriched in cell cycle functions than those identified by the competing method.

Finally, we have applied GibbsOS to two breast cancer cell line data sets to identify regulatory modules that are associated with two estrogen-related conditions: estrogen-induced [86] and long-term estrogen-deprived (LTED) [87]. To understand the role that estrogen signaling plays in transcriptional regulation, estrogen independent growth, and ultimately breast cancer progression and resistance, we select 26 estrogen receptor (ER or ESR1) related transcription factors to study their actions in breast cancer cells under different environmental conditions (E2 status). Computational results have revealed both different and common transcriptional modules between estrogen-induced and estrogen-deprived breast cancer cell lines, suggesting potential complications of regulatory rewiring in a condition-specific manner.

4.1.2 Infer pathway modularization and crosstalk using IMPACT protocol

Detecting aberrant signal transduction pathways from high-throughput data using GIST algorithm

We propose a novel computational model to identify aberrant pathways from gene expression data and PPI data. We bring forward the concept of 'node potential', 'edge potential' and 'flow potential' to evaluate the quality of the a pathway configuration and its association with disease phenotype. We use the Gibbs distribution formulation to convert the potential functions into a probability distribution, which is interpreted as being equivalent to the posterior distribution of a pathway configuration given the observed gene expression. To reduce the computational cost of enumerating pathway samples from large sample space, we use a Gibbs sampling technique to estimate the joint distribution from conditional distributions. Pseudo edges are introduced to connect proteins that do not have an edge in original PPI network to ensure some important Markov properties of the Gibbs sampler (i.e., irreducibility and reversibility).

We have compared the GIST algorithm with existing computational methods in terms of identifying pathway nodes and edges under different noise settings. We have observed superior and robust performance of GIST algorithm to recover both pathway

nodes and edges in the presence of experimental noise, which may be attributed to incorporation of prior knowledge such as sub cellular locations. We also add different levels of false positive connections into the original protein-protein interaction network to investigate the robustness of methods against structural perturbations. Simulation results have demonstrated that despite a slightly better performance in retrieving pathway nodes, GIST achieved significantly higher capability to estimate edge directions when false positive connections are introduced to the PPI.

We have applied the proposed GIST algorithm on yeast protein-protein interaction data to identify MAPK signaling pathways. Our proposed method shows comparable or slightly better performance than competing method, such as random color coding and integer linear programming, in yeast studies while it is more successful in retrieving sub-optimum paths that has competing scores as the global optimum.

Untangle signaling pathway interactions to identify pathway crosstalk

The GIST algorithm generates a huge list of pathway samples with largest potentials. To further investigate the interaction and structural linkage between sampled pathways, we developed a novel computational method called Structural Organization Uncovers pathway Landscape (SOUL). By exploring the 'pathway landscape', which is reconstructed through structural clustering of pathway states, SOUL re-organizes pathways states into local modules with different functionalities. SOUL prioritizes proteins that are topologically important to bridge signal transduction among multiple pathway modules as nodal proteins, which may be potential drug targets in cancer therapeutic research.

We have applied IMPACT to two breast cancer patient datasets to identify pathway modules that are related to endocrine resistance [64, 65]. Computational results have revealed aberrant signal transduction linking ER signaling pathways with downstream machinery through nodal proteins such as HSP90AA1. Validation of identified proteins from patient datasets on breast cancer cell line models shows that proteins with elevated expression in early-recurrent group of patients have significant up-regulation in resistant

cell lines. Cell viability analysis and siRNA experiments have confirmed potential importance of HSP90AA1 in the development of drug resistance of breast cancer cells.

4.2. Future work

Methods and studies discussed in this dissertation work can be extended and improved in several aspects to better tackle realistic problems associated with cancer research. We have summarized the main points of future work as follows:

4.2.1. Transcriptional regulatory network identification

Comprehensive modeling of structural error in regulatory network identification

The proposed GibbsOS method models false positive connections in initial binding matrix (from ChIP-on-chip data, motif sequence data, etc.) in forms of 'background' genes. However, we consider the concept of 'background genes' only describing part of the problem; a more comprehensive summarization about how false connections complicate the reconstruction of regulatory networks is listed as follows: 1) genes that are irrelevant to regulation may be introduced into the initial topology ('background genes'). 2) Genes that are true regulatory targets may be excluded from analysis due to a stringent threshold. 3) Identification of co-regulatory modules is further compromised in the presence of both false positive and false negative connections among genes and multiple transcription factors in binding data. As an alternative, computational methods were proposed to convert information from binding data into prior probabilities without using a 'hard' threshold [78, 155]. However, how to map p-value of original ChIP-on-chip data into prior probability in real biological studies usually lacks convincing justification, while the performance of these algorithms is sensitive to the mapping function. An *ad hoc* strategy that minimizes false negatives is to loosen the threshold of p-value, at the cost of introducing more false positives. Therefore, it is of significant importance to develop effective computational methods that are not only robust to noise in gene expression data, but can also counteract 'structural error' in binding data, especially in the context of identifying condition-specific regulatory networks for cancer research.

One potential solution is to propose a novel approach built upon the NCA model, which estimates the posterior distribution of connectivity matrix using a two-stage Gibbs sampler. Instead of defining the elements of connectivity matrix as fixed binaries (0 for

'no binding' or 1 for 'binding'), we model them as Bernoulli random variables with certain prior probabilities. The algorithm iterates between the estimation of hidden transcription factor activities (TFAs) and that of posterior distribution of connectivity matrix. By probabilistically perturbing the connectivity matrix, we evaluate the confidence of each protein-DNA interaction from raw binding data as true regulatory interaction in a condition-specific manner.

Improve the efficiency of the GibbsOS method on large scale applications

In the current design of GibbsOS, we calculate the outlier sum of all the candidate genes for one transcription factor (TF) at a time, which requires pairwise regression analysis. Therefore the general computational cost for one iteration is $N^2 \times L$ where N is the average number of genes in the candidate pool of one transcription factor and L is the total number of TFs. One major concern is that when applied to human cancer studies, where the initial topology usually contains thousands of genes and up to hundreds of TFs, the proposed framework may be computationally inefficient. To improve the applicability of the GibbsOS method on large scale datasets, we have adopted the method proposed by Chang *et al*. to use subspace decomposition instead of pairwise regression analysis. The Gibbs sampler now iterates between the estimation of transcription factor activities (TFAs) using eigenvalue decomposition and resampling of binding matrix. The new method, which we call probabilistic network analysis (pNCA), reduces the computational cost to $N \times L$.

Clustering analysis of whole transcriptome binding data to uncover co-operative modules

One critical challenge to apply regulatory network identification methods to human cancer research is the huge size of binding motifs (>500) and un-specificity of the available binding data. Human regulatory network is complicated in that: 1) multiple transcription factors may work co-operatively (or competitively) in transcriptional regulation. 2) Transcription regulation is dynamic and condition-specific, where cells adapt to external stimuli or environmental change through regulatory rewiring. In our

previous application of GibbsOS to breast cancer cell lines, we purposely selected 26 breast cancer related motifs as a pilot study. When scaling up the study to whole human transcriptome, we are facing practical challenges to directly apply GibbsOS to high throughput datasets; for example, the maximum number of TF-gene interactions that can be evaluated with confidence is limited by the total number of samples. In order to precisely uncover the co-operative regulations in cancer transcriptome, we need computational methods to pre-define regulatory modules where multiple binding motifs in the same module show the potential of co-regulation. Gong *et al*. [156] has proposed a computational method to cluster genes according to gene expression and binding motif, which can be used to dissect the hidden hierarchy embedded in the initial binding matrix. By using such a clustering method as a preprocessing tool to hypothesize potential co-operative regulatory modules, GibbsOS can better evaluate the regulatory interactions among transcription factors and their target genes through a "divide and conquer" strategy.

4.2.2. Identify aberrant signal transduction and pathway crosstalk through data integration

Two-stage framework to decipher signal transduction from global to local

We have developed a computational method (i.e., GIST) to reconstruct signal transduction pathways between given source proteins and target proteins. We are also able to infer signal transduction directions by pooling sampled pathways from GIST. Both the estimated node scores and edge probabilities are conditioned on sampled paths that connect pre-defined source and targets of certain length. This raises some new challenges when applying GIST to real biological applications: in most cases we have to rely on our domain knowledge to propose source and target proteins. Prior knowledge from literature or database is helpful as preliminary exploration towards understanding data, but may not be effective to identify novel signaling interactions. To resolve the abovementioned technical challenges and show more extensive and representative aberrant signal transduction in cancer study, we propose to use a two-stage pathway identification framework: 1) initial exploration of signal transduction is conducted among given proteins with a wide range of pathway length. The sampled pathways of different

lengths are aggregated to estimate pathway node distribution, which we refer to as the 'global' distribution of pathway nodes. Prioritize significant proteins in each cellular compartment (e.g., membrane, cytoplasm or nucleus). 2) Build 'local' signaling pathways between different cellular compartments and estimate the edge directions [45, 157-159]. This two-stage framework provides multiple alternatives of signal transduction by opening several 'windows' in different compartments of the cell, which offers a more complete picture of possible aberrant signal transduction within the cancer system.

Investigate pathway crosstalk among multiple functional modules

We have developed a computational method (i.e., SOUL) to identify pathway modules and nodal proteins by exploring pathway landscape. To better study pathway crosstalk and scale to pathways with variant lengths, we can improve the proposed method in the following aspects: 1) the current version of SOUL uses number of overlapped genes in the pathway to measure the similarity between two individual pathways. To measure the similarity between two pathways of different length, new statistics or measures need to be devised. For example, Li *et al.* have used shortest path to measure the distance between two individual pathways from knowledge database [129]. 2) SOUL uses betweenness centrality to measure the topological importance of the nodes in the network, which does not consider either the node score or edge probability estimated from the GIST algorithm [160]. Neither does SOUL associate statistical significance with the identified nodal proteins or pathway modules [161]. 3) Currently the selection of pathway modules (clusters) is mainly based on users' preference. It is important to incorporate functional enrichment analysis into the method so that pathway modules of different functionalities can be highlighted. This is especially important to understand the role of nodal protein in supervising multiple functional pathway modules [162].

Develop software with GUI for IMPACT

We are planning to implement a Cytoscape plug-in for IMPACT method using java [163]. The software will offer user-friendly interface for data uploading, parameter

setting and result visualization. The software will provide two key graphical features: a hierarchical display of sampled signaling pathways and visualization of the pathway landscape (including structural heatmap and potential distribution). For the sampling algorithm, two computational options will be provided: exhaustive search and Gibbs sampling. The exhaustive search is useful when working on relatively small network size to identify pathways of small or medium length; Gibbs sampling will be effective estimate of the pathway distribution when working on large PPI networks for long pathways.

From distribution learning to structural learning

We have developed the IMPACT method to learn pathway distribution by integrating gene expression data, protein-protein interaction data and other biological knowledge. Initial applications to yeast datasets and human cancer datasets demonstrated the advantage of the proposed distribution learning method to reconstruct alternative pathways and identify crosstalk among multiple pathways. Preliminary exploration of aberrant signal transductions in human cancer study also revealed new insights into the problem: protein-protein interaction data, which serve as the base for pathway sampling, is highly redundant and non-specific. Here the redundancy is mainly about edge connections (protein-protein interactions), where some of the proteins may have up to 200 interactions in raw PPI data. Therefore the result identified by GIST always contains densely intertwined linear paths that are hard for biological interpretation. The current version of the GIST algorithm selects edges based on overall association strength of gene expressions using correlation test, which does not consider higher order dependency among multiple proteins. To model more complicated interactions in PPI network in order to screen non-specific edges, we propose to incorporate structural learning into pathway identification. Wang *et al*. have proposed a Gibbs sampling framework with embedded Metropolis-Hastings sampling on PPI network to select edges for subnetwork identification [164]. Zhang *et al*., have developed a method called DDN to investigate dependency among edges in differential networks[165]. Structural learning will tremendously increase solution space where the problem itself is usually non-convex and

does not have analytical solution. In this case, we may again resort to Monte Carlo methods to sample network or pathway structures according to the proposed model.

4.3. Conclusion

In this dissertation work, we have developed computational methods based on Markov Chain Monte Carlo techniques to identify aberrant signal transduction in human cancer research. A complete pathway identification protocol has been proposed to integrate regulatory network identification with subnetwork identification to infer intracellular signal transduction. Targeted at resolving real biological problems in human cancer study, e.g., to propose potential therapeutic targets for combating breast cancer endocrine resistance, we tailor our computational methods to deliver results that echo the latest advance in systems biology. We have demonstrated the effectiveness and robustness of the proposed methods over existing methods on synthetic datasets. We have also applied the proposed pathway identification protocol to breast cancer patient datasets to identify potential drug targets, which have been validated by cell line data and wet-lab experiments.

Appendix A. Addendum of chapter 2 on target gene distribution and convergence

A1. Distribution of Target Genes

For any function defined over a set of candidate genes, e.g., the conditional outlier sum statistic [73] of $\theta_i \in \Theta_i = \{\theta_{i1}, \theta_{i2} ... \theta_{iK}\}$ given $\theta_1 ..., \theta_{i-1}, \theta_{i+1}, ... \theta_L$, has the following form:

$$OS|_{\theta_i} = f_{os}\left(\theta_i \middle| \theta_1 ..., \theta_{i-1}, \theta_{i+1}, ... \theta_L\right). \tag{A.1}$$

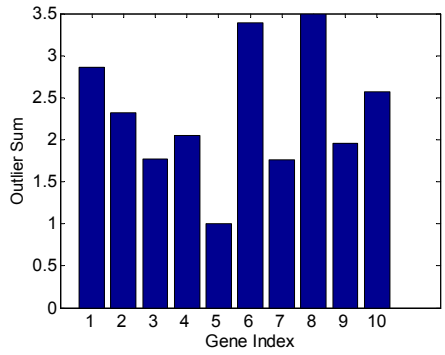
The domain Θ_i is the set of all possible candidate genes for TF *i*, which are discrete values (gene indices) rather than real valued numbers (Figure A.1 (a)). By normalizing the outlier sum using constant $K_0 = \sum_{k=1}^{K} OS|_{\theta_i = \theta_{ik}}$, we interpret a general function over Θ_i into a probabilistic function *p* as follows (Figure A.1 (b)):

$$p\left(\theta_i \middle| \theta_1 ..., \theta_{i-1}, \theta_{i+1}, ... \theta_L\right) = \frac{1}{K_0} \cdot f_{os}\left(\theta_i \middle| \theta_1 ..., \theta_{i-1}, \theta_{i+1}, ... \theta_L\right). \tag{A.2}$$

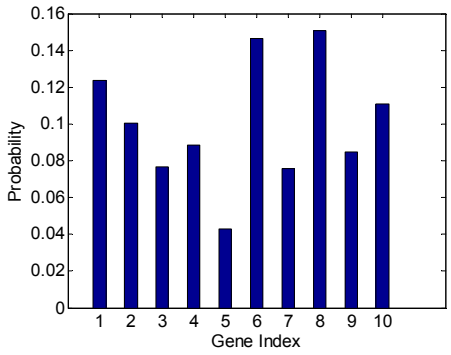
As a property of the probability distribution, we have:

$$\sum_{k=1}^{K} p\left(\theta_i = \theta_{ik} \middle| \theta_1 ..., \theta_{i-1}, \theta_{i+1}, ... \theta_L\right) = 1. \tag{A.3}$$

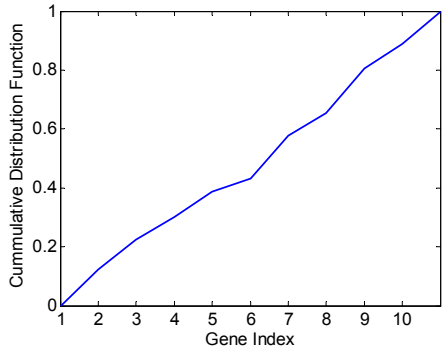
We can plot the Cumulative Distribution Function (CDF) of probability function *p* according to original gene alignment, which is shown in Figure A.1 (c). For a better visualization of the CDF plot, we can first sort the genes according to their probability and then aggregate them to obtain a smooth version of CDF (Figure A.1 (d)).



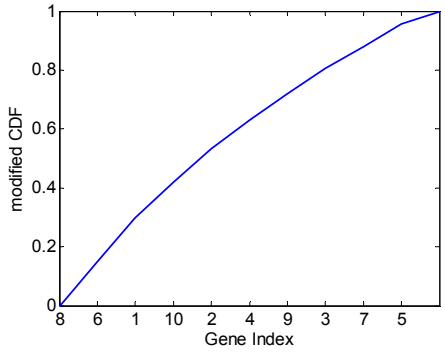
a. outlier sum function



b. probability distribution of target genes



c. original CDF of target genes



d. modified CDF

Figure A.1. Illustration of Gibbs distribution on target genes

A2. Convergence of Gibbs Sampler

Suppose the current state of Gibbs sampler is $\bar{\Theta} = (\theta_1, ..., \theta_i, ..., \theta_L)$ (with probability $\pi(\bar{\Theta})$). The transition probability from the current state $\bar{\Theta}$ to the next state $\bar{\Theta}' = (\theta_1, ..., \theta_i', ..., \theta_L)$ (with probability $\pi(\bar{\Theta}')$) is denoted as $q(\bar{\Theta} \to \bar{\Theta}')$. In order to assure that our Gibbs algorithm shall reach a unique stationary distribution, the following detailed balance condition must be satisfied:

$$\pi(\bar{\Theta}) \cdot q(\bar{\Theta} \to \bar{\Theta}') = \pi(\bar{\Theta}') \cdot q(\bar{\Theta}' \to \bar{\Theta}) . \tag{A.4}$$

If we define the transition probability as

$$q(\bar{\Theta} \to \bar{\Theta}') = \frac{p(\theta_1, ..., \theta_i', ..., \theta_L)}{\sum_{\theta'} p(\theta_1, ..., \theta_i', ..., \theta_L)}, \tag{A.5}$$

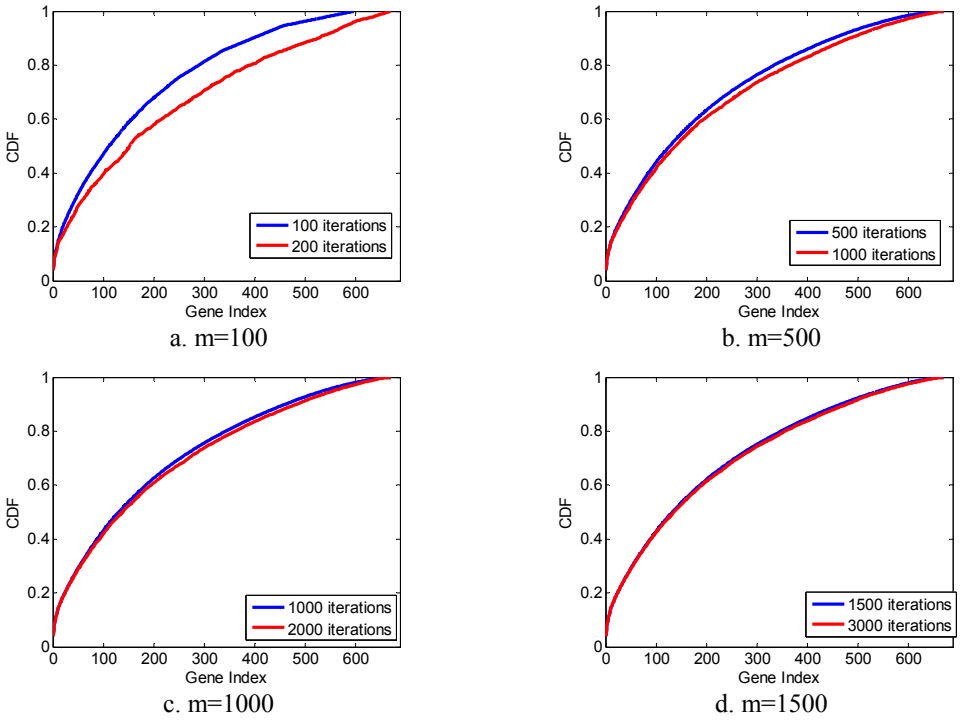
we have

$$\begin{aligned}\pi(\bar{\Theta})\cdot q(\bar{\Theta}\to\bar{\Theta}') &= p(\theta_1,...,\theta_i,...,\theta_L)\cdot\frac{p(\theta_1,...,\theta_i',...,\theta_L)}{\sum_{\theta_i}p(\theta_1,...,\theta_i,..,\theta_L)}\\ &= p(\theta_1,...,\theta_i',...,\theta_L)\cdot\frac{p(\theta_1,...,\theta_i,..,\theta_L)}{\sum_{\theta_i}p(\theta_1,...,\theta_i,..,\theta_L)} \qquad \text{(A.6)}\\ &= \pi(\bar{\Theta}')\cdot q(\bar{\Theta}'\to\bar{\Theta})\end{aligned}$$

Thus, the detailed balance is satisfied. As long as the sampling chain is long enough, the final distribution will converge to a stationary distribution.

To test whether our Gibbs sampler is converged in real biological studies we use the Cauchy's convergence criterion defined as follows. The sequence of random variable $X_n(\varsigma)$ is converged in the mean square sense if and only if: $E\left[\left(X_n(\varsigma)-X_m(\varsigma)\right)^2\right]\to 0$ when $n,m\to\infty$, where in our case the random variable $X_n(\varsigma)$ is the CDF of the target gene distribution after *n* iterations. For the breast cancer cell line study in estrogen induced condition (early up-regulated case), we ran the Gibbs sampler for thousands of iterations and compared the CDF of target genes after *m* iterations and that after *2m* iterations. We show the cases where *m*=100, 500, 1000, and *m*=1500, respectively.

From Figure. A.2 we can see as the number of iterations *m* goes larger the estimated distribution gradually converges to a stationary distribution. After more than one thousand iterations, we have sufficient samples to estimate the target distribution and the correlation between the distributions estimated from 1,000 iterations and that of 2,000 iterations is larger than 0.98.



a. m=100 b. m=500

c. m=1000 d. m=1500

Figure A.2. Convergence check using CDF comparison

Appendix B. Addendum of chapter 3 on biological experiments

B.1. Pathway modules identified from Loi data

Table B.1. Proteins in four identified pathway modules from Loi data

Module 1	Module 2	Module 3	Module 4
BCR	CAV1	BRCA1	BRCA1
CREBBP	CSF1R	BRCA2	CCNA2
CSNK2A1	ERBB2	CCNA2	CDC2
EGR1	ESR1	CCNB1	CDC25C
ESR1	FYN	CDC2	CHUK
FOS	HCK	CDC25A	CSNK2A1
HDAC2	INPP5D	CDC25C	E2F1
HSP90AA1	JAK1	CHEK1	FAS
HSPA1A	LYN	E2F1	FYN
IGF1R	PECAM1	PBK	HSP90AA1
IRS1	PTPRC	TGFB1	LCK
IRS2	STAT3	TGFBR2	PRKCA
JUN	STAT5A	TOP2A	TNFRSF1A
PTPN11	WAS	TP53	WAS
RPS6KA1			YWHAQ
SRC			
STAT3			
STMN1			
TOP2A			
TSC2			
YWHAQ			
YWHAZ			

For Loi data, we identified four pathway modules using IMPACT (Figure 3.16). Table B.1 lists the proteins in each pathway modules. Among all 52 proteins in Table B.1, the top three proteins with highest betweenness centrality (bc) [131] are: CSNK2A1(bc=1,277), YWHAQ (bc=823) and HSP90AA1 (bc=644), implying a

significant role of these cytoplasm signaling proteins to bridge membrane receptors with nucleus activities.

B.2. IMPACT case study on Symmans dataset

We applied the IMPACT protocol to Symmans data to identify aberrant pathways associated with drug resistance in breast cancer. Three case studies were performed to investigate different functional pathways related ER signaling, cell cycle and apoptosis. For ER signaling pathways, we put more weights on pathways in the membrane and cytoplasm by setting $\psi_k = 4$ for $k < 4$ and $\psi_k = 0$ for $k = 4$, while for cell cycle and apoptosis pathways, we set $\psi_k = 0$ for $k < 4$ to explore more on nucleus signaling. Figure B.1 is the assembled ER signaling pathway network using top 200 pathway samples obtained from the GIST algorithm. Functional annotation analysis reveled that the assembled pathway network was highly enriched in functional terms including: ErbB signaling (Benjamin p-value=2.8E-10), MAPK signaling (Benjamin p-value=1.6E-8), Jak-STAT signaling (Benjamin p-value=3.4E-5) and insulin signaling (Benjamin p-value=1.2E-4), etc. The result is very consistent with the functional/pathway enrichment for the Loi dataset. Notably, two nodal proteins HSP90AA1 and MAPK1 have elevated gene expression in the 'early recurrence' group of patients, which is also in agreement with the expression of the two proteins in Loi data (log2 fold-change>0, node color red). This suggests that these two proteins potentially be important signaling hubs overseeing one or more pathways associated with early recurrence of breast cancer.

Figure B.2 (a) and (b) shows the assembled network from cell cycle and apoptosis pathways identified from Symmans data. Genes in Figure B.2 have significant enrichment in functions such as positive regulation of biosynthetic process (Benjamin p-value=8.4E-21), regulation of apoptosis/programmed cell death (Benjamin p-value<2.1E-11), regulation of mitotic cell cycle (Benjamin p-value<1.2E-5), etc. By comparing three pathway networks we hypothesize that cytoplasmic nodal proteins HSP90AA1, CSNK2A1 and MAPK1 may have unique topological importance to intracellular signal transduction by bridging membrane proteins (such as estrogen receptor (ESR1), EGFR, or canonical death receptors), with cell cycle and apoptosis machinery in the nucleus.

We further applied SOUL to pooled samples from three case studies to identify pathway modules that corresponded to local peak potentials. Figure B.3 (a) shows the structural heatmap of the pathway samples. ER signaling pathways form one big cluster with apparently higher within-cluster similarities compared to cell cycle and apoptosis. This is concordant with Figure B.1 that ER signaling pathways are highly tangled through some major hub proteins (e.g., HSP90AA1, CSNK2A1, MAPK1, LCK, LYN and STAT3, etc), implying a potential complexity to determine signaling directions. We also notice that one cell cycle pathway cluster on the upper-left corner of heatmap showed quite unique structural pattern compared to the rest of cell cycle pathways, suggesting the possibility of multiple distinct pathway modules. By re-shuffling the pathway potential according to structural heatmap, we obtained the ordered multimodal potential distribution in Figure B.3 (b). Four local clusters with peak potentials were identified as four pathway modules. Figure B.3 (c) shows the final pathway network assembled from four pathway modules. We observe that signal transduction in the membrane has to pass through cytoplasmic proteins such as MAPK1, HSP90AA1 and CSNK2A1 in order to reach the downstream. A Venn diagram analysis (Figure B.3 (d), Table B.2) has confirmed that pathway module 1 has quite deterministic signal transduction (with 12 unique proteins), which goes almost surly from IGFR1 or INSR through cytoplasm signaling hubs (e.g., SRC, CHUK and HSP90AA1, etc.) and finally converges to nucleus. In contrast, modules 2 and 4 have highlighted potential diversities in signal transduction between membrane protein and Jak-STAT pathways. Signaling can be initiated by either estrogen receptor (ESR1) via canonical Jak-STAT pathway members (PIK3R1, SOS1 and PTPN6), or by various membrane receptors (INSR, EGFR), death receptors (FAS, TNFRSF1A) through PTPN6, SHC1 or LYN. For module 3, though its nodes are mostly included in modules 2 and 4, has offered an alternative path where signal transduction is finally bridged to cell cycle progression (CDC2, E2F1).

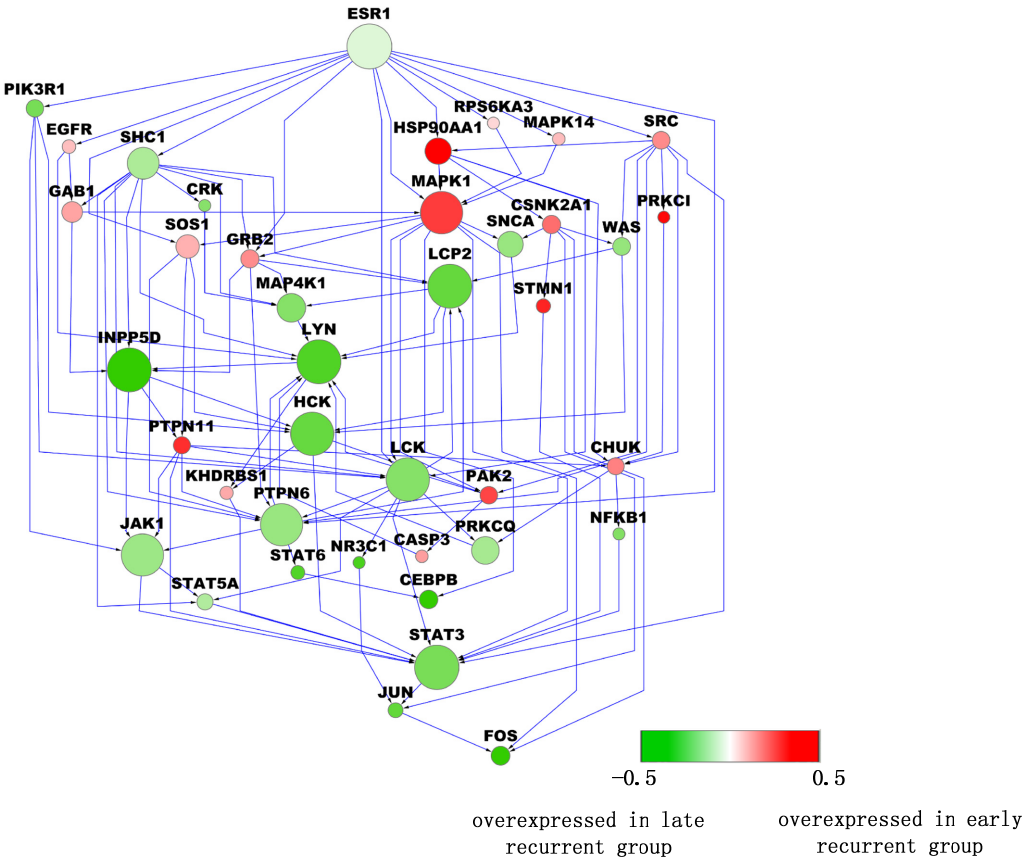
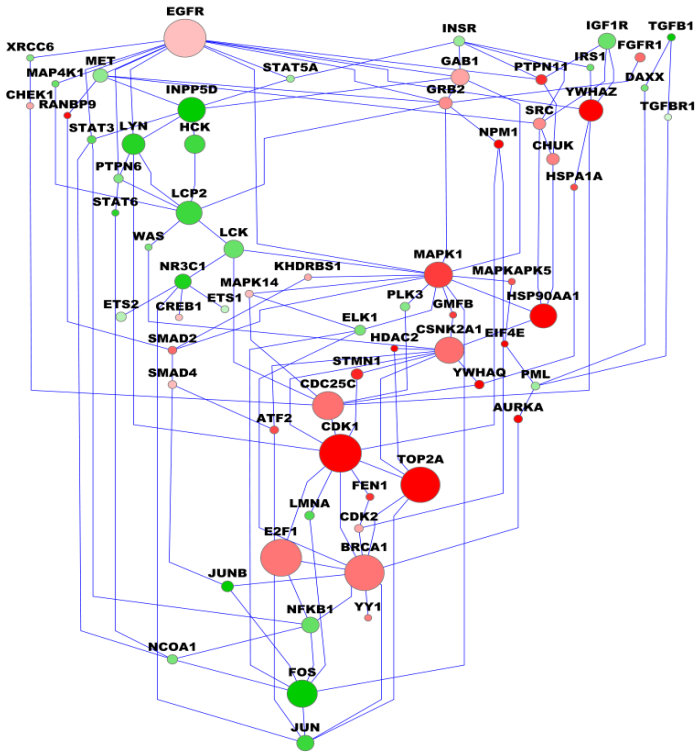
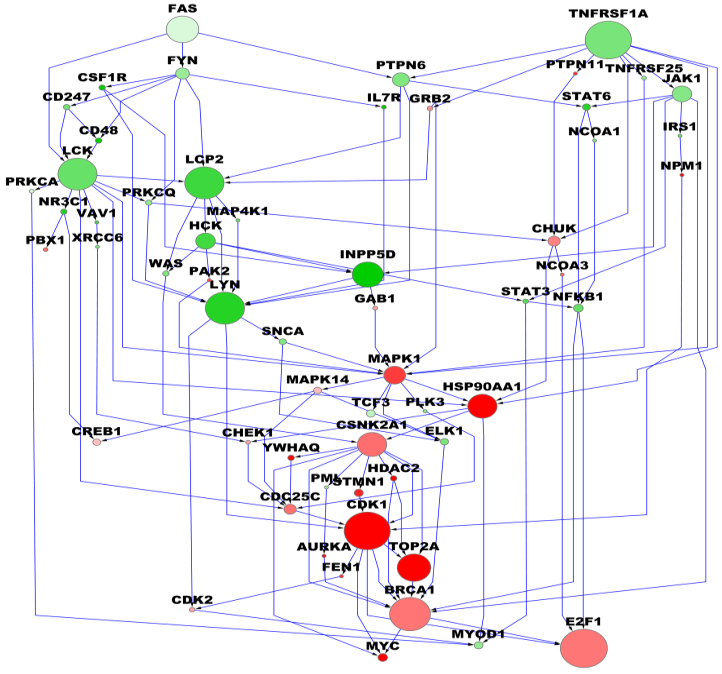


Figure B.1. ER signaling pathways identified from Symmans data



(a)



(b)

Figure B.2. Cell cycle and apoptosis pathways identified from Symmans data: (a) cell cycle pathways identified from Symmans data; (b) apoptosis pathways identified from Symmans data. Color map is the same as Figure B.1.

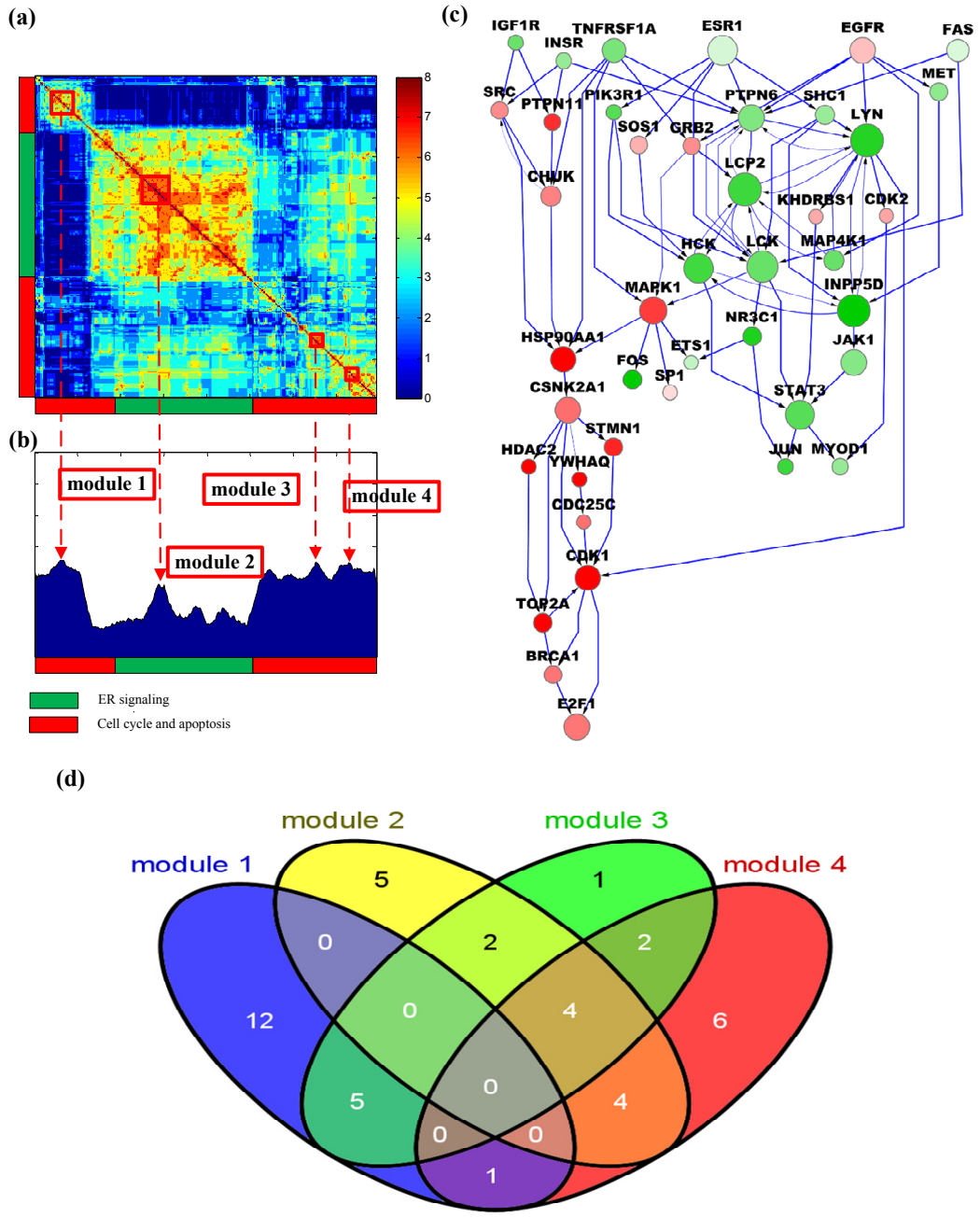


Figure B.3. Pathway landscape and identified modules from Symmans data. (a) Structural heatmap of sampled pathways from the GIST algorithm on Symmans dataset. (b) Potential distribution of pathway samples. Pathway clusters that correspond to locally enriched potentials are identified as pathway modules. Four pathway modules have been identified from Symmans data. (c) Layout of pathway network from four identified pathway modules. The color map is the same as in Figure B.1. (d) 4-way Venn diagram showing the overlap among four pathway modules.

Table B.2. Proteins in identified pathway modules from Symmans data

Module 1	Module 2	Module 3	Module 4
BRCA1	ESR1	CDC2	CDK2
CDC2	HCK	E2F1	EGFR
CDC25C	INPP5D	EGFR	ETS1
CHUK	JAK1	FAS	FAS
CSNK2A1	JUN	GRB2	FOS
E2F1	KHDRBS1	HCK	HCK
GRB2	LCK	INPP5D	INPP5D
HDAC2	LCP2	INSR	JAK1
HSP90AA1	LYN	LCP2	JUN
IGF1R	MAP4K1	LYN	LCK
INSR	PIK3R1	MAP4K1	LCP2
MAPK1	PTPN6	MET	LYN
PTPN11	SHC1	PTPN6	MAPK1
SRC	SOS1	TNFRSF1A	MYOD1
STMN1	STAT3		NR3C1
TNFRSF1A			SP1
TOP2A			STAT3
YWHAQ			

Bibliography

[1] A. Ma'ayan, "Introduction to network analysis in systems biology," *Sci Signal,* vol. 4, p. tr5, Sep 13 2011.

[2] S. Navlakha and Z. Bar-Joseph, "Algorithms in nature: the convergence of systems biology and computational thinking," *Mol Syst Biol,* vol. 7, p. 546, 2011.

[3] M. L. Metzker, "Sequencing technologies - the next generation," *Nat Rev Genet,* vol. 11, pp. 31-46, Jan 2010.

[4] A. Git*, et al.*, "Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression," *RNA,* vol. 16, pp. 991-1006, May 2010.

[5] H. Kitano, "Computational systems biology," *Nature,* vol. 420, pp. 206-10, Nov 14 2002.

[6] J. P. Mathew*, et al.*, "From bytes to bedside: data integration and computational biology for translational cancer research," *PLoS Comput Biol,* vol. 3, p. e12, Feb 23 2007.

[7] P. C. Ng*, et al.*, "An agenda for personalized medicine," *Nature,* vol. 461, pp. 724-6, Oct 8 2009.

[8] "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature,* vol. 455, pp. 1061-8, Oct 23 2008.

[9] "Integrated genomic analyses of ovarian carcinoma," *Nature,* vol. 474, pp. 609-15, Jun 30 2011.

[10] K. R. Christie*, et al.*, "Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms," *Nucleic Acids Res,* vol. 32, pp. D311-4, Jan 1 2004.

[11] S. Weng*, et al.*, "Saccharomyces Genome Database (SGD) provides biochemical and structural information for budding yeast proteins," *Nucleic Acids Res,* vol. 31, pp. 216-8, Jan 1 2003.

[12] S. S. Dwight*, et al.*, "Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO)," *Nucleic Acids Res,* vol. 30, pp. 69-72, Jan 1 2002.

[13] M. Kuhn*, et al.*, "STITCH 2: an interaction network database for small molecules and proteins," *Nucleic Acids Res,* vol. 38, pp. D552-6, Jan 2010.

[14] M. Safran*, et al.*, "GeneCards Version 3: the human gene integrator," *Database (Oxford),* vol. 2010, p. baq020, 2010.

[15] G. Stelzer*, et al.*, "GeneDecks: paralog hunting and gene-set distillation with GeneCards annotation," *OMICS,* vol. 13, pp. 477-87, Dec 2009.

[16] S. Djuranovic*, et al.*, "miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay," *Science,* vol. 336, pp. 237-40, Apr 13 2012.

[17] J. R. Buchan and R. Parker, "Molecular biology. The two faces of miRNA," *Science,* vol. 318, pp. 1877-8, Dec 21 2007.

[18] B. S. Shastry, "SNP alleles in human disease and evolution," *J Hum Genet,* vol. 47, pp. 561-6, 2002.
[19] J. Staaf*, et al.*, "High-resolution genomic and expression analyses of copy number alterations in HER2-amplified breast cancer," *Breast Cancer Res,* vol. 12, p. R25, 2010.
[20] P. M. Haverty*, et al.*, "High-resolution genomic and expression analyses of copy number alterations in breast tumors," *Genes Chromosomes Cancer,* vol. 47, pp. 530-42, Jun 2008.
[21] S. Temam*, et al.*, "Epidermal growth factor receptor copy number alterations correlate with poor clinical outcome in patients with head and neck squamous cancer," *J Clin Oncol,* vol. 25, pp. 2164-70, Jun 1 2007.
[22] P. A. Jones, "Cancer. Death and methylation," *Nature,* vol. 409, pp. 141, 143-4, Jan 11 2001.
[23] S. de Assis*, et al.*, "High-fat or ethinyl-oestradiol intake during pregnancy increases mammary cancer risk in several generations of offspring," *Nat Commun,* vol. 3, p. 1053, 2012.
[24] J. Dopazo*, et al.*, "Methods and approaches in the analysis of gene expression data," *J Immunol Methods,* vol. 250, pp. 93-112, Apr 2001.
[25] V. G. Tusher*, et al.*, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc Natl Acad Sci U S A,* vol. 98, pp. 5116-21, Apr 24 2001.
[26] V. Pedraza*, et al.*, "Gene expression signatures in breast cancer distinguish phenotype characteristics, histologic subtypes, and tumor invasiveness," *Cancer,* vol. 116, pp. 486-96, Jan 15 2010.
[27] M. Mutarelli*, et al.*, "Time-course analysis of genome-wide gene expression data from hormone-responsive human breast cancer cells," *BMC Bioinformatics,* vol. 9 Suppl 2, p. S12, 2008.
[28] M. J. Fullwood*, et al.*, "Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses," *Genome Res,* vol. 19, pp. 521-32, Apr 2009.
[29] I. Hajirasouliha*, et al.*, "Detection and characterization of novel sequence insertions using paired-end next-generation sequencing," *Bioinformatics,* vol. 26, pp. 1277-83, May 15 2010.
[30] S. K. Lou*, et al.*, "ABMapper: a suffix array-based tool for multi-location searching and splice-junction mapping," *Bioinformatics,* vol. 27, pp. 421-2, Feb 1 2011.
[31] M. Kinsella*, et al.*, "Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs," *Bioinformatics,* vol. 27, pp. 1068-75, Apr 15 2011.
[32] B. F. Cravatt*, et al.*, "The biological impact of mass-spectrometry-based proteomics," *Nature,* vol. 450, pp. 991-1000, Dec 13 2007.
[33] T. I. Lee*, et al.*, "Chromatin immunoprecipitation and microarray-based analysis of protein location," *Nat Protoc,* vol. 1, pp. 729-48, 2006.
[34] A. Visel*, et al.*, "ChIP-seq accurately predicts tissue-specific activity of enhancers," *Nature,* vol. 457, pp. 854-8, Feb 12 2009.
[35] W. W. Wasserman and A. Sandelin, "Applied bioinformatics for the identification of regulatory elements," *Nat Rev Genet,* vol. 5, pp. 276-87, Apr 2004.

[36] V. Matys, *et al.*, "TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes," *Nucleic Acids Res,* vol. 34, pp. D108-10, Jan 1 2006.
[37] B. Causier and B. Davies, "Analysing protein-protein interactions with the yeast two-hybrid system," *Plant Mol Biol,* vol. 50, pp. 855-70, Dec 2002.
[38] S. Mathivanan, *et al.*, "An evaluation of human protein-protein interaction data in the public domain," *BMC Bioinformatics,* vol. 7 Suppl 5, p. S19, 2006.
[39] D. Szklarczyk, *et al.*, "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Res,* vol. 39, pp. D561-8, Jan 2011.
[40] M. S. Skrzypek and J. Hirschman, "Using the Saccharomyces Genome Database (SGD) for analysis of genomic information," *Curr Protoc Bioinformatics,* vol. Chapter 1, pp. Unit 1 20 1-23, Sep 2011.
[41] W. Huang da, *et al.*, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nat Protoc,* vol. 4, pp. 44-57, 2009.
[42] W. Huang da, *et al.*, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Res,* vol. 37, pp. 1-13, Jan 2009.
[43] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Res,* vol. 28, pp. 27-30, Jan 1 2000.
[44] M. P. Brynildsen, *et al.*, "Versatility and connectivity efficiency of bipartite transcription networks," *Biophys J,* vol. 91, pp. 2749-59, Oct 15 2006.
[45] T. Aittokallio and B. Schwikowski, "Graph-based methods for analysing networks in cell biology," *Brief Bioinform,* vol. 7, pp. 243-55, Sep 2006.
[46] M. Y. Becker and I. Rojas, "A graph layout algorithm for drawing metabolic pathways," *Bioinformatics,* vol. 17, pp. 461-7, May 2001.
[47] G. E. Zinman, *et al.*, "Biological interaction networks are conserved at the module level," *BMC Syst Biol,* vol. 5, p. 134, 2011.
[48] C. Wang, "From network to pathway: integrative network analysis of genomic data," *Virginia tech PhD dissertation,* 2011.
[49] L. Chen, "Integrative Modeling and Analysis of High-throughput Biological Data," *Virginia Tech Dissertation,* Dec. 15 2010.
[50] C. Andrieu, *et al.*, "An introduction to MCMC for machine learning," *Machine Learning,* vol. 50, pp. 5-43, Jan-Feb 2003.
[51] G. Casella and E. I. George, "Explaining the Gibbs Sampler," *American Statistician,* vol. 46, pp. 167-174, Aug 1992.
[52] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *Ieee Transactions on Pattern Analysis and Machine Intelligence,* vol. 6, pp. 721-741, 1984.
[53] E. Segal, *et al.*, "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data," *Nat Genet,* vol. 34, pp. 166-76, Jun 2003.
[54] N. Friedman, *et al.*, "Using Bayesian networks to analyze expression data," *J Comput Biol,* vol. 7, pp. 601-20, 2000.

[55] M. A. Mahdavi and Y. H. Lin, "False positive reduction in protein-protein interaction predictions using gene ontology annotations," *BMC Bioinformatics,* vol. 8, p. 262, 2007.
[56] J. P. Ioannidis, "Microarrays and molecular research: noise discovery?," *Lancet,* vol. 365, pp. 454-5, Feb 5-11 2005.
[57] P. Mendes*, et al.*, "Artificial gene networks for objective comparison of analysis algorithms," *Bioinformatics,* vol. 19 Suppl 2, pp. ii122-9, Oct 2003.
[58] W. L. Allen*, et al.*, "A systems biology approach identifies SART1 as a novel determinant of both 5-fluorouracil and SN38 drug resistance in colorectal cancer," *Mol Cancer Ther,* vol. 11, pp. 119-31, Jan 2012.
[59] J. Y. Chen*, et al.*, "A systems biology case study of ovarian cancer drug resistance," *Comput Syst Bioinformatics Conf,* pp. 389-98, 2006.
[60] J. Y. Chen*, et al.*, "A systems biology approach to the study of cisplatin drug resistance in ovarian cancers," *J Bioinform Comput Biol,* vol. 5, pp. 383-405, Apr 2007.
[61] T. Ito*, et al.*, "Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins," *Proc Natl Acad Sci U S A,* vol. 97, pp. 1143-7, Feb 1 2000.
[62] D. C. Altieri, "Survivin, cancer networks and pathway-directed drug discovery," *Nat Rev Cancer,* vol. 8, pp. 61-70, Jan 2008.
[63] A. C. Mita*, et al.*, "Survivin: key regulator of mitosis and apoptosis and novel target for cancer therapeutics," *Clin Cancer Res,* vol. 14, pp. 5000-5, Aug 15 2008.
[64] S. Loi*, et al.*, "Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen," *BMC Genomics,* vol. 9, p. 239, 2008.
[65] W. F. Symmans*, et al.*, "Genomic index of sensitivity to endocrine therapy for breast cancer," *J Clin Oncol,* vol. 28, pp. 4111-9, Sep 20 2010.
[66] O. Alter*, et al.*, "Singular value decomposition for genome-wide expression data processing and modeling," *Proc Natl Acad Sci U S A,* vol. 97, pp. 10101-6, Aug 29 2000.
[67] S. I. Lee and S. Batzoglou, "Application of independent component analysis to microarrays," *Genome Biol,* vol. 4, p. R76, 2003.
[68] I. Shmulevich*, et al.*, "Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks," *Bioinformatics,* vol. 18, pp. 261-74, Feb 2002.
[69] J. C. Liao*, et al.*, "Network component analysis: reconstruction of regulatory signals in biological systems," *Proc Natl Acad Sci U S A,* vol. 100, pp. 15522-7, Dec 23 2003.
[70] C. Chang*, et al.*, "Fast network component analysis (FastNCA) for gene regulatory network reconstruction from microarray data," *Bioinformatics,* vol. 24, pp. 1349-58, Jun 1 2008.
[71] M. P. Brynildsen*, et al.*, "A Gibbs sampler for the identification of gene expression and network connectivity consistency," *Bioinformatics,* vol. 22, pp. 3040-6, Dec 15 2006.

[72] D. C. Montgomery, "Introduction to linear regression analysis," *Wiley Interscience,* 2006.
[73] R. Tibshirani and T. Hastie, "Outlier sums for differential gene expression analysis," *Biostatistics,* vol. 8, pp. 2-8, Jan 2007.
[74] G. Casella and E. I. George, "Explaining the Gibbs sampler," *The American Statistician,* vol. 46, p. 8, 1992.
[75] T. Van den Bulcke*, et al.*, "SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms," *BMC Bioinformatics,* vol. 7, p. 43, 2006.
[76] S. Hempel*, et al.*, "Unraveling gene regulatory networks from time-resolved gene expression data - a measures comparison study," *BMC Bioinformatics,* vol. 12, p. 292, 2011.
[77] B. Efron*, et al.*, "Least angle regression," *Annals of Statistics,* vol. 32, 2004.
[78] G. Chen*, et al.*, "Clustering of genes into regulons using integrated modeling-COGRIM," *Genome Biol,* vol. 8, p. R4, 2007.
[79] Z. Chen, "Assessing sequence comparison methods with the average precision criterion," *Bioinformatics,* vol. 19, pp. 2456-60, Dec 12 2003.
[80] R. Boscolo*, et al.*, "A generalized framework for network component analysis," *IEEE/ACM Trans Comput Biol Bioinform,* vol. 2, pp. 289-301, Oct-Dec 2005.
[81] L. M. Tran*, et al.*, "gNCA: a framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation," *Metab Eng,* vol. 7, pp. 128-41, Mar 2005.
[82] P. T. Spellman*, et al.*, "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," *Mol Biol Cell,* vol. 9, pp. 3273-97, Dec 1998.
[83] C. T. Harbison*, et al.*, "Transcriptional regulatory code of a eukaryotic genome," *Nature,* vol. 431, pp. 99-104, Sep 2 2004.
[84] Y. L. Yang*, et al.*, "Inferring yeast cell cycle regulators and interactions using transcription factor activities," *BMC Genomics,* vol. 6, p. 90, 2005.
[85] A. Sutton*, et al.*, "Yeast ASF1 protein is required for cell cycle regulation of histone gene transcription," *Genetics,* vol. 158, pp. 587-96, Jun 2001.
[86] C. J. Creighton*, et al.*, "Genes regulated by estrogen in breast tumor cells in vitro are similarly regulated in vivo in tumor xenografts and human breast tumors," *Genome Biol,* vol. 7, p. R28, 2006.
[87] H. Aguilar*, et al.*, "Biological reprogramming in acquired resistance to endocrine therapy of breast cancer," *Oncogene,* vol. 29, pp. 6071-83, Nov 11 2010.
[88] S. Masamura*, et al.*, "Estrogen deprivation causes estradiol hypersensitivity in human breast cancer cells," *J Clin Endocrinol Metab,* vol. 80, pp. 2918-25, Oct 1995.
[89] L. A. Martin*, et al.*, "Enhanced estrogen receptor (ER) alpha, ERBB2, and MAPK signal transduction pathways operate during the adaptation of MCF-7 cells to long term estrogen deprivation," *J Biol Chem,* vol. 278, pp. 30458-68, Aug 15 2003.
[90] L. Chen*, et al.*, "Multilevel support vector regression analysis to identify condition-specific regulatory networks," *Bioinformatics,* vol. 26, pp. 1416-22, Jun 1 2010.

[91] R. B. Riggins, *et al.*, "The nuclear factor kappa B inhibitor parthenolide restores ICI 182,780 (Faslodex; fulvestrant)-induced apoptosis in antiestrogen-resistant breast cancer cells," *Mol Cancer Ther,* vol. 4, pp. 33-41, Jan 2005.
[92] M. A. Pratt, *et al.*, "Estrogen withdrawal-induced NF-kappaB activity and bcl-3 expression in breast cancer cells: roles in growth and hormone independence," *Mol Cell Biol,* vol. 23, pp. 6887-900, Oct 2003.
[93] H. Yamashita, *et al.*, "Naturally occurring dominant-negative Stat5 suppresses transcriptional activity of estrogen receptors and induces apoptosis in T47D breast cancer cells," *Oncogene,* vol. 22, pp. 1638-52, Mar 20 2003.
[94] R. B. Riggins, *et al.*, "Physical and functional interactions between Cas and c-Src induce tamoxifen resistance of breast cancer cells through pathways involving epidermal growth factor receptor and signal transducer and activator of transcription 5b," *Cancer Res,* vol. 66, pp. 7007-15, Jul 15 2006.
[95] E. M. Fox, *et al.*, "Signal transducer and activator of transcription 5b, c-Src, and epidermal growth factor receptor signaling play integral roles in estrogen-stimulated proliferation of estrogen receptor-positive breast cancer cells," *Mol Endocrinol,* vol. 22, pp. 1781-96, Aug 2008.
[96] D. Hungermann, *et al.*, "Influence of whole arm loss of chromosome 16q on gene expression patterns in oestrogen receptor-positive, invasive breast cancer," *J Pathol,* vol. 224, pp. 517-28, Aug 2011.
[97] E. H. Han, *et al.*, "Prostaglandin E2 induces CYP1B1 expression via ligand-independent activation of the ERalpha pathway in human breast cancer cells," *Toxicol Sci,* vol. 114, pp. 204-16, Apr 2010.
[98] W. G. Angus, *et al.*, "Expression of CYP1A1 and CYP1B1 depends on cell-specific factors in human breast cancer cell lines: role of estrogen receptor status," *Carcinogenesis,* vol. 20, pp. 947-55, Jun 1999.
[99] F. Fang, *et al.*, "Quantification of PRL/Stat5 signaling with a novel pGL4-CISH reporter," *BMC Biotechnol,* vol. 8, p. 11, 2008.
[100] B. He, *et al.*, "A repressive role for prohibitin in estrogen signaling," *Mol Endocrinol,* vol. 22, pp. 344-60, Feb 2008.
[101] B. H. Kang, *et al.*, "Combinatorial drug design targeting multiple cancer signaling networks controlled by mitochondrial Hsp90," *J Clin Invest,* vol. 119, pp. 454-64, Mar 2009.
[102] B. H. Kang and D. C. Altieri, "Compartmentalized cancer drug discovery targeting mitochondrial Hsp90 chaperones," *Oncogene,* vol. 28, pp. 3681-8, Oct 22 2009.
[103] L. Rajendran, *et al.*, "Subcellular targeting strategies for drug design and delivery," *Nat Rev Drug Discov,* vol. 9, pp. 29-42, Jan 2010.
[104] M. Kanehisa, *et al.*, "KEGG for integration and interpretation of large-scale molecular data sets," *Nucleic Acids Res,* vol. 40, pp. D109-14, Jan 2012.
[105] A. Subramanian, *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proc Natl Acad Sci U S A,* vol. 102, pp. 15545-50, Oct 25 2005.
[106] S. Efroni, *et al.*, "Identification of key processes underlying cancer phenotypes using biologic pathway analysis," *PLoS One,* vol. 2, p. e425, 2007.

[107] A. L. Tarca, *et al.*, "A novel signaling pathway impact analysis," *Bioinformatics,* vol. 25, pp. 75-82, Jan 1 2009.
[108] C. J. Vaske, *et al.*, "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM," *Bioinformatics,* vol. 26, pp. i237-45, Jun 15 2010.
[109] R. Amanchy, *et al.*, "A curated compendium of phosphorylation motifs," *Nat Biotechnol,* vol. 25, pp. 285-6, Mar 2007.
[110] X. M. Zhao, *et al.*, "Uncovering signal transduction networks from high-throughput data by integer linear programming," *Nucleic Acids Res,* vol. 36, p. e48, May 2008.
[111] S. Mathivanan, *et al.*, "Human Proteinpedia enables sharing of human protein data," *Nat Biotechnol,* vol. 26, pp. 164-7, Feb 2008.
[112] M. Steffen, *et al.*, "Automated modelling of signal transduction networks," *BMC Bioinformatics,* vol. 3, p. 34, Nov 1 2002.
[113] J. Scott, *et al.*, "Efficient algorithms for detecting signaling pathways in protein interaction networks," *J Comput Biol,* vol. 13, pp. 133-44, Mar 2006.
[114] F. Huffner, *et al.*, "Faspad: fast signaling pathway detection," *Bioinformatics,* vol. 23, pp. 1708-9, Jul 1 2007.
[115] A. Lan, *et al.*, "ResponseNet: revealing signaling and regulatory networks linking genetic and transcriptomic screening data," *Nucleic Acids Res,* vol. 39, pp. W424-9, Jul 2011.
[116] E. Yeger-Lotem, *et al.*, "Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity," *Nat Genet,* vol. 41, pp. 316-23, Mar 2009.
[117] N. Yosef, *et al.*, "Toward accurate reconstruction of functional protein networks," *Mol Syst Biol,* vol. 5, p. 248, 2009.
[118] J. Supper, *et al.*, "BowTieBuilder: modeling signal transduction pathways," *BMC Syst Biol,* vol. 3, p. 67, 2009.
[119] G. Bebek and J. Yang, "PathFinder: mining signal transduction pathway segments from protein-protein interaction networks," *BMC Bioinformatics,* vol. 8, p. 335, 2007.
[120] W. Liu, *et al.*, "Proteome-wide prediction of signal flow direction in protein interaction networks based on interacting domains," *Mol Cell Proteomics,* vol. 8, pp. 2063-70, Sep 2009.
[121] C. H. Yeang, *et al.*, "Physical network models," *J Comput Biol,* vol. 11, pp. 243-62, 2004.
[122] A. Gitter, *et al.*, "Discovering pathways by orienting edges in protein interaction networks," *Nucleic Acids Res,* vol. 39, p. e22, Mar 2011.
[123] A. Gitter, *et al.*, "Linking the signaling cascades and dynamic regulatory networks controlling stress responses," *Genome Res,* vol. 23, pp. 365-76, Feb 2013.
[124] J. Gu, *et al.*, "Robust identification of transcriptional regulatory networks using a Gibbs sampler on outlier sum statistic," *Bioinformatics,* vol. 28, pp. 1990-7, Aug 1 2012.
[125] W. M. Bolstad, "Understanding Computational Bayesian Statistics," 2009.
[126] D. R. Cox and D. Oakes, *Analysis of survival data*. London: Chapman & Hall, 1984.

[127] S. Z. Li, *Markov random field modeling in computer vision*: Spinger-Verlag Tokyo, 1995.
[128] J. Gu*, et al.*, "GIST: a Gibbs sampler to identify intracellular signal transduction pathways," *Conf Proc IEEE Eng Med Biol Soc,* vol. 2011, pp. 2434-7, 2011.
[129] Y. Li*, et al.*, "A global pathway crosstalk network," *Bioinformatics,* vol. 24, pp. 1442-7, Jun 15 2008.
[130] A. W. Rives and T. Galitski, "Modular organization of cellular networks," *Proc Natl Acad Sci U S A,* vol. 100, pp. 1128-33, Feb 4 2003.
[131] G. M. Ibrahim*, et al.*, "Network analysis reveals patterns of antiepileptic drug use in children with medically intractable epilepsy," *Epilepsy Behav,* vol. 28, pp. 22-5, Jul 2013.
[132] U. Brandes, "On variants of shortest-path betweenness centrality and their generic computation," *Social Networks,* vol. 30, pp. 136-145, May 2008.
[133] Y. Gong and Z. Zhang, "Alternative signaling pathways: when, where and why?," *FEBS Lett,* vol. 579, pp. 5265-74, Oct 10 2005.
[134] H. Saito, "Regulation of cross-talk in yeast MAPK signaling pathways," *Curr Opin Microbiol,* vol. 13, pp. 677-83, Dec 2010.
[135] G. Dennis, Jr.*, et al.*, "DAVID: Database for Annotation, Visualization, and Integrated Discovery," *Genome Biol,* vol. 4, p. P3, 2003.
[136] I. Xenarios*, et al.*, "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions," *Nucleic Acids Res,* vol. 30, pp. 303-5, Jan 1 2002.
[137] I. Xenarios*, et al.*, "DIP: The Database of Interacting Proteins: 2001 update," *Nucleic Acids Res,* vol. 29, pp. 239-41, Jan 1 2001.
[138] I. Xenarios*, et al.*, "DIP: the database of interacting proteins," *Nucleic Acids Res,* vol. 28, pp. 289-91, Jan 1 2000.
[139] R. Sharan*, et al.*, "Conserved patterns of protein interaction in multiple species," *Proc Natl Acad Sci U S A,* vol. 102, pp. 1974-9, Feb 8 2005.
[140] N. Ogawa*, et al.*, "New components of a system for phosphate accumulation and polyphosphate metabolism in Saccharomyces cerevisiae revealed by genomic expression analysis," *Mol Biol Cell,* vol. 11, pp. 4309-21, Dec 2000.
[141] S. Rupp*, et al.*, "MAP kinase and cAMP filamentation signaling pathways converge on the unusually large promoter of the yeast FLO11 gene," *EMBO J,* vol. 18, pp. 1257-69, Mar 1 1999.
[142] W. E. Johnson*, et al.*, "Adjusting batch effects in microarray expression data using empirical Bayes methods," *Biostatistics,* vol. 8, pp. 118-27, Jan 2007.
[143] E. L. Kaplan, "Citation Classic - Nonparametric-Estimation from Incomplete Observations," *Current Contents/Life Sciences,* pp. 14-14, 1983.
[144] P. Sonego*, et al.*, "ROC analysis: applications to the classification of biological sequences and 3D structures," *Brief Bioinform,* vol. 9, pp. 198-209, May 2008.
[145] M. A. Martinez-Gonzalez*, et al.*, "[What is hazard ratio? Concepts in survival analysis]," *Med Clin (Barc),* vol. 131, pp. 65-72, Jun 14 2008.
[146] B. Pucci*, et al.*, "Cell cycle and apoptosis," *Neoplasia,* vol. 2, pp. 291-9, Jul-Aug 2000.
[147] Q. Wang*, et al.*, "Cyclin dependent kinase 1 inhibitors: a review of recent progress," *Curr Med Chem,* vol. 18, pp. 2025-43, 2011.

[148] Y. Matthess, *et al.*, "Cdk1/cyclin B1 controls Fas-mediated apoptosis by regulating caspase-8 activity," *Mol Cell Biol,* vol. 30, pp. 5726-40, Dec 2010.
[149] R. Clarke, *et al.*, "Cellular and molecular pharmacology of antiestrogen action and resistance," *Pharmacol Rev,* vol. 53, pp. 25-71, Mar 2001.
[150] N. Brunner, *et al.*, "MCF7/LCC2: a 4-hydroxytamoxifen resistant human breast cancer variant that retains sensitivity to the steroidal antiestrogen ICI 182,780," *Cancer Res,* vol. 53, pp. 3229-32, Jul 15 1993.
[151] R. Nehra, *et al.*, "BCL2 and CASP8 regulation by NF-kappaB differentially affect mitochondrial function and cell fate in antiestrogen-sensitive and -resistant breast cancer cells," *FASEB J,* vol. 24, pp. 2040-55, Jun 2010.
[152] R. Clarke, *et al.*, "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,* vol. 86, pp. 3649-53, May 1989.
[153] F. U. Hartl, "Molecular chaperones in cellular protein folding," *Nature,* vol. 381, pp. 571-9, Jun 13 1996.
[154] C. Wong and S. Chen, "Heat shock protein 90 inhibitors: new mode of therapy to overcome endocrine resistance," *Cancer Res,* vol. 69, pp. 8670-7, Nov 15 2009.
[155] Z. Bar-Joseph, *et al.*, "Computational discovery of gene modules and regulatory networks," *Nat Biotechnol,* vol. 21, pp. 1337-42, Nov 2003.
[156] T. Gong, *et al.*, "Motif-guided sparse decomposition of gene expression data for regulatory module identification," *BMC Bioinformatics,* vol. 12, p. 82, 2011.
[157] G. Schramm, *et al.*, "Regulation patterns in signaling networks of cancer," *BMC Syst Biol,* vol. 4, p. 162, 2010.
[158] O. Kuchaiev, *et al.*, "GraphCrunch 2: Software tool for network modeling, alignment and clustering," *BMC Bioinformatics,* vol. 12, p. 24, 2011.
[159] T. Milenkovic, *et al.*, "GraphCrunch: a tool for large network analyses," *BMC Bioinformatics,* vol. 9, p. 70, 2008.
[160] T. Opsahl, *et al.*, "Node centrality in weighted networks: Generalizing degree and shortest paths," *Social Networks,* vol. 32, pp. 245-251, Jul 2010.
[161] H. Y. Chuang, *et al.*, "Network-based classification of breast cancer metastasis," *Mol Syst Biol,* vol. 3, p. 140, 2007.
[162] F. Emmert-Streib and G. V. Glazko, "Pathway analysis of expression data: deciphering functional building blocks of complex diseases," *PLoS Comput Biol,* vol. 7, p. e1002053, May 2011.
[163] P. Shannon, *et al.*, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Res,* vol. 13, pp. 2498-504, Nov 2003.
[164] X. Wang, *et al.*, "Sampling-based Subnetwork Identification from Microarray Data and Protein-protein Interaction Network," in *Proc. ICMLA 2012: Machine Learning in Health Informatics Workshop*, Boca Raton, Florida, 2012.
[165] B. Zhang, *et al.*, "DDN: a caBIG(R) analytical tool for differential network analysis," *Bioinformatics,* vol. 27, pp. 1036-8, Apr 1 2011.