Module-based Analysis of Biological Data for Network Inference and Biomarker Discovery

Yuji Zhang

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, Chair

Habtom W. Ressom, Co-Chair

Chang-Tien Lu

Christopher L. Wyatt

William T. Baumann

Yue J. Wang

July 20th, 2010

Arlington, Virginia

Keywords: Data Integration, Network Modeling, Gene Module Identification, Gene Regulatory Module, Biomarker Discovery

Module-based Analysis of Biological Data for Network Inference and Biomarker Discovery

Yuji Zhang

ABSTRACT

Systems biology comprises the global, integrated analysis of large-scale data encoding different levels of biological information with the aim to obtain global insight into the cellular networks. Several studies have unveiled the modular and hierarchical organization inherent in these networks. In this dissertation, we propose and develop innovative systems approaches to integrate multi-source biological data in a modular manner for network inference and biomarker discovery in complex diseases such as breast cancer.

The first part of the dissertation is focused on gene module identification in gene expression data. As the most popular way to identify gene modules, many cluster algorithms have been applied to the gene expression data analysis. For the purpose of evaluating clustering algorithms from a biological point of view, we propose a figure of merit based on Kullback-Leibler divergence between cluster membership and known gene ontology attributes. Several benchmark expression-based gene clustering algorithms are compared using the proposed method with different parameter settings. Applications to diverse public time course gene expression data demonstrated that fuzzy c-means clustering is superior to other clustering methods with regard to the enrichment of clusters for biological functions. These results contribute to the evaluation of clustering outcomes and the estimations of optimal clustering partitions.

The second part of the dissertation presents a hybrid computational intelligence method to infer gene regulatory modules. We explore the combined advantages of the nonlinear and dynamic properties of neural networks, and the global search capabilities of the hybrid genetic algorithm and particle swarm optimization method to infer network interactions at modular level. The proposed computational framework is tested in two biological processes: yeast cell cycle, and human *Hela* cancer cell cycle. The identified gene regulatory modules were evaluated using several validation strategies: 1) gene set enrichment analysis to evaluate the gene modules derived from clustering results; (2) binding site enrichment analysis to determine enrichment of

the gene modules for the cognate binding sites of their predicted transcription factors; (3) comparison with previously reported results in the literatures to confirm the inferred regulations. The proposed framework could be beneficial to biologists for predicting the components of gene regulatory modules in which any candidate gene is involved. Such predictions can then be used to design a more streamlined experimental approach for biological validation. Understanding the dynamics of these gene regulatory modules will shed light on the related regulatory processes.

Driven by the fact that complex diseases such as cancer are “diseases of pathways”, we extended the module concept to biomarker discovery in cancer research. In the third part of the dissertation, we explore the combined advantages of molecular interaction network and gene expression profiles to identify biomarkers in cancer research. The reliability of conventional gene biomarkers has been challenged because of the biological heterogeneity and noise within and across patients. In this dissertation, we present a module-based biomarker discovery approach that integrates interaction network topology and high-throughput gene expression data to identify markers not as individual genes but as modules. To select reliable biomarker sets across different studies, a hybrid method combining group feature selection with ensemble feature selection is proposed. First, a group feature selection method is used to extract the modules (subnetworks) with discriminative power between disease groups. Then, an ensemble feature selection method is used to select the optimal biomarker sets, in which a double-validation strategy is applied. The ensemble method allows combining features selected from multiple classifications with various data subsampling to increase the reliability and classification accuracy of the final selected biomarker set. The results from four breast cancer studies demonstrated the superiority of the module biomarkers identified by the proposed approach: they can achieve higher accuracies, and are more reliable in datasets with same clinical design.

Based on the experimental results above, we believe that the proposed systems approaches provide meaningful solutions to discover the cellular regulatory processes and improve the understanding about disease mechanisms. These computational approaches are primarily developed for analysis of high-throughput genomic data. Nevertheless, the proposed methods can also be extended to analyze high-throughput data in proteomics and metablomics areas.

Acknowledgments

I would like to express my sincere appreciation to my respected supervisors Dr. Jason J. Xuan and Dr. Habtom W. Ressom, for their invaluable guidance, encouragement, and support during my Ph.D. program as well as the critical reviews and comments on my publications and this dissertation. They are not only insightful advisors, but also good friends. I am happy to meet and work together with them in past five years.

I would like to especially thank Dr. Yue J. Wang for granting me the opportunity to work in the Computational Bioinformatics and Bio-imaging Laboratory five years ago. Without his invaluable help and advice, I couldn't have achieved so much progress over last five years.

I would like to extend my appreciation to my advisory committee members: Dr. Chang-Tien Lu, Dr. Christopher L. Wyatt and Dr. William T. Baumann, for their valuable examinations and constructive suggestions to improve the present work. They have provided valuable discussions with me about many technical aspects in this dissertation, suggested improvements in my presentation, and shared their insightful feedback.

I also want to thank Dr. Robert Clarke and Dr. Benildo G. de los Reyes for their generous support, guidance and resources designed to improve the quality of my research and publications during my graduate study. As biologists, they helped me propose the biological problems to be studied and provided biological interpretations of the inferred computational results.

I would also like to acknowledge my colleagues and friends for their friendship, discussion and help to make my study life easy and interesting. I have benefitted greatly from discussions with them in the group meetings of discussing challenging and interesting bioinformatics problems. Their suggestions also contributed to the formation of the research solutions presented in this dissertation.

DEDICATION

To

my dear husband Jian and my lovely daughter Lyra,

for

their love, encouragement, and devoted support

TABLE OF CONTENTS

LIST OF FIGURES

List of Tables

List of Abbreviations

ACO	Ant Colony Optimization
AUC	Area Under the ROC Curve
BCG	Breast Cancer Gene
BPTT	Backpropagation Through Time
BSEA	Binding Site Enrichment Analysis
cDNA	complementary DeoxyriboNucleic Acid
ChIP-chip	Chromatin Immunoprecipitation (ChIP) on a Microarray (chip)
CORG	Condition-Responsive Gene
CNS	Central Nervous System
CV	Cross Validation
DBN	Dynamic Bayesian Network
DNA	Deoxy Ribonucleic Acid
EST	Expressed Sequence Tag
FCM	Fuzzy c-means
FDR	False Discovery Rate
FP	False Positive
GA	Genetic Algorithm
GEM	Gene Expression Microarray
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
GWLA	Genome Wide Location Analysis
HC	Hierarchical Clustering
IPA	Ingenuity Pathway Analysis
KCM	K-means Clustering
KL	Kullback-Leibler

KNN	K-nearest Neighbor
mRNA	Messenger RNA
NM	Network Motif
NMR	Nuclear Magnetic Resonance
OMIM	Online Mendelian Inheritance in Man
PDI	Protein-DNA Interaction
PPI	Protein-Protein Interaction
PSO	Particle Swarm Optimization
RMSE	Root Mean Square Error
RNA	Ribonucleic Acid
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristics
rRNA	Ribosomal Ribonucleic Acid
RT-PCR	Reverse Transcriptase Polymerase Chain Reaction
SOM	Self-Organized Map
SVM	Support Vector Machine
TP	True Positive
TRN	Transcriptional Regulatory Network
tRNA	Transfer Ribonucleic Acid
Y2H	Yeast-Two-Hybrid

1 Introduction

Traditional molecular biology research usually focuses on the study of one or a few genes or proteins. Although an impressive number of biological principles have been discovered, these "from the bottom up" approaches are limited in providing us a full picture of cells, tissues and organisms. All biological organisms exhibit systems properties, which arises from many parts interacting as a whole. Because of the potential complexity of a biological system, the interactions among different components must be studied together to understand a whole biological organism. Systems biology aims at improving the understanding of how the different biological components (i.e., DNA, RNA, proteins and metabolites) combine to make the organism viable "from the top down" [1]. A gene is therefore no longer studied as an isolated entity but as part of a complex network or system. The network of cellular components and their interactions mediates signal transduction by translating dynamically changing environmental cues into the observed behavior. Modeling the dynamic action of these networks to predict cellular behavior is the ultimate goal of systems biology research [2, 3].

The advent of high throughput technologies has allowed large-scale identification of cellular entities, their active patterns, and their biochemical and genetic interactions. These biotechnologies include microarrays [4], ChIP-chip experiments [5], yeast-two-hybrid (Y2H) [6] and protein chips [7], which have enabled biologists to experimentally detect gene expression levels, protein-DNA interactions (PDIs), protein-protein interactions (PPIs) at the whole genome level for many organisms [8, 9]. These generated genome-wide data are called "omics" data. Top-down systems biology is increasingly relying on the integration of heterogeneous omics Utilizing distinct yet complementary data sources instead of a single data set to understand the spatio-temporal interactions within cells yields several advantages. First, different omics data (e.g., genomics, transcriptomics, proteomics, interactiomics, and metabolomics) unveil distinct aspects of cell activities; integrating these data leads to a more complete insight into these networks. Second, experimental and biological noise in individual data sources could lead to limited predictive power. The integration of multiple data sources provides an effective means to deal with the high noise level in individual data souse by lowering the false discovery rate (FDR)

and increasing the sensitivity [10]. Thus, molecular biology have been extended to genome-wide analyses, leading to a "paradigm shift": research has evolved from studying a few genes into a system thinking in which high throughput data are analyzed in an integrative way.

Modularity has been determined as the design principle of biological systems [11]. The omics data usually contain thousands of variables (i.e., semi-global or global gene expression data) and limited sample sizes (i.e., several to hundred samples in one experiment), suffering from curse of dimensionality [12]. Dissecting the cellular activities at modular level instead of gene level can capture the "high-level" regulatory relationships in cells. In this dissertation, we will focus on module-based approaches to integrate multi-source biological data for transcriptional regulatory network (TRN) inference and biomarker discovery in diseases.

1.1 Basic Molecular Biology

The roots of bioinformatics are in molecular biology – developing computational techniques to analyze the structure and function of molecules based on their chemical composition (sequence). Nowadays bioinformatics focuses on broader questions such as deciphering the interplay among molecules and the evolution of biological systems, yet the basic objects under study are mostly molecular entities. In this section, we briefly describe these entities and other biological terms mentioned throughout the dissertation.

DNA

The DNA molecule is the blueprint of life – a material found in every cell of each organism, which codes for the set of functional units and their relationships within the organism. It consists of two strands which are base-paired to form a double helix. The replication of DNA is carried out by DNA polymerases producing two daughter double strands from a parent double strand according to the base-pair complementarity. Since the daughter double strands are identical to the parent double strand, the information is preserved. DNA can be stained with certain dyes. This process can be used to visualize chromosome structure. The characteristic bands produced are called cytobands and can be used to identify chromosomal abnormalities such as translocations, where parts of chromosomes have been rearranged.

RNA

RNA is produced from a DNA template by RNA polymerases during transcription. The base-pair complementarity ensures the conservation of information. RNA is generally found as a single stranded molecule. By base-pairing, it can exhibit higher order structures such as stem-loops, hairpins or pseudo-knots which can act as recognition signals for the binding of proteins. Most of the RNA in a cell is bound to proteins, forming ribonucleoproteins. The ribosome is responsible for translation of RNA into proteins. There are three types of RNA:

- messenger RNA (mRNA) encodes proteins. It serves as a messenger for information from the DNA to the protein level.
- transfer RNA (tRNA) assigns the codons in the mRNA to amino acid residues forming the protein. Therefore tRNA acts as a translator between the nucleic acid and the proteins in the cell.
- ribosomal RNA (rRNA) and over 50 ribosomal proteins constitute the ribosome, the location of translation.

Protein

Proteins are the 'workhorses' of a cell. Proteins are involved in synthesis, repair and replication of DNA, control of ion flow across the membrane, catalysis of reactions, signaling pathways in the cell, and immunoreactions of the body against invaders. To achieve all this, they show a remarkable richness in their structure. Although all polypeptides consist of a simple linear chain of amino acids linked by peptide bindings, proteins exhibit a variety of higher order structures that can be described in four hierarchical levels:

1) The primary structure is the linear sequence of amino acids that constitute the polypeptide chain determined by protein sequencing.
2) The secondary structure is the local spatial arrangement of amino acid residues. The following secondary structures are commonly found in nature: α helix, β sheets, turns and random coils.
3) The tertiary structure of a protein describes the three dimensional conformation of the amino acids residuals. It can be experimentally studied by X-ray crystallography or by nuclear magnetic resonance (NMR). Single proteins further bind to form multimeric proteins.

4) The spatial arrangement of multimeric proteins specifies the quaternary structure. Even larger complexes could be formed by proteins constituting `macromolecular' assemblies.

Most higher order structures are based on non-covalent bindings between residues and may be stabilized by interaction with the local environment. The three dimensional structure of a protein specifies its functions.

Central Dogma of Molecular Biology

The flow of information from DNA to RNA to proteins is called the central dogma of molecular biology (Figure 1.1). DNA is the primary source of information which codes for the set of functional units and their relationships within the organism. This information is encoded in nucleotide triplets called codons. DNA is transcribed into RNA, which is then translated into proteins. The overall process of transcription and subsequent translation is referred to as gene expression. Although the information flows strictly from nucleic acids to proteins, the relation between DNA, RNA and protein is circular as proteins carry out, or at least support, the synthesis of DNA and RNA.

A gene (subsequence of DNA that codes for a functional unit within the cell) is considered active or expressed when it has been transcribed, and a protein is considered as expressed when it has been synthesized. A protein may be inhibited or activated by chemical modifications or the binding of other molecules such as proteins or metabolites.

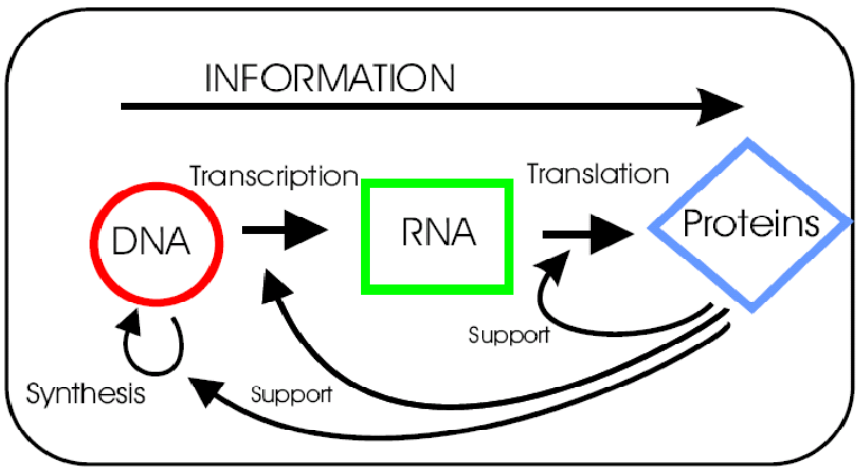


Figure 1.1 Central dogma of molecular biology.

1.2 Biological Data Sources

In this section we describe the integration of multiple biological data sources. This will help us better understand how to incorporate these different types of "omics" data in order to bring us a complete picture of cell activities, therefore inferring TRNs and identifying disease biomarkers.

1.2.1 Gene Expression Data

Gene expression data is divided into two levels: mRNA level and protein level. In this dissertation, we focus on the gene expression data on mRNA level, including cDNA microarrays, oligonucleotide chips, and RT-PCR.

cDNA microarrays

Developed at Stanford University, cDNA microarrays are glass slides on which cDNA has been deposited by high-speed robotic printing. They are ideally suited for expression analysis of up to 10,000 cDNA clones per array from expressed sequence tag sequencing projects (e.g., the private effort at Incyte Pharmaceuticals and the public Washington University project).

cDNA microarrays use small glass slides on which pre-synthesized DNA is spotted (Figure 1.2). Each spot represents one gene. Spotted arrays use two samples: a reference and a test sample, for instance normal versus malignant tissue. These samples are labeled using distinct fluorescent molecules, Cy3 (green) and Cy5 (red), and simultaneously hybridized to the microarray. Relative amounts of a particular gene transcript in the two samples are determined by measuring the signal intensities detected at both fluorescence wavelengths and calculating the ratios of the two signal intensities.

Oligonucleotide chips

The oligonucleotide chips, most widely produced by Affymetrix, consist of small glass plates with thousands of short synthetic oligonucleotide probes attached to their surface. For example on Affymetrix arrays, the oligonucleotides are synthesized directly onto the surface using a combination of semiconductor-based photolithography and light-directed chemical synthesis. Due to the combinatorial nature of the process, very large numbers of mRNAs are probed at the same time. However, manufacturing and reading of the chips requires expensive equipment. Current chips have over 65,000 different probes, with typically several probes for each mRNA. In contrast to cDNA arrays, each chip measures one biological sample or condition (Figure 1.2).

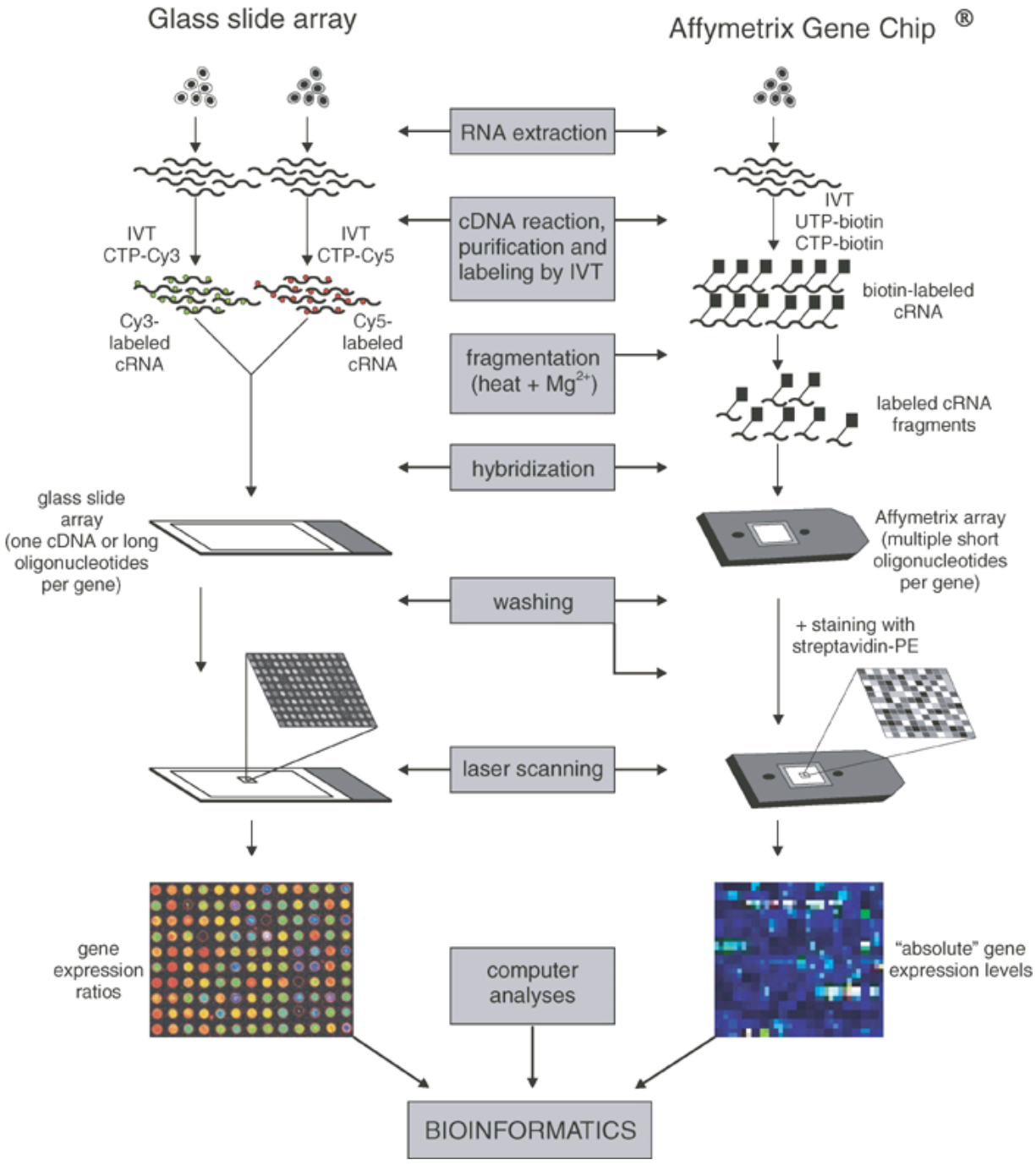


Figure 1.2 cDNA microarrays versus Oligonucleotide chips. The main difference between the two types of microarrays, is that two biological samples are hybridized on the spotted arrays, while one sample is hybridized on Affymetrix arrays (Figure taken from Staal et al. [13]).

RT-PCR

To measure gene expression using RT-PCR (Reverse Transcriptase Polymerase Chain Reaction), the mRNA is first reverse-transcribed into cDNA, and the cDNA is then amplified to measurable levels using PCR. Using built-in calibration techniques, RT-PCR achieves high accuracy coupled with an exceptional sensitivity of 10 molecules/10μl assay volume and a dynamic range covering 6-8 orders of magnitude. The method does require PCR primers for all the genes of interest, and is not inherently parallel like the previous three, so automation is crucial to scale up.

This method has been used to measure the expression levels of hundreds of genes. For example in [14], RT-PCR was used to measure the expression levels of 112 genes at nine different time points during the development of rat cervical spinal cord, and 70 genes during development and following injury of the hippocampus.

1.2.2 Protein-protein Interaction Data

Protein-protien interactions are essential for a wide range of cellular processes and form a network of astonishing complexity. Until recently, our knowledge of this complex network was rather limited. The emergence of large-scale PPI maps has given us new possibilities to systematically survey and study the underlying biological system. First attempts to collect PPIs on a large scale were initiated for model organisms such as *S.cerevisiae*, *D.melanogaster* and *C.elegans* [8, 9, 15-17]. Evidently, the generated interaction maps offered a rich resource for systematic studies of molecular networks.

After these initial efforts, the focus moved towards deciphering the human *interactome*. Recently, the first large-scale human PPI networks have been constructed using alternative strategies. Human interaction maps can be divided into three classes: (i) maps obtained from literature search [18-20], (ii) maps derived from interactions between orthologous proteins in other organisms [21-23] and (iii) maps based on large scans using Y2H assays [24, 25]. All of these different mapping strategies have their advantages as well as disadvantages. For example, Y2H based mapping approaches offer rapid screens between thousands of proteins, but might be compromised by large false-positive rates. The extent, however, how much the resulting interaction maps are influenced by the choice of mapping strategy, is not clear. Thus, it is important to critically assess and compare quality and reliability of produced maps.

Protein-protein interaction networks are commonly represented in a graph format, with vertices corresponding to proteins and edges corresponding to PPIs. An example of a PPI network constructed in this way is presented in Figure 1.3 [26]. The network consists of many small subnets (groups of proteins that interact with each other but not interact with any other protein) and one large connected subnet comprising more than half of all interacting proteins. The volume of experimental data on PPIs is rapidly increasing thanks to high-throughput techniques, which are able to produce large batches of PPIs. For example, yeast contains over 5000 proteins, and currently about 18000 PPIs have been identified between the yeast proteins, with hundreds of labs around the world adding to this list constantly [27]. The analogous networks for mammals are expected to be much larger. For example, humans are expected to have around 12000 proteins and about 10^6 interactions.

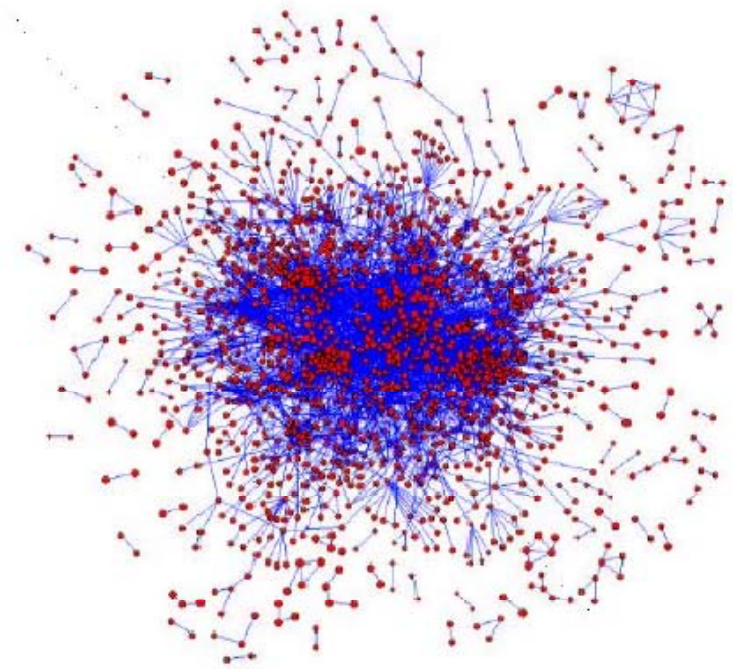


Figure 1.3 A PPI network constructed on 11000 yeast interactions involving 2401 proteins (Figures obtained from Przulj et al. [26]). The network consists of many small subnets (groups of proteins that interact with each other but not interact with any other protein) and one large subnet comprising more than half of all interacting proteins.

1.2.3 Protein-DNA Interaction Data

Three different sources of PDI information are available: (i) experimental data from genome wide location analysis (GWLA); ii) curated binding information in the TRANSFAC database [28]; and (iii) putative binding sites based on computational prediction algorithms. As an *in vivo* study, GWLA (i) is biologically most significant, but provides only the roughest (binary) information about possible binding sites and no precise location information. Database information (ii), on the other hand, presents a compilation of mostly *in vitro* studies and provides more accurate information, but at the expense of only small coverage of all intergenic regions. The third method is based on *in silico* predictions and provides the most detailed information on DNA-binding site locations, but the highest rate of false positives.

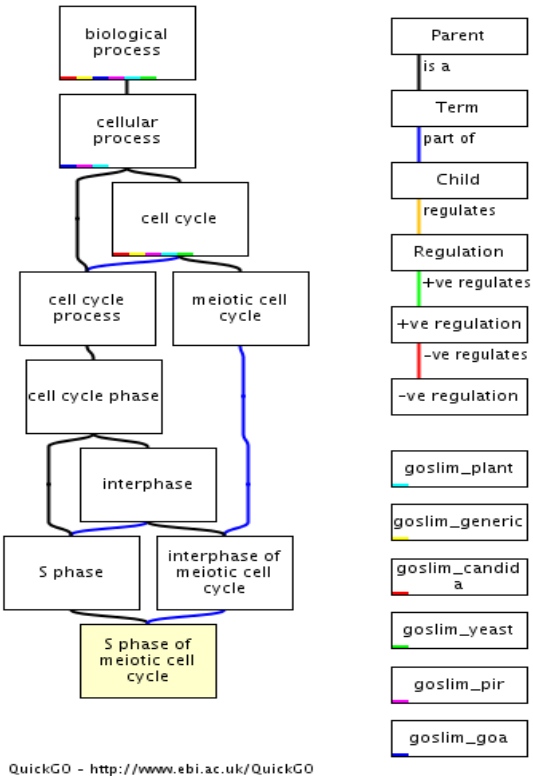


Figure 1.4 The directed acyclic graph induced from the GO term S phase of meiotic cell cycle (GO: 0051332), wherein at the bottommost level is the GO term of interest itself, and at the upper levels are all its ancestors, adapted from QuickGO Go Browser (http://www.ebi.ac.uk/ego/).

1.2.4 Gene Ontology

The gene ontology (GO) Consortium [29] has developed three separate ontologies - molecular function, biological process and cellular component—to describe the attributes of gene products, where molecular function defines what a gene product does at the biochemical level without specifying where or when the event actually occurs or its broader context; biological process describes the contribution of a gene product to a biological objective; and cellular component refers to where in the cell a gene product functions. Each GO is structured as a directed acyclic graph, wherein each term is a child of one or multiple parents, and child terms are instances or components of parent terms. For example, in Figure 1.4, the term S phase of meiotic cell cycle (GO: 0051332) is an instance of the term S phase (GO: 0051320) as well as an instance of the term interphase of meiotic cell cycle (GO: 0051328).

1.2.5 Complementarity of the data

Genome-wide expression data sets have typically been used in cluster analysis to detect sets of genes with a similar expression profile. Although these co-expressed genes are potentially co-regulated, co-expressed genes could also have different regulatory programs.

Protein-protein interaction and protein-DNA interaction data provide another way of studying regulation: physical interactions between two proteins or a transcription factor and DNA regions are identified. However, one should keep in mind that such interactions in a specific condition do not necessarily imply that regulatory relationships exist in this condition. In the case of combinatorial control, for instance, an additional regulator might be required. Alternatively the presence of a particular ligand or metabolite may be needed for the activation of the regulator. For this reason, ChIP-chip data contain false-positives. On the other hand, some interactions of a regulator with its target genes only occur in very specific conditions that may not (yet) be present in the ChIP-chip compendium.

Gene ontology information represents the gene annotations from already validated biological evidences [29]. Three GO categories (biological process, molecular function, and cellular component) can be used as a basis to determine/evaluate the inferred relationships from other high-throughput biological data sources.

All these approaches are useful for studying the transcriptional network but their individual power is limited because they only provide partial information on the network: expression data only provide indirect evidence for regulation, interaction data indicate regulator binding but the binding might not be functional, and GO information only correspond to known biological evidences. However, these data types offer complementary information useful for TRN inference and disease biomarker discovery. The goal of this dissertation is therefore to integrate them and thus exploit as much available information as possible to obtain a more comprehensive view on the TRN and discovery of disease biomarker.

1.3 Systems Biology in Bioinformatics

Systems biology comprises the global and integrative analysis of multi-source information encoding different levels of biological information. It is an integrative research strategy that aims to: i) tackle the complexity of biological systems and their behaviors at different levels of organization; ii) understand the biological phenomena at "systems level", which cannot be elucidated by the functions of individual components defining the system; iii) integrate multi-source information generated by different high-throughput experimental techniques; and iv) combine data-driven (top-down) and model-driven (bottom-up) approaches into a question driven approach capable to discover basic principles.

The origin of systems biology can be traced back to many years of research in physiology, biochemistry, and molecular and cellular biology. However, its current development is in bioinformatics field, where high-throughput experimental technologies and computational tools have been continuously developed to generate, analyze and integrate different types of 'omics' data. In this context, systems biology can be defined as an integrative research: "integration of 'omic' data and knowledge resources, integration of different levels of biological complexity, integration of computational and experimental technologies, and integration of scientific disciplines" [30]. Over the last decade, researchers have applied systems biology based approaches to gain deeper understandings of biological functions and properties in different model organisms. Transcriptional regulatory network is so complex that its activities cannot be reflected solely by gene expression data. To address this challenge, other data sources such as

regulatory sequence motif information have been integrated to discover the co-expressed gene modules [31].

1.4 Modular Architecture of Biological Network

Regulatory relationships in one cell are represented in the form of a network (e.g., TRN) defined as a graph: the nodes represent the genes, a directed edge from one node to another indicates that the first gene codes for a transcription factor that regulates the second gene, and an undirected edge between two nodes represent their interactions at protein level, which further directs them to regulate common downstream genes. The architecture of biological network is described by means of graph feature such as distance, diameter or degree. We briefly introduce these concepts, followed by a description of network analysis at different network levels.

The shortest path between two nodes is called the distance. The graph diameter is the maximum distance between any two nodes in the network. If this diameter is small, the network is called a small world network. Many interaction networks, for instance the TRN, are known to exhibit this small world property [32].

The degree of a node is defined as the number of edges that connect to this node. The number of incoming edges is called the in-degree, while the number of outgoing edges is called the out-degree. If a node has a high degree, this means that this node is connected to many other nodes in the network. Interaction networks are not randomly organized but seem to have a scale-free architecture with the typical power law degree distribution: a limited number of nodes have a high degree while most nodes have a small degree. This means that there are some hubs present in the network: these are nodes connected to many other nodes. The PPI network, TRN network and metabolic network are all scale-free networks. The advantage of this kind of organization is that the loss of one nonhub link is not as disruptive in scale-free networks as in random networks. In other words, scale-free networks are generally more robust. The hubs are of course extremely important and usually they play essential roles in biological systems [33].

In addition to being scale-free, Shen-orr et al. [34] discovered the presence of network motifs (NMs) in the TRN of *E. coli*. Network motifs are topologically distinct regulatory interaction patterns that are present more frequently in true interaction networks than in random networks. Therefore these NMs must have a specific biological function: they are postulated to be the basic

signal transduction elements, each with their own characteristic properties. Examples of NMs are the single input, multiple input and feed-forward loop motifs (Figure 1.5).

Previous studies have unveiled the modular and hierarchical organization of interaction networks [35-39]. Indeed, biological processes consist of pathways that mainly act on their own although communication exists between these pathways. Therefore one might expect that the distinct biological processes are organized in discrete and separable modules (Figure 1.5). A module in a network can be defined in various ways [40]:

1) One popular definition of a module involves co-expressed genes, with or without environmental context dependence and assigning a regulatory motif or regulator to these genes. This definition of a module will refer to a gene module in the remainder of this dissertation.

2) Another definition of topological module is by means of graph-based techniques. NMs are one type of such modules, called gene regulatory modules in the remainder of this dissertation.

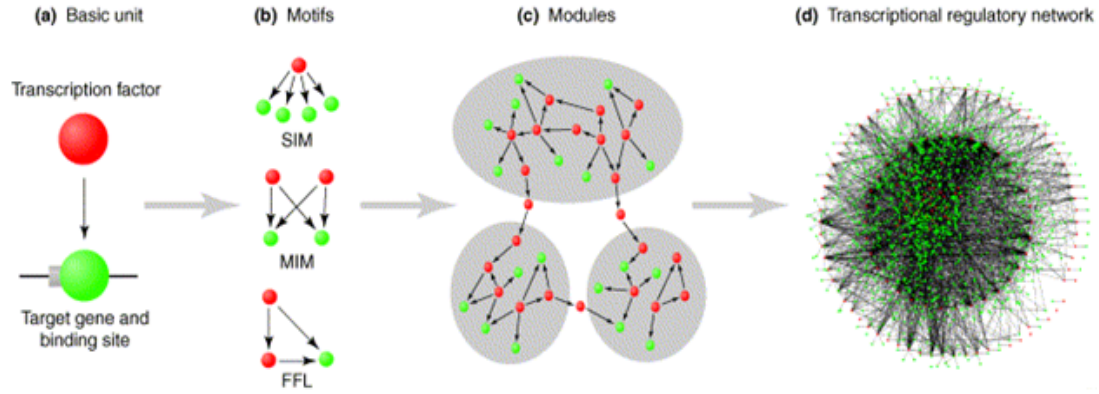


Figure 1.5 The different layers in the transcriptional regulatory network. The basic unit of the regulatory network consists of a transcription factor that regulates a target gene (a). These basic units are organized into NMs (b). Three examples are shown: a single input motif, a multiple input motif and a feed-forward loop. The transcription factors are indicated in red, the target genes in green. The NMs are further organized in modules (c). The modules themselves are also linked by hubs. Finally the complete TRN is shown (d). (Figure taken from Babu et al. [41].)

1.5 Reconstruction of Transcriptional Regulatory Network

Revealing the complete interaction network underlying the system's behavior is the ultimate goal of systems biology. In this dissertation, we focus on the TRN at modular level. The purpose

of TRN inference is to determine for all transcription factors the regulatory motifs they the conditions in which they are active, the regulators they cooperate within these conditions, and their target genes in these conditions. By using a top-down data-driven approach we aim at reconstructing the transcriptional network on a large scale.

Top-down inference can be contrasted with bottom-up inference [10, 42]:

- *Bottom-up* inference starts from a comprehensive expert model of known interactions between molecular entities as described in literature and curated databases. Such models are used to simulate cellular behavior or to predict the outcome of a perturbation experiment. Inconsistencies between observed data and simulations point at deficiencies in the current network structure and outline hypotheses of novel interactions that better explain the observations.
- *Top-down* network inference methodologies start from a global view of the behavior of the system by using high-throughput data. This kind of inference method does not rely on expert knowledge on the relationships between the molecular components. Top-down inference is data-driven, thus it is data-demanding. Given the current data availability, top-down network inference problems are often underdetermined (i.e. the network that is reconstructed from the data is not unique, because many equally likely solutions can explain the observations). However, the top-down inference is made increasingly tractable by integrating data from different sources, and holds great promise for future bioinformatics research.

One possible solution to the above underdetermined problem in top-down network inference is to integrate the multi-complementary high-throughput datasets. For instance, transcriptional regulation is a process that needs to be understood at multiple levels of description [43, 44] (Figure 1.6) including (1) the factor-target gene interaction, in which transcription factors activated under certain conditions interact with their conserved binding site sequences; and (2) transcriptional regulation, which explains how the bindings of transcription factors to their unique recognition sites regulate the expression of specific genes. A single source of information such as gene expression data is aimed at only one level of description (transcriptional regulation level), thus it is limited in its ability to obtain a full understanding of the entire regulatory Other types of information such as PPI [6, 7] and PDI [5] data provide complementary

constraints on the models of regulatory processes. By integrating limited but complementary data sources, we realize a mutually consistent hypothesis bearing stronger similarity to the underlying causal structures [44]. Among the various types of high-throughput biological data available nowadays, time course gene expression profiles and genome-wide location analysis data are two complementary sets of information that is used to infer regulatory components. Time course gene expression data are advantageous over typical static expression profiles as time can be used to disambiguate causal interactions. Genome-wide location analysis data, on the other hand, provide high-throughput quantitative information about *in vivo* binding of transcription factors to the target regulatory regions of the DNA. Prior biological knowledge generated by geneticists will also help guide inference from the above data sets and integration of multiple data sources offers insights into the cellular system at different levels.

Another way to reduce the complexity of the TRN inference problem is to decompose it into small units of commonly used network structures, call gene regulatory modules. TRNs are made of repeated occurrences of simple patterns – NMs. Since the establishment of the first NM in *Escherichia coli* [34], similar NMs have also been found in eukaryotes including yeast [45], plants, and animals [46-48], suggesting that the general structure of NMs are evolutionarily conserved. One well known family of NMs is the feed-forward loop [49], which appears in hundreds of gene systems in *E. coli* [34, 50] and yeast [45, 51], as well as in other organisms [46-48, 52-54]. A comprehensive review on NM theory and experimental approaches is available in [55]. Knowledge of the NMs to which a given transcription factor belongs facilitates the identification of downstream target gene clusters. In yeast, a genome-wide location analysis was carried out for 106 transcription factors and five NMs were considered significant: auto regulatory motif, feed-forward loop, single input module, multi-input module and regulator cascade [45]. The NMs will be considered as candidate gene regulatory modules in the remainder of the dissertation.

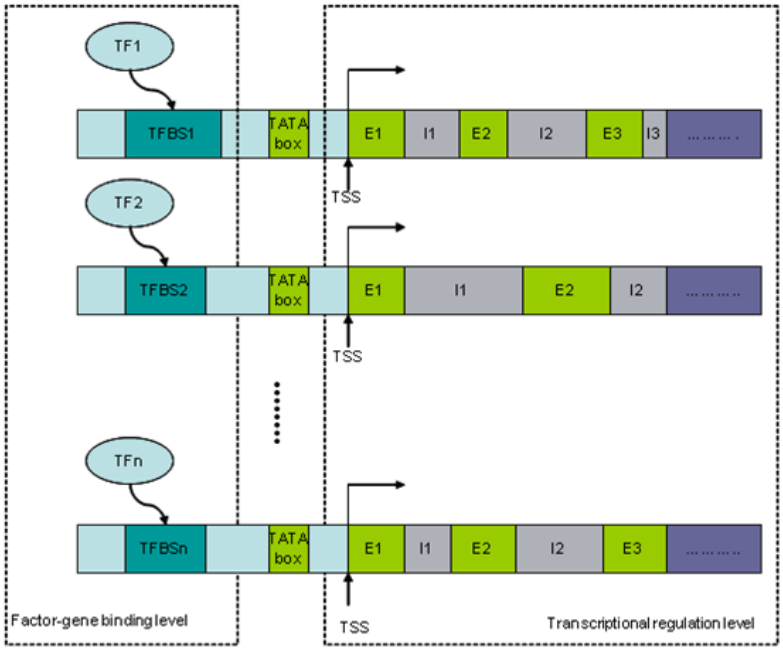


Figure 1.6 The gene transcriptional regulatory program. The gene transcriptional regulatory program is simplified in two levels. At the factor-gene binding level, the "activated" transcription factors bind to their specific conserved sequence motifs, called transcription factor binding sites. When the binding process is completed, the regulation mechanism instructs the gene transcription from transcriptional start site (DNA to mRNA); first part of the central dogma in molecular biology. (Figure taken from Zhang et al. [56].)

1.6 Problem Statement

Understanding transcriptional regulation is a leading problem in contemporary biology. The availability of multi-source biological data help elucidating transcriptional regulatory mechanisms. In this dissertation, we develop several integration approaches to help understand biological problems at modular level.

1.6.1 Gene Module Identification

Many clustering algorithms have been suggested to analyze high throughput gene expression data. Clustering methods generally aim to identify subsets (clusters) in the data based on the similarity between single objects. Similar objects should be assigned to the same cluster, while objects which are not similar to each other, should be assigned to different clusters.

Different cluster algorithms have been applied to the analysis of expression data: hierarchical clustering (HC), k-means clustering (KCM), and self-organized maps (SOM) to name just a few.

Choosing the most appropriate clustering algorithm with appropriate parameters for a given dataset is not straightforward. Different clustering algorithms or different runs of the same algorithm may generate different partitions for the same dataset. The prediction of the optimal number of clusters is a challenge in clustering (unsupervised classification) problems. Many methods for estimating the number of clusters have been proposed [57]. Basically, cluster validation techniques are divided into two main categories: internal and external validation measures. Internal validation measures use information from within the given data set to represent the goodness of fit between the input data set and clustering results. However, clusters are validated using the same information from which they are derived, and the values of gene expression data include various types of intrinsic and extrinsic experimental noise. Also, in gene expression data, genes are often highly correlated with more than one cluster. These problems make such internal validation measures difficult. On the other hand, external validation measures evaluate a clustering result by comparing it to a given "golden standard" (e.g. GO) which is another partition of the objects. An underlying assumption of cluster analysis is the "guilt by association" rule, i.e. genes with similar expression patterns are more likely to have similar biological function. The best clustering method for a particular data set is the one that has the strongest tendency to bring genes of similar function together when applied to diverse expression data sets. External validation measures have the strong benefit of providing an independent assessment.

In this dissertation, we propose to compare different gene-based clustering algorithms by a figure of merit based on Kullback-Leibler (KL) divergence between cluster membership and known GO attributes. In particular, we will focus on two questions in the following:

i) What choice of cluster number generally yields the most information about gene function (where function is known)?

ii) Which clustering parameters (e.g., distance measure) generally yields the most information about gene function (where function is known)?

1.6.2 Gene Regulatory Module Inference

Enormous amount of biological data has been generated by the use of high-throughput analytical methods in biology during the last two decades. However, reverse engineering of a global TRN remains challenging because of the underdetermined problem.

One approach to address the curse of dimensionality is to integrate multiple large data sets with prior biological knowledge. This approach offers a solution to tackle the challenging task of inferring TRN. A number of researches have explored the integration of multiple data sources (e.g., time course expression data and molecular interaction data) for TRN inference [58-61]. A typical approach for exploiting two or more data sources uses one type of data to validate the results generated independently from the other (i.e., without data fusion). Computational methodologies that allow systematic integration of data from multiple resources are needed to fully use the complementary information available in those resources.

Another way to reduce the complexity of the TRN inference problem is to decompose it into simple units of commonly used network structures. Transcriptional regulatory network is a network of interactions between transcription factors and the genes they regulate, governing many of the biological activities in cells. Cellular networks like TRNs are composed of many small but functional modules or units [62]. Breaking down TRNs into these functional modules will help understanding regulatory behaviors of the whole networks and study their properties and functions.

In this dissertation, we tackle the TRN inference problem by combining the two aspects above into one computational framework. This framework focuses on two aspects in TRN inference:

1) Integrating data from different sources will reduce the amount of false positive observations, as the independent observations from various data sources contribute to the overall probability of a relationship between genes being true.
2) Inference of gene regulatory modules instead of the whole network significantly reduces the computational complexity, which is the main problem in TRN inference.

1.6.3 Module-based Biomarker Discovery

Biomarkers are needed for diagnosis and early detection of complex diseases such as cancer. In recent years, an increasing number of disease biomarkers have been identified through analysis of genome-wide gene expression profiles. Gene biomarkers for classification of cancer patients are typically identified by scoring or ranking individual genes with regard to their capacity to distinguish between clinically relevant classes. However, the reliability and reproducibility of these gene biomarkers have been challenged because of the biological heterogeneity and noise within and across patients. These biomarkers are mainly "downstream" reflector of the perturbations defining clinical outcomes through the complex interplay of interaction networks. They may not directly account for the activity, perturbations or roles that disease-related cellular networks show.

The availability of genome-wide interaction network data opens up new possibilities to discover potential biomarkers and elucidate cancer-related complex mechanisms at network level. Researchers have explored the idea of incorporating known pathway knowledge into the identification of genes and subnetworks that are related to disease [63-65]. The availability of large PPI, PDI and pathway data in public databases enables new opportunities for elucidating pathways involved in major diseases and pathologies [66]. However, these types of networks are typically analyzed separately [67-69], which hides the full complexity of the cellular circuitry since many processes involve combinations of these two types of interactions.

In this dissertation, we propose a module-based biomarker discovery approach by integrating gene expression data and the combined molecular interaction network to identify modules that are related to diseases [63-65]. Compared to conventional gene-based approach, the module biomarkers identified by this proposed approach are expected to have the following properties:

1) They are more reliable and robust within/across patient samples;
2) They achieve comparable prediction accuracies in independent studies;
3) They are more enriched in disease "driver" genes.

1.7 Summary of Contributions

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

- Knowledge-based approach to compare expression-based gene clustering algorithms. A figure of merit, based on the Kullback-Leibler (KL) divergence between cluster membership and known gene attributes, is used to select the optimal cluster number with biological significance. The optimal clusters are considered as gene modules - co-regulated genes by common transcription factors in the TRN. (Chapter 2)
- Reconstruction of a TRN by inferring gene regulatory modules. Our method integrates multiple data sources to facilitate the TRN inference task. By breaking down the TRN into gene regulatory modules, we not only address the computational challenges of reverse engineering TRNs but also we create modules for better understanding of complex interaction networks. The gene regulatory modules are used to generate new hypotheses that can be tested through actual laboratory experimentation and can be assembled together to construct a TRN. (Chapter 3)
- Module-based biomarker discovery approach to identify biomarkers in breast cancer research. We have demonstrated that the module biomarkers are more reliable reproducible than individual marker genes selected without interaction network information, and that they achieve higher accuracy in the classification of different cancer patient groups. (Chapter 4)

1.8 List of Relevant Publications

1.8.1 Journal Papers

[1] **Y. Zhang**, J. Xuan, B. G. de los Reyes, R. Clarke, and H. W. Ressom, "Module-based biomarker discovery for breast cancer," In preparation for *Bioinformatics*.

[2] **Y. Zhang**, J. Xuan, B. G. de los Reyes, R. Clarke, and H. W. Ressom, " Inferring transcription factor-target gene relationships in cancer cell cycle from multiple-source biological data," ***PLoS ONE,*** vol. 5: e10268, 2010.

[3] **Y. Zhang**, J. Xuan, B. G. de los Reyes, R. Clarke, and H. W. Ressom, "Reverse engineering module networks by PSO-RNN hybrid modeling," ***BMC Genomics,*** vol. 10, pp. S15, 2009.

[4] **Y. Zhang**, J. Xuan, B. G. de los Reyes, R. Clarke, and H. W. Ressom, "Network motif-based identification of transcription factor-target gene relationships by integrating multi-source biological data," ***BMC Bioinformatics***, vol. 9, pp. 203, 2008.

[5] **Y. Zhang,** J. Zhu, X. Sun and Z. Lu, "A method of oligonucleotide synthesis optimization," ***Biotechnolgy***, vol. 12, pp. 26-28, 2002. (in Chinese)

1.8.2 Conference Papers

[6] **Y. Zhang**, J. Xuan, B. G. de Los Reyes, R. Clarke, and H. W. Ressom, "Network motif-based identification of breast cancer susceptibility genes," in Proc. of ***Conf Proc IEEE Eng Med Biol Soc (EMBC)***, Vancouver, British Columbia, Canada, 2008:5696-5699.

[7] **Y. Zhang**, J. Xuan, B. G. de los Reyes, R. Clarke, and H. W. Ressom, "Reverse engineering module networks by PSO-RNN hybrid modeling," in Proc. of ***The 2008 International Conference on Bioinformatics & Computational Biology (BIOCOMP'08)***, Las Vegas, NV, 2008:401-407.

[8] H. W. Ressom, **Y. Zhang**, J. Xuan, Y. Wang, and R. Clarke, "Integrating Multi-Source Biological Data for Transcriptional Regulatory Module Discovery," in Proc. of ***Third IEEE/NIH Life Science Systems and Applications (LISA) Workshop***, Bethesda, MD, 2007.

[9] H. W. Ressom, **Y. Zhang**, J. Xuan, Y. Wang, and R. Clarke, "Inferring network interactions using recurrent neural networks and swarm intelligence," in Proc. of ***Conf Proc IEEE Eng Med Biol Soc (EMBC)***, New York City, New York, USA, 2006:4241-4244.

[10] H. W. Ressom, **Y. Zhang**, J. Xuan, J. Wang, and R. Clarke, "Inferring network interactions using recurrent neural networks and particle swarm optimization," in Proc. of ***Proceedings of the First International Conference on Computational Systems Biology***, Shanghai, China, 2006.

[11] H. W. Ressom, **Y. Zhang**, J. Xuan, J. Wang, and R. Clarke, "Inference of gene regulatory networks from time course gene expression data using neural networks and swarm intelligence," in Proc. of ***the 2006 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2006)***, Toronto, ON, Canada, 2006:435-442.

1.9 Outline of Dissertation

The main research topics discussed in this dissertation are the module-based analysis of biological data for network inference and biomarker discovery. We conduct our study at different levels of biological data integration based on the biological problems presented. Among them, gene module identification, gene regulatory module inference, and module biomarker discovery are studied in detail.

Chapter 2 addresses knowledge-based clustering evaluation of gene expression data. Existing expression-based clustering methods are reviewed first. The common clustering validation techniques are further discussed. Based on that, we propose a knowledge-based evaluation strategy to compare diverse clustering algorithms based on the KL divergence between cluster membership and known gene attributes. The proposed figure of merit is used to indicate the biological significance between different clustering partitions. We compare the results of various clustering algorithms with different parameter settings with the proposed method on three gene expression datasets. The optimal clustering partition is evaluated by binding site enrichment analysis (BSEA) when ground-truth is not available.

Chapter 3 addresses gene regulatory module inference by hybrid computational intelligence modeling. Existing TRN inference methods are reviewed first. A computational framework that integrates information from time course gene expression experiment, molecular interaction data, and GO category information is proposed to infer the relationship between transcription factors and their potential target genes at NM level. The identified gene regulatory modules are evaluated using the various biological validation strategies. The proposed method is applied on two gene expression datasets and compared with some existing benchmark methods. The experimental results are discussed in detail in terms of their biological functional relevance.

Chapter 4 addresses module-based biomarker discovery in cancer research. Existing methods on biomarker selection are reviewed first. Then, we present module-based biomarker discovery approach by integrating gene expression data and interaction network information. We derive the candidate modules from interaction network based on their discriminative power between two patient groups. A mutual information based discriminative score is proposed for module selection. Two statistical analyses are carried out to identify significant modules. Following that, an ensemble feature selection method based on ant colony optimization (ACO) is proposed to select the optimal module biomarker set. The proposed method is applied on four breast cancer datasets, including one in house dataset. The results show that the proposed method has outstanding feature selection capability and better reproducibility on independent dataset. The identified module biomarkers are enriched in disease “driver” genes compared to the gene selected through the traditional biomarker selection methods.

Chapter 5 summarizes the original contribution of the dissertation research, proposes several problems and tasks for the future work, and presents conclusion for the conducted research work.

2 Gene Module Identification by Knowledge-based Cluster Evaluation

Many clustering algorithms have been suggested to analyze high throughput gene expression data. Clustering methods generally aim to identify subsets (clusters) in the data based on the similarity between single objects. Similar objects should be assigned to the same cluster, while objects which are not similar to each other, should be assigned to different clusters.

Since different clustering algorithms or different runs of the same algorithm may generate different partitions for the same dataset, how to choose appropriate algorithm with appropriate parameters for one dataset becomes a critical challenge. An underlying assumption of cluster analysis is the “guilt by association” rule, i.e. genes with similar expression patterns are more likely to have similar biological function. With this goal in mind, we propose a figure of merit based on KL divergence between cluster membership and known GO attributes. Several benchmark expression-based gene clustering algorithms were compared using the proposed method with different parameter settings. Applications to diverse public available time course gene expression data demonstrated that fuzzy c-means (FCM) clustering is superior to other clustering methods with regard to the enrichment of clusters for biological functions. These results contribute to the evaluation of clustering outcomes and the estimations of optimal clustering partitions, which present an effective tool to support biomedical knowledge discovery in gene expression data analysis.

2.1 Introduction

DNA microarray technology has revolutionized the study of complex biological network by measuring the simultaneous expressions of thousands of genes across different experimental conditions. To reveal the underlying structures of these gene expression data, a central step of the analysis is the identification of groups of gene that exhibit similar expression patterns. Clustering gene expression data into homogeneous groups has been shown to be instrumental in functional annotation, tissue classification, regulatory motif identification, and other applications. This is based on the observation that genes showing similar expression patterns (i.e. co-expressed genes) are often functionally related and are controlled by the same regulatory mechanisms (i.e. co-regulated genes). Therefore, expression clusters are frequently enriched by genes of certain functions, e.g. DNA repair, or M phase of cell cycle. If a gene of unknown function falls into such a cluster, it is likely to serve the same functions as other members of the cluster. This "guilt-by-association" rule enables assigning possible functions to a large number of genes by clustering of co-expressed genes [70]. It is especially valuable for organisms and cell types where little previous knowledge about their biology exists.

Different cluster algorithms have been applied to the analysis of expression data: hierarchical clustering (HC), k-means clustering (KCM), and self-organized maps (SOM) to name just a few [61, 71, 72]. Assessing the clustering results and interpreting clusters found by various clustering methods are as important as generating the clusters [73]. Given the same data set, different clustering algorithms can potentially generate very different clusters. The prediction of the optimal number of clusters is a big challenge in clustering (unsupervised classification) To address this challenge, many methods for estimating the number of clusters have been proposed in the literature [57]. Halkidi et al. [57] divided cluster validation techniques into two main categories: internal and external validation measures. Internal validation measures use information from within the given data set to represent the goodness of fit between the input data set and clustering results. Compactness, connectedness and separation of clusters are possible measures of goodness of fit. Several enhanced approaches that combine the above different types of measurements have been proposed, including Dunn Index [74], Dunn-like Index [75], the Davies-Bouldin Index [76], and the Silhouette Width [77]. An alternative validation technique is to assess the predictive power or stability of a clustering partition. Yeung et al. proposed an

adjusted figure of merit to estimate the predictive power of a clustering algorithm [78]. Ben-Hur et al. repeatedly drew overlapping subsamples of the same dataset to assess the degree of stability of a partitioning [79]. One problem with such internal measures is that clusters are validated using the same information from which clusters are derived. However, the values of gene expression data include various types of intrinsic and extrinsic experimental noise [80, 81], which makes such internal validation measures even more difficult. Another problem with gene expression data is that genes are often highly correlated with more than one cluster [70]. This might be expected, since gene products frequently participate in more than one regulatory mechanism to different degrees. Additional prior biological knowledge may help evaluate the partitioning of various clustering algorithms.

Different from internal validation measures, external validation measures evaluate a clustering result by comparing it to a given "golden standard" which is another partition of the objects. The golden standard must be obtained by an independent process based on information other than the given data set. Gibbons and Roth proposed a figure of merit based on the mutual information between cluster membership and a set of filtered gene attributes from GO database [82]. Datta and Datta introduced two performance indices to quantify the performance of clustering results in grouping genes with similar biological functions compared with a reference collection of relevant functional classes [83]. Gat-Viks et al. projected the vectors of biological attributes of the clustered genes onto the real line and evaluated the quality of gene clusters with ANOVA test [84]. Loganantharaj et al. constructed a metric to evaluate the clustering performance by computing inter-cluster cohesiveness and the intra-cluster separation with respect to biological features [85].

In clustering of gene expression data, our goal is to bring genes of similar function together. The best clustering method is the one that has the strongest tendency to bring genes of similar function together when applied to diverse expression datasets [82]. Therefore, external validation measures have the strong benefit of providing an independent and hopeful assessment. Recently developed biological ontologies aim to annotate biological entities with a consistent, controlled and structured vocabulary. Gene ontology is one of the most widely used ontology databases seeking to capture information about the role of gene products within an organism. Although GO databases are necessarily incomplete and evolving, they represent the best computable summary

of our present state of knowledge. In this chapter, we propose a figure of merit based on KL divergence to investigate the relationship between groups of genes generated by data-driven clustering and their respective GO attributes. This external validation measure is applied for biological evaluation of clustering obtained from a set of clustering algorithms to facilitate gene annotation of high quality. In particular, we would like to tackle two problems in clustering:

i). What choice of cluster number generally yields the most information about gene function (where function is known)?

ii). Which distance measure generally yields the most information about gene function (where function is known)?

The organization of this chapter is as follows. Section 2.2 briefly reviews the four clustering algorithms used for performance comparison. Section 2.3 introduces the proposed method to quantify the similarity between gene clusters and their functional categories in GO databases. Section 2.4 will present the evaluation and comparison results on three real data sets: the rat central nervous system (CNS) data set [14], the yeast cell cycle data set [60], and the human *Hela* cell cycle data set [86]. Finally, Section 2.5 is devoted to the summary.

2.2 Review of Competing Clustering Methods

In this section, we review the clustering algorithms widely used to cluster genes. For simplicity, we use genes instead of data points to illustrate the algorithms.

2.2.1 Hierarchical Clustering

Hierarchical clustering is the first method used for clustering genes in gene expression data [87]. There are two approaches for hierarchical clustering: agglomerative approach which groups small clusters into larger ones, and divisive approach which splits big clusters into smaller ones. In this chapter, we focus on agglomerative hierarchical clustering. As a bottom-up approach, it starts by considering the n genes as n nodes in the sample set. At each iterative stage, a pair of nodes with the shortest distance between them will be merged to form a node. Thus a hierarchical tree is constructed after $(n-1)$ steps. The pseudocode of this algorithm is presented here:

1) Start with each gene in a cluster of its own;

2) Calculate the pair-wise distances between all clusters;

3) Find the closest pair of clusters and merge them;

4) If all genes are merged into one single cluster, go to Step 5; otherwise, go Step 3;

5) Return the dendrogram representing the merging process.

To define the distance between two clusters, several distance measures are available, including single linkage (shortest pair-wise distance between genes in two clusters), complete linkage (largest pair-wise distance), and average linkage (average pair-wise distance). Suppose we have two clusters $C1$ and $C2$, three distance measures are calculated by

$$D_{min}(C1, C2) = \min_{\substack{g_i \in C1 \\ g_j \in C2}} d(g_i, g_j) \tag{2.1}$$

$$D_{max}(C1, C2) = \max_{\substack{g_i \in C1 \\ g_j \in C2}} d(g_i, g_j) \tag{2.2}$$

$$D_{ave}(C1, C2) = \frac{\sum_{g_i \in C1} \sum_{g_j \in C2} d(g_i, g_j)}{\mathrm{N_{C1} N_{C2}}} \tag{2.3}$$

where d is a distance metric; $N_{C1}$and N_{C2} are the sizes of the two clusters $C1$ and $C2$, respectively. In the comparison experiment, we implemented hierarchical clustering with three distance measures in MATLAB 2009a version.

2.2.2 K-means Clustering

K-means clustering is a classical clustering method [88] also widely used in microarray data. The algorithm splits the data into K clusters optimizing a given objective function. If the data is given as a set of d dimensional vectors (e.g., $x_i, i = 1,2, \dots, n$), a common objective function is the square error function:

$$E = \sum_i \sum_j d(x_i, c_j)^2 \tag{2.4}$$

where d is the distance metric and c_j is the center of cluster j. The square error function E describes the within-cluster variation. Minimization of E results in clusters with a minimal sum of the distances between the data vectors x and the cluster centers c. The total variation of the data is split into within- and between-cluster variation. Since the total variation is fixed, minimizing the within-cluster variation leads to maximizing the between-cluster variation.

A popular approach for minimizing the square error function E in k-means clustering is to iteratively partition the data using following algorithm [73]:

1) Initiation: Choose K random vectors as cluster centers $\overline{x_j}$;
2) Partitioning: Assign x_i to $\overline{x_j}$ if $d(x_i, \overline{x_j}) < d(x_i, \overline{x_k})$ for all k with $k \neq j$;
3) Calculation of cluster centers $\overline{x_j}$ based on the partition derived in step 2: The cluster center $\overline{x_j}$ is defined as the mean value of all vectors within the cluster;
4) Calculation of the square error function E;
5) If the chosen stop criterion is met, stop; otherwise continue with step 2.

The stop criterion is set by a maximal number T_{max} of iterations or by a minimal threshold ε for decrease of E. Alternatively, k-means clustering is stopped if no data object is reassigned to another cluster in consecutive iterations. For the distance metric D, several distance measures can be chosen. In the comparison experiment, we investigated the performance of k-means clustering with two distance measures including Euclidean distance and Pearson correlation.

2.2.3 Self-organizing Maps

Self-organizing maps (SOM) [89, 90] has been applied in many microarray analysis. It first maps K nodes in a low-dimensional grid space (usually two-dimensional) from the d-dimensional space that the data set is situated and the nodes are adjusted iteratively. Each time, a node in the data is randomly selected. Other nodes closest to the chosen one are identified. These nodes are adjusted to look similar to the chosen point depending on their distance to the chosen point and the two-dimensional geometry of the nodes. This process is repeated until the nodes converge and serves as cluster centers to form clustering. Clusters generated from nodes close to each other in the two-dimensional grid geometry will have similar expression patterns. Tamayo et al. pointed out that this clustering structure is favorable for interpretation [90]. No recommendation, however, was given on how to select the initial structure of the SOM. A similar clustering approach based on SOMs was presented by Toronen et al. analyzing yeast gene expression data [72].

2.2.4 Fuzzy c-means Clustering

The above three clustering methods are called hard clustering, assigning data objects to exactly one cluster. The underlying assumption for this strict assignment is that clusters are well separated. In many situations, however, this might not be the case. Clusters may be overlapping with data objects between clusters sharing attributes of several clusters. For instance in gene expression data, it has been pointed out that genes were often highly correlated with the patterns of more than one cluster [70]. This might be expected, since gene products frequently participate in more than one regulatory mechanism to different degrees. The regulation of a gene is generally not an "on-off", but gradual manner which allows a finer control of the gene's functions.

To accommodate this situation, fuzzy clustering generalizes the partitioning of hard clustering. In contrast to hard clustering, a data object can be member of several clusters. The membership values μ_{ij} takes any value between zero and one. This results in fuzzy partitions that take the form of

$$M_{fc} = \left\{ U_{ij} \in R^{c*N} \middle| \begin{matrix} \mu_{ij} \in [0,1] & \forall i,j \\ \sum_{i=1}^{C} \mu_{ij} = 1 & \forall j \\ 0 < \sum_{j=1}^{N} \mu_{ij} < N & \forall i \end{matrix} \right\} \tag{2.5}$$

where c is the number of partitions, and N is the number of data points. Hard k-means clustering can be seen as a special case of fuzzy clustering where the membership values are either zero or one.

An important objective function for fuzzy clustering is the c-means function J_m which is similar to the square error function E for k-means clustering. It weights, however, the distances of the data vector x_i to the cluster center c_j according to the membership values of x_i:

$$J_m = \sum_i \sum_j (\mu_{ij})^m \left\| x_i - c_j \right\|_D^2 \tag{2.6}$$

where m is the fuzzification parameter ($m > 1$) and D is a distance norm of the quadratic form $\|X\|_A = \sqrt{xAx}$. In contrast to E, the objective function J_m contains, beside the number of a further parameter m. The fuzzification parameter m determines the influence of data objects x_i on the clustering process depending on their membership values μ_{ij}. If the fuzzification parameter m is increased, poorly classified objects which have small membership values μ_{ij}

contribute less to the calculation of the cluster centers c_i. Data objects with a large noise content thus have a reduced influence on the outcome of the clustering process. This makes fuzzy clustering especially suitable for noisy data sets such as gene expression data.

Several methods for minimizing the fuzzy objective function J_m have been proposed [91-93]. Fuzzy c means clustering is the most common algorithm for solving this problem. It is based on the first order conditions for a minimum of J_m for c cluster centers and N data vectors:

$$\mu_{kl} = \frac{1}{\left(\frac{\|x_l - c_k\|_D}{\sum_{i=1}^{c}\|x_l - c_i\|_D}\right)^{\frac{2}{m-1}}}, \ \forall k \in [1, c], l \in [1, N] \tag{2.7}$$

$$c_k = \frac{\sum_{j=1}^{N}(\mu_{kj})^m x_j}{\sum_{j=1}^{N}(\mu_{kj})^m}, \forall k \in [1, c] \tag{2.8}$$

A Picard iteration alternating between the evaluations of the equations above adjusts μ_{ij} and c_k until the change in J_m falls below a threshold ε or a maximal number of iterations T_{max} is reached. Note that c_k is the weighted mean of all x_i in cluster k. When $m \to 1$, the fuzzy clustering turns into hard k-means clustering. The cluster centers c_k are then simply the means of the clusters k. In the comparison experiment, we used an empirical method [94] to determine an adequate value for m based on the distribution of distances between genes in the sample space. A MATLAB implementation of FCM clustering was used for the following comparisons.

2.3 Proposed Method

The proposed evaluation methodology is based on the premise that the best clustering algorithm for expression data is that which tends to bring genes of similar function together, where function is known. The flowchart of the proposed approach is outlined in Figure 2.1. Genes are clustered by several benchmark clustering algorithms based on their expression profiles. Various clustering parameters and distance measures are investigated. These clustering results are compared by a figure of merit based on the Kullback-Leibler (KL) divergence between cluster membership and known gene attributes from public databases. We provide a more detailed description of the proposed method in the following sections.

2.3.1 Data Preprocessing of Gene Ontology

The GO database is organized in a rooted directed acyclic graph, with three branches corresponding to the three categories: cellular component, molecular function, and biological process. Each gene is annotated by one or multiple GO terms along the graph. The hierarchical nature of GO implies that genes annotated with a specific node are also annotated with every ancestor of that node. Nodes closer to the root of the graph usually correspond to more abstract functional descriptions and cover more genes, while nodes farther away from the root correspond to more detailed functional descriptions. Note that the structure of GO is not necessary a tree since each node may have multiple parents and may have multiple paths to the root of the graph [29].

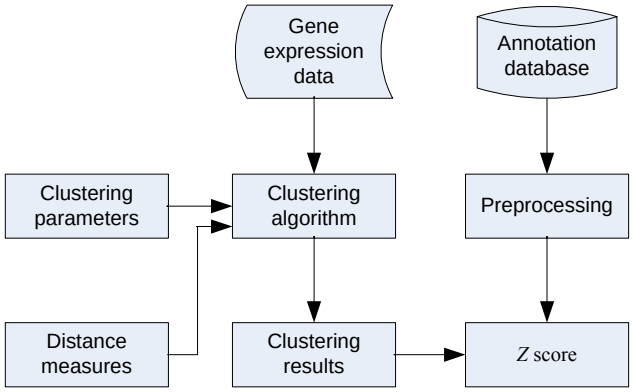


Figure 2.1 The flowchart of the proposed knowledge-based validation method in clustering.

For numerical analysis, each gene is assigned to a vector consisting of binary attributes $A = \{A_1, A_2, \dots, A_{N_A}\}$ with $A_i = \{0,1\}$, where $A_i = 1$ indicates that the gene has been annotated with the GO attribute A_i or one of its descendents, and $A_i = 0$ indicates our lack of knowledge about whether this gene possesses GO attribute A_i.

2.3.2 The Kullback-Leibler Divergence

The KL divergence, or relative entropy between two probability mass function $p(x)$ and $q(x)$ over a random variable X, proposed by Kullback and Leibler [95] is defined as

$$D(p||q) = \sum_{x \in X} p(x) log_2 \frac{p(x)}{q(x)} \tag{2.9}$$

The KL divergence $D(p||q)$ is a measure of the distance between two probability distributions or equivalently, it is the inefficiency of assuming that the distribution of X is q when the true distribution is p. In the context of clustering validity assessment, the KL divergence is interpreted as measuring the relative dissimilarity between the gene clusters and GO attributes being compared. The KL divergence is not a "distance metric" since it is asymmetric and does not satisfy the triangle inequality. However, numerical experiments show that this property improves the sensitivity of the measure [96]. The KL divergence has several important and useful properties, a few of which are listed below:

1) The KL divergence remains well-defined for continuous distributions;
2) It is invariant under parameter transformation;
3) Convergence in the KL sense implies convergence in the L1 norm sense but no proof is known for the reverse.
4) The χ^2 statistics is twice the first term in the Taylor expansion of the KL divergence, and
5) $D(p \| q)$ is convex in the pair (p, q).

2.3.3 A Figure of Merit based on KL Divergence

Suppose that n genes in one dataset are annotated by m GO attributes, and they are grouped into J clusters based on their gene expression profiles. Let p_{jr} be the probability of selecting a gene of GO attribute r from a cluster j. Thus the frequency distribution of GO attributes in a cluster j is defined as

$$P_j = (p_{j1}, p_{j2}, \dots, p_{jm}), where\ \sum_{r=1}^{m} p_{jr} = 1 \tag{2.10}$$

Let p_r^b be the probability of selecting a gene of GO attribute r from the total gene set. The background frequency distribution of GO attributes in the dataset is

$$P_b = (p_{b1}, p_{b2}, \dots, p_{bm}), where \sum_{r=1}^{m} p_{br} = 1 \tag{2.11}$$

Therefore, to quantitatively measure the biological significance of the cluster j, the KL divergence between p_{jr} and p_{br} is calculated by

$$D(P_j||P_b) = \sum_{r=1}^{m} p_{jr} log_2 \frac{p_{jr}}{p_{br}} \tag{2.12}$$

By summing over all the J clusters, the mean KL divergence is

$$\bar{D} = \frac{1}{J}\sum_{j=1}^{J} D(P_j||P_b) \tag{2.13}$$

To normalize $\bar{D}$, we calculate $\bar{D}_{random}$ for 1000 clustering partitions by randomly assigning genes to J clusters with the same gene numbers as the original one. A z score is calculated by

$$z\ score = \frac{\bar{D} - mean(\bar{D}_{random})}{std(\bar{D}_{random})} \tag{2.14}$$

The $z\ score$ is interpreted as a standardized distance between the $\bar{D}$ obtained by clustering analysis and those $\bar{D}_{random}$ obtained by random assignment of genes to clusters.

2.3.4 Selection Criteria of GO Attributes

It is reasonable to assume that clustering methods are used to seek a fine structure in the dataset to be clustered. Genes can be broadly classified according to the cell cycle phases in which they peak, yielding five clusters in yeast cell cycle dataset [60]. However, such clustering results generate little new knowledge. People are more interested in finding those groups which sharing rather specific biological functions. On the other hand, it is not helpful if we group each gene into one cluster. Base on these considerations, we propose two criteria to select GO attributes from the database:

1) Independency U: GO attributes should be as independent as possible with each other. The uncertainty coefficient U_{ij} between A_i and A_j is defined as

$$U_{ij} = \frac{MI(A_i, A_j)}{\max(MI)} \tag{2.15}$$

where MI is the mutual information between A_i and A_j. The U is defined as the upper bound of the U_{ij}. When U_{ij} is larger than U, it indicates that two GO attributes A_i and A_j are shared by substantial the same collection of genes. One of them should be removed from the analysis to avoid counting essentially the same attribute twice.

2) Sharedness N_{min}: GO attributes shared by small number of genes will opt to generate the optimal clusters with too specific biological functions.

We will investigate the sensitivity of GO attribute filtering process in the Results section.

2.4 Results

2.4.1 Datasets

In this section, we demonstrated the proposed method via three public gene expression datasets: rat CNS dataset [14], yeast cell cycle dataset [60], and human *Hela* cell cycle dataset [86].

Rat CNS dataset

This case study is based on the dataset published in [14], consisting of gene expression levels for 112 genes during the development of the CNS of rats. Each gene was measured at nine different points in time (of which the last, measured for the adult animal, was not used here). The first measurement was made 10 days before birth, and the intervals between measurements were 2 or 3 days in the period before birth and 7 days after birth.

Yeast cell cycle dataset

The yeast cell cycle data presented in [60] consist of six time series (*cln3*, *clb2*, *alpha*, *cdc15*, *cdc28*, and *elu*) expression measurements of the transcript (mRNA) levels of *S. cerevisiae* genes. 800 genes were identified as cell cycle regulated based on cluster analysis in [60]. Here, we used the *cdc15* time course data of the 800 genes since it has the largest number of time points (24).

Human *Hela* cancer cell cycle dataset

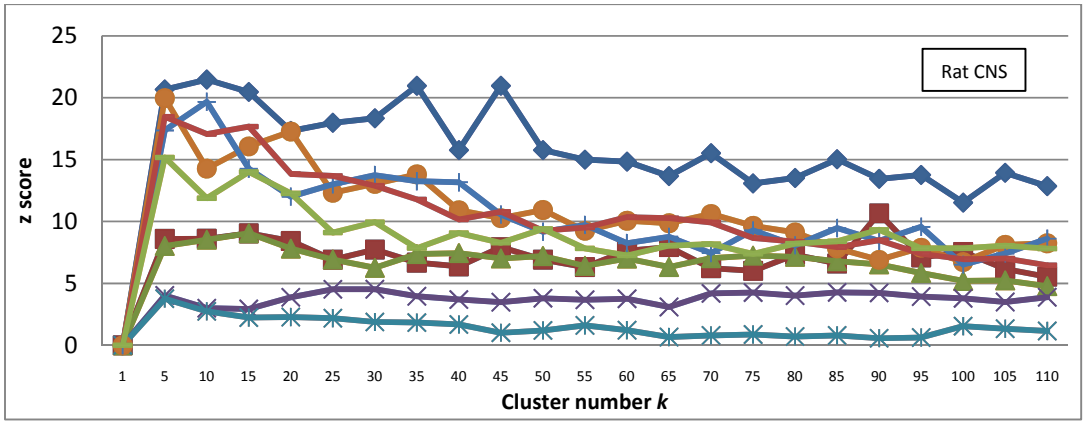
The human *Hela* cell cycle dataset [86] consists of five time courses (114 total arrays). RNA samples were collected for points (typically every 1-2 h) for 30 h (Thy- Thy1), 44 h (Thy-Thy2), 46 h (Thy-Thy3), 36 h (Thy-Noc), or 14 h (shake) after the synchronous arrest. The cell cycle related gene set contains 1,134 clones corresponding to 874 UNIGENE clusters (UNIGENE build 143). Of these, 1,072 have corresponding Entrez gene IDs, among which 226 have more than one mapping to clones. In total, 846 genes were used for TRN inference. We chose the Thy-Thy3 time course gene expression pattern for 846 genes, since it has the largest number of time points (47).

To preprocess these datasets, we first imputed the missing values in the data by using K-nearest neighbor (KNN) imputation [97]. Following that, we standardized the data between zero and one.

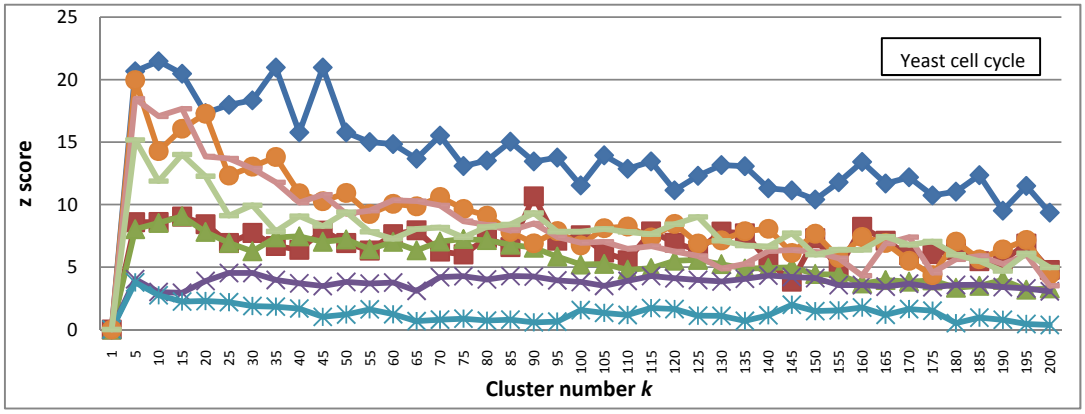
2.4.2 Performance Evaluation of Clustering Methods

The results are shown in Figure 2.2. For all three datasets, we calculated *z* scores for three hierarchical clustering methods (single, complete, and average linkage, all using the centered Pearson correlation distance), k-means clustering methods performed with three distance metrics (Euclidean, Pearson correlation, and Cosine), FCM with two different fuzzification parameter settings (FCM with default m =2 , and FCM-O with optimal m derived from the clustering dataset [94]), and SOM.

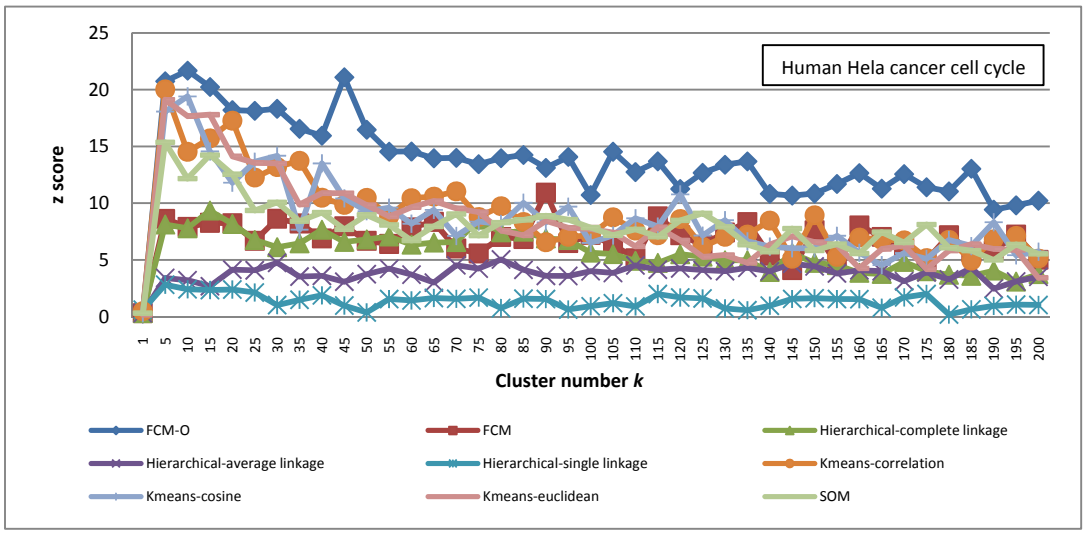
We have observed some common trends in the results from three different datasets. First of all, FCM clustering has the highest *z* score across different cluster numbers. We argue that there are two main reasons: first, since gene products frequently participate in more than one regulator mechanism to different degrees, they often have expression patterns correlated with more than one cluster. Fuzzy c means is suitable for this task since it can assign gene degrees of membership to a cluster. This feature enables FCM to provide more information about the structure of gene expression data. Second, compared to hard clustering, FCM is more noise robust. Compared to that, the hierarchical clustering with single linkage has the lowest z scores across different cluster numbers. After revisiting the clustering results, we observed that single-linkage hierarchical clustering tends to produce one single large clade and several singletons. This division necessarily separates genes that have attributes in common. On the other hand, the single clade will contain most genes, yielding almost no information. Overall, in all three datasets, the performance of FCM-O is better than other clustering methods. The single-linkage hierarchical clustering performs the worst, and average linkage is slightly better. Complete-linkage hierarchical clustering is slightly worse than k-means clustering with different distance metrics. SOM appears to perform as well as k-means approaches, but doesn't decrease *z* scores with increasing cluster number *k*, which is characteristic of hierarchical and k-means based methods. The k-means based methods have comparable results among three datasets. As shown in Figure 2.2, the optimal *z* scores for all methods lie in the region of small cluster number *k*.


Rat CNS
z score
Cluster number k

(A)


Yeast cell cycle
z score
Cluster number k

(B)


Human Hela cancer cell cycle
z score
Cluster number k
FCM-O
FCM
Hierarchical-complete linkage
Hierarchical-average linkage
Hierarchical-single linkage
Kmeans-correlation
Kmeans-cosine
Kmeans-euclidean
SOM

(C)

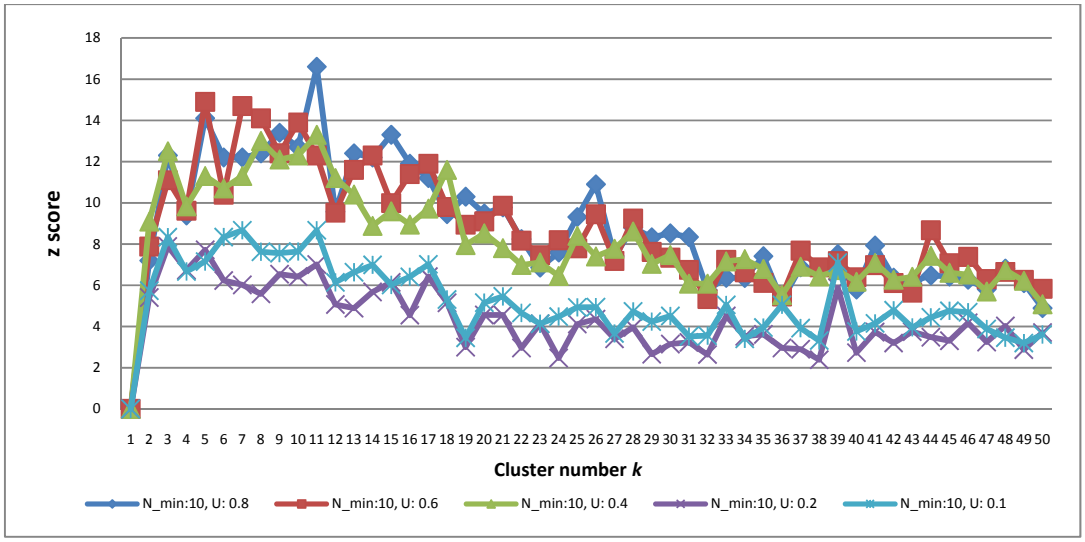
Figure 2.2 Three datasets clustered using different clustering algorithms. The horizontal axis shows the number of clusters desired, and the vertical axis shows z scores. Datasets are (A) rat CNS, (B) yeast cell cycle, and (C) human *Hela* cancer cell cycle.

2.4.3 Sensitivity of GO Attribute Selection

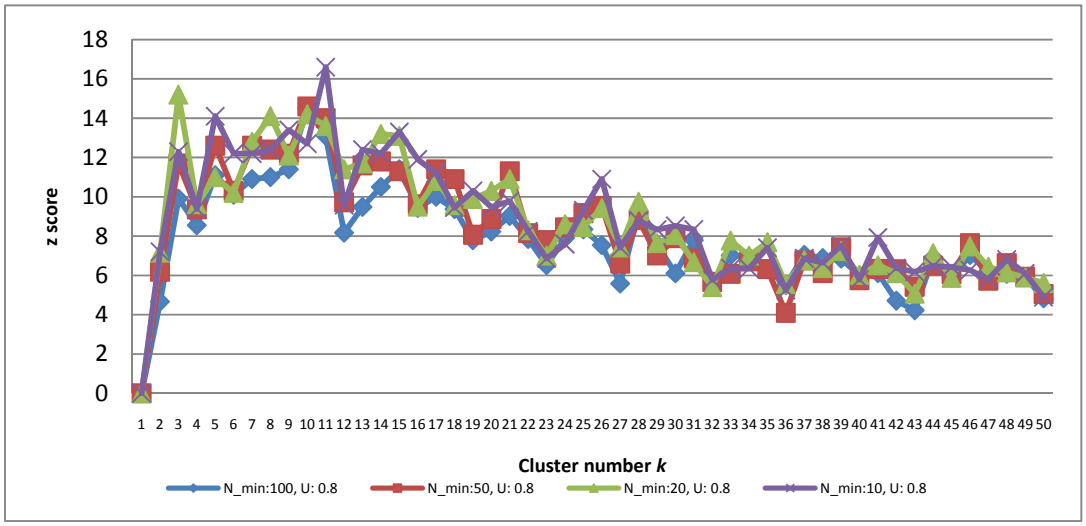
We also investigated whether the selection of GO attributes will affect the clustering evaluation process. As shown in Figure 2.3, different values of U and N_{min} were applied to yeast cell cycle dataset together with k-means clustering (centered Pearson correlation). For same N_{min} ($N_{min} = 10$), different values of U produced different sets of z scores. The larger U is, the higher z scores are produced (Figure 3.2(A)). For same U ($U = 0.8$), however, different values of N_{min} produced similar sets of z scores (Figure 3.2(B)). These results indicate that the independency U is an important factor in estimating the biological significance of clusters. One interesting finding is that although different U values may generate different z scores, the basic shape (e.g., location of the peak) does not.

2.5 Summary

A knowledge-based evaluation of clustering algorithms is presented in this chapter. This proposed approach applies a figure of merit based on KL divergence between cluster membership and known gene attributes. Well-known gene expression datasets from the rat CNS, yeast cell cycle and human *Hela* cancer cell cycle were analyzed to illustrate the applications. Several cluster partitions were used to evaluate the optimum number of clusters for the datasets. The results show that FCM clustering with optimal fuzzification parameter m is most suitable for gene expression data analysis which contains large amount of noise and no clear boundaries between clusters. This may facilitate the identification of experimental conditions for which genes are co-regulated and, thus, the identification of underlying regulatory processes.



(A)



(B)

Figure 2.3 Comparison of different parameter settings: (A) different values of $\boldsymbol{U}$ with same value of $\boldsymbol{N_{min}}$; (B) different values of $\boldsymbol{N_{min}}$ with same value of $\boldsymbol{U}$.

3 Gene Regulatory Module Inference by Hybrid Computational Intelligence Modeling

Integrating data from multiple global assays and curated databases is essential to understand the spatio-temporal interactions within cells. Different experiments measure cellular processes at various widths and depths, while databases contain biological information based on established facts or published data. Integrating these complementary datasets helps infer a mutually consistent TRN with strong similarity to the structure of the underlying gene regulatory modules. Decomposing the TRN into a small set of recurring regulatory patterns, called NMs, facilitates the inference. Identifying NMs defined by specific transcription factors establishes the modular framework structure of a TRN and allows the inference of transcription factor-target gene relationship. This chapter introduces a computational framework for utilizing data from multiple sources to infer transcription factor-target gene relationships on the basis of NM regulatory modules. The data include time course gene expression profiles, molecular interaction data, and GO information.

The proposed computational framework was tested in two biological processes: yeast cell cycle progression process, and human *Hela* cancer cell cycle. The identified gene regulatory modules were evaluated using the following validation strategies: (1) gene set enrichment analysis (GSEA) to evaluate the gene modules derived from clustering results; (2) binding site enrichment analysis to determine enrichment of the gene modules for the cognate binding sites of their predicted transcription factors; (3) comparison with previously reported results in the literatures to confirm the inferred regulations. The proposed framework could be beneficial to biologists for predicting the components of gene regulatory modules in which any candidate gene is involved. Such predictions can then be used to design a more streamlined experimental approach for biological validation. Understanding the dynamics of these gene regulatory modules will shed light on the processes that occur in cancer cells resulting from errors in cell cycle regulation.

3.1 Introduction

Enormous amount of biological data has been generated by the use of high-throughput analytical methods in biology during the last two decades. However, reverse engineering of a global TRN remains challenging because of several limitations including the following:

1) Tens of thousands of genes act at different temporal and spatial combinations in living cells;
2) Each gene interacts virtually with multiple partners either directly or indirectly, thus possible relationships are dynamic and non-linear;
3) Current high-throughput technologies generate data that involve a substantial amount of noise; and
4) The sample size is extremely low compared with the number of genes [98].

These inherited properties create significant problems in analysis and interpretation of these data. Standard statistical approaches are not powerful enough to dissect data with thousands of variables (i.e., semi-global or global gene expression data) and limited sample sizes (i.e., several to hundred samples in one experiment). These properties are typical in microarray and proteomic datasets [99] as well as other high dimensional data where a comparison is made for biological samples that tend to be limited in number, thus suffering from curse of dimensionality [12].

One approach to address the curse of dimensionality is to integrate multiple large data sets with prior biological knowledge. This approach offers a solution to tackle the challenging task of inferring TRN. Gene regulation is a process that needs to be understood at multiple levels of description [43, 44]. A single source of information (e.g., gene expression data) is aimed at only one level of description (e.g., transcriptional regulation level), thus it is limited in its ability to obtain a full understanding of the entire regulatory process. Other types of information such as various types of molecular interaction data by yeast two-hybrid analysis or genome-wide location analysis [100] provide complementary constraints on the models of regulatory By integrating limited but complementary data sources, we realize a mutually consistent hypothesis bearing stronger similarity to the underlying causal structures [44]. Among the various types of high-throughput biological data available nowadays, time course gene expression profiles and molecular interaction data are two complementary sets of information

that are used to infer regulatory components. Time course gene expression data are advantageous over typical static expression profiles as time can be used to disambiguate causal interactions. Molecular interaction data, on the other hand, provide high-throughput qualitative information about interactions between different entities in the cell. Also, prior biological knowledge generated by geneticists will help guide inference from the above data sets and integration of multiple data sources offers insights into the cellular system at different levels.

A number of researches have explored the integration of multiple data sources (e.g., time course expression data and sequence motifs) for TRN inference [58-61]. A typical approach for exploiting two or more data sources uses one type of data to validate the results generated independently from the other (i.e., without data fusion). For example, cluster analysis of gene expression data was followed by the identification of consensus sequence motifs in the promoters of genes within each cluster [60]. The underlying assumption behind this approach is that genes co-expressed under varying experimental conditions are likely to be co-regulated by the same transcription factor or sets of transcription factors. Holmes et al. constructed a joint likelihood score based on consensus sequence motif and gene expression data and used this score to perform clustering [101]. Segal et al. built relational probabilistic models by incorporating gene expression and functional category information as input variables [102]. Gene expression data and GO data were combined for TRN discovery in B cell [103]. Computational methodologies that allow systematic integration of data from multiple resources are needed to fully use the complementary information available in those resources.

Another way to reduce the complexity of the TRN inference problem is to decompose it into simple units of commonly used network structures. TRN is a network of interactions between transcription factors and the genes they regulate, governing many of the biological activities in cells. Cellular networks like TRNs are composed of many small but functional modules or units [62]. Breaking down TRNs into these functional modules will help understanding regulatory behaviors of the whole networks and study their properties and functions. One of these functional modules are called NM [54]. Since the establishment of the first NM in *Escherichia coli* [34], similar NMs have also been found in eukaryotes including yeast [45], plants [62], and animals [46-48], suggesting that the general structure of NMs are evolutionarily conserved. One well known family of NMs is the feed-forward loop [49], which appears in hundreds of gene

systems in *E. coli* [34, 50] and yeast [45, 51], as well as in other organisms [46-48, 52-54]. A comprehensive review on NM theory and experimental approaches could be found in Ref. [55]. Knowledge of the NMs to which a given transcription factor belongs facilitates the identification of downstream target gene modules. In yeast, a genome-wide location analysis was carried out for 106 transcription factors and five NMs were considered significant: autoregulation, feed-forward loop, single input module, multi-input module and regulator cascade.

In this Chapter, we propose a computational framework that integrates information from time course gene expression experiment, molecular interaction data, and GO category information to infer the relationship between transcription factors and their potential target genes at NM level. This was accomplished through a three-step approach outlined in the following: first, as introduced in Chapter 2, we applied cluster analysis of time course gene expression profiles to reduce dimensionality and used the GO category information to determine biologically meaningful gene modules, upon which a model of the gene regulatory module is built. This step enables us to address the scalability problem that is faced by researchers in inferring TRNs from time course gene expression data with limited time points. Second, we detected significant NMs for each transcription factor in an integrative molecular interaction network consisting of PPI and PDI data (hereafter called molecular interaction data) from thirteen publically available databases. Finally, we used neural network (NN) models that mimic the topology of NMs to identify gene modules that may be regulated by a transcription factor, thereby inferring the regulatory relationships between the transcription factor and gene modules. A hybrid of genetic algorithm and particle swarm optimization (GA-PSO) methods was applied to train the NN models.

The organization of this chapter is as follows. Section 3.2 briefly reviews the related methods. Section 3.3 introduces the proposed method to infer TRNs by integrating multiple sources of biological data. Section 3.4 will present the results on two real data sets: the yeast cell cycle data set [60], and the human *Hela* cell cycle data set [86]. Finally, Section 3.5 is devoted to the summary and discussions.

3.2 Review of Related Methods

In recent years, high throughput biotechnologies have made large-scale gene expression surveys a reality. Gene expression data provide an opportunity to directly review the activities of thousands of genes simultaneously. However, computational methods that can handle the complexity (e.g., noisy, substantial amount of variables, high dimensionality) of these biological data are often unavailable [104]. Powerful computational methods and data mining tools are needed for biologically meaningful inferences from gene expression data.

A variety of continuous or discrete, static or dynamic, quantitative or qualitative models have been proposed for inference of interaction networks. These include biochemically driven methods [105], linear models [106, 107], Boolean networks [108], fuzzy logic [109, 110], Bayesian networks [111], and recurrent neural networks (RNNs) [112-114]. Biochemically inspired models are developed on the basis of the reaction kinetics between different components of a network. However, most of the biochemically relevant reactions under participation of proteins do not follow linear reaction kinetics, and the full network of regulatory reactions is very complex and hard to unravel in a single step. Linear models attempt to solve a weight matrix that represents a series of linear combinations of the expression level of each gene as a function of other genes, which is often underdetermined since gene expression data usually have far fewer dimensions than the number of genes. In a Boolean network, the interactions between genes are modeled as Boolean function. Boolean networks assume that genes are either “on” or “off” and attempt to solve the state transitions for the system. The validity of the assumptions that genes are only in one of these two states has been questioned by a number of researchers, particularly among those in the biological community. In [109], an approach was proposed based on fuzzy rules of a known activator/repressor model of gene interaction. This algorithm transforms expression values into qualitative descriptors that is evaluated by using a set of heuristic rules and searches for regulatory triplets consisting of activator, repressor, and target gene. This approach, though logical, is a brute force technique for finding gene relationships. It involves significant computational time, which restricts its practical usefulness. In [110], we proposed the use of clustering as an interface to a fuzzy logic–based method to improve the computational efficiency. In a Bayesian network model, each gene is considered as a random variable and the edges between a pair of genes represent the conditional dependencies entailed in

the network structure. Bayesian statistics are applied to find certain network structure and the corresponding model parameters that maximize the posterior probability of the structure given the data. Unfortunately, this learning task is NP-hard, and it also has the underdetermined problem. The RNN model has received considerable attention because it can capture the nonlinear and dynamic aspects of gene regulatory interactions. Several algorithms have been applied for RNN training in network inference tasks, such as fuzzy-logic [113] and genetic algorithm [114].

3.3 Proposed Method

3.3.1 Overview of the Proposed Framework

In the proposed framework, we consider two different layers of networks in the TRN. One is the molecular interaction network at the factor-gene binding level. The other is the functional network that incorporates the consequences of these physical interactions, such as the activation or repression of transcription. We used three types of data to reconstruct the TRN, namely PPIs derived from a collection of public databases, PDIs from the TRANSFAC database [28], and time course gene expression profiles. The first two data sources provided direct network information to constrain the TRN model. The gene expression profiles provided an unambiguous measurement on the causal effects of the TRN model. Gene ontology annotation describes the similarities between genes within one network, which facilitates further characterization of the relationships between genes. The goal is to discern dependencies between the gene expression patterns and the physical inter-molecular interactions revealed by complementary data sources.

The framework model for TRN inference is illustrated in Figure 3.1. Besides data pre-processing, three successive steps are involved in this framework as outlined in the following:

***Gene module selection*:** genes with similar expression profiles are represented by a gene module to address the scalability problem in TRN inference [110]. The assumption is that a subset of genes that are related in terms of expression (co-regulated) can be grouped together by virtue of a unifying *cis*-regulatory element(s) associated with a common transcription factor regulating each and every member of the cluster (co-expressed) [115]. Gene ontology information is used to define the optimal number of clusters with respect to certain broad functional categories. Since each gene module identified from clustering analysis mainly

represents one broad biological process or category as evaluated by *FuncAssociate* [116], the regulatory network implies that a given transcription factor is likely to be involved in the control of a group of functionally related genes [117]. This step is implemented by the method proposed in Chapter 2.

***Network motif discovery*:** to reduce the complexity of the inference problem, NMs are used instead of a global TRN inference. The significant NMs in the combined molecular interaction network are first established and assigned to at least one transcription factor. These associations are further used to reconstruct the regulatory modules. This step is implemented using FANMOD software [118]. We briefly describe it in the following section.

***Gene regulatory module inference*:** for each transcription factor assigned to a NM, a NN is trained to model a TRN that mimics the associated NM. Genetic algorithm generates the candidate gene modules, and particle swarm optimization (PSO) is used to configure the parameters of the NN. Parameters are selected to minimize the root mean square error (RMSE) between the output of the NN and the target gene module's expression pattern. The RMSE is returned to GA to produce the next generation of candidate gene modules. Optimization continues until either a pre-specified maximum number of iterations are completed or a pre-specified minimum RMSE is reached. The procedure is repeated for all transcription factors. Biological knowledge from public databases is used to evaluate the predicted results. This step is the main focus of this chapter.

3.3.2 Network Motif Discovery

The NM analysis is based on network representation of PPIs and PDIs. A node represents both the gene and its protein product. A PPI is represented by a bi-directed edge connecting the interacting proteins. A PDI is an interaction between a transcription factor and its target gene and is represented by a directed edge pointing from the transcription factor to its target gene.

All connected subnetworks containing three nodes in the interaction network are collated into isomorphic patterns, and the number of times each pattern occurs is counted. If the number of occurrences is at least five and significantly higher than in randomized networks, the pattern is considered as a NM. The statistical significance test is performed by generating 1000 randomized networks and computing the fraction of randomized networks in which the pattern

appears at least as often as in the interaction network, as described in [119]. A pattern with $p \leq 0.05$ is considered statistically significant. This NM discovery procedure was performed using the FANMOD software [118]. For different organisms, different NMs may be identified. As shown in Figure 3.2 and Figure 3.3, different sets of NMs were detected in yeast and human. Both NM sets shared some similar NM structures. For example, Figure 3.2(B) and Figure 3.3(C) were both feed-forward loops. This NM has been indentified and studied in many organisms including *E. coli*, yeast, and human. Knowledge of these NMs to which a given transcription factor belongs facilitates the identification of downstream target gene modules.

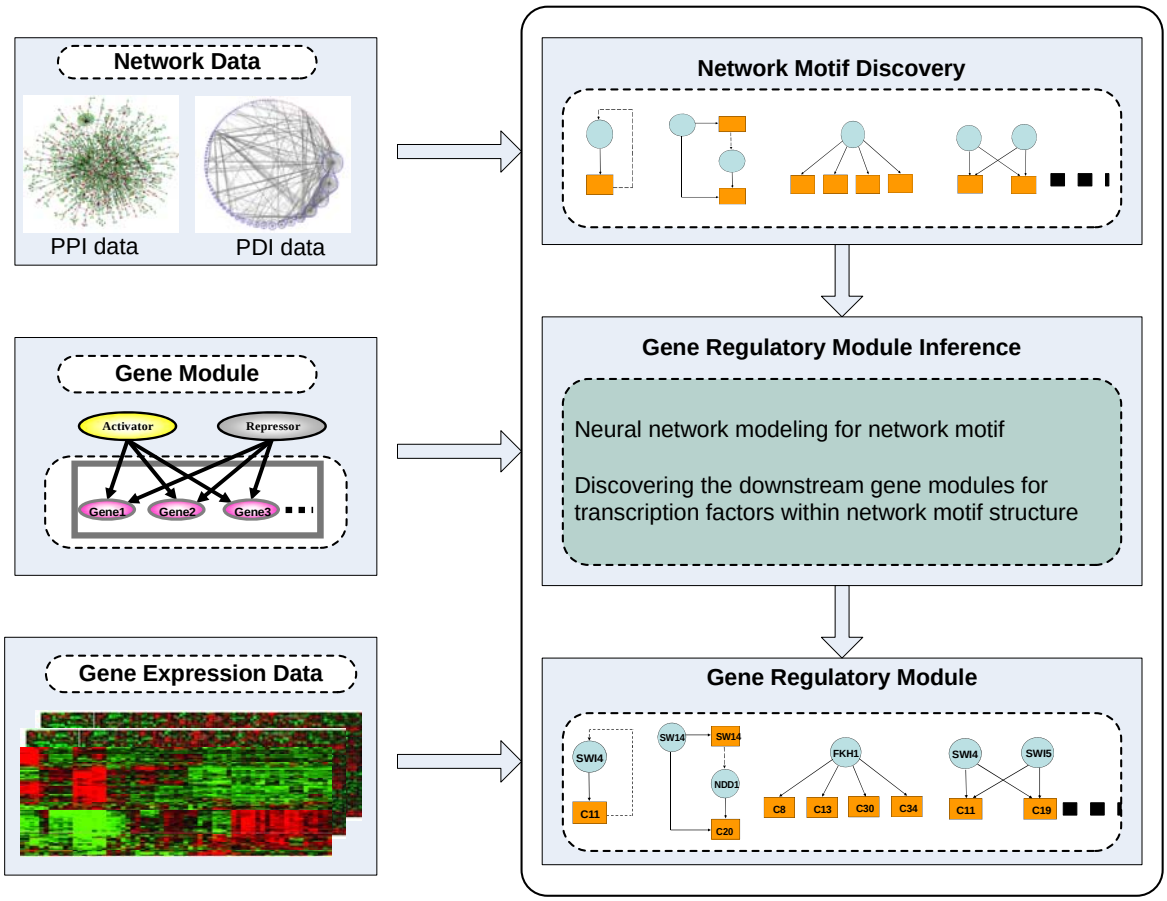


Figure 3.1 Schematic overview of the computational framework used for the gene regulatory module inference.

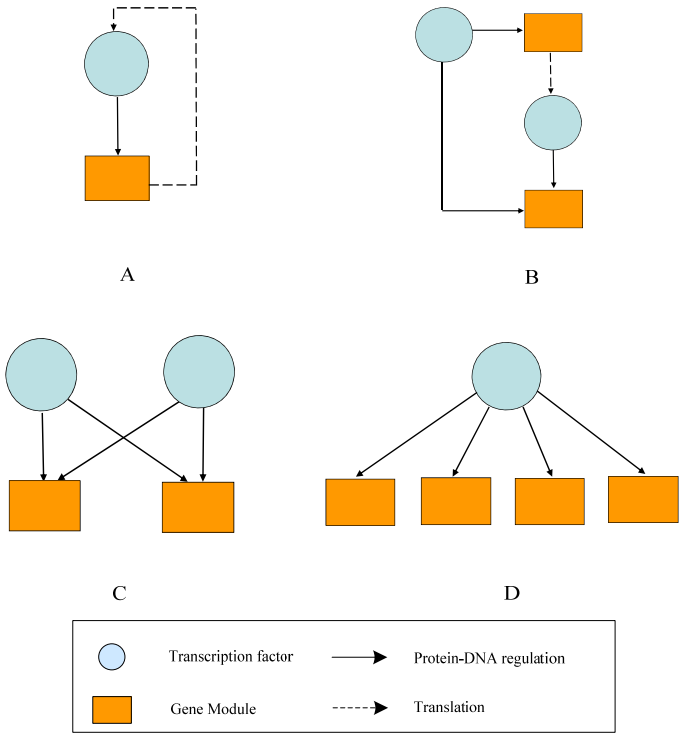


Figure 3.2 Four NMs discovered in yeast: (A) auto-regulatory motif; (B) feed-forward loop; (C) single input module; and (D) multi-input module.

3.3.3 Transcriptional Regulatory Module Inference

In building NNs for inferring TRNs, the identification of the correct downstream gene modules and determination of the free parameters (weights and biases) to mimic the real data is a challenging task given the limited available quantity of data. For example, in inferring a GRN from microarray data, the number of time points is considerably low compared to the number of genes involved. Considering the complexity of the biological system, it is difficult to adequately describe the pathways involving a large number of genes with few time points. We addressed this challenge by inferring TRNs at NM modular level instead of gene level. Neural network models were built for all NMs detected in the molecular interaction data. A hybrid search algorithm, called GA-PSO, was proposed to select the candidate gene modules (output node) in a NN and to update its free parameters simultaneously. We illustrate the models and training algorithm in detail in the following sections.

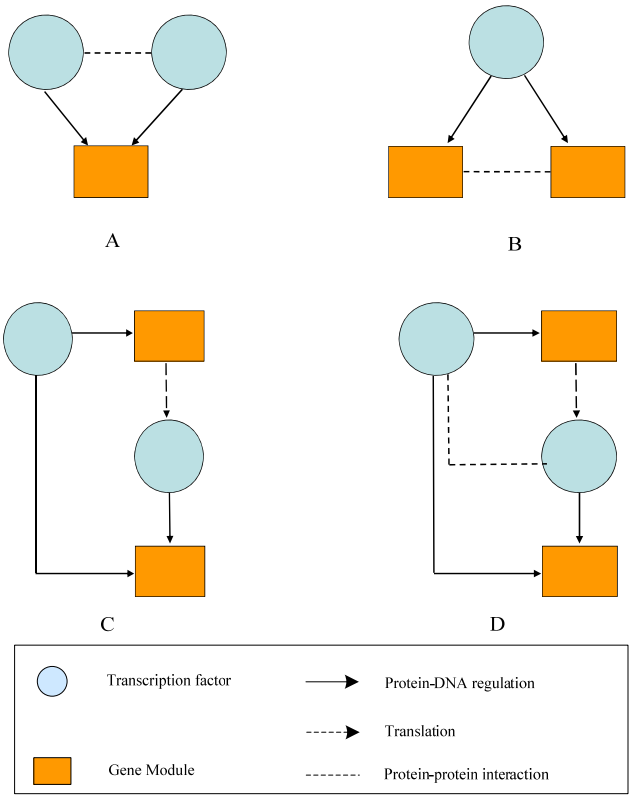
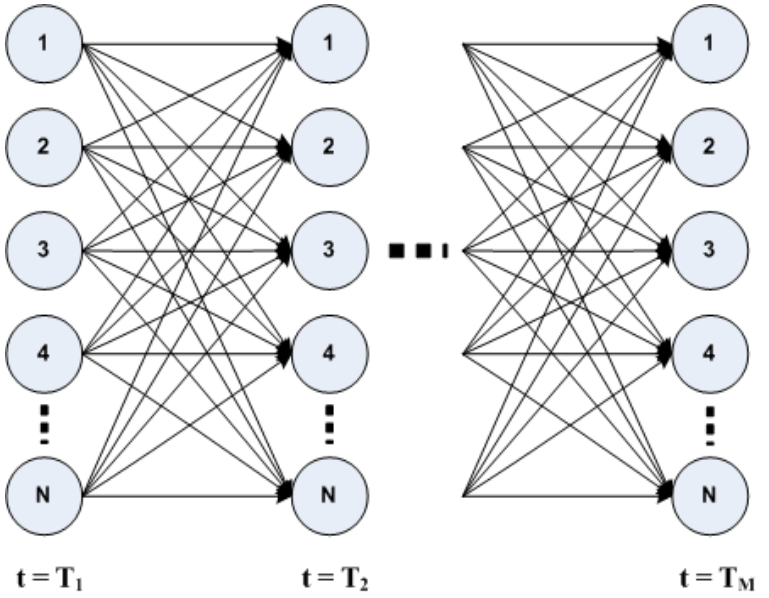


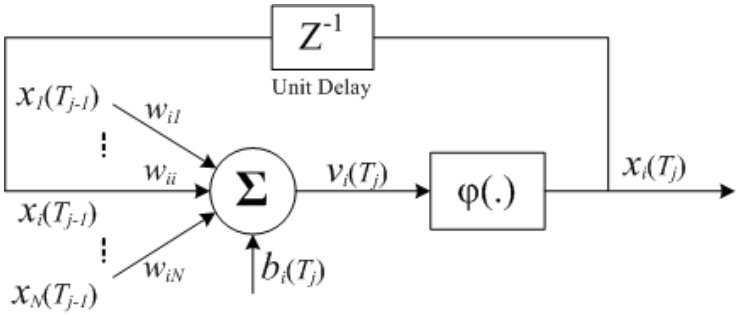
Figure 3.3 Four NMs discovered in human: (A) multi-input module; (B) single input module; (C) feed-forward loop - 1; and (D) feed-forward loop - 2.

3.3.3.1 Neural Network Model

The neural network model is based on the assumption that the regulatory effect on the expression of a particular gene can be expressed as a neural network (Figure 3.4(A)), where each node represents a particular gene and the wirings between nodes define regulatory interactions. Each layer of the network represents the expression level of genes at time *t*. The output of a node at time $t + \Delta t$ is derived from the expression levels at the time t and the connection weights of all genes connected to the given gene. As shown in the figure, the output of each node is fed back to its input after a unit delay and is connected to other nodes. The network is used as a model to a TRN: every gene in the network is considered as a neuron; NN model considers not only the interactions between genes but also gene self-regulations.



(A)



(B)

Figure 3.4 Architecture of a fully connected NN (A) and details of a single recurrent neuron (B).

Figure 3.4 (B) illustrates the details of the *i*th self-feedback neuron (e.g. *i*th gene in the TRN), where v_i, known as the induced local field (activation level), is the sum of the weighted inputs (the regulation of other genes) to the neuron (*i*th gene); and $\varphi(.)$ represents an activation function (integrated regulation of the whole NN on *i*th gene), which transforms the activation level of a neuron into an output signal (regulation result). The induced local field and the output of the neuron, respectively, are given by:

$$v_i(T_j) = \sum_{k=1}^{N} w_{ik} x_k(T_{j-1}) + b_i(T_j) \tag{3.1}$$

$$x_i(T_j) = \varphi(v_i(T_j)) \tag{3.2}$$

where the synaptic weights w_{i1}, w_{i2},…, w_{iN} define the strength of connections between the *i*th neuron (e.g. *i*th gene) and its inputs (e.g. expression level of genes). Such synaptic weights exist between all pairs of neurons in the network. $b_i(T_j)$ denotes the bias for the *i*th neuron at time T_j. We denote $\vec{w}$ as a weight vector that consists of all the synaptic weights and biases in the network. $\vec{w}$ is adapted during the learning process to yield the desired network outputs. The activation function $\varphi(.)$ introduces nonlinearity to the model. When information about the complexity of the underlying system is available, a suitable activation function can be chosen (e.g. linear, logistic, sigmoid, threshold, hyperbolic tangent sigmoid or Gaussian function.) If no prior information is available, our algorithm uses the hyperbolic tangent sigmoid function.

As a cost function, we use the RMSE between the expected output and the network output across time and neurons in the network. The cost function is written as:

$$E(\vec{w}) = \sqrt{\frac{1}{Nm}\sum_{t=T_1}^{T_m}\sum_{i=1}^{N}[\,x_i(t) - \hat{x}_i(t)]^2} \tag{3.3}$$

where $x_i(t)$ and $\hat{x}_i(t)$ are the true and predicted values (expression levels) for the *i*th neuron at time *t*. The goal is to determine weight vector $\vec{w}$ that minimize this cost function. This is a challenging task if the size of the network is large and only few samples (time points) are available.

For each NM shown in Figure 3.2 and 3.3, a corresponding NN is built. Figure 3.5 presents the detailed NN model for each NM in Figure 3.2. For auto-regulatory motif, the NN model is the same as single input module except that its downstream gene module contains the transcription factor itself. Since the expression level of gene module is the mean of expression level of its member genes, the input node and output node have different expression profiles. This avoids an open-loop problem which may cause the stability of the model. For single input and multiple input modules, instead of selecting multiple downstream gene modules simultaneously, we build NN models to find candidate gene modules one at one time. The selected gene modules are merged together to build the final NM gene regulatory modules.

Based on NMs to which a given transcription factor belongs, the next process is to find out the target gene modules and the relationships between them. To resolve this problem, we propose a hybrid GA-PSO training algorithm for NN model.

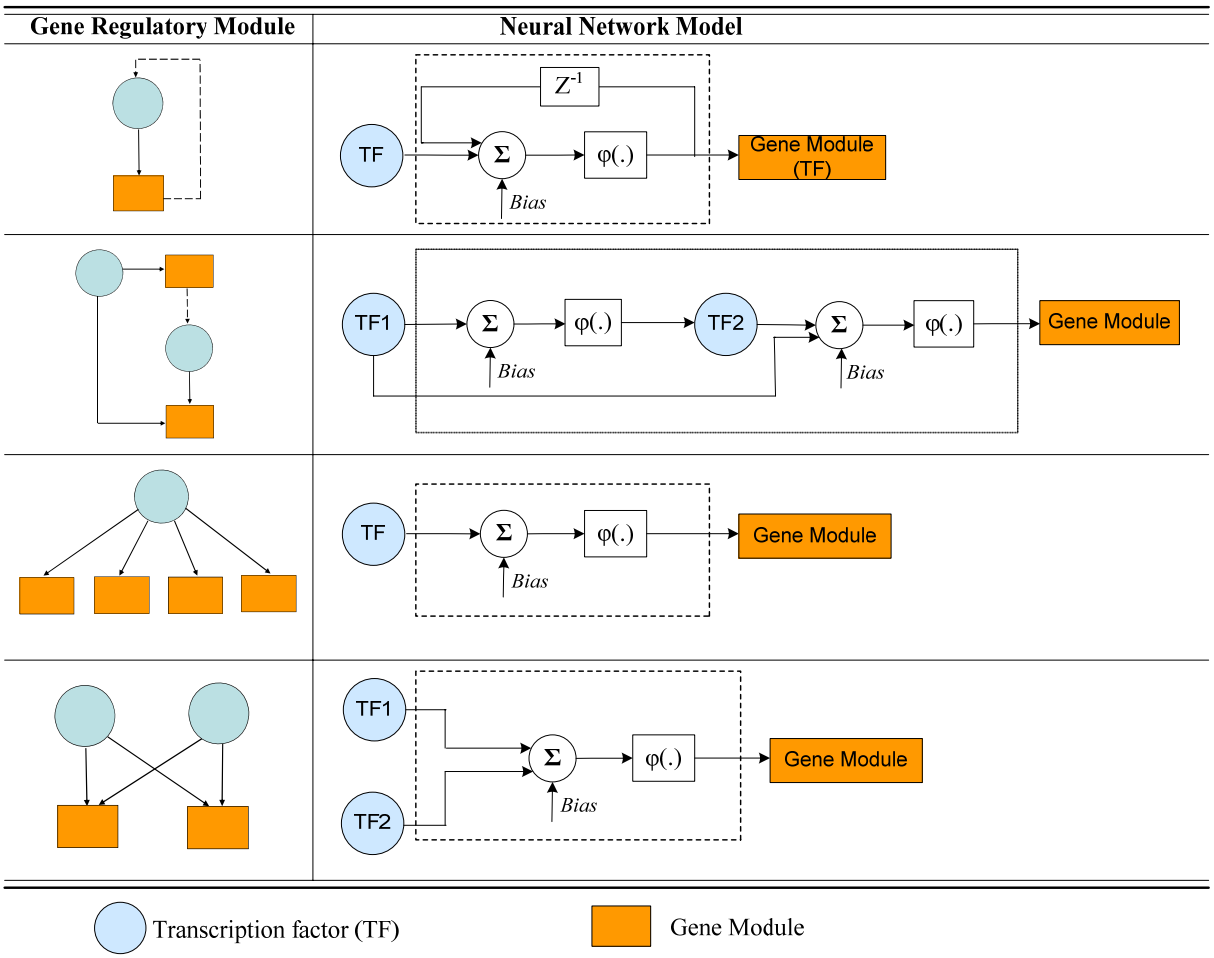


Figure 3.5 NN models mimicking the topologies of the four NMs shown in Figure 3.2. Z^{-1} denotes a unit delay and $\varphi(.)$ is a logistic sigmoid activation function.

3.3.3.2 Genetic Algorithm

Genetic algorithms are stochastic optimization approaches which mimic representation and variation mechanisms borrowed from biological evolution, such as *selection*, *crossover*, and *mutation* [120]. In a GA, a candidate solution is represented as a linear string analogous to a biological *chromosome*. The general scheme of GAs starts from a population of randomly generated candidate solutions (chromosomes). Each chromosome is then evaluated and given a value which corresponds to a *fitness* level in the objective function space. In each generation, chromosomes are chosen based on their fitness to reproduce offspring. Chromosomes with a high level of fitness are more likely to be retained while the ones with low fitness tend to be This process is called *selection*. After selection, offspring chromosomes are constructed from

parent chromosomes using operators that resemble crossover and mutation mechanisms in evolutionary biology. The crossover operator, sometimes called *recombination*, produces new offspring chromosomes that inherit information from both sides of parents by combining partial sets of elements from them. The mutation operator randomly changes elements of a chromosome with a low probability. Over multiple generations, chromosomes with higher fitness values are left based on the survival of the fitness. A detailed description of GA is shown in Figure 3.6.

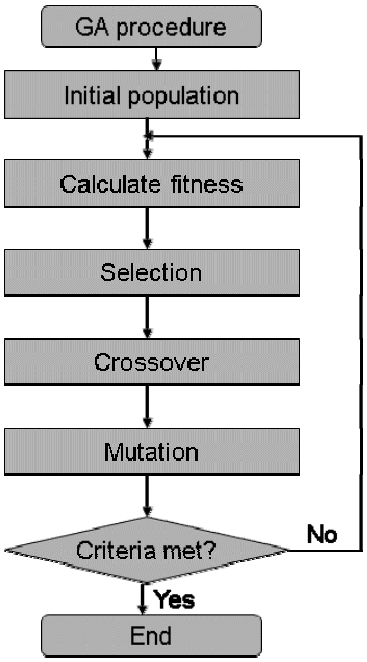


Figure 3.6 A detailed description of GA.

In this chapter, we propose to apply to GA to select the best suitable downstream gene module(s) for each transcription factor and the NN model(s) that mimic its NM(s). The Genetic Algorithm and Direct Search Toolbox (Mathworks, Natick, MA) is used for implementation of GA.

3.3.3.3 Particle Swarm Optimization

After the candidate downstream gene modules are selected by GA, PSO is proposed to determine the parameters in the NN model.

Particle swarm optimization is motivated by the behavior of bird flocking or fish blocking, originally intended to explore optimal or near-optimal solutions in sophisticated continuous spaces [121]. Its main difference from other evolutionary algorithms (e.g., GA) is that PSO

relies on cooperation rather than competition. Good solutions in the problem set are shared with their less-fit ones so that the entire population improves.

Particle swarm optimization consists of a swarm of particles, each of which represents a candidate solution. Each particle is represented as a D-dimensional vector $\vec{w}$, with a corresponding D-dimensional instantaneous trajectory vector $\Delta\vec{w}(t)$, describing its direction of motion in the search space at iteration t. The index i refers to the ith particle. The core of the PSO algorithm is the position update rule (Eq. (3.4)) which governs the movement of each of the n particles through the search space.

$$\vec{w}_i(t+1) = \vec{w}_i(t) + \Delta\vec{w}_i(t+1) \tag{3.4}$$

$$\Delta\vec{w}_i(t+1) = \chi[\Delta\vec{w}_i(t) + \Phi_1\left(\vec{w}_{i,best}(t) - \vec{w}_i(t)\right) + \Phi_2\left(\vec{w}_{G,best}(t) - \vec{w}_i(t)\right)] \tag{3.5}$$

where $\Phi_1 = c_1\begin{bmatrix} r_{1,1} & 0 & 0 & 0 \\ 0 & r_{1,2} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & r_{1,D} \end{bmatrix}$ and $\Phi_2 = c_2\begin{bmatrix} r_{2,1} & 0 & 0 & 0 \\ 0 & r_{2,2} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & r_{2,D} \end{bmatrix}$

At any instant, each particle is aware of its individual best position, $\vec{w}_{i,best}(t)$, as well as the best position of the entire swarm, $\vec{w}_{G,best}(t)$. The parameters c_1 and c_2 are constants that weight particle movement in the direction of the individual best positions and global best positions, respectively; and $r_{1,j}$ and $r_{2,j}$, $j=1,2,\ldots D$ are random scalars distributed uniformly between 0 and 1, providing the main stochastic component of the PSO algorithm. Figure 3.7 shows a vector diagram of the contributing terms of the PSO trajectory update. The new change in position, $\Delta\vec{w}_i(t+1)$, is the resultant of three contributing vectors: (i) the inertial component, $\Delta\vec{w}_i(t)$, (ii) the movement in the direction of individual best, $\vec{w}_{i,best}(t)$, and (iii) the movement in the direction of the global (or neighborhood) best, $\vec{w}_{G,best}(t)$.

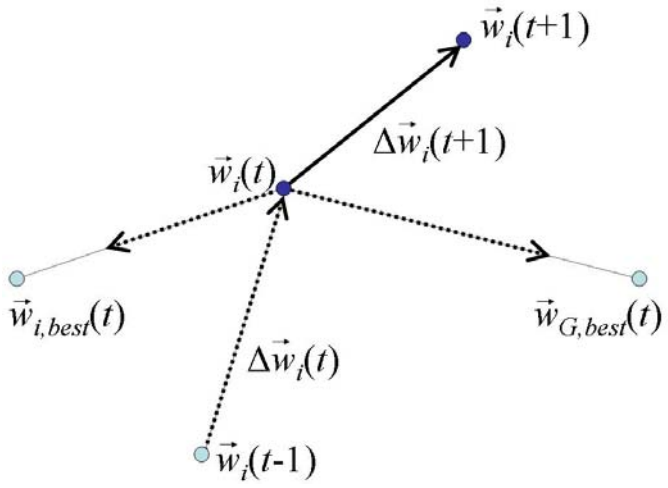
The constriction factor, χ, may also help to ensure convergence of the PSO algorithm, and is set according to the weights c_1 and c_2 as in Eq.(3.6).

$$\chi = \frac{2}{\left|2-\varphi-\sqrt{\varphi^2-4\varphi}\right|}, \varphi = c_1 + c_2, \varphi > 4 \tag{3.6}$$

The key strength of the PSO algorithm is the interaction among particles. The second term in Eq. (3.5), $\Phi_2\left(\vec{w}_{G,best}(t) - \vec{w}_i(t)\right)$, is considered to be a "social influence" term. While this

term tends to pull the particle towards the globally best solution, the first term, $\Phi_1\left(\overrightarrow{w}_{i,best}(t) - \overrightarrow{w}_i(t)\right)$, allows each particle to think for itself. The net combination is an algorithm with excellent trade-off between total swarm convergence, and each particle's capability for global exploration. Moreover, the relative contribution of the two terms is weighted stochastically.

The algorithm consists of repeated application of the velocity and position update rules presented above. Termination occurs by specification of a minimum error criterion, maximum number of iterations, or alternately when the position change of each particle is sufficiently small as to assume that each particle has converged.



A pseudo-code description of the PSO algorithm is provided below:

1) Generate initial population of particles, $\overrightarrow{w}_i$, i=1,2,…,n, distributed randomly (uniform) within the specified bounds.
2) Evaluate each particle with the objective function, $f(\overrightarrow{w}_i)$; if any particles have located new individual best positions, then replace previous individual best positions, $\overrightarrow{w}_{i,best}$, and keep track of the swarm global best position, $\overrightarrow{w}_{G,best}$.
3) Determine new trajectories, $\Delta\overrightarrow{w}_i(t+1)$, according to Eq. (3.5).
4) Update each particle position according to Eq. (3.4).

5) Determine if any $\overrightarrow{w_i}(t+1)$ are outside of the specified bounds; hold positions of particles within the specified bounds.

If termination criterion is met (for example completed maximum number of iterations), then $\overrightarrow{w_{G,best}}$ is the best solution found; otherwise, go to step (2).

Selection of appropriate values for the free parameters of PSO plays an important role in the algorithm's performance. In our study, the parameters setting is presented in Table 3.1, and the constriction factor χ is determined by Eq.(3.6).

Table 3.1 PSO parameter setting

Parameter	Value
Maximum velocity, Vmax	2
Maximum search space range, |Wmax|	[-5,5]
Acceleration constants, c1 & c2	2.05, 2.05
Size of Swarm	20

The PSOt Toolbox [122] was used for implementation of PSO.

3.3.3.4 GA-PSO Training Algorithm

A hybrid of GA and PSO methods (GA-PSO) is applied to determine the gene modules that may be regulated by each transcription factor. Genetic algorithm generates candidate gene modules, while the PSO algorithm determines the parameters of a given NN represented by a weight vector $\overrightarrow{w}$. The RMSE between the NN output and the measured expression profile is returned to GA as a fitness function and to guide the selection of target genes through reproduction, cross-over, and mutation over hundreds of generations. The stopping criteria are pre-specified minimum RMSE or maximum number of generations. The GA-PSO algorithm is run for each transcription factor to train a NN that has the architecture mimicking the identified known NM(s) for the transcription factor. Thus, for a given transcription factor (input), the following steps are carried out to identify its likely downstream gene modules (output) based on their NM(s):

1. Assign the NM to the transcription factor it belongs to.

2. Use the following GA-PSO algorithm to build a NN model that mimics the NM to identify the downstream gene modules.

    2.1. Generate combinations of *M* gene modules to represent the target genes that may be regulated by the transcription factor. Each combination is a vector/chromosome. The initial set of combinations is composed of the initial population of chromosomes.

    2.2. Use the PSO algorithm to train a NN model for each chromosome, where the input is the transcription factor and the outputs are gene modules. The goal is to determine the optimized parameters of the NN that maps the measured expression profiles of the transcription factor to the gene modules.

    2.3. For each chromosome, calculate the RMSE between the predicted output of the NN and measured expression profiles for the target gene modules.

    2.4. Apply GA operators (reproduction, cross-over, mutation) based on the RMSE calculated in Step 2.3 as a fitness value. This will generate new vectors/chromosomes altering the choice of output gene module combinations.

    2.5. Repeat steps 2.1 – 2.4 until stop criteria are met. The stopping criteria are numbers of generations or minimum RMSE, depending on which one is met first.

    2.6. Repeat Steps 2.1 – 2.5 for each NM the transcription factor is assigned to.

3. Repeat Steps 1 and 2 for each transcription factor.

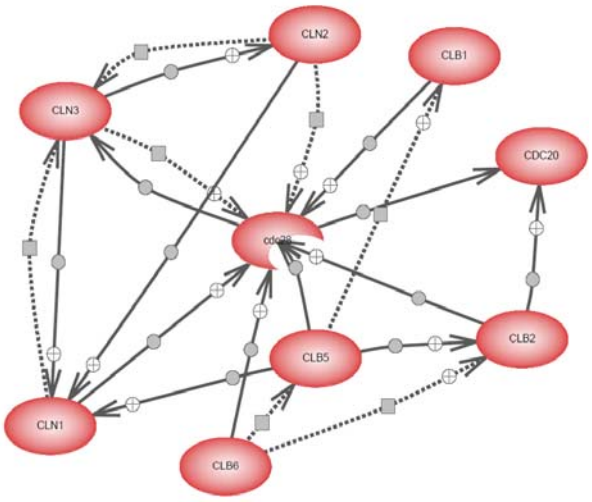
When the process is completed, NM regulatory modules are constructed between transcription factors and their regulated gene modules.

3.4 Results

3.4.1 PSO Performance Evaluation

To perform large scale analysis using the proposed framework, we first evaluated the performance of PSO in parameter training of NN models for a well-known TRN.

A small TRN consisting of nine well-studied cyclin genes (*CLB1*, *CLB2*, *CLB5*, *CLB6*, *CDC20*, *CLN1*, *CLN2*, *CLN3* and *CDC28*) were reconstructed from Pathway Studio pathway analysis software (Figure 3.8) (http://www.ariadnegenomics.com/products/pathway-studio/). Table 3.2 illustrates the gene relationships depicted in Figure 3.8. The expression levels of these genes were extracted from Spellman et al [60]. In the dataset, the biological samples were synchronized by three different methods: α factor arrest, arrest of a cdc15, and cdc28 temperature-sensitive mutant. We used the cdc15 part, which has 24 experimental conditions, as training dataset. The other two parts, alpha and cdc28, which have 18 and 17 time points, were used as validation and testing datasets, respectively.



A fully connected NN model was constructed for nine genes. The PSO algorithm was run 10 times to determine the parameters between each gene. In each run, Eq. (3.3) was evaluated 1000 times to identify the parameter vector $\vec{w}$ that leads to the least RMSE. To improve the prediction accuracy, only the connections obtained in at least 50% of the runs were selected. Figure 3.9 shows the outputs of the true network and the predicted RNN for the testing dataset.

As the true TRN that governs the interaction among the nine genes is not available, the accuracy of the network is determined by how well it fits the measured gene expression data. To get insight into the performance of the PSO method, we randomly generated 100 NN structures and optimized their parameters using PSO. The average and standard deviation of the RMSE for the 100 randomly generated RNN were 0.25 and 0.14, respectively, while the optimal RNN model found through the PSO method yielded an average MSE = 0.1 and standard deviation = 0.08 in 10 runs (after normalization).

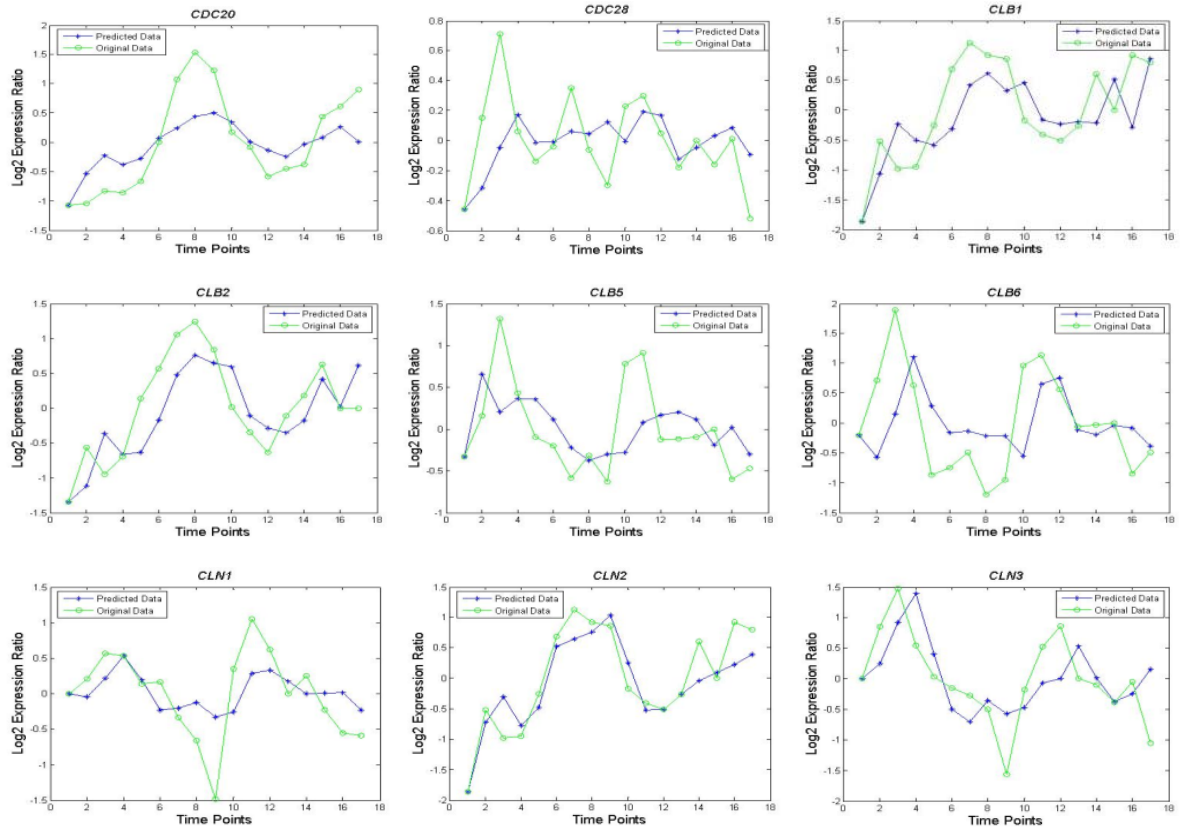


Figure 3.9 Original and predicted outputs of the testing set.

As shown in Table 3.2, among the 20 relations predicted by PSO, 13 of them concur with the known interactions obtained by PathwayStudio. Two gene self-regulations (*CLB1* and *CLB2*) were found by PSO, but not by PathwayStudio. One other relation is also found by PathwayStudio, but with reversed direction of regulation. We further compared the performance of the proposed method, NN trained with the backpropagation through time (BPTT) method [123], and dynamic Bayesian networks (DBNs) [124] (Table 3.3). We calculated the precision as TP/(TP+FP), where TP and FP denote true positive and false positive, respectively. The

computational time is the average over 10 runs for each method in the MATLAB It can be seen that PSO-NN is able to unveil more true relationships in this cyclin TRN, with less false positives and computational time than the other two methods. The results demonstrate that the proposed PSO-NN method is very promising in capturing the nonlinear dynamics of TRNs and unveiling the gene interaction relation. However, the limited sample size limited the application of PSO in larger TRN inference. The results for BPTT-NN and DBN are similar and they both consider the calculated mean and standard deviation and the weights to decide the existence of regulations. It is expected that the properties of the methods can be more effectively investigated with more data available. In the following sections, we will demonstrate the inference capability of PSO-NN method at modular level using the computational framework presented in Figure 3.1.

Table 3.2 Comparison of "true" interactions and predicted interactions by the PSO-NN method

Relation Type	Prior knowledge (PathwayStudio)	PSO-NN
Regulation	CLB6 ---> CLB5	CLB6 ---> CLB5
Regulation	CLN2 --+> cdc28	CLN2 --+> cdc28
Regulation	CLB6 --+> CLB2	CLB6 --+> CLB2
Regulation	CLB1 <+-- CLB5	CLB1 <+-- CLB5
Regulation	CLN2 ---> CLN3	--
Regulation	CLN1 ---> CLN3	CLN1 ---> CLN3
Regulation	cdc28 <+-- CLN3	cdc28 <+-- CLN3
DirectRegulation	CDC20 <+-- CLB2	--
DirectRegulation	CDC20 <--- cdc28	--
DirectRegulation	CLB6 --+> cdc28	CLB6 --+> cdc28
DirectRegulation	CLN2 <+-- CLN3	--
DirectRegulation	CLB5 --+> cdc28	CLB5 --+> cdc28
DirectRegulation	CLB2 --+> cdc28	CLB2 --+> cdc28
DirectRegulation	CLN1 <+-- CLN3	CLN1 --+> CLN3
DirectRegulation	CLN1 <+-- CLN2	CLN1 <+-- CLN2
DirectRegulation	CLN1 <+-- CLB5	--
DirectRegulation	CLB2 <+-- CLB5	CLB2 <+-- CLB5
DirectRegulation	cdc28 ---> CLN3	cdc28 ---> CLN3
DirectRegulation	CLB1 --+> cdc28	CLB1 --+> cdc28
DirectRegulation	CLN1 --+> cdc28	CLN1 <+-- cdc28
Self Regulation	--	CLB1 ---> CLB1
Self Regulation	--	CLB2---> CLB2

Table 3.3 Comparison of the performance of computational methods

Method	TP	FP	Precision	Computational time (sec)
PSO-NN	13	2	87%	213
BPTT-NN	9	4	69%	562
DBN	10	3	76%	654

3.4.2 Yeast Cell Cycle Dataset

3.4.2.1 Data Sources

Gene expression dataset: the yeast cell cycle dataset presented in [60] consists of six time series (*cln3*, *clb2*, *alpha*, *cdc15*, *cdc28*, and *elu*) expression measurements of the mRNA levels of *S. cerevisiae* genes. 800 genes were identified as cell cycle regulated based on cluster analysis in [60]. We used the *cdc15* time course data of the 800 genes since it has the largest number of time points (24). Missing values in the data were imputed using KNN imputation [97]. The expression pattern of each gene was standardized between 0 and 1.

Molecular interaction data: data of transcription factors and their target genes were extracted from the SCPD database [125], from the YPD database [126], and from recent publications on genome-wide experiments that locate binding sites of given transcription factors [45, 59, 100, 127]. For data extraction from the latter we used the same experimental thresholds as in the original papers. Protein-protein interaction data was extracted from the DIP database [27], from the BIND database [18], and from the MIPS database [128]. In total the molecular interaction dataset consisted of 8184 protein pairs connected by PPIs and 5976 protein pairs connected by PDIs.

Table 3.4 Candidate transcription factors among the yeast cell cycle related genes

Gene Name	ORF	Gene Name	ORF	Gene Name	ORF
ACE2	YLR131C	KAR4	YCL055W	RFA3	YJL173C
ADA2	YDR448W	MATALPHA1	YCR040W	RFC4	YOL094C
ARP7	YPR034W	MCM2	YBL023C	RFC5	YBR087W
ASH1	YKL185W	MCM3	YEL032W	RLF2	YPR018W
CAC2	YML102W	MCM4	YPR019W	RME1	YGR044C
CBF2	YGR140W	MCM5	YLR274W	SFG1	YOR315W
CDC45	YLR103C	MCM6	YGL201C	SMC1	YFL008W

Gene Name	ORF	Gene Name	ORF	Gene Name	ORF
CDC6	YJL194W	MCM7	YBR202W	SPT16	YGL207W
CHA4	YLR098C	MEC3	YLR288C	STB1	YNL309W
CRP1	YHR146W	MET28	YIR017C	STB5	YHR178W
CTF4	YPR135W	MIF2	YKL089W	STP2	YHR006W
EST1	YLR233C	MIG2	YGL209W	SUT1	YGL162W
FKH1	YIL131C	MSH2	YOL090W	SWI4	YER111C
GAT3	YLR013W	MSH6	YDR097C	SWI5	YDR146C
GCR1	YPL075W	NDD1	YOR372C	TAF2	YCR042C
HCM1	YCR065W	NRM1	YNR009W	TBF1	YPL128C
HHF1	YBR009C	ORC1	YML065W	TEA1	YOR337W
HHF2	YNL030W	PHD1	YKL043W	TEC1	YBR083W
HHO1	YPL127C	PLM2	YDR501W	TEL2	YGR099W
HHT1	YBR010W	PMS1	YNL082W	TOP3	YLR234W
HHT2	YNL031C	POG1	YIL122W	TOS4	YLR183C
HMLALPHA1	YCL066W	RAD5	YLR032W	VHR1	YIL056W
HOP1	YIL072W	RAD53	YPL153C	WHI5	YOR083W
HST3	YOR025W	RAD54	YGL163C	WTM1	YOR230W
HST4	YDR191W	RAP1	YNL216W	WTM2	YOR229W
HTA1	YDR225W	RDH54	YBR073W	YHP1	YDR451C
HTA2	YBL003C	RDS2	YPL133C	YOX1	YML027W
HTB1	YDR224C	RFA1	YAR007C	--	--
HTB2	YBL002W	RFA2	YNL312W	--	--

3.4.2.2 Gene Module Identification

We grouped 800 cell cycle-regulated genes into clusters by FCM, where genes with similar expression profiles are represented by a gene module. The optimal cluster number was determined by the proposed method in Chapter 2. The highest z score was obtained when the number of clusters was 34 by FCM clustering with optimal parameter m = 1.1573. The detailed clustering information is presented in Appendix A. We evaluated the resulting clusters through the GSEA method. All clusters except 10, 18, 21, 22, 25 and 26 are enriched in some GO

categories (Appendix B). We used these clusters as candidate gene modules in our subsequent analyses to reduce the search space for gene regulatory module inference.

3.4.2.3 Network Motif Discovery

Among the 800 cell cycle related genes, 85 have been identified as DNA-binding transcription factors (Table 3.4). Four NMs were considered significant: auto-regulatory motif, feed-forward loop, single input module, and multi-input module (shown in Figure 3.2). These NMs were used to build NN models for corresponding transcription factors.

3.4.2.4 Gene Regulatory Module Inference

Neural network models that mimic the topology of the NMs were constructed to identify the relationships between transcription factors and putative gene modules. The NN models were trained to select for all 85 transcription factors the downstream targets from the 34 gene modules. Table 3.5 presents the experimental results obtained for various numbers of generations that GA was used. The PSO generation for NN is set to 1000 [129]. As illustrated in the table, the minimum value of RMSE decreases as the number of generations increases. The minimum RMSE for GA generations 600 and 800 are 0.077 and 0.075 respectively. We chose 600 for generations of GA. Our inference method mapped all 85 transcription factors to the target gene modules and inferred the most likely NMs.

Table 3.5 The experimental results of GA-PSO with NN

GA generations	Average RMSE	Minimum RMSE
100	1.27	0.78
200	0.84	0.40
400	0.62	0.12
600	0.35	0.077
800	0.31	0.075

We evaluated the predicted gene regulatory modules for the following eight well known cell cycle related transcription factors: *SWI4*, *SWI5*, *FKH1*, *NDD1*, *ACE2*, *KAR4*, *MET28* and *RAP1*. Since the "true" TRN is not available, the accuracy of putative regulatory relationship was determined by searching known gene connections in databases. Based on the results of the NM module prediction, we collected literature evidences from SGD [130] and BIND [18] databases.

We examined the inferred relationships for each of the eight transcription factors. An inferred relationship is assumed to be biologically significant if the transcription factors are correlated with the biological functions associated with the critical downstream cluster(s). Figure 3.10 lists the significant relationships; the eight transcription factors yielded an average precision of NMs for four of these transcription factors were identified in Chiang et al. [114] together with other four transcription factors. The eight transcription factors in [114] yielded an average precision of 80.1%.

The regulatory relationships inferred by the proposed method are expected to correspond more closely to biologically meaningful regulatory systems and naturally lead themselves to optimum experimental design methods. The results presented in Figure 3.10 are verified from previous biological evidences. For example, *FKH1* is a gene whose protein product is a fork head family protein with a role in the expression of G2/M phase genes. It negatively regulates transcriptional elongation, and regulates donor preference during switching. To further investigate the possibilities that the predicted downstream gene clusters are truly regulated by *FKH1*, we applied the motif discovery tool, WebMOTIFS [131] to find shared motifs in these gene clusters. The results revealed that a motif called Fork_head, GTAAACAA, is identified as the most significant motif among these gene clusters [132]. This finding strongly supports our NM inference results. The details of the BSEA results are shown in Table 3.6. Another example is the FFL involving *SWI5*, *GAT3* and Gene Cluster 10. *SWI5* has been identified as the upstream regulator of *GAT3* [45, 59, 133]. Genes in cluster 10 are mostly involved in DNA helicase activity and mitotic recombination, both of which are important biological steps in the regulation of cell cycle. Although no biological evidences have shown that *SWI5* and *GAT3* are involved in these processes, there are significant numbers of genes in cluster 10 which are characterized (according to yeastract.com) as genes regulated by both transcription factors (24 for *GAT3* and 23 for *SWI5* out of 44 genes in cluster 10, respectively).

Table 3.6 Binding site enrichment analysis for gene modules identified in yeast cell cycle dataset

Cluster ID	Predicted motifs
1	bZIP
	HSF_DNA-bind
	Zn_clus

Cluster ID	Predicted motifs
2	E2F_TDP
	IRF
	Runt
4	Fork_head
	homeobox
	TBP
5	SRF-TF
	HLH
	Fork_head
6	E2F_TDP
	bZIP
	Myc_N_term
7	SRF-TF
	RHD
	bZIP_Maf
8	Fork_head
	myb_DNA-binding
	zf-C4
9	CBFB_NFYA
	HNF-1_N
	RHD
10	Fork_head
	HSF_DNA-bind
	PAX
11	bZIP_Maf
	zf-C4
	AP2-domain
12	E2F_TDP
	bZIP
	HSF_DNA-bind
13	TF_AP-2

Cluster ID	Predicted motifs
	Fork_head
	RHD
14	RHD
	bZIP
	zf-C4
15	TF_AP-2
	HLH
	Myc_N_term
16	homeobox
	IRF
	bZIP_Maf
17	CBFB_NFYA
	HLH
	bZIP
18	SRF-TF
	bZIP
	TF_Otx
19	HLH
	HNF-1_N
	SRF-TF
20	E2F_TDP
	HSF_DNA-bind
	zf-C4
22	Myc_N_term
	E2F_TDP
	HSF_DNA-bind

3.4.3 Human *Hela* Cell Cycle Dataset

3.4.3.1 Data Sources

Gene expression dataset: the human *Hela* cell cycle dataset [86] consists of five time courses (114 total arrays). RNA samples were collected for points (typically every 1-2 h) for 30 h (Thy-Thy1), 44 h (Thy-Thy2), 46 h (Thy-Thy3), 36 h (Thy-Noc), or 14 h (shake) after the

synchronous arrest. The cell-cycle related gene set contains 1,134 clones corresponding to 874 UNIGENE clusters (UNIGENE build 143). Of these, 1,072 have corresponding Entrez gene IDs, among which 226 have more than one mapping to clones. In total, 846 genes were used for TRN inference. We chose the Thy-Thy3 time course gene expression pattern for 846 genes, since it has the largest number of time points (47). Missing values in the data were imputed using KNN imputation [97]. The expression pattern of each gene was standardized between 0 and 1.

Molecular interaction data: PPIs were extracted from twelve publicly available large-scale protein interaction maps, seven of which are based on information from scientific literature literature-based, three on orthology information, and two on results of previous yeast two-hybrid (Y2H) analyses. The analysis was restricted to binary interactions in order to make consistent between Y2H-based interactions and the remaining maps. Detailed information about the twelve maps was shown in Table 3.7. To merge twelve interaction maps into one combination map, all proteins were mapped to their corresponding Entrez gene IDs. The PDI data was extracted from the TRANSFAC database [134]. In total the molecular interaction data consisted of 20,473 protein pairs connected by PPIs and 2,546 protein pairs connected by PDIs.

3.4.3.2 Gene Module Identification

A total of 846 genes associated with the control of cell cycle have been identified previously in *Hela* cells [86]. We further partitioned these genes into more specific functional groups by FCM [92]. The optimal value of m for the dataset used in this study was 1.1548 [94]. The highest z score was obtained with 39 clusters, indicating an optimal condition to reduce the search space for TRN inference. The detailed clustering result is presented in Appendix C. To evaluate the optimal clusters selected based on GO, GSEA was applied using the optimal value. The total set of genes involved in cell cycle regulation was further subdivided into 39 clusters. Of these clusters, 31 were clearly associated with GO categories that imply a more specific function that unifies the members of one but not other clusters, thereby establishing more direct relationships among certain smaller sub-groups of genes. For example, clusters 8 and 29 are both associated with pre-mitotic, mitotic and post-mitotic events (M-phase). However, members of cluster 8 is distinguished from the members of cluster 29 by virtue of their specific roles in chromosome doubling (DNA replication) and cytokinesis. Conversely, members of cluster 29 is distinguished from the members of cluster 8 by virtue of their specific roles in spindle fiber assembly and

disassembly. The detailed enriched biological categories for each cluster is presented in Appendix D.

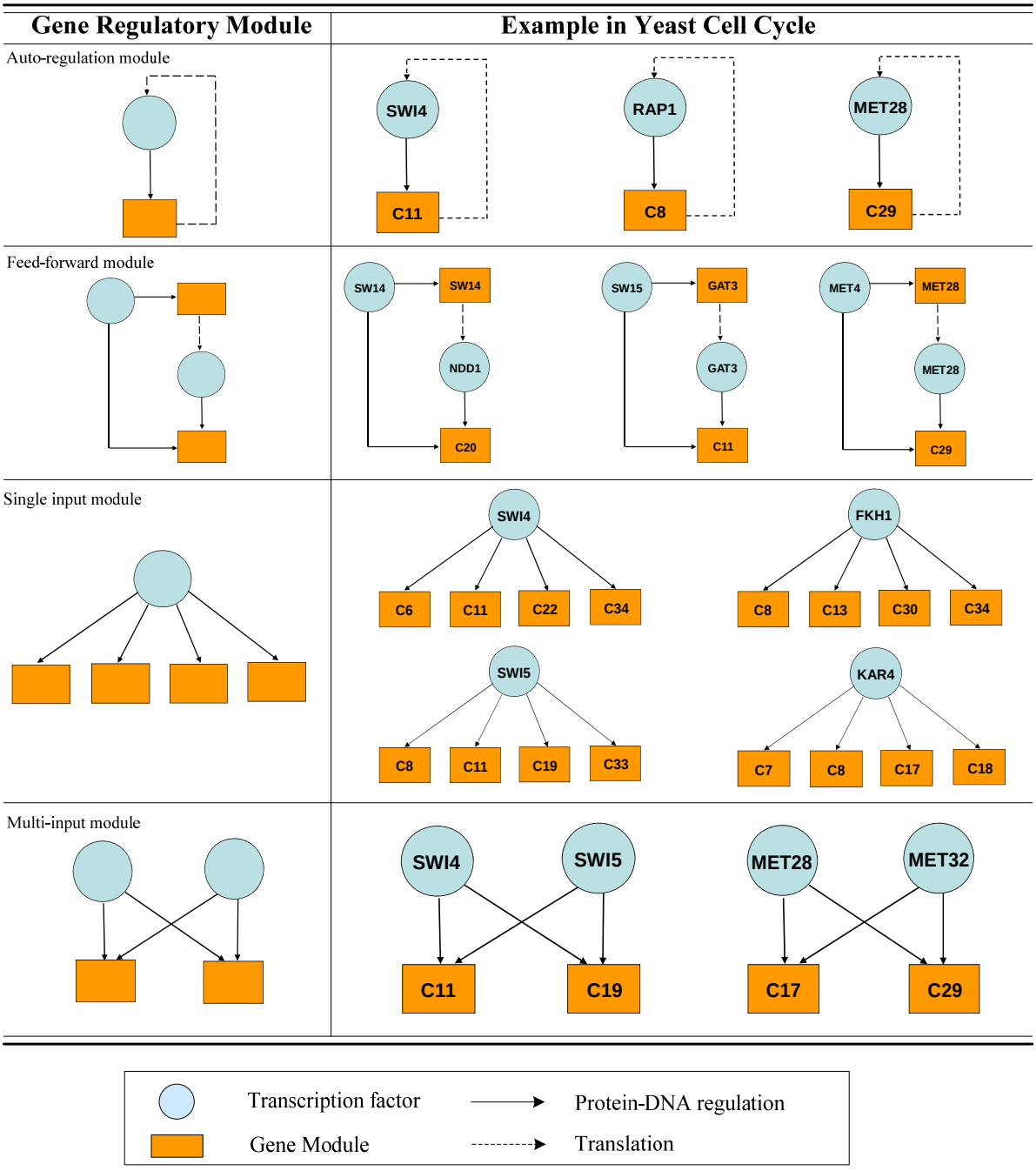


Figure 3.10 Predicted gene regulatory modules from eight known cell cycle dependent transcription factors in yeast cell cycle dataset. The left panel presents the four gene regulatory modules, and the right panel depicts inferred gene regulatory modules for eight known cell cycle dependent transcription factors.

Biological significance of these highly specific functional relationships, established by our clustering scheme, can further be extended in terms of relationships within the regulatory context. For instance, members of both gene modules 8 and 29 have been identified previously as direct downstream targets of *E2F* factors [135]. Similar relationships are established with other clusters such as gene module 32, which is comprised of genes with biochemical roles of a DNA ligase. Thus, the genes in gene module 32 are involved in processes associated with gap repair or Okazaki fragment processing during DNA replication and chromosome doubling. Previous studies have established that genes associated with this function are under the regulatory control of *E2F1* and *PCNA* [136].

Based on all these relationships, we demonstrated that one specific strength of the proposed method is its ability to distinguish genes that are related by function in a broad sense and sub-categorize them into highly specific (narrow) functional categories, resulting in the prediction of regulatory relationships that are consistent with biologically validated relationships.

Table 3.7 PPI network database information

Networks	Proteins	Interactions	Methods[a]	References	Version[b]
MDC-Y2H	1703	3186	Y2H-ASSAY	Stelzl et al. 2005 [25]	23.09.2005
CCSB-Y2H	1549	2754	Y2H-ASSAY	Rual et al. 2005 [24]	31.10.2005
HPRD	8788	32776	LITERATURE	Peri et al. 2003 [19]	22.08.2008
DIP	1085	1397	LITERATURE	Salwinski et al. 2004 [137]	01.03.2007
BIND	5286	7394	LITERATURE	Bader et al. 2003 [18]	01.03.2007
BioGrid	7953	24624	LITERATURE	Stark et al. 2006 [138]	22.08.2008
IntAct	7273	19404	LITERATURE	Hermjakob et al. 2004 [139]	22.08.2008
COCIT	3737	6580	TEXT-MINING	Ramani et al. 2005 [20]	18.11.2005
REACTOME	1554	37332	LITERATURE	Joshi-Tope et al. 2005 [140]	01.03.2007
ORTHO	6225	71466	ORTHOLOGY	Lehner and Fraser 2004 [21]	17.11.2005
HOMOMINT	4127	10174	ORTHOLOGY	Persico et al. 2005 [23]	01.06.2006
OPHID	4785	24991	ORTHOLOGY	Brown and Jurisica 2005 [22]	14.12.2005

The table displays the number of proteins and the number of interactions derived from each map.

[a]Methods refers to the approach taken from the construction of the corresponding map.

[b]Version describes the date of data downloaded for each dataset.

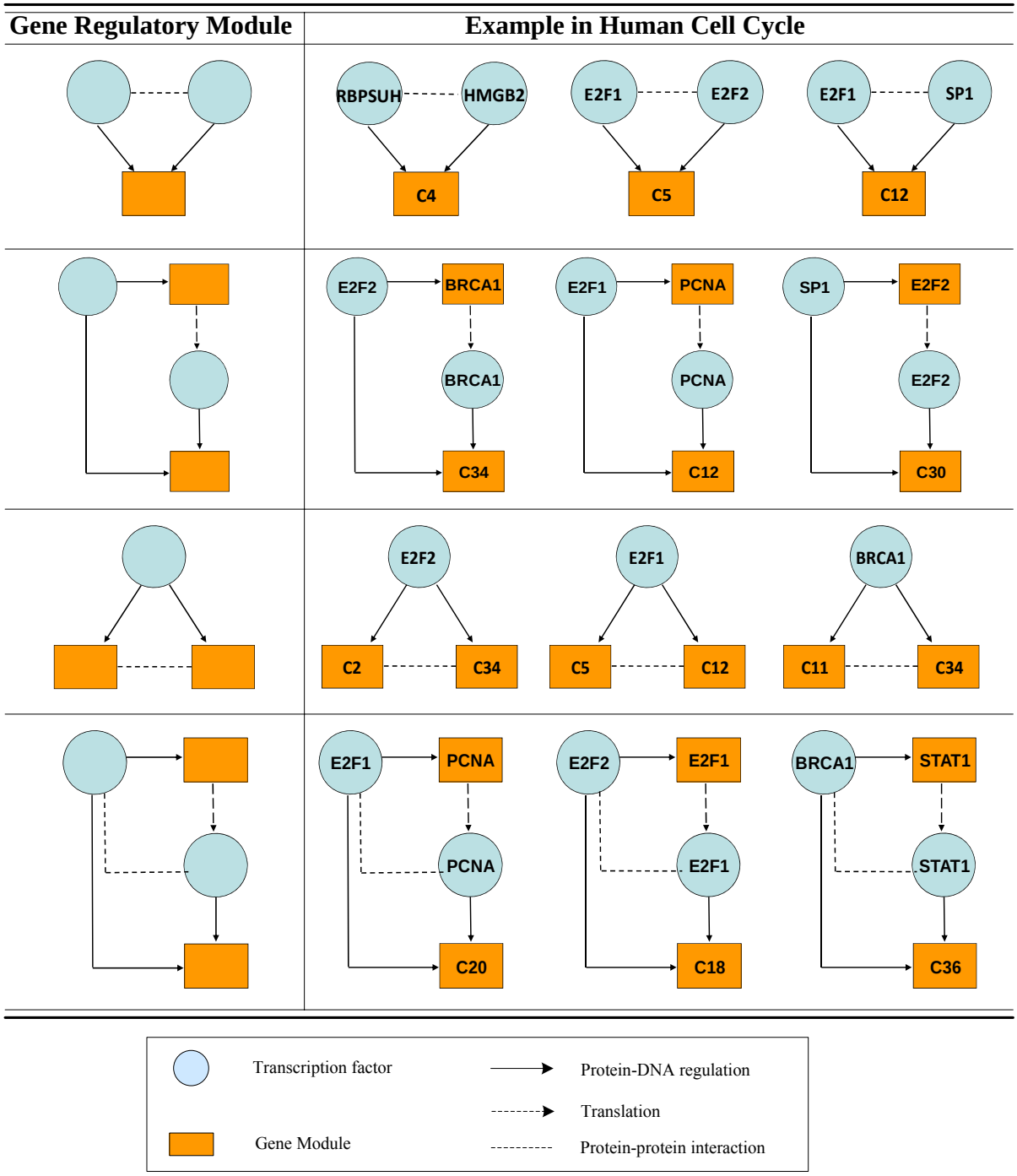


Figure 3.11 Predicted gene regulatory modules from known human cell cycle dependent genes. The left panel presents the four gene regulatory modules, and the right panel depicts inferred transcription factor-target gene relationships for eight cell cycle dependent transcription factors.

3.4.3.3 Network Motif Discovery

All genes with either direct or indirect roles in the regulation of transcription were first identified from the total set of 846 cell cycle associated genes according to GO categories that denote possible roles in transcription [29]. Candidate genes that remained after filtering other gene function categories are those that are assigned to the following putative functions: transcription factor activity (GO: 0003700), regulation of transcription (GO: 0061019), and transcription factor complex (GO: 0005667). Since GO information alone may not be sufficient to identify the genes with bona fide roles as transcription factors, we further filtered our list of candidate transcription factors by adding another layer of confirmatory information based on the results of PubMed database searches. This additional annotation allowed us to validate the GO classification of our candidate genes. The detailed descriptions of GO terms and specific roles in transcription of candidate transcription factors used in this study are presented in Table 3.8. Among the 846 cell cycle related genes, 46 were annotated with functions related to transcriptional regulation based on both GO and PubMed databases. These genes were considered as putative transcription factors.

In the microarray gene expression data, genes are often represented by multiple oligonucleotide probes. Genes represented by probe sets with larger variance were further considered in this study. We decomposed the collected human molecular interaction network into several NMs, with each NM potentially associated with a given transcription factor(s). A total of four NMs were found to be significant in the combined molecular interaction network (Figure 3.3), thus each transcription factor was assigned to at least one of these NMs.

3.4.3.4 Gene Regulatory Module Inference

The relationships between transcription factors and gene modules were determined based on NN models. For each of the four NMs (Figure 3.3), a suitable NN was built as we previously described [56]. The NN models were trained using the GA-PSO algorithm to find the downstream gene clusters for all 46 putative transcription factors. Associations between each transcription factor and 39 gene modules was determined by training the NN model that mimics the specific NM for a given transcription factor. Due to a reduction in the computational complexity (mapping between 46 transcription factors and 39 gene clusters instead of 846 the numbers of GA and PSO generations needed to reach the pre-specified minimum RMSE was

significantly reduced. The proposed inference method successfully assigned all 46 putative transcription factors to their target gene modules and inferred the most likely gene regulatory modules (see Figure 3.11 for representative gene regulatory modules).

The validity and accuracy of the network depicted by the gene regulatory modules are assessed by comparison with a network model constructed based on actual biological data. In the absence of such information, we performed an initial validation of the network by searching for known gene connections in databases. Based on the GRM prediction results, we collected literature evidence from the NCBI and TRANSFAC databases [28]. We reviewed each predicted GRM and examined the relationships between the transcription factor and its target gene module(s). Subsequent analysis was performed under the basic assumption that the inferred NM is more likely to be biologically meaningful if the transcription factors therein are correlated with the enriched biological functions in the downstream clusters.

Significant NMs resulting from the survey of available literature cell cycle dependent genes such as *E2F1, E2F2, SP1, BRCA1, STAT1, PCNA, RBPSUH*, and *HMGB2* are listed in Figure 3.11. Based on the combined information, the biological implication of the network is further explained. For instance, *E2F* is a transcription factor that plays a crucial role in cell-cycle progression in mammalian cells [141]. *E2F1*, which contains two overlapping *E2F*-binding sites in its promoter region, is activated at the G1/S transition in an E2F-dependent manner. *E2F2* interacts with certain elements in the *E2F1* promoter and both genes are involved in DNA replication and repair [142], cytokinesis, and tumor development [143]. According to the GSEA results, gene module 8 is enriched with genes involved in mitosis and cytokinesis, and gene module 34 is enriched with genes involved in several functional categories associated with tumor development. As shown in Figure 3.11, both gene module 8 and 34 are predicted to be regulated by *E2F1* and *E2F2*, and these results are in agreement with previous reports based on biological data [141, 143].

Our analysis predicts that *E2F1* and *PCNA* are components of the same network. Both of these genes are involved in the regulation of gene modules 32 and 34. The best understood molecular function of the *PCNA* protein is its role in the regulation of eukaryotic DNA polymerase delta processivity, which ensures the fidelity of DNA synthesis and repair [144]. However, recent studies have provided evidence that the *PCNA* protein also functions as a direct

repressor of the transcriptional coactivator p300 [145]. Another study shows that *PCNA* represses the transcriptional activity of retinoic acid receptors (*RAR*s) [146]. Thus, the involvement of these genes in the same network, as predicted by our network inference algorithm, is strongly supported by knowledge of regulatory relationships already established in experimental data. The results of our prediction are in agreement with these reports since both gene modules 8 and 32 are enriched with genes involved in DNA synthesis and regulatory processes.

We proposed three approaches to investigate further whether the genes predicted to be regulated by *E2F* genes in gene modules 8, 32 and 34 are validated in classical non-genome wide methods. First, we investigated how many "known" *E2F1* and *E2F2* targets are predicted by our proposed method. According to Bracken et al. [147], 130 genes were reviewed as *E2F* targets, 44 of which were originally identified by classical, non-genome-wide approaches. Since we restricted our analysis to the 846 cell cycle related genes, 45 genes matched the *E2F* target genes listed in ref. [147], 21 of which were known from studies using classical molecular biology analyses. The gene targets predicted by our method match 15 of 45 genes, all 15 of which are among those found originally using standard molecular biology experiments. One possible reason is that genome-wide approaches are usually highly noisy and inconsistent across different studies. The detailed information of these genes is presented in Table 3.9.

Second, we wanted to see whether our predicted gene target clusters are enriched in the corresponding binding sites for the transcription factors in their upstream region. For both *E2F1* and *E2F2*, 7 out of 17 genes in gene module 8 contain binding sites in their upstream regions as confirmed by data in the SABiosciences database (http://www.sabiosciences.com/chipqpcr search.php?app=TFBS).

Finally, we determined how many genes in the gene clusters have *E2F* binding sites. We applied WebMOTIFS [131] to find shared motifs in the gene modules predicted to the *E2F* targets using BSEA (Table 3.10). The results revealed that a motif called E2F_TDP, GCGSSAAA, is identified as the most significant motif among gene modules 2, 8, 29, 31, 32 and 34. Unfortunately, for gene modules 30 and 36 the number of genes in these clusters is too small for WebMOTIFS analysis. All these gene modules are predicted to the downstream targets of *E2F*. For instance, 43 out of 52 genes in gene module 2 have putative *E2F* binding sites in their

upstream regions. The detailed information of BSEA results is shown in Table 3.10. For those gene regulatory modules where two transcription factors are involved in, the downstream gene modules are found to be enriched in both the binding site sequence motifs. For instance, gene module 32 is enriched in both E2F_TDP and MH1 motifs, corresponding to the two transcription factors in the gene regulatory module: *E2F1* and *SP1*. These BSEA results strongly support our inference results.

We also performed an additional analysis of the results presented in Figure 3.11 using the Ingenuity Pathway Analysis (IPA) software (Ingenuity® Systems, www.ingenuity.com). This tool uses a knowledge base of over one million known functional relationships among proteins. Results of the analysis of the *BRCA1*, *STAT1*, *E2F1*, and *E2F2*-related networks are shown in Figure 3.12. These networks were reconstructed based of the putative transcription factors and genes in the predicted NMs. All the networks confirmed the inferred relationships between transcription factors and some of the genes in their downstream target clusters. For example, as shown in Figure 3.12 (A), *BRCA1* regulates two clusters that interact with each other and with the network reconstructed by IPA. Some genes in the clusters show indirect regulations through intermediate genes, such as *BRCA1* acting through *MLLT4* and *RAD18*. Figure 3.12 (B) depicts a predicted NM in which *BRCA1* and *STAT1* regulate all three genes in gene module 36. Figure 3.12 (C) shows a predicted NM with *E2F1* and *E2F2* interacting with each other and regulating the genes in gene module 34. Figure 3.12 (D) presents a motif where *E2F2* and *PCNA* bind together to activate expression of downstream genes in gene module 34. The notable consistency between IPA and the results from our method indicates that our approach can generate realistic hypotheses for further biological experimental validation.

3.5 Summary and Discussions

Reconstruction of TRNs is one of the major challenges in the post-genomics era of biology. In this chapter, we focused on two broad issues in TRN inference: (1) development of an analysis method that uses multiple types of data and (2) network analysis at the NM modular level. Based on the information available nowadays, we proposed a data integration approach that effectively infers the gene networks underlying certain patterns of gene co-regulation in yeast cell cycle and human *Hela* cell cycling. The predictive strength of this strategy is based on

Table 3.8 List of 46 transcription factors in human *Hela* cell cycle dataset

Gene Name	IMAGE ID	GO number	Associated GO category	Specific role during transcription
BCLAF1	**173309**	GO:0045449	Regulation of transcription	Transcriptional repressor
BRCA1	**241474**	GO:0045449	Regulation of transcription	Transcriptional activator/repressor
BRD8	**815287**	GO:0003700	Transcription factor activity	Transcriptional coactivator
CDK7	**1915416/**130242	GO:0045449	Regulation of transcription	Transcriptional activator
CIITA	**1536451**	GO:0045449	Regulation of transcription	Transcriptional activator/repressor
CTCF	**240367**	GO:0003700	Transcription factor activity	Transcriptional activator/repressor
DMTF1	**490728**	GO:0003700	Transcription factor activity	Transcriptional activator
DR1	**487797/**566760	GO:0045449	Regulation of transcription	Transcriptional repressor
DSCR1	**884462**	GO:0003700	Transcription factor activity	Transcriptional repressor
E2F1	**236142/**768260	GO:0003700	Transcription factor activity	Transcriptional activator
E2F2	**293331**	GO:0003700	Transcription factor activity	Transcriptional activator
FOXM1	**564803**	GO:0003700	Transcription factor activity	Transcriptional activator
GATA2	**149809/**135688	GO:0003700	Transcription factor activity	Transcriptional activator
GTF3C4	**780958/**291827	GO:0005667	Transcription factor complex	Transcriptional activator
HCFC1	**344049**	GO:0003700	Transcription factor activity	Transcriptional activator/repressor
HIF1A	**897806**	GO:0003700	Transcription factor activity	Transcriptional activator
HMG20B	**878184**	GO:0003700	Transcription factor activity	Transcriptional repressor
HMGB2	**1842250/**363103	GO:0003700	Transcription factor activity	Transcriptional repressor
ILF2	**242952**	GO:0045449	Regulation of transcription	Transcriptional activator
KDM5B	**838829**	GO:0003700	Transcription factor activity	Transcriptional repressor
KLF6	**510381**	GO:0003700	Transcription factor activity	Transcriptional activator
KLF9	**302549**	GO:0003700	Transcription factor activity	Transcriptional activator/repressor
MAPK13	**590774**	GO:0045449	Regulation of transcription	Transcriptional repressor
MNT	**809731**	GO:0003700	Transcription factor activity	Transcriptional repressor
NCOA3	**197520**	GO:0045449	Regulation of transcription	Transcriptional coactivator
NFE2L2	**884438**	GO:0003700	Transcription factor activity	Transcriptional activator
NFIC	**1455463/**265874	GO:0003700	Transcription factor activity	Transcriptional activator
NR3C1	**271198**	GO:0003700	Transcription factor activity	Transcriptional coactivator
NR5A2	**245517**	GO:0003700	Transcription factor activity	Transcriptional activator
PCNA	**43229/**789182	GO:0006275	Regulation of transcription	Transcriptional repressor
PHTF2	**30114**	GO:0045449	Regulation of transcription	Transcriptional activator/repressor

PKNOX1	**1947972**	GO:0003700	Transcription factor activity	Transcriptional activator
PTTG1	**2018976/**781089	GO:0003700	Transcription factor activity	Transcriptional activator/repressor
RBPJ	**845502**	GO:0045449	Regulation of transcription	Transcriptional repressor

the combined constraints arising from multiple biological data sources including time course gene expression data, combined molecular interaction network data, and GO category information.

This computational framework allows us to fully exploit the partial constraints that can be inferred from each data source. First, to reduce the inference dimensionalities, the genes were grouped into clusters by FCM, where the optimal fuzziness value was determined by statistical properties of gene expression data. The optimal cluster number was identified by integrating GO category information. Second, the NM information established from the combined molecular interaction network was used to assign NM(s) to a given transcription factor. Once the NM(s) for a transcription factor was identified, a hybrid GA-PSO algorithm was applied to search for target gene modules that may be regulated by that particular transcription factor. This search was guided by the successful training of a NN model that mimics the regulatory NM(s) assigned to the transcription factor. The effectiveness of this method was illustrated via well-studied cell cycle dependent transcription factors (Figure 3.10 and 3.11). The upstream BSEA indicated that the proposed method has the potential to identify the underlying regulatory relationships between transcription factors and their downstream genes at the modular level. This demonstrates that our approach can serve as a method for analyzing multi-source data at the modular level.

Compared to the approach developed in [148], our proposed method has several advantages. First, our method performs the inference of TRNs from genome-wide expression data together with other biological knowledge. It has been shown that mRNA expression data alone cannot reflect all the activities in one TRN. Additional information will help constrain the search space of causal relationships between transcription factors and their downstream genes. Second, we decompose the TRN into well characterized functional units - NMs. Each transcription factor is assigned to specific NM(s), which is further used to infer the downstream target genes. We not only reduce the search space in the inference process, but also provide experimental biologists the regulatory modules for straightforward validation, instead of one whole TRN containing thousands of genes and connections as is often generated by IPA. Third, we group the genes into

functional groups that are potentially regulated by one common transcription factor. The proposed approach reduces the noise in mRNA expression data by incorporating gene functional annotations.

In summary, we demonstrate that our method can accurately infer the underlying relationships between transcription factor and the downstream target genes by integrating multi-sources of biological data. As the first attempt to integrate many different types of data, we believe that the proposed framework will improve data analysis, particularly as more data sets become available. Our method could also be beneficial to biologists by predicting the components of the TRN in which their candidate gene is involved, followed by designing a more streamlined experiment for biological validation.

Table 3.9 *E2F* target genes inferred the proposed method in human *Hela* cell cycle dataset

E2F Target Genes Listed in [147] and Present in the List of Genes Analyzed	Included in the Gene Modules Identfied (Y/N)	E2F Target Genes Listed in [147] and Present in the List of Genes Analyzed	Included in the Gene Modules Identfied (Y/N)
AURKB	N	E2F2	Y
BARD1	N	FEN1	N
BMP2	N	MAD2L1	N
BRCA1	N	MCM2	Y
BUB1	N	MCM4	Y
BUB1B	N	MCM5	Y
BUB3	N	MCM6	Y
CASP3	N	MSH2	N
CCNA2	Y	NPAT	Y
CCND1	Y	ORC1L	N
CCNE1	Y	PCNA	N
CCNE2	N	PMS2	N
CDC2	Y	PRC1	N
CDC20	N	RAD51	N
CDC25A	Y	RAD54L	N
CDC45L	Y	RFC2	N
CDC6	Y	RFC4	N
CDKN2C	Y	RPA2	N
CDKN2D	N	RRM1	N
CENPE	N	RRM2	N
CKS2	N	TOP2A	N
DHFR	N	TTK	N
E2F1	Y	TYMS	N

Table 3.10 Binding site enrichment analysis for gene modules identified in human *Hela* cell cycle dataset

Cluster #	Sequence logo[a]	Binding domain (Pfam ID)	Corresponding transcription factor	Conserved binding motif[b]
Cluster 2		E2F_TDP (PF02319)	E2F1 E2F2	GCGssAAa
Cluster 8		E2F_TDP (PF02319)	E2F1 E2F2	GCGssAAa
Cluster 29		zf-C4 (PF00105)	BRCA1	TGACCTTTGAC Cyy
		E2F_TDP (PF02319)	E2F1 E2F2	GCGssAAa
Cluster 31		HMG_box (PF00505)	HMGB2	AACAAwRr
Cluster 32		MH1 (PF03165)	SP1	TGGc.....gCCA
		E2F_TDP (PF02319)	E2F1 E2F2	GCGssAAa
Cluster 34		E2F_TDP (PF02319)	E2F1 E2F2	GCGssAAa
Cluster 38		zf-C4 (PF00105)	BRCA1	TGACCTTTGAC Cyy
		E2F_TDP (PF02319)	E2F1 E2F2	GCGssAAa

[a]Sequence logos represent the motif significantly overrepresented in individual gene cluster associated with their predicted upstream transcription factors, according to the WebMOTIFS discovery algorithm [131]. Individual base letter height indicates level of conservation within each binding site position.

[b]Conserved binding motifs are the conserved binding sequences used in the WebMOTIFS discovery algorithm.

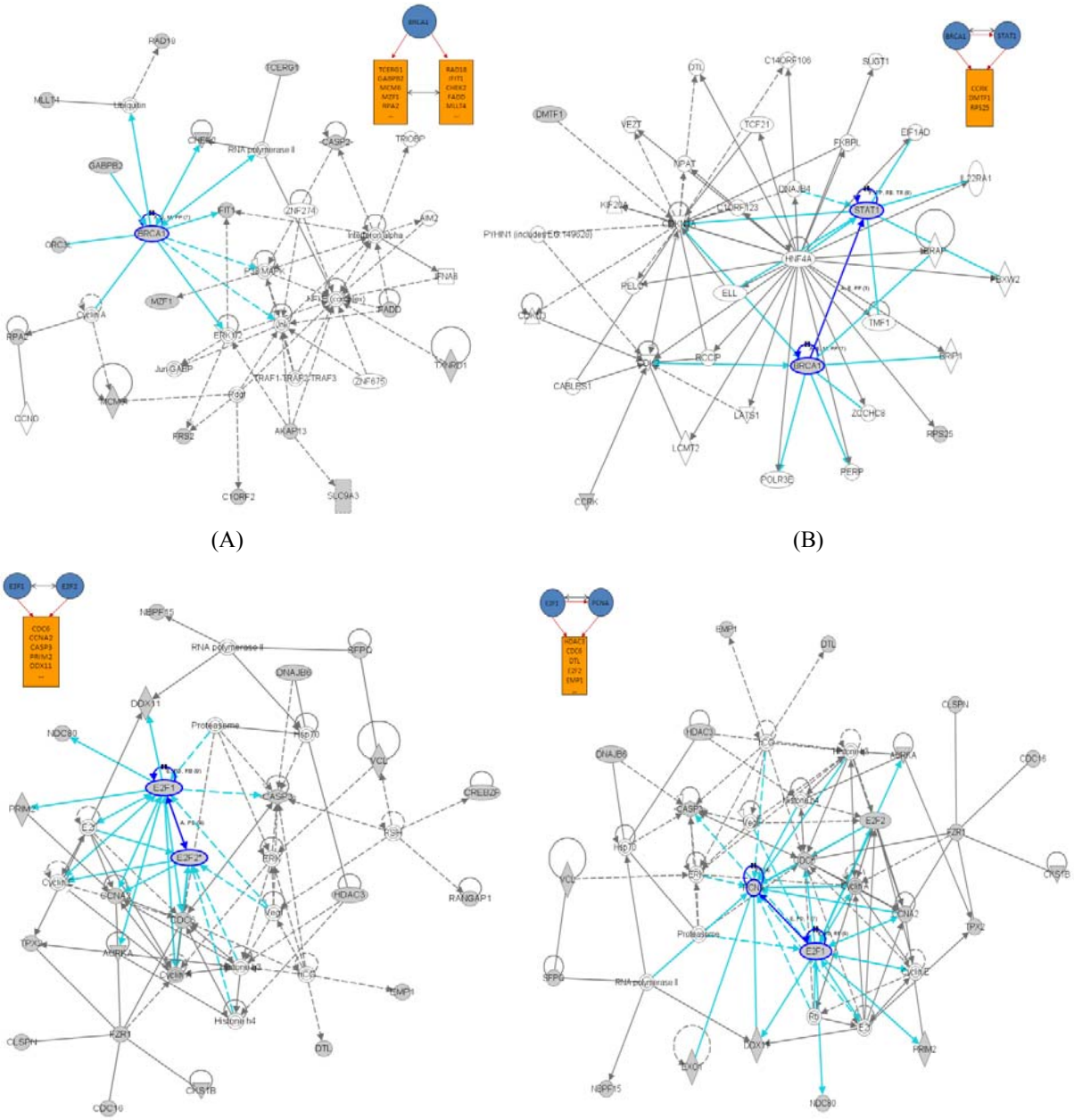


Figure 3.12 Figure Ingenuity analysis for gene regulatory modules. (A) BRCA related gene regulatory modules; (B) BRCA1 and STAT1-related gene regulatory modules; (C) E2F related gene regulatory modules; (D) E2F and PCNA-related gene regulatory modules. Shaded genes are genes identified in the gene regulatory module and others are those associated with the identified genes based on IPA analysis.

4 Module-based Disease Biomarker Discovery

Computational identification of disease biomarkers capable of withstanding independent validation efforts is still an open challenge in cancer research. Gene biomarkers for classification of cancer patients are typically identified by scoring or ranking individual genes with regard to their capacity to distinguish between clinically relevant classes. However, the reliability and reproducibility of these gene biomarkers have been challenged because of the biological heterogeneity and noise within and across patients. The availability of genome-wide molecular interaction network data opens up new possibilities to discover potential biomarkers and elucidate cancer-related complex mechanisms at network level. In this chapter, we propose a module-based biomarker discovery framework to integrate molecular interaction network information and gene expression data to identify biomarkers, not as individual genes but as functional modules. To select reliable biomarker sets across different studies, a hybrid method combining group feature selection with ensemble feature selection is proposed. First, a group feature selection method is used to extract the functional modules (subnetworks) with discriminative power between subclasses of diseases or between cases and controls. Then, an ensemble feature selection method is used to select the optimal module biomarker sets, in which a double-validation strategy is applied. The first patient dataset is used for within-dataset validation strategy. The trained classifier built by the first dataset is then evaluated by a second independent patient dataset, called cross-dataset validation. The ensemble method allows combining features selected from multiple runs with various data subsampling to increase the reliability and classification accuracy of the final selected biomarker set. The results from real datasets demonstrate that the identified module biomarkers: i) can achieve higher prediction accuracy in independent validation datasets; ii) are more reproducible than individual gene markers; iii) improve the biological interpretability of results; and iv) are enriched in cancer "disease drivers". The proposed approach is illustrated through four patient datasets related to breast cancer metastasis.

4.1 Introduction

In the last few decades, high-throughput genomic and proteomic techniques have generated a large number of diagnostic, prognostic and predictive molecular signatures related to many diseases [98, 149-154]. Traditional biomarker discovery approaches for disease classification are typically selected by scoring individual genes for how well their expression patterns discriminate between different subclasses of disease or between cases and controls. However, there are several disadvantages of these approaches including the following:

1) Lack of adequate biological interpretation: the genes selected by traditional biomarker discovery methods are mainly "downstream" reflectors of the perturbations defining clinical outcomes through the complex interplay of interaction networks. They may not directly account for the activity, perturbations or roles that disease-related cellular networks show [155].
2) Oversimplified assumption of gene independence: traditional biomarker discovery approaches assume gene independence, i.e., gene biomarkers are typically selected independently, although proteins are well known to function coordinately within protein complexes, signaling pathways, and higher-order cellular processes. Thus, the resulting classifiers may contain biomarker genes with redundant information that may lead to decreased classification performance.
3) Low reproducibility/reliability: biomarker sets identified from different labs share very few genes in common. This is well illustrated by two prominent studies of survival prediction in breast cancer. van't Veer et al. [156] generated a list of 70 genes from 96 patient samples, which were subsequently tested successfully on a larger cohort of 295 patients [149]. Wang et al. [157] analyzed the gene expression profiles of 286 patients and reported a gene biomarker set of 76 genes. Each gene set was trained and tested within its own samples and achieved good prediction performance. However, these two gene biomarker sets share only three genes. As a result, the predictive power of a classifier developed from one study could not be adequately reproduced when testing it on samples of another study, although both studies contain patients with similar phenotypes. Cellular heterogeneity within tissues and genetic heterogeneity across

patients in complex diseases (e.g., breast cancer) may weaken the discriminative power of individual genes, even within a clinically homogeneous patient group [158].

4) Inadequate focus on genes that are "disease drivers": oncogenes and tumor suppressors are disease drivers whose mutations result in a detrimental change of function that leads to cancer. These genes are generally more conserved than other proteins, and tend not to be most highly differentially expressed between different clinical groups of patients [62]. These genes would not be selected by traditional statistical ranking methods, such as *TP53* and *MYC*. However, their expression patterns are more stable in patients of the same clinical subgroups and more robust across different studies. Search for biomarkers that may represent upstream regulators with potential causal roles in the determination of differential phenotypes may help us define more reliable and reproducible biomarker sets.

The above limitations of traditional biomarker discovery approaches have received great attention by the community of cancer research [158-161]. We argue that the fundamental reason for these limitations is that these traditional biomarker identification methods lead to genes whose roles are mostly "passengers" rather than "drivers" of the phenotypic differences between sample groups (e.g., poor versus good outcomes). Regulatory networks often act as amplification cascade, where highly differentially expressed genes tend to be further downstream from the somatic or inherited determinants of the clinical outcomes. Since the regulatory networks comprise the complex interactions of multiple potential casual factors and sources of biological noise [30], these downstream genes are more prone to be most unstable across and within samples. On the other hand, oncogenes and tumor suppressors are generally not the most differentially expressed genes although they may show an outlier behavior in some samples [162]. The biomarkers enriched in these disease drivers may represent upstream regulators with potential causal roles in the determination of differential phenotypes, which will improve the reliability and reproducibility of the prediction model in unknown samples.

The organization of this chapter is as follows. Section 4.2 briefly reviews related systems approaches for biomarker discovery. Section 4.3 discusses the integrative network analyses of cancer-related genes. Section 4.4 introduces the proposed module-based biomarker discovery approach. In Section 4.5, the proposed methodology for module biomarker discovery and

classification evaluation is illustrated with the case studies of breast cancer. Section 4.6 is devoted to the summary and discussions.

4.2 Review of Related Methods for Network-based Biomarker Discovery

Chuang et al. [69] were among the first to show that the network-based biomarker discovery approach offers promising results toward finding better biomarkers for cancer prognosis. This novel strategy was based on the identification of protein interaction subnetworks with coherent expression patterns of their component genes, which distinguishes the samples of patients that developed distant metastasis after surgery from those that did not. This was achieved by overlaying PPI data on the gene expression profiles, generating combined activity scores for subnetworks across all patient samples, and computing the discriminative score based on the mutual information between activity scores and metastasis potential. The resulting subnetwork biomarkers were more reproducible than individual biomarker genes, which were selected without protein interaction information. These subnetwork markers also achieved higher accuracy in classification metastatic versus non-metastatic tumors. In addition, the vast majority of these biomarkers contained highly interconnected proteins encoded by genes that were not discriminative themselves. These included a significant number of previously identified breast cancer susceptibility genes, such as *TP53*, *BRCA1*, and *ERBB2*. This approach offered the advantage of identifying more reliable biomarkers significantly enriched in common biological processes. However, the subnetwork biomarkers selected from the two data sets overlapped only very partially, and contained only a minority of the cancer susceptibility genes. One possible reason is that the level and modulation of expression of many functionally relevant genes are of low magnitude and beyond the detection capabilities of current microarray technologies [163]. Also, the PPI network used was far from being complete [164], which hindered from detecting some potential subnetwork biomarkers.

Partially based the works of Chuang et al. [69], Lee et al. [165] investigated another approach based on pathway activities inferred for each patient. An activity level was summarized from the gene expression levels of its condition-responsive genes (CORGs), defined as the subset of genes in the pathway whose combined expression delivers optimal discriminative power for the disease phenotypes. The goal was to integrate the expression levels of the CORGs to estimate “pathway

activities" and perform patient classification. The resulting classifier achieved better performance than classifiers that were based on individual gene expression, as well as other related techniques for the quantitative characterization of biological pathways [166, 167]. However, the majority of human genes have not been assigned to a definitive pathway. Limited coverage of human pathway information excludes potential disease biomarker discovery not included in these pathways. In addition, the proteins, which make up one individual pathway, rarely operate in isolation but "cross-talk" with other pathways' proteins to process signal information [168]. A network-level view of signaling events will be more helpful in identifying disease related activities through cellular networks.

Edelman et al. [169] proposed an integrative hierarchical approach to analyze multiple biological pathway relationships to model cancer progression. Different from the approach in Lee et al. [165], Edelman et al. identified significant biological pathways in advance. They first detected pathways implicated in distinct progression phases, and inferred pathway interaction networks among these relevant pathways over the steps in tumor progression by regularized multi-task learning: from normal tissue to primary tumor, from primary tumor to metastasis. After that, they applied learning gradients and inverse regression to refine the relevant pathways to those genes most differentially expressed over progression, and a gene interaction network was constructed for these refined gene sets. The transformation of expression data sets into the space of enrichment scores of gene sets extended previous research to gain insight into disease processes at the pathway level. The case studies in prostate cancer and melanoma indicated findings that are consistent with previous research. However, their approach was limited to well-studied pathways, such as the *P53* pathway and the *RAS* pathway. For those genes not included in such pathways, it is yet unrealistic to understand their roles in the mechanisms of tumorgenesis by such approaches.

Hua et al. demonstrated another integrative approach to investigating a transcriptional regulatory cascade involved in the progression of breast cancer using gene expression, ChIP-chip data, PPI and epigenomic information [170]. A candidate biomarker *H2A.Z*, associated with breast cancer patient survival, was detected and validated by different experimental techniques. Hua et al. demonstrated that the expression of *H2A.Z* can be used to estimate metastasis

occurrence. This integrative framework enabled an accelerated identification of a molecule linked to breast cancer progression, which can be applied to a wide range of complex diseases.

Different from the approaches reviewed above, Mani et al. [171] proposed an integrative framework to discover disease biomarkers based on the analysis of perturbed or deregulated interactions (network edges) instead of perturbed genes (network nodes). They first generated a comprehensive network of interactions in B cells using different types of interaction evidence, including PPIs, PDIs, and signaling pathways. Then, they searched for interactions in this network that showed perturbations at the gene correlation level related to specific phenotypes. Gene pairs with significant gain and loss of correlations between control and phenotype classes were detected. Finally, such gene sets were combined into a statistical analysis that identified subsets of genes with a large number of deregulated interactions in their neighborhoods. However, since this method is based on the estimation of quantitative dependencies between genes, relatively larger data sets are needed.

Another systems biology approach was applied to prion disease by Hwang and his colleagues [172]. In their study, network dynamics referred to specific time-dependent changes at the mRNA level that were mapped onto PPI networks. Global gene expressions were measured in the brains of different mouse strain/prion strain combinations for multiple time points during the progression of the disease. Differentially expressed genes with potential crucial rules in prion disease progression were detected by an integrative statistical analysis of gene expression profiles. These genes were then mapped to different PPI networks associated with specific disease processes, and modules that may be involved in genetic effects on incubation time and in prion strain specificity were further characterized.

In Chang et al. [173], a module-based approach was reported for discovering module-based signatures relevant to disease detection and drug response estimation. Chang et al. started by looking for biologically relevant modules of gene expression data using statistical factor This method allows the deconstruction of the original gene expression data set into a set of informative components of data variation linked to different subsets of genes. These module signatures were analyzed in the context of different indicators of pathway activity as measured in the gene expression data sets. Chang et al. demonstrated that these signaling modules can dissect the complexity of oncogenic states that define disease outcomes as well as response to pathway-

specific therapeutics. The insights obtained from the latter may be of particular importance in the design of personalized therapeutic strategies based on the identification of patients most likely to benefit from a treatment.

Lim et al. succeeded to detect candidate biomarkers by identifying "upstream regulators" causally related to the phenotypic differences [174]. The transcription factors were determined if they caused the up/down-regulation of genes linked to poor outcome in patient samples through ARACNe algorithm [175]. The inferred sets of "master regulators" were shown to be more powerful and robust than the signatures proposed by original investigations based on standard gene-based analysis.

Most approaches reviewed here aim primarily to identify disease biomarkers based on integration of gene expression data and interaction network information. The results have allowed researchers to gain new knowledge of: i) network-based mechanisms underlying complex common diseases; ii) strategies to improve disease classification performance; iii) approaches to enhance robustness in biomarker selection and classification across independent data sets; and iv) the identification of potential disease drivers or causal agents at different levels of biological organization. However, no methods have succeeded in integrating different types of interaction network information into biomarker discovery, i.e., PPIs, PDIs, and signaling networks. Different types of interaction networks reflect cellular activities at different levels. Integrative analysis of these networks will help us better understand the information exchange between genes/proteins and find significant changes between different disease statuses in a network context.

4.3 Integrative Network Analysis of Cancer-associated Genes

The extraction and interpretation of biological insights of the differentially expressed genes in high-throughput gene expression profiling studies are challenging tasks. Gene set enrichment analysis is widely used to detect significant functional categories in these genes, but the in depth relationships between genes in different functional categories cannot be easily illustrated. Also, a particular phenotype is the result of collaborations of a group of genes, which do not necessarily belong to the same functional category. Therefore, integration of microarray gene expression data sets into interaction networks could help analyzing and interpreting the biological

significance of the genes in a network and their gene-interdependent context. It has been shown in many studies that different types of the interaction network are all "scale-free" networks: a small group of nodes act as highly connected hubs (high degree), whereas most nodes have only a few links (low degree). In PPI networks, hub proteins are involved in a large number of interactions, meaning that these proteins will take part in many biological processes and therefore would have higher dynamics in expression. In PDI networks, hub genes are usually global transcription factors, which govern a large number of genes in response to internal and external signals. In signaling networks, hub proteins are focal nodes that are shared by many signaling pathways, called "information exchanging and processing center". The integration of known cancer genes onto these interaction networks indicate that cancer genes are enriched in network hub proteins. In [176], Johsson and Bates mapped known mutated cancer genes onto a human protein interaction network constructed from the entire human genome using an orthology-based method. These cancer proteins had on average twice as many interaction partners as other proteins in the network, which implied that these cancer genes is evolutionarily conserved. Other studies [177-179] also confirmed that proteins whose mutation results in a detrimental change of function that leads to cancer may generally be more conserved than other proteins.

To investigate the cancer related activities in an interaction network context, first we need to understand that interaction networks are complex systems in which a gene does not independently perform a single task. Instead, individual genes tend to collaborate to carry out some specific biological function, in which these genes are called a functional module. We assume that a complex network can be broken up into many small but functional modules. For instance, interaction networks can be decomposed into NMs. Cancer-associated genes were demonstrated to be enriched in particular NM types, called hotspots in the mammalian cellular signaling network [180]. It was suggested that some regulatory NMs are critical to induce cancer or metastasis and these genes may work together to govern cell behaviors. These hotspots are potentially biomarker clusters or drug target clusters for curing cancer. If a cancer-related gene is mutated in one phenotype, this mutation will influence its surrounding interaction partners at modular level. As discussed above, the cancer-related genes are usually upstream disease "drivers", which tend not to be highly differentially expressed compared their downstream

"passenger" genes. However, the downstream genes with most differentially expression patterns are most prone to be unstable across samples because of the complex interplay of interaction network, especially in cancer. Systems approaches to biomarker discovery using functional modules instead of individual genes will improve the reliability of predictions using independent data sets, as these functional modules include disease "drivers", whose expression patterns are more stable in independent data sets although they are not mostly highly differentially expressed.

The availability of large PPI, PDI and signaling pathway data enables new opportunities for elucidating modules involved in major diseases and pathologies [66]. Several approaches have been demonstrated to extract the relevant functional modules based on coherent expression patterns of their genes [67, 68]. However, PPIs, PDIs, and signaling pathways are typically analyzed separately in previous studies [67-69], which hides the full complexity of the cellular circuitry since many processes involve combinations of different types of interactions. In this chapter, we proposed a module-based approach to identify module biomarkers by integrating interaction network data and patient gene expression profiles. The proposed method could be used to identify genetic alteration and to predict the likelihood of disease status in unknown samples. The biomarkers here are not encoded as individual genes or proteins, but as modules of interacting proteins within a large-scale human interaction network. The proposed method has several advantages over previous analyses of differential expression. First, the resulting module biomarkers provide models of the molecular mechanisms underlying disease mechanisms. Second, module-based classification achieves higher accuracy in prediction, which is ascertained by selecting biomarkers from one data set and applying them to a second independent validation data set. Third, the identified module biomarkers are likely to be more reproducible between different disease-related experiments than individual biomarker genes selected without network information. Finally, it provides the capability to detect genes with known disease mutations that are typically not detected through gene-based differential expression analysis. These uniquely identified genes are called "disease drivers" that are causally responsible for the determinations of differential phenotypes.

4.4 Materials and Methods

Figure 4.1 illustrates the steps that we propose to identify disease module biomarkers by integrating gene expression profiles and interaction network information. In the following sections, we briefly describe each of the steps.

4.4.1 Gene Expression Data

We obtained three mRNA expression datasets from three breast cancer studies [149, 157, 181] and one in house dataset. We divided these datasets into two groups: (1) prognosis group and (2) endocrine treatment prediction group. The prognosis group includes the van de Vijer and the Wang datasets that consist of patients with either poor or good outcomes. Poor outcome is defined as all patients with time of metastasis within five years of surgery, and good outcome as those with time of metastasis greater than or equal to five years after surgery. The endocrine treatment prediction group includes the Loi and our in house datasets consisting of patients with either early recurrence or non-recurrence. Early recurrence is defined as patients with recurrence within three years of endocrine treatment, and non-recurrence refers to those with time of recurrence greater than fifteen years after endocrine treatment. Table 4.1 presents the number of patients in each dataset and the microarray platform used to generate gene expression data. Since the four studies were performed on different microarray platforms, we restrict our analysis to the common genes present in all datasets. For simplicity, we used the terms “gene” and “protein” interchangeably in this work.

We normalize the expression of each gene across all samples in every dataset separately. For the dataset generated by Agilent platform, we use log ratio (base 2) between the measured and control samples. For datasets generated by Affymetrix chips, we use log (base 2) to transform the original expression values of each gene in each array, and normalize the log-space gene expression values by

$$g_{ij} \rightarrow log_2(g_{ij}) - log_2(\overline{g_i}) = log_2\left(\frac{g_{ij}}{\overline{g_i}}\right) \tag{4.1}$$

where g_{ij} is the intensity of gene i on a particular sample j, and $\overline{g_i}$ is the mean intensity of gene g_i over all samples. This normalization mimics a two channel microarray where the reference channel is a pool of all samples under consideration [182].

Table 4.1 The four datasets used in method evaluation

Name	Microarray platform	Number of samples
van de Vijver dataset	Agilent oligonucleotide Hu25K	Poor outcome: 78 samples Good outcome: 217 samples
Wang dataset	Affymetrix HG-U133a	Poor outcome: 106 samples Good outcome: 180 samples
Loi dataset	Affymetrix HG-U133	Early recurrence: 12 samples Non recurrence: 12 samples
In house dataset	Affymetrix HG-U133	Early recurrence: 24 samples Non recurrence: 40 samples

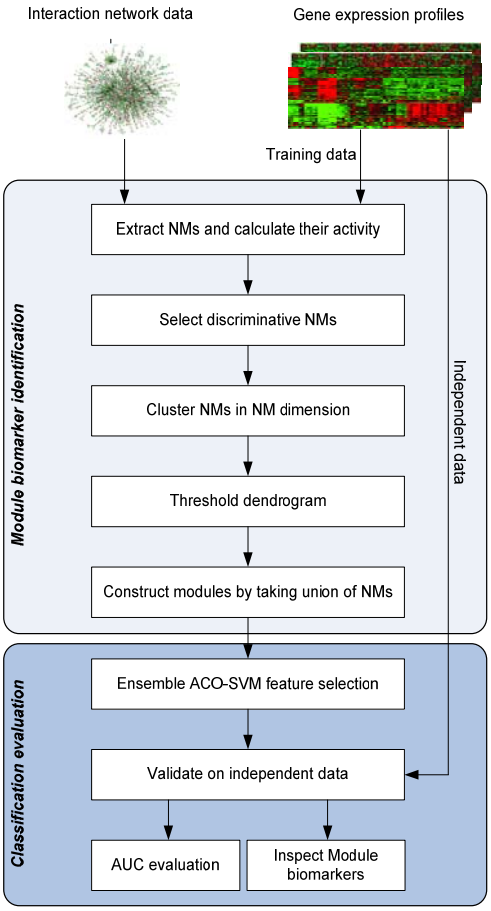


Figure 4.1 Schematic overview of module-based biomarker discovery.

4.4.2 Molecular Interaction Network Data

Protein-protein interaction data were extracted from eight protein interaction databases [18, 19, 27, 139, 140, 183-185] and two high-throughput yeast two-hybrid studies [24, 25]. Protein-DNA interaction data were extracted from the TRANSFAC database [28]. Signaling pathway data were extracted from the following three sources: i) the most comprehensive manually curated signaling pathway database, BioCarta (http://www.biocarta.com/); ii) a literature-mined signaling network [186]; and iii) 10 manually curated signaling pathways for cancer from the Cancer Cell Map (http://cancer.cellmap.org/cellmap/). To construct a corresponding human interaction network for all gene expression datasets, we extracted available interactions among common genes in four datasets. Totally, we found 63,113 protein-protein interactions, 1789 protein-DNA interactions and 3,862 signaling interactions among 10650 common genes in four datasets.

4.4.3 Module Biomarker Identification

To detect the modules in the collected interaction network, we first extract the significant NMs in the integrated cellular network as previously described in Chapter 3 [51]. We assume that NMs in an interaction network are enriched in "disease driver" genes which are more conserved than other downstream "passenger" genes. These NMs could form large aggregated modules that perform specific functions by forming collaborations among a large number of NMs. In this study, we focus on three-node NMs since larger size NMs (number of nodes > 3) are composed of three-node ones in most cases [119].

All the identified NMs are then examined by calculating their activity scores via gene expression data. Each NM is considered as a subnetwork. We assume that in a subnetwork A, there are M genes with expression levels across N patient samples:

$$G_k = \{g_{ij} | i = 1,2,\dots,M, j = 1,2,\dots,N\} \tag{4.2}$$

Given a particular gene i, the expression values g_{ij} are normalized to z-transformed scores z_{ij} so that the z score vector z_i has mean $\mu = 0$ and standard deviation $\sigma = 1$ over all samples j. The z score is defined by

$$z_{ij} = \frac{g_{ij} - \hat{\mu}_i}{\hat{\sigma}_i} \tag{4.3}$$

where $\hat{\mu}_i$ is mean expression value of gene *i* across samples, and $\hat{\sigma}_i$ is standard deviation of expression value of gene *i* across samples.

Let z represent the corresponding vector of class labels (e.g., tumor metastatic or non-metastatic). The discriminative score of gene *i* is defined as the mutual information $MI_i(x; y)$ between the expression levels of gene i and sample labels c:

$$MI_i(x; y) = \sum_{x \in z_i} \sum_{y \in c} p(x, y) log \frac{p(x, y)}{p(x)p(y)} \tag{4.4}$$

where x is the discretized value of z_i, and y is the sample lables, $p(x, y)$ is the joint probability density function of z_i and c, and $p(x)$ and $p(y)$ are the marginal pdf's of z_i and c. A histogram technique is applied to transform the continuous gene expression values to discrete ones for the calculation of the mutual information [187].

The activity score of a subnetwork *A* is then calculated by combining the transformed z scores derived from the expression of its individual genes. The individual z_{ij} of each member gene in one subnetwork are combined into the activity of a z_{A_j} by

$$z_{A\_j} = \frac{1}{\sqrt{\sum_{i=1}^{M} w_i^2}} \sum_{i=1}^{M} w_i \, z_{ij} \tag{4.5}$$

where w_i denotes the weight that is defined as

$$w_i = \frac{MI_i(x; y)}{\sum_{i=1}^{M} MI_i(x; y)} \tag{4.6}$$

The weighted z score is intended to emphasize the hub genes which are surrounded by many highly discriminative genes although they are not highly differently expressed themselves.

The discriminative score of subnetwork *A* is calculated similarly as defined in Eq. (4.4):

$$MI_A(x; y) = \sum_{x \in z_A} \sum_{y \in c} p(x, y) log \frac{p(x, y)}{p(x)p(y)} \tag{4.7}$$

where x is the discretized value of z_A, and y is the sample lables.

We perform two permutation tests to evaluate the significance of the identified NMs. For the first test, we test whether the mutual information with the disease class is stronger than that obtained by random assignments of classes to patients [64]. For the random model, we permute the sample labels for 100000 trials, yielding a null distribution of mutual information scores for each trial, and the real score of each NM is indexed on this null distribution. For the second test, we test if the mutual information with network interactions is stronger than that obtained by random assignments of gene expression vectors to individual genes. The mutual information for each NM is calculated over 100000 random trials in which the expression vectors of individual genes are permuted over the network. The score of each NM is indexed on the "global" null distribution of all random NM activity scores. In this study, significant NMs are selected that have both permutation test *P* values less than 0.0001.

The NMs that passed the significance tests are clustered in the NM dimension using the hierarchical clustering method. This result in a tree in which each internal leaf node is associated with a vector representing the average of all of the NM vectors at its decent leaves. We annotate each interior node with the Pearson correlation between the vectors associated with its two children in the hierarchy. We define as a NM cluster in which each interior node whose Pearson correlation differs by more than 0.05 from the Pearson correlation of its parent node in the hierarchy. The module is then formed by taking the union of the clustered NMs.

4.4.4 Classification Evaluation with Ensemble Feature Selection

After the modules are formed, the reliability of these module biomarkers is evaluated across different datasets. An ensemble feature selection strategy is proposed to increase the stability of feature selection algorithm. A wrapper approach ant colony optimization-support vector machine (ACO-SVM) is used as a baseline method for comparison purpose. We illustrate each step in the classification process in the following sections.

4.4.4.1 Ant Colony Optimization

Ant colony optimization studies artificial systems that takes inspiration from the behavior of real ant colonies [188]. The basic idea of ACO is that a large number of simple artificial agents are able to build good solutions to solve hard combinatorial optimization problems via low-level

based communications. Real ants cooperate in their search for food by depositing chemical traces (pheromones) on the ground. Artificial ants cooperate by using a common memory that corresponds to the pheromone deposited by real ants. The artificial pheromone is accumulated at runtime through a learning mechanism. Artificial ants are implemented as parallel processes whose role is to build problem solutions using a constructive procedure driven by a combination of artificial pheromone and a heuristic function to evaluate successively constructive steps.

A simple example is shown in Figure 4.2. There is a path along which ants walk from the food source to the nest (Figure 4.2 (A)). However, an obstacle appears and cut off the path (Figure 4.2 (B)). When the ants walk from the food source to the nest, they have to decide whether to take the up or down paths (Figure 4.2 (C)). The choice is influenced by the intensity of the pheromone trails by preceding ants. The higher level of pheromone one path has, the higher probability it will be chosen. The first ant reaching the obstacle has the same probability to take up or down paths since there is no previous pheromone on these two alternative paths. Because the up path is shorter than the down path, the first ant following it will reach the nest before the first ant following the down path (Figure 4.2 (C)). Other ants starting from nest to the food will find a stronger pheromone trail on the up path, since half of the ants that by chance decide to approach the food via the up path and by the already arrived ones coming via the up path will prefer (in probability) the up path to the down path. As a consequence, the number of ants following the up path per unit of time will be higher than the number of ants following the down path. This causes the level of pheromone trail on the up path to grow faster than on the down one, and therefore the probability with which any single ant chooses the path to follow is quickly biased towards the up one. Very quickly all ants will choose the up path (Figure 4.2 (D)).

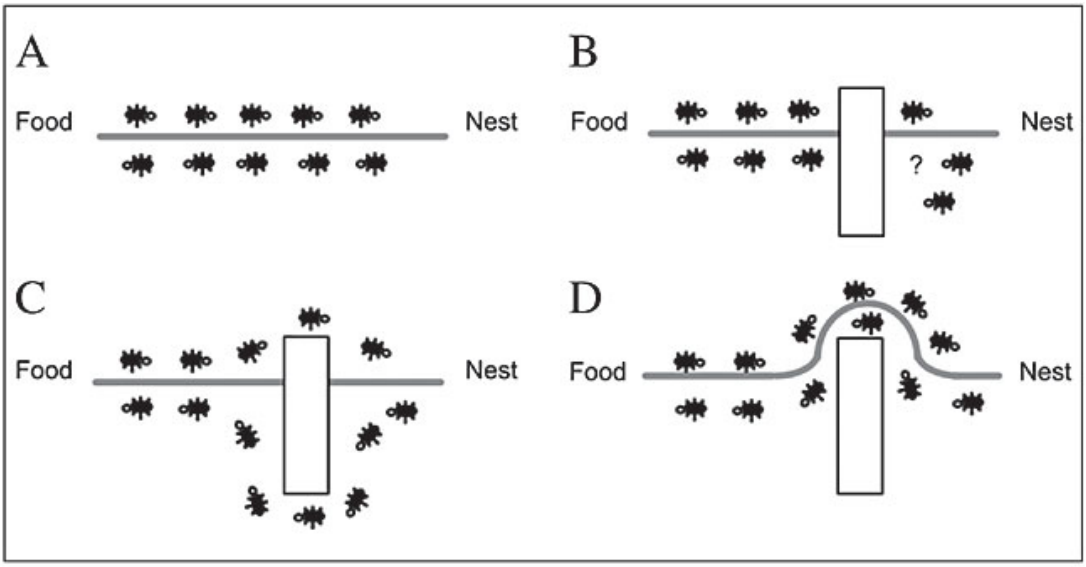


Figure 4.2 An example with real ants: (A) ants in a pheromone trail between nest and food; (B) an obstacle interrupts the trail; (C) ants find two paths to go around the obstacle; (D) a new pheromone trail is formed along the shorter path. Figure taken from [189].

In this chapter, we propose to use ACO for feature selection because of its efficiency and capability in identifying a set of interacting variables that are useful for classification. Also, ACO allows the integration of prior information into the algorithm for improved feature selection.

Through the probability function given below, each ant picks *n* sets of distinct features from *L* candidate features:

$$P_i(t) = \frac{(\tau_i(t))^{\alpha}\eta_i^{\beta}}{\sum_i(\tau_i(t))^{\alpha}\eta_i^{\beta}} \tag{4.5}$$

where $\tau_i(t)$ is the amount of pheromone trail at time *t* for the feature represented by index *i*; η_i represents prior information (e.g. univariate *t*-statistic) for the feature represented by index *i*; α and β are parameters that determine the relative influence of pheromone trail and prior information.

At $t = 0$, $\tau_i(t)$ is set to a constant for all features. Thus, at the first iteration, each ant chooses *n* distinct features (a trail) from *L* features with probabilities proportional to the existing prior knowledge. Let S_j be the *j*th ant consisting of *n* distinct features. Depending on the performance of S_j, the amount of pheromone trail for S_j is updated. The performance function is evaluated on the basis of disease state classification capability of each S_j. We use the features in

S_j to build a classifier and estimate the classification accuracy through the cross validation (CV) method. The amount of pheromone trail for each feature in S_j is updated in proportion to the corresponding classification accuracy using

$$\tau_i(t+1) = \rho * \tau_i(t) + \Delta\tau_i(t) \tag{4.6}$$

where ρ is a constant between 0 and 1, representing the evaporation of pheromone trails; $\Delta\tau_i(t)$ is an amount proportional to the classification accuracy of S_j. $\Delta\tau_i(t)$ is set to zero if the *i*th feature $f_i \notin S_j$. This update is made for all *N* ants $(S_1, S_2, \dots, S_N)$. Note that at $t = 0$, $\Delta\tau_i(t)$ is set to zero for all features. The updating rule allows trails that yield good classification accuracy to have their amount of pheromone trail increased, while others gradually evaporate. As the algorithm progresses, features with larger amounts of pheromone trails and strong prior information influence the probability function to lead the ants towards them.

Compared to PSO we used in Chapter 2, which is mostly used for continuous optimization problem, ACO is more suitable for discrete optimization problem. The reasons are listed below:

1) Ant colony optimization is driven by two parameters: heuristic value and pheromone value. Mostly these values are derived from parameters having discrete values.
2) Particle swarm optimization is driven by neighbor's velocity. Velocity is continuous parameter. As one of the parameters used for deriving velocity is time and time is continuous.

Ant colony optimization fits better at graph searching problems while PSO fits better at NN learning as well as pattern recognition. Reason behind this is, parameters used for graph searching are mostly discrete parameters while parameters used for learning/recognition are continuous parameters.

4.4.4.2 Support Vector Machine

Support Vector Machines (SVMs) are learning kernel-based systems that use a hypothesis space of linear functions in high-dimensional feature spaces [190]. In classification problems that involve two classes, linear SVMs search for the optimal hyperplane that maximizes the margin of separation between the hyperplane and the closest data points on both sides of the hyperplane. Thus, parameters of SVMs are determined on the basis of structural risk minimization, not error-risk minimization. Thus, they have the tendency to overcome the overfitting problem. In high

dimensional data classification problems, SVMs have proven themselves as one of the pattern classification algorithms with great generalization ability. We will use a linear SVM as the reference classifier for feature selection in module space.

4.4.4.3 ACO-SVM Feature Selection Algorithm

Ant colony optimization-support vector machine combines ACO and SVM to select features that are useful for SVM classification of two disease groups. ACO starts with a population of N module sets, where each module set consists of a pre-specified number (n) of distinct modules. Each module is selected from a given set of candidate modules (L) based on its probability function described previously in Eq. (4.5). SVM classifiers are then built for each module set and the performance of the module set in distinguishing the two groups is evaluated through the five-fold cross-validation method. Using Eq. (4.6), we update the amount of pheromone trail for each module in proportion to the classification accuracy of the module set, in which the module is involved. The goal is to provide those modules that can lead to improved classification accuracy with better probability of being selected in subsequent iterations. The block diagram of the ACO-SVM feature selection approach is outlined in Figure 4.3.

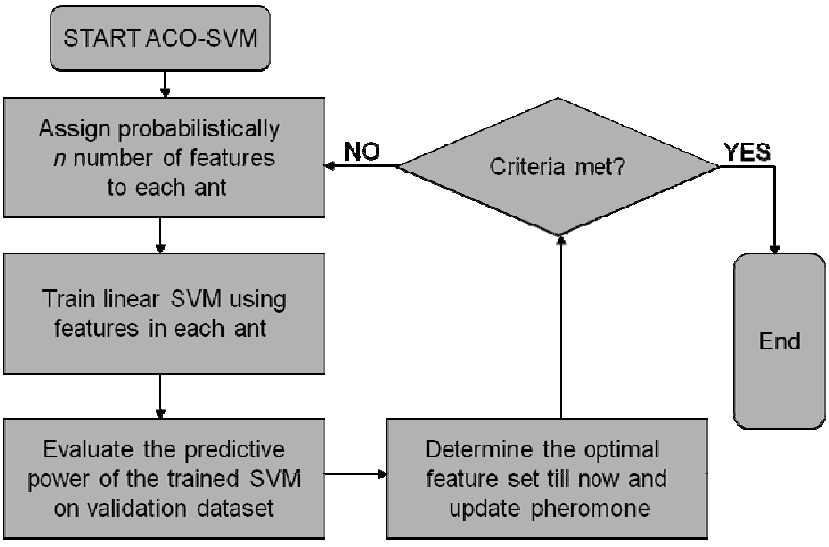


Figure 4.3 The block diagram of ACO-SVM algorithm.

4.4.4.4 Ensemble Feature Selection based on ACO-SVM

In order to select robust module biomarkers for classification in unknown patient samples, we propose an ensemble feature selection technique to select module subsets in training dataset

and validate their discriminative power in an independent validation dataset. Similar to ensemble learning for classification, ensemble feature selection techniques use a two-step procedure:

i) A number of different feature selectors are created;

ii) The outputs of these component feature selectors are aggregated to generate the final ensemble results.

We focus on the analysis of ensemble feature selection techniques using ACO-SVM feature selection approach. The ACO-SVM approach is used to select the best features in terms of their ability to distinguish between two patient phenotypes in a validation dataset which is not involved in the feature selection step.

To build a robust module biomarker set in one dataset, we generate slight variations of the original dataset, and aggregate the outputs of the ACO-SVM feature selection method using these variant samples. The rationale behind this is that for a stable biomarker set, training datasets with small change should generate biomarker sets with high similarities. The biomarkers with high frequencies in these biomarker sets are presumed to be most relevant to sample distinction and used to predict the class membership of independent samples. A subsampling approach is proposed to generate the training datasets with slight variations: a large number (e.g., 500) of datasets are generated by stratified subsampling the original dataset without replacement. As gene expression datasets generally contain only tens of samples, we generate subsamplings containing 90% of the samples of the original dataset, and the remaining 10% of the samples are used as internal validation dataset to estimate the performance of a classifier, called *within-dataset validation*. Since we consider typically 500 independent partitions in 90% training and 10% validation, we reduce the risk of overoptimistic results of traditional cross-validation experiments on small sample domains [191].

The biomarker sets generated from 500 subsampling datasets using the ACO-SVM approach are evaluated through a frequency plot, where we compute the frequency with which modules are selected. The most frequently selected set of modules is then validated by using it to classify an independent validation dataset. This approach is referred to as *cross-dataset validation*.

The double-validation procedure stated above is designed to provide an unbiased evaluation of the generalization error in an independent dataset. Since both prognostic and treatment outcome prediction groups contain two datasets, we evaluate the classification performance of

the module biomarker set generated from one dataset on the other dataset in the same group, or vice versa. The above ensemble feature selection procedure is illustrated in Figure 4.4.

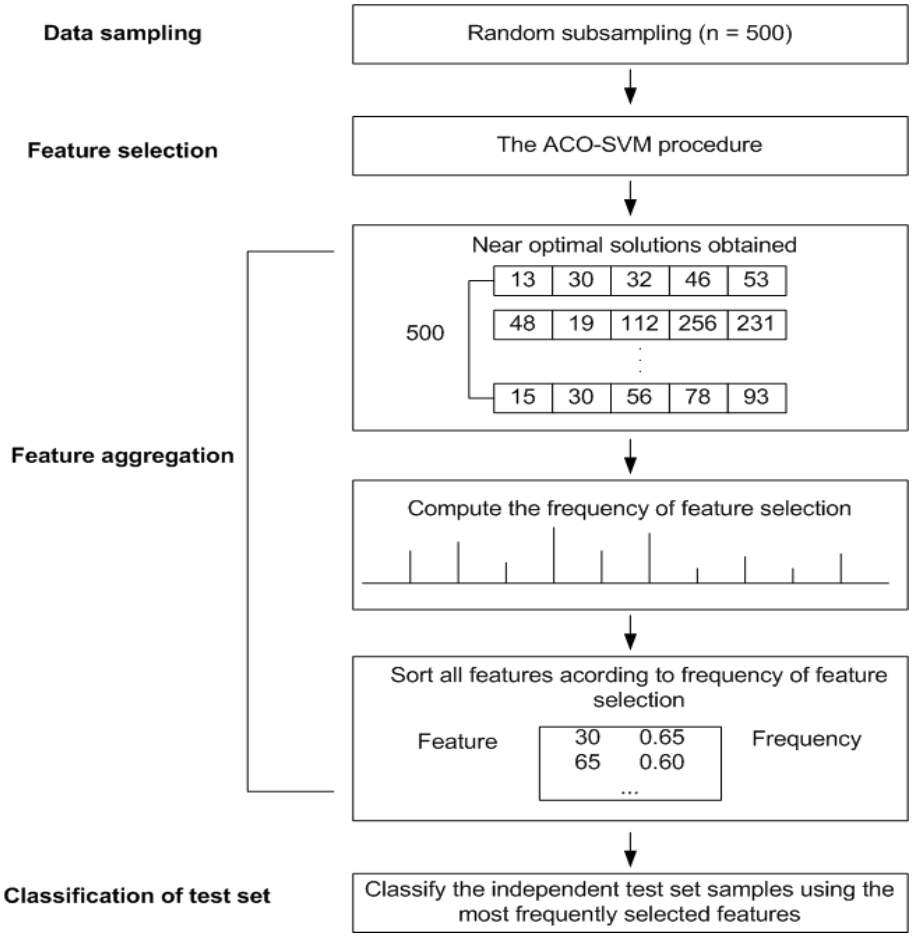


Figure 4.4 The flowchart of ensemble ACO-SVM approach.

4.5 Results

In this section, we report the experimental evaluations of our methods to search for module biomarkers with discriminative power between different subgroups of breast cancer patients in an interaction network context. Four breast cancer datasets were used to identify biomarkers for prognosis and treatment prediction purposes. In the following subsections, we present our results and comparison to previously proposed methods applied to the same public datasets.

4.5.1 Biological Interpretability of Module Biomarkers

The collected interaction network involved 72,562 three-node NMs detected using FANMOD tool [118]. Totally, 1017, 752, 696 and 908 NMs were identified in the four breast cancer datasets (van de Vijer, Wang, Loi, and in house datasets, respectively). This was based on two permutation tests for statistical significance consisting of 581, 707, 793, and 886 genes, respectively. By hierarchical clustering analysis, 162, 313, 270 and 343 module markers were constructed as candidate module biomarkers of the four datasets, respectively. Each module may be viewed as a putative biomarker for breast cancer. The modules are not based on individual detected genes, but rather on the aggregate behavior of genes connected in a functional module. This approach is indeed a departure from conventional gene-based expression analysis, which does not provide biological insight into the identified biomarkers.

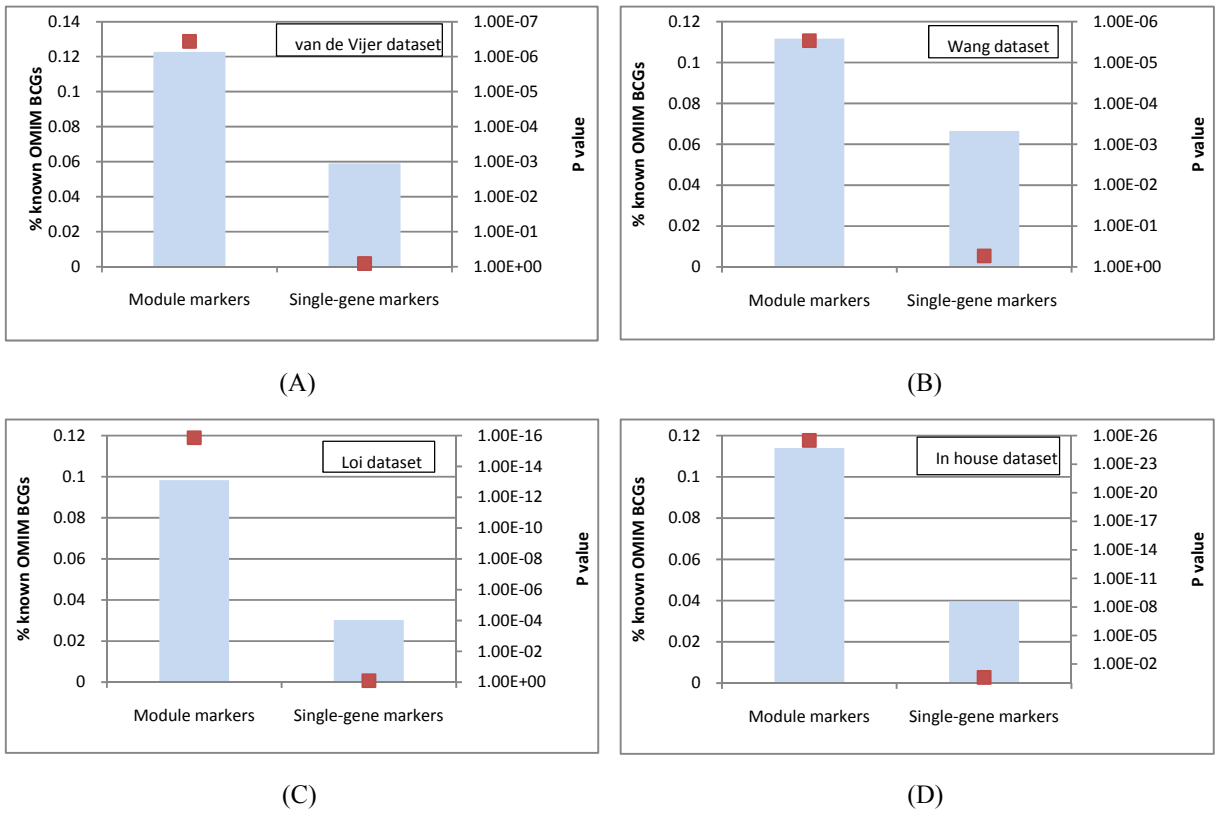


Figure 4.5 Detection of BCGs in module biomarkers of four datasets. The enrichment of disease genes is shown for modules or individual genes selected from van de Vijer dataset (A), Wang dataset (B), Loi dataset (C) and our in house dataset (D). Blue bars chart the percentage of BCGs among all genes covered in the markers on the left axis; the red dots chart the hypergeometric *P* values of enrichment on the right axis.

We further investigated whether the proposed module-based analysis can implicate upstream disease driver genes with relative low discriminative potential (e.g., those with larger *P* value in two-tailed t-test). Such proteins arise within a significant module if they are essential for maintaining its integrity. Moreover, these disease driver genes are mostly in the upstream of the gene regulatory cascade, regulating their downstream genes to be differentially expressed under different disease status. Detecting modules containing these disease driver genes is expected to improve the reliability and robustness of these module biomarkers across different datasets. To evaluate the power of a module-based method to identify disease driver genes, we assembled a list consisting of 711 breast cancer related genes (BCGs) extracted from the Online Mendelian Inheritance in Man (OMIM) database. The genes in the module biomarkers identified from four datasets are more enriched with these BCGs than the ones from a conventional gene expression based analysis without network information (Figure 4.5). In particular, we found that 69 out of 162, 123 out of 313, 120 out of 270, and 136 out of 343 module biomarkers in four datasets contained at least one known BCG, among which 31, 26, 41 and 44 module biomarkers contained two or more known BCGs, respectively. Most of these BCGs are not significantly differently expressed (Table 4.2). Disease genes that can be only detected by the proposed approach include *BRCA1*, *ESR1*, *TP53*, etc.

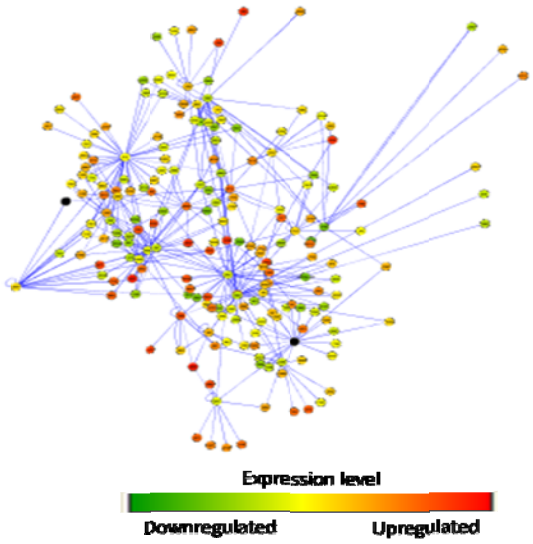


Figure 4.6 Module biomarkers enrichment in "disease drivers".

The "disease driver" genes are usually hub genes in the interaction data, i.e., genes with more than ten surrounding genes. We retrieved the existing interactions surrounding hub genes (genes

with more than ten interactions) from the collected molecular interaction data. As shown in Figure 4.6, in module biomarkers identified from van de Vijer dataset, only 4 out of 23 hub genes showed discriminative potential (P value <0.01). However, these hub genes are important biological markers than other members in one module since they are the center to gather its surrounding differential genes into one module biomarkers.

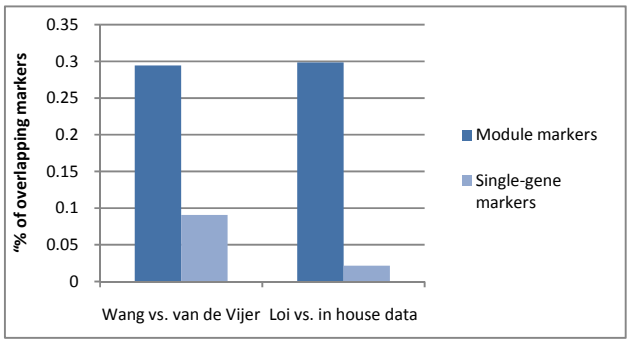


Figure 4.7 Agreement in markers selected from one dataset versus those selected from the other dataset in the same clinical group.

We also examined the agreement between module biomarkers identified from different cohorts of patients. The same selection process was also run for gene biomarkers selected by conventional methods. For comparison purpose, we select same number of top discriminative genes as the ones in module biomarkers for four datasets, i.e. the top 581, 707, 793, and 886 discriminative genes in four datasets respectively. As shown in Figure 4.6, the module biomarkers are more reproducible between datasets than individual gene biomarkers selected without network information (e.g. t-test).

Table 4.2 BCGs in module markers derived from four datasets

BCG	van de Vijer dataset	Wang dataset	Loi dataset	In house dataset
Differentially expressed (P value<0.05)	10	15	16	13
Not differentially expressed	61	64	62	88
Total	71	79	78	131

4.5.2 Classification Evaluation of Module Biomarkers

We tested the classification ability of the identified module biomarkers from four datasets using the proposed ACO-SVM approach. To use module information for classification, the weighted *z* score of module biomarkers were used as input feature values to a classifier based on

SVM. An ensemble ACO-SVM approach was used to select the optimal features based on Area Under the ROC Curve (AUC) scores in a double-validation procedure, as described in Methods section. We used a baseline ACO-SVM approach for comparison purpose. To perform ensemble feature selection for gene biomarkers, the *z* score of candidate gene biomarkers were used as input feature values to a classifier based on SVM. The AUC scores of the second independent validation dataset by the classifier built from both module and gene biomarkers selected from the first dataset are shown in Figure 4.7. Through the double-validation strategy, we showed that the module biomarkers outperformed the gene biomarkers in all four experiments. This implies that the module biomarkers are more robust across different datasets generated on different

4.5.3 Comparison to Existing Methods

Several studies have been reported to integrate interaction network information and other biological data (e.g., microarray data) for identification of genetic mediators of disease progression [69, 165, 192, 193]. However, only individual interaction layers, such as the transcriptional layer or the protein complex layer, were modeled by these methods. We propose an integrative approach for the identification of module-based biomarkers associated with the presentation of a specific tumor phenotype. In our approach, we choose to use an interaction network containing PPI, PDI and signaling pathway information. By adopting a genome-wide, mixed-interaction network, instead of the individual interaction layers of previous studies, we cover a far greater range of processes within the cell. This integration allows the method to capture several different mechanisms of action associated cancer progression and metastasis.

Compared to Chuang et al. [69] and Lee et al. [165], besides larger coverage of biological processes in our analysis, our approach uses an ensemble feature selection method to improve the classification accuracy and reliability of the module biomarkers. Both Chuang et al. and Lee et al. applied five-fold CV for one single dataset, which would generate overoptimistic results that do not adequately reproduce in independent datasets. In this study, a strict double validation strategy was used to estimate the classification performance. Such strategy leads to better assessment of classification accuracy in applying the resulting module biomarker set to classify previously unseen samples.

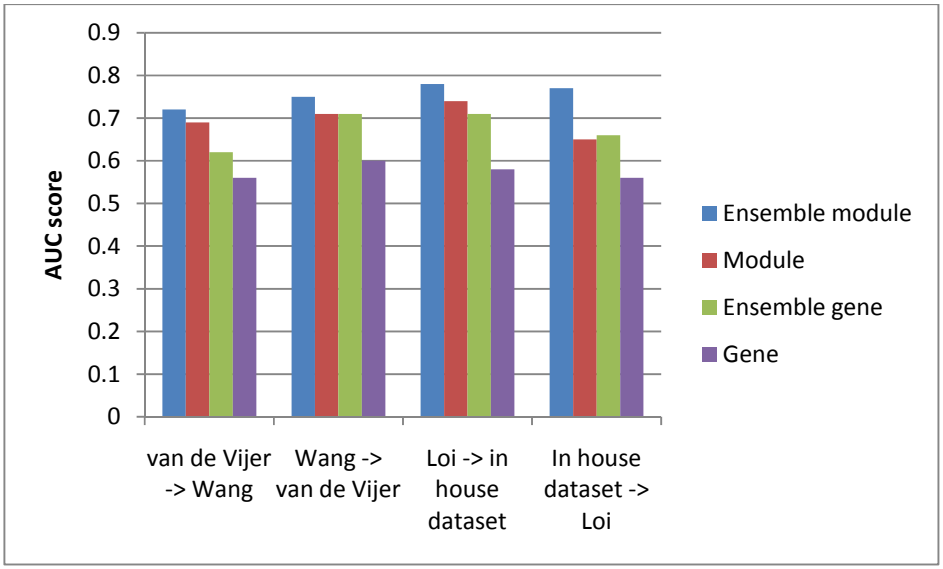


Figure 4.8 AUC classification performances of modules, genes with ensemble feature selection strategy, and without the ensemble strategy.

4.6 Summary and Discussions

In this chapter, we introduce a module-based feature selection framework to identify module biomarkers with high reproducibility and classification accuracy. This was accomplished by a hybrid feature selection approach that identifies groups of associated genes by incorporating interaction network information, called module biomarkers. Different from traditional data-driven group feature selection methods, we identified the "active subnetworks" within an interaction network context. The motivation is that a disease or clinical response may be viewed as an emergent behavior of interaction network that is altered by the complex interplay of genetic and environmental stimuli. Individual genes tend to collaborate to carry out some specific biological function, in which these genes are called a functional module. In our study, we decomposed interaction networks into NMs. Cancer-associated genes have been shown to be enriched in particular NM types, called hotspots in the mammalian cellular signaling network [180]. These hotspots are potentially biomarker clusters or drug target clusters for curing cancer. If a cancer-related gene is mutated in one phenotype, this mutation will influence its surrounding interaction partners at a functional module level. Since the cancer-related genes are usually upstream disease "drivers", they tend not to be highly differentially expressed compared to their downstream "passenger" genes. On the other hand, these disease driver genes are more stable

and study-independent than their downstream "passenger" genes across and within patient samples. We hypothesize that these functional modules are expected to have higher stability and reproducibility in unknown samples, since they have disease "drivers" as the hubs surrounded by their downstream "passengers". Our findings demonstrate that module biomarkers are more enriched with these disease driver genes such as *TP53*, which could not be identified by gene-based univariate methods (e.g., t statistic). The reliability of the module biomarkers from different dataset are compared to gene-based approach. The overlaps of module biomarkers from different datasets are largely improved, further confirming that the module biomarkers we identified are likely to be involved in cancer related mechanisms.

Also, we introduce in this chapter a strategy in which a set of ensemble feature selection methods is applied to improve biomarker stability and classification performance. In module space, the ensemble feature selection methods are combined with a double-validation strategy to select the optimal module biomarkers according to their classification accuracy on independent datasets. The stability of our ensemble feature selection approach is improved compared to non-ensemble method (Figure 4.7). This is particularly convenient since it corresponds to sizes of practical interest for the design of a diagnosis/predictive model. As high-quality interaction data become available, such hybrid feature selection methods will help us exploit more disease related information in these and other similar datasets available in human diseases. The proposed work in this chapter is supported in part by grant from the National Institutes of Health (R21CA139246).

5 Conclusion and Future Work

5.1 Summary of Original Contributions

We conclude this dissertation with a summary of the contributions of the proposed research. In this dissertation, we proposed and developed a series of computational methods to integrate and analyze multi-source biological data in a modular manner. The original contributions of this research work are summarized as follows.

5.1.1 Gene Module Identification by Knowledge-based Cluster Evaluation

We developed a knowledge-based evaluation approach for clustering algorithms to identify gene modules – a set of co-regulated genes by common transcription factor(s). The best clustering method is the one that has the strongest tendency to bring genes of similar function together when applied to diverse gene expression datasets. By investigating the KL divergence between cluster membership and known gene attributes, the z scores generated from different clustering partitions are used to indicate their enrichment for biological functions. Different clustering algorithms can be compared using the proposed method. The optimal cluster number is evaluated by both GSEA and BSEA analyses.

We have applied the proposed method to three gene expression datasets: the rat central nervous system, yeast cell cycle and human *Hela* cancer cell cycle datasets. Several clustering partitions were used to evaluate the optimum number of clusters for these datasets. The results demonstrated that FCM clustering with optimal fuzzification parameter m is most suitable for time course gene expression data analysis which contains large amounts of noise and no clear boundaries between clusters. This may facilitate the identification of experimental conditions for which genes are co-regulated and the underlying regulatory processes.

5.1.2 Gene Regulatory Module Inference by Hybrid Computational Intelligence Modeling

We have developed a computational framework for utilizing biological data from multiple sources to infer transcription factor-target gene relationships on the basis of gene regulatory modules, including time course gene expression profiles, molecular interaction data, and GO information. First, we applied cluster analysis of time course gene expression profiles to reduce dimensionality and used the GO category information to determine biologically meaningful gene modules, upon which a model of the gene regulatory module is built. This step enables us to address the scalability problem that is faced by researchers in inferring TRNs from time course gene expression data with limited time points. Second, we detected significant NMs for each transcription factor in an integrative molecular interaction network. Finally, we used NN models that mimic the topology of NMs to identify gene modules that may be regulated by a transcription factor. A hybrid of GA-PSO methods was applied to train the NN models.

The proposed computational framework was tested in two biological processes: yeast cell cycle, and human *Hela* cancer cell cycle. The identified gene regulatory modules were evaluated using the following validation strategies: (1) gene set enrichment analysis to evaluate the gene modules derived from clustering results; (2) binding site enrichment analysis to determine enrichment of the gene modules for the cognate binding sites of their predicted transcription factors; (3) comparison with previously reported results in the literatures to confirm the inferred regulations. The evaluations confirmed the consistency between the inferred gene regulatory relationships and the known biological evidences. The proposed framework could be beneficial to biologists for predicting the components of gene regulatory modules in which any candidate gene is involved. Such predictions can then be used to design a more streamlined experimental approach for biological validation. Understanding the dynamics of these gene regulatory modules will shed light on the underlying cellular regulatory processes.

5.1.3 Module-based Biomarker Discovery

We have developed a module-based biomarker discovery framework that integrates biological network information and gene expression data to identify biomarkers, not as

individual genes but as functional modules. The proposed framework presented a hybrid method combining group feature selection with ensemble feature selection to improve the reliability of disease biomarkers for samples in the same clinical group. First, a group feature selection method was proposed to extract the modules (subnetworks) with significant discriminative power between subclasses of diseases or between cases and controls. Then, an ensemble feature selection method was used to select the optimal biomarker sets, in which a double-validation strategy was applied. The ensemble method allows combining features selected from multiple runs with various data subsampling to increase the reliability and classification accuracy of the final selected biomarker set.

We have applied the proposed approach to four breast cancer datasets. Compared to gene-based approaches, the identified module biomarkers have the following important features: i) achieve higher prediction accuracy in independent validation datasets; ii) are more reproducible than individual gene markers; iii) improve the biological interpretability of results; and iv) are enriched in cancer "disease drivers".

5.2 Future Work

Large transformations of biology are expected with major influences on our society in this century to come. The integration of information science and molecular biology will intensify as faster computers and internet connections facilitate biological research. This section outlines several remaining problems/topics to be further explored, which have emerged for consideration during the research work of this dissertation. The discussion presented here can be viewed as a starting point for future research.

5.2.1 Possible Improvements on Knowledge-based Clustering Validation

The proposed cluster validation method is an external validation measure to select the clustering algorithm with the strongest tendency to bring genes of similar function together when applied to diverse expression datasets. Compared to the internal validation techniques, the external validation technique provides additional feedback on the quality of the data and previous preprocessing steps. A further improvement of cluster validation technique will be expected to combine validity under both internal and external measures, i.e., it will exhibit a

distinct underlying cluster structure while being consistent with prior biological knowledge. We are also aware that an objective cluster validation is only possible on the data with known well-defined cluster structures. Development of simulation datasets that mimic the properties of real biological data are particular important for evaluating a clustering validation method with respect to specific data properties.

5.2.2 Integration of More Types of Biological Data

Biological systems are characterized by many highly interconnected levels. As we have done in this dissertation, we analyzed the networks in a cell at transcriptional level. The combined PPI and PDI network may be augmented with additional types of relations, which provide further insight into other types of cellular mechanisms. One of the major tasks ahead is therefore the integration of more sources of information. One intriguing dataset to add is that of signaling transduction pathways, including directed PPIs such as those between kinases and their substrates, and interactions between signaling molecules (e.g., pheromone) and their targets. Assuming these data were available, it could enable the characterization of signal transduction pathways and their control mechanisms. Attempts to collect genome-wide signaling data are underway, e.g., using protein chips designed to test kinase phosphorylation interactions as performed by Snyder and colleagues [194]. In the future, we propose to incorporate the pathway information to complement the combined PPI and PDI network towards a more complete *interactome* in the cell. Cancer related pathways will be collected from multiple databases to build a signaling pathway network [195, 196], serving as another source of information.

5.2.3 Integration Method for Module Biomarker Discovery

The method developed for module biomarker discovery in this dissertation generates reliable biomarker sets across different datasets generated on different microarray platforms. However, one problem with these biomarker sets is that they are not always enriched in cancer-related annotations. One possible explanation is that we only consider the two conditions in this study (e.g., metastasis versus non-metastasis). In the future research, more cancer types/conditions could be included for analysis. The activity of module biomarkers will be calculated across multiple conditions, and the most significant ones will be considered as biomarker sets. Another

improvement could be the calibration of activity score against the background distribution. In order to properly capture the connection between expression and network topology, we need determine whether the activity score of a module is higher than expected relative to the modules from a randomized network (drawn from the same gene expression data but independently of the network). Randomization methods in network topology will be used for this purpose. For example, to keep the randomized networks as close as possible to the real network in terms of their network properties, we preserve the topology of each network and permute over the network nodes, generating 1000 random isomorphic network. Some possible analyses could be: (1) the PPI network is kept as is and the PDI network is randomized; (2) the PDI network is kept as is and the PPI network is randomized; (3) both networks are randomized. On the other hand, from a biological point of view, it has been shown that multiple diseases converge in common networks in an evolutionary way. Such networks are potentially crucial drug targets in multiple diseases. It is anticipated that more unexploited information will be discovered as more high-quality interaction data become available.

5.3 Conclusions

Different sources of biological data reflect cellular activities at different interaction levels. Integrated analysis of these large-scale datasets will help us unveil the global insight in the cellular networks. In this dissertation, we proposed three major research topics for integration and analysis of multi-source biological data, including gene expression data, various molecular interaction data, and GO information. A comprehensive computational framework was presented to tackle three biological problems: gene module identification, gene regulatory module inference, and module biomarker discovery. To address these problems, we developed computational models to identify gene modules, infer regulatory relationships between transcription factors and gene modules, and discover discriminative modules as potential disease biomarkers. Experimental results obtained by our proposed methods demonstrate their ability to extract the underlying regulatory structure and discover potential disease biomarkers from multi-source biological data. These findings stimulate the novel hypotheses for future research. Finally, several related tasks are proposed for future work.

Appendix A. Optimal gene modules in yeast cell cycle dataset

ORF	Cluster #	ORF	Cluster #	ORF	Cluster #	ORF	Cluster #	ORF	Cluster #
YBL002W	0	YBR243C	7	YGL008C	12	YBR158W	18	YNR066C	24
YBL003C	0	YCL062W	7	YGL255W	12	YDL117W	18	YOL058W	24
YBR009C	0	YCL063W	7	YGR240C	12	YDL179W	18	YOL088C	24
YBR010W	0	YDR130C	7	YGR279C	12	YGL055W	18	YOR250C	24
YDL055C	0	YDR150W	7	YIL011W	12	YGR041W	18	YOR332W	24
YDR224C	0	YDR261C	7	YLR084C	12	YGR044C	18	YOR391C	24
YDR225W	0	YDR302W	7	YLR100W	12	YGR086C	18	YPL025C	24
YNL030W	0	YDR346C	7	YLR169W	12	YIL009W	18	YPL054W	24
YNL031C	0	YER032W	7	YLR353W	12	YIL104C	18	YPR107C	24
YPL127C	0	YGL101W	7	YML052W	12	YJL078C	18	YPR155C	24
YBR094W	1	YGL216W	7	YML064C	12	YJL217W	18	YBR054W	25
YDR089W	1	YGR138C	7	YML072C	12	YKL116C	18	YBR092C	25
YEL025C	1	YHR098C	7	YML116W	12	YLR079W	18	YDR033W	25
YER145C	1	YHR108W	7	YNL172W	12	YLR194C	18	YNL160W	25
YGR176W	1	YHR135C	7	YOL030W	12	YNL046W	18	YPR149W	25
YGR177C	1	YHR205W	7	YOR153W	12	YNL078W	18	YCL027W	26
YGR259C	1	YKL004W	7	YOR298W	12	YNL192W	18	YCL055W	26
YGR260W	1	YKL048C	7	YPR128C	12	YOR263C	18	YGL089C	26
YHL040C	1	YKL096W-A	7	YPR138C	12	YOR264W	18	YGL090W	26
YHR151C	1	YKR037C	7	YBL061C	13	YPL158C	18	YJR004C	26
YIL094C	1	YLR180W	7	YBR087W	13	YDR451C	19	YKL177W	26
YIL119C	1	YLR209C	7	YDL103C	13	YFL037W	19	YKL178C	26
YJR003C	1	YLR437C	7	YDR053W	13	YIL123W	19	YLR452C	26
YKR079C	1	YML065W	7	YDR190C	13	YIL129C	19	YNR044W	26
YLR056W	1	YML125C	7	YDR279W	13	YJL158C	19	YBL030C	27
YLR214W	1	YMR002W	7	YDR440W	13	YKL096W	19	YCL025C	27
YLR413W	1	YMR163C	7	YEL076C-A	13	YLR300W	19	YCLX09W	27
YLR438W	1	YMR198W	7	YER149C	13	YMR215W	19	YDR380W	27
YML123C	1	YNL043C	7	YGL185C	13	YMR307W	19	YHR137W	27
YMR015C	1	YNL216W	7	YGR042W	13	YOR247W	19	YJR048W	27
YMR202W	1	YOL012C	7	YGR234W	13	YOR248W	19	YKL035W	27
YNL111C	1	YOL112W	7	YGR276C	13	YBR204C	20	YKL043W	27
YNR050C	1	YOR323C	7	YHR106W	13	YCR018C	20	YKR039W	27
YOR383C	1	YOR324C	7	YHR126C	13	YGR238C	20	YKR046C	27
YPL036W	1	YOR337W	7	YIL076W	13	YLR347C	20	YLL028W	27

ORF	Cluster #	ORF	Cluster #	ORF	Cluster #	ORF	Cluster #	ORF	Cluster #
YPL075W	1	YPL128C	7	YJR043C	13	YNL326C	20	YLR058C	27
YPL111W	1	YPR034W	7	YKL172W	13	YOR023C	20	YLR142W	27
YER124C	2	YPR111W	7	YKR091W	13	YAR003W	21	YML120C	27
YGL028C	2	YBR038W	8	YLR135W	13	YDL010W	21	YMR058W	27
YHR143W	2	YDR146C	8	YML020W	13	YDR307W	21	YMR145C	27
YLR286C	2	YGL021W	8	YMR246W	13	YDR481C	21	YMR189W	27
YNR067C	2	YGR108W	8	YNL231C	13	YEL064C	21	YNL037C	27
YBR133C	3	YLR190W	8	YOL094C	13	YFL060C	21	YOL119C	27
YDR247W	3	YML119W	8	YOR075W	13	YGL060W	21	YOR256C	27
YEL017W	3	YMR001C	8	YOR115C	13	YGL163C	21	YOR273C	27
YER018C	3	YMR032W	8	YOR242C	13	YGL207W	21	YPL021W	27
YGL125W	3	YNL058C	8	YOR283W	13	YGR153W	21	YPL061W	27
YGR113W	3	YPR119W	8	YOR307C	13	YHR123W	21	YPL250C	27
YHR086W	3	YPR156C	8	YOR308C	13	YHR154W	21	YPL265W	27
YHR146W	3	YBL052C	9	YOR355W	13	YHR159W	21	YAL067C	28
YHR178W	3	YBR242W	9	YPL232W	13	YJL015C	21	YER042W	28
YIL050W	3	YCRX05W	9	YBR073W	14	YJL072C	21	YER091C	28
YIL131C	3	YDL048C	9	YCR065W	14	YJR054W	21	YFR030W	28
YIL144W	3	YDL180W	9	YDL018C	14	YJR155W	21	YGL184C	28
YIR010W	3	YDR011W	9	YDL101C	14	YKL089W	21	YGR055W	28
YJL119C	3	YDR029W	9	YDL163W	14	YKL182W	21	YIR017C	28
YJL134W	3	YDR149C	9	YDL164C	14	YLR050C	21	YJL060W	28
YJL137C	3	YFR039C	9	YDR400W	14	YLR151C	21	YJR010W	28
YJR001W	3	YGL195W	9	YDR507C	14	YLR233C	21	YJR137C	28
YKL069W	3	YGL209W	9	YHR110W	14	YML021C	21	YKL001C	28
YKR010C	3	YGR035C	9	YHR149C	14	YML133C	21	YKR069W	28
YKR041W	3	YHR029C	9	YIL141W	14	YNL181W	21	YLL061W	28
YLL032C	3	YIL056W	9	YJL073W	14	YOL034W	21	YLL062C	28
YLR288C	3	YIL122W	9	YJL074C	14	YOR176W	21	YLR302C	28
YLR455W	3	YJL099W	9	YJL187C	14	YPL014W	21	YLR303W	28
YMR003W	3	YJR110W	9	YKL113C	14	YPL057C	21	YNL276C	28
YMR295C	3	YKL130C	9	YLL002W	14	YBR108W	22	YPL274W	28
YNL176C	3	YKL183W	9	YLL022C	14	YBR256C	22	YPR167C	28
YNL197C	3	YKR021W	9	YLR103C	14	YBR273C	22	YBL009W	29
YOR073W	3	YLR034C	9	YLR313C	14	YDL089W	22	YBL063W	29
YOR083W	3	YLR057W	9	YLR383W	14	YDL169C	22	YDR113C	29
YOR372C	3	YLR098C	9	YML060W	14	YDR085C	22	YDR297W	29
YPL116W	3	YML035C-A	9	YMR078C	14	YDR342C	22	YDR355C	29
YPL253C	3	YMR183C	9	YNL072W	14	YDR493W	22	YEL042W	29

ORF	Cluster #	ORF	Cluster #	ORF	Cluster #	ORF	Cluster #	ORF	Cluster #
YPL264C	3	YMR254C	9	YNL102W	14	YEL060C	22	YEL061C	29
YPL269W	3	YNL056W	9	YNL233W	14	YEL068C	22	YER003C	29
YCL012W	4	YNL171C	9	YNL300W	14	YGR219W	22	YGL225W	29
YCL013W	4	YOL014W	9	YNL312W	14	YHR067W	22	YGR014W	29
YCL014W	4	YOL069W	9	YOR074C	14	YHR094C	22	YGR099W	29
YCR024C-A	4	YOL114C	9	YPL208W	14	YHR113W	22	YHR061C	29
YHL028W	4	YOR058C	9	YPL221W	14	YIL114C	22	YHR173C	29
YHR023W	4	YOR104W	9	YPR135W	14	YIR036C	22	YJL092W	29
YIL106W	4	YOR105W	9	YPR174C	14	YJL044C	22	YJL118W	29
YIL158W	4	YOR129C	9	YBR157C	15	YJL067W	22	YJL201W	29
YJL051W	4	YOR152C	9	YCL065W	15	YJL196C	22	YKL008C	29
YJR092W	4	YOR235W	9	YCL066W	15	YKL104C	22	YLL012W	29
YLR131C	4	YOR320C	9	YCR040W	15	YKL151C	22	YLR045C	29
YML033W	4	YPL058C	9	YCR041W	15	YLR231C	22	YLR342W	29
YML034W	4	YPL133C	9	YDL037C	15	YLR273C	22	YLR372W	29
YML058W	4	YPR013C	9	YDR461W	15	YML110C	22	YLR373C	29
YOR025W	4	YPR014C	9	YER150W	15	YMR011W	22	YLR380W	29
YOR315W	4	YPR045C	9	YFL026W	15	YNL015W	22	YMR144W	29
YPL141C	4	YPR157W	9	YFL044C	15	YNL208W	22	YNL126W	29
YPL155C	4	YBL111C	10	YGL162W	15	YOL016C	22	YNL166C	29
YPL242C	4	YBL112C	10	YGR146C	15	YOR018W	22	YNL283C	29
YDL093W	5	YBL113C	10	YGR284C	15	YOR052C	22	YNR009W	29
YDL095W	5	YBR071W	10	YHL026C	15	YOR317W	22	YOR188W	29
YDL096C	5	YDL127W	10	YIL167W	15	YPL187W	22	YPL032C	29
YDR353W	5	YDR545W	10	YIL168W	15	YAR008W	23	YPR141C	29
YEL047C	5	YEL040W	10	YJL079C	15	YBR070C	23	YPR159W	29
YER016W	5	YEL075C	10	YKR042W	15	YBR296C	23	YBL064C	30
YER152C	5	YEL076C	10	YLR041W	15	YCL023C	23	YBR053C	30
YGL027C	5	YER111C	10	YLR297W	15	YCL061C	23	YBR067C	30
YHR127W	5	YER189W	10	YML050W	15	YDL011C	23	YCL040W	30
YJL173C	5	YER190W	10	YML066C	15	YDL102W	23	YCL042W	30
YJR148W	5	YFL064C	10	YMR253C	15	YDL105W	23	YDR368W	30
YKL066W	5	YFL065C	10	YOL150C	15	YDL156W	23	YFR015C	30
YKL067W	5	YFL066C	10	YOR229W	15	YDL227C	23	YGL037C	30
YKL101W	5	YFL067W	10	YOR258W	15	YDR503C	23	YKL103C	30
YKL127W	5	YGR296W	10	YAL053W	16	YDR528W	23	YML100W	30
YKR090W	5	YHL049C	10	YBR007C	16	YER170W	23	YNL134C	30
YLR121C	5	YHL050C	10	YBR161W	16	YFL008W	23	YNL173C	30
YLR154C	5	YHR218W	10	YDL157C	16	YFR027W	23	YOR230W	30
YLR234W	5	YHR219W	10	YDL197C	16	YGL038C	23	YPR160W	30

ORF	Cluster #	ORF	Cluster #	ORF	Cluster #	ORF	Cluster #	ORF	Cluster #
YLR326W	5	YIL177C	10	YDL211C	16	YHR153C	23	YAR071W	31
YML012W	5	YJL225C	10	YDR144C	16	YIL025C	23	YBL023C	31
YML109W	5	YKR077W	10	YDR356W	16	YIL026C	23	YBR093C	31
YMR238W	5	YLL066C	10	YDR488C	16	YJL181W	23	YBR202W	31
YNL169C	5	YLL067C	10	YDR501W	16	YJR030C	23	YCR042C	31
YNL263C	5	YLR049C	10	YER019W	16	YJR154W	23	YDR191W	31
YOL019W	5	YLR462W	10	YER118C	16	YKL042W	23	YEL032W	31
YOR084W	5	YLR463C	10	YGL061C	16	YKL108W	23	YHR005C	31
YOR114W	5	YLR464W	10	YGL093W	16	YLR032W	23	YHR215W	31
YOR321W	5	YLR465C	10	YGL200C	16	YLR212C	23	YJL157C	31
YPR075C	5	YLR466W	10	YGR140W	16	YLR235C	23	YJL194W	31
YPR076W	5	YLR467W	10	YGR188C	16	YLR236C	23	YLR274W	31
YPR106W	5	YNL339C	10	YHR172W	16	YLR457C	23	YNL145W	31
YAL040C	6	YOL011W	10	YIL132C	16	YLR458W	23	YOR066W	31
YBR287W	6	YPL283C	10	YJL018W	16	YML102W	23	YPR019W	31
YDL039C	6	YPR202W	10	YJL019W	16	YMR048W	23	YDR055W	32
YDL138W	6	YPR203W	10	YJL091C	16	YNL225C	23	YJL159W	32
YDR001C	6	YPR204W	10	YJR006W	16	YNL273W	23	YKL163W	32
YDR077W	6	YBR088C	11	YKL165C	16	YNL289W	23	YKL164C	32
YDR157W	6	YDL003W	11	YLR343W	16	YNL304W	23	YKL185W	32
YDR337W	6	YER001W	11	YML061C	16	YOL101C	23	YNL327W	32
YFL006W	6	YER070W	11	YMR076C	16	YOR033C	23	YAR007C	33
YFL011W	6	YGR189C	11	YNL165W	16	YOR144C	23	YBL035C	33
YFR002W	6	YIL066C	11	YOR195W	16	YOR342C	23	YBR089W	33
YGL032C	6	YIL140W	11	YPL124W	16	YPR018W	23	YCL022C	33
YGL201C	6	YKR012C	11	YPL209C	16	YCR002C	24	YCL024W	33
YGR065C	6	YKR013W	11	YPL241C	16	YDR448W	24	YCL060C	33
YHR022C	6	YLR183C	11	YPL255W	16	YGL212W	24	YDR097C	33
YHR092C	6	YML027W	11	YAL022C	17	YGR124W	24	YDR309C	33
YIL072W	6	YMR199W	11	YAR018C	17	YHR006W	24	YER095W	33
YIL121W	6	YMR305C	11	YBR139W	17	YHR030C	24	YGR109C	33
YIL162W	6	YOL007C	11	YBR200W	17	YHR208W	24	YGR151C	33
YJL195C	6	YOL090W	11	YDL128W	17	YIL074C	24	YGR152C	33
YKL044W	6	YPL163C	11	YGL116W	17	YIL117C	24	YGR221C	33
YKL209C	6	YPL256C	11	YGR092W	17	YIL135C	24	YJL115W	33
YKR103W	6	YBL098W	12	YGR143W	17	YIL138C	24	YKL045W	33
YLR013W	6	YBL100C	12	YGR230W	17	YKL052C	24	YMR179W	33
YLR040C	6	YBR069C	12	YHR152W	17	YKL065C	24	YNL082W	33
YNL328C	6	YBR086C	12	YLR254C	17	YLR095C	24	YNL262W	33
YOL132W	6	YBR138C	12	YMR031C	17	YLR099C	24	YNL309W	33

ORF	Cluster #	ORF	Cluster #	ORF	Cluster #	ORF	Cluster #	ORF	Cluster #
YOR049C	6	YCL038C	12	YNL057W	17	YLR210W	24	YOL017W	33
YOR126C	6	YCL043C	12	YOL070C	17	YLR225C	24	YPL153C	33
YOR127W	6	YCR069W	12	YOL158C	17	YML117W	24	YPL267W	33
YOR270C	6	YDR276C	12	YOR313C	17	YMR055C	24	YPR120C	33
YPL095C	6	YDR432W	12	YOR314W	17	YMR278W	24	YPR175W	33
YBR008C	7	YEL065W	12	YBR083W	18	YNL003C	24		

Appendix B. GSEA for optimal clusters in yeast cell cycle dataset

Cluster ID	# of genes in cluster	Enriched Functional Category	Total Genes in the category	Clustered Genes	P value
1	10	nucleosome	10	9	1.8E-27
		DNA binding	229	9	7.2E-13
2	25	steroid metabolic process	43	5	4.9E-07
		steroid biosynthetic process	32	4	5.8E-06
3	5	cytokinesis,completion of separation	11	5	4.9E-15
		cell separation during cytokinesis	13	5	1.4E-14
4	33	kinetochore	54	5	6.6E-06
		mitotic cell cycle	271	8	0.000047
5	16	cellular bud	150	6	9.2E-07
		cytoskeletal part	180	6	2.7E-06
6	29	dolichyl-phosphate-mannose-protein mannosyltransferase activity	7	3	2.8E-06
		protein amino acid O-linked glycosylation	16	3	0.000043
7	28	transporter activity	338	8	0.000063
		primary active transmembrane transporter activity	53	4	0.000071
8	37	microtubule	35	4	0.000042
		cytoplasmic microtubule	14	3	0.00006
9	11	cellular bud neck	115	5	6.9E-07
		site of polarized growth	152	5	2.7E-06
10	44	N/A	N/A	N/A	N/A
11	37	DNA helicase activity	75	14	1.1E-18
		mitotic recombination	41	7	2.1E-09
12	16	ribonucleoside-diphosphate reductase activity	4	2	0.000034
		cell cycle process	440	7	0.000043
13	28	plasma membrane	261	9	8.6E-07
		transmembrane transporter activity	246	7	0.000063
14	31	leading strand elongation	14	3	0.000035
15	30	response to DNA damage stimulus	233	10	4.8E-08
		DNA replication	131	8	9.1E-08

Cluster ID	# of genes in cluster	Enriched Functional Category	Total Genes in the category	Clustered Genes	P value
		DNA metabolic process	710	15	1.1E-07
16	22	L-serine ammonia-lyase activity	3	2	0.000033
17	32	microtubule-based process	101	12	1.9E-14
		microtubule cytoskeleton	94	11	3.3E-13
18	15	N/A	N/A	N/A	N/A
19	20	cellular bud	150	6	4.1E-06
		site of polarized growth	152	5	0.000078
20	10	cell wall	114	7	5E-11
		glucanosyltransferase activity	6	2	0.000032
21	6	N/A	N/A	N/A	N/A
22	28	N/A	N/A	N/A	N/A
23	29	pentose transmembrane transporter activity	4	3	3.2E-07
		fructose transmembrane transporter activity	15	3	0.000035
24	35	chromosome	231	12	1.3E-09
		mitotic sister chromatid cohesion	22	5	8.4E-08
		DNA replication	131	8	3.4E-07
25	30	N/A	N/A	N/A	N/A
26	5	N/A	N/A	N/A	N/A
27	8	response to pheromone	94	7	8.6E-13
		conjugation with cellular fusion	119	7	4.7E-12
28	24	amine transmembrane transporter activity	16	4	2.6E-07
		polyamine transmembrane transporter activity	10	3	5.3E-06
29	17	sulfur metabolic process	67	11	7.3E-19
		methionine metabolic process	24	7	7E-14
30	30	cytoskeletal part	180	9	7E-08
		spindle	80	6	1.4E-06
31	14	energy reserve metabolic process	36	3	0.000055
		cellular carbohydrate metabolic process	213	5	0.000058
32	15	MCM complex	6	5	1.9E-13
		pre-replicative complex	15	6	2.4E-13
		DNA replication preinitiation complex	21	6	2.6E-12

Cluster ID	# of genes in cluster	Enriched Functional Category	Total Genes in the category	Clustered Genes	P value
33	6	cell wall	114	5	9.2E-09
		structural constituent of cell wall	19	3	4.3E-07
34	22	DNA-dependent DNA replication	97	10	5.8E-14
		replisome	15	5	4.8E-10

Appendix C. Optimal gene modules in human *Hela* cell cycle dataset

Gene Symbol	Cluster #	Gene Symbol	Cluster #	Gene Symbol	Cluster #	Gene Symbol	Cluster #
RNPC3	0	HSD17B11	12	ZNF414	21	AOC3	31
BIRC2	1	KLF6	12	ZNFX1	21	APEX2	31
BLM	1	MAN1A2	12	ZRANB2	21	ARHGDIB	31
C14ORF106	1	MID1	12	ZWINT	21	BTBD3	31
C20ORF117	1	MRPL19	12	ANKRD10	22	C14ORF130	31
CDCA7L	1	NCAPH	12	ARHGAP11A	22	CALM3	31
CDCA8	1	NNMT	12	ARHGAP19	22	CENPE	31
CDKN3	1	NT5DC1	12	ARL6IP2	22	CNN2	31
CENPA	1	NUSAP1	12	ATF7IP	22	CTSD	31
CNIH4	1	PDGFA	12	BRD8	22	ERN2	31
COL7A1	1	RAD21	12	BUB1	22	FOXM1	31
CTCFL	1	RCBTB2	12	BUB3	22	HCP5	31
ARGLU1	1	SEPN1	12	CCDC88A	22	HELLS	31
ERN2	1	SFPQ	12	CDC7	22	HMGB3	31
HELLS	1	SLC17A2	12	CDCA5	22	IFIT2	31
HORMAD1	1	STIL	12	CENPM	22	KIAA0586	31
HP1BP3	1	TIMP1	12	CFD	22	KIF5B	31
LMNB1	1	TOP2A	12	CROP	22	KIFC1	31
MAN1A2	1	TSC22D1	12	CTNNA1	22	LNPEP	31
MAP3K2	1	WWC1	12	DNAJC3	22	MAD2L1	31
FBXL20	1	ZWINT	12	ESCO2	22	ME3	31
KIAA0101	1	ARGLU1	13	FAM83D	22	MEGF9	31
MPHOSPH1	1	TTLL7	13	FKBP1A	22	MLF1IP	31
NIPBL	1	CEP55	14	ILF2	22	MRP63P6	31
NSUN5C	1	FAM60A	14	INADL	22	NUP43	31
NUCKS1	1	HCFC1	14	INSM1	22	OGT	31
PACS1	1	MELK	14	INTS7	22	ORMDL1	31
PIF1	1	SORL1	14	ITFG1	22	OSGIN2	31
RGS3	1	TOP2A	14	ITPR1	22	PLCXD1	31
SEPHS1	1	WDR90	14	JMJD2A	22	PSCD2	31
SH3GL2	1	ABCC2	15	KBTBD2	22	RNPC3	31
STAT5B	1	ADAMTS1	15	KIAA1712	22	RRP1	31
THRAP3	1	BIRC5	15	MAN1A2	22	SERPINB4	31

Gene Symbol	Cluster #	Gene Symbol	Cluster #	Gene Symbol	Cluster #	Gene Symbol	Cluster #
TNPO1	1	C14ORF68	15	MAP3K15	22	SKIP	31
TOMM70A	1	CAPS	15	MBD4	22	SLIT2	31
TRA2A	1	CDC27	15	MCM4	22	TSKU	31
WDR68	1	CEP57	15	MEPCE	22	ACD	32
APOA1BP	2	CHAF1B	15	MRPS2	22	ANKRD40	32
ARMC1	2	DNAJB4	15	MTMR15	22	ARHGAP8	32
ASXL1	2	FAM64A	15	NR5A2	22	BBS2	32
C12ORF32	2	FEN1	15	NUF2	22	BMP2	32
C6ORF166	2	FRZB	15	ORC1L	22	C13ORF3	32
CDC25A	2	FZR1	15	PBK	22	CCDC90B	32
CDC25C	2	GINS3	15	PLCXD1	22	CCNE1	32
CDKN1B	2	HIST2H4B	15	POLA1	22	CDC45L	32
CIITA	2	HIST2H4B	15	PTGER3	22	CKS1B	32
CIT	2	HRSP12	15	PTMS	22	DONSON	32
CKAP2	2	IL18BP	15	RAD51	22	DYNC1LI2	32
DNAJB9	2	KIAA0182	15	RBBP8	22	ERN2	32
FABP1	2	KIAA1586	15	RCCD1	22	FANCA	32
FLJ13231	2	MCM8	15	RFC4	22	FLJ20699	32
GINS2	2	NCOA3	15	RNF141	22	GOT1	32
GOLGA8A	2	NEK2	15	SETD8P1	22	HIST1H2BB	32
GSPT1	2	NUP98	15	SLBP	22	HIST3H2A	32
HERPUD2	2	PCF11	15	SMTN	22	HLA-DOA	32
HSPC157	2	POLD3	15	SP1	22	HMGB2	32
IFNAR1	2	PPP1R2	15	TXNDC9	22	HMGB3	32
IVNS1ABP	2	RANGAP1	15	UBE2D3	22	HTF9C	32
KDELC1	2	SGCD	15	ZC3HC1	22	INTS7	32
KIAA1147	2	STAT5B	15	ZNHIT2	22	ITGB3	32
KIAA1370	2	UBE2T	15	AHI1	23	KIFC2	32
KIAA1524	2	ZSCAN5	15	BARD1	23	LMBRD2	32
KIF5B	2	ATAD2	16	C15ORF23	23	MAP3K15	32
LARP7	2	CCNB2	16	INCENP	23	MDM2	32
MNX1	2	DHFR	16	LRCH1	23	MED31	32
MZF1	2	E2F5	16	BTBD3	24	NCOA5	32
NFIA	2	ENOSF1	16	C1ORF103	24	OSGIN2	32
NUDT4	2	FAM115A	16	C3ORF60	24	PDXP	32
NUSAP1	2	GADD45A	16	CCDC99	24	PHTF1	32
PANK2	2	HIF1A	16	CCNE2	24	RFC4	32

Gene Symbol	Cluster #	Gene Symbol	Cluster #	Gene Symbol	Cluster #	Gene Symbol	Cluster #
PLK2	2	HSPA8	16	CDH24	24	RNF126	32
POLQ	2	INDO	16	CENPF	24	RNF141	32
PRKAR1A	2	INSR	16	CENTB5	24	SAP30	32
PSEN1	2	KIAA0841	16	CKAP5	24	SLC38A2	32
PTPN4	2	KIAA1598	16	CTNND1	24	STCH	32
PVRIG	2	KLHDC5	16	CXCL14	24	TFF3	32
QRICH1	2	KPNB1	16	DNAJB1	24	TMEM132A	32
RAD51C	2	LBR	16	DSP	24	TPX2	32
RAN	2	MNT	16	ELP3	24	TRIP13	32
RGS3	2	NEK2	16	FREQ	24	TTK	32
RHEB	2	NEK4	16	FYN	24	TUBB2C	32
RHOBTB3	2	PMS2	16	FZR1	24	UBE2C	32
RNF113A	2	RPL13A	16	GLI1	24	UBE2C	32
SGK1	2	SDC1	16	HIST2H4B	24	BRCA1	33
TLE3	2	SFRS5	16	HS2ST1	24	C20ORF111	33
TOP1	2	SLC25A45	16	HSPB8	24	CIT	33
UACA	2	SPAG5	16	INTS8	24	DCC1	33
VPS37C	2	TFAP2A	16	KIAA1333	24	DIS3	33
WDR76	2	TGIF1	16	KIF14	24	DTL	33
PPP1R10	3	TLOC1	16	KRAS	24	DUSP4	33
AMD1	4	ZNF24	16	LMNA	24	EFHC1	33
APOA1BP	4	C9ORF100	17	MLF1IP	24	GPSM2	33
BAG3	4	CKS2	17	MSH2	24	KPNA2	33
CCDC14	4	DYNLL1	17	MUC1	24	MCM4	33
FGA	4	EIF4EBP2	17	NCOA5	24	PHF15	33
HIPK2	4	ESCO2	17	NKTR	24	ZNF281	33
KIAA1333	4	GCLM	17	OSBPL6	24	ZRANB2	33
LENG8	4	GOLGA8A	17	PCNA	24	AK3	34
NDE1	4	HIST2H4B	17	PHTF2	24	AMD1	34
NR3C1	4	HOXB4	17	PKNOX1	24	ANP32E	34
NSUN3	4	MDC1	17	POM121	24	AOC2	34
ODF2	4	MGAT2	17	PRIM1	24	ARL6IP2	34
PCNA	4	NOS1	17	PRPSAP1	24	ASAM	34
RMI1	4	NUP37	17	PRR11	24	AURKA	34
WSB1	4	PIK3CD	17	RRM2	24	CASP3	34
BRD8	5	PPP2CB	17	SCYL1	24	CCDC14	34
CDKN2D	5	PRR11	17	SMARCD1	24	CCNA2	34

Gene Symbol	Cluster #	Gene Symbol	Cluster #	Gene Symbol	Cluster #	Gene Symbol	Cluster #
DMXL2	5	RANGAP1	17	SMC4	24	CDC16	34
FANCD2	5	RNPS1	17	SUV420H1	24	CDC6	34
LYAR	5	SHCBP1	17	TIPIN	24	CKS1B	34
MATN2	5	SMARCB1	17	TOP2A	24	CLSPN	34
ABCA5	6	SNUPN	17	TUBD1	24	COQ9	34
BRD7	6	SRF	17	UBQLN2	24	CREBZF	34
C13ORF34	6	TLE3	17	UNG	24	CRK	34
C21ORF15	6	TMEM138	17	VCAM1	24	DCC1	34
CDC20	6	TNPO2	17	ANKRD10	25	DDX11	34
CDC42	6	TREX1	17	C14ORF142	25	DHX8	34
CFLAR	6	UBE2T	17	C4B	25	DNA2L	34
CIC	6	UHRF1	17	CDCA3	25	DNAJB6	34
DR1	6	WIBG	17	CRYBA1	25	DTL	34
EBI3	6	ZNF207	17	DLG7	25	E2F2	34
FAM60A	6	ABCC5	18	FUSIP1	25	EMP1	34
GATA2	6	BUB1B	18	GH1	25	EXO1	34
HELLS	6	CNOT10	18	GMNN	25	FAM110A	34
HJURP	6	DCTN6	18	GOLGA8A	25	FZR1	34
HP1BP3	6	DNAJA1	18	GTSE1	25	GOLGA8A	34
JMJD1C	6	DZIP3	18	HIST1H2AC	25	HCG_2026038	34
KIAA1524	6	FLAD1	18	KMO	25	HDAC3	34
MAPK13	6	GOLGA8A	18	PCDH7	25	HIST3H2A	34
MKI67	6	HN1	18	PPP3CA	25	HP1BP3	34
MSH2	6	KCTD2	18	PRPS2	25	KIAA1641	34
MYCBP2	6	M6PRBP1	18	PTPN9	25	KIF11	34
NLRP2	6	MCM6	18	RDH11	25	KIFC1	34
PAK1IP1	6	NEIL3	18	SRD5A1	25	LRRC17	34
PBK	6	NUP160	18	STK17B	25	LYRM7	34
PSMD11	6	OLR1	18	USP1	25	MPHOSPH1	34
TSG101	6	RRM2	18	ADAM22	26	N4BP3	34
ADCK2	7	SFRS3	18	AFAP1	26	NBPF15	34
HSPA7	7	SGK1	18	AGPAT3	26	NDC80	34
RCAN1	7	SYNCRIP	18	AMD1	26	NDE1	34
ANLN	8	TMEM138	18	ARGLU1	26	PRIM2	34
AURKB	8	TNS4	18	ARL4A	26	RANGAP1	34
C15ORF29	8	USP6NL	18	BAIAP2	26	SFPQ	34

Gene Symbol	Cluster #	Gene Symbol	Cluster #	Gene Symbol	Cluster #	Gene Symbol	Cluster #
C21ORF87	8	WDR51A	18	C14ORF57	26	SLC22A3	34
CCND1	8	WDR63	18	C9ORF100	26	SS18	34
CDC25A	8	WSB1	18	CDC25B	26	TPX2	34
CDC27	8	ANKRD25	19	CDCA7	26	TROAP	34
CIC	8	AP3D1	19	CEP70	26	UBE2S	34
DCP1A	8	ARL6IP2	19	CKAP2	26	UBL3	34
FYN	8	EFHC1	19	COQ6	26	UBXD5	34
HMMR	8	FLJ13231	19	CRLS1	26	VCL	34
ITPR3	8	IFIT1	19	G3BP1	26	WDR68	34
MASTL	8	KIAA1641	19	GCSH	26	Y18H1A.11	34
MCM2	8	MKI67	19	GRK6	26	ZMYM1	34
PRR16	8	PKMYT1	19	HIST1H2AC	26	ABCA7	35
ROCK1	8	PPP2R2A	19	HIST2H4B	26	AURKA	35
SLBP	8	PWP1	19	INSIG2	26	B4GALT1	35
SPTBN1	8	SSR3	19	KAZALD1	26	BIVM	35
ANKRD10	9	TMPO	19	KRAS	26	C4BPB	35
BUB3	9	TUBB2C	19	MCAM	26	CALD1	35
C4ORF30	9	USP53	19	MET	26	CCNF	35
C6	9	ZNF24	19	MITF	26	CDKL5	35
C9ORF140	9	ANTXR1	20	NAB1	26	CDKN2AIP	35
CALM3	9	ARHGAP19	20	PHIP	26	CENPL	35
CCND1	9	ARL6IP1	20	SAPS3	26	CHML	35
CD97	9	ASIP	20	SLC4A1AP	26	CIC	35
CDC25B	9	ATAD2	20	SMC4	26	CKS2	35
CDH24	9	BMI1	20	TLE3	26	DEPDC1	35
CHGN	9	CADM1	20	TOP3A	26	FAM113A	35
DEPDC7	9	CAPN7	20	TPX2	26	GMNN	35
DIAPH3	9	CENPQ	20	TTC31	26	HLA-DRA	35
DONSON	9	ESCO2	20	TUBA1A	26	HMGCR	35
G3BP1	9	ESD	20	TYMS	26	HTF9C	35
GDF15	9	FKBP1A	20	WDR62	26	IFIT1	35
GNB3	9	H2AFX	20	WDR68	26	KIAA0802	35
GPR126	9	HIST1H2AM	20	ZBTB7A	26	NCAPD2	35
GRPEL1	9	ITGB3BP	20	CASP2	27	NDE1	35
HDAC3	9	KIF2C	20	CBX3	27	NMB	35
HELLS	9	LMO4	20	CHAF1A	27	NUSAP1	35
HRB	9	LOC199800	20	CKS1B	27	PCF11	35

Gene Symbol	Cluster #	Gene Symbol	Cluster #	Gene Symbol	Cluster #	Gene Symbol	Cluster #
HSPA2	9	NCOA5	20	DHFR	27	PSCD3	35
HSPA8	9	RAD51AP1	20	DKC1	27	RERE	35
IDI2	9	RECQL4	20	DUSP4	27	SPAG5	35
KCTD9	9	SFRS3	20	EIF2A	27	SREBF1	35
KIF23	9	SFRS7	20	HIST2H3PS2	27	STAT5B	35
KIF5B	9	SLC25A36	20	HMG20B	27	SVIP	35
KLF6	9	SUCLG2	20	HPS4	27	TAF9	35
KLF9	9	TOB2	20	KIAA1529	27	TRIM26	35
KMO	9	TUBA3C	20	KIF22	27	TUBA4A	35
LPP	9	ZFX	20	KIF5B	27	UBE2C	35
MCM2	9	ALKBH1	21	PEBP4	27	UBE2C	35
MCM8	9	AP3M2	21	PNN	27	VANGL1	35
MCM5	9	ASF1B	21	POLD3	27	VEGFC	35
NFE2L2	9	ASPHD2	21	RAB23	27	ZCCHC10	35
NFIC	9	ATF7IP	21	RBM8A	27	ZNF521	35
NUF2	9	BRIP1	21	SLC38A2	27	CCRK	36
PASK	9	C15ORF29	21	SV2B	27	DMTF1	36
PCAF	9	CDC2	21	AP4B1	28	RPS25	36
PLK1	9	CDK7	21	ASXL1	28	UHRF1	37
PSMG3	9	CTCF	21	BRD8	28	AKAP13	38
PTTG1	9	CYB5R2	21	C14ORF106	28	ANP32B	38
RRM1	9	DEPDC1B	21	CCNB1	28	ANP32E	38
RUNX1	9	DHFR	21	CDK7	28	C16ORF57	38
SCML1	9	E2F1	21	CEP350	28	C1ORF2	38
SHC1	9	EIF4E	21	CNTROB	28	CASP2	38
SLC25A27	9	ESPL1	21	DEXI	28	CDKN2AIP	38
TOPBP1	9	FAM105B	21	KIAA1333	28	CDR2	38
TTC31	9	FXR1	21	RSRC2	28	CHEK2	38
USP1	9	GAS1	21	TRAIP	28	CTR9	38
USP16	9	GAS6	21	ZPBP	28	DNAJB1	38
VANGL1	9	GATA2	21	CCNA2	29	ECT2	38
WSB1	9	GNB1	21	CDC42EP1	29	FADD	38
YWHAH	9	HIST2H4B	21	CDKN2C	29	FEM1B	38
ZNF593	9	HMG20B	21	DLG7	29	FRS2	38
ACYP1	10	HSPB8	21	GABPB2	29	GPR126	38
CD24	10	JARID1B	21	HELLS	29	GTSE1	38
CDC42EP4	10	MAP3K7IP2	21	KATNA1	29	HMGB2	38

Gene Symbol	Cluster #	Gene Symbol	Cluster #	Gene Symbol	Cluster #	Gene Symbol	Cluster #
CREBZF	10	MCM6	21	KIAA1586	29	HSF2	38
HIST2H2AA3	10	MDM1	21	MAP3K15	29	IFIT1	38
KIAA0182	10	NASP	21	MCM6	29	KCNC4	38
KPNA2	10	NFKBIL2	21	MND1	29	MLLT4	38
TAF15	10	NUCKS1	21	MZF1	29	NY-SAR-48	38
TSN	10	ODF2	21	NPAT	29	ORC3L	38
WSB1	10	OGT	21	PPP1R10	29	PLAG1	38
ANLN	11	PTP4A1	21	PRC1	29	RAD18	38
HJURP	11	RAB3A	21	RPA2	29	RFC2	38
ACPP	12	RAD54L	21	STAG1	29	RPS25	38
ASXL1	12	REEP1	21	TCERG1	29	SLC39A10	38
ATF7IP	12	SAP30BP	21	TMEM140	29	SLC44A2	38
AURKB	12	SGK1	21	VANGL1	29	SLC9A3	38
BCLAF1	12	SP1	21	ANP32E	30	STAG3L2	38
C6ORF166	12	TRIM45	21	CFLAR	30	TACC3	38
CASP8AP2	12	TUBB2A	21	E2F1	30	TOMM34	38
CLSPN	12	USP13	21	KIF23	30	TULP4	38
DET1	12	VANGL1	21	MCM4	30	TXNRD1	38
FANCG	12	VPS25	21	MCM5	30	USP1	38
GOLGA8A	12	XPO4	21	WISP1	30	ZNF217	38
HN1	12	ZBED5	21	ADH4	31	ZNF587	38

Appendix D. GSEA for optimal clusters in human *Hela* cell cycle dataset

Cluster #	Number of genes in this cluster	Number of genes in this GO category	P-value	GO category	Enriched functional category
1	8	323	1.60E-08	GO:0000278	mitotic cell cycle
	3	5	2.70E-08	GO:0004756	selenide, water dikinase activity
	3	5	2.70E-08	GO:0016781	phosphotransferase activity, paired acceptors
	10	810	1.50E-07	GO:0007049	cell cycle
	8	452	2.20E-07	GO:0022402	cell cycle process
	7	359	7.10E-07	GO:0022403	cell cycle phase
	3	20	3.00E-06	GO:0004571	mannosyl-oligosaccharide 1,2-alpha-mannosidase activity
	3	22	4.10E-06	GO:0015924	mannosyl-oligosaccharide mannosidase activity
	3	29	9.70E-06	GO:0015923	mannosidase activity
	12	2263	5.00E-05	GO:0005524	ATP binding
2	8	323	5.20E-08	GO:0000278	mitotic cell cycle
	8	359	1.20E-07	GO:0022403	cell cycle phase
	13	1436	3.40E-07	GO:0006793	phosphorus metabolic process
	13	1436	3.40E-07	GO:0006796	phosphate metabolic process
	10	810	6.30E-07	GO:0007049	cell cycle
	8	452	6.70E-07	GO:0022402	cell cycle process
	6	292	8.20E-06	GO:0000279	M phase
	10	1187	1.90E-05	GO:0016310	phosphorylation
	31	10329	2.70E-05	GO:0044237	cellular metabolic process
	15	2823	2.70E-05	GO:0032553	ribonucleotide binding
	15	2823	2.70E-05	GO:0032555	purine ribonucleotide binding
	5	218	2.90E-05	GO:0007067	mitosis
	24	6657	3.00E-05	GO:0043283	biopolymer metabolic process
	5	223	3.30E-05	GO:0000087	M phase of mitotic cell cycle
	15	2934	4.30E-05	GO:0017076	purine nucleotide binding
	16	3353	5.10E-05	GO:0000166	nucleotide binding
4	3	5	1.00E-09	GO:0006597	spermine biosynthetic process
	3	5	1.00E-09	GO:0008215	spermine metabolic process
	3	6	2.00E-09	GO:0004014	adenosylmethionine decarboxylase activity
	3	8	5.60E-09	GO:0008295	spermidine biosynthetic process
	3	9	8.40E-09	GO:0008216	spermidine metabolic process

Cluster #	Number of genes in this cluster	Number of genes in this GO category	P-value	GO category	Enriched functional category
	3	12	2.20E-08	GO:0006596	polyamine biosynthetic process
	3	15	4.50E-08	GO:0006595	polyamine metabolic process
	3	32	4.90E-07	GO:0042401	biogenic amine biosynthetic process
	3	36	7.10E-07	GO:0042398	amino acid derivative biosynthetic process
	3	51	2.10E-06	GO:0016831	carboxy-lyase activity
	3	69	5.10E-06	GO:0016830	carbon-carbon lyase activity
	3	71	5.60E-06	GO:0006576	biogenic amine metabolic process
	3	86	1.00E-05	GO:0006575	amino acid derivative metabolic process
	3	94	1.30E-05	GO:0009309	amine biosynthetic process
	3	122	2.90E-05	GO:0044271	nitrogen compound biosynthetic process
5	2	103	0.0002	GO:0009314	response to radiation
	4	31	2.50E-08	GO:0004707	MAP kinase activity
	5	157	4.70E-07	GO:0031497	chromatin assembly
	4	70	7.10E-07	GO:0004702	receptor signaling protein serine/threonine kinase activity
	5	186	1.10E-06	GO:0006323	DNA packaging
	5	209	1.90E-06	GO:0006333	chromatin assembly or disassembly
	4	100	3.00E-06	GO:0000786	nucleosome
	5	250	4.60E-06	GO:0000785	chromatin
	6	470	6.00E-06	GO:0051276	chromosome organization and biogenesis
6	4	139	1.10E-05	GO:0006334	nucleosome assembly
	20	8814	1.70E-05	GO:0043170	macromolecule metabolic process
	4	168	2.30E-05	GO:0005057	receptor signaling protein activity
	5	386	3.70E-05	GO:0044427	chromosomal part
	5	387	3.80E-05	GO:0006325	establishment and/or maintenance of chromatin architecture
	10	2263	4.00E-05	GO:0005524	ATP binding
	27	17624	4.30E-05	GO:0005488	binding
	10	2286	4.40E-05	GO:0032559	adenyl ribonucleotide binding
	10	2392	6.40E-05	GO:0030554	adenyl nucleotide binding
7	2	38	1.80E-05	GO:0019722	calcium-mediated signaling
	3	32	9.20E-07	GO:0000910	cytokinesis
	3	36	1.30E-06	GO:0006270	DNA replication initiation
8	4	223	8.70E-06	GO:0000087	M phase of mitotic cell cycle
	3	73	1.10E-05	GO:0008094	DNA-dependent ATPase activity
	3	136	7.30E-05	GO:0006261	DNA-dependent DNA replication
	9	292	7.60E-09	GO:0000279	M phase
9	8	223	1.70E-08	GO:0000087	M phase of mitotic cell cycle
	10	452	2.40E-08	GO:0022402	cell cycle process

Cluster #	Number of genes in this cluster	Number of genes in this GO category	P-value	GO category	Enriched functional category
	9	359	4.50E-08	GO:0022403	cell cycle phase
	12	810	6.90E-08	GO:0007049	cell cycle
	11	679	1.10E-07	GO:0006259	DNA metabolic process
	8	323	3.00E-07	GO:0000278	mitotic cell cycle
	7	218	3.00E-07	GO:0007067	mitosis
	6	141	4.40E-07	GO:0007218	neuropeptide signaling pathway
	8	470	4.80E-06	GO:0051276	chromosome organization and biogenesis
	26	5551	9.70E-06	GO:0005634	nucleus
	4	73	1.50E-05	GO:0008094	DNA-dependent ATPase activity
	49	16682	1.50E-05	GO:0009987	cellular process
	4	75	1.70E-05	GO:0000775	chromosome, pericentric region
	3	36	5.50E-05	GO:0006270	DNA replication initiation
12	4	14	1.40E-09	GO:0008191	metalloendopeptidase inhibitor activity
	4	60	6.60E-07	GO:0007059	chromosome segregation
	3	20	1.50E-06	GO:0004571	mannosyl-oligosaccharide 1,2-alpha-mannosidase activity
	3	22	2.10E-06	GO:0015924	mannosyl-oligosaccharide mannosidase activity
	3	29	4.90E-06	GO:0000070	mitotic sister chromatid segregation
	3	29	4.90E-06	GO:0015923	mannosidase activity
	3	30	5.40E-06	GO:0000819	sister chromatid segregation
15	7	333	5.20E-08	GO:0006260	DNA replication
	7	679	6.10E-06	GO:0006259	DNA metabolic process
	3	36	7.80E-06	GO:0006270	DNA replication initiation
	17	5551	9.40E-06	GO:0005634	nucleus
	4	132	1.20E-05	GO:0005813	centrosome
	4	136	1.30E-05	GO:0006261	DNA-dependent DNA replication
	4	146	1.80E-05	GO:0005815	microtubule organizing center
	7	810	1.90E-05	GO:0007049	cell cycle
	5	323	2.20E-05	GO:0000278	mitotic cell cycle
	2	7	2.30E-05	GO:0045931	positive regulation of mitotic cell cycle
	11	2607	4.70E-05	GO:0043234	protein complex
16	10	1312	1.30E-06	GO:0003700	transcription factor activity
	2	3	3.80E-06	GO:0004146	dihydrofolate reductase activity
	2	4	7.60E-06	GO:0006545	glycine biosynthetic process
	25	11558	2.50E-05	GO:0008152	metabolic process
	10	1870	2.90E-05	GO:0030528	transcription regulator activity

Cluster #	Number of genes in this cluster	Number of genes in this GO category	P-value	GO category	Enriched functional category
	17	5551	2.90E-05	GO:0005634	nucleus
	2	10	5.70E-05	GO:0009070	serine family amino acid biosynthetic process
17	5	76	1.80E-08	GO:0005643	nuclear pore
	5	85	3.10E-08	GO:0046930	pore complex
	6	188	4.40E-08	GO:0005635	nuclear envelope
	5	104	8.60E-08	GO:0044453	nuclear membrane part
	5	144	4.40E-07	GO:0031965	nuclear membrane
	9	1090	1.40E-06	GO:0044428	nuclear part
	12	2607	7.40E-06	GO:0043234	protein complex
	5	260	7.90E-06	GO:0019932	second-messenger-mediated signaling
	2	5	1.10E-05	GO:0046934	phosphatidylinositol-4,5-bisphosphate 3-kinase activity
	6	598	3.60E-05	GO:0031967	organelle envelope
	6	600	3.70E-05	GO:0031975	envelope
	16	5551	4.80E-05	GO:0005634	nucleus
	12	3215	6.20E-05	GO:0032991	macromolecular complex
	6	679	7.30E-05	GO:0006259	DNA metabolic process
18	10	2263	7.90E-05	GO:0005524	ATP binding
19	2	6	4.60E-06	GO:0005521	lamin binding
	2	33	0.00016	GO:0000159	protein phosphatase type 2A complex
20	12	1388	1.30E-06	GO:0006996	organelle organization and biogenesis
	7	386	2.60E-06	GO:0044427	chromosomal part
	5	153	4.70E-06	GO:0007018	microtubule-based movement
	7	454	7.60E-06	GO:0005694	chromosome
	5	174	8.80E-06	GO:0030705	cytoskeleton-dependent intracellular transport
	4	100	2.10E-05	GO:0000786	nucleosome
	15	2871	2.50E-05	GO:0016043	cellular component organization and biogenesis
	2	5	2.50E-05	GO:0004963	follicle-stimulating hormone receptor activity
	4	121	4.30E-05	GO:0003777	microtubule motor activity
21	14	810	1.70E-09	GO:0007049	cell cycle
	29	5551	1.10E-06	GO:0005634	nucleus
	7	298	3.90E-06	GO:0051726	regulation of cell cycle
	7	359	1.30E-05	GO:0022403	cell cycle phase
	2	3	1.40E-05	GO:0004146	dihydrofolate reductase activity
	34	8221	1.60E-05	GO:0043231	intracellular membrane-bounded organelle

Cluster #	Number of genes in this cluster	Number of genes in this GO category	P-value	GO category	Enriched functional category
	34	8223	1.60E-05	GO:0043227	membrane-bounded organelle
	2	4	2.80E-05	GO:0006545	glycine biosynthetic process
22	6	15	8.20E-13	GO:0004957	prostaglandin E receptor activity
	6	19	4.40E-12	GO:0004955	prostaglandin receptor activity
	6	23	1.60E-11	GO:0004953	icosanoid receptor activity
	6	23	1.60E-11	GO:0004954	prostanoid receptor activity
	13	679	8.50E-09	GO:0006259	DNA metabolic process
	9	333	1.20E-07	GO:0006260	DNA replication
	9	359	2.20E-07	GO:0022403	cell cycle phase
	6	136	1.00E-06	GO:0006261	DNA-dependent DNA replication
	9	452	1.50E-06	GO:0022402	cell cycle process
	8	354	2.30E-06	GO:0006281	DNA repair
	11	810	3.90E-06	GO:0007049	cell cycle
	3	13	3.90E-06	GO:0006268	DNA unwinding during replication
	3	15	6.10E-06	GO:0032392	DNA geometric change
	3	15	6.10E-06	GO:0032508	DNA duplex unwinding
	7	292	7.10E-06	GO:0000279	M phase
	8	414	7.40E-06	GO:0006974	response to DNA damage stimulus
	7	323	1.40E-05	GO:0000278	mitotic cell cycle
	3	20	1.50E-05	GO:0004571	mannosyl-oligosaccharide 1,2-alpha-mannosidase activity
	2	3	1.70E-05	GO:0006271	DNA strand elongation during DNA replication
	2	3	1.70E-05	GO:0022616	DNA strand elongation
	3	22	2.10E-05	GO:0015924	mannosyl-oligosaccharide mannosidase activity
	32	6657	2.40E-05	GO:0043283	biopolymer metabolic process
	6	245	3.00E-05	GO:0051301	cell division
	4	73	3.10E-05	GO:0008094	DNA-dependent ATPase activity
	8	505	3.10E-05	GO:0009719	response to endogenous stimulus
	3	29	4.80E-05	GO:0015923	mannosidase activity
	36	8319	5.40E-05	GO:0005515	protein binding
	2	5	5.80E-05	GO:0004963	follicle-stimulating hormone receptor activity
24	27	5551	7.00E-07	GO:0005634	nucleus
	17	2346	1.10E-06	GO:0043228	non-membrane-bounded organelle
	17	2346	1.10E-06	GO:0043232	intracellular non-membrane-bounded organelle
	7	333	3.50E-06	GO:0006260	DNA replication

Cluster #	Number of genes in this cluster	Number of genes in this GO category	P-value	GO category	Enriched functional category
	21	3862	4.40E-06	GO:0044446	intracellular organelle part
	21	3872	4.60E-06	GO:0044422	organelle part
	32	8221	4.90E-06	GO:0043231	intracellular membrane-bounded organelle
	32	8223	4.90E-06	GO:0043227	membrane-bounded organelle
	35	9656	5.50E-06	GO:0043229	intracellular organelle
	35	9661	5.60E-06	GO:0043226	organelle
	9	679	5.60E-06	GO:0006259	DNA metabolic process
	32	8319	6.40E-06	GO:0005515	protein binding
	7	386	9.20E-06	GO:0044427	chromosomal part
	3	28	2.20E-05	GO:0006275	regulation of DNA replication
	7	454	2.60E-05	GO:0005694	chromosome
	2	5	3.70E-05	GO:0005652	nuclear lamina
	10	1084	3.70E-05	GO:0005198	structural molecule activity
	3	34	3.90E-05	GO:0006284	base-excision repair
25	2	3	1.80E-06	GO:0031616	spindle pole centrosome
	4	162	7.30E-06	GO:0005179	hormone activity
26	3	3	2.70E-09	GO:0004799	thymidylate synthase activity
	3	3	2.70E-09	GO:0006231	dTMP biosynthetic process
	3	3	2.70E-09	GO:0009157	deoxyribonucleoside monophosphate biosynthetic process
	3	3	2.70E-09	GO:0009162	deoxyribonucleoside monophosphate metabolic process
	3	3	2.70E-09	GO:0009176	pyrimidine deoxyribonucleoside monophosphate metabolic process
	3	3	2.70E-09	GO:0009177	pyrimidine deoxyribonucleoside monophosphate biosynthetic process
	3	3	2.70E-09	GO:0042083	5,10-methylenetetrahydrofolate-dependent methyltransferase activity
	3	3	2.70E-09	GO:0046073	dTMP metabolic process
	3	5	2.70E-08	GO:0006597	spermine biosynthetic process
	3	5	2.70E-08	GO:0008215	spermine metabolic process
	3	5	2.70E-08	GO:0009129	pyrimidine nucleoside monophosphate metabolic process
	3	5	2.70E-08	GO:0009130	pyrimidine nucleoside monophosphate biosynthetic process
	3	5	2.70E-08	GO:0009221	pyrimidine deoxyribonucleotide biosynthetic process
	3	6	5.40E-08	GO:0004014	adenosylmethionine decarboxylase activity
	3	6	5.40E-08	GO:0009263	deoxyribonucleotide biosynthetic process
	3	8	1.50E-07	GO:0008295	spermidine biosynthetic process
	3	9	2.30E-07	GO:0008216	spermidine metabolic process

Cluster #	Number of genes in this cluster	Number of genes in this GO category	P-value	GO category	Enriched functional category
	3	12	5.90E-07	GO:0006596	polyamine biosynthetic process
	3	14	9.80E-07	GO:0009219	pyrimidine deoxyribonucleotide metabolic process
	3	15	1.20E-06	GO:0006595	polyamine metabolic process
	3	19	2.60E-06	GO:0009262	deoxyribonucleotide metabolic process
	3	32	1.30E-05	GO:0042401	biogenic amine biosynthetic process
	3	33	1.40E-05	GO:0009123	nucleoside monophosphate metabolic process
	3	33	1.40E-05	GO:0009124	nucleoside monophosphate biosynthetic process
	3	36	1.90E-05	GO:0042398	amino acid derivative biosynthetic process
	3	38	2.20E-05	GO:0006221	pyrimidine nucleotide biosynthetic process
	3	51	5.40E-05	GO:0006220	pyrimidine nucleotide metabolic process
	3	51	5.40E-05	GO:0016831	carboxy-lyase activity
27	3	9	5.30E-08	GO:0016538	cyclin-dependent protein kinase regulator activity
	9	1388	1.80E-06	GO:0006996	organelle organization and biogenesis
	2	3	2.30E-06	GO:0004146	dihydrofolate reductase activity
	2	4	4.60E-06	GO:0006545	glycine biosynthetic process
	5	386	2.10E-05	GO:0044427	chromosomal part
	2	10	3.40E-05	GO:0009070	serine family amino acid biosynthetic process
	5	454	4.50E-05	GO:0005694	chromosome
	5	470	5.30E-05	GO:0051276	chromosome organization and biogenesis
	6	810	6.40E-05	GO:0007049	cell cycle
28	4	245	3.90E-06	GO:0051301	cell division
29	6	292	8.00E-08	GO:0000279	M phase
	6	323	1.50E-07	GO:0000278	mitotic cell cycle
	4	75	3.40E-07	GO:0005819	spindle
	5	223	7.30E-07	GO:0000087	M phase of mitotic cell cycle
	3	24	8.50E-07	GO:0000922	spindle pole
	2	3	1.80E-06	GO:0031616	spindle pole centrosome
	4	245	3.70E-05	GO:0051301	cell division
30	2	44	8.70E-05	GO:0030693	caspase activity
31	8	638	1.90E-06	GO:0007010	cytoskeleton organization and biogenesis
	4	75	3.30E-06	GO:0000775	chromosome, pericentric region
	10	1388	1.20E-05	GO:0006996	organelle organization and biogenesis
	6	386	1.30E-05	GO:0044427	chromosomal part
	5	238	1.70E-05	GO:0030036	actin cytoskeleton organization and biogenesis

Cluster #	Number of genes in this cluster	Number of genes in this GO category	P-value	GO category	Enriched functional category
	5	256	2.40E-05	GO:0030029	actin filament-based process
	6	452	3.10E-05	GO:0022402	cell cycle process
	6	454	3.20E-05	GO:0005694	chromosome
	12	2346	4.10E-05	GO:0043228	non-membrane-bounded organelle
	12	2346	4.10E-05	GO:0043232	intracellular non-membrane-bounded organelle
	5	292	4.60E-05	GO:0000279	M phase
32	3	9	1.30E-06	GO:0016538	cyclin-dependent protein kinase regulator activity
	6	202	1.20E-05	GO:0019787	small conjugating protein ligase activity
	6	239	3.10E-05	GO:0016881	acid-amino acid ligase activity
	7	386	5.20E-05	GO:0044427	chromosomal part
33	7	679	2.60E-08	GO:0006259	DNA metabolic process
	5	810	4.80E-05	GO:0007049	cell cycle
	4	414	5.70E-05	GO:0006974	response to DNA damage stimulus
	11	6657	7.10E-05	GO:0043283	biopolymer metabolic process
	10	5551	0.00011	GO:0005634	nucleus
34	11	292	1.70E-10	GO:0000279	M phase
	10	218	1.80E-10	GO:0007067	mitosis
	10	223	2.20E-10	GO:0000087	M phase of mitotic cell cycle
	10	245	5.60E-10	GO:0051301	cell division
	11	359	1.50E-09	GO:0022403	cell cycle phase
	10	323	8.00E-09	GO:0000278	mitotic cell cycle
	11	452	1.60E-08	GO:0022402	cell cycle process
	3	5	1.50E-07	GO:0006597	spermine biosynthetic process
	3	5	1.50E-07	GO:0008215	spermine metabolic process
	3	6	3.00E-07	GO:0004014	adenosylmethionine decarboxylase activity
	3	8	8.40E-07	GO:0008295	spermidine biosynthetic process
	32	5551	8.90E-07	GO:0005634	nucleus
	3	9	1.30E-06	GO:0008216	spermidine metabolic process
	3	9	1.30E-06	GO:0016538	cyclin-dependent protein kinase regulator activity
	15	1388	1.40E-06	GO:0006996	organelle organization and biogenesis
	9	470	2.60E-06	GO:0051276	chromosome organization and biogenesis
	3	12	3.30E-06	GO:0006596	polyamine biosynthetic process
	43	9656	4.00E-06	GO:0043229	intracellular organelle
	5	100	5.30E-06	GO:0000786	nucleosome
	3	15	6.70E-06	GO:0006595	polyamine metabolic process

Cluster #	Number of genes in this cluster	Number of genes in this GO category	P-value	GO category	Enriched functional category
	18	2346	1.30E-05	GO:0043228	non-membrane-bounded organelle
	18	2346	1.30E-05	GO:0043232	intracellular non-membrane-bounded organelle
	8	454	1.80E-05	GO:0005694	chromosome
	47	11718	1.80E-05	GO:0044424	intracellular part
	6	225	2.20E-05	GO:0065004	protein-DNA complex assembly
	5	139	2.60E-05	GO:0006334	nucleosome assembly
	4	71	3.10E-05	GO:0006576	biogenic amine metabolic process
	37	8221	3.40E-05	GO:0043231	intracellular membrane-bounded organelle
	23	3862	4.00E-05	GO:0044446	intracellular organelle part
	23	3872	4.10E-05	GO:0044422	organelle part
	5	157	4.70E-05	GO:0031497	chromatin assembly
	9	679	4.90E-05	GO:0006259	DNA metabolic process
35	4	19	4.00E-08	GO:0007051	spindle organization and biogenesis
	8	292	6.50E-08	GO:0000279	M phase
	8	323	1.40E-07	GO:0000278	mitotic cell cycle
	7	218	1.60E-07	GO:0007067	mitosis
	7	223	1.90E-07	GO:0000087	M phase of mitotic cell cycle
	8	359	3.20E-07	GO:0022403	cell cycle phase
	7	283	9.20E-07	GO:0007017	microtubule-based process
	8	452	1.80E-06	GO:0022402	cell cycle process
	10	810	2.10E-06	GO:0007049	cell cycle
	5	114	2.30E-06	GO:0000226	microtubule cytoskeleton organization and biogenesis
	6	245	6.20E-06	GO:0051301	cell division
	3	22	9.10E-06	GO:0030261	chromosome condensation
	2	3	1.00E-05	GO:0045136	development of secondary sexual characteristics
	3	32	2.90E-05	GO:0045787	positive regulation of cell cycle
	5	202	3.70E-05	GO:0019787	small conjugating protein ligase activity
36	3	608	9.90E-05	GO:0004674	protein serine/threonine kinase activity
38	5	141	3.20E-06	GO:0007218	neuropeptide signaling pathway
	2	7	5.30E-05	GO:0005123	death receptor binding

Bibliography

[1] D. Greenbaum, et al., "Interrelating different types of genomic data, from proteome to secretome: 'oming in on function," *Genome Res*, vol. 11, pp. 1463-8, 2001.

[2] M. Arita, et al., "All systems go: launching cell simulation fueled by integrated experimental biology data," *Curr Opin Biotechnol*, vol. 16, pp. 344-9, 2005.

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

[4] K. K. Jain, "Tech.Sight. Biochips for gene spotting," *Science*, vol. 294, pp. 621-3, 2001.

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

[6] S. Fields and O. Song, "A novel genetic system to detect protein-protein interactions," *Nature*, vol. 340, pp. 245-6, 1989.

[7] H. Zhu, et al., "Global analysis of protein activities using proteome chips," *Science*, vol. 293, pp. 2101-5, 2001.

[8] T. Ito, et al., "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proc Natl Acad Sci U S A*, vol. 98, pp. 4569-74, 2001.

[9] P. Uetz, et al., "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae," *Nature*, vol. 403, pp. 623-7, 2000.

[10] S. C. De Keersmaecker, et al., "Integration of omics data: how well does it work for bacteria?," *Mol Microbiol*, vol. 62, pp. 1239-50, 2006.

[11] U. Alon, "Biological networks: the tinkerer as an engineer," *Science*, vol. 301, pp. 1866-7, 2003.

[12] E. Wit and J. McClure, *Statistics for Microarrays: Design, Analysis and Inference*: John Wiley & Sons, 2006.

[13] F. J. Staal, et al., "DNA microarrays for comparison of gene expression profiles between diagnosis and relapse in precursor-B acute lymphoblastic leukemia: choice of technique and purification influence the identification of potential diagnostic markers," *Leukemia*, vol. 17, pp. 1324-32, 2003.

[14] X. Wen, et al., "Large-scale temporal gene expression mapping of central nervous system development," *Proc Natl Acad Sci U S A*, vol. 95, pp. 334-9, 1998.

[15] A. C. Gavin, et al., "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, vol. 415, pp. 141-7, 2002.

[16] L. Giot, et al., "A protein interaction map of Drosophila melanogaster," *Science*, vol. 302, pp. 1727-36, 2003.

[17] S. Li, et al., "A map of the interactome network of the metazoan C. elegans," *Science*, vol. 303, pp. 540-3, 2004.

[18] G. D. Bader, et al., "BIND: the Biomolecular Interaction Network Database," *Nucleic Acids Res*, vol. 31, pp. 248-50, 2003.

[19] S. Peri, et al., "Development of human protein reference database as an initial platform for approaching systems biology in humans," *Genome Res*, vol. 13, pp. 2363-71, 2003.

[20] A. K. Ramani, et al., "Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome," *Genome Biol*, vol. 6, pp. R40, 2005.

[21] B. Lehner and A. G. Fraser, "A first-draft human protein-interaction map," *Genome Biol*, vol. 5, pp. R63, 2004.

[22] K. R. Brown and I. Jurisica, "Online predicted human interaction database," *Bioinformatics*, vol. 21, pp. 2076-82, 2005.

[23] M. Persico, et al., "HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms," *BMC Bioinformatics*, vol. 6 Suppl 4, pp. S21, 2005.

[24] J. F. Rual, et al., "Towards a proteome-scale map of the human protein-protein interaction network," *Nature*, vol. 437, pp. 1173-8, 2005.

[25] U. Stelzl, et al., "A human protein-protein interaction network: a resource for annotating the proteome," *Cell*, vol. 122, pp. 957-68, 2005.

[26] N. Przulj, et al., "Functional topology in a network of protein interactions," *Bioinformatics*, vol. 20, pp. 340-8, 2004.

[27] 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, 2002.

[28] V. Matys, et al., "TRANSFAC: transcriptional regulation, from patterns to profiles," *Nucleic Acids Res*, vol. 31, pp. 374-8, 2003.

[29] M. Ashburner, et al., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet*, vol. 25, pp. 25-9, 2000.

[30] F. Azuaje, "What does systems biology mean for biomarker discovery?," *Expert Opinion on Medical Diagnostics*, vol. 4, pp. 1-10, 2010.

[31] C. Huttenhower, et al., "Detailing regulatory networks through large scale data integration," *Bioinformatics*, vol. 25, pp. 3267-74, 2009.

[32] A. Wagner and D. A. Fell, "The small world inside large metabolic networks," *Proc Biol Sci*, vol. 268, pp. 1803-10, 2001.

[33] H. Yu, et al., "Genomic analysis of essentiality within protein networks," *Trends Genet*, vol. 20, pp. 227-31, 2004.

[34] S. S. Shen-Orr, et al., "Network motifs in the transcriptional regulation network of Escherichia coli," *Nat Genet*, vol. 31, pp. 64-8, 2002.

[35] N. Guelzim, et al., "Topological and causal structure of the yeast transcriptional regulatory network," *Nat Genet*, vol. 31, pp. 60-3, 2002.

[36] L. H. Hartwell, et al., "From molecular to modular cell biology," *Nature*, vol. 402, pp. C47-52, 1999.

[37] E. Ravasz, et al., "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, pp. 1551-5, 2002.

[38] A. W. Rives and T. Galitski, "Modular organization of cellular networks," *Proc Natl Acad Sci U S A*, vol. 100, pp. 1128-33, 2003.

[39] A. Tanay, et al., "Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data," *Proc Natl Acad Sci U S A*, vol. 101, pp. 2981-6, 2004.

[40] D. M. Wolf and A. P. Arkin, "Motifs, modules and games in bacteria," *Curr Opin Microbiol*, vol. 6, pp. 125-34, 2003.

[41] M. M. Babu, et al., "Structure and evolution of transcriptional regulatory networks," *Curr Opin Struct Biol*, vol. 14, pp. 283-91, 2004.

[42] F. J. Bruggeman and H. V. Westerhoff, "The nature of systems biology," *Trends Microbiol*, vol. 15, pp. 45-50, 2007.

[43] J. D. Han, et al., "Evidence for dynamically organized modularity in the yeast protein-protein interaction network," *Nature*, vol. 430, pp. 88-93, 2004.

[44] A. Blais and B. D. Dynlacht, "Constructing transcriptional regulatory networks," *Genes Dev*, vol. 19, pp. 1499-511, 2005.

[45] T. I. Lee, et al., "Transcriptional regulatory networks in Saccharomyces cerevisiae," *Science*, vol. 298, pp. 799-804, 2002.

[46] D. T. Odom, et al., "Control of pancreas and liver gene expression by HNF transcription factors," *Science*, vol. 303, pp. 1378-81, 2004.

[47] L. A. Boyer, et al., "Core transcriptional regulatory circuitry in human embryonic stem cells," *Cell*, vol. 122, pp. 947-56, 2005.

[48] G. Swiers, et al., "Genetic regulatory networks programming hematopoietic stem cells and erythroid lineage specification," *Dev Biol*, vol. 294, pp. 525-40, 2006.

[49] S. Mangan and U. Alon, "Structure and function of the feed-forward loop network motif," *Proc Natl Acad Sci U S A*, vol. 100, pp. 11980-5, 2003.

[50] S. Mangan, et al., "The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks," *J Mol Biol*, vol. 334, pp. 197-204, 2003.

[51] R. Milo, et al., "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, pp. 824-7, 2002.

[52] L. A. Saddic, et al., "The LEAFY target LMI1 is a meristem identity regulator and acts together with LEAFY to regulate expression of CAULIFLOWER," *Development*, vol. 133, pp. 1673-82, 2006.

[53] N. Iranfar, et al., "Transcriptional regulation of post-aggregation genes in Dictyostelium by a feed-forward loop involving GBF and LagC," *Dev Biol*, vol. 290, pp. 460-9, 2006.

[54] R. Milo, et al., "Superfamilies of evolved and designed networks," *Science*, vol. 303, pp. 1538-42, 2004.

[55] U. Alon, "Network motifs: theory and experimental approaches," *Nat Rev Genet*, vol. 8, pp. 450-61, 2007.

[56] Y. Zhang, et al., "Network motif-based identification of transcription factor-target gene relationships by integrating multi-source biological data," *BMC Bioinformatics*, vol. 9, pp. 203, 2008.

[57] H. Maria, et al., "On Clustering Validation Techniques," *J. Intell. Inf. Syst.*, vol. 17, pp. 107-145, 2001.

[58] A. J. Hartemink, et al., "Combining location and expression data for principled discovery of genetic regulatory network models," *Pac Symp Biocomput*, pp. 437-49, 2002.

[59] I. Simon, et al., "Serial regulation of transcriptional regulators in the yeast cell cycle," *Cell*, vol. 106, pp. 697-708, 2001.

[60] 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, 1998.

[61] S. Tavazoie, et al., "Systematic determination of genetic network architecture," *Nat Genet*, vol. 22, pp. 281-5, 1999.

[62] E. Wang, et al., "Cancer systems biology: exploring cancer-associated genes on cellular networks," *Cell Mol Life Sci*, vol. 64, pp. 1752-62, 2007.

[63] 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, 2005.

[64] L. Tian, et al., "Discovering statistically significant pathways in expression profiling studies," *Proc Natl Acad Sci U S A*, vol. 102, pp. 13544-9, 2005.

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

[66] S. E. Calvano, et al., "A network-based analysis of systemic inflammation in humans," *Nature*, vol. 437, pp. 1032-7, 2005.

[67] T. Ideker, et al., "Discovering regulatory and signalling circuits in molecular interaction networks," *Bioinformatics*, vol. 18 Suppl 1, pp. S233-40, 2002.

[68] J. Chen and B. Yuan, "Detecting functional modules in the yeast protein-protein interaction network," *Bioinformatics*, vol. 22, pp. 2283-90, 2006.

[69] H. Y. Chuang, et al., "Network-based classification of breast cancer metastasis," *Mol Syst Biol*, vol. 3, pp. 140, 2007.

[70] S. Chu, et al., "The transcriptional program of sporulation in budding yeast," *Science*, vol. 282, pp. 699-705, 1998.

[71] M. B. Eisen, et al., "Cluster analysis and display of genome-wide expression patterns," *Proc Natl Acad Sci U S A*, vol. 95, pp. 14863-8, 1998.

[72] P. Toronen, et al., "Analysis of gene expression data using self-organizing maps," *FEBS Lett*, vol. 451, pp. 142-6, 1999.

[73] A. Jain and R. Dubes, *Algorithms for clustering data*. New Jersey: Prentice Hall, 1998.

[74] J. Dunn, "Well separated clusters and optimal fuzzy partitions," *Journal of Cybernetics*, vol. 4, pp. 95-104, 1974.

[75] J. C. Bezdek and N. R. Pal, "Some new indexes of cluster validity," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 28, pp. 301-315, 1998.

[76] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 1, pp. 224-227, 1979.

[77] J. Handl, et al., "Computational cluster validation in post-genomic data analysis," *Bioinformatics*, vol. 21, pp. 3201-12, 2005.

[78] K. Y. Yeung, et al., "Validating clustering for gene expression data," *Bioinformatics*, vol. 17, pp. 309-18, 2001.

[79] A. Ben-Hur, et al., "A stability based method for discovering structure in clustered data," *Pac Symp Biocomput*, pp. 6-17, 2002.

[80] J. M. Raser and E. K. O'Shea, "Noise in gene expression: origins, consequences, and control," *Science*, vol. 309, pp. 2010-3, 2005.

[81] L. Klebanov and A. Yakovlev, "How high is the level of technical noise in microarray data?," *Biol Direct*, vol. 2, pp. 9, 2007.

[82] F. D. Gibbons and F. P. Roth, "Judging the quality of gene expression-based clustering methods using gene annotation," *Genome Res*, vol. 12, pp. 1574-81, 2002.

[83] S. Datta and S. Datta, "Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes," *BMC Bioinformatics*, vol. 7, pp. 397, 2006.

[84] I. Gat-Viks, et al., "Scoring clustering solutions by their biological relevance," *Bioinformatics*, vol. 19, pp. 2381-9, 2003.

[85] R. Loganantharaj, et al., "Metric for Measuring the Effectiveness of Clustering of DNA Microarray Expression," *BMC Bioinformatics*, vol. 7 Suppl 2, pp. S5, 2006.

[86] M. L. Whitfield, et al., "Identification of genes periodically expressed in the human cell cycle and their expression in tumors," *Mol Biol Cell*, vol. 13, pp. 1977-2000, 2002.

[87] R. J. Cho, et al., "A genome-wide transcriptional analysis of the mitotic cell cycle," *Mol Cell*, vol. 2, pp. 65-73, 1998.

[88] J. B. Macqueen, "Some methods for classification and analysis of multivariate observations," presented at Procedings of the Fifth Berkeley Symposium on Math, Statistics, and Probability, 1967.

[89] T. Kohonen, "Self-Organizing Maps," 1997.

[90] P. Tamayo, et al., "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," *Proc Natl Acad Sci U S A*, vol. 96, pp. 2907-12, 1999.

[91] I. Gath and A. B. Geva, "Unsupervised optimal fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 773-781, 1989.

[92] C. B. James, *Pattern Recognition with Fuzzy Objective Function Algorithms*: Kluwer Academic Publishers, 1981.

[93] E. G. Donald and C. K. William, "Fuzzy clustering with a fuzzy covariance matrix," presented at Decision and Control including the 17th Symposium on Adaptive Processes, 1978 IEEE Conference on, 1978.

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

[95] S. Kullback and R. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, pp. 79-86, 1951.

[96] J. Kasturi, et al., "An information theoretic approach for analyzing temporal patterns of gene expression," *Bioinformatics*, vol. 19, pp. 449-58, 2003.

[97] O. Troyanskaya, et al., "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, pp. 520-5, 2001.

[98] R. Clarke, et al., "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data," *Nat Rev Cancer*, vol. 8, pp. 37-49, 2008.

[99] W. Bubitzky, et al., *Fundamentals of Data Mining in Genomics and Proteomics*. New York: Springer, 2007.

[100] B. Ren, et al., "Genome-wide location and function of DNA binding proteins," *Science*, vol. 290, pp. 2306-9, 2000.

[101] I. Holmes and W. J. Bruno, "Evolutionary HMMs: a Bayesian approach to multiple alignment," *Bioinformatics*, vol. 17, pp. 803-20, 2001.

[102] E. Segal, et al., "Rich probabilistic models for gene expression," *Bioinformatics*, vol. 17 Suppl 1, pp. S243-52, 2001.

[103] K. Tuncay, et al., "Transcriptional regulatory networks via gene ontology and expression data," *In Silico Biol*, vol. 7, pp. 21-34, 2007.

[104] K. Nasmyth and L. Dirick, "The role of SW14 and SW16 in the activity of Gl cyclins in yeast," *Cell*, vol. 66, pp. 995-1013, 1991.

[105] M. Naraghi and E. Neher, "Linearized buffered Ca2+ diffusion in microdomains and its implications for calculation of [Ca2+] at the mouth of a calcium channel," *J Neurosci*, vol. 17, pp. 6961-73, 1997.

[106] T. Chen, et al., "Modeling gene expression with differential equations," *Pac Symp Biocomput*, pp. 29-40, 1999.

[107] P. D'Haeseleer, et al., "Linear modeling of mRNA expression levels during CNS development and injury," *Pac Symp Biocomput*, pp. 41-52, 1999.

[108] I. Shmulevich, et al., "Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks," *Bioinformatics*, vol. 18, pp. 261-74, 2002.

[109] P. J. Woolf and Y. Wang, "A fuzzy logic approach to analyzing gene expression data," *Physiol Genomics*, vol. 3, pp. 9-15, 2000.

[110] H. Ressom, et al., "Increasing the efficiency of fuzzy logic-based gene expression data analysis," *Physiol Genomics*, vol. 13, pp. 107-17, 2003.

[111] N. Friedman, et al., "Using Bayesian networks to analyze expression data," *J Comput Biol.*, vol. 7, pp. 601-620, 2000.

[112] H. W. Ressom, et al., "Inferring network interactions using recurrent neural networks and particle swarm optimization," presented at Proceedings of the First International Conference on Computational Systems Biology, Shanghai, China, 2006.

[113] I. Maraziotis, et al., "Gene networks inference from expression data using a recurrent neuro-fuzzy approach," presented at Conf Proc IEEE Eng Med Biol Soc., 2005.

[114] J. H. Chiang and S. Y. Chao, "Modeling human cancer-related regulatory modules by GA-RNN hybrid algorithms," *BMC Bioinformatics*, vol. 8, pp. 91, 2007.

[115] K. Y. Yeung, et al., "From co-expression to co-regulation: how many microarray experiments do we need?," *Genome Biol*, vol. 5, pp. R48, 2004.

[116] G. F. Berriz, et al., "Characterizing gene sets with FuncAssociate," *Bioinformatics*, vol. 19, pp. 2502-4, 2003.

[117] M. J. De Hoon, et al., "Statistical analysis of a small set of time-ordered gene expression data using linear splines," *Bioinformatics*, vol. 18, pp. 1477-85, 2002.

[118] S. Wernicke and F. Rasche, "FANMOD: a tool for fast network motif detection," *Bioinformatics*, vol. 22, pp. 1152-3, 2006.

[119] E. Yeger-Lotem, et al., "Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction," *Proc Natl Acad Sci U S A*, vol. 101, pp. 5934-9, 2004.

[120] M. Mitchell, *An introduction to genetic algorithm*: MIT Press, 1998.

[121] J. Kennedy and R. C. Eberhart, "Particle swarm optimization," *Proceedings of the 1995 IEEE International Conference on Neural Networks (Perth, Australia)*, vol. IV, pp. 1942-1948, 1995.

[122] B. Birge, "PSOt - a particle swarm optimization toolbox for use with Matlab," presented at Swarm Intelligence Symposium, 2003. SIS '03. Proceedings of the 2003 IEEE, 2003.

[123] P. Werbos, "Backpropagation through time: what it does and how to do it," presented at Proceedings of the IEEE In Proceedings of the IEEE, 2002.

[124] B. E. Perrin, et al., "Gene networks inference using dynamic Bayesian networks," *Bioinformatics*, vol. 19 Suppl 2, pp. ii138-48, 2003.

[125] J. Zhu and M. Q. Zhang, "SCPD: a promoter database of the yeast Saccharomyces cerevisiae," *Bioinformatics*, vol. 15, pp. 607-11, 1999.

[126] M. C. Costanzo, et al., "YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information," *Nucleic Acids Res*, vol. 29, pp. 75-9, 2001.

[127] V. R. Iyer, et al., "Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF," *Nature*, vol. 409, pp. 533-8, 2001.

[128] H. W. Mewes, et al., "MIPS: a database for genomes and protein sequences," *Nucleic Acids Res*, vol. 30, pp. 31-4, 2002.

[129] H. W. Ressom, et al., "Inferring network interactions using recurrent neural networks and swarm intelligence," *Proceedings of the 28th IEEE Engineering in Medicine and Biology Society Annual International Conference, New York City, NY*, pp. 4241-4244, 2006.

[130] J. M. Cherry, et al., "Genetic and physical maps of Saccharomyces cerevisiae," *Nature*, vol. 387, pp. 67-73, 1997.

[131] K. A. Romer, et al., "WebMOTIFS: automated discovery, filtering and scoring of DNA sequence motifs using multiple programs and Bayesian approaches," *Nucleic Acids Res*, vol. 35, pp. W217-20, 2007.

[132] D. Weigel and H. Jackle, "The fork head domain: a novel DNA binding motif of eukaryotic transcription factors?," *Cell*, vol. 63, pp. 455-6, 1990.

[133] C. T. Harbison, et al., "Transcriptional regulatory code of a eukaryotic genome," *Nature*, vol. 431, pp. 99-104, 2004.

[134] E. Wingender, et al., "The TRANSFAC system on gene expression regulation," *Nucleic Acids Res*, vol. 29, pp. 281-3, 2001.

[135] B. Ren, et al., "E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints," *Genes Dev*, vol. 16, pp. 245-56, 2002.

[136] S. T. Shibutani, et al., "Intrinsic negative cell cycle regulation provided by PIP box- and Cul4Cdt2-mediated destruction of E2f1 during S phase," *Dev Cell*, vol. 15, pp. 890-900, 2008.

[137] L. Salwinski, et al., "The Database of Interacting Proteins: 2004 update," *Nucleic Acids Res*, vol. 32, pp. D449-51, 2004.

[138] C. Stark, et al., "BioGRID: a general repository for interaction datasets," *Nucleic Acids Res*, vol. 34, pp. D535-9, 2006.

[139] H. Hermjakob, et al., "IntAct: an open source molecular interaction database," *Nucleic Acids Res*, vol. 32, pp. D452-5, 2004.

[140] G. Joshi-Tope, et al., "Reactome: a knowledgebase of biological pathways," *Nucleic Acids Res*, vol. 33, pp. D428-32, 2005.

[141] Y. Takahashi, et al., "Analysis of promoter binding by the E2F and pRB families in vivo: distinct E2F proteins mediate activation and repression," *Genes Dev*, vol. 14, pp. 804-16, 2000.

[142] S. Ishida, et al., "Role for E2F in control of both DNA replication and mitotic functions as revealed from DNA microarray analysis," *Mol Cell Biol*, vol. 21, pp. 4684-99, 2001.

[143] J. W. Zhu, et al., "E2F1 and E2F2 determine thresholds for antigen-induced T-cell proliferation and suppress tumorigenesis," *Mol Cell Biol*, vol. 21, pp. 8547-64, 2001.

[144] J. Essers, et al., "Nuclear dynamics of PCNA in DNA replication and repair," *Mol Cell Biol*, vol. 25, pp. 9350-9, 2005.

[145] R. Hong and D. Chakravarti, "The human proliferating Cell nuclear antigen regulates transcriptional coactivator p300 activity and promotes transcriptional repression," *J Biol Chem*, vol. 278, pp. 44505-13, 2003.

[146] P. J. Martin, et al., "The proliferating cell nuclear antigen regulates retinoic acid receptor transcriptional activity through direct protein-protein interaction," *Nucleic Acids Res*, vol. 33, pp. 4311-21, 2005.

[147] A. P. Bracken, et al., "E2F target genes: unraveling the biology," *Trends Biochem Sci*, vol. 29, pp. 409-17, 2004.

[148] C. W. Li, et al., "Construction and Clarification of Dynamic Gene Regulatory Network of Cancer Cell Cycle via Microarray Data," *Cancer Inform*, vol. 2, pp. 223-41, 2007.

[149] M. J. van de Vijver, et al., "A gene-expression signature as a predictor of survival in breast cancer," *N Engl J Med*, vol. 347, pp. 1999-2009, 2002.

[150] S. Ramaswamy, et al., "Multiclass cancer diagnosis using tumor gene expression signatures," *Proc Natl Acad Sci U S A*, vol. 98, pp. 15149-54, 2001.

[151] T. R. Golub, et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531-7, 1999.

[152] T. Hastie, et al., "'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns," *Genome Biol*, vol. 1, pp. RESEARCH0003, 2000.

[153] A. Ben-Dor, et al., "Clustering gene expression patterns," *J Comput Biol*, vol. 6, pp. 281-97, 1999.

[154] K. B. Duan, et al., "Multiple SVM-RFE for gene selection in cancer classification with expression data," *IEEE Trans Nanobioscience*, vol. 4, pp. 228-34, 2005.

[155] M. P. Brynildsen and J. J. Collins, "Systems biology makes it personal," *Mol Cell*, vol. 34, pp. 137-8, 2009.

[156] L. J. van 't Veer, et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, pp. 530-6, 2002.

[157] Y. Wang, et al., "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *Lancet*, vol. 365, pp. 671-9, 2005.

[158] L. Ein-Dor, et al., "Outcome signature genes in breast cancer: is there a unique set?," *Bioinformatics*, vol. 21, pp. 171-8, 2005.

[159] S. Michiels, et al., "Prediction of cancer outcome with microarrays: a multiple random validation strategy," *Lancet*, vol. 365, pp. 488-92, 2005.

[160] J. P. Ioannidis, "Microarrays and molecular research: noise discovery?," *Lancet*, vol. 365, pp. 454-5, 2005.

[161] P. E. Lonning, et al., "Genomics in breast cancer-therapeutic implications," *Nat Clin Pract Oncol*, vol. 2, pp. 26-33, 2005.

[162] S. A. Tomlins, et al., "Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer," *Science*, vol. 310, pp. 644-8, 2005.

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

[164] M. E. Futschik, et al., "Comparison of human protein-protein interaction maps," *Bioinformatics*, vol. 23, pp. 605-11, 2007.

[165] E. Lee, et al., "Inferring pathway activity toward precise disease classification," *PLoS Comput Biol*, vol. 4, pp. e1000217, 2008.

[166] A. H. Bild, et al., "Oncogenic pathway signatures in human cancers as a guide to targeted therapies," *Nature*, vol. 439, pp. 353-7, 2006.

[167] Z. Guo, et al., "Towards precise classification of cancers based on robust gene functional expression profiles," *BMC Bioinformatics*, vol. 6, pp. 58, 2005.

[168] I. D. Fraser and R. N. Germain, "Navigating the network: signaling cross-talk in hematopoietic cells," *Nat Immunol*, vol. 10, pp. 327-31, 2009.

[169] E. J. Edelman, et al., "Modeling cancer progression via pathway dependencies," *PLoS Comput Biol*, vol. 4, pp. e28, 2008.

[170] S. Hua, et al., "Genomic analysis of estrogen cascade reveals histone variant H2A.Z associated with breast cancer progression," *Mol Syst Biol*, vol. 4, pp. 188, 2008.

[171] K. M. Mani, et al., "A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas," *Mol Syst Biol*, vol. 4, pp. 169, 2008.

[172] D. Hwang, et al., "A systems approach to prion disease," *Mol Syst Biol*, vol. 5, pp. 252, 2009.

[173] J. T. Chang, et al., "A genomic strategy to elucidate modules of oncogenic pathway signaling networks," *Mol Cell*, vol. 34, pp. 104-14, 2009.

[174] W. K. Lim, et al., "Master regulators used as breast cancer metastasis classifier," *Pac Symp Biocomput*, pp. 504-15, 2009.

[175] K. Basso, et al., "Reverse engineering of regulatory networks in human B cells," *Nat Genet*, vol. 37, pp. 382-90, 2005.

[176] P. F. Jonsson and P. A. Bates, "Global topological features of cancer proteins in the human interactome," *Bioinformatics*, vol. 22, pp. 2291-7, 2006.

[177] S. Wuchty, "Evolution and topology in the yeast protein interaction network," *Genome Res*, vol. 14, pp. 1310-4, 2004.

[178] H. B. Fraser, et al., "Evolutionary rate in the protein interaction network," *Science*, vol. 296, pp. 750-2, 2002.

[179] R. Saeed and C. M. Deane, "Protein protein interactions, evolutionary rate, abundance and age," *BMC Bioinformatics*, vol. 7, pp. 128, 2006.

[180] A. Awan, et al., "Regulatory network motifs and hotspots of cancer genes in a mammalian cellular signalling network," *IET Syst Biol*, vol. 1, pp. 292-7, 2007.

[181] S. Loi, et al., "Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen," *BMC Genomics*, vol. 9, pp. 239, 2008.

[182] E. Segal, et al., "A module map showing conditional activity of expression modules in cancer," *Nat Genet*, vol. 36, pp. 1090-8, 2004.

[183] P. Pagel, et al., "The MIPS mammalian protein-protein interaction database," *Bioinformatics*, vol. 21, pp. 832-4, 2005.

[184] A. Chatr-aryamontri, et al., "MINT: the Molecular INTeraction database," *Nucleic Acids Res*, vol. 35, pp. D572-4, 2007.

[185] T. Beuming, et al., "PDZBase: a protein-protein interaction database for PDZ-domains," *Bioinformatics*, vol. 21, pp. 827-8, 2005.

[186] A. Ma'ayan, et al., "Formation of regulatory patterns during signal propagation in a Mammalian cellular network," *Science*, vol. 309, pp. 1078-83, 2005.

[187] G. D. Tourassi, et al., "Application of the mutual information criterion for feature selection in computer-aided diagnosis," *Med Phys*, vol. 28, pp. 2394-402, 2001.

[188] M. Dorigo, et al., "Ant algorithms for discrete optimization," *Artif Life*, vol. 5, pp. 137-72, 1999.

[189] M. Perretto and H. S. Lopes, "Reconstruction of phylogenetic trees using the ant colony optimization paradigm " *Genet. Mol. Res.*, vol. 4, pp. 581-589, 2005.

[190] B. E. Boser, et al., "A training algorithm for optimal margin classifiers," presented at Proceedings of the Fifth Annual Workshop on Computational Learning Theory, 1992.

[191] U. M. Braga-Neto and E. R. Dougherty, "Is cross-validation valid for small-sample microarray classification?," *Bioinformatics*, vol. 20, pp. 374-80, 2004.

[192] K. Lage, et al., "A human phenome-interactome network of protein complexes implicated in genetic disorders," *Nat Biotechnol*, vol. 25, pp. 309-16, 2007.

[193] A. Ergun, et al., "A network biology approach to prostate cancer," *Mol Syst Biol*, vol. 3, pp. 82, 2007.

[194] H. Zhu, et al., "Analysis of yeast protein kinases using protein chips," *Nat Genet*, vol. 26, pp. 283-9, 2000.

[195] "http://www.sonycsl.co.jp/person/tetsuya/Pathway/Cancer-related/cancer-related.html."

[196] "http://www.abcam.com/index.html?pageconfig=resource&rid=11069&pid=10628."