2014-07-10

Community Structure, Inference and Network-Based Markers

Gao, Shang

UNIVERSITY OF CALGARY

Network Biology:

Community Structure, Inference and Network-Based Markers

by

Shang Gao

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

CALGARY, ALBERTA

June, 2014

Abstract

In the core of system biology, it is believed that molecules within the cell act collaboratively in an organized behavior. Researchers are studying the interactions and mainly concentrate on identifying malfunctioning molecules as potential disease biomarkers. Thus, a network has become an important means to represent biological systems, and network approaches have shown substantial promise due to the simplicity in data representation and associated rich analytical apparatus. Generally speaking, the workflow of a computational system biology study means: 1.) Investigating certain elements of biological networks and their interactions, which depends on the purpose of the study. 2.) Collecting experimental high-throughput and genome-wide data and integrating computational methods to analyze the data and validate findings. In this thesis, we frame the investigations by first asking a system biology question, and then provide computational means to answer the question.

My thesis consists of three major interrelated components, as the title suggests, we first study the network structure by a novel strategy of bridging together social and biological networks based on our argument that there exist a strong analogy between humans and molecules. As social network analysis is gaining popularity in modeling real world problems, the task of applying the social network model concepts and notions to biological data is still one of the most attractive research problems to be addressed. We design computational means to find community structures and design efficient algorithms to dynamically analyze gene boundaries using geometric convexity. Our approach contributes to the new branch of applying social network mechanisms in biological data analysis, leading to new data mining strategies implied by witnessing social behaviors in gene expression analysis.

Further into the topology study of biological networks, we investigate the relationship between the multi-scalability of community structures of metabolic networks and the distributional effect of network motifs, i.e., the inference problem. We observe several patterns

through studying three organisms, including the effect of directionality of networks, homogeneity of motif-enriched communities, and motif type-specific distributions across scales. We also provide methods to quantify motif influence under the community context. Overall, our work suggests that the theoretic evolvability of modularity tightly correlates with motif distributional effect and vice versa. In this regard, we design computational tools to analyze community structure of very large networks of arbitrary types. The Multi-scale Community Finder (MCF) is the first tool in this area.

Finally we arrive at the question of how to design efficient bio-markers for complex diseases, e.g., cancer. First, it is important to understand the complexity of cancer. We believe that to understand individualized gene behavior across patients, relational status of genes needs to be considered because complex disease phenotype is often caused by cascaded failures of genetic interactions in cancer cells. We implement a framework to quantify the molecular heterogeneity of tumors from gene-gene relational perspective using co-expression networks and interactome data. Next, we present a method to reverse engineer integrative gene networks. The main advantage of our method is the integration of different quantitative and qualitative data sets in order to reconstruct a multiplex network, without necessarily imposing data constraints, such as each genomic datum needs to have the same number of entities. Another advantage of our method is that from the integrated networks, predictions can be made by propagating beliefs from seed nodes representing known knowledge. Thus, we combine data integration and network-based prediction into a single framework. We demonstrate our method through case studies using breast cancer data. Our approaches present promising results and new ways of thinking and mining complex genomic datasets.

Overall, this thesis presents a comprehensive study of biological networks and the novel application of computational means to implement the biomarker detection problem in the era of big genomic data. Finally it is important to highlight the fact that our study considers the challenges due to data heterogeneity and the diversity in the sources producing the data.

Acknowledgements

This thesis could not have been accomplished without the gracious help and supervision of my advisor, Dr. Reda Alhajj. I would like to thank him for his time and patience and also for his encouragement. I am also grateful to Dr. Jon Rokne and Dr. Douglas Demetrick, my committee members, for their time and effort in reviewing this work and their valuable feedback during the past. I also thank Dr. Mohamed Helaoui and Dr. Zongmin Ma for sitting in my defense committee and for their time to review this thesis.

I would like to acknowledge my wife, Gaozhu Wu, who has never had any hesitations to support my academic career.

I owe thanks to my parents, who have always encouraged me and been supportive of me in pursuing my goals.

Table of Contents

List of Tables

List of Figures and Illustrations

List of Symbols, Abbreviations and Nomenclature

Symbol	Definition
GEO	Gene Expression Omnibus
PPI	Protein-Protein Interaction
KEGG	Kyoto Encyclopedia of Genes and Genomes
FDR	False Discovery Rate
TSG	Tumor Suppressor Gene
PR	Precision-Recall
AUPR	Area Under the Precision Recall curve
AUPR	Area Under the Precision Recall curve
ROC	Receiver Operating Characteristic

Chapter 1

Introduction

"A big challenge for computer scientists who are considering getting involved in Systems Biology, in addition to the requirement of good level of biological foundation, is to keep open-minded and be creative in the design of modern methodologies to make a contribution to biological and computer science domain." – Eberhard O. Voit

1.1 Networks and System Biology

System biology emerged as a new interdisciplinary area, thanks to unprecedented technological and scientific progress in the last century. As the name suggests, system biology aims to study the behavior of biological systems and ultimately to predict the behavior of a system by understanding elementary constituents (i.e., biological entities) and a wide variety of individual interactions between them [79, 80], examples include protein-protein interactions and metabolic pathways [54, 51]. As such, the area of system biology provides exciting opportunities for biologists, mathematicians, and computer scientists to unify their methodologies and the understanding of biological processes from both theoretic and experimental aspects.

The motivation of studying biological systems as a whole roots in the fact that biological states and behaviors are often extremely complex so that the reductionist approach makes it difficult if not impossible to pinpoint dynamic predictors from collective biological entities [110]. The hope is to use systemic approaches, via globally representing and analyzing biological interactions in the system, to obtain further knowledge of system properties and dynamics.

In the core of system biology, network has become an important medium to represent biological systems, and network approaches have shown substantial promise to achieve the aforementioned goals with its simplicity in data representation and associated rich analytical apparatus. A network consists of vertices (a.k.a nodes) and edges (a.k.a links) that connect nodes, which effectively abstract biological entities and their interactions in a connected map [41]. For example, genetic regulations can be represented by networks of interacting genes and proteins. In computer science and mathematics, networks are also known as graphs, emphasizing the underlying data structure [42, 135]. Although system biology studies can focus on various biological systems to answer specific questions like how does the modular structure in gene regulatory networks predict cellular states, "network science" provides theoretic tools and therefore becomes the driving force of system biology. For this reason, "network biology" focuses on the study of networks in system biology, i.e., their topological structure, dynamics, and visual analytics, etc.

Biological networks serve as a powerful representation in modeling many areas of biological data such as cancer: cancer as a complex disease can be better understood by analyzing communities of heterogenous datasets [157, 159, 28, 27], as community mining in computational literature empowers the analysis of the evolution of nodes and further unravels the mystery of cancer as a dynamic process in multi-dimensional contexts. More importantly, many network principles like the social network theories can bring fresh angles of viewing biological networks [26, 19, 48], i.e., the perspective of viewing genes or proteins as actors that can influence others inside cells under different environments.

Biological networks can represent various types and levels of molecular organizations in a cell [54, 74]. Different types of biological networks include protein interaction network [183], metabolic network [51], co-expression network [185], and transcriptional network [100], etc. Each type is driven by a large amount of data accumulated to facilitate the analysis of functional significance and topological properties of the cellular system in question. In this thesis, we studied various types of biological networks for different purposes.

1.2 Background

1.2.1 Community Structure in Networks

Network science serves as a theoretical engine overall and community detection techniques become the main focus for the dynamic modularity analysis that can reveal functional significance. Cancer biology, on the other hand, is an application domain for which dynamic behaviors are to be interpreted. In this section, we shall discuss the background and related work by these two components, with recent highlights of the connection between network science and system biology.

A community is a concept originated from social networks, i.e., group organizations can often be found in a society such as friendship networks and families [111], and communities are subunits of graphs. Even though community structures are analyzed in many applications such as biological and social networks, the current literature lacks a precise definition [41]. A common topological property of communities is that links between communities are sparse whereas much densely connected nodes are expected within communities. The degree to which this topological property is quantified depends largely on the context. It is worth mentioning that the prerequisite for communities to manifest is that the graph needs to be sparse, or $n >> m$, where m is the number of edges in the graph and n is the number of nodes.

Communities are common to both social and biological networks. Further, many concepts and techniques can be applied to both networks [48, 26], so one should note their differences. A social network is an abstract representation of social structures consisting of individuals and connections or ties among individuals, such as collaboration network where individuals are scientists and two individuals are connected if they have a joined publication. The analysis of abstract structures of social networks sheds light on the understanding of real social behaviors. Biological networks on the other hand, are mostly constructed indirectly, such as experimentally validated interactions between two proteins in the PPI network, or gene-gene interactions inferred from gene expression profiles. Apart from the obvious connection between social and biological networks that both are graph-based structures, various biological networks are analogous to social networks. For example, in PPI networks, proteins can be viewed as individuals and interactions can be viewed as social connections. The advantage of such perspective is that properties or behaviors in the language of social networks can also be "mapped" to biological networks, for example, it is found that both PPI networks and collaboration networks share the "scale-free" property [68, 10].

One of the most important properties shared by both social and biological networks is the community structure, i.e., both types of networks tend to have clustered structures. This suggests that further analysis and interpretations of communities can be useful to investigating disease and genomic data. However, social networks possess properties that other types of networks do not have, for example, social networks tend to have assortative mixing patterns and high levels of clustering [119].

1.2.2 The View of Cancer

Cancer is a complex disease. Over the past decade, researchers had gained substantial understanding of the genetic causes and impacts of cancer, leading to clinical prognosis and therapeutics endeavors. In particular, the molecular and phenotypic basis of cancer have been comprehensively investigated, with the hope to discover common disease patterns in terms of biomarkers [39, 145]. Unfortunately, as all cancers are different from patient to patient, the work has not been successfully parallelized with biological development and medical practice on cellular level. Researchers either lack the data, e.g., genetic interactions in mammalian cells or the means to find hidden links in the data, i.e., associations between changes and causes under different cellular contexts and external factors.

Cancer research is revolutionized by the advent of advanced sequencing technologies: cancer genomes are being analyzed by various research communities around the world. Understanding in oncogenicity along with other aspects of cancer has been tremendously advanced. However, with the vast amount of data ever available, the conclusions seem to be dispersive. In other words, data is not well integrated in understanding the biological systems in question. With this mindset, system biology has gained popularity in many areas of computational biology for analyzing genomic data [79, 80]. The system level thinking prompts the network thinking, which serves as the core in modeling complex data and the interconnections explicitly or implicitly stored in them.

Speaking of the network thinking of cancer, cells interact with each other in tissue compartments, respond to extracellular cues, and produce outputs that ultimately determine their fate, i.e., cell chooses to suicide (programmed cell death or apoptosis) or further divide (mitosis). What is complicated in the process is the dynamics in the cancer development: how cells change and respond under specific environments, how do processes reach a balanced or stable state, and what are the consequences if such states are perturbed. In order to understand cancer as an evolutionary and context-dependent process, effective computational methods need to be in place. In this thesis, network-based approaches are proposed to model the complexity of cancer.

The simple view of cancer consists of two sets of genes: oncogenes and tumor suppressor genes (TSGs), they act as accelerator and decelerator towards and against cancer, respectively [55]. Cancer is a genetic disease that undergoes evolutionary processes: while cells undergo different phases in cell cycle regulation, different genes exert their functions in concerted manner, e.g., their products bind, activate, and signal other cellular functional units. If harmonious states were disrupted, cells turn cancerous. This is depicted in Figure 1.1.

With the loss of cell cycle control, cancers share a common phenotype: dedifferentiated (cells that are less specialized in functioning), uncontrollable cell growth (mitosis is nonstoppable or cells do not commit programmed cell death) and proliferation (dysfunction of cell-cell interactions causing tumorigenesis) [167].

Proto-oncogenes are initiators of cancer: they are genes that turn healthy/normal cells

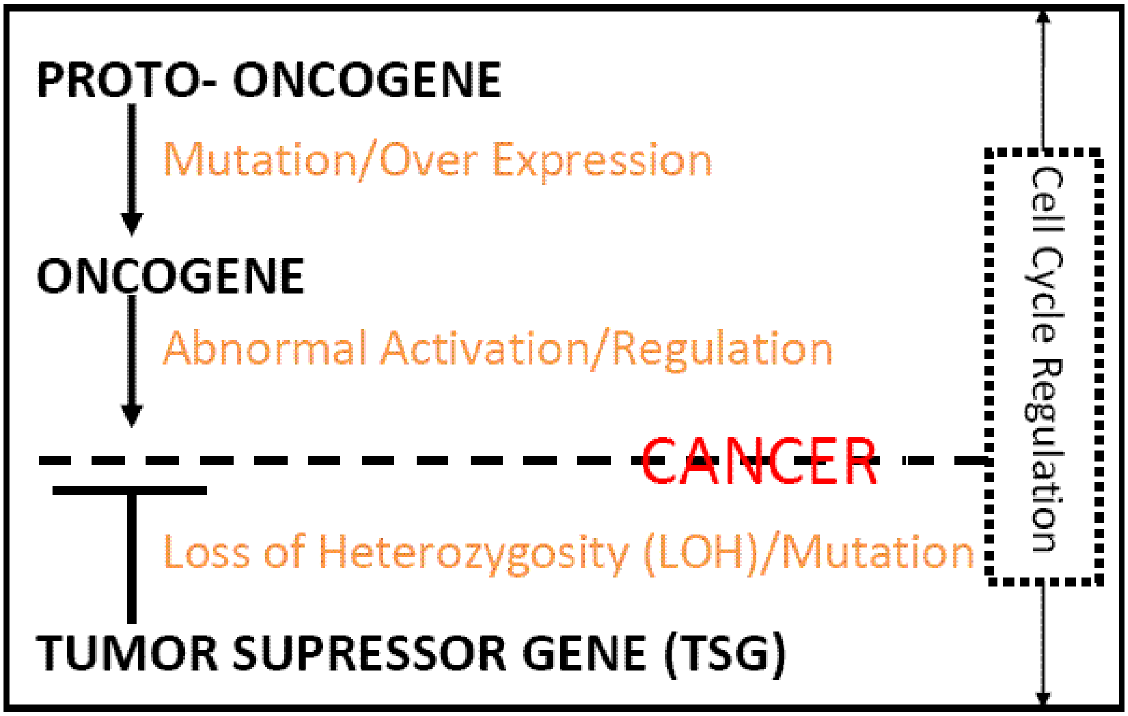

PROTO- ONCOGENE
Mutation/Over Expression
ONCOGENE
Abnormal Activation/Regulation
CANCER
Loss of Heterozygosity (LOH)/Mutation
TUMOR SUPRESSOR GENE (TSG)
Cell Cycle Regulation

Figure 1.1: The overview of cancer in cellular environment

into cancerous state. Oncogenes are mutated proto-oncogenes. It is worth noting that proto-oncogenes are not at all times execrable; as a matter of fact some genes are useful in normal cell cycle regulation. The proteins produced usually stimulate mitosis, inhibit cell differentiation, and halt apoptosis, all of which are essential for normal development of multicellular organs and tissues. When mutated to oncogenes, the protein levels that drive these functions are elevated, causing dysfunctions in cells by breaking the balancing force between mitosis and apoptosis. Because of the critical role of proto-oncogenes and oncogenes, they can be clinical drug targets.

Tumor suppressors on the other hand, act in the opposite way: they typically prevent cells from reaching cancerous state. The first TSG was found in 1986 in retinoblastoma. Different from dominant nature of oncogenes, TSGs are usually recessive: a normal cell has two copies of a gene, called alleles. The famous “two hit hypothesis” states that the inactivation of one allele does not turn cells to cancerous, but the “second” hit, i.e., the inactivation of second allele may affect the fitness of the cell, leading to uncontrollable cell growth and proliferation.

From the above discussion, we can see that the dysfunction or deregulation of the cell cycle are caused by mutations. Mutations are fundamentally the evil of all cancers [153, 17, 50, 175]. In the simplest sense, the evolutionary aspect of cancer can be viewed as a mutation-selection process, genes mutate in cells, occupy cellular compartments with increased reproduction rate. Some mutations confer growth advantage and are termed “driver” mutations whereas others that are not selected are called “passenger” mutations. Mutations are evolving infrastructures of cancer and they have received a great interest in recent research communities, thanks to the cost efficient sequencing technologies [129].

1.2.3 Network Construction

The simplest model of a social network consists of a set of actors (interchangeably called individuals) linked by certain type of relationship. For instance, in pharmacology the actors could be drugs, and two actors are connected if it is not possible for them to appear together in the same prescription. Analyzing such network will lead to communities of drugs never used together. The links may reflect either binary relationships (a missing link indicates the absence of relationship) or weighted connections to indicate the strength or degree of the relationship (which may be negative or positive). It is also possible to have more than one set of actors. For instance, drugs and diseases may be two sets of actors such that a link between a drug and a disease indicates the usage of the drug for treating the disease. The network could be analyzed to discover the most important drugs used in treating most of the diseases. The number of actors’ groups in the model specifies the degree of the mode for the social network. The two versions described above are known as one-mode and two-mode social networks. It is possible to derive two one-mode networks from a two-mode network by applying a process known as folding, which operates directly on the adjacency matrix of the two-mode network [87]. Folding is simply the multiplication of a matrix by its transpose.

Since gene expression data has become a main source to quantitatively measure the abundance of mRNA transcripts, many “reverse engineering” methods are proposed to reconstruct regulatory networks [32, 9]. However, such network reconstruction methods bring a new challenge: how to integrate different sources of biological datasets which may include multiple gene expression profiles, interaction networks, or literature-based evidence of gene-gene associations?

1.2.4 The Complexity of Community Structures

Given the networks (obtained either directly or indirectly), the goodness of community structures gives rise to the “right” decomposition of the complexity of biological processes. To exemplify, consider the cell cycle control mechanism, different cyclins and cyclin-dependent kinases regulate cellular activities both separately and collaboratively, turning cells to highly flexible (responding to signals from extracellular environment) yet harmonious (controlled division and growth) state. Decomposing these functional units in protein interaction networks therefore provides a guide to the wide variety of activities/events in the cell [138, 71].

The problem is not that simple, because cancer is an evolutionary process. This is to say that patients do not get cancer overnight, the damage to the multi-cellular tissues somehow accumulates, i.e., through deleterious mutations in the case of smoking in causing lung cancer. For this reason, snapshot of how protein clusters look like at a particular time provides little value towards prognosis [108, 17, 121].

The static mining approach is useful for understanding structural properties of communities, e.g., hubs and cliques, but less useful in the constantly changing cellular environment. With this in mind, the static mining strategies need to be extended with time resolutions, a new parameter in modularity-maximization based community mining [5, 31, 113, 85].

Speaking of community evolution alone, one can target on two scales. The first is the global scale, i.e., how do communities evolve as a whole in the network background that represents biological process (cell cycle regulation) over time [125, 57]. For example, do certain communities shrink, expand or stay unchanged relative to others? The second is the local scale, i.e., how do communities evolve if individual nodes are removed or muted by accumulative mutations in future generations? Further, what role does the specificity of genes (i.e., oncogene or TSG) play in community dynamics?

1.3 Goals of the Thesis

There are two major challenges in bioinformatics research: the first is data complexity, which means data can be noisy, incomplete and subject to different experimental protocols and laboratory conditions. On top of this, the dynamic nature of cellular processes adds another major obstacle in many endeavors. For example, current protein interaction network data is far from complete and accurate. It is apparent that such interactome data cannot be directly used for network-based biomarker detection. The system biology approaches typically study variables of interests while keep others constant. This leads to the second major challenge of how to achieve predictability given the data complexity. Although system biology approaches are promising in parallel development of experimental technologies and protocols, we are still facing data complexity problems and as a consequence, inconsistent conclusions are inevitable.

In this thesis, we noticed the inherited limitations of system biology approaches and probed to address the data complexity and the predictability issues in two directions:

1. We used social network analogies. Since many network approaches originate from social network findings, we could treat genes as social actors and analogously think social interactions as regulatory relationships between genes and their products. The data abstractions are very similar between biological and social networks, for example, in co-expression networks, we could observe important structural properties like scale-free and hub effects. Following this direction of thinking, we analyzed boundary genes from co-expression networks

in Chapter 2, and in Chapter 3, we further studied the structural relationship between motifs and community scales, which is universal in both social and biological networks. We further developed a software called ”Multi-scale Community Finder (MCF)” for detecting community structures in large networks.

2. To deal with predictability issues in biological systems, especially in the context of cancer, we followed the principled design approach. Given system-wide assumptions, we designed biomarkers based on structural heterogeneity in gene co-expression networks in Chapter 5. The principled designed approach reduces the data complexity. We followed up in Chapter 6 to design a method to evaluate the predictive performance of network-based markers by considering nodal connectedness. In the proposed method, we made the proximity assumptions in the network.

To handle data complexity and predictability simultaneously, in Chapter 7 we studied an approach to infer gene networks by integrating multiple sources of biological data, such as pathway data and multiple gene expression profiles. By the integrative reconstruction, we aimed to effectively reduce data complexity and improve predictive accuracy. As demonstrated in the case studies and the comparison with benchmark data sets, we were able to achieve better predictability using integrative networks.

1.4 Contributions

This thesis presents several contributions in the area of computational system biology. Overall, we have provided computational means to analyze high-throughput data sets and different biological networks.

We have dealt with data complexity. The computational challenge in mining high-throughput data has become a bottleneck for researchers. On one hand, the advance in profiling technologies opened up many possibilities to quantitatively analyze the data, on the other hand, the meta-analysis of such data has, in some way, led to inconsistent conclusions. For example, the inconsistent predictive performance reported for different gene sets with different patient cohorts is one example. The underlying reason is that biological data is far from comprehensive and ideal; therefore making many claims difficult, if not impossible, to prove. For instance, gene expression data sets for cancer metastasis are the "snapshot data" in the dynamic cellular environment. The conclusions solely based on such data sets inevitably inherit the nature of inconsistencies. In this thesis, we approached the data complexity from social and network perspective.

On top of the data complexity challenges, we aimed to address the predictability problem, which means to make designed predictions as if the data complexity was partially resolved. For example, we designed biomarkers based on co-expression heterogeneity and made network-based predictions using integrative networks. Despite system biology limitations, the thinking driven by the (over-)simplified model could offer fruitful lessons in biological problems. In other words, there is a tradeoff between making assumptions to reduce the complexity and making consistent and comprehensive conclusions. We next detail the contributions of each chapter.

In Chapter 2, we adapted the social network model to study genes and investigated the social inspiration by concentrating on boundary genes in expression data, because they resemble boundary nodes in social communities. We proposed three procedures for mining dynamic social communities. Our approach contributes to the new branch of the social thinking in biological systems.

In Chapter 3, we studied the relationship between multi-scale community structures of metabolic network and motif distributions. We used three model organisms to address the relationship between community scalability and motif distributional effect in metabolic networks. We investigated the question of "who drives whom", at least topologically, and further provided methods to quantify the motif distributional effect. Our study deepens the understanding of organizational principles in biological networks. The tool we used to analyze network communities was described in Chapter 4, which is the first tool, to the best of our knowledge, to deal with different types of large network in detecting community structures.

In Chapter 5, we contribute to the study of breast heterogeneity by designing biomarkers based on relational heterogeneity in co-expression networks. We found that different categories of genes stratified by the level of co-expression heterogeneity behave differently in terms of predictive performance. Our study exemplified the way to design efficient biomarkers using biological and clinical principles.

Given the biomarkers as the output of designed principle, like the ones based on the relational heterogeneity information from Chapter 5, we asked the basic question: how to evaluate the markers in the network context. Most existing methods only provide simple aggregation but ignore the connectedness in the network topology. In Chapter 6, we provided a method to quantify the predictive performance of network-based markers using an optimization method, and further added the line of evidence that most of the network-based markers are not robust.

In Chapter 7, we aimed to reconstruct integrative networks. We provided a method with a single parameter for the integration. The framework is flexible and can deal with different types of biological data. We showed the effectiveness and efficiency of the integrative reconstruction using benchmark datasets and breast cancer data. The outcome of the study showed the possibility of reducing the data complexity for more effective network-based predictions.

1.5 Organization of the Thesis

Generally speaking, the workflow of a computational system biology study means:

- Investigating certain elements of biological networks and their interactions, which depends on the purpose of the study (i.e., *what problems are we trying to address using the chosen networks?*).

- Collecting experimental high-throughput/genome-wide data and using computational methods to analyze the data and validate findings.

In the following chapters, we shall frame the investigations by first asking a system biology question, and then provide computational means to answer the question. Since system biology crosses a plethora of subfields in biological and medical sciences, we focus on community structures (a.k.a modular patterns) and the use of topological properties and information in biological networks to study complex diseases in this thesis. As examples, we studied community/modularity based methods to decompose and analyze breast cancer data and investigated methods to infer gene regulatory networks that integrate multiple sources of genomic data.

As social network analysis is gaining popularity in modeling real world problems, the task of applying the social network model concepts and notions to biological data is one of the most attractive research problems to be addressed. In **Chapter 2**, we focus on a particular set of genes that reside on the community boundaries in gene co-expression networks. Stemmed from community mining problem in social networks, peripheries of communities (i.e., boundaries) can be used to aid certain biological analysis. The proposed method consists of three parts:

1. Finding communities of gene co-expression networks through clustering.

2. Analyzing stability of community structures by Monte Carlo method.

3. Designing of dynamic adoption of boundaries using geometric convexity.

We validate our findings using breast cancer gene expression data from various studies. Our approach contributes to the new branch of applying social network mechanisms in biological data analysis, leading to new data mining strategies that associate social behaviors to biological network analysis.

In **Chapter 3**, we study the relationship between multi-scale community structures and network motifs using metabolic networks. Metabolism is a set of fundamental processes that play important roles in a plethora of biological and medical contexts. It is understood that the topological information of reconstructed metabolic networks, such as modular organization, has crucial implications on biological functions. Recent interpretations of modularity in network settings provide a view of multiple network partitions induced by different resolution parameters. Here we ask the question: How do multiple network partitions affect the organization of metabolic networks? Since network motifs are often interpreted as the superfamilies of evolved units, we further investigate their impact under multiple network partitions and investigate how does the distribution of network motifs influences the organization of metabolic networks. We study *Homo sapiens*, *Saccharomyces cerevisiae* and *Escherichia coli* metabolic networks, and analyze the relationship between different community structures and motif distribution patterns. Further, we quantify the degree to which motifs participate in the modular organization of metabolic networks.

In **Chapter 4**, we present a tool to find community structures of different types of networks. **M**ulti-scale **C**ommunity **F**inder (MCF) is a tool to profile network communities (i.e., clusters of nodes) with the control of community sizes. The controlling parameter is referred to as the scale of the network community profile. MCF is able to find communities in all major types of networks including directed, signed, bipartite, and multi-slice networks. The fast computation promotes the practicability of the tool for large-scaled analysis (e.g., protein-protein interaction and gene co-expression networks).

In **Chapter 5**, we study cancer heterogeneity using breast cancer data. It is well known that cancer is a highly heterogeneous disease, and the predictive capability of targeted gene signature approach suffers from the inter-tumor heterogeneity. Here we propose a framework to quantify the molecular heterogeneity of tumors from gene-gene relational perspective using co-expression networks and interactome data. We believe that to understand individualized gene behavior across patients, relational status of genes needs to be considered because complex disease phenotype is often caused by cascaded failures of genetic interactions in cancer cells. We quantify gene-gene relational heterogeneity from a benchmark dataset using co-expression networks inferred from microarray data, and show that genes related to breast cancer metastasis can be stratified to different classes based on their relational status obtained from pairwise comparisons of co-expression networks. Further we use the relational heterogeneity information to predict patient survival and found that relationally heterogeneous gene set is less predictive than relatively conserved cancer genes and weekly co-expressed genes in terms of metastasis. We explore heterogenous gene sets using interactome data and identified densely connected components that are causal to inter-tumor heterogeneity, and independently validate our approach with two patient cohorts. Our results demonstrate the efficiency of using heterogeneity information to design network-based markers.

In **Chapter 6**, we argue that it is necessary to use the network structures to evaluate performance of biomarkers. To address this, we aim to learn a weight coefficient for each node in the network from the quantitative measure such as gene expression data. The weight coefficients are computed from an optimization problem which minimizes total weighted difference between nodes in a network structure; this can be expressed in terms of graph Laplacian. After obtaining the coefficient vector for the network-based markers, we can then compute the corresponding network predictor. We demonstrate the effectiveness of the proposed method by conducting experiments using published breast cancer biomarkers with three patient cohorts. Network-based markers are firstly grouped based on GO terms related to cancer hallmarks. We compare the predictive performance of each network marker group across gene expression data sets. We also evaluate the network predictor against the average method for feature aggregations. The reported results show that predictive performance of network markers is generally not consistent across patient cohorts.

In **Chapter 7**, we conclude the thesis by solving the inference problem, since rapidly accumulating genomic data have posed a challenge to integrate multiple data sources and to analyze the integrated networks globally. In this chapter, we present a method to reverse engineer integrative gene networks. The main advantage of our method is the integration of different quantitative and qualitative datasets in order to reconstruct a multiplex network, without imposing data constraints, such as each genomic datum needs to have the same number of entities. The computation boils down to solving small quadratic programs based on local neighborhood of nodes. Another advantage of our method is that from the integrated networks, predictions can be made by propagating beliefs from seed nodes representing known knowledge via weighted edges. Thus, we combined data integration and network-based prediction into a single framework. We applied the method to DREAM5 dataset, and compared the results with the community networks from the challenge. Further, we demonstrate our method through case studies using breast cancer data, including the integration of metastasis gene expression data with interactome data and biological pathway data. Network-based predictions are compared between interactome-integrated and pathway-integrated networks. Overall, our method has the potential to be applied in many settings of network system biology.

Chapter 2

A Closer Look at "Social" Boundary Genes Reveals Knowledge to Gene Expression Profiles

[1] *As social network analysis is gaining popularity in modeling real world problems, the task of applying the social network model concepts and notions to biological data is one of the most attractive research problems to be addressed. Here we focus on a particular set of genes that reside on the community boundaries in gene co-expression networks. Stemmed from community mining problem in social networks, peripheries of communities (i.e., boundaries) can be used to aid certain biological analysis. The proposed method consists of three parts: 1) Finding communities of gene co-expression networks through clustering. 2) Analyzing stability of community structures by Monte Carlo method. 3) Designing of dynamic adoption of boundaries using geometric convexity. We validate our findings using breast cancer gene expression data from various studies. Our approach contributes to the new branch of applying social network mechanisms in biological data analysis, leading to new data mining strategies that associate social behaviors to biological network analysis.*

2.1 Introduction

Recently, researchers have started to realize the effectiveness of social network analysis mechanisms in understanding group behaviors and network dynamics. For example, Centola [19] has demonstrated that social behaviors (how individuals adopt health recommendations) spread in a counter-intuitive fashion, i.e., information exchanged through social interactions

[1] The content of this chapter is based on the following article:
Gao, S., Zeng, J., ElSheikh, A., Naji, G., Alhajj, R., Rokne, J., & Demetrick, D. (2011). A Closer Look at "Social" Boundary Genes Reveals Knowledge to Gene Expression Profiles. *Current Protein and Peptide Science, 12*(7), 602-613.

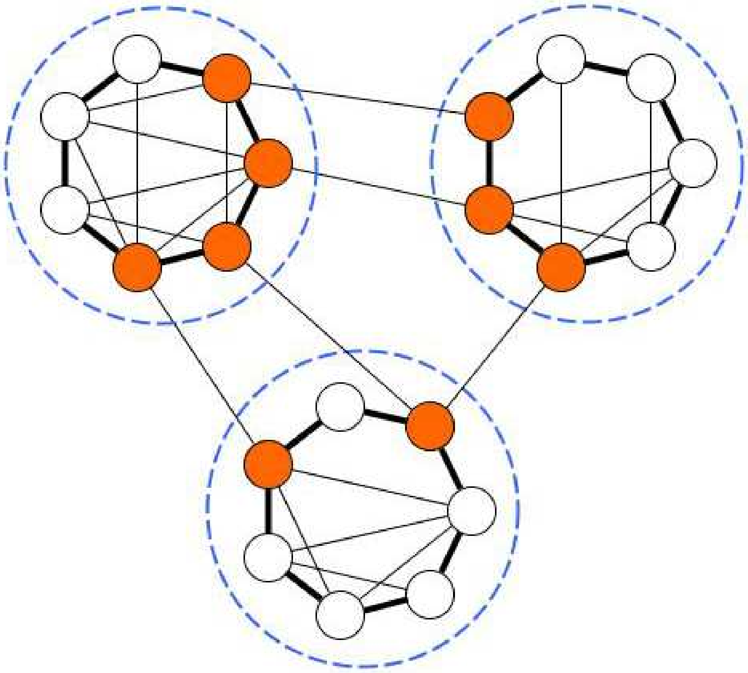


Figure 2.1: Boundary nodes in communities. Dashed circles represent virtual communities inferred from data. Darkened lines represent connections or links between boundary nodes. Orange colored represent boundary nodes interacting with other communities in the network

spread faster in clustered-lattice networks. We argue that this type of social network analysis should motivate biologists to model the behavior and evolution of human genome in a wide variety of approaches and perspectives, with the hope to account for less well-characterized phenomena such as genetic variations [84]. The reason why following the social notions is reasonable is certainly underlined by formal theoretical models [58]. Accordingly, our work described in this chapter adopts social network mechanisms to model gene expression data and to understand the behavior of genes. The motivation for this work is that although clustered gene co-expression modules are well understood [92], the behavior of certain subsets from those modules is less understood. In the literature, there is substantial research on the construction and interpretation of co-expression networks, but only little work has been done on the elucidation of the boundary subset. More importantly, the behavioral aspects of the boundary genes in co-expression context are not thoroughly discussed.

In this chapter, we adapt the social network model to study genes; and hence, we investigate social inspiration by concentrating on boundary genes in microarray expression data since they resemble boundary nodes in social communities. The boundary of a social community is a set of individuals that reside on its borders and is separate from neighbor communities. Intuitively, compared to interior nodes boundary nodes are more volatile in that individuals residing on them are more likely to leave the local community and join a neighbor community instead. Further, the boundary nodes are less interactive compared with other nodes in the community. Therefore, mining community boundaries can lead to profound implications in social network analysis.

In this work, we hypothesized that the same behavior applies to gene expression profiles in which co-expression modules are analogous to communities and boundary genes are analogous to unstable individuals. We introduce clustering based methods to find boundary gene set in dynamic settings using statistical methods. To visualize the boundary, we benefit from the fact that boundaries can be represented as convex hulls in a geometric setting in the clustering process. The idea is illustrated in Figure 2.1, the darkened lines are geometric boundaries of a convex hull and we are interested in the behavior of the boundary nodes, i.e., genes and their strength of connections in gene expression data.

2.2 Related Work

To describe the literature pertaining to this study, we categorize related works according to the mining workflow (see the experimental study section) into three categories as follows.

2.2.1 Clustering of Gene Expression Data

Since clustering plays a fundamental role in data mining by partitioning the observations into subsets, its importance in grouping gene expression profiles has been intensively studied [69, 143]. The unsupervised nature of the clustering process allows subjective interpretations of the result, which is a double-edge sword in data analysis [34], because the lack of domain specific calibrations may lead to undesirable results. To account for this, different computational methods and metrics are employed in clustering; see reference [69] for a comprehensive review. In this chapter, we use clustering as a pre-processing step in building co-expression networks from gene expression data sets, as the clusters represent the functional modules of similarly expressed genes [18].

2.2.2 Social and Cellular Networks

To better understand the interactions between genes, networks can be built by inferring from the gene expression data. This reverse engineering approach prompts a global view of structural knowledge [103, 102]. The notion of social networks (originally stemmed from sociology and anthropology) permits behavioral studies between nodes, e.g., genes in cellular networks, such as learning mechanisms, as well as community mining that is analogous to the concept of modularity in cellular networks [48, 136]. Within the context of gene expression profiles, co-expression modules are often analyzed in conjunction with networks [92, 20]. The connection between gene clusters and co-expression modules is apparent, and the latter is used in the context of gene expression analysis.

2.2.3 Boundary Mining

To mine the boundary of clusters, several methods have been proposed in several directions: the work described in [15] discusses the stability of clusters in association with cluster boundaries; on the other hand, the works described in [59, 150, 123] focus on the estimation of the boundaries from the clustering results; and the work described in [161] introduces a method to visualize high-dimensional clusters, therefore implicitly depicted cluster boundaries. In this chapter, we identify community boundaries and investigate their social behaviors through randomized clustering process and Monte Carlo methods. Boundaries are represented as convex hulls in a geometric perspective having benefits in terms of visualization and accuracy. Our approach contributes to a new branch of social community mining by effectively identifying and representing social boundaries under dynamic settings. Although we have focused on gene expression data in this chapter, the discussed approach in general enough and can be applied to study other types of molecules within the body as well as to tackle various real world applications. For instance, in terrorism networks our approach can be used to find crucial nodes in terrorist organizations.

2.3 Background

2.3.1 Social Network

Social network mining and analysis is a relatively new field that combines sociology perspectives and data analysis techniques [87]. The emerging interest attributes to the fact that many real world problems can be modeled as social networks. For instance, different genes and their functions can be modeled as two-mode social networks in drug development. After a social network is constructed, one of the important tasks is to identify social communities [119]. A community can be described as a group of individuals who share similar interests. The sociological implication is that individuals within the same community tend to preferably interact with each other more frequently.

Mathematically, a social network can be defined as a graph $G = (V, E)$, where $V = \{v_1, v_2, ..., v_n\}$ represents individuals (interchangeably called actors) and $E = \{e_1, e_2, ..., e_m\}$ represents links or interactions between individuals in V. Communities can often be identified using sub-graph extraction and similarity based methods [179]. However, static graph-based methods do not account for the dynamic nature of social networks in real world problems, i.e., the question of how does communities structure change and evolve over time is a more complex problem. Therefore, we concentrate on the behavior and design of the dynamics of social network analysis, which would lead to more practical models for various real world problems [156], and consequently raises the need for more powerful computational methods and analytical means.

2.3.2 Convex Hull Construction

In the Euclidean space, let C be a set in a vector space; C is said to be *convex* if for any $x, y \in C$ and any $t \in [0, 1]$, the point $(1 - t)x + ty$ is also in C. The convex hull for a set of points X in a real vector space is the minimal convex set containing X. Formally, the convex hull of a set of points X in n dimensions is the intersection of all convex sets containing X. For N points $p_1, p_2, ..., p_N$ the convex hull $H(X)$ of set X is defined as:

$$H(X) = \{\sum_{j=1}^{N} \lambda_j p_j | \lambda_j \geq 0, \sum_{j=1}^{N} \lambda_j = 1\}$$

In computational geometry, convex hull often refers to the *boundary of the minimal convex set* containing a given non-empty finite set of points in the plane. Since convex hull computation is a fundamental problem in computation geometry, numerous efficient algorithms have been proposed in the literature. Theoretically, convex hull construction can be implemented in multi-dimensional spaces; however, in real practice, 2D and 3D convex hulls are the most commonly used and discussed in the literature. Recent dominant algorithms for finding convex hull in 2D and 3D have the complexity $O(nlogn)$ [133]. The reason we are interested in convex hulls is the fact that they geometrically represent boundaries of a set which can be viewed as community boundaries. In other words, if we define $p_1, p_2, ..., p_N$ as nodes or individuals in a set X, then $H(X)$ describes the boundary of X or confines X. A question following the above argument could be articulated as follows: how to define such X in the first place in order to compute $H(X)$? We handle this partitioning process by using a clustering method [118].

2.4 Methods

Our proposed method for mining community boundaries can be divided into two phases. In the first phase, we cluster the gene expression data where each cluster is viewed as a community. In the second phase, we construct convex hulls that represent community boundaries. The procedure is described in Algorithm 1. Algorithm 1 estimates the optimal number of clusters by first constructing hierarchical clustering tree from different linkage functions. To determine the best linkage function, we compute correlation coefficients among a pool of options, such as average, complete, single linkages, etc. and the optimal linkage is given by the maximal correlation coefficient of the hierarchical tree from the pool. The best number of clusters is given by the highest average silhouette value with the optimal linage function.

Algorithm 1 Find Community Boundaries

```
Input: Data set $D$, rows are genes and columns are samples.
Output: Convex hulls representing community boundaries.
  1. Compute pair-wise distance of genes in $D$, and compute agglomerative clustering tree
     using different linkage functions.
  2. Pick the linkage function with the highest correlation coefficient.
  3. Find the optimal number of clusters using silhouette value; cluster $D$.
  4. Repeat: If cluster $i$ has less than 3 objects, or objects are co-linear, mark those objects
     as obsolete, and terminate the convex hull construction for cluster $i$.
       $i \leftarrow i + 1$
       Until: all clusters are examined
  5. Construct convex hulls for each non-obsolete cluster.
  6. Record total boundary distances for each community in a vector.
```

The silhouette value measures average silhouette width for each cluster as well as overall silhouette width for the entire dataset. The silhouette value is computed as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where $a(i)$ is average dissimilarity of the i^{th} object to all other objects in the same cluster; $b(i)$ is the minimum average dissimilarity of the i^{th} object to all objects in other clusters, and $-1 \leq s(i) \leq 1$. Having $s(i)$ close to 1 implies the sample is well clustered; on the other hand, having $s(i)$ close to -1 implies that the sample is poorly clustered.

Pair-wise distance is an important measure in the clustering process, and the choice depends on the data domain. For analyzing gene expression profiles, three measures are frequently used: Euclidian distance, Pearson correlation and mutual information. Each has advantages and disadvantages compared to others in the literature, e.g., [18, 123]. In this chapter, we use the Pearson correlation as a measure of similarity to avoid the computational cost from mutual information. Algorithm 1 essentially uses clustering to identify community structures and further construct convex hulls to describe community boundaries. This initial construction process will be used in adapting changes in dynamic social networks as discussed in the following subsections.

2.4.1 Monte Carlo Based Method in Stabilizing the Community Structure

An important question raised due to the clustering method is: how to ensure that the community structures extracted are stable? In practice, this concern is crucial because the communities are sometimes formed based on a subset of representative samples that are subject to bias and noise with different clustering methods like k-means and SOM. Therefore, the ideal boundaries should be "stabilized" so that objects on the boundaries are not biased to the clustering method, i.e., they should be robust subject to bias and noise. For example, a community of patients having the same set of expressed genes might be subject to sampling error and/or measurement noise, and this consequently destabilizes the community structures.

Algorithm 2 Monte Carlo Simulation to Stabilize Community Structures

```
Input: Clusters formed from Algorithm 1; simulation error bound $\varepsilon$; selection threshold $\tau$.
Output: Convex hulls representing stabilized community structures.
  1. Generate Gaussian noise variables $\sim N(0, \sigma^2)$; and compute $T = \frac{1}{N} \sum_{0}^{N-1} x[n]$ for one realization of noise, where $N$ is the number of samples.
  2. Repeat step 1 to obtain a set of $M$ realizations, denoted as $T = \{T_1, T_2, ..., T_M\}$.
  3. Count the number of realizations that exceed a threshold, called $\gamma$.
  4. The number of trials needed to estimate the probability of number of $T_i$'s that exceed $\gamma$ is at least $\frac{[Q^{-1}(\alpha/2)]^2(1-P)}{\varepsilon^2 P}$, where $Q(.)$ is the right-tail probability and $\varepsilon$ is the absolute error bound defined by the user.
  5. Construct convex hulls $M$ times (to replace step 4 of Algorithm 1).
  6. For $M$ set of convex hulls in cluster $i$, find the most frequent vertices that define the boundaries, demarcated by a threshold $\tau$.
  7. Output selected vertices that define the community structures for cluster $i$.
```

To resolve this issue, we resort to the computerized Monte Carlo simulation. The Monte Carlo method allows us to estimate the probability of random events described by statistical randomization [3]. In this chapter, we use the Monte Carlo simulation to determine the number of trials needed to stabilize the convex hulls if randomization is involved in the clustering process. Practically, if the number of simulations satisfies pre-defined conditions in Monte Carlo methods, the "common" vertices of convex hulls from those trials can be used to construct stabilized communities. In other words, statistical measures and their properties can eliminate the effect of noisy components from imperfect measurements. In view of the aforementioned reasoning, we present Algorithm 2 that uses the Monte Carlo simulation to stabilize the community structures in noisy environments.

The proof of step 4 in Algorithm 2 is given by [75]. The noise is assumed to be Gaussian in this Monte Carlo simulation, which is common in real circumstances. The choice of ε and τ is configured by users; one such possible configuration is $\varepsilon = 0.01(1\%)$, and for $100(1-\alpha)\% = 95\%(\alpha = 0.05)$ of time that the community structures are stable.

```
Algorithm 3 Determine Dynamic Boundaries
Input: Convex hulls representing community boundaries at time $t_1$, a set of new individuals $V = p_1, p_2, ..., p_s$.
Output: Adjusted convex hulls representing changed community boundaries at time $t_2$ including $V$.
  1. For each $p_i$, $1 \leq i \leq s$, in set $V$, if it falls within the boundary of existing convex hulls, the community structure does not change; otherwise go to step 2.
  2. For each object $p_i$ in set $V$, find the line segment $l$ of a convex hull from $H(X_1), H(X_2), ..., H(X_{k'})$ in Algorithm 1 that minimizes $\|p - l\|$, then $l$ is the line segment that should be replaced by two new line segments connecting the new individual.
  3. Delete the edge $l$ whose left endpoint is $l_L$ and right endpoint is $l_R$, and add two edges joining $l_L$, $p_i$ and $l_R$, $p_i$, respectively. Denote the new convex hull $H(X_{l*})$.
  4. Replace $H(X_l)$ with $H(X_{l*})$, forming new community structures.
  5. Update total boundary distances $d_{i*}$ associated with $H(X_{l*})$ in vector $D = [d_1, d_2, ..., d_{i*}, ..., d_{k'}]$.
```

As Algorithm 3 is the simulation to stabilize the communities' structures, the Gaussian noise can be seen as a predisposition that affects the robustness of communities. In step 4, if $x[n]$ is the Gaussian noise, then variable T follows Gaussian distribution with standard deviation σ^2/N and mean 0; T is a vector of M realizations, the intended meaning of $P(T > \gamma)$ is the probability of the number of T_i's that exceed γ.

2.4.2 Ad Hoc Design of Dynamic Adaption of Boundaries

Given the convex hulls, constructed from Algorithm 1, describing community boundaries, the distance vector D_b can be used as cost vectors or weights of the virtual links between nodes on the boundaries. D_b can be calculated as the sum of distances between convex hulls and the element in D_b denotes the current weight that the cluster convex hull represents. The use of D_b is to gauge the change of distance after introducing new nodes into the communities as demonstrated in Figure 2.2.

We can use this information for changing the social networks at the later stages. We formally define the problem as follows: Given a sequential time ordering $t_0, t_1, ..., t_n$ convex hulls at t_0 and distance vector D_b; for each time stamp t_i, $1 \leq i \leq n$, new individuals join the social network sequentially in vector form $N_{ti} = [n_{ti1}, n_{ti2}, ..., x_{tik}]$. Determine community memberships for $n_{ti1}, n_{ti2}, ..., x_{tik}$.

The basic idea for determining community membership of a new node or individual is to use the distance metric to gauge the nearest boundaries between the new node and the constructed convex hull. This is illustrated in Figure 2.2. At time t_1, convex hulls are constructed from Algorithm 1, and from Figure 2.2 A we see that two communities are formed. At time t_2, a new individual represented as a point in Euclidean space is introduced, and from Figure 2.2 B it is clear that the new point is "closer" to the triangle community on the left. Therefore, the new individual is included in the triangle community. The boundary of the changed community with the new member is adjusted by adding two edges (in light green) and deleting one, i.e., new convex hull $H(X_i \cup [\{\textit{new point}\}])$ is updated from $H(X_i)$.

Note that if the new point falls within boundaries of the convex hull, the communitys boundaries remain unchanged, i.e., $H(X_i \cup [\{\textit{new point}\}])$. The procedure is depicted in Algorithm 3. It is worth noting that $\|p - l\|$ in step 2 is the perpendicular distance between

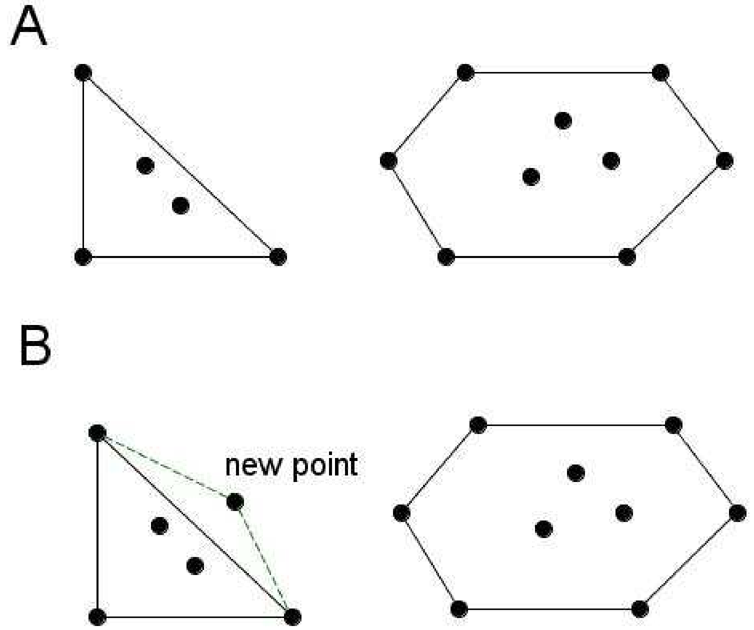


Figure 2.2: Dynamic hull construction representing changing communities: A) convex hulls from Algorithm 1 at t_1. B) adjusted convex hulls with a new point at t_2

a point p and the boundary line segment l. Another advantage of our proposed approach in determining dynamic community boundaries is that there is room for outlier detection, i.e., what if some individual does not belong to any existing community? In order to answer this question, we need to define a positive value as the threshold, ε, and we say that p_i is non-joinable to the existing communities if $|d_{i*} - d_i| \leq \varepsilon$; such p_i is problematic because it may itself form a new community or it is abnormally unusual (outlier). In either case, further processing needs to be done; however, in most circumstances, community structures are likely to be stable given enough time.

The ad hoc design of the dynamic adaption of boundaries is not context dependent since as illustrated in Figure 2.2, the geometric meaning of the graph is being simplified. In the context of gene expression data set, each point maps to a gene expressed in the microarray consisting of different samples (measurements of expression levels). The biological meaning of the addition of new nodes corresponds to the effect of newly considered genes to the

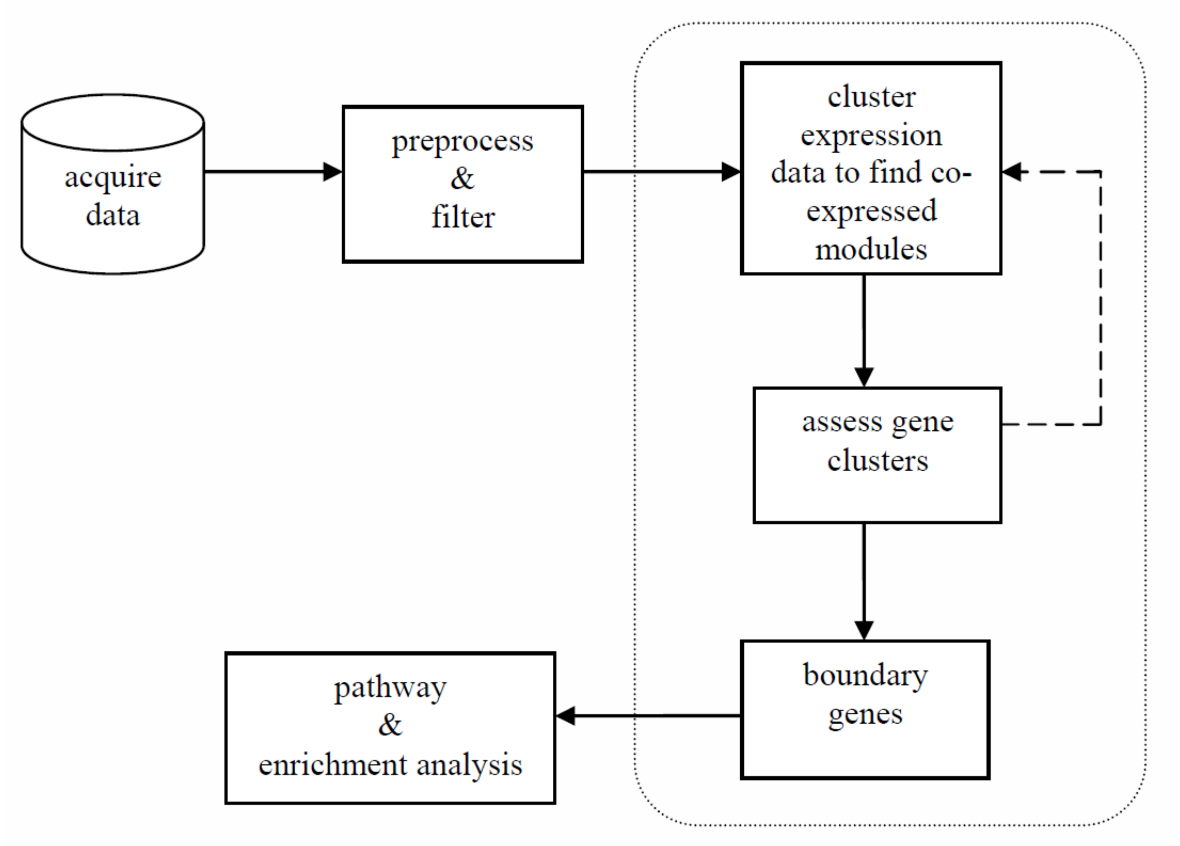


Figure 2.3: Work flow in finding boundary genes

functional modules represented by communities. As we lack the data to demonstrate this process at the moment, it is left as future work and the design is ad hoc.

2.5 Experiments

2.5.1 Data Sets and Tools

To demonstrate the applicability and effectiveness of the proposed approach, we have used the five gene expression datasets investigated in [159] to study the impact of social boundary genes in human breast cancer. These gene expression data sets were obtained from the Gene Expression Omnibus (GEO) (`http://www.ncbi.nlm.nih.gov/geo/`), and they are briefly described in Table 2.1.

To carry out the experiments, we preprocessed and visualized the gene expression data

Table 2.1: Breast cancer gene expression data sets from GEO

GSE Accession No.	**Description**	**Number of Samples**
GSE5116	17β-estradiol (E2) transformation	12
GSE5764	Ductal and lobular carcinomas	30
GSE6548	Activity of ESR1 and BMI1	8
GSE6885	Breast epithelial cell types (HMLER and BPLER)	21
GSE8597	Estrogen receptors (ERs) targets in MCF7 cells	16

sets with Genesis [154]; and we implemented the proposed approach using MATLAB 7. The framework is described in Figure 2.3. In this chapter, we focus on the steps in the dashed box concerning co-expression boundaries, and we discuss options for the enrichment analysis.

In our experiments, we focused on genes with most significant degree of variation measure computed by the coefficient of variation (CV) statistic as:

$$C_v = \frac{\sigma}{\mu}$$

where σ is the standard deviation and μ is the sample mean. We ranked genes based on the CV measure and used the top 3051 genes (CV-gene hereafter) for our study, as the Human Genome U133 Plus 2.0 Array contains a large number of genes and the majority of which are not ascribed to breast cancer phenotype. It is therefore reasonable to assume that genes expressed with high degree of variation are more likely to be involved in tumorigenesis [33]. The proposed method is not gene set or expression profile specific since we are interested in analyzing the impact and behavior of boundaries with unsupervised learning.

To illustrate the effect of using different similarity measures to CV-gene, graphical data terrains are depicted in Figure 2.4 & Figure 2.5. The Euclidean distance pair-wise links are sparse, and the peaks representing high expression levels are all pervading compared with the data terrain with the Pearson Correlation.

This sheds light on our investigation of the clustering process in finding co-expressions between genes since we are interested in condensed co-expression modules where a compact

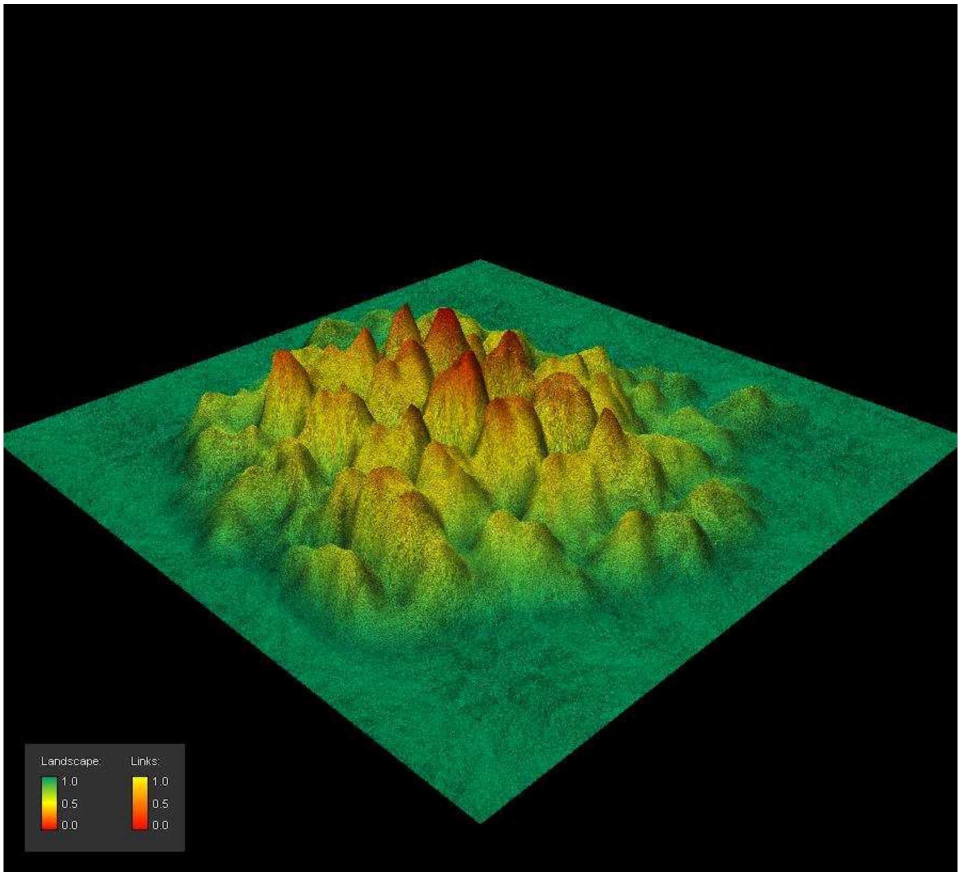


Figure 2.4: Data Terrain Euclidean Distance

set of boundary genes interact with majority of other genes to higher degree, hence possibly exert more functionality to biological processes for which the impact of gene-gene interactions underlying diseases is not overt.

2.5.2 Co-Expression Modules and Boundary Genes

In order to find co-expression modules based on the breast cancer gene expression profile, we cluster the data set as described in Algorithm 1. We choose the best number of clusters or modules to form. We use the hierarchical clustering method to build the agglomerative tree. This leads to the choice of the linkage function and, consequently, the optimal clustering that gives the maximal average silhouette measure.

Figure 2.6 shows the mean and coefficient of variation (CV) measures of silhouette value

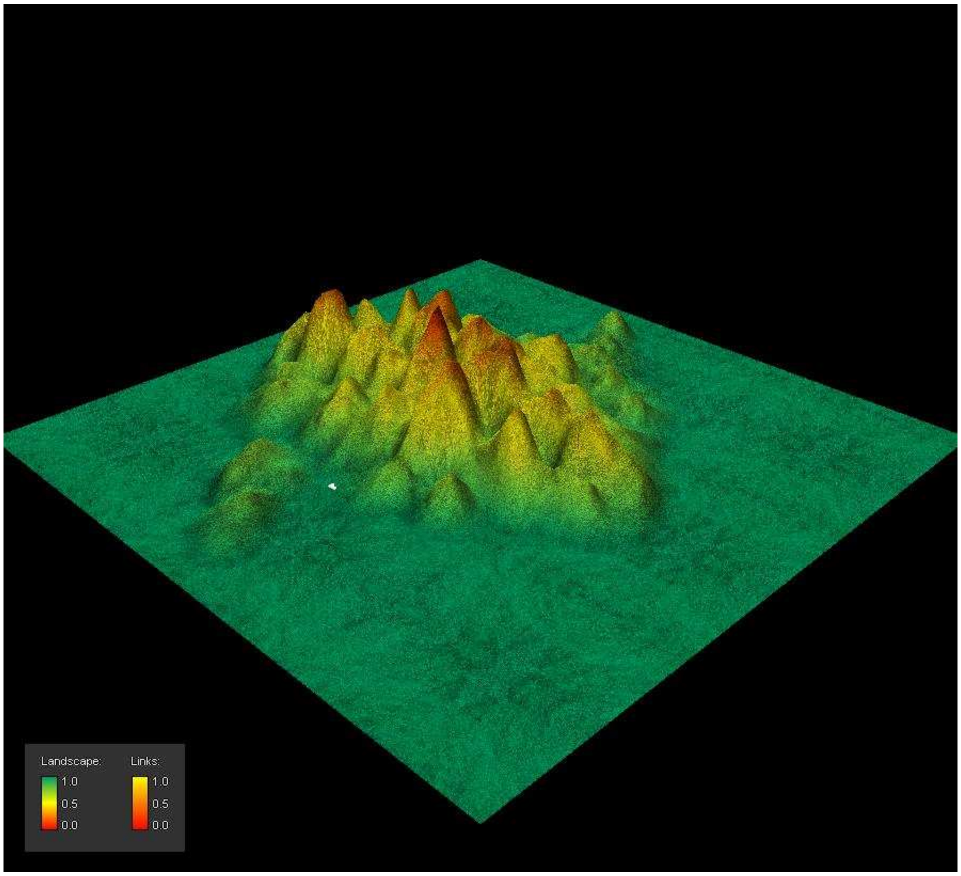


Figure 2.5: Data Terrain Pearson Correlation

with different number of clusters representing breast gene co-expression modules. The average linkage function is used in Figure 2.4 with cophenetic correlation coefficient 0.8196; other choices include complete, single, ward and weighted linkages by specifying the coefficient measure as 0.7807, 0.2733, 0.5926 and 0.7806, respectively. As shown in Figure 2.6, the mean and CV measures tend to fluctuate less with the number of clusters greater than ~ 45, rendering a rather stable pattern of near-horizontal lines. This observation suggests that the number of co-expression modules that accounts for the cancerous state is steadily demarcated. Further, the mean and coefficient of variation measures of silhouette value coincide and show that the optimal number of clusters with CV-gene is 26 with mean and CV as 0.4089 and 1.1967, respectively. The most significant normalized gene clusters are shown in Figure A.1.

Table 2.2: Boundary genes for the six most significant co-expression modules; repeated genes are bolded and genes with mutations are underlined

Significant Modules	**Number of Genes**	**Boundary Genes**	**Significance**
Module 5	166	*IL12RB1*,*LHFPL5*,*MIPOL1*,***AKR1CL2***,***EARS2***,*CD80*,*C6orf114*,*DTWD1*,*ZNF655*,***CNOT6L***,***CPA6***,*MPP4*,*COBL*,*CXorf52*,***DKFZp761H2121***,***SVEP1***,***GPR128***,*LUZP1*,***C1S***	$p < 0.005$
Module 18	180	*GLYATL2*,*NHEDC1*,*FBXL14*,*ITIH5*,***SCML4***,*POLR2B*,*ZNF586*,*GIMAP1*,*ZNF663*,*SEPSECS*,*ST7OT1*,***GPR128***,*ABCC13*,*C20orf12*,*TMLHE*,***EARS2***	$p < 0.005$
Module 10	186	***SCML4***,***CNOT6L***,*C12orf66*,*IGFBP5*,***AKR1CL2***,***SETMAR***,*C3orf15*,*FAM154B*,*ZNF441*,*TRNT1*,*ZNF641*,***MRAP***,***C1orf210***,***SVEP1***	$p < 0.005$
Module 4	201	*LOC253039*,***SCML4***,*TIGD4*,*C1S*,*AFG3L1*,*MUC19*,*GAPT*,*C9orf96*,***MRAP***,*LILRA5*,*WBP2NL*,***CPA6***,*C3orf23*,***TERF2***	$p < 0.005$
Module 8	212	*FLJ30672*,*C1S*,*LOC100134445*,*ZNF585A*,*TTBK2*,*PDE4D*,*CASC2*,*CXorf52*,*C1orf210*,*ZNF385B*,*FLJ42709*,*ATP6V1C2*,*ZAP70*,***DKFZp761H2121***,*RFX6*,***VSTM2A***,*ADAMTS9*	$p < 0.005$
Module 13	386	*VSTM2A*,*PRPSAP1*,*WDR78*,*FLJ37035*,***SETMAR***,*PARD3B*,*ITPKB*,*LOC283861*,***TERF2***,*CNR1*,*ZNF837*,*LUC7L*,*STAP1*, *ZNF831*	$p < 0.005$

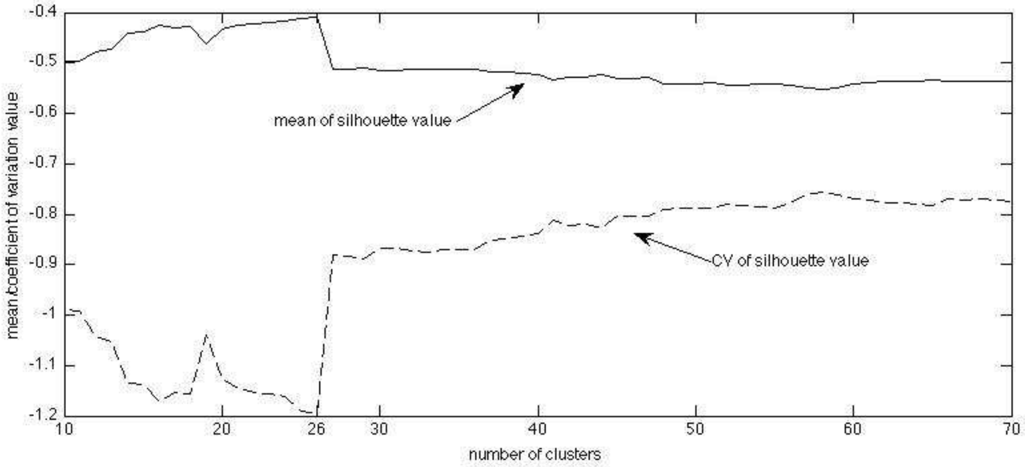


Figure 2.6: Mean and Coefficient of Variation(CV) values against number of clusters

The expression profiles from Figure A.1. show similar behavioral patterns of gene modules, depicted by pink contours containing high-expressed (peaks) and low-expressed genes (horizontal lines). Since contours of significant gene co-expression modules are similar, the quality assessment of co-expression modules can be reduced by a set of representative genes sampled from each cluster. As aforementioned, the boundary genes from social perspective that confine the gene co-expressions modules can be used to achieve this interpretation. The boundaries can be geometrically portrayed by convex hulls in Algorithm 2. The boundary genes are listed in Table 2.2.

The significance of boundary gene set is assessed with hypergeometric distributions, in other words, the probability of being a boundary set of genes in each co-expression module. Interestingly, Table 2.2 shows that boundary genes from different co-expression modules can overlap (bolded). The most repeated genes are *C1S* (modules 5, 4, 8), and *SCML4* (modules 18, 10, 4). The repetition of boundary genes means that co-expression modules share common genes residing on their borders. The most repeated genes such as *C1S* and *SCML4*, therefore, may serve as indicators of co-expression modules in association with cancer phenotype under investigation and further signaling pathway [73] or a gene ontology enrichment analysis [6] can be conducted. For example, the repeated *C1S* gene participates

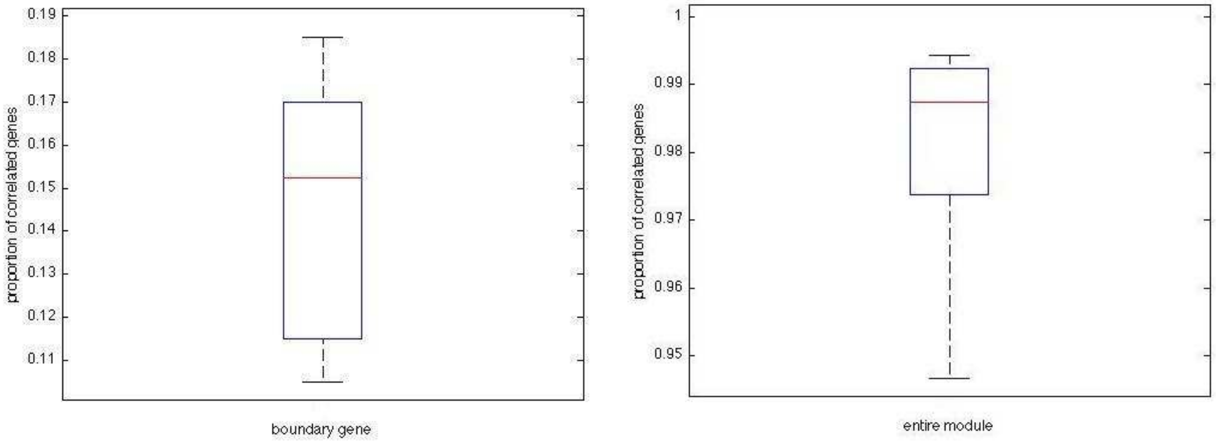


Figure 2.7: Correlation comparison between entire gene module and boundary genes, the y-axis is the proportion of correlated genes

in the complement and coagulation cascades pathway in KEGG database that is a defense system against pathogens (see Figure A.2 in Appendix 1) in biological processes relate to membrane organization in GO, while there is no pathway found for less-characterized *SCML4* gene whose protein product Sex comb on midleg-like protein 4 functions by forming protein complexes to regulate transcription activity and belongs to the group of histone modification proteins.

As a further exploration, since cancer arises from somatic mutations [175, 153], we map the boundary genes to somatic mutations in COSMIC database [7, 182], and observe that most boundary genes are not catalogued with genetic mutations. However, genes with mutations tend not to be shared on co-expression boundaries and they tend to aggregate in modules (module 8 in Table 2.2). This knowledge can be further used to assess co-expression modules by extracting the boundary as a gene set and evaluate the probability based on assumed distribution as described in [182].

2.5.3 Connectivity of Boundary Genes and Pathway Discussions

To examine the co-expression levels of boundary genes, we compare the boundary genes with the corresponding co-expression modules. This is done because the relationship between a certain gene set and its genetic context (which is the co-expression network) is an important aspect in assessment. We compute the correlation coefficient statistic:

$$\frac{\mathrm{cov}(g_i, g_j)}{\sqrt{\mathrm{cov}(g_i)\mathrm{cov}(g_j)}}$$

where g_i and g_j are expression levels of gene i and gene j, respectively, and $cov(.)$ is the covariance between g_i and g_j.

The boundary genes from the breast cancer data set show substantially lower correlation proportions in terms of expression levels measured (∼17% for boundary genes against ∼95% for the entire co-expression module), which is worth further investigation from biological perspective. The computational method based on co-expressions hinges on the fact that observations (i.e., gene expression levels) residing on the boundary are least relevant to the centroids representing most influential genes. This finding is in agreement with the social intuition that boundary individuals of communities are weekly linked. However, the underlying biological reason for the boundary genes being weekly connected in significant co-expression modules as a general trend in reference to perturbation data analysis such as CV-gene is unclear.

There is a variety of approaches to assess the clusters, for instance, to map the genes to pathways and compute the probability as a measure of significance, which is used in Table 2.2 above in the similarly vein with hypergeometric distributions [182]. The mapping is done by querying a pathway database such as KEGG (`http://www.genome.jp/kegg/`) or PANTHER (`http://www.pantherdb.org/`). Alternatively, Gene Ontology (`http://www.geneontology.org/`) can be used to assess the components of boundary genes in biological processes in pursuit of further common ontological intersections, if any [47]. Recent

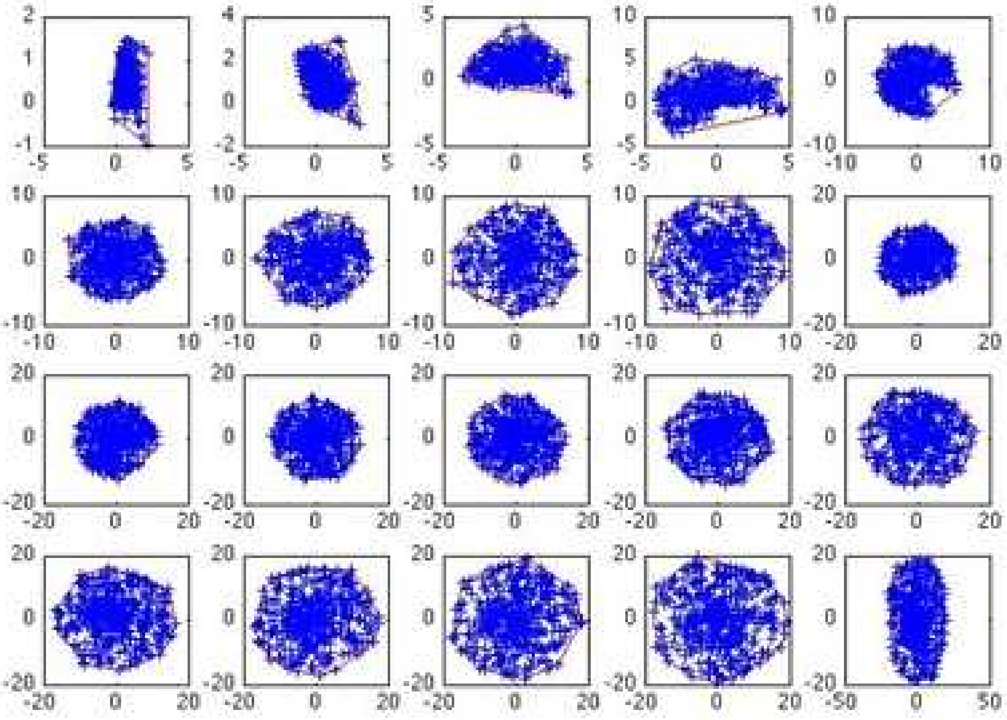


Figure 2.8: Communities snapshots by first 20 Monte Carlo simulations; horizontal and vertical axes are polar coordinates

approaches in computation system biology provide further flexibility of this task. In the work described in [104] topological pathways are converted to compact graphs for mining and analysis. As one of the future works, we intend to elucidate boundary gene set with topological information using system biology approaches.

2.5.4 Stability of Clusters

In this section, we experiment with the yeast time series data set to demonstrate the stability effect of clusters. The data set can be found at the following repository: `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28`. The data set contains 6400 genes expressed at 7 different time stamps. In this experiment, we treat genes as nodes in the social network. The focus of this experiment is to show community structure stabilization with randomized clustering methods using Monte Carlo methods. As a preprocessing step, we deleted all genes with unknown expression levels including empty values that might be caused by measurement errors. The total number of genes is 6276 after preprocessing. We used the correlation pair-wise measure to cluster the data set. The correlation is defined as:

$$d_{corr} = 1 - \frac{(x_s - 1/n\sum_j x_{sj})(x_t - 1/n\sum_j x_{tj})'}{\sqrt{(x_s - 1/n\sum_j x_{sj})(x_s - 1/n\sum_j x_{sj})'}\sqrt{(x_t - 1/n\sum_j x_{tj})(x_t - 1/n\sum_j x_{tj})'}}$$

where x_s and x_t are row vectors representing the expression of genes at different time stamps. We generate community structures by Monte Carlo trials as discussed, and the first 20 sample communities are shown in Figure 2.8. By using a sufficiently large number of trials, the gene community structures stabilize or converge. We do not eliminate duplicates from different communities because the occurrences of genes are given by a large number of Monte Carlo trials in which randomized likelihood is computed (Algorithm 2). Therefore, the order of genes that define communities implies ranking of likelihoods. For example, gene i appears before gene j, this means that gene i has higher likelihood of being a boundary gene than gene j does. The statistics can be kept along with the computations so that unique genes can be found by a simple ranking.

2.6 Chapter Discussion and Conclusions

Our motivation of this study is to provide a new perspective in the analysis of gene expression data by looking at the community boundaries inspired by theoretical network science [48, 12], and the results were analyzed via cancer data sets and mutation data. Other biological analysis problems can be further aided using this strategy and thinking.

The workflow described in Figure 2.3 does not tie to specific type of biological networks such as gene regulatory networks (GRN) or protein-protein interaction (PPI) networks. One of the reasons we instantiated the study with gene co-expression network is that pair-wise comparisons between expression level of genes can be easily computed quantitatively, as opposed to the computation using GRN and PPI networks. For example, in PPI networks, proteins are linked with physical reactions observed, and therefore the global network structure such as scale-free properties or functional modularity are of more interests to biologists [68]. Having said so, we realized that GRN and PPI networks can be inferred computationally from gene expression data sets [147, 178]. However, as our study focuses on the strategic analysis summed to boundary genes, the interposition of network inference is out of the scope of the chapter (in fact, the topic is intensively discussed in bioinformatics). As aforementioned, we have included in the chapter the proposed method and real experiments to demonstrate some use of the workflow by linking gene expression data analysis to cancer mutations, and this direction is studied by recent publications using gene co-expression networks [159]. Our new probe is the detailed look at boundary genes that are not discussed in this regard, being one of many possible practicabilities. We regard this as an important future direction to pursue. The second reason we only focused on the gene co-expression network in this study is the following: with reference to the workflow presented in Figure 2.3, in GRN of PPI networks, perhaps the key point in terms of mining strategy pertains to graph clustering or partitioning in order to find communities, and this is another popular topic heavily studied [41]. Finding boundary vertices in graphs is a much harder task in terms of computation, whereas in gene co-expression networks quantitative measures are at front end [56]. Topology of graphs affects boundaries to a great extent, therefore to find boundary vertices, constraints need to be imposed or proximity measures can be utilized such as centrality ranking.

As demonstrated by the study and experiments, there is much room to further explore this boundary-based approach. For example, the separation of communities can also be measured using the boundaries and it is possible to combine two closely-related communities in real settings if time series data is used. As discussed, probabilistic framework can also be developed to elucidate boundaries with pathway information. Overall, the boundaries of community structures can lead to profound implications in networks with different application domains.

Chapter 3

Multi-scale modularity and motif distributional effect in metabolic networks

[1]

We study the relationship between multi-scale community structures and network motifs using metabolic networks. Metabolism is a set of fundamental processes that play important roles in a plethora of biological and medical contexts. It is understood that the topological information of reconstructed metabolic networks, such as modular organization, has crucial implications on biological functions. Recent interpretations of modularity in network settings provide a view of multiple network partitions induced by different resolution parameters. Here we ask the question: How do multiple network partitions affect the organization of metabolic networks? Since network motifs are often interpreted as the superfamilies of evolved units, we further investigate their impact under multiple network partitions and investigate how does the distribution of network motifs influences the organization of metabolic networks. We study Homo sapiens, Saccharomyces cerevisiae and Escherichia coli metabolic networks, and analyze the relationship between different community structures and motif distribution patterns. Further, we quantify the degree to which motifs participate in the modular organization of metabolic networks.

[1] The content of this chapter is based on the following articles:
Gao, S., Chen, A., Rahmani, A., Jarada, T., Alhajj, R., Demetrick, D., & Zeng, J. (2013). MCF: A tool to find multi-scale community profiles in biological networks. *In Submission.*
Gao, S., Addam, O., Chen, A., Rahmani, A., Zeng, J., Tan, M., Alhajj, R., Rokne, J., & Demetrick, D. (2013). Multi-scale Modularity and Motif Distributional Effect in Metabolic Networks. *In Proceedings of 5th International Conference on Bioinformatics and Computational Biology.* International Society for Computers and Their Applications.

3.1 Introduction

Compared with other major types of cellular networks (e.g., transcription networks that capture genetic regulatory mechanism, and signaling networks that depict signal transductions in a cellular environment under different contexts such as diseases and extracellular stimuli), metabolic networks are well studied biological systems with rich data resources such as bio-chemical databases [67, 134, 152]. During the past decade, substantial work has been devoted to the understanding of the metabolism using reconstructed metabolic networks. Researchers have studied how the topological information of metabolic networks implies or relates to important cellular processes and biological functions [88, 177, 11]. From this perspective, topological hallmarks of complex networks, including clustering coefficients [134], hierarchical organization [51, 142], and local interaction patterns [165, 126] have been heavily investigated. By viewing networks as abstract representations of systems, topological properties such as clustering coefficient and skewed degree distributions are regarded as network phenomena underpinned by complex networks rather than casual individual patterns. In particular, modular and hierarchical organization is one of the well understood characteristics of metabolic networks; and it has been further demonstrated that metabolic networks exhibit functional cartography in organization [134, 51, 38, 188].

The analytical backbone for studying modularity and organization of networks boils down to community detection problems in graphs, which is one of the most important trends in network mining and analysis [41].

In a sparse network such as the metabolite-centric network considered here, the task of finding clusters of nodes (a.k.a. communities and modularity) is often considered as a one-step process based on network topological information. However, it is necessary to control the sizes of communities in different network partitions since community detection is subject to resolution limits (which means certain small node groups cannot be detected as communities) leading to multiple ways of tracking functional compositions for different sizes

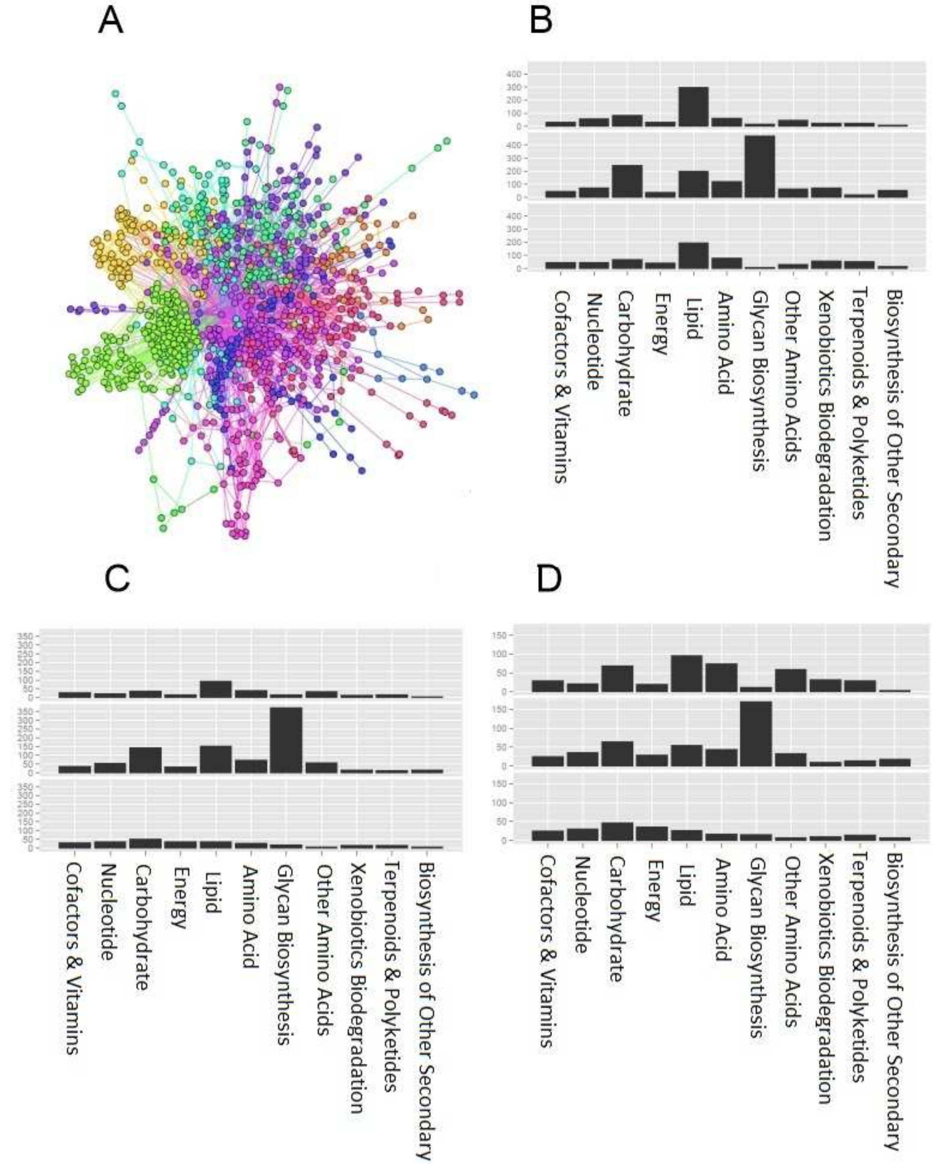


Figure 3.1: A) The reconstructed metabolic network for the *E. coli* data; each color of the nodes represents a community. B) Distribution of metabolite links in functional classes with resolution parameter 0.5 for the *E. coli* network. C) Distribution of metabolite links in functional classes with resolution parameter 1.5 for the *H. sapiens* network. D) Distribution of metabolite links in functional classes with resolution parameter 2.5 for the *S. cerevisiae* network

of node groups (Figure 3.1 (B-D)). The reason for controlling the resolution of community structures underlies the complex relational structure of large networks. In other words, the modular structure offers little value to understanding the network organization if we are dealing with a small number of nodes. However, this is not the case when we consider large-scale metabolic networks that include entire metabolic activities of different organisms. Here we want to understand how different community structures can be mined and controlled in complex networks. For this purpose, we consider a Markov process in computing the modularity that was shown to be robust and flexible in the network settings. Under this context, we study different community structures (known as multi-scalability of network communities) via a parameter that controls the process, called scale or resolution (γ) of the associated community structure. Figure 3.1 (A) shows a reconstructed metabolic network and its natural community structure from one-step clustering process (i.e., with $\gamma = 1$).

In terms of the metabolic networks studied in this chapter, the multi-scale organization of the community structures provides insight into the cellular mechanisms of how certain metabolites are grouped together, and in what physical proximity (i.e., the relationship with cellular compartmentalization) they exist and function. Furthermore, by considering different levels of metabolic organizations, we are able to computationally track the degree to which scale fluctuations affect functional cartography [165, 126, 113]. For example, Figure 3.1 (B–D) shows the multi-scalability effect and distribution of the functional classes given by links within the communities of metabolic networks in different partitions. Each bar of the histogram represents the largest number of links in a community that belongs to a functional class.

Network motifs are frequently occurring subgraphs compared with an ensemble of random networks. They are seen as the superfamilies of evolved units [109, 131, 74]. It is suggested that network motifs such as feed-forward loops (FFL) and single-input modules (SIM) act as fundamental units that drive evolving networks. Based on z-scores, motifs are classified into motifs (z-score>0) and anti-motifs (z-score<0) [109]. In a sense, both modular structure and network motifs are viewed as evidence of the evolving organization in complex cellular networks. However, the relationship between these two entities is not well understood; "who drives whom" at least topologically remains an open question. For example, Goemann *et al.* used a pair-wise disconnectivity measure to evaluate network motifs from the topological information of transcriptional networks [49] in the entirety, rather than peeking into community-dependencies of motifs. Interestingly, although it is hypothesized that network motifs to some extent play evolutionary roles, it has been disputed whether various types of motifs have tight connections to biological functions. From these arguments, it is suggestive that although motifs reflect network evolution, the way of interpreting them is perhaps more important. In other words, the existence of network motifs and their functional roles are context dependent—an implication of concern in recent research efforts [66, 81]. We suspect that the complexity of modular structures in metabolic networks is connected to such contextual dependency [82]. In this chapter, to instantiate our findings we investigate the coupling effect between these two seemingly parallel causes with metabolic networks of three organisms, namely (*Homo sapiens*, *Saccharomyces cerevisiae* and *Escherichia coli*).

In a nutshell, given a reconstructed metabolic network we investigate multiple network partitions with different resolutions and we study their effect on how network motifs are distributed. We first show how the direction of edges affects multiple network partitions and motif distributions. Further, with different network partitions, we analyze the distribution patterns of motifs in individual communities. We then focus on specific motif types in metabolic networks and communities that contain them in order to understand the effect of motif and anti-motif distributions. To deepen the analysis, we compare the physical cellular compartments with different community profiles and quantify their pair-wise variations. We finally use two different methods to quantify the degree to which motifs are dependent in different community structures. This allows us to rank network motifs in the presence of community background.

3.2 Materials and methods

3.2.1 Construction of metabolic networks

Metabolic networks are constructed using BiGG [146], which is a curate database for exploring biochemical information of metabolism. We constructed metabolite-centric networks in our study. The choice of metabolite-centric networks originated from the existing literature [67, 51], they are natural representations of the biochemical reactions available in studying metabolism. In a metabolic network, nodes i and j represent metabolites and two nodes are linked if they participate in some biochemical reactions where i is substrate (product) and j is product (substrate). The advantage of the network representation mainly lies on the global view of complex metabolic systems, and the ease of studying mesoscopic properties and organization of networks. There are two issues concerning the construction process:

1. Reversibility information of biochemical reactions: In the BiGG database if a reaction is reversible, then nodes are connected by bidirectional edges; otherwise nodes are linked with unidirectional edges.

2. Topological reductions of a metabolic network: In order to enhance the structural compactness in multi-scale modularity analysis, we implemented the process described in [187, 134, 166]. We excluded the following 15 common metabolites, i.e., cofactors: ATP, ADP, AMP, NAD, NADH, NADP, NADPH, NH_3, CoA, O_2, CO_2, Orthophosphate, Glutamate, Pyrophosphate, and H. To make sure that the constructed networks are not defragmented, we checked connected components of networks before and after the removal of metabolites, and only 3 nodes in total became isolated after the removal. The overall con-

nectivity of the constructed networks is therefore not substantially affected. To obtain organism specific metabolic networks, we first find genes that encode enzymes in that organism (enzyme-gene network and enzyme-reaction network); then, two binary matrices corresponding to enzyme-gene and enzyme-reaction networks are multiplied to get the organism specific network [98].

3.2.2 Dynamic multi-scale modularity profiling

Compared with other views of graph clustering such as Normalized Cut [151], this formulation of community detection allows probabilistic and dynamic interpretations [139, 36, 31]. The transition matrix P of random walks can be computed by $P = D^{-1}A$, where A is the adjacency matrix of the network with $A_{ij} = 1$; if there is an edge between nodes i and j, $i, j \in V$, 0 otherwise, and diagonal matrix $D = diag(d_i)$, with $1 \leq i \leq |V|$ and $d_i = \sum_j A_{ij}$ denoting the degree of node $i \in V$. The probability of arriving at node j after t steps $p_{ij}^{(t)}$ can be obtained by:

$$p_{ij}^{(t)} = (P^t)_{ij} \tag{3.1}$$

It is easy to show that the stationary distribution π_i of the Markov chain for the undirected network when $t \to \infty$ is $d_i / \sum_{i \in V} d_i$, satisfies the condition $\pi P = \pi$.

Traditionally, the modularity Q of a network can be seen as a quality measure [116, 51] that takes the form: $Q = \sum_{ij} (C_{intra} - C_{exp})$, where $i, j \in V$, C_{intra} is the fraction of intra-community links and C_{exp} is the expected fraction of intra-community links. The computation of expected weights relies on the choice of the null model through which the graph dynamics, e.g., normalized Laplacian dynamics $\frac{\partial p_i}{\partial t} = \sum_j \frac{1}{d_j} A_{ij} p_j - p_i$ over Eq.(3.1) can be used. The stationary distribution or steady state of the normalized Laplacian dynamics is given by $p_i^* = d_i / \sum_{i \in V} d_i$. The dynamic view of modularity offers many useful insights in interpreting the connection between theoretic processes and network topological information [139, 31, 113]. Assuming that random walkers start from a steady state in undirected networks, the null model can be derived as the likelihood of two independent random walkers remaining in the same community after time t. The modularity is then defined as: $Q = \frac{1}{2m}\sum\limits_{ij}\left(A_{ij} - p_i^* p_j^*\right)$. The problem with this modularity maximization is that it is subject to a resolution limit, meaning that it is possible that some small communities cannot be found by optimizing the above equation [42]. This calls for the attention to the fact that simply optimizing Q may lead to incomplete community structures. For this reason the control of resolution in finding communities interplays with many features of modularity analysis [36]. Taking the control of resolution into account, modularity can be written as:

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \gamma p_i^* p_j^*\right), \tag{3.2}$$

where γ is the resolution parameter [113]. Eq. 3.2 is a combinatorial view for choosing different ways of membership assignments to optimize the modularity. In other words, we are seeking the best way to assign community membership to nodes in the network in order to achieve the highest Q measure, and such search is often based on some heuristic. The second term $\gamma p_i^* p_j^*$ in Eq. 3.2 refers to the expected fraction of intra-community links; by varying the parameter γ the expected fraction is re-weighted hence changing the overall modularity measure (Q). One recent interpretation for scales of modularity, is inverse time ($\gamma \propto \frac{1}{t}$) relating to the stability of communities structures [85, 31].

For fast community detection, we adopted the Louvain method which is a heuristic based approach for finding communities structures [16]. The Louvain method involves the calculation of change in modularity, $\triangle Q$, in searching for the best improvement of modularity in each iteration. Since we constructed directed metabolic networks, we derived $\triangle Q$ for directed networks using the Louvain method.

3.2.3 Network motifs as evolutionary units

The significance of a network motif i is measured by the z-score:

$$Z_i = \frac{Nreal_i - < Nrand_i >}{std(Nrand_i)},$$

where $Nreal_i$ is the count of the occurrences of subgraph i in the network; $< Nrand_i >$ and $std(Nrand_i)$ are the mean and the standard deviation of the number of times subgraph i occurs in randomly generated networks, respectively. There were disputes about the origin of motifs in networks, i.e., whether they arise due to functional reasons or topological information [66, 82, 81]. We think that motifs need to be calibrated within the community context. Therefore, our contribution is different from the recent works described in [163, 49]. For fast motif detection and detailed statistics, we used the FANMOD package [174].

3.3 Results

In this section, we discuss the relationship between multi-scale community profiles and motif distributional effect. In this chapter, we only study 3-node and 4-node motifs found by the FANMOD software (i.e., connected subgraphs of size 3 and 4).

3.3.1 Motifs tend to distribute synchronously regardless of directionality

We first investigate the interplay between network directionality and multi-scalability of community structures because the direction of edges in networks affects community detection in computational methods. In particular, null models are different for directed and undirected networks in computing the modularity measure (Q) in the Markov process [85, 113]. To do this, we first collect motifs in the directed networks while optimizing the modularity with different scales (γs). We then repeat the process with directed networks but by ignoring the direction of edges (by treating the directed edges as undirected edges and by eliminating redundant links. For example, the directed edges $A \rightarrow B$ and $A \leftarrow B$ are replaced by an undirected edge A–B) when optimizing Q using the null model associated with undirected networks. To collect network motifs, we first apply a motif detection algorithm to discover all motifs in the network, then we run the community detection algorithm with various values of the resolution parameter (γ) to produce a set of community profiles, each corresponds to a different value of γ. For each community formation P_γ induced by the parameter γ, we count the total number of motifs belonging to individual communities of P_γ . We say a motif m belongs to a community C if and only if all the nodes of m belong to C.

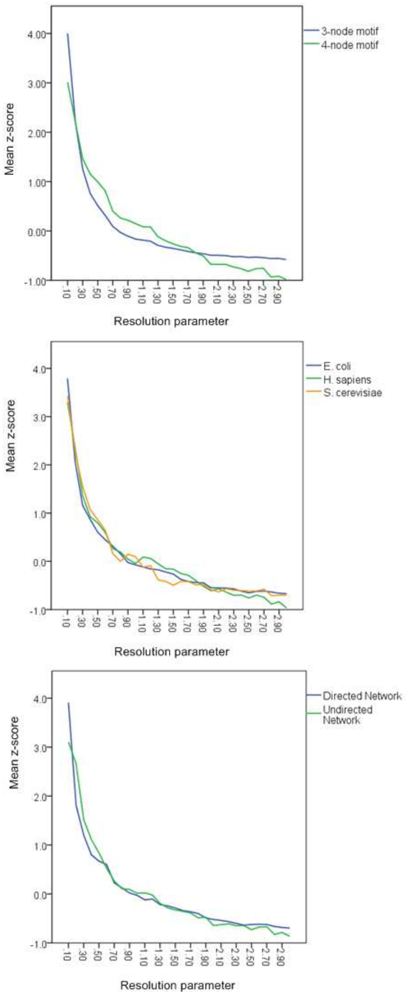


Figure 3.2: Motif distributional effect. Mean $z-$score for the total number of motifs with increasing resolutions parameter (γ), grouped by motif category (top), organism (middle) and directionality (bottom)

Here we aim to check how the directionality of motifs (regardless of motif types) as a factor affects the distribution of motifs in different community profiles. In Figure 3.2 , we show the mean z-score of the total number of 3-node and 4-node motifs with increasing γ. As γ increases, generally we expect the number of motifs to drop monotonically because the number of communities increases with increasing scale (γ), resulting in more motifs getting cut. Figure 3.2 shows that the total number of directed and undirected 3-node and 4-node motifs decreases with similar patterns. This suggests that directionality plays a minor role in motif distributional effect. We note that in Figure 3.2 (top) the mean z-score of the 3-node motifs decreases faster than the 4-node motifs when the resolution parameter is below 0.3; however, this is reversed when the resolution parameter is between 0.3 and 1.7.

We then check to what extent different motif occurrences are over or less represented in different network partitions. In other words, communities in different partitions form the background and network motifs are entities of interest in the presence of such background [190]. In a sense, the motif occurrences are decided at the class level, which is the level of magnitude of the z-scores (similar to quantifying the degree to which a subset of genes out of the background set are reverent to different biological processes). Suppose we have M motifs in the metabolic network of organism o, and assume N out of the M motifs have been observed in the counting process as described above, M_s of the M motifs are significant in terms of the z-score (i.e., z-score exceeds the predefined threshold), and N_s of the N motifs are significant in terms of the z-score. We computed the p-value to quantify the statistical significance of motifs, which is a probability of at least N_s motifs have z-score above threshold, if we were to select M_s motifs at random in the metabolic network. The p-value is given by:

$$p^o(M_s, N_s, M, N) = \sum_{N_s \leq x \leq \min\{M_s, N\}} \frac{\binom{N}{x}\binom{M-N}{M_s-x}}{\binom{M}{N}}$$

where $\binom{a}{b}$ is the binomial coefficient. We observed subtle p-value differences. This indicates that the enrichment effect of motifs under different partitions is slightly different but not exactly the same. For example, with $\gamma = 2.3$, the p-values for the *S. cerevisiae* and the *E. coli* metabolic networks are ≤ 0.0001, whereas the p-value is 0.0002 (z-score threshold is set to $|z| > 3$, Bonferroni correction applied) for the *H. sapiens* metabolic network with the same γ.Interestingly, this finding is counter-intuitive compared with Figure 3.2 (middle) where the

z-score for the number of motifs decreases monotonically for all organisms. The results lead to an interpretation that different community formations endorse different network motifs. In other words, motif significance is contextualized by the scale of community structures provided that two algorithms (community detection and motif finding) are independent. This interpretation offers an alternative evidence on the role of motifs that is different from those which argue that motifs arise solely based on the topology of networks [66, 81, 82]. As the occurrences of motifs in different community profiles show (Figure 3.2), it is suggested that only a proportion of significant motifs (based on the z-score) arise from the network topology; for others to play functional roles, they need to be considered in communities with certain resolution.

Another observation is that the occurrences of directed motifs tend to juxtapose with the occurrences of undirected motifs (Figure 3.2 (bottom)). This suggests that community profiling is robust against motif occurrences in general. With the view that motifs are favorably chosen by communities, the directionality of metabolic networks is nevertheless irrelevant in "hosting" different network motifs. Although the number of motifs tends to be larger in undirected networks, which attributes to topological properties (types of directed motifs is larger than undirected motifs), their distribution patterns are similar. As a result, considering one type of networks, either directed or undirected, suffices for analyzing motif occurrences in subsequent sections.

Comparison with random networks

We performed the simulation study to check how likely are motifs to be generated, given a certain partition. Given a metabolic network, we first obtain the community profile with a resolution parameter; we then generate degree-distribution-preserving random networks for each community in the partition. Treating communities as subnetworks, we simulate the global random network by preserving degree distributions of individual subnetworks. We observed that as community profiles are coarsened and individual communities become relatively dense, motifs are found to have different significance values (some of them are no longer significant from the motif finding algorithm). For example, for motif 238 in the yeast metabolic network, the z-score decreases from 51.47 to 19.89 (resolution = 2.5) and to 22.99 (resolution = 3.0), whereas for motif 102, the z-score decreases from 5.57 to 3.43 (resolution = 2.5) and 4.5 (resolution = 3.0). This suggests that motifs are likely explained by different network partitions.

Reversely, given that we have a number of motifs, we create random networks that approximate this particular profile of motifs and compare the average modularity value with the original network. We first start with the same number of different motifs in the original network, and randomly connect the nodes of motifs with random directions so that the total degree distribution approximates the original network. This process was repeated 10 times for each resolution parameter. We observed that the average modularity generally deceases when the resolution parameter increases for the simulated random networks. For example, for the yeast metabolic network, the average modularity value was 0.41 (resolution parameter = 2.0) compared to 0.56 in the original network, whereas with resolution parameter 2.5 the average modularity value was 0.39 compared to 0.54 in the original network. This again suggests the cause-and-effect relationship between network partitions and motif distributions.

3.3.2 Communities enriched with most motifs show homogeneity across scales

In light of the above lessons, we assume that motifs drive evolving networks, perhaps not globally but locally. We then raise the question: to what degree do motifs drive network partitions? Communities at different resolutions can be interpreted as the primary organization principle of metabolic networks, whereas motifs can be regarded as the secondary organization principle that depends on network partitions. Here we define intra- and inter-community motifs as follows: intra-community motifs are subgraphs where all edges occur within a community, and inter-community motifs are subgraphs with edges straddling different communities.

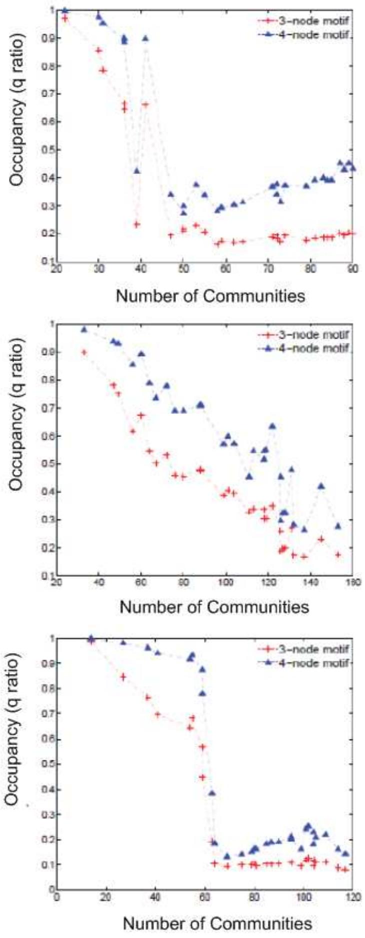


Figure 3.3: Motif enrichment effect. The relationship between the number of communities in different partitions and the occupancy ratio (in terms of largest communities across partitions) q for 3-node and 4-node motifs in the directed networks of the three organisms

To answer the aforementioned question, we analyzed the relationship between the community profiles and the number of occurrences of 3-node or 4-node motifs in one single community under the given resolution. For 3-node motifs we define

$$q_3 = \frac{\text{number of 3-node motifs in the largest community in a partition}}{\text{total number of 3-node motifs in the network}}$$

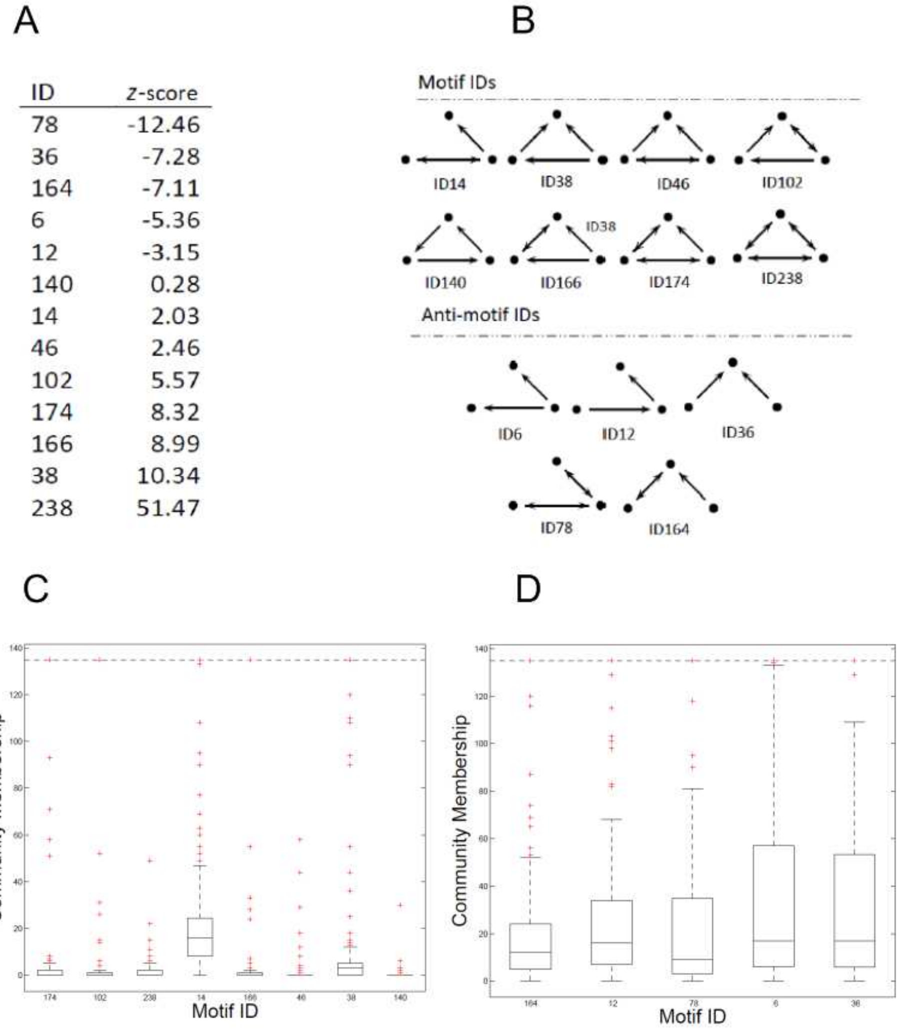
; q_4 can be defined similarly for 4-node motifs. The q ratio can be interpreted as the rate of occupancy of 3-node/4-node motifs in the largest community of a network partition. The ratio ranges from 0 to 1, with 0 being the situation where motifs are not found in the community structure induced by γ, and for 1 motifs are all found in a community with the same resolution. Figure 3.3 shows that under different community structures 4-node motifs have the higher occupancy rate than 3-node motifs in three organisms. Since the q-ratio measures the proportion of occurrences of 3-node/4-node motifs in the largest communities of different partitions, their decreasing patterns against the number of communities reflect that the occurrences of network motifs are related to modular structures for varying resolution parameters. This observation consequently mirrors the fact that subsets of metabolites participating in various chemical reactions and regulating cellular activities tend to self-organize in community structures.

We also notice that individual distributions of 3-node and 4-node motifs are distinct in the three organisms. For example, in the *S. cerevisiae* network there is a sudden drop at $s \approx 40$, and the same effect is observed in the *E. coli* network around $s \approx 60$. With relatively coarsened communities in *S. cerevisiae* and *E. coli*, the the q ratio reveals a pattern of a steady increase followed by a sharp descend at $s \approx 48$ and $s \approx 61$, respectively. However, this pattern is not observed in the *H. sapiens* network, possibly due to the higher complexity of mammalian cellular mechanisms and relatively incomplete collection of ORFs.

3.3.3 Type-specificity of motifs reveals different distributional patterns

We now turn to type-specific motifs and study their distributional effect in communities across scales, and here we focus on 3-node motifs. We categorized 3-node motifs by their z-scores [109, 49, 163] into motifs (where z-score > 0), and anti-motifs (where z-score < 0, Figure 3.4 (A & B)).

We observed that both motifs and anti-motifs occur in communities from different partitions as seen from Figure 3.4 (C & D). In order to check how motifs and anti-motifs occur in different community profiles (i.e., with different resolution parameters), we plotted the number of occurrences of motifs and anti-motifs by considering all community profiles (γ ranges from 0.1 to 3). We excluded communities with less than 20 nodes. From Figure 3.4 (C)



ID	*z*-score
78	-12.46
36	-7.28
164	-7.11
6	-5.36
12	-3.15
140	0.28
14	2.03
46	2.46
102	5.57
174	8.32
166	8.99
38	10.34
238	51.47

Figure 3.4: Motif significance in communities. A) z-scores (z) of motifs ($z > 0$) and anti-motifs ($z < 0$) in the directed *S. cerevisiae* network. B) Motif and Anti-motif graphs and their IDs. C) Distribution of motifs in boxplots for the directed *S. cerevisiae* network. D) Distribution of anti-motifs in boxplots for the directed *S. cerevisiae* network. The y-axis refers to the total number of communities a motif belongs to

& (D), we have observed that motifs tend to occur less than anti-motifs across community profiles, and the distribution does not correlate with the z-score. For example, motif 14 has z-score 2.03, but occurs in more communities (when varying the resolution parameters) than that of motif 238 with substantially higher z-score (= 51.47, Figure 3.4 (A, C & D)).

To investigate type-specific distribution of motifs across different network partitions, we have used the two-sided χ^2 test with the null hypothesis H_0 stated as: motifs of different types are similarly distributed, i.e., similar statistical proportions in different network partitions. We observed that both anti-motifs and motifs give two-sided p-value < 0.0001 on average, with the exception of relatively small or large γ's. This suggests that both under and over-fragmentation ($p = 0.002, \gamma \leq 0.6$ and $0.0013, \gamma \geq 2.5$, respectively) results in the disappearances of the distribution pattern of significant motifs.

3.3.4 Compartmentalized view and multi-scalability

Given a metabolic network, it is interesting to know which metabolites belong to which compartments because compartments of metabolites give rise to a physical partition of the network. We have used the compartmentalized data of metabolites from the BiGG database [146]. For example, for the *Homo sapiens Recon 1* data set, there are eight compartments: Peroxisome, Extraorganism, Golgi Apparatus, Cytosol, Nucleus, Endoplasmic Reticulum (ER), Mitochondria, and Lysosome. We manually partitioned the reconstructed network with each of the compartments as one community, therefore, creating a physical partition of the metabolites, denoted as P_c. P_c can be seen as the reference community formation that corresponds to the physical metabolic organization in the cells. In the case that metabolite i belongs to several compartments, we simply assign i to the community (compartment) to which it has the largest number of links.

A natural question is: How does the physical compartmentalization of metabolites compare with the multi-scale community structures? To answer this, we compare P_c with other community formations induced by γ, $0.1 \leq \gamma \leq 3$ using variation of information (VI) [106],

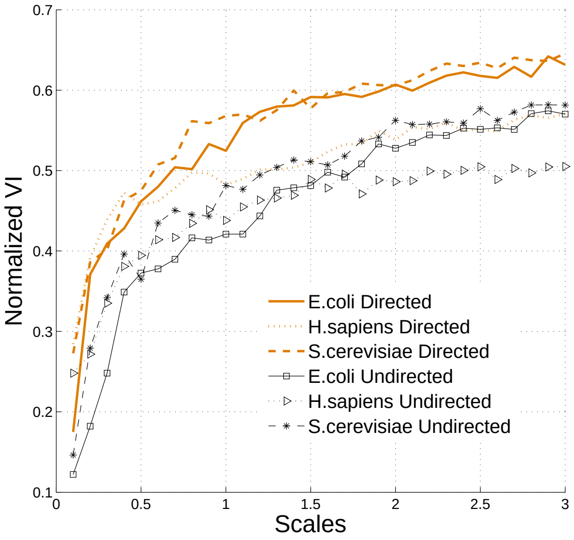


Figure 3.5: Comparing partitions with different scales. Normalized variation of information for both directed and undirected networks of the three organisms

which is shown to be a true metric in comparing partitions. The normalized VI can be written as

$$VI_{norm}(P_\gamma, P_c) = \frac{H(P_\gamma|P_c) + H(P_c|P_\gamma)}{\log N}$$

, where $H(P_\gamma|P_c)$ is the conditional entropy associated with the partition with resolution parameter γ for the given the physical partition, N is the total number of nodes in the network, and $VI_{norm}(P_\gamma, P_c)$ ranges from 0 to 1: the smaller the value of VI_{norm}, the more agreement exists between P_γ and P_c.

Figure 3.5 shows that as the scale increases, VI_{norm} for both directed and undirected metabolic networks increases logarithmically, which means that coarsened community profiles render greater uncertainty in terms of information theoretic measures in general; as the Markov time approaches infinity, the community formation gets closer to the compartmentalized view of metabolic networks. It is argued by Delvenne *et al.* that the stability of community structures increases when the Markov time approaches infinity, and this conclusion coincides with the observation from Figure 3.5: more stable community structures give less uncertainty compared with the compartmentalized view of metabolic networks, which in turn are enriched with more motifs (Figure 3.2) [31].

This fact suggests that functional cartographies provided by particular network partitions cannot guarantee to comprehensively depict physical metabolite organizations. Therefore, control of the resolution is needed for the analysis of metabolic networks in general. Taken together with the motifs distributional effect previously discussed, the variation of partitions justifies why different motifs as superfamilies of evolved networks are presented more or less in communities across different partitions, as motifs exist in community contexts which are more or less sensitive to the variation of information.

3.3.5 Quantifying the context-dependency of motifs in multiple scales

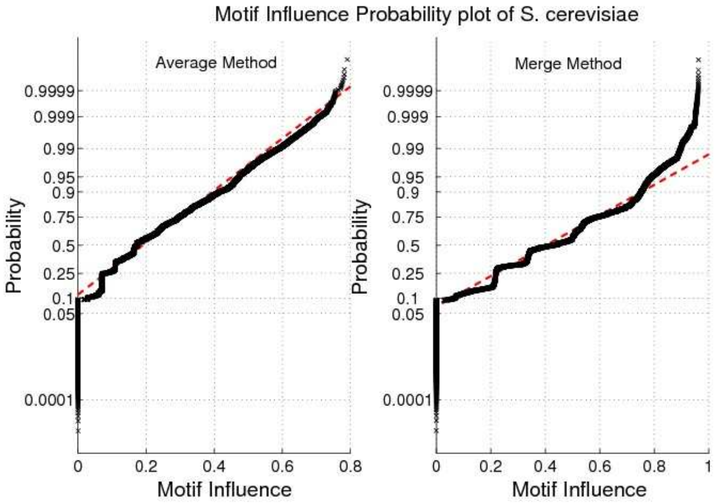


Figure 3.6: Motif influence probability. The probability plot of motif influences with multiple scales using the average and merge method for the *S. cerevisiae* metabolic network. The red dashed reference line is for judging whether the data follows a normal distribution

Given patterns of motif distribution in a community structure, we seek a measure to quantify the influence of motifs in community context. There are several terminologies in characterizing similar measures such as participation coefficient of nodes within a community structure [51], social influence of actors in social networks [77] and disconnectedness index [49] for groups of nodes. Here, we discuss two methods to quantify the context-dependency of motifs based on participation coefficients [51].

1. Average method: We average the participation coefficients of participating nodes to represent the influence of motifs. Formally, suppose we have a motif instance $M_i = (V_i, E_i)$, with participating vertices $\{v_i^1, v_i^2, ..., v_i^k\} \subset V_i$, the influence of motifs in a community structure is defined as: $\wp_{avg} = (P_{v_i^1} + P_{v_i^2} +, ..., + P_{v_i^k}) \big/ k$, where $P_{v_i^1}$ is the participation coefficient of node v_i^1, defined as $P_{v_i} = 1 - \sum_{c=1}^{k} \left(\frac{l_{ic}}{l_i}\right)^2$, where l_{ic} is the total degree of node i to all nodes in community c and l_i is the total degree of node i [51].

2. Merge method: We view a motif as a single node in metabolic networks while preserving all links of participating vertices, that is, we replace $\{v_i^1, v_i^2, ..., v_i^k\}$ by a new node v_i^{new} and all participating edges of the motif are made incident to v_i^{new}. The participation value of the motif is defined as: $\wp_{merge} = P_{v_i^{new}}$, where $P_{v_i^{new}}$ is the participation coefficient as defined in the average method. In the merge method, when two motifs share a node, we make two separate new nodes to represent each motif and link them by an edge.

The average method takes the inter-connectivity of participating vertices of motifs into consideration, whereas the merge method totally ignores it. Figure 3.6 shows the influence of motifs in the *S. cerevisiae* network. The merge method gives a higher probability of motifs in general (maximum ~ 0.96 for $\wp_{merge}$). In contrast, the average method gives the highest probability of ~ 0.79 for $\wp_{avg}$. This indicates that the inter-connectivity patterns of motifs within their corresponding communities affect the magnitude of participation coefficients overall. This suggests that the directionality between participating nodes of motifs is topologically important. From Figure 3.6, there is a noticeable set of motifs having lower participation probability in both methods (below ~ 0.05), indicating that some participating

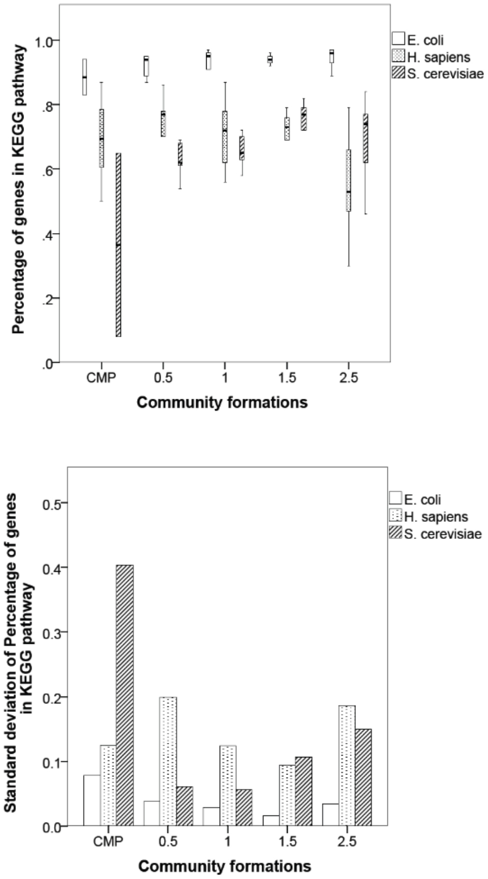


Figure 3.7: Links between community and function. Top: relationship between the percentage of genes in the KEGG pathway and different community formations. Bottom: standard deviation of the percentages of genes in the KEGG pathway with different community formations. CMP: compartmentalized partition

metabolites have substantially lower individual participation coefficients than others. However, low participation does not imply the same magnitude of functional significance of motifs as some biochemical compounds function to maintain cellular activity at low levels, such as Glucose with different biosynthesis rates and its transporter in nucleotide biosynthesis [70].

3.3.6 Linking communities with biological functions

In order to link communities with functions, we quantified the connection between communities at different network partitions with gene function overlaps using the tool DAVID [61] (http://david.abcc.ncifcrf.gov/). DAVID performs functional annotation of genes, enrichment analysis, and gene functional classification. For each organism, we chose five largest communities with five different resolutions (with $\gamma = 0.5$, 1, 1.5, 2.5 and the compartmentalized formation), and we summarized the degree of overlapping between genes in individual communities and in the KEGG pathways (Figure 3.7 (top & bottom)).

It is interesting to observe that the compartmentalized partition of the *S. cerevisiae* metabolic network gives the highest standard deviation and wider range of pathway overlapping compared with the other two organisms. At different community partitions, the degree of community-function overlapping is organism-specific, for example, *E. coli* shows higher degree of overlapping across partitions, however, the over-fragmentation ($\gamma = 2.5$) for *H. sapiens* causes more scattered overlapping than that of others (Figure 3.7 (top)). In general at moderate γ's, the functional overlapping tends to vary less (Figure 3.7 (bottom)).

3.4 Chapter Discussion and Conclusions

We observed that 8 out of 10 motifs with the highest $\wp_{avg}$ are feed-forward loops (FFL) in the *E. coli* metabolic network. The metabolites that appear most in FFLs are Glycerol 3-phosphate, bicarbonate, hexadecanoate and 2-Acyl-sn-glycero-3-phosphoethanolamine. These metabolites take part in important functions such as fatty acid biosynthesis, central metabolism, amino acid metabolism, nucleotide metabolism and lipopolysaccharide biosynthesis. The other two motifs with the highest $\wp_{avg}$ are like single input modules (SIM) without autoregulation. Though the network motifs for *H. sapiens* are similar to that of *E. coli*, with the FFLs being the largest in count, a new one which is like a FFL without the effect of the input on output is also observed. The most frequently occurred metabolites are chondroitin sulfate E, acyl carrier protein, N-Acetyl-D-galactosamine, Malonyl coenzyme A and Propionyl coenzyme A. These appear in glycan, lipid, carbohydrate, and amino acid metabolism of human. The network motifs in the *S. cerevisiae* metabolic network are similar to *H. sapiens*, but on the contrary to others, FFLs are not the largest in number. The most frequently appearing metabolites are water, malonyl coenzyme A, octanoyl-CoA, bicarbonate, deodecanoate and dodecanoyl-CoA. The functions these metabolites take part in are relatively less diverse, with mostly lipid and amino acid metabolism.

As the network structure has become a popular means to model complex relationships in biological systems (examples include genome-wide co-expression studies, gene regulatory networks, and protein-protein interaction networks etc.), modular structure is deemed important organization principles. Recent developments in complex graph clustering methods have implicated the practical applications with biological networks in different settings. In this chapter, we focused on multiple network partitions induced by different resolutions and performed comprehensive analysis to examine the relationship between multi-scale modularity and motif distributional effect in the metabolic networks using three model organisms. We observed several interesting patterns involving the effect of directionality of network in distributing 3-node and 4-node motifs, homogeneity of motif-enriched communities, and motif type-specific distribution across various partitions. We also provided a general method to quantify the community context of motifs. Overall, our work suggests that network partitions are tightly connected to motif distributional effect, and this added to the line of evidence that both modularity and network motifs could potentially evolve from modularly varying environments.

Chapter 4

MCF: a tool to find multi-scale community profiles in biological networks

[1] *We present a tool to find community structures of different types of networks.* ***M****ulti-scale* ***C****ommunity* ***F****inder (MCF) is a tool to profile network communities (i.e., clusters of nodes) with the control of community sizes. The controlling parameter is referred to as the scale of the network community profile. MCF is able to find communities in all major types of networks including directed, signed, bipartite, and multi-slice networks. The fast computation promotes the practicability of the tool for large-scaled analysis (e.g., protein-protein interaction and gene co-expression networks).*

4.1 Introduction

A network has become a popular means to model complex relationships in biological systems. Examples include genome-wide co-expression studies, gene regulatory networks, and protein-protein interaction networks, etc [11, 120, 107, 158, 172]. Often, these networks require clustering analysis, in which groups of densely connected nodes are identified (see [41] for a detailed review). In fact, modular structure is deemed important characteristic in biological networks [138, 142]. Unlike traditional clustering methods, communities (i.e., clusters) in network representation are subject to resolution limit, which means some smaller communities cannot be detected by simply optimizing the modularity measure [41]. This may cause inaccurate or misleading functional annotations of groups of nodes based on modular

[1] The content of this chapter is based on the following article:
Gao, S., Chen, A., Rahmani, A., Jarada, T., Alhajj, R., Demetrick, D., & Zeng, J. (2013). MCF: A tool to find multi-scale community profiles in biological networks. *Computer methods and programs in biomedicine, 112*(3), 665-672.

structures of networks. Considering this, statistical methods were developed to deal with multi-scale community profiles [85, 5, 113] in complex networks. In particular, Mucha *et al.* proposed a systemic approach to unfold multi-scale multiplex community structures. In multiplex networks that involve multiple time or context dependent networks slices, the same controlling parameter (referred to as the scale of the community profile) in single-slice networks is generalized to multi-slice networks [113]. Such advance in community detection is useful in studying many biological problems, including the study of time-coursed data and integrative network analysis of high-throughput data [89], as well as social networks [122].

Despite the static view of multi-scaled community profiles in controlling community sizes, recent interpretation of the scale parameter as inverse time in the random walk process can be regarded as a factor that impacts the stability of community structures across partitions, i.e., time intervals for certain communities to emerge or disappear [31, 85, 139]. This viewpoint provides many theoretical properties of graph clustering dynamics based on statistical mechanics [135]. Taken together, a tool that extends traditional graph clustering methods to allow the multi-scale capability is needed.

Existing tools for finding network communities (a.k.a., graph clusters) such as 'jClust' [127] and 'GLay' [155] only implement graph clustering methods without consideration of multiple scales. 'igraph' includes a primary version multi-scale community detection method for only undirected networks [25]. To cover the gap and satisfy the need, we developed a fast tool, Multi-scale Community Finder (MCF), based on modularity improvement heuristic in finding multi-scale community structures in all major types of networks, including (un)directed, signed, bipartite, multi-slice networks, etc. We implemented two different methods for controlling scales of networks from recent studies.

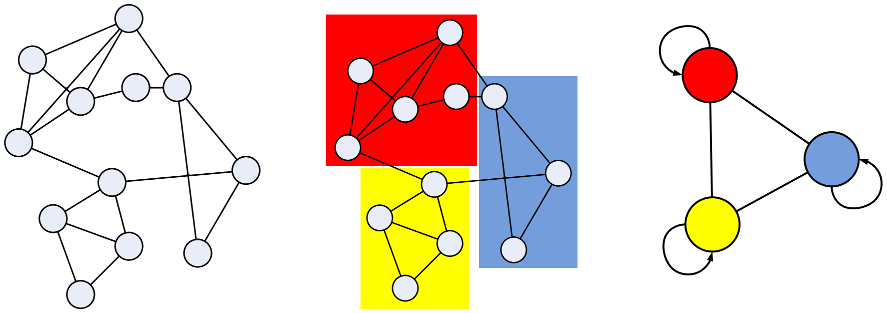


Figure 4.1: Illustration of the Louvain method, colored regions represent communities

4.2 Methods and Implementation

4.2.1 The Louvain Method

MCF is based on the existent 'Louvain' method [16], which is a heuristic based approached based on the improvement of modularity measure in each iteration. Louvain method begins by considering each node being in its own group, and then the method moves nodes to the neighboring groups that give the most increase in modularity. The second stage generates a new network by aggregating the nodes in the groups into one node, and combining the links. These two steps are repeated until the modularity score stops increasing; the modularity score provides information on the strength of the groupings produced based on the structure of the network (Figure 4.1).

The main advantage of the Louvain method is the fast computation time. We chose the "Louvain" method mainly to efficiently deal with large-scaled analysis of biological networks, e.g., the gene co-expression networks that may contain thousands of links and the rapidly growing protein interaction data. We compared the computational time with Newman method [116], detailed in Table 4.1. We observed that when dealing with large networks, MCF tends to find communities faster than Newman method.

MCF can take both single-sliced and multi-sliced networks as input. In both cases, we extend the 'Louvain' method to adapt for different types of networks by deriving modularity

Table 4.1: Comparison of MCF with Newman method, computational time is measured in seconds

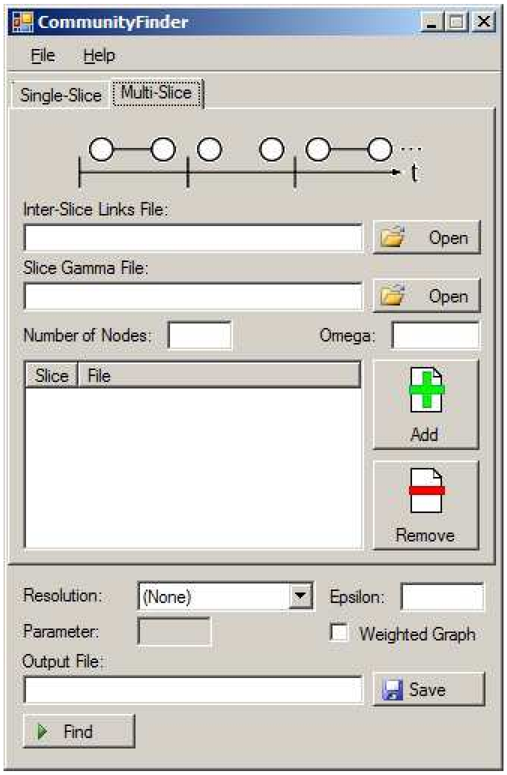
No. of nodes	No. of links	MCF(sec)	Newman(sec)
10	20	2.34	2.3
100	952	2.54	2.56
1000	9610	3.53	4.1
10000	96028	26.26	30.2
100000	1048435	7024.18	8182.32

update computations (see the Appendix 2 for details), since each network type has its own physical properties in computing modularity measure. For example, we incorporated two methods to find communities in bipartite networks in MCF.

4.2.2 Multi-slice Modularity

The multi-slice network community detection implementation is based on the recent method described in [113]. In multiplex networks, each network slice is subject to a scale parameter, and users are expected to provide link strength of node i between slice s and slice r, assuming there is some evidence of evolutionary connection of node replica in different slices. This external link can be derived in different manner, depending on applications or can be simply set to constant value as demonstrated in [113]. For example, when examining time-series data, each protein functional class known a priori can be artificially set to have one inter-slice strength.

To deal with multi-scaled computations, we implemented two methods, which are shown to be equivalent in theory. In Arenas *et al.*'s method, the adjacency matrix is modified by adding weights to diagonal entries [5], so the method requires the weight as an additional input. In Lambiotte *et al.*'s method, the scale parameter is needed to control community scales [85]. Arenas *et al.*'s method requires priori examination of the adjacency matrix, which may be difficult for users to accomplish. We have included tutorial files along with sample data sets for MCF in the project website (Figure 4.2, see `http://bsdxd.cpsc.ucalgary.`



Resolution: (None)

□ Weighted Graph

Figure 4.2: MCF for multi-slice networks. Single-sliced networks of different types are similar under single-slice tab

ca/MCF).

Finally, it is worth noting that MCF is not intended for comprehensive community detection like ‘jClust’ and ‘GLay’. Rather it is specialized for multi-scaled community finding using a fast and robust heuristic method for large graph data analysis.

Overall, MCF is a novel tool which implements an efficient method to find communities in all major types of networks. The tool is useful for large-scale network analysis in finding communities with controlled sizes.

Chapter 5

Quantifying gene co-expression heterogeneity in cancer towards efficient network biomarker design

[1]

We study cancer heterogeneity using breast cancer data. It is well known that cancer is a highly heterogeneous disease, and the predictive capability of targeted gene signature approach suffers from the inter-tumor heterogeneity. Here we propose a framework to quantify the molecular heterogeneity of tumors from gene-gene relational perspective using co-expression networks and interactome data. We believe that to understand individualized gene behavior across patients, relational status of genes needs to be considered because complex disease phenotype is often caused by cascaded failures of genetic interactions in cancer cells. We quantify gene-gene relational heterogeneity from a benchmark dataset using co-expression networks inferred from microarray data, and show that genes related to breast cancer metastasis can be stratified to different classes based on their relational status obtained from pairwise comparisons of co-expression networks. Further we use the relational heterogeneity information to predict patient survival and found that relationally heterogeneous gene set is less predictive than relatively conserved cancer genes and weekly co-expressed genes in terms of metastasis. We explore heterogenous gene sets using interactome data and identified densely connected components that are causal to inter-tumor heterogeneity, and independently validate our approach with two patient cohorts. Our results demonstrate the efficiency of using heterogeneity information to design network-based markers.

[1] The content of this chapter is based on the following article:
Gao, S., Sarhan, A., Alhajj, R., Rokne, J., Demetrick, D., & Zeng, J. (2014) Quantifying gene co-expression heterogeneity in cancer towards efficient network biomarker design. *In Submission.*

5.1 Introduction

Cancer is commonly regarded as a complex disease caused by intertwining or cascaded failures of gene products in diversified environments. In recent years, some endeavors of understanding genomic characteristics of cancer or other complex diseases focus on integrative methods with interaction networks [11], thanks to the rapidly evolving array based technologies (to reverse engineer gene-gene/gene-DNA networks) and protein hybridization protocols (to map protein-protein interactions in large scale). Networks provide unique advantages in modeling multiplex relationships between genes, proteins, and disease types etc. In fact, the notion of 'Network Medicine' has become a promising direction to identifying disease biomarkers and functional modules [23, 64, 124]. In cancers, identifying pairwise disordered relationships between genes and proteins naturally fits the goals of network modeling, that is, to study relationships and find connection patterns between nodes in a global map. To this end, many integrative methods using network data have been proposed to track differential regions of network (i.e., subnetworks) predictive of disease phenotype [94]. In targeted gene-signature strategies, many gene sets are believed to provide predictive power to cancer diagnosis and prognosis. Chen *et al.* proposed to reconcile different gene sets that are poorly overlapped using protein interaction networks [21].

In most integrative methods, a fundamental assumption is that networks reflect accurate connectedness of data. This became a critical concern in protein interaction networks due to different experimental protocols and noises [30]. For gene co-expression networks, where two genes are connected if correlated, link inaccuracies mostly arise due to data heterogeneity, given that the measuring method of co-expression level remains unchanged. Since cancer is a highly heterogeneous disease, which means tumors may exhibit different genomic landscapes, the task of finding a common target for disease prevention is extremely difficult and most gene signatures found with data heterogeneity are non-robust [93]. To this end, an ensemble of networks from the same source and a comparison mechanism for these networks are needed to reduce the amount of data heterogeneity.

The flood of gene expression data sets from disease patients, on the other hand, has provided many possibilities for studying gene expression patterns, such as finding differential expressed genes between case and control samples, or genes related to metastasis for predicting patient survival [170]. In most models, the relational patterns between genes are largely overlooked in co-expression settings. In other words, the linked structure of network is not integrated into methods to identify disease markers. This gap impedes our understanding of disease patterns, whilst in fact diseases are caused by a compendium of disordered links between functional units [35, 157]. Here we developed a method to quantify gene-gene relational differences using co-expression networks and used such information to identify key links and subnetworks related to disease phenotype.

The main strategy of this chapter is to compare different network replica of disease samples from gene expression profiles and integrate the co-complexity to protein-protein interaction (PPI) networks. The computational backbone we used is graph matching. Graph matching can be interpreted as finding nodal correspondences between a pair of networks and has been used in biological network settings such as aligning protein interaction networks of different species [76, 184]. Here, pairs of networks are inferred from gene expression data and we are to identify if gene A in one network matches gene B in the other network. By matching, we are able to quantify differential co-complex associations between gene pairs and to find conserved subnetworks (i.e., frequently matched regions) that are robust against inter-tumor heterogeneity.

To exemplify and see why the matching of two networks is challenging, let us consider the two small networks in Figure 5.1, where we have two small networks abstracted as graphs, G and G'. G' differs from G by having two additional edges (darkened links). G and G' can be seen as two gene co-expression networks inferred from a disease data set, i.e., D' is the replica of D. The orange lines are result of matching computation. For example, we found

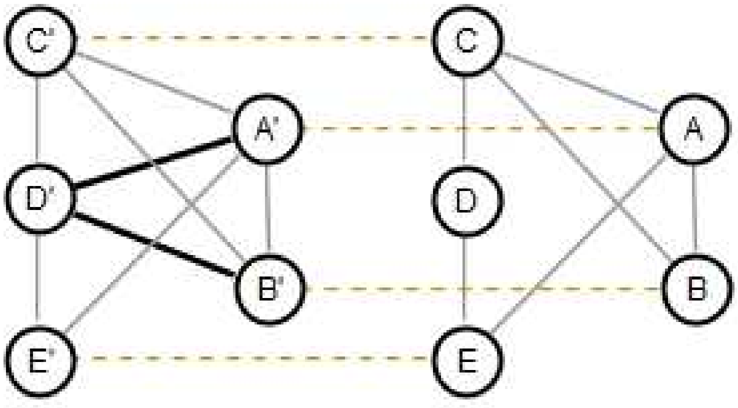


Figure 5.1: An example of comparing two networks G with nodes $\{A, B, C, D, E\}$ and G' with nodes $\{A', B', C', D', E'\}$. Orange dash lines represent correspondences of nodes, thick edges in G' represent change of connections between G and G'. In co-expression networks, nodes represent genes and edges represent gene-gene correlations

that node C matches C' because their connection status in two networks are not changed relative to unmatched nodes D and D' (D' has two additional edges). Note that since nodes are connected in networks (ignore isolated nodes), altering connection status of one node affects the rest of nodes in the network (such as A and B), but to different degree. In small networks like G and G' in Figure 5.1, it is easy to track the matching patterns of nodes, however, when dealing with medium or large size networks, tracking the relational changes of all nodes becomes a challenging task. In fact graph matching is known as a NP-hard problem [184].

In this chapter we study gene co-expression networks in breast cancer metastasis. Breast cancer is known to be a heterogeneous disease and tumor heterogeneity affects prognosis and individualized treatments [130]. Therefore, accurate predictions of patient survival become important, for example, to avoid over-treating patients with chemotherapy who would be better off without it. We detail our objectives as follows:

1. To identify gene-gene relational differences across heterogeneous cancer sam-

ples, and to provide a computational method to quantify the co-complexity of expression patterns. By this, we could be able to identify heterogeneous or unstable biological processes and functions associated with cellular phenotypes.

2. To map genes with different relational heterogeneity to interaction networks and to infer associated subnetworks. On the roadmap of designing and searching efficient network biomarkers, these subnetworks could lead to better targets for cancer prognosis and therapeutics.

5.2 Methods

5.2.1 Network setting and encoding graph features

Suppose we have two graphs $G = (V, E)$ and $G' = (V', E')$ with same number of nodes, n, the task is to find correspondences between node v_i in G and $v_{i'}$ in G', i.e., if v_i matches $v_{i'}$. We first computed different graph features for G and G' and assembled feature maps by concatenating column-wise individual feature vectors, that is, $f = [f_1|f_2|...|f_k]$ and $f' = [f'_1|f'_2|...|f'_k]$, where k is the number of features, f_i and f'_i are i^{th} feature vectors of graph G and G', respectively. We used 8 graph features: node degree, clustering coefficient, within-module degree, participation coefficient, node betweenness, subgraph centrality, average shortest path and eccentricity [40, 51, 117]. Each row in f_i and f'_i is a descriptor of nodal properties; therefore the feature map essentially encodes important topological information of nodes in a graph. Based on the feature map, we can derive the affinity matrix for graph matching problem.

5.2.2 Topology matching of nodes in two networks

The matching of two graphs is represented by a permutation matrix P, with all elements being 1's or 0's and an additional property that each row and column has exactly a single 1; $P_{ii'} = 1$ if node i in G (denote as ${v_i}^G$) matches node i' in G' (denote as ${v_{i'}}^{G'}$), and only one match is allowed for each node. Graph matching problem involves pair-wise constraints, which means if we have two candidate matches between ${v_i}^G$ and ${v_{i'}}^{G'}$, ${v_j}^G$ and ${v_{j'}}^{G'}$ the compatibility of two simultaneous matching (i.e., the connection pattern between node i and j in graph G, and between node i' and j' in graph G' are the same or similar) is encoded in the pair-wise affinity matrix M with dimension $|E| \times |E'|$ by $|E| \times |E'|$, where $|.|$ denote the cardinality of a set. Finding the optimal P for graph matching is known to be a NP-hard problem. When matching two gene networks, the problem is simplified to labeled graph matching, in which we know the labels of the nodes (i.e., gene symbols in networks). To find optimal matching between two networks in labeled settings, we aim to find correspondences between nodes by using high dimensional graph features between two networks previously described. M is computed by column-wise Euclidean distance between feature matrices f and f', the permutation matrix is then computed by finding $\mathbf{argmax}(x^T M x)$, with x being the indicator of clusters of nodes with pairwise matching constraints [90]. The matching problem in the quadratic form is formulated as an eigenvalue problem and is known as the spectral method in graph matching, which is essentially a heuristic method (by using principle eigenvectors). When dealing with large graphs like biological networks, spectral method is more efficient than other optimization-based methods.

5.2.3 Data preprocessing and network construction

Breast cancer data sets used were retrieved from NCBI GEO database (`http://www.ncbi.nlm.nih.gov/geo/`) with the following IDs: GSE2034 (Wang data), GSE1456 (Pawitan data), and GSE6532 (Loi data). To avoid data bias, only Affymetrix chips of the same platform (HG-U133A/B) are used. Gene expression data were analyzed with MAS5.0 algorithm, log2 transformed and then median-centered across arrays.

Gene co-expression networks have become a popular means and a system-wide proxy in modeling complex gene-gene relationships. Here we constructed co-expression networks by using the rank-based method [141]. Briefly, Pearson correlation coefficient is first computed for gene A with all others genes, and then ranked based on the magnitude of correlation. Ranked-based method reconstruction is shown to be more robust and accurate in large scale co-expression networks. We used the correlation threshold 0.7 and maximal number of neighbors (top ranked genes) 10 in our study.

5.2.4 Network data and inference

The protein interaction data is obtained from Reactome [24], KEGG [73] and IntAct databases [78]. Network inference algorithm is based on the 'neighboring approach': finding linker nodes to connect the gene set of interests by allowing different number of genes external to the set in the protein interaction network [4]. The p-value for the significance of inferred networks is estimated as the probability of obtaining the network with the same or larger number of nodes from random gene sets with the same number of mapped genes (from the gene set of interests to the reference network).

5.3 Results

5.3.1 Quantifying co-expression heterogeneity of breast cancer patients

We performed genome-wide survival screening by univariate Cox regression on Wang data set [170] with total 286 patients, among which 93 tumors metastasized during follow-up visits within 5 years of surgery therefore were categorized as 'metastatic' and 183 tumors showed no evidence of distant-metastasis and were categorized as 'non-metastatic' in the experiment. 10 patients were censored at the last follow-up. We obtained 1102 probes related to patient survival (p-value < 0.05) and we added 324 genes known related to breast cancer from Network of Cancer Genes 3.0 (NCG) database (which maps to 468 probe sets) [29]. We stratified patients by Estrogen receptor (ER) status, and then constructed the rank-based gene co-expression networks on 1000 bootstrapped samples with genes from survival screening and NCG with duplicated probes removed. We represented genes at the probe

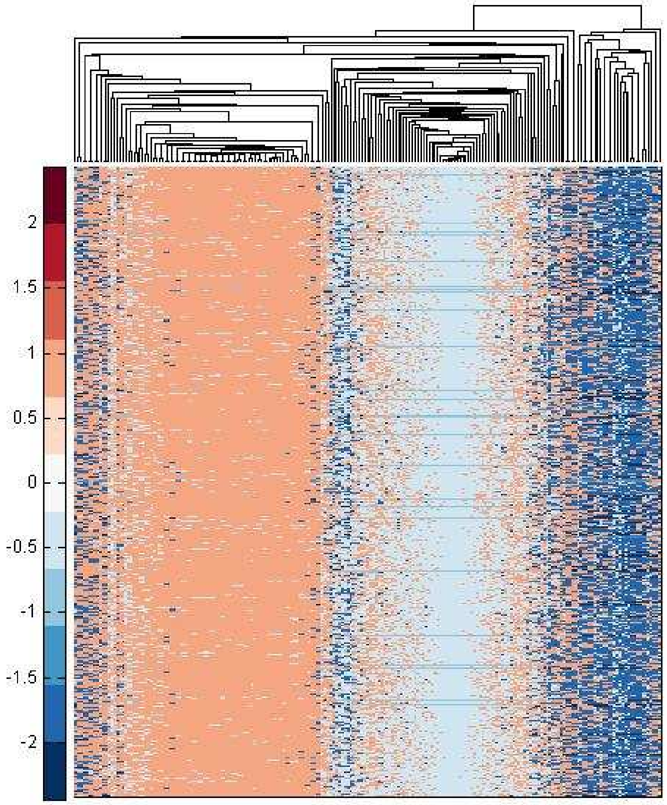


Figure 5.2: Hierarchical clustering results of 1000 pairwise co-expression network matching using city block distance. For each pairwise network matching, a gene is assigned value 1 if matched, 0 if unmatched and -1 if genes are isolated in either network

level in co-expression networks in order to infer more accurate relational status.

Figure 5.2 shows three distinctive clusters for genes in pairwise co-expression network matching, corresponding to different relational heterogeneity levels. We obtained three gene classes based on this information: Co-expression Conserved Genes (CCGs), Co-expression Heterogeneous Genes (CHGs) and Isolated Genes (IGs), containing 49, 55 and 24 genes, respectively. We summarized matching results for each cluster in Table 5.1, which shows that genes in each cluster do not exhibit monotonic behavior, i.e., they could be matched,

Table 5.1: Summary for average number of matched, mismatched and isolated genes in each gene cluster based on 1000 pairwise matching of co-expression networks for ER+ patients. The maximal number for each gene cluster is bolded

	Avg. Matched	**Avg. Mismatched**	**Avg. Isolated**
CCG	**830**	119	51
CHG	271	**656**	73
IG	245	125	**630**

mismatched or isolated in different pairs of co-expression networks. These results suggest that:

1. Relational gene expression patterns of breast cancer patients are highly heterogeneous. Therefore, co-complexity information is an important factor that may improve the predictive power of patient survival.

2. Modular patterns of gene networks (or gene clusters in general) obtained by simply applying graph clustering algorithms are non-robust in developing cancer biomarkers because of the inter-tumor heterogeneity of gene-gene associations (Table 5.1), an added line of evidence from Li *et al.* who showed that most gene signatures from 'one-step clustering process' are non-robust in predicting patient survival [93].

5.3.2 Heterogeneous gene sets are predictive of patient survival

To show that gene classes based on relational heterogeneity level, CCG, CHG and IG, have impact on the prediction of patient survival, we define the Relational Heterogeneity Score (RHS)for each patient as follows:

$$RHS_{CHG} = \sum_{i=1}^{|CHG|} I_{ER} w_i x_i + \sum_{i=1}^{|CHG|} (1 - I_{ER}) w_i x_i$$

where $|CHG|$ is the number of genes in CHG, I_{ER} is 1 if the tumor is ER positive as measured in the original experiment, w_i is the Cox's regression coefficient in the survival screening for the corresponding probe i with expression value x_i on a log2 scale. RHS for CCG and IG can be defined similarly. RHS essentially measures the level of heterogeneity of patients for genes that are found to be relationally different in co-expression networks.

We first calculate $RHS_{CHG}, RHS_{CCG}, and RHS_{IG}$ individually for each patient and then apply logistic regression to predict distant metastasis of tumors; it is interesting to observe that CCG and IG are more predictive in the regression model (with coefficients 0.40 and 0.48, p-values 0.0447 and 0.0001, respectively) than CHG. This suggests that although being relationally heterogeneous in co-expression networks, genes that are more relationally conserved in breast cancer are more predictive of disease outcome. These genes are interpreted as ‘all-time’ genes whose change of interaction dynamics is crucial in breast cancer prognosis.

In addition to CCG being predictive of metastasis, we surprisingly find that IG has similar predictive power; however, IGs are more probable ($> 60\%$) to be isolated in co-expression networks (Table 5.1) therefore are not expected to affect functional interactions. Here we provide two explanations:

1. Co-expression data is not entirely the casual factor for breast cancer metastasis. Interaction dynamics at the protein network level could be more predictive in a sense to account for inter-tumor heterogeneity, i.e., genes that are subject to various physical interactions or associations (rather than statically co-expressed in reverse engineerings view) must be considered and PPI data needs to be integrated into the prediction model in search of network markers.

2. 2. In the ensemble of co-expression networks, we imposed strong correlation threshold (> 0.7) and considered at most 10 correlated partners for each gene. Therefore, IGs are obtained in a strong correlation setting. The apparent predictability of IG for patient survival suggests that weakly co-expressed genes or transient interacting proteins are crucial ‘stabilizers’ to the disease [30].

Since a gene can possibly be matched, mismatched or isolated in the ensemble of pairwise

Table 5.2: Classification table for combined and non-combined CCG, CHG, and IG for ER+ patients in the logistic regression model. Combined statistics are in parenthesis

Actual group Metastatic or not	**Predicted group Metastatic or not**		**Percent correct %**
	No	Yes	
No	112(109)	17(20)	86.82(84.50)
Yes	33(38)	47(42)	58.75(52.50)
Percent of cases correctly classified			76.08(72.25)

matching, we test if the combined gene class, i.e., the union of CCG, CHG, and IG, provide better classification performance than that of individual gene classes in the regression model (Table 5.2), with combined RHS defined as the sum of $RHS_{CHG}, RHS_{CCG}, andRHS_{IG}$. We observed that the combined gene class gives an impaired classification performance in general for both metastatic and non-metastatic patients (area under the ROC curve, AUC = 0.756, 95% CI 0.692-0.813) compared with individual gene classes (AUC = 0.780, 95% CI 0.718-0.835). When we combined only CCG and IG, we found that the classification accuracy slightly dropped in each patient group with total percent of cases correctly classified 73.68% (AUC = 0.779, 95% CI 0.716-0.833). This in turn suggests that CHG may play a role to predict cancer metastasis. The reason for CHG being less significant than CCG and IG in terms of predictive power can be caused by the nature of relational heterogeneity: when adding up RHS of relationally conserved genes, the magnitude of RHS reflects the level of conserveness. In contrast, with CHG the weighted score is less predictable in terms of the magnitude.

5.3.3 Comparison with existing gene signatures

To further explore the effect of gene-gene relational heterogeneity in predicting patient survival, we compared the performance of CCG, CHG and IG with the 76-gene signature originated from the same data set for ER+ tumors (which contains 60 genes). We divided the total 209 ER+ tumors into high, intermediate, and low risk groups based on the Relapse Score (RS) of the 60-gene signature (RS60 predictor) and previously defined RHS for the logistic regression based predictor (RHS predictor) of CCG, CHG and IG.

Table 5.3: Gene Ontology (GO) annotations of RHS classes. FDR: false discovery rate

	Term	***p*-value**	**FDR**
CCG	GO:0051726-regulation of cell cycle	0.0022	0.0022
	GO:0042325-regulation of phosphorylation	0.00933	13.23
IG	GO:0051249-regulation of lymphocyte activation	8.63E-05	0.13
	GO:0002694-regulation of leukocyte activation	1.35E-04	0.20
	GO:0050865-regulation of cell activation	1.65E-04	0.25
	GO:0050863-regulation of T cell activation	8.61E-04	1.28
	GO:0042493-response to drug	0.004941	7.13
	GO:0002684-positive regulation of immune system process	0.006468	9.23
CHG	GO:0005829-cytosol	9.26E-05	0.11
	GO:0005198-structural molecule activity	3.64E-06	0.00
	GO:0007155-cell adhesion	8.05E-04	1.24
	GO:0022610-biological adhesion	8.13E-04	1.25
	GO:0005578-proteinaceous extracellular matrix	1.14E-05	0.01
	GO:0031012-extracellular matrix	1.97E-05	0.02

The risk stratification for each predictor is created from simple quintiles of RS and RHS. Figure 5.3 shows the result of the comparison between RS60 and RHS predictor. We observed that although two predictors were derived from different approaches, the proportions of non-metastatic and metastatic patients are comparable (left column of Figure 5.3). Kaplan-Meier curves for tumors based on the relational heterogeneity level and risk group further support this observation (right column of Figure 5.3). The only notable difference is patients with low gene-gene heterogeneity profile tend to have relatively lower survival probability than patients with low metastasis risk (orange curves, right column of Figure 5.3). This is caused by post-processing of gene signatures: Wang *et al.* used ROC analysis to find the optimal gene signature predictive of metastasis. Our aim is to view the predictor and the proximity with others from the co-complexity perspective. In fact when performed the same analysis the differences in survival curves are negligible (data not shown).

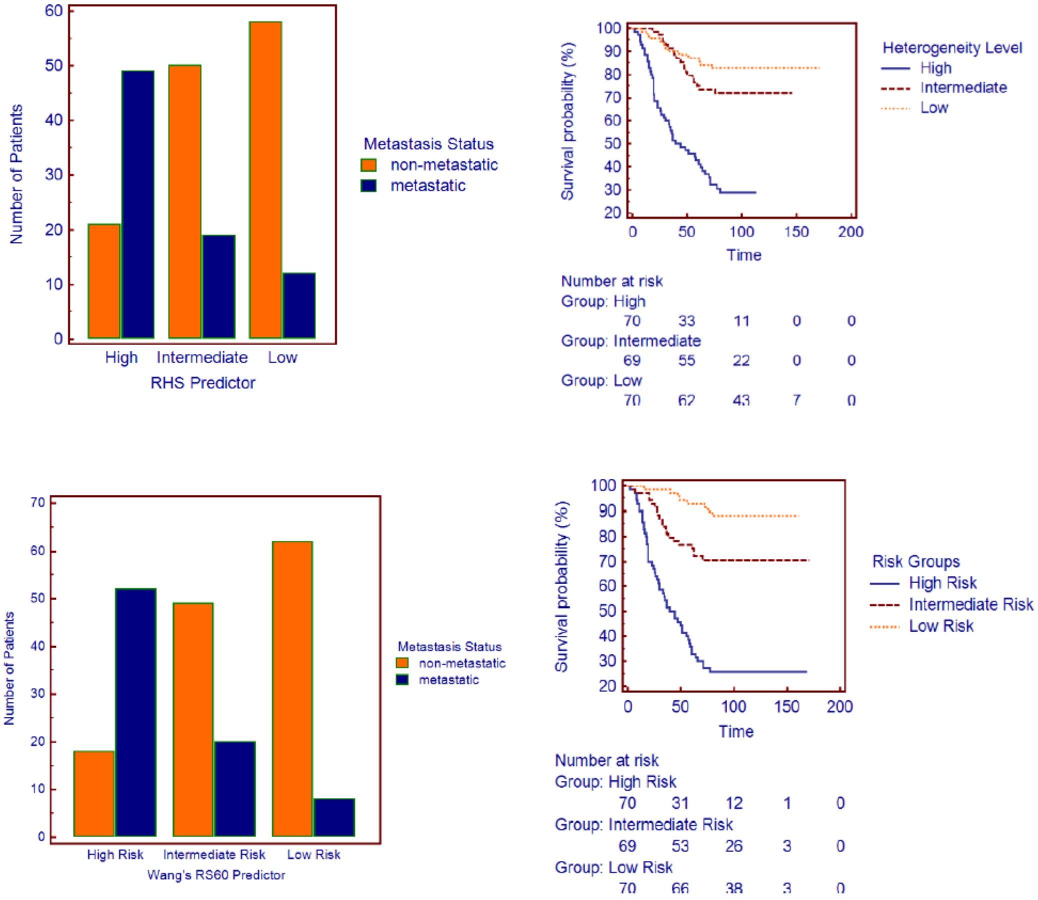


Figure 5.3: Proportions of metastatic and non-metastatic patients groups and corresponding Kaplan-Meier survival curves with 5 year distant metastasis as endpoint for gene-gene heterogeneity information (RHS predictor, first row) and Wang's 60 gene signature (RS60 predictor, second row) for ER+ tumors

5.3.4 Functional analysis of CCH, CHG and IG

From the above analysis, we see that inter-tumor heterogeneity based gene classes are predictive of patient survival, at least comparable with existing gene signatures. We performed functional annotation of gene classes used in the RHS predictor (Table 5.3). Relationally conserved and isolated genes (CCG and IG) related to metastasis correspond to hallmarks of cancer (e.g., cell cycle regulation) and post-translational modification processes (e.g. phosphorylation). In contrast, GO terms related to relationally heterogeneous genes (CHG) are more diversified. The heterogeneous property of the CHG genes from network matching coincides with diversified functional complexity of the markers. This also provides an explanation for the fact that CHG alone is not as predictive as CCG and IG. Taking this observation into account, we assume that in order to track inter-tumor heterogeneity and make efficient cancer prognosis, the gene-gene relational heterogeneity information is subsumed under the framework of targeted gene signature approaches. This assumption provides a challenging opportunity to design efficient network biomarkers from the relational perspective of genomic entities.

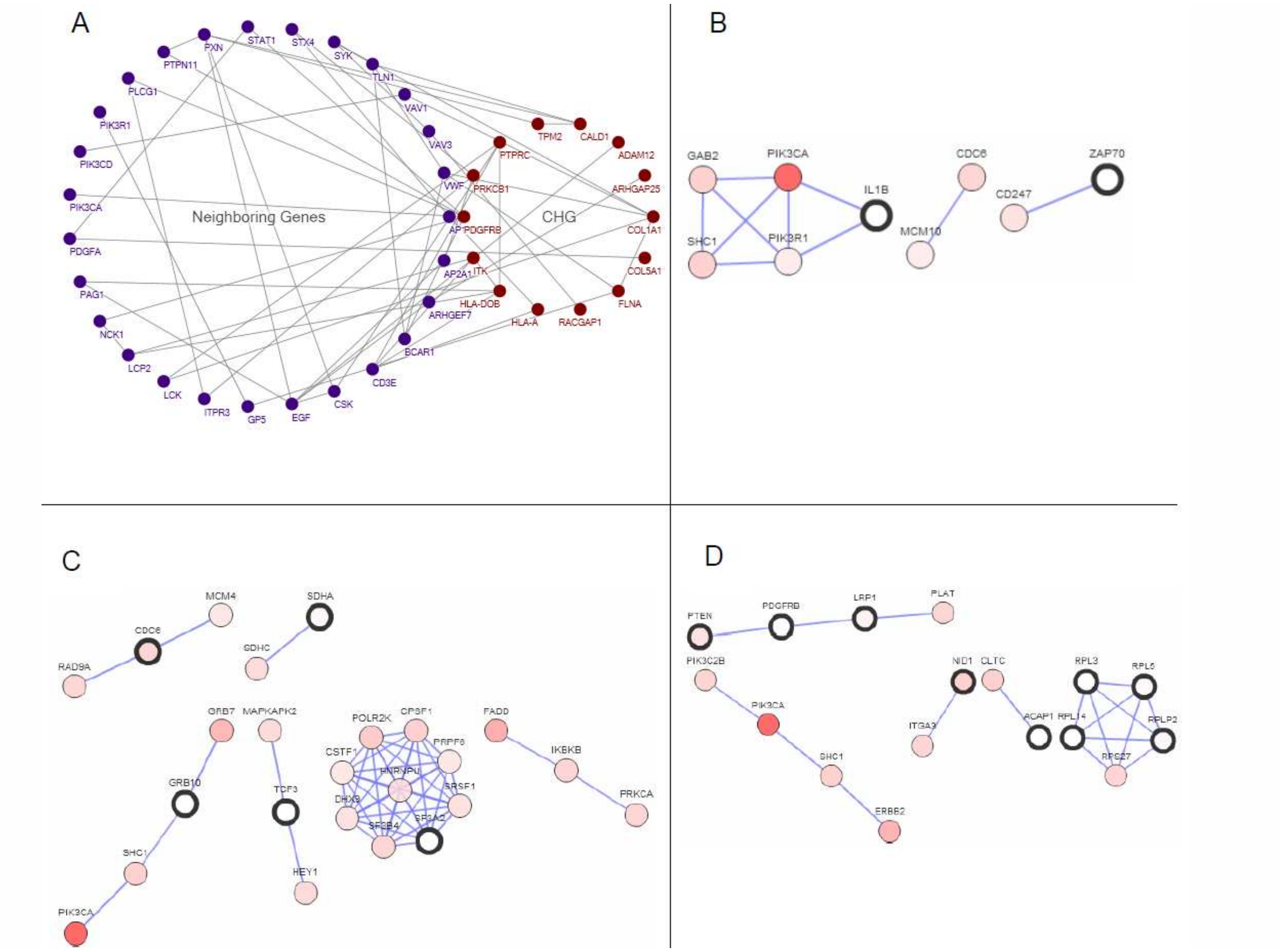


Figure 5.4: Subnetworks inferred using heterogeneity classes. A) CHG subnetwork with maximum 3 linker nodes allowed. CHG genes and added nodes are colored differently. B-D) Significant subnetworks from cross-checking with TCGA data. Black circles are genes in our study. Color of neighboring genes reflects percentage of cases being altered in TCGA breast cancer cases. B) IG subnetwork. C) CCG subnetwork. D) CHG subnetwork

5.3.5 Mapping to Interactome for network markers

We further study the role of heterogeneity genes in PPI networks to identify useful markers from the interactome data. To do this, we first map CHG, CCG and IG to PPI networks. While there are many network inference methods for this task, we used the 'connected neighbor' approach (see Methods). Here we aim to identify densely connected subnetworks from the heterogeneity gene sets. Our working assumption is that if genes exhibit different level (as quantified by co-expression matching) of inter-tumor heterogeneity, the connected portion of them in PPI networks may be the hinge to depict or represent key casual disordered interactions related to disease phenotype.

Table 5.4: Inferred Interactions from different maximal number missing genes allowed for CCG, IG and CHG separately.

No. of Inferred Interactions

	M1	**M2**	**M3**	***p*-value**
CCG	1	2	3	< 0.1
IG	-	-	6	< 0.1
CHG	2	9	57	< 0.1

We vary the maximal number of missing genes (denoted as M_i for i missing genes needed) to connect CHG, CCG and IG individually in the PPI network (Table 5.4). Only CHG class can be extracted from PPI network (Figure 5.4. A), considerably different from CCG and IG. This result suggests that metastatic genes that do not show co-expression inter-tumor heterogeneity (or most of the time being isolated) tend to be separated far away in terms of shortest paths in the PPI network and likely to act as standalone effectors in causing cancer metastasis. Interestingly, this challenges the view that cascading behavior leads to disease outcome. Table 5.4 shows that even though CCG and IG are predictive of patient survival (Figure 5.3), the primary cause is not cascading interactions but the detrimental effect in producing essential all-time proteins. We further checked the CCG, CHG and IG with TCGA data for breast invasive carcinoma and obtained significant subnetworks with

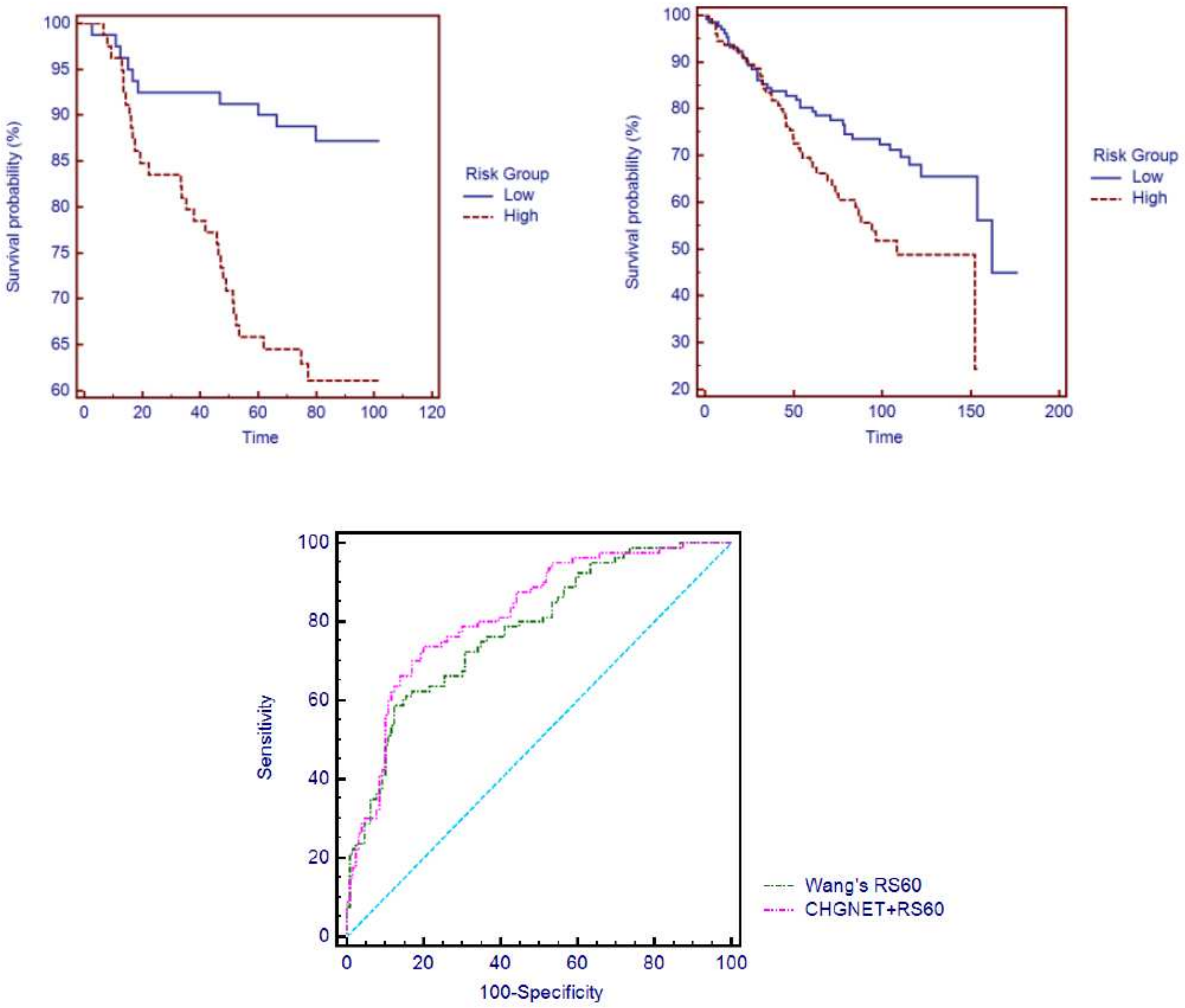


Figure 5.5: Top left : Kaplan-Meier survival curves for Pawitan cohort . Top right: Kaplan-Meier survival curves for Loi cohort. Bottom: ROC curves for CHGNET and Wang's 60 gene signature

neighboring genes being altered most frequently across tumors, as shown in Figure 5.4. B - D.

To test the predictability of markers based on the relational heterogeneity and network neighbors, we performed survival analysis on two independent patient cohorts (Figure 5.5 Top left and right). We combined low risk and intermediate risk groups into one risk group and used the same method of risk grouping as previously described. Figure 5.5 is only based on CHG and its network neighbors (Figure 5.4 A), denoted as CHGNET. Without known conserved or 'all-time' genes and regardless of biased marker selection based on GO terms (such as cell cycle and apoptosis genes), CHGNET provides reasonable performance to predict patient survival. When combined with known markers, the predictability is consistently improved. For example, we combined Wangs RS60 predictor with CHGNET and compared with the original RS60 for ER+ tumors, AUC was increased (Figure 5.5 Bottom).

5.4 Chapter Discussion and Conclusions

In a sense, we followed the working assumption (that has become increasing evident) that fast accumulating interactome data can be used to aid efficient biomarker design, and we demonstrated as a principle from inter-tumor heterogeneity perspective to target it. We speculate that there may be other biological initiatives to design biomarkers using the same framework we developed here.

A caveat should be brought up front: Although integration has become a major theme and challenge in the biomarker problem, the underlying reason should be carefully examined. In PPI networks, the relative betweenness of CHG class is small compared with an ensemble of randomly simulated networks (p-value cutoff 0.005, 500 random networks) [46]. In contrast, CCH and IG genes that are relatively more predictive of patient survival (Figure 5.3) do not show significant differences in betweenness in PPI networks. When comparing pairwise betweenness of CCG, IG and CHG with same p-value cutoff, no differences are found. These results suggest that topology of interactome data (which is subject to noise) alone is insufficient for obtaining good markers. Therefore, the integration of biological data to predict complex disease phenotype should be guided by a practical valid principle.

Chapter 6

Evaluating predictive performance of network biomarkers with network structures

[1]

We argue that it is necessary to use the network structures to evaluate performance of biomarkers. To address this, we aim to learn a weight coefficient for each node in the network from the quantitative measure such as gene expression data. The weight coefficients are computed from an optimization problem which minimizes total weighted difference between nodes in a network structure; this can be expressed in terms of graph Laplacian. After obtaining the coefficient vector for the network-based markers, we can then compute the corresponding network predictor. We demonstrate the effectiveness of the proposed method by conducting experiments using published breast cancer biomarkers with three patient cohorts. Network-based markers are firstly grouped based on GO terms related to cancer hallmarks. We compare the predictive performance of each network marker group across gene expression data sets. We also evaluate the network predictor against the average method for feature aggregations. The reported results show that predictive performance of network markers is generally not consistent across patient cohorts.

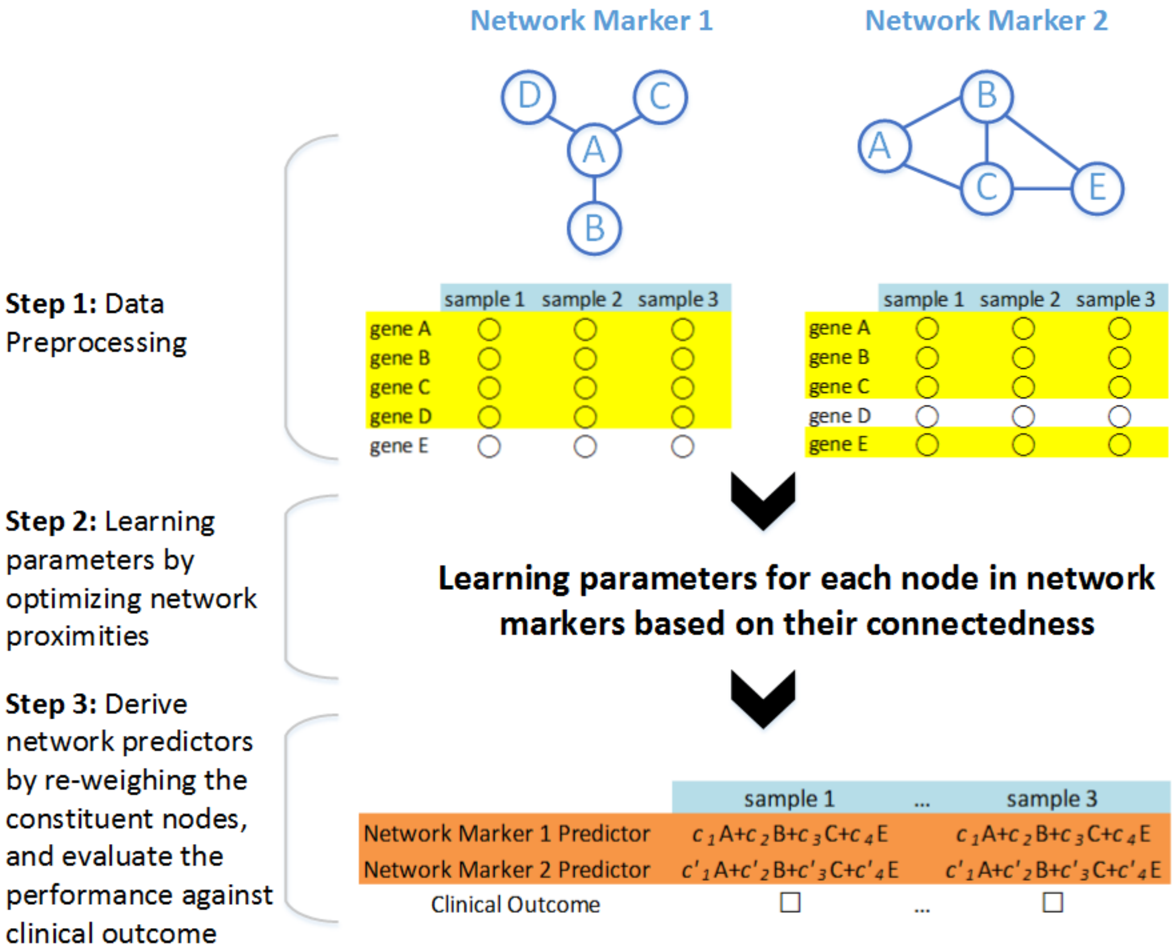
6.1 Introduction

Networks provide unique advantages in modeling multiplex relationships between genes, proteins, and diseases. In recent years, network-based approaches became promising in the disease biomarker detection problem. A notion of 'network medicine' has received consider-

[1] The content of this chapter is based on the following article under revision:
Gao, S., Afra, S., Alhajj, R., Zeng, J.,Rokne, J. & Demetrick, D. (2014) Evaluating predictive performance of network biomarkers with network structures. *In Submission.*

able spotlights centered at the principle that genes and gene products act highly interactively to cause complex diseases [11, 64, 23]. For example in cancers, identifying pair-wise disordered relationships between genes and proteins naturally fits the goals of network modeling, that is, to study relationships and find connection patterns between nodes and modules in a global map [94, 149]. To this end, many integrative methods using network data have been proposed to track differential regions of network (i.e., subnetworks) predictive of disease phenotype [94]. In this direction, a fundamental assumption is that networks reflect accurate connectedness of data. This became a critical concern in protein interaction networks due to different experimental protocols and noises [30]. For gene co-expression networks, where two genes are connected if correlated, link inaccuracies mostly arise due to data heterogeneity. Many methods are proposed to quantify the interconnectedness and the topological overlapping of networks [181, 91, 60]. Here our aim is to use nodal connectedness to evaluate network-based markers (by aggregating genes into network predictors) against certain outcome such as clinical variables. Indeed, network modules (a.k.a. subnetworks) were deemed a fundamental medium to understand and to naturally represent biological pathways and cellular processes [23, 53, 55]. For this reason, many believe that network modules could provide useful directions for finding key components attributable to disease phenotypes. Although obtaining network markers had been the major focus since van't Veer's pioneer work in the network-based thinking [164], (for example, Chuang *et al.* derived biomarkers to predict breast cancer metastasis [23]), little work have been done to evaluate the predictive power of the derived network markers, that is, to evaluate the predictability of a gene set against some phenotype given the connectedness of constitutive genes or gene products. In this chapter, we introduce a lightweight, parameter-free method for evaluating network-based markers, called Interconnectedness Network Score (INS), using clinical outcome and gene expression data.

The motivation of designing an effective method to gauge the predictive power of network



Step 1: Data Preprocessing

	sample 1	sample 2	sample 3
gene A	○	○	○
gene B	○	○	○
gene C	○	○	○
gene D	○	○	○
gene E	○	○	○

	sample 1	sample 2	sample 3
gene A	○	○	○
gene B	○	○	○
gene C	○	○	○
gene D	○	○	○
gene E	○	○	○

Step 2: Learning parameters by optimizing network proximities

Learning parameters for each node in network markers based on their connectedness

Step 3: Derive network predictors by re-weighing the constituent nodes, and evaluate the performance against clinical outcome

	sample 1	...	sample 3
Network Marker 1 Predictor	$c_1A+c_2B+c_3C+c_4E$		$c_1A+c_2B+c_3C+c_4E$
Network Marker 2 Predictor	$c'_1A+c'_2B+c'_3C+c'_4E$		$c'_1A+c'_2B+c'_3C+c'_4E$
Clinical Outcome	□	...	□

Figure 6.1: Overall workflow of INS for evaluating predictive level of network-based markers/modules

biomarkers is demanding. After extensive efforts for finding targeted disease biomarkers, for example from our previous results [44, 45], one needs to retrospectively check to see how predictive derived network markers are against clinical outcome, especially to compare the predictive results with singleton markers (i.e., individual genes that are known related to the diseases). The usual approach to aggregate the network connectedness is by averaging gene expressions of constitutive genes in a network [23, 171, 22, 97], and then using Receiver Operating Characteristic (ROC) to measure the performance of the network-based markers. This way each network module is essentially transformed to a pseudo-feature. The upside of such aggregation is that we can utilize standard ROC curves to interpret the predictive level (in this context, predictability refers to the performance of classifiers); the downside of it is that when averaging, the connectedness information of network modules is lost. For example, consider two network markers with different nodal connectedness as shown in Figure 6.1, with simple average aggregation the derived new features are indistinguishable between graph structures of two network modules (because both derived features equal to average gene expressions over nodes A, B, C and D). Therefore the question here we ask is how to derive effective features that better describe network markers/modules given their graph structures?

Here we design a method to derive module-based network features (Figure 6.1). The main idea is to learn a weight coefficient for each node in the network modules from the quantitative measure such as gene expression data. The weight coefficients are computed from an optimization problem [13, 140], since each pair of nodes connected by an edge in a network module has different strength of associations (computed as edge weights), and we are seeking a coefficient vector that preserves network connectedness. This is obtained by minimizing the total weighted difference between coefficients associated with nodes (Figure 6.1, Step 2), which can be written in terms of graph Laplacian (see Material and Methods). After obtaining the coefficient vector for the network marker, we can then compute the corresponding network predictor (Figure 6.1, Step 3). The method effectively takes network proximity into consideration, therefore the derived network predictors are more reliable for plotting ROC curves.

To demonstrate the method, here we evaluate published breast cancer biomarkers with four patient cohorts, network markers are firstly grouped based on GO terms related to cancer hallmarks. We compare the predictive performance of each network marker group across gene expression data sets. We also evaluate the network predictor against the average method for the feature aggregation aforementioned.

6.2 Material and Methods

6.2.1 Learning network coefficients

Input: Suppose we have a collection of m networks that are indicative of certain cellular phenotype, $\Theta = \{\mathbf{A}_1, \mathbf{A}_1, ..., \mathbf{A}_m\}$, where $\mathbf{A}_p, 1 \leq p \leq m$ is the adjacency matrix of network p in Θ with $\mathbf{A_p} := [a]_{ij} = 1$ if node i and j are connected, 0 otherwise. We have a gene expression data set $\mathbf{R} := [r]_{gs}$ for which we want to evaluate the predictive power of network markers in Θ, where $[r]_{gs}$ is the expression value of gene g in sample s, and denote the clinical variable (e.g., metastasis outcome) as $\mathbf{o} = (o_1, o_1, ..., o_{|s|})$ where $|s|$ denotes the number of samples in $\mathbf{R}$.

Output: Let the coefficient vector for network $\mathbf{A_p}$ with k nodes be $\mathbf{c} = (c_1, c_2, ..., c_k)$. Our goal is to derive $\mathbf{c}$ for $\mathbf{A_p}$ whose pair-wise magnitude preserves the neighborhood connectivity of $\mathbf{A_p}$.

To preserve the local connectivity by coefficient vector $\mathbf{c}$, the problem reduces to minimize

$$\sum\nolimits_{ij} (c_i - c_j)^2 w_{ij} \tag{6.1}$$

where w_{ij} is the weight between node i and j in $\mathbf{A_p}$. $\mathbf{W} := [w]_{ij}$ refers to the weighted adjacency matrix for $\mathbf{A_p}$ [13]; w_{ij} represents the weight (a similarity measure) between gene

i and j if connected. We used the heat kernel to compute w_{ij}. Putting Eq. 6.1 in matrix form (see ref. [13] for details),

$$\sum\nolimits_{ij} (c_i - c_j)^2 w_{ij} = 2\mathbf{c^T}(\mathbf{D} - \mathbf{W})\mathbf{c} = 2\mathbf{c^T}\mathbf{L}\mathbf{c} \tag{6.2}$$

Where D is the diagonal matrix with $\mathbf{D} := [d]_{ii} = \sum_j w_{ij}$, $\mathbf{L}$ is the graph Laplacian $\mathbf{L} := \mathbf{D} - \mathbf{W}$. The problem is then reduced to finding:

$$\underset{\mathbf{c^T D c = 1}}{\arg\min}\, \mathbf{c^T L c} \tag{6.3}$$

The constraint $\mathbf{c^T D c = 1}$ removes the arbitrary scaling factor to the solution, which is given by the second smallest eigenvector of the generalized eigenvalue problem:

$$\mathbf{Lc} = \lambda \mathbf{Dc}$$

Coefficient vector c represents the relative importance of nodes due to the network topology measured by w_{ij} that is, if two nodes are far apart in the network, w_{ij} incurs a heavy penalty from Eq. 6.1 After solving for c, we obtained the coefficient vector which is subsequently used to weigh gene expression levels of constituent genes. After re-weighing (Step 3 of Figure 6.1), we obtained corresponding network predictors.

6.2.2 Breast cancer biomarkers

Breast cancer biomarkers were retrieved from the Cell Circuits database (`http://www.cellcircuits.org`) [99]. We search Gene Ontology terms (p-value < 0.001) related to cancer hallmarks from Chuang *et al.*'s work [23, 55] , and collect network markers for each GO group [93]. Totally we obtained 62 network-based biomarkers from 7 GO groups (Table 6.1). Gene symbols are mapped using UniProt ID Mapping (`http://www.uniprot.org/`) (The UniProt Consortium) and DAVID (`http://david.abcc.ncifcrf.gov/`) [61].

Table 6.1: GO terms searched and number of network markers obtained

GO Term	**GO ID**	**Number of Network Modules (gene groups)** ($p < 0.001$)
Apoptosis	GO:0006915	9
Cell adhesion	GO:0007155	4
Cell cycle	GO:0007049	28
Immune response	GO:0006955	3
Phosphorylation	GO:0016310	8
Response to external stimulus	GO:0009605	7
Cell growth	GO:0016049	3

6.2.3 Gene expression data preprocessing and normalization

Gene expression data sets were retrieved from NCBI GEO database (`http://www.ncbi.nlm.nih.gov/geo`) with the accession ID GSE2034 ($n = 286$) [170], GSE1456 ($n = 159$) [128] and GSE6532 ($n = 327$) [96]. All three data sets use HG-U113A platform, we did so in order to avoid bias in cross-platform validation. Gene expression data were processed with MAS5.0 algorithm, and subsequently log2 transformed and median-centered across samples.

6.3 Results

6.3.1 Nodal connectedness affects predictive performance

We used DREAM5 (Dialogue for Reverse Engineering Assessments and Methods) gene expression data sets, described in detail in [100]. The input data includes a compendium of 805 microarray experiments for E.coli, consisting of 4511 genes (including 214 decoy genes). To see if the nodal connectedness affects the predictive performance against genetic perturbations, we used the gold standard benchmark provided by the DREAM5 challenge. The benchmark data includes experimentally validated 2066 transcriptional interactions retrieved from RegulonDB. We created two sets of network modules: the first set includes network modules with at least one hub gene and its immediate neighbors (hub set). The hub gene

Table 6.2: Predictive performance of two different set of network modules

	Average AUC	**95% Confidence Interval**
Hub Set	0.67	0.614 - 0.701
Random Set 1	0.43	0.393 - 0.472
Random Set 2	0.51	0.481 - 0.545
Random Set 3	0.41	0.382 - 0.455
Random Set 4	0.54	0.481 - 0.575
Random Set 5	0.53	0.480 - 0.561
Random Set 6	0.47	0.422 - 0.511
Random Set 7	0.39	0.347 - 0.426
Random Set 8	0.43	0.393 - 0.472
Random Set 9	0.49	0.455 - 0.531
Random Set 10	0.37	0.362 - 0.445

is identified as nodes with degrees greater than average node degree of benchmark network plus 2 standard deviations of the total node degree distribution; the second set includes randomly selected network modules without any hub genes (random set). We collected 26 network modules from the first set with average network size 6 and we randomly selected the same number of nodes to form the second set. If a network module from the random set has size less than 6 we randomly add neighbors from one of the constituent genes. We compare the ROC curves for these two sets of network modules (Table 6.2). Hub set modules have higher average AUC than random set modules, which indicates that network topology affects the predictive performance.

6.3.2 Retro-perspective validation using Wang's data

We use the retrieved network markers to predict metastatic and non-metastatic samples in Wang's cohort where the network markers were derived from [23]. We tested the predictive performance over the entire range of sensitivity and specificity values of network markers against Wang's data set, and compared the AUC with average aggregation. In apoptosis (Figure 6.2), cell growth, immune response, and response to external stimulus groups, our method reports better performance over the average method (6 out of 9, 2 out of 3, 3

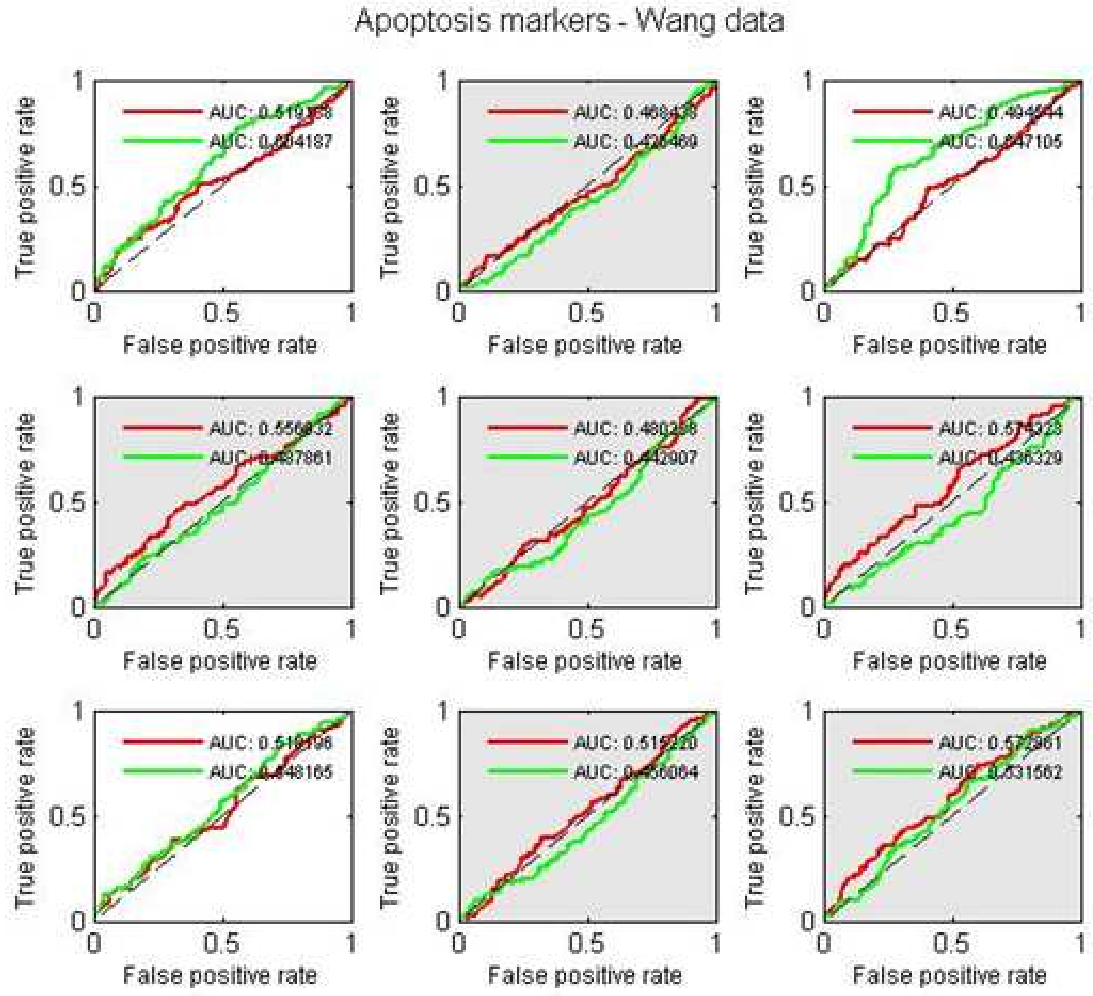


Figure 6.2: ROC curves for the apoptosis GO term against Wang's data, 6 out of 9 markers show better predictive performance with our method (in grey), other groups show similar trend except Cell Cycle group

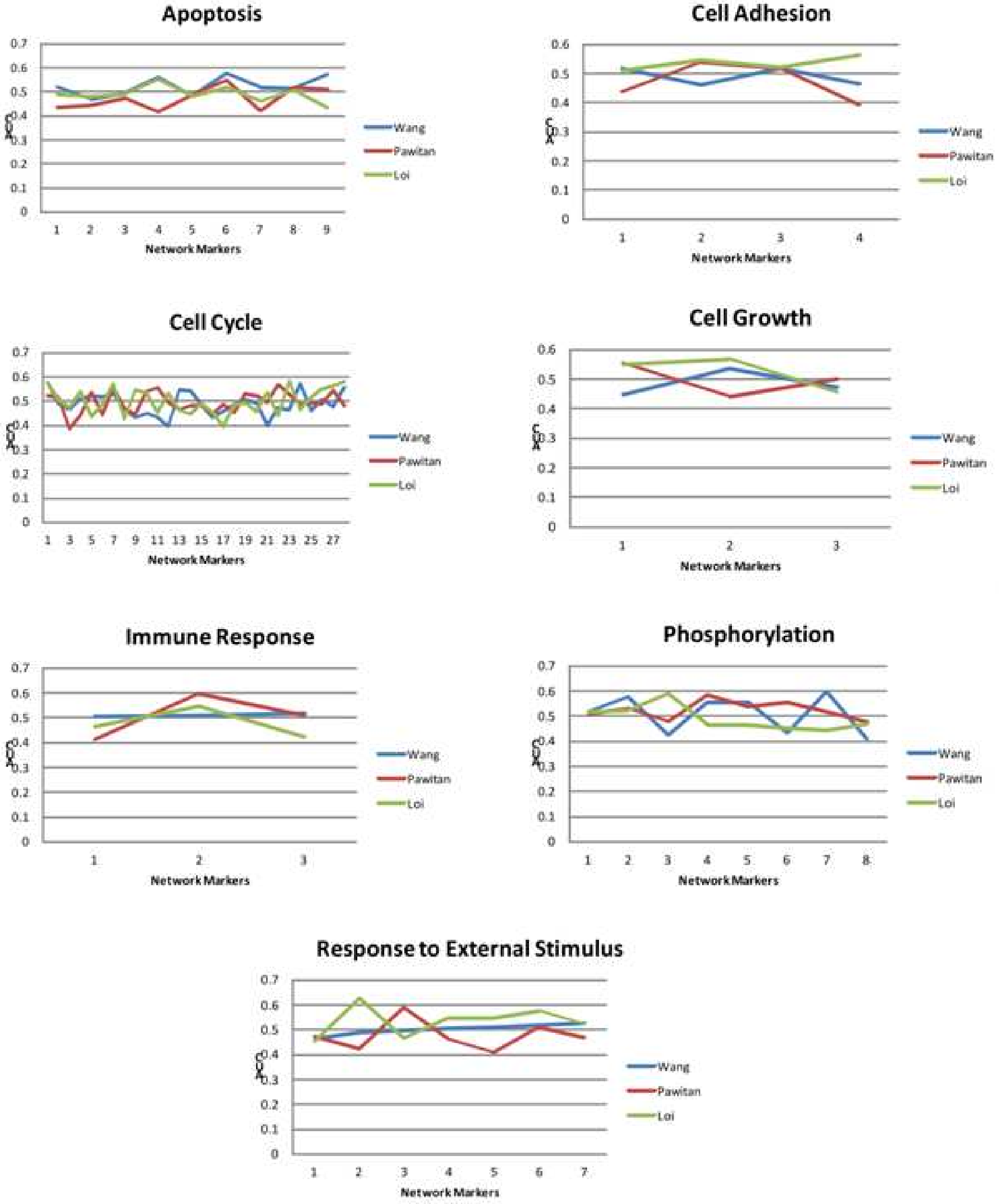


Figure 6.3: Area Under the Curve (AUC) for 62 network markers in three different patient cohorts

out of 3, 7 out of 7 network markers with higher AUC). Other GO groups (cell adhesion, phosphorylation) show similar performance for both methods (see supplementary material). This suggests that by taking the network connectivity into account, predictive performance can be improved in classifying breast cancer metastasis.

Interestingly, in the cell cycle group only 1 out of 28 markers shows higher AUC with our method. The implication is that edge connectedness in the network markers are not predictive of metastasis in general for this group of markers, because totally ignoring it (using average aggregation) leads to better classification performance.

For apoptosis makers shown in Figure 6.2, 95% CIs show moderate overlapping between INS (average upper bound and lower bound are 0.463 and 0.561 respectively, binomial exact test) and simple average method (average upper bound and lower bound are 0.423 and 0.488 respectively, binomial exact test) for network-based markers with better performance using INS. Similar effects are observed in other GO groups. From the above study with E. coli data, the INS method shows consistent better performance. It is worth noting that our aim is not to propose a method that produces better predictive performance using existing network-based markers, as the way of identifying network-based markers differs, the predictive results differ; this is the similarly true when evaluating gene signatures: there is a big pool of gene signatures but very few of them produce consistent predictive performance [93]. For example using Chuang's data we observed that 6 out of 9 markers show better performance for the apoptosis group, this is likely due to the inconsistent predictive performance for individual network-based markers from Chuang's data. In fact, from Chuang's method, the network-based markers are derived from a greedy approach, which does not find the optimal solution in general.

6.3.3 Cross validation with other gene expression data sets

To cross check the predictive performance of network biomarkers, we compare the ROC curves for each GO group with Loi and Pawitan's cohort [128, 96] using network predictors.

From Figure 6.3, we did not observe unique trend for all GO groups. For example, in the Apoptosis group 6 out of 9 markers are more predictive in Wang's cohorts, whereas in the Cell Cycle group Wang's cohort are not as predictive as the other two cohorts (6 out of 28 markers have higher AUC compared with Loi and Pawitan's cohort). The results show that network biomarkers are not consistently predictive across patient cohorts, facing the same dilemma of the gene set approach where most gene sets are not robust in predicting cellular phenotypes [93].

6.4 Chapter Discussion and Conclusions

Although network approaches have become promising in predicting disease phenotypes like breast cancer recurrence, the way of making them predictive is problematic. On one hand, most work have focused on obtaining the network modules and the argument that those are more predictive than the gene set approach; on the other hand, when evaluating the predictive performance of network markers, simple aggregation is employed and therefore network connectedness is totally ignored. Here we offer a simple approach in taking network connectedness into account when evaluating network biomarkers against clinical variables. Using this method, we showed that the network markers are not consistently predictive when compared with the simple aggregation approach. The crucial problem is that methods that identify network markers generally do not include network connectedness as the factor when "scoring" subnetworks. Similar to most of the gene signatures, network markers do not show robust predictive performance across gene expression profiles in different GO groups, making them non-robust when predicting breast cancer metastasis.

With the aforementioned, the conclusion is not solely due to absence of nodal connectedness in evaluating network modules, but rather in identifying truly predictive network-based markers. Our focus in this chapter is to evaluate network-based markers when considering network topology. Our observations that network-based markers are not consistently predictive across patient cohorts reflect the fact that network-based markers are far from being predictive in clinical studies.

Chapter 7

Multiplex network reconstruction: integrating multiple sources of genomic data to predict network-based knowledge

We present a method to reverse engineer integrative gene networks. The main advantage of our method is the integration of different quantitative and qualitative datasets in order to reconstruct a multiplex network, without imposing data constraints, such as each genomic datum needs to have the same number of entities. The computation boils down to solving small quadratic programs based on local neighborhood of nodes. Another advantage of our method is that from the integrated networks, predictions can be made by propagating beliefs from seed nodes representing known knowledge via weighted edges. Thus, we combined data integration and network-based prediction into a single framework. We applied the method to DREAM5 dataset, and compared the results with the community networks from the challenge. Further, we demonstrate our method through case studies using breast cancer data, including the integration of metastasis gene expression data with interactome data and biological pathway data. Network-based predictions are compared between interactome-integrated and pathway-integrated networks. Overall, our method has the potential to be applied in many settings of network system biology.

7.1 Introducction

As networks are becoming ubiquitous in modeling complex biological systems, we are still facing two major challenges underlying (computational) system biology: the integration of various types of genomic data and the prediction of unknown knowledge using networks

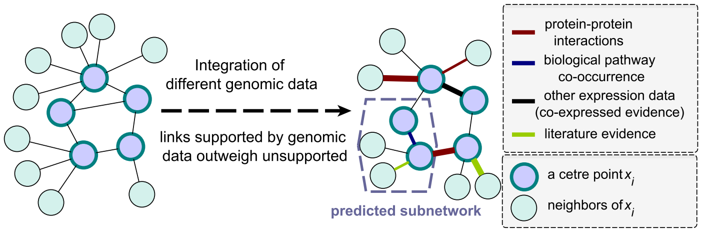

Integration of
different genomic data
links supported by genomic
data outweigh unsupported
predicted subnetwork
protein-protein interactions
biological pathway co-occurrence
other expression data (co-expressed evidence)
literature evidence
a cetre point x_i
neighbors of x_i

Figure 7.1: Integration of multiple genomic data sets to reconstruct gene networks and network-based predictions to extract relevant subnetworks by graph learning

[169, 64, 11]. These two problems are closely related: in order to make accurate network-based predictions, one must integrate different sets of evidence from rapidly accumulating biological data and knowledge. For example, in the network biomarker detection problem, the aim is to find compact network modules predictable to disease outcome or cellular phenotypes [64, 11, 23, 21]. To achieve this goal, a number of genomic information can be useful, including sequencing data, gene expression profiles, known protein-protein interactions and gene regulations etc., and non-genomic evidence like literature-based data mining for unraveling hidden regulatory relationships between genes and their products. Such intertwined problem calls for a robust and integrative method to reconstruct cellular networks that are able to reflect different types of evidence between nodes. We refer to a network with multiple link types between nodes as the 'multiplex network' [113], which is very natural in system biology settings to model complex data.

Here we propose a flexible method to reconstruct multiplex networks. We first categorize the pertaining data sets into two types: quantitative and qualitative data, where the former includes gene expression profiling (RNA-Seq data or Affymetrix Genome-wide Arrays for instance) and the latter contains pathway and biological process data, protein-protein interactions, GO term similarities, etc. Then we construct the local neighborhood of each node (i.e., the 'center node') by minimizing the reconstruction error in terms of distance or similarity measures from its nearest neighbors inferred from quantitative data [144, 168]. The integration of other data sets, either quantitative or qualitative, is done by setting simple weight constraints on the links against the objective of minimizing reconstruction errors (Figure 7.1). This way, we solve a small quadratic programming problem (as the objective is quadratic) for constructing a small subnetwork corresponding to a central node and its nearest neighbors [144, 140]. By posting supporting evidences from different data sources as constraints, the optimization problem effectively eliminates links established solely based on a single data set, e.g., false positive co-expressed genes lacking biological meaning [89], hence accentuates the links only supported by evidence. After all center nodes are considered, subnetworks can be combined into an integrative single network.

There are three main advantages of our proposed multiplex network reconstruction. Firstly, the process is flexible with different data types. We view data sources as different types of evidences, and convert it to the constraint in the optimization problem by simply stating that supported links outweigh unsupported links. This way, we do not impose that evidential data sets have exactly the same set of nodes (such as genes in the context of constructing a regulatory network), as assumed by others [86, 112]. Secondly, the integration of constraints can be contextualized. For example, if we were to integrate gene expression profiles with protein-protein interaction (PPI) data, one could use protein interactions observed in PPI data to support co-expressed links, or one could relax the supporting strength by allowing non-direct neighbors if shortest path between co-expressed gene products falls below certain threshold in the PPI network. Thirdly, the local neighborhood assembly allows for efficient learning and prediction on the network. This is because each central node is a linear combination of its neighbors, and each individual neighbor in turn links to its own neighbors in the same way and so forth; therefore information is relied in the network (that is also sparse). Such reconstruction enables network-based predictions by propagating partially known information of nodes to the whole network. This is known as semi-supervised learning [191].

To demonstrate the applicability and effectiveness of the proposed methodology, here we first apply the network reconstruction algorithm to DREAM5 dataset, and further study breast cancer metastasis using the integrative reconstruction method through three case studies:

1. Integrating interactome data and gene expression profiles.
2. Analyzing integrated pathway co-occurrence networks and identifying differential genes between metastatic and non-metastatic networks.
3. Comparing network-based predictions made from integrated networks (i.e., pathway-integrated and interactome-integrated) using seed nodes and extracting relevant subnetworks with different data sources.

7.1.1 Related Work

Recent existing works for integrative network construction can be divided into two categories. 1.) In statistical-based approaches, Lo *et al.* modeled the external knowledge in the form of prior distributions and then applied supervised framework to train and calibrate the distribution [95]. This way, a directed acyclic graph (DAG) is constructed for gene regulatory network. Haibe-Kains *et al.* integrated literature data from PubMed with expression profiles and interpreted gene interactions using MeSH terms and Gene Ontology (GO) [53]. To infer a predictive network, seeded Bayesian inference with similar prior modeling is employed [33]. For charting interaction maps with a large number of genes, regression-based technique was used to improve the performance of network construction [53, 33]. As the authors pointed out, integration with prior information suffers from large networks and is computationally expensive [53]; on the other hand, in regression-based methods it is difficult to determine the number of candidate regulators for each gene. In addition, some regularization [180, 186] or filtering techniques [62, 65, 114, 148] are often needed . 2.) In kernel-based approaches, different types of genomic data sets are represented in the form of networks resulting from kernel computations, and the pool of different networks are combined to get an integrative network. For example, GeneMANIA analyzes gene lists using genomics and proteomics evidence in network settings and computes weights for each data source for their relative importance [173]. The computation framework is described in [112], which involves a regression model to learn weights of different networks by minimizing the least square error between the target network and the composite network. In general, combining networks can be modeled as a multiple kernel learning problem [105] with each kernel encoding relationships between data points [86, 160]. The computational challenge of these methods is to solve optimization problems like Semi-Definite Programming (SDP) that are often costly [105]. Another constraint of the kernel-based method is that it requires the same set of nodes in each network [86], and genomic data must be represented with kernels in the first stage. Such limitation makes the integration of some genomic evidence difficult (if not impossible), because not all genomic data can be kernelized (e.g., the similarity between genes in different PubMed abstracts is hard to define and interpret) and are of the same dimension.

7.2 Material & Methods

7.2.1 Local Neighborhood Construction

We introduce the network reconstruction method based on local neighborhood of individual nodes as follows. Given a quantitative data set consisting of n data points $\{x_1, x_2, \ldots \ldots \ldots, x_n\}$, we aim to find the reconstruction weight $w_{i.}$ of each node i by minimizing the cost function [144]:

$$\varepsilon(w_{ij}) = \left\| x_i - \sum_{j:x_j \in \eta(x_i)} w_{ij} x_j \right\|^2 \tag{7.1}$$

Where x_i is a vector containing quantitative measure of a data point (e.g., gene expression across samples); $\eta(x_i)$ refers to the neighbors of gauged by some metric (e.g., mutual information); $||.||$ denotes the Euclidean norm. Assuming that data points reside on a linked structure (i.e., manifold), the cluster assumption states two data points connected by paths that pass through high-density region are likely to be in the same class. The reconstruction error $\varepsilon(w_{ij})$ seeks optimal weight assignments w_{ij} between data point i and its nearest neighbors in the structure. Additional constraints to the objective function of Eq.7.1 include $\sum_{j:x_j\in\eta(x_i)} w_{ij} = 1$ and $w_{ij} \geq 0$ [144, 168, 140].

To integrate other data sets, the idea is to give prominence to links that are advocated by other qualitative evidence in terms of edge weights. Formally, we impose an additional constraint for w_{ij} as follows.

$$\sum_d \sum_{j:x_j\in\eta(x_i)} w_{ij}^{s} \geq \sqrt{k} \sum_d \sum_{k:x_k\in\eta(x_i)} w_{ik}^{\sim s}, k \geq 1 \tag{7.2}$$

Where w_{ij}^{s} refers to links between node i and j that are supported by qualitative datum d; $w_{ik}^{\sim s}$ refers to links that are not supported by d; $\sqrt{k}$ is a scaling factor. By Eq. 7.2, we are forcing data points to reflect different types of evidence in the underlying network by edge weights. Put together the optimization problem is formulated as follows.

$$\begin{aligned} \text{minimize} \quad & \varepsilon(w_{ij}) = \left\| x_i - \sum_{j:x_j\in\eta(x_i)} w_{ij} x_j \right\|^2 \\ \text{subject to} \quad & \sum_{j:x_j\in\eta(x_i)} w_{ij} = 1,\ w_{ij} \geqslant 0 \\ & \sum_d \sum_{j:x_j\in\eta(x_i)} w_{ij}^{s} \geqslant \sqrt{k} \sum_d \sum_{k:x_k\in\eta(x_i)} w_{ik}^{\sim s} \\ & k \geqslant 1 \end{aligned}$$

It is worth noting that in general $w_{ij} \neq w_{ji}$, i.e., the reconstructed network $W = [w_{ij}]$ is not symmetric, because nodes i and j are not necessarily mutual nearest neighbors to each other. To make W symmetric, one can simply replace W by $\frac{W+W^T}{2}$.

7.2.2 Quadratic Programming, Fast Assembly and Integration

The above optimization problem can be written as a quadratic program as follows.

$$\begin{aligned}\varepsilon(w_{ij}) &= \left\| x_i - \sum\nolimits_{j:x_j\in\eta(x_i)} w_{ij}x_j \right\|^2 \\ &= \left\| \sum\nolimits_{j:x_j\in\eta(x_i)} w_{ij}(x_i - x_j) \right\|^2 \\ &= \sum\nolimits_{j,k:x_j,x_k\in\eta(x_i)} w_{ij}w_{ik}(x_i - x_j)^T(x_i - x_k) \\ &= \sum\nolimits_{jk} w_{ij}w_{ik}G_{jk} \qquad (7.3)\end{aligned}$$

Where in Eq. 7.3, we used the fact $\sum_{j:x_j\in\eta(x_i)} w_{ij} = 1$; and G_{jk} is the local (neighborhood) gram matrix defined as $G_{jk} = \langle x_i - x_j, x_i - x_k\rangle$, $\langle.\rangle$ denotes inner product. Thus the reconstruction weights can be written in quadratic programming formulation.

$$\begin{aligned}\text{minimize} \quad & \varepsilon(w_{ij}) = \sum\nolimits_{j,k:x_j,x_k\in\eta(x_i)} w_{ij}G_{jk}w_{ik} \\ \text{subject to} \quad & \sum\nolimits_{j:x_j\in\eta(x_i)} w_{ij} = 1, w_{ij} \geq 0 \\ & \sum\nolimits_d \sum\nolimits_{j:x_j\in\eta(x_i)} w_{ij}^s \geq \sqrt{k} \sum\nolimits_d \sum\nolimits_{k:x_k\in\eta(x_i)} w_{ik}^{\sim s} \\ & k \geqslant 1\end{aligned}$$

Each data point x_i incurs a quadratic program, and so we need to solve n small quadratic programming problems, because $\eta(x_i)$ is relatively small compared with n. Another benefit of such local reconstruction is that each quadratic programming problem can be solved independently of others, thus parallel computation methods can be used to efficiently reconstruct the network. Such computational strategy is a crucial factor for reverse engineering of gene networks in practice [2]. In our experiments, we used MATLAB Parallel Computing Toolbox® (`http://www.mathworks.com`). It is worth mentioning that our focus is different from Aluru *et al.*'s work where they emphasized the parallel computation of the metric in terms of mutual information, while we focused on the parallel assembly process of local neighborhoods and the otherwise costly integration process.

7.2.3 Data Preprocessing and Normalization

Breast cancer data sets were retrieved from NCBI GEO database (`http://www.ncbi.nlm.nih.gov/geo`) with the accession ID GSE2034 [170]. Gene expression data were processed with MAS5.0 algorithm, and subsequently log2 transformed and median-centered across samples.

To gather metastasis related genes, we performed genome-wide survival screening by univariate Cox regression (p-value < 0.05) on gene expression data with metastatic and non-metastatic patients [23, 93]. For genes known related to breast cancer, we collected 324 genes from Network of Cancer Genes 3.0 (NCG) database [29]. Differentially expressed genes between metastatic and non-metastatic samples were gathered by using two-sample t-test (p-value < 0.05). We eliminated redundant genes from different collections for analysis.

7.2.4 DREAM5 Dataset

We used the well-established **D**ialogue for **R**everse **E**ngineering **A**ssessments and **M**ethods (DREAM5) dataset to evaluation our method [100] (`http://wiki.c2b2.columbia.edu/dream/index.php/D5c4`). The challenge is designed to reverse-engineer complete transcriptional regulatory networks from gene expression data. The dataset consists of four networks: *E. coli*, *S. cerevisiae*, *in silico*, and *S. aureus*, and the first three are used for evaluation [100]. To compare the results, we used the same Precision-Recall (PR) statistic and gold standards (i.e., true regulatory edges) to evaluate the predictive performance as describe in the original publication.

7.2.5 Biological Pathway and Protein Interaction Data

We collected biological pathway and protein-protein interaction data from multiple databases using ConsensusPathDB [72] (`http://cpdb.molgen.mpg.de`). ConsensusPathDB integrates a wide range of network and complex interaction data for *Homo sapiens*, including 31 data sources and literature curated interactions. For the interactome data, we mapped gene symbols to Entrez Gene IDs to avoid naming ambiguity using Gene ID Conversion tool in DAVID (`http://david.abcc.ncifcrf.gov`) [61], and we excluded the self-interactions.

7.2.6 Distance Metric for Computing Nearest Neighbors

One remaining question to find the reconstruction weight is to choose a distance metric for computing the neighborhood of data points, $\eta(x_i), i \in [1, n]$. To measure the strength of association between two points or variables, many metrics can be used. Popular choices are, among others, Pearson's correlation coefficient [89, 37, 1] and information-theoretic method like mutual information [2, 132, 101]. Although simple and fast to compute, Pearson's correlation fails to identify non-linear associations between variables, therefore leading to inaccurate reconstruction of gene regulatory networks; mutual information on the other hand, captures non-linear associations but is computational expensive. Another computational overhead arises due to the fact that mutual information is not able to distinguish direct interactions from indirect ones subject to Data Processing Inequality (DPI), resulting in a separate computational overhead to filter out indirect links [2, 101].

Here we wish to use a distance metric to leverage the need of identifying non-linear associations and reducing the DPI computational overhead between data points. For this purpose, we chose the mutual information based measure and further make it a distance metric, which means that it has to satisfy the triangle inequality, non-negativity, symmetry, and indiscernability criteria. To this end, the mutual information-based distance between data point x_i and x_j, $d(x_i, x_j)$ is defined as:

$$d(x_i, x_j) = H(x_i, x_j) - I(x_i; x_j) = H(x_i|x_j) + H(x_j|x_i) \tag{7.4}$$

Where $H(x_i, x_j)$ is the joint entropy, $I(x_i; x_j)$ is the mutual information between x_i and x_j; $H(x_i|x_j)$ is the conditional entropy of x_i given x_j. $d(x_i, x_j)$ is a distance metric satisfying metric properties, for example triangle inequality, i.e., $d(x_i, x_k) \leq d(x_i, x_j) + d(x_j, x_k)$ for another point x_k . This effectively reduces the cost of post-processing DPIs.

7.2.7 Computing $d(x_i, x_j)$ by Freedman-Diaconis Binning

Denote $x_{i,s}$ as the measure of data point i in sample s and number of samples $|s|$. Since quantitative data sets are often continuous-valued, estimation of $x_{i,s}$ is difficult for computing $I(x_i; x_j)$. To address this, we apply the homeomorphic transformations [2, 83] by binning the vector $x_i = \left\langle x_{i,1}, x_{i,2}, ..., x_{i,|s|} \right\rangle$ using Freedman-Diaconis method [43]. The bin width h of Freedman-Diaconis method is calculated as $h = 2\text{IQR}(x_i) \big/ |s|^{1/3}$, where IQR refers to interquartile range of vector x_i. The number of bins is then computed as $n_{bin} = \lceil \max x_i - \min x_i \rceil / h$, where $\lceil . \rceil$ is the ceiling function. After binning, x_i becomes $x_i^{bin} = \langle x_{i,1}, x_{i,2}, ..., x_{i,k} \rangle$. By plotting data points on the manifold using the distance metric (Eq. 7.4), the reconstruction process effectively meets the cluster assumption, i.e., local neighborhoods are patched with important interactions (those observed from other data sources and whose endpoints are quantitatively close).

7.2.8 Prediction based on Reconstructed Network

After network reconstruction, the next task is to predict relevant network regions or subnetworks given certain partial knowledge. Such setting is not uncommon in biological problems like predicting novel disease-causing genes from regulatory networks, because the knowledge of disease related markers are being renewed rapidly. We cast the network prediction into the graph-based semi-supervised learning framework [191, 189, 14], which entails the following question: Given a small number of known labels (hence 'semi-supervised') of nodes (i.e., genes known related to disease) what can we infer about other nodes in the network map (hence 'graph-based')?

Formally, given a set of data points $\chi = \{x_1, ..., x_l, x_{l+1}, ..., x_n\}$, where first l points, i.e., $x_1, ..., x_l$ points are labeled as $y_1, ..., y_l$ and the rest $(x_{l+1}, ..., x_n)$ are unlabeled, the aim is to score the unlabeled points such that higher score implies higher relevance to the labeled points in the network. The objective function $Q(f)$ to be minimized consists of two terms:

$$Q(f) = \sum\nolimits_{i=1}^{l} \|y_i - f_i\|^2 + \lambda f^T L f \tag{7.5}$$

The first term is the loss function and measures how well do points vary from known labels (i.e., consistency), if $x_i \in \{x_{l+1}, ..., x_n\}$ then the corresponding label is 0; the second term measures the smoothness, which means that labels of nearby points in the network should not vary much weighted by regularization parameter λ . L is the graph Lapacian with $L = D - W$, where D is the diagonal matrix with node degrees on the diagonal and W is the matrix containing reconstructed weights from quadratic programming, f^T denotes the transpose of vector f. Eq. 7.5 is a regularized least square problem and can be solved efficiently with analytical solutions [189, 137]. By minimizing $Q(f)$ we compute predicted scores in vector f. It is worth noting that such learning mechanism relies on the underlying network structure: in the way of computing W each node is a linear combination of its neighbors, which enables the propagation of predicted scores and regularization of smoothness on the graph. In our experiment we set $\lambda = 0.1$.

7.2.9 Complexity Analysis

The mutual information computation runtime (including the binning process) is bounded by O($n\,|s|$) where $|s|$ is the number of samples and n is the number of nodes. The cost for finding the reconstruction weights is $O(|s|\,nk^3)$ [144], where k is the number of nearest neighbors. n and $|s|$ are denoted the same as the above. The integration of data sources is linear in terms of their cardinalities (e.g., number of interactions in PPI data).

7.3 DREAM5 Experiments

7.3.1 Effect Of Number Of Nearest Neighbors k

In our integrative network reconstruction model, the only required parameter is the number of nearest neighbors k. We here investigate the choice of k by using networks of three organisms

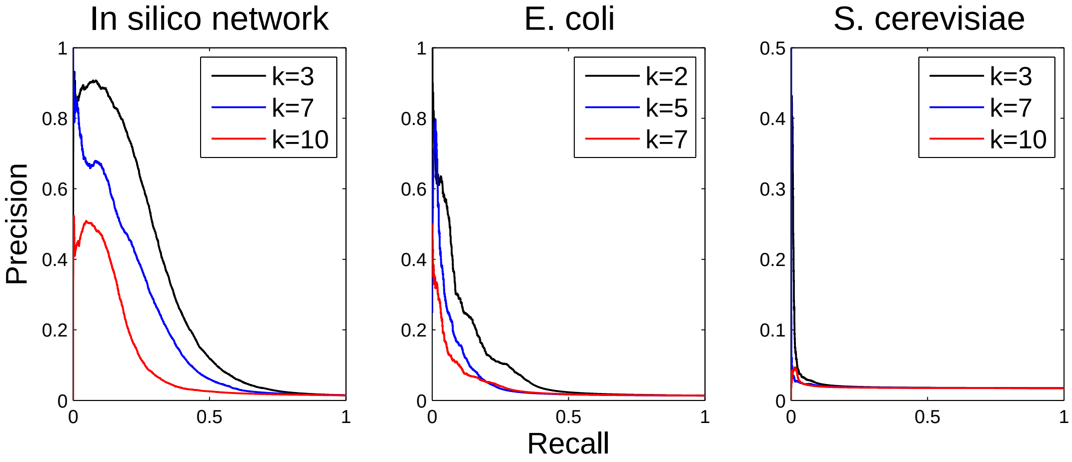


Figure 7.2: Precision-Recall curves for three networks with different values of k

from DREAM5 dataset. Intuitively, the choice of k depends on the underlying regulatory relationship between genes in "true" network structures, which by far is not accurately known. We use the provided gold standards of three organisms (*E. coli*, *S. cerevisiae*, and *in silico* networks) from the challenge to evaluate the effect of k in network reconstruction.

We interrogate the Precision-Recall (PR) curves for the three networks for a various settings of k, shown in Figure 7.2. We observed that while increasing k improves the Area Under the Precision-Recall (AUPR) curve to a certain level, the margin of such improvement decreases. This suggests that on average the number of regulators for the network (i.e., the average node degree in gene regulatory networks) controlled by k is likely to approximate the unknown "true" network by an upper bound, characterized by complex cellular regulatory mechanisms of different species.

7.3.2 The Effect Of Integration And Comparison With Community Networks

To demonstrate how the network reconstruction method effectively improves the prediction of regulatory networks, we first construct the neighborhood networks as based networks for each microorganism from mutual information based distance described previously, and then choose the best performing inference algorithms from each category (regression, mutual information, correlation, Bayesian networks, miscellaneous methods, and Meta predictors,

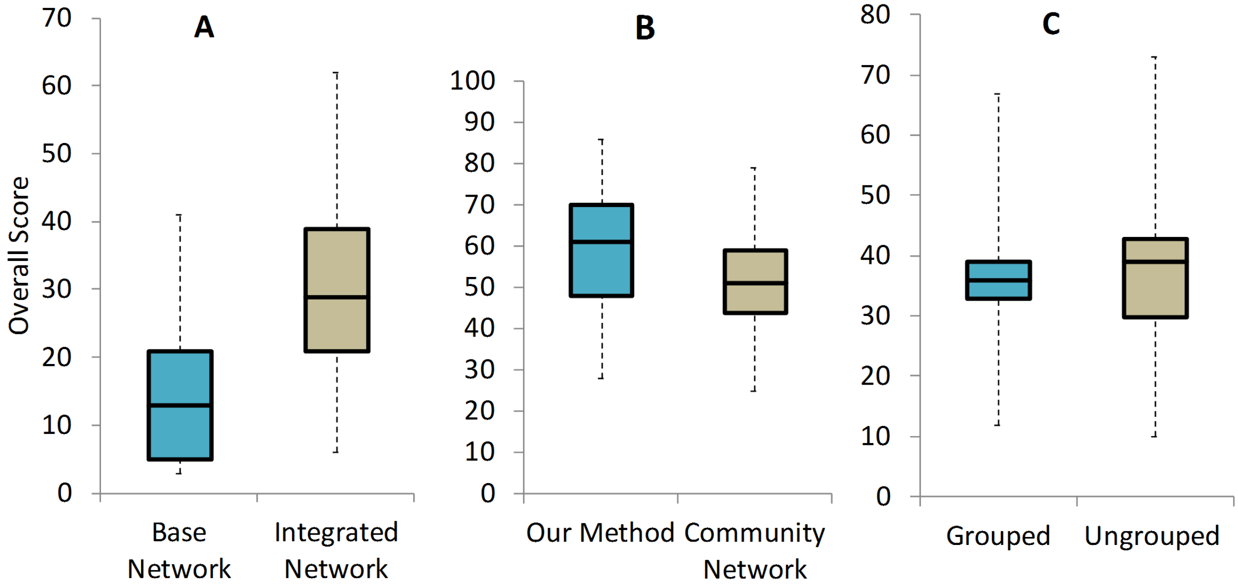


Figure 7.3: A) Effect of integrating other network data. B) Comparison with rank-based community networks from [100] using randomly sampled 20 inference algorithms from DREAM5 dataset. C) Overall scores from using grouped evidence sets from 5 algorithm categories and ungrouped evidence sets

see [100] for details) as the evidence sets. We integrate each evidence set (randomly ordered) incrementally, and observed that the overall score increased from Figure 7.3 A. We stress that, the network reconstruction method we studied here does not merely serve the purpose of integrating multiple networks (such as the rank-based method described in [100]), rather it is a general method in that biologist could select relevant evidence sets, such as pathway co-occurrence evidence from curated databases, as shown in the following case studies.

To compare integrative networks using our reconstruction method with the rank-based community networks, we treat top 100 predicted edges in randomly sampled 20 inference methods as evidence sets and build integrative network incrementally (Figure 7.3 B). The flexibility of such network reconstruction relies on the fact that one can arbitrarily combine the evidence sets for different analysis. For example, in the above comparison, we could reconstruct the network 20 times, each time adding one evidence set. Alternatively, one can choose to group evidence sets by the type of inference algorithms (Figure 7.3 C), with each

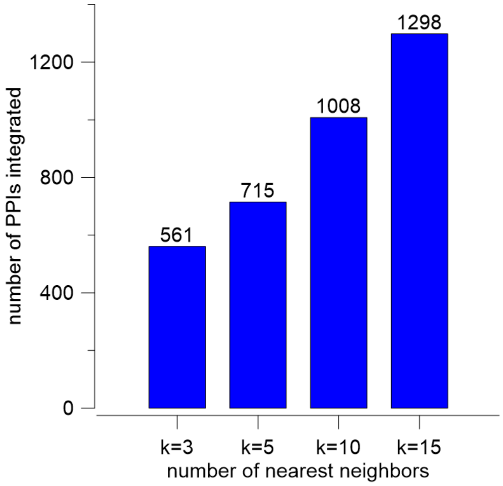


Figure 7.4: Number of protein-protein interactions recalled from network reconstructions

group containing the union of predicted edges from the pertaining methods.

7.4 Case Study: Integrating Gene Expression Data with Interactome Data

We collected 122410 non-self interacting protein interactions from various sources, and 4237 unique probe sets from Wang cohort [170] from differential expression, NCG database, and genome-wide survival screening (see Material & Methods) for Estrogen receptor positive (ER+) tumors. We reconstructed integrated networks by combining gene expression data (quantitative data) and protein interaction data (qualitative data).

To see whether the computed links using the mutual information based distance correctly recall the protein-protein interactions during the integrated reconstruction, we chose $k = 3, 5, 10, 15$ for nearest neighbors and recorded the number of protein interactions being integrated (Figure 7.4). The number of protein interactions recalled from the interactome data increases with the number of nearest neighbors selected. This is not surprising because protein interaction data are often noisy and collected from various sources and experimental protocols. Consequently some of the observed interactions are actually irrelevant in the context of some specific phenotypes like cancer [176], and therefore they should be excluded in the reconstructed network for prediction.

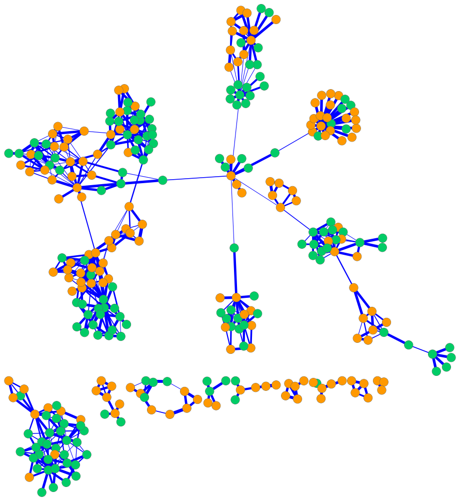


Figure 7.5: Topological structure of extracted network modules using ClusterONE algorithm. Thickness of edges is visualized using reconstruction weight. Overlapped nodes are colored green and non-overlapped nodes are colored orange

To further analyze the integrated network, we performed the clustering analysis using ClusterONE algorithm [115]. ClusterONE identifies overlapped clusters by iteratively maximizing the cohesiveness of networks. We extracted 19 clusters (Figure 7.5) containing total 314 genes (p-value < 0.05) with edges weights being w_{ij} from the network reconstruction. We observed that edges connecting dense regions are less weighted. This suggests that edges in dense regions are more likely to have external qualitative evidence (e.g., protein-protein interactions) supporting them. We further performed Gene Ontology (GO) (`http://www.geneontology.org`) analysis for the integrated network with ontology term level ranging from 7 to 15 (Kappa score = 0.6). For ER+ cancers, we found that the regulation of lymphocyte proliferation, positive regulation of lymphocyte differentiation and activation, peptidyl-tyrosine phosphorylation terms are most enriched by the clustering profile (Figure A.3), along with known cancer-related pathways such as ERK1/ERK2 cascade. The association graph between specific GO terms for the significant clusters is shown in Figure A.3

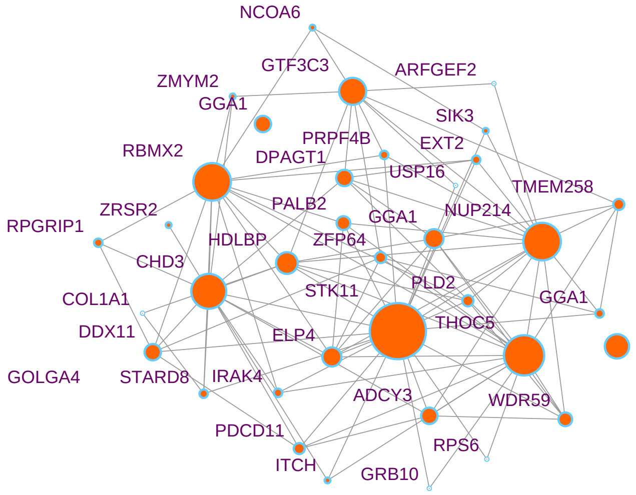


Figure 7.6: Thirty-five genes mapped to differential network, node labeled with gene symbols, node size is proportional to degree

7.5 Case Study: Mining Pathway Co-occurrence Data

7.5.1 Differential Network Analysis

In this case study, we integrate genes expression data with biological pathways. The qualitative evidence is that two genes are supportively connected if they co-occur in the same pathway. We mapped 324 NCG cancer genes to 431 probe sets and collected 4387 biological pathways (see Material & Methods). In cancer studies, we are often interested in comparing different phenotypic networks. Here, we constructed integrated networks using metastatic and non-metastatic patient samples with Wang's data [170]. Let $W_{\text{metastatic}}$ and $W_{\text{non-metastatic}}$ denote the weighted adjacency matrix from the reconstruction method. We reconstructed the differential network as $W_{\text{diff}} = W_{\text{metastatic}} - W_{\text{non - metastatic}}$. W_{diff} contains the weight difference between edges in metastatic and non-metastatic networks, therefore by analyzing this differential network, we could find genes discriminative between two phenotypes. Table 7.1 summarizes the overall topological statistics of 3 filtered networks ($w > 0.01$). $W_{\text{metastatic}}$ recalls 848 co-occurrences between node neighbors, whereas $W_{\text{non-metastatic}}$ recalls 750 co-occurrences. This suggests that discriminative weights are assigned in different phenotypic networks reconstructed with our method.

7.5.2 Identifying Differential Network Genes for Breast Cancer Metastasis

We computed the weighted node degree d_i of $|W_{\text{diff}}| := \left|[W_{\text{diff}}]_{ij}\right|$, where $|.|$ refers to the absolute value. d_i is the absolute total weight difference of node i between $W_{\text{metastatic}}$ and $W_{\text{non-metastatic}}$. We found 35 genes with $d_i >$ average weighted degree (Figure A.4), which are weighted hubs in the differential network [11, 157]. These genes are the ones that vary the most between metastatic and non-metastatic networks in terms of edge weights. We extracted a subnetwork with 35 genes from the differential network (Figure 7.6). *GNAS*, *ADCY3* and *PLD2* are found in GnRH signaling pathway, which is coupled to G-proteins to activate phospholipase C in human. The downstream of GnRH signaling pathway transactivates the epidermal growth factor receptor (EGFR) and activates the mitogen-activated protein kinases (MAPKs). Our results suggest that GnRH signaling maybe the source of cellular instability (as evidenced from differential network analysis) that triggers breast cancer metastasis. *RPS6* and *STK11* participate in mTOR signaling pathway that integrates intra- and extra- cellular signals to regulate cell growth and proliferation. *STK11* acts as a hub in the extracted subnetwork. In contrast, *RPS6* only interacts with *STK11* and *WDR59*.

Table 7.1: Topological comparisons between metastatic, non-metastatic and differential networks

	Metastatic	**Non-metastatic**	**Differential**
network clustering coefficient	0.575	0.471	0.36
number of edges	1564	1618	1301
avg. shortest path	2.344	2.421	2.468
avg. no of neighbors	7.258	7.508	6.137
network density	0.017	0.017	0.015
network heterogeneity	2.846	2.563	2.857

7.6 Case Study: Network-based Predictions using Partial Knowledge

In the above two case studies, we integrated gene expression data (quantitative evidence) with protein interactome and pathway data (qualitative evidence). We now use the integrated networks to predict relevant network-based markers using the semi-supervised framework (see Material & Methods). To check if the network-based predicted markers are distinct by integrating different qualitative evidence, i.e., protein interaction data (NCG+PPI) and pathway data (NCG+PATH), we use the same NCG genes in Wang's cohort as in the previous case study, with $k = 10$ (number of nearest neighbors) to reconstruct NCG+PPI and NCG+PATH networks. The semi-supervised learning requires labels (representing a priori known knowledge) of nodes. We refer to these labeled nodes as seed nodes. Obviously, one can define seed nodes based on different views, for example, using several experimentally verified disease genes as seed nodes. For our purpose here, we simply use weighted hubs (defined as nodes with degrees > average node degree of a network + 2 standard deviations of the total node degree distribution) as the seed nodes for prediction.

NCG+PPI and NCG+PATH networks recall 175 and 645 supported edges in the network reconstruction, respectively, this suggests that different data sources used in the integration substantially affect the result of predicted markers; therefore network-based prediction of markers should be accompanied by purpose-specific contexts. Figure 7.7 shows subnetworks extracted with 10 seed nodes and top 20 predicted nodes for NCG-PPI and NCG+PATH.

7.7 Chapter Discussion and Conclusions

The optimization problem has close relationship with dimensionality reduction: the objective function (Eq. 7.1) is often used to find reconstruction weights for mapping the data points to a low-dimensional coordinate system [162]. The optimal weights w_{ij} are invariant to translation, rotation and rescaling, and can be used to further approximate low-dimensional representation of data points.

The working assumption of the reconstruction process described is that each center node can be represented as a linear combination of its neighboring nodes as measured by certain metric. This linearity assumption is fundamental in many regression models widely used in multivariate predictions. In the reconstruction process, the assumption simplifies the optimization problem to quadratic programming problems which can be efficiently solved, in contrast to kernel-based methods that involve much harder SDP. Compared with existing works, the method described in this chapter does not require complex prior modeling like in the statistical-based approaches and mitigates the limitations of kernel-based approaches by allowing flexible genomic data representation (i.e., partial or incomplete). Given different types of genomic evidence, we aim to recover a weighted network that best depicts associations between nodes in high dimensional spaces. Further, the nearest-neighbor fashion of relaying strength of associations can be used in efficiently making predictions and extracting subnetworks in multiplex networks.

The only parameter for the integrated reconstruction and subsequent network-based prediction is the number of nearest neighbors, k. Although we simply chose k empirically in this chapter based on AUPR and the nature of integrated datasets, flexibility can be sought to fit customized settings. For example, one could use different k for each node based on their ranked importance for a phenotype, which can be deemed as another way of integrating known knowledge. More importantly, such refinement does not incur additional computational cost for network reconstruction, because nearest neighbors of each node are independently computed and then assembled into a multiplex network. In general, our method of posing data integration as constraints is flexible and computationally efficient.

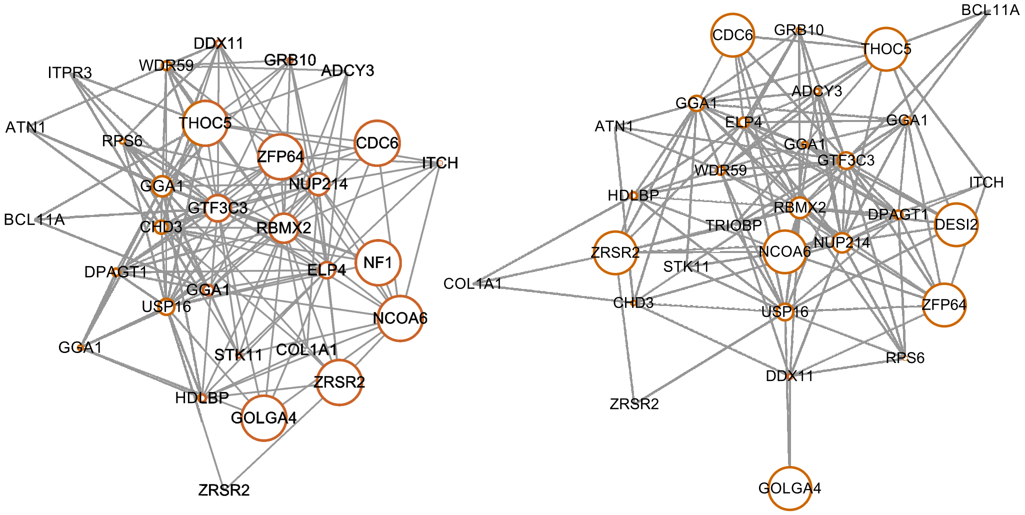


Figure 7.7: Subnetworks with 10 seed nodes and top 20 nodes in terms of predicted scores. Left: NCG+PPI. Right: NCG+PATH. Node size is proportional to degree

Chapter 8

Conclusion and Future Directions

In the past, it has become increasingly evident that networks are ubiquitous in the real world: from social networks to the increasing potential of network biology and medicine. Linked structure (possibly hidden) from data fundamentally renovates the method of computation modeling and knowledge discovery so as to comprehend big data and their embedded relationships.

8.1 Conclusions

The network serves as an effective data media that inspires network medicine and network biomarker detection. The rationale embraces the underlying biological mechanism in how biological entities interact in a complex but concerted way to carry out biological functions. Cancer, in particular, appears to be a mysterious endpoint encompassing disease therapeutics, diagnostics and many other bio-medical branches. The cellular complexity naturally caters for network approaches of understanding how genes and their products communicate and interact in a scalable fashion: a network consisting of thousands of nodes can be effectively used to describe a disease cellular state. The genetic causations can therefore be better understood by the network map.

In Chapter 2, we have analyzed social communities from another angle, i.e., we identify the boundaries of social communities through convex hull constructions, and demonstrate the usability of the method through breast cancer genes expression data sets. We argue that social mechanisms can be applied to biological data, and from the experiments we can conclude that boundary genes are less interactive in co-expression modules and more volatile in community memberships subject to noise. The fragility can be resolved by Monte Carlo trials. Three algorithms are proposed for mining dynamic social communities. Our approach contributes to the new branch of mining social communities by identifying and representing social community boundaries. The proposed approach is also well suited to dynamic settings.

In Chapter 3, we regarded the metabolic networks as dynamic entities with evolving properties and explored how this new perspective can further refine our understanding of the underlying biochemical functions given the topologies of the reconstructed metabolic networks. We investigated the relationship between the multi-scalability of community structures of metabolic networks and the distributional effect of network motifs. We observed several patterns through studying three organisms, including the effect of directionality of networks, homogeneity of motif-enriched communities, and motif type-specific distributions across scales. We also provide methods to quantify motif influence under the community context. Our work suggests that the theoretic evolvability of modularity tightly correlates with motif distributional effect.

In Chapter 4, we developed a tool based on the "Louvain" method to detect community structures for arbitrary network types. The heuristic-based method tremendously reduces the computational time when detecting community in very large networks.

In Chapter 5, we proposed a method to quantify relational heterogeneity from gene co-expression networks. We first stratified genes based on the level of relational heterogeneity and showed that such classification is predictive of patient survival in breast cancer metastasis. We further explored the network markers obtained from highly relational heterogeneous gene set and demonstrated the improved performance to predict patient survival. Such design of disease network biomarkers may offer some new opportunities for targeted cancer treatment and personalized medicine, because inter-tumor heterogeneity is posed as an unsolved challenge for many cancers and further understanding of distinct molecular features from interaction dynamics is a key to design network biomarkers.

In Chapter 6, we provided a method to evaluate network-based markers. The conclusion that network-based markers are not consistently predictive across patient cohorts are not surprising. Given that most gene signatures are not robust in presence of cellular complexity, experimental noise, and incomplete data collection, the added dimension, i.e., from single genes that form gene signatures to interacting genes that form a much more complex network modules, further confounds predictive signals. For this reason, network-based personalized medicine are not clinically deployed, although the network-based thinking is pervasive. Our approach in evaluating network-based markers is an added line of evidence.

In Chapter 7, we designed an algorithm to efficiently reconstruct gene-gene networks by integrating multiple sources of genomic data. The method does not impose data constraints, and more importantly, we improved the performance of network-based predictions using reconstructed networks. Sub-networks can be subsequently extracted. In a sense, we reduced the data complexity and achieved better predictive performance by a single computational framework.

Overall, network approaches have become useful and effective in system biology studies. In this thesis, we discussed several computational approaches to deal with data complexity and to improve the predictability using different biological networks.

8.2 Future Directions

The network promise comes with two notable challenges to the computational infrastructure, among others.

1. With a flood of biological data sets abreast with the advance of experimental protocols, flexible integration of heterogeneous data is a problem that is largely undealt with due to case-to-case variation, specificity, and complexity. Our future goal is to resolve such a challenge by building an effective solution to integrate the huge pool of data into network representation. The flexibility and efficiency boil down to a centralized design to manage acquisition, validation,

storage and distribution of data in a consistent, editable environment.

2. In general, effective apparatus to mine the big network data is lacking, which prevents the network realizing its practical value. A well-designed analytical engine to query and perform large-scale, multiplex network analysis of data is far from efficient and flexible for heterogeneous, and possibly noisy, cancer data. A graph-based data mining capability, including a network analytical engine and services, thus becomes a key function requirement in the network data repository. My target is, therefore, to engineer graph data storage with an analytical engine. Existing graph databases provide a starting point to achieve this goal; however, it is limited to data storage and data retrieval.

Although in recent years network approaches to medicine and biomarker detection has shed light on solving the mystery of cancer, the urge to perform integrative, scalable and differential network analysis is bottlenecked by state-of-the-art information provisions and hosting. Only when such a bottleneck is addressed, can robust and reproducible network methods be advanced and practically deployed. In the system biology paradigm, the design and implementation of the information solution is essential in answering cancer biology questions.

In the future study, we plan to explore the differential network biology approach in complex disease analysis [94, 63]. So far, the network-based markers are derived from static network structures, and as noted previously, such assumption is not realistic in most circumstances. As an important step further, we could derive differential networks, which contain nodal and link differences between two phenotypic networks in question. In fact, the notion of "differential network biology" is applied successfully in previous literature to study network responses to DNA damage [8]. In the context of cancer biomarkers, let us consider the illustration in Figure 8.1.

Given two networks, possibly inferred from genomic data sets or obtained from known networks such as human signalling pathways after overlaying high-throughput data (Figure 8.1 A & B), differential links could be identified between two phenotypic networks (Figure 8.1 B) by thresholding interaction strengths. Using those differential links (Figure 8.1 C), we could prioritize sub-networks that are relevant to the the disease phenotype (Figure 8.1 D). Clearly again, data complexity and predictability are two sides of the same coin.

We conclude the thesis with George E. P. Box's remark, "All models are wrong, but some are useful." We felt the same way in modeling complex computational biology problems. There are huge challenges in the years to come, including data management, predictive models and technological advances, etc. However, given the success of the system biology, we believe that multi-disciplinary studies and efforts become promising than ever to solve complex problems in the filed of biological and medical research.

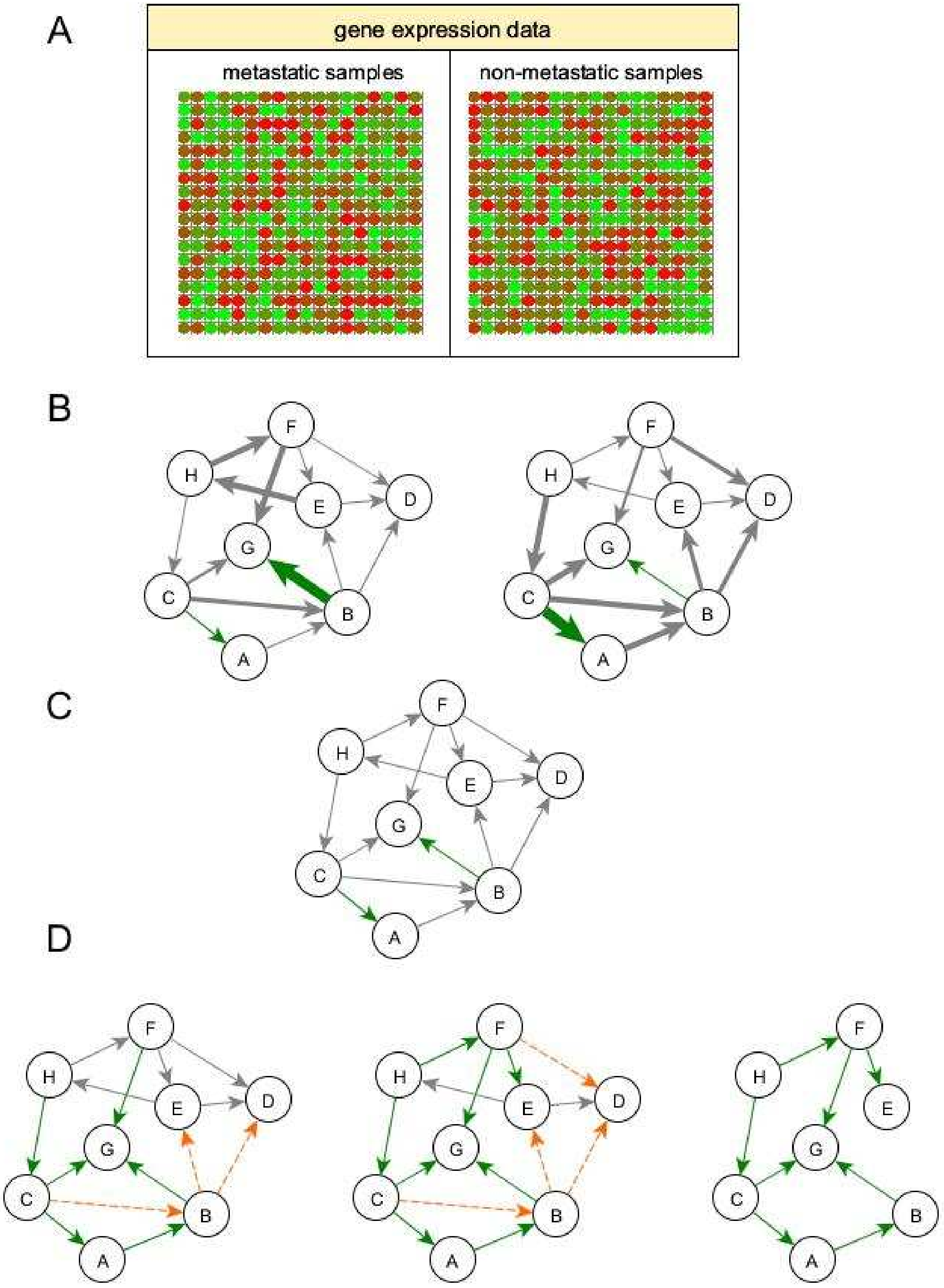


Figure 8.1: An example of differential network biological approaches; thickness of links represents association/interaction strengths; orange links represent interactions that are pruned iteratively (from left to right in D).

Appendix A

Supplementary Figures

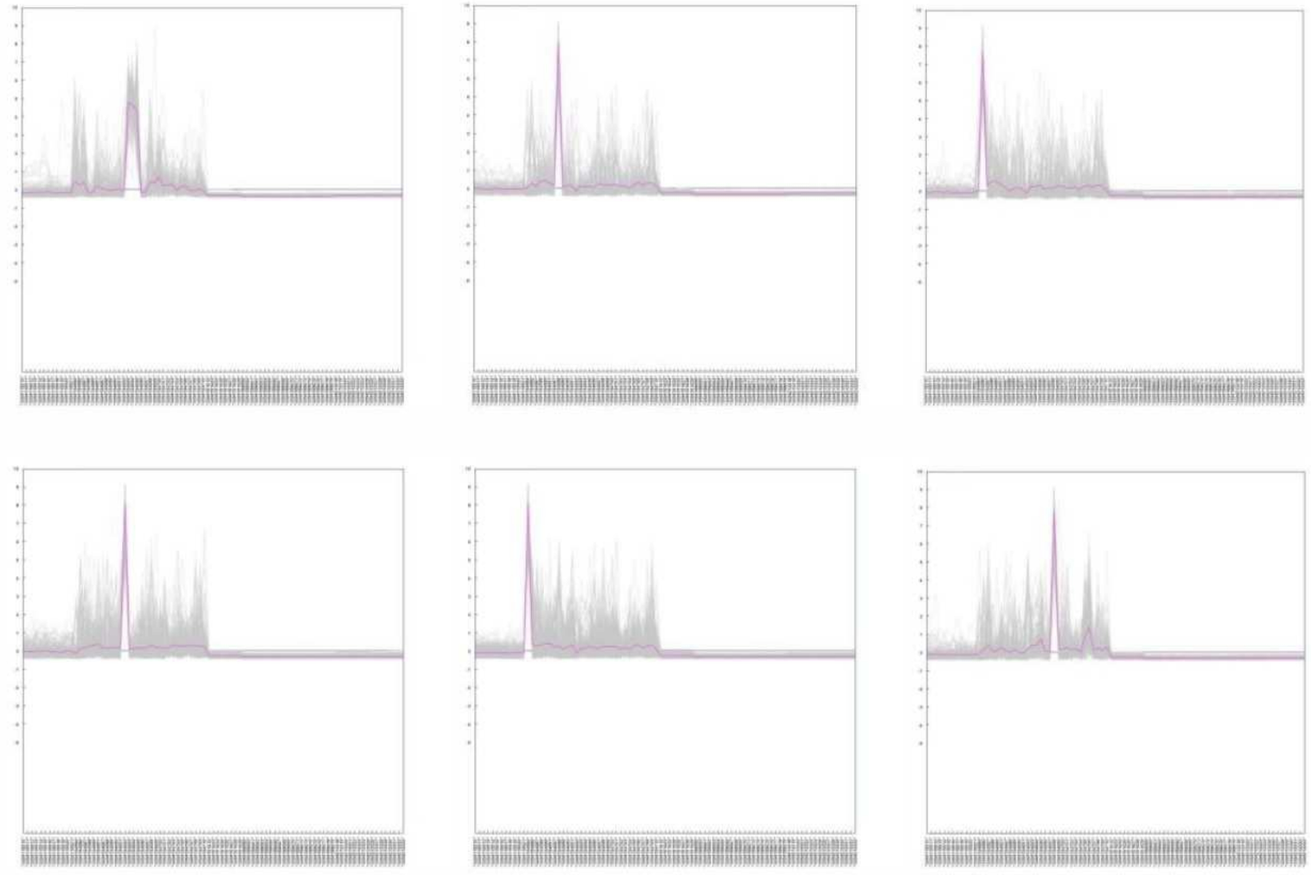


Figure A.1: Six most significant gene co-expression clusters

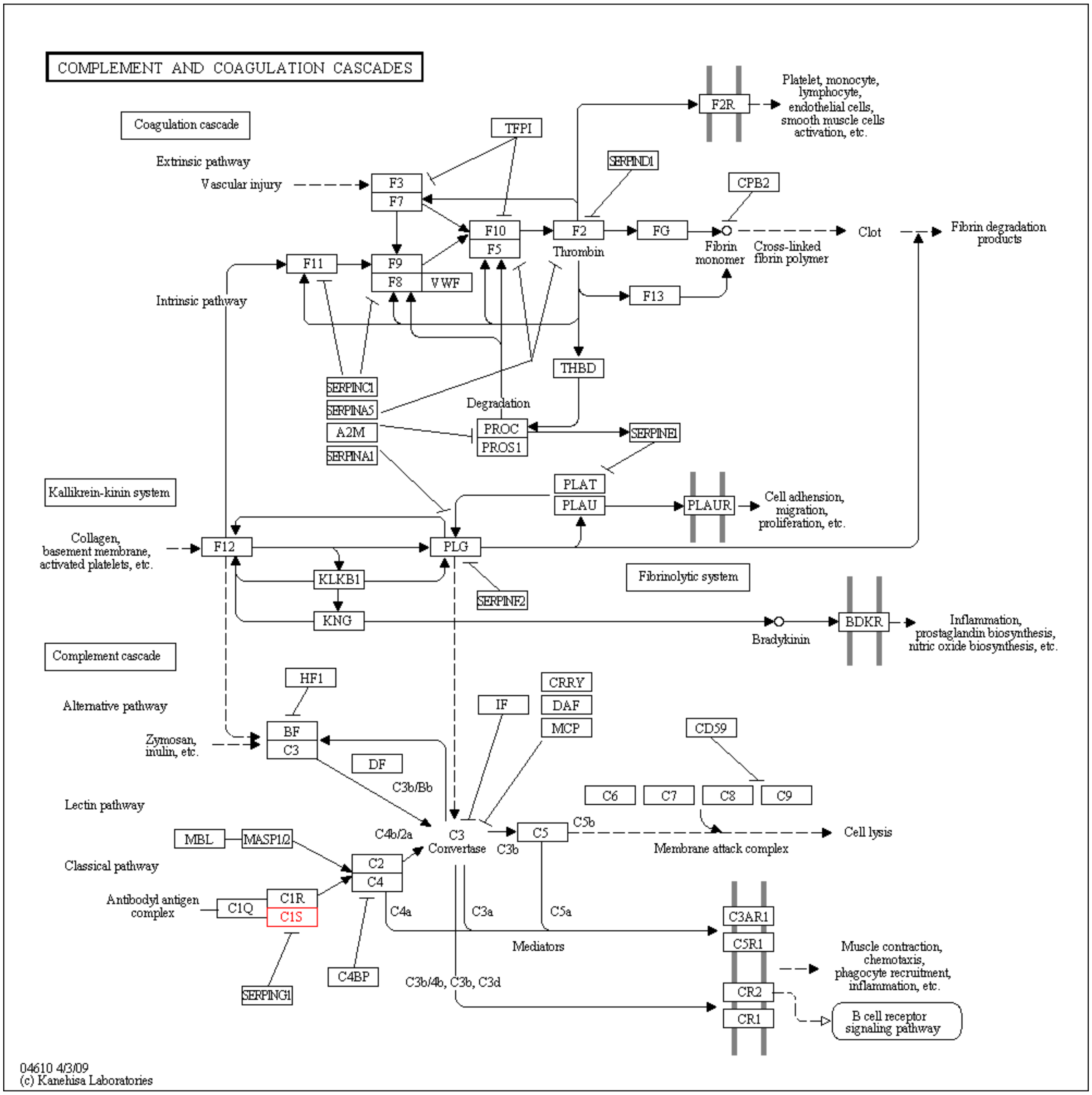


Figure A.2: C1S from KEGG

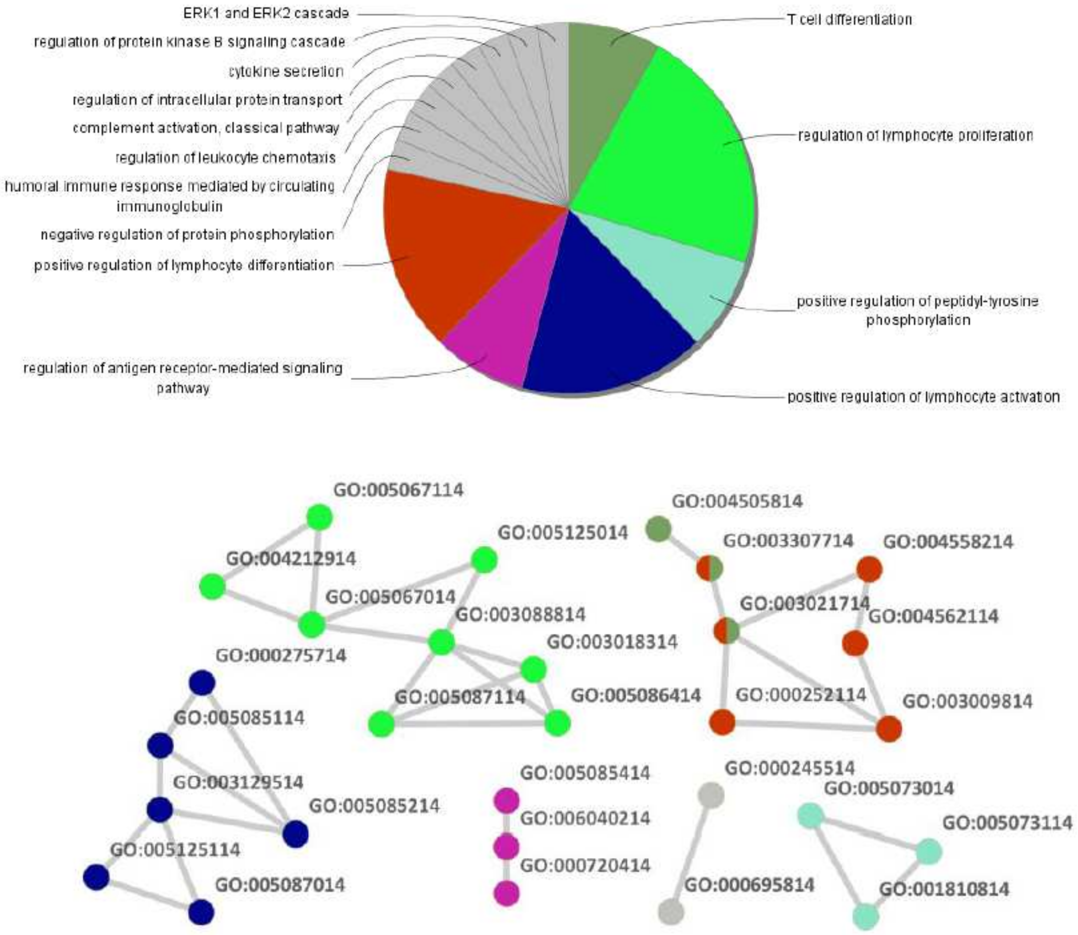
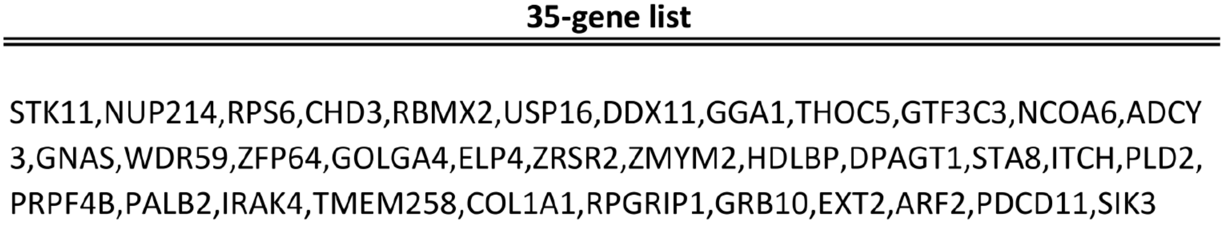


Figure A.3: GO summary of genes in extracted modules (top) and term-term associations (bottom) for the integrated PPI network (Kappa Score = 0.6)



35-gene list

STK11,NUP214,RPS6,CHD3,RBMX2,USP16,DDX11,GGA1,THOC5,GTF3C3,NCOA6,ADCY3,GNAS,WDR59,ZFP64,GOLGA4,ELP4,ZRSR2,ZMYM2,HDLBP,DPAGT1,STA8,ITCH,PLD2,PRPF4B,PALB2,IRAK4,TMEM258,COL1A1,RPGRIP1,GRB10,EXT2,ARF2,PDCD11,SIK3

Figure A.4: Most varied genes in W_{diff} in terms of weighted node degree.

Appendix B

Modularity Gain Computation

Here we provide the details for modularity gain (ΔQ) computation with resolution parameter (γ) for all major types of networks (Fig. B.1).

B.1 Calculating modularity gain

In order to be able to apply the Louvain method to maximize the modularity measure, we need to calculate the gain in modularity (ΔQ) resulting from moving node x from its current community U to any other community V. For all types of networks we more or less take the following approach:

First, we break ΔQ into two terms ΔQ_{add} and ΔQ_{remove}, where ΔQ_{add} is the change in modularity caused by adding x to V while ΔQ_{remove} represents the change in modularity caused by removing x from U. Thus, we have:

$$\Delta Q = \Delta Q_{add} + \Delta Q_{remove} \tag{B.1}$$

We further define:

$$\Delta Q_{add} = Q_{joint}(x, V) - Q_{disjoint}(x, V) \tag{B.2}$$

and

$$\Delta Q_{remove} = Q_{disjoint}(x, U) - Q_{joint}(x, U) \tag{B.3}$$

where $Q_{disjoint}(x, G)$ is the modularity of node x and community G when x is not assigned to G and $Q_{joint}(x, G)$ is the modularity when x is assigned to G. Therefore, to get ΔQ we need to work out the joint and disjoint modularities of x and U and V.

In the following subsections we apply this approach to calculate the modularity gain in different types of networks. We only present ΔQ_{add} computations as ΔQ_{remove} computations are symmetric.

B.2 Undirected Network

The formulation of Modularity (Q) for undirected graphs is as follows:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \gamma \frac{k_i k_j}{2m} \right] \delta(g_i, g_j) \tag{B.4}$$

where A_{ij} represents the weight of the edge between i and j, $k_i = \sum_j A_{ij}$, γ is the resolution parameter, g_i is the community to which vertex i is assigned, $\delta(a, b)$ is 1 if $a = b$ and 0 otherwise, and $m = \frac{1}{2} \sum_{ij} A_{ij}$.

Following the approach presented for calculating the modularity gain, we have:

$$\begin{aligned} Q_{disjoint}(x, V) =& \frac{1}{2m} \left(A_{xx} + \textstyle\sum_{inside}(V) \right) \\ & - \frac{\gamma}{4m^2} \left(k_x^2 + \sum_{i \in V} k_i \sum_{j \in V} k_j \right) \\ =& \frac{1}{2m} \left(A_{xx} + \textstyle\sum_{inside}(V) \right) \\ & - \frac{\gamma}{4m^2} \left(k_x^2 + \textstyle\sum_{tot}(V)^2 \right) \end{aligned} \tag{B.5}$$

where $\sum_{inside}(G) = \sum_{i,j \in G} A_{ij}$, and $\sum_{tot}(G)$ is the sum of the weights of the links incident to nodes in G. For $Q_{joint}(x, V)$, we have:

$$\begin{aligned} Q_{joint}(x, V) =& \frac{1}{2m} \left(A_{xx} + 2k_{x,in}(V) + \textstyle\sum_{inside}(V) \right) \\ & - \frac{\gamma}{4m^2} \left((k_x + \textstyle\sum_{tot}(V))^2 \right) \\ =& \frac{1}{2m} \left(A_{xx} + 2k_{x,in}(V) + \textstyle\sum_{inside}(V) \right) \\ & - \frac{\gamma}{4m^2} (k_x^2 + \textstyle\sum_{tot}(V)^2 + 2k_x \textstyle\sum_{tot}(V)) \end{aligned} \tag{B.6}$$

where $k_{x,in}(G)$ is the sum of the weights of the links from x to nodes in C.

According to Equations B.2, B.5, and B.6:

$$\Delta Q_{add} = \frac{1}{m}\left(k_{x,in}(V)\right) - \frac{\gamma}{2m^2}\left(k_x \textstyle\sum_{tot}(V)\right) \tag{B.7}$$

Using the notation presented above, we redefine the modularity as the summation of the modularities of clusters:

$$Q = \frac{1}{2m}\sum_{1}^{N_M}\left[\textstyle\sum_{inside}(G) - \frac{\gamma}{2m}\textstyle\sum_{tot}(G)^2\right] \tag{B.8}$$

where N_M is the number of communities. One advantage of Equation B.8 over Equation B.4 is that it gives us a more efficient way of calculating Q when we already have $\sum_{inside}$ and $\sum_{tot}$ of all communities.

B.3 Directed Network

Equation B.9 defines the modularity measure for directed networks.

$$Q_{directed} = \frac{1}{m}\sum_{i,j}\left[A_{ij} - \gamma\frac{k_i^{out}k_j^{in}}{m}\right]\delta(g_i, g_j) \tag{B.9}$$

where A_{ij} is the weight of the directed edge from i to j, $k_i^{out} = \sum_j A_{ij}$, and $k_j^{in} = \sum_i A_{ij}$.

Following the gain calculation approach, we have:

$$\begin{aligned} Q_{disjoint}(x, V) =& \frac{1}{m}\left(A_{xx} + \textstyle\sum_{inside}(V)\right) \\ &- \frac{\gamma}{m^2}\left(k_x^{in}k_x^{out} + \textstyle\sum_{tot}^{in}(V)\textstyle\sum_{tot}^{out}(V)\right) \end{aligned} \tag{B.10}$$

where $\sum_{tot}^{in}(G) = \sum_{i\in G} k_i^{in}$. For $Q_{joint}(x, V)$, we have:

$$\begin{aligned} Q_{joint}(x, V) = \frac{1}{m}&\left(A_{xx} + \textstyle\sum_{inside}(V) + \sum_{i\in V} A_{ix} + \sum_{i\in V} A_{xi}\right) \\ &- \frac{\gamma}{m^2}\Big(k_x^{in}k_x^{out} + \textstyle\sum_{tot}^{in}(V)\textstyle\sum_{tot}^{out}(V) \\ &\qquad + k_x^{in}\textstyle\sum_{tot}^{out}(V) + k_x^{out}\textstyle\sum_{tot}^{in}(V)\Big) \end{aligned} \tag{B.11}$$

Now we work out ΔQ_{add} based on Equations B.2, B.10 and B.11.

$$\Delta Q_{add} = \frac{1}{m}\left(\sum_{i\in V} A_{ix} + \sum_{i\in V} A_{xi}\right) - \frac{\gamma}{m^2}\left(k_x^{in}\sum_{i\in V} k_i^{out} + k_x^{out}\sum_{i\in V} k_i^{in}\right) \tag{B.12}$$

Akin to the undirected networks, we can redefine ΔQ for directed networks.

$$Q = \frac{1}{m}\sum_{1}^{N_M}\left[\textstyle\sum_{inside}(G) - \frac{\gamma}{m}\sum_{tot}^{in}(G)\sum_{tot}^{out}(G)\right] \tag{B.13}$$

B.4 Signed Network

Modularity in signed networks is defined as:

$$\begin{aligned} Q = & \overbrace{\frac{1}{2m}\sum_{ij}\left[A_{ij}^{+} - \gamma\frac{k_i^{+}k_j^{+}}{2m^{+}}\right]\delta(g_i, g_j)}^{\text{positive}} \\ & - \overbrace{\frac{1}{2m}\sum_{ij}\left[A_{ij}^{-} - \gamma\frac{k_i^{-}k_j^{-}}{2m^{-}}\right]\delta(g_i, g_j)}^{\text{negative}} \end{aligned} \tag{B.14}$$

where $m = m^{+} + m^{-}$.

To calculate ΔQ_{add} in a signed graph we consider the positive and negative parts of Equation B.14 separately and work out ΔQ_{add} for each part in a similar way to an undirected graph.

$$\begin{aligned} \Delta Q_{add} = & \Delta Q_{add}^{+} - \Delta Q_{add}^{-} \\ = & \frac{1}{m}k_{x,in}^{+}(V) - \gamma\frac{1}{2(m^{+})m}k_x^{+}\textstyle\sum_{tot}^{+}(V) \\ & - \frac{1}{m}k_{x,in}^{-}(V) + \gamma\frac{1}{2(m^{-})m}k_x^{-}\textstyle\sum_{tot}^{-}(V) \end{aligned} \tag{B.15}$$

B.5 Bipartite Network

In this subsection, we study two different definitions of modularity for bipartite networks and work out modularity gain in both cases.

B.5.1 First Method

The first formulation of Modularity (Q) for two-mode networks is as follows:

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \gamma b_{ij}\frac{k_i k_j}{m}\right]\delta(g_i, g_j) \tag{B.16}$$

where b_{ij} is one if i and j belong to different modes and 0 otherwise.

For $Q_{disjoint}(x, V)$, we have:

$$\begin{aligned} Q_{disjoint}(x,V) \ =& \frac{1}{2m}{\textstyle\sum}_{inside}(V) \\ & - \frac{1}{2m^2}{\textstyle\sum}_{tot}^{M_x}(V){\textstyle\sum}_{tot}^{\neg M_x}(V) \end{aligned}$$

where M_x is the mode to which x belongs while $\neg M_x$ is the other mode, and ${\textstyle\sum}_{tot}^{M}(G)$ is the sum of the weights of the links incident to nodes of mode M in G. For $Q_{joint}(x, V)$, we have:

$$\begin{aligned} Q_{joint}(x,V) =& \frac{1}{2m}\left(2k_{x,in}(V) + {\textstyle\sum}_{inside}(V)\right) \\ & - \frac{1}{2m^2}\left(k_x + {\textstyle\sum}_{tot}^{M_x}(V)\right){\textstyle\sum}_{tot}^{\neg M_x}(V) \end{aligned}$$

Now we get ΔQ_{add} from Equation B.2.

$$\Delta Q_{add} = \frac{1}{m}k_{x,in}(V) - \frac{1}{2m^2}k_x{\textstyle\sum}_{tot}^{\neg M_x}(V) \tag{B.17}$$

The following equation defines modularity as a summation over all communities.

$$Q = \frac{1}{2m}\sum_{1}^{N_M}\left[{\textstyle\sum}_{inside}(G) - \frac{2\gamma}{m}{\textstyle\sum}_{tot}^{M_1}(G){\textstyle\sum}_{tot}^{M_2}(G)\right] \tag{B.18}$$

B.5.2 Second Method

Guimerá *et al.* [52] consider the nodes in a bipartite network as a number of *actors* and *teams*. They define the bipartite modularity of partition $\mathcal{P}$ as:

$$\mathcal{M}_{\mathcal{B}}(\mathcal{P}) = \sum_{s=1}^{N_M}\left(\frac{\sum\limits_{i\neq j\in s} c_{ij}}{\sum\limits_{a} m_a\left(m_a - 1\right)} - \frac{\sum\limits_{i\neq j\in s} t_i t_j}{\left(\sum\limits_{a} m_a\right)^2}\right) \tag{B.19}$$

where t_i is the total number of teams to which actor i belongs, m_a is the number of actors belonging to team a, and c_{ij} is the actual number of teams in which i and j are together. Here, we use $\mathcal{M}_{\mathcal{B}}(\mathcal{P})$ interchangeably with Q which is the general notation of modularity.

The disjoint modularity of actor x and community V is as follows:

$$\begin{aligned} Q_{disjoint}(x,V) =& \frac{1}{\sum\limits_a m_a\left(m_a-1\right)}\left(\sum_{i\neq j\in V} c_{ij}\right) \\ & - \frac{1}{\left(\sum\limits_a m_a\right)^2}\left(\sum_{i\neq j\in V} t_i t_j\right) \end{aligned}$$

For $Q_{joint}(x,V)$, we have:

$$\begin{aligned} Q_{joint}(x,V) =& \frac{1}{\sum\limits_a m_a\left(m_a-1\right)}\left(\sum_{i\neq j\in V} c_{ij} + 2\sum_{i\in V} c_{xi}\right) \\ & - \frac{1}{\left(\sum\limits_a m_a\right)^2}\left(\sum_{i\neq j\in V} t_i t_j + 2\sum_{i\in V} t_x t_i\right) \end{aligned}$$

Finally, we work out ΔQ_{add} based on Equation B.2.

$$\Delta Q_{add} = \frac{2}{\sum\limits_a m_a\left(m_a-1\right)}\left(\sum_{i\in V} c_{xi}\right) - \frac{2}{\left(\sum\limits_a m_a\right)^2}\left(\sum_{i\in V} t_x t_i\right) \tag{B.20}$$

The authors of [52] also give a formulation for modularity in directed networks by transforming them into a undirected bipartite networks. The calculation of modularity gain presented here is also valid in such case.

B.6 Multi-slice Network

The formulation of Modularity (Q) in multi-slice networks is defined by the following equation:

$$Q = \frac{1}{2\mu}\sum_{ijsr}\left[\overbrace{\left(A_{ijs} - \gamma_s\frac{k_{is}k_{js}}{2m_s}\right)\delta_{sr}}^{\text{intra-slice}} + \overbrace{\delta_{ij}C_{jsr}}^{\text{inter-slice}}\right]\delta(g_{is}, g_{jr}) \tag{B.21}$$

where, at time slice s, A_{ijs} represents the weight of the edge between i and j, $k_{is} = \sum_j A_{ijs}$, $2m_s = \sum_{ij} A_{ijs}$, and g_{is} is the community to which vertex i is assigned at time slice s. $Cjsr$ is the weight of the inter-slice edge linking node j at time slice r to the same node at time slice s. $2\mu = \sum_{jr} \kappa_{jr}$, where κ_{jr} is defined by $\kappa_{jr} = k_{jr} + c_{jr}$ in which $k_{jr} = \sum_i A_{ijr}$ and $c_{jr} = \sum_s C_{jsr}$ [113].

In order to simplify the calculation of ΔQ_{add} in multi-slice networks, we break it into ΔQ_{intra} and ΔQ_{inter} based on the intra- and inter-slice portions of Equation B.21. ΔQ_{intra} is calculated in a manner similar to ΔQ in an undirected single-slice network and ΔQ_{inter} is given by:

$$\Delta Q_{inter} = \sum_s C_{xst} \delta(g_{xs}, V)$$

Thus we have:

$$\Delta Q_{add} = \frac{1}{2\mu} \left[\overbrace{\left(2k_{x,in}(V_t) - \frac{\gamma_t}{m_t} \left(k_{xt} \textstyle\sum_{tot}(V_t) \right) \right)}^{\text{intra-slice}} + \overbrace{\sum_s C_{xst} \delta(g_{xs}, V)}^{\text{inter-slice}} \right] \tag{B.22}$$

where $k_{x,in}(V_t)$ is the sum of the weights of the intra-slice links from x_t to nodes in V, and $\sum_{tot}(V_t)$ is the sum of the weights of the intra-slice links incident to nodes in V at time slice t.

Equation B.21 can be rewritten in community format as follows:

$$Q = \frac{1}{2\mu} \sum_1^{N_M} \left[\sum_s \left(\textstyle\sum_{inside}(G_s) - \frac{\gamma_s}{2m_s} \textstyle\sum_{tot}(G_s)^2 \right) + \textstyle\sum_{inside}^{inter}(G) \right] \tag{B.23}$$

where $\sum_{inside}^{inter}(G) = \sum_{i \in G} \sum_{rs} C_{irs}$.

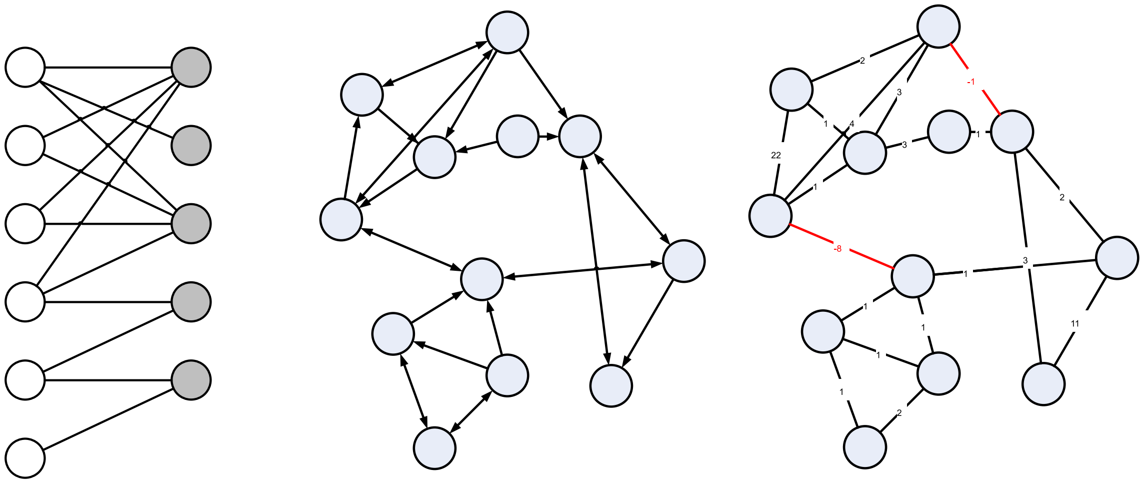


Figure B.1: Different types of networks which MCF can handle. From left to right: bipartite network, directed network, signed network (negative links in red).

Bibliography

[1] A. Aggarwal, D. L. Guo, Y. Hoshida, S. T. Yuen, K. M. Chu, S. So, A. Boussioutas, X. Chen, D. Bowtell, H. Aburatani, S. Y. Leung, and P. Tan. Topological and functional discovery in a gene coexpression meta-network of gastric cancer. *Cancer Res*, 66(1):232–41, 2006.

[2] M. Aluru, J. Zola, D. Nettleton, and S. Aluru. Reverse engineering and analysis of large genome-scale gene networks. *Nucleic acids research*, 41(1):e24, 2013.

[3] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to mcmc for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

[4] Alexey V Antonov, Sabine Dietmann, Igor Rodchenkov, and Hans W Mewes. Ppi spider: a tool for the interpretation of proteomics data in the context of protein–protein interaction networks. *Proteomics*, 9(10):2740–2749, 2009.

[5] Alex Arenas, Alberto Fernandez, and Sergio Gomez. Analysis of the structure of complex networks at different resolution levels. *New Journal of Physics*, 10(5):053039, 2008.

[6] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[7] S Bamford, E Dawson, Simon Forbes, J Clements, R Pettett, A Dogan, A Flanagan, J Teague, P Andrew Futreal, MR Stratton, et al. The cosmic (catalogue of somatic mutations in cancer) database and website. *British journal of cancer*, 91(2):355–358, 2004.

[8] Sourav Bandyopadhyay, Monika Mehta, Dwight Kuo, Min-Kyung Sung, Ryan Chuang, Eric J Jaehnig, Bernd Bodenmiller, Katherine Licon, Wilbert Copeland, Michael Shales, et al. Rewiring of genetic networks in response to dna damage. *Science*, 330(6009):1385–1389, 2010.

[9] Mukesh Bansal, Vincenzo Belcastro, Alberto Ambesi-Impiombato, and Diego Di Bernardo. How to infer gene networks from expression profiles. *Molecular systems biology*, 3(1), 2007.

[10] Albert-László Barabási. Scale-free networks: a decade and beyond. *Science*, 325(5939):412–413, 2009.

[11] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011.

[12] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[13] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, volume 14, pages 585–591, 2001.

[14] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

[15] Shai Ben-David and Ulrike Von Luxburg. Relating clustering stability to properties of cluster boundaries. In *COLT*, volume 2008, pages 379–390, 2008.

[16] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *J Statistical Mechanics: Theory and Experiment*, 2008(10):P10008+, 2008.

[17] Ivana Bozic, Tibor Antal, Hisashi Ohtsuki, Hannah Carter, Dewey Kim, Sining Chen, Rachel Karchin, Kenneth W Kinzler, Bert Vogelstein, and Martin A Nowak. Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences*, 107(43):18545–18550, 2010.

[18] Atul J Butte and Isaac S Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Pac Symp Biocomput*, volume 5, pages 418–429, 2000.

[19] Damon Centola. The spread of behavior in an online social network experiment. *science*, 329(5996):1194–1197, 2010.

[20] Aaron N Chang. Prioritizing genes for pathway impact using network analysis. In *Protein Networks and Pathway Analysis*, pages 141–156. Springer, 2009.

[21] James Chen, Lee Sam, Yong Huang, Younghee Lee, Jianrong Li, Yang Liu, H Rosie Xing, and Yves A Lussier. Protein interaction network underpins concordant prognosis among heterogeneous breast cancer signatures. *Journal of biomedical informatics*, 43(3):385–396, 2010.

[22] Lina Chen, Xiaoli Qu, Mushui Cao, Yanyan Zhou, Wan Li, Binhua Liang, Weiguo Li, Weiming He, Chenchen Feng, Xu Jia, et al. Identification of breast cancer patients based on human signaling network motifs. *Scientific reports*, 3, 2013.

[23] Han-Yu Chuang, Eunjung Lee, Yu-Tsueng Liu, Doheon Lee, and Trey Ideker. Network-based classification of breast cancer metastasis. *Molecular systems biology*, 3(1), 2007.

[24] David Croft, Gavin OKelly, Guanming Wu, Robin Haw, Marc Gillespie, Lisa Matthews, Michael Caudy, Phani Garapati, Gopal Gopinath, Bijay Jassal, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, 39(suppl 1):D691–D697, 2011.

[25] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[26] Peter Csermely. Creative elements: network-based predictions of active centres in proteins and cellular and social networks. *Trends in biochemical sciences*, 33(12):569–576, 2008.

[27] Qinghua Cui. A network of cancer genes with co-occurring and anti-co-occurring mutations. *PLoS One*, 5(10):e13180, 2010.

[28] Qinghua Cui, Yun Ma, Maria Jaramillo, Hamza Bari, Arif Awan, Song Yang, Simo Zhang, Lixue Liu, Meng Lu, Maureen O'Connor-McCourt, et al. A map of human cancer signaling. *Molecular systems biology*, 3(1), 2007.

[29] Matteo D'Antonio, Vera Pendino, Shruti Sinha, and Francesca D Ciccarelli. Network of cancer genes (ncg 3.0): integration and analysis of genetic and network properties of cancer genes. *Nucleic acids research*, 40(D1):D978–D983, 2012.

[30] Jishnu Das, Jaaved Mohammed, and Haiyuan Yu. Genome-scale analysis of interaction dynamics reveals organization of biological networks. *Bioinformatics*, 28(14):1873–1878, 2012.

[31] J-C Delvenne, Sophia N Yaliraki, and Mauricio Barahona. Stability of graph communities across time scales. *Proceedings of the National Academy of Sciences*, 107(29):12755–12760, 2010.

[32] Patrik Dhaeseleer, Shoudan Liang, and Roland Somogyi. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16(8):707–726, 2000.

[33] A. Djebbari and J. Quackenbush. Seeded bayesian networks: constructing genetic networks from microarray data. *BMC systems biology*, 2:57, 2008.

[34] Edward R Dougherty and Marcel Brun. A probabilistic theory of clustering. *Pattern Recognition*, 37(5):917–925, 2004.

[35] Janusz Dutkowski and Trey Ideker. Protein networks as logic functions in development and cancer. *PLoS computational biology*, 7(9):e1002180, 2011.

[36] Weinan E, Tiejun Li, and Eric Vanden-Eijnden. Optimal partition and effective dynamics of complex networks. *Proc Natl Acad Sci*, 105(23):7907–7912, 2008.

[37] L. L. Elo, H. Jarvenpaa, M. Oresic, R. Lahesmaa, and T. Aittokallio. Systematic construction of gene coexpression networks with applications to human t helper cell differentiation process. *Bioinformatics*, 23(16):2096–103, 2007.

[38] Young H. Eom, Soojin Lee, and Hawoong Jeong. Exploring local structural organization of metabolic networks using subgraph patterns. *J Theor Biol*, 241(4):823–9, 2006.

[39] Janine T Erler and Rune Linding. Network-based drugs and biomarkers. *The Journal of pathology*, 220(2):290–296, 2010.

[40] Ernesto Estrada and Juan A Rodriguez-Velazquez. Subgraph centrality in complex networks. *Physical Review E*, 71(5):056103, 2005.

[41] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

[42] Santo Fortunato and Marc Barthelemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007.

[43] David. Freedman and Persi. Diaconis. On the histogram as a density estimator:l2 theory. *Probability Theory and Related Fields*, 57(4):453–476, 1981.

[44] Shang Gao, Alan Chen, Ali Rahmani, Tamer Jarada, Reda Alhajj, Doug Demetrick, and Jia Zeng. Mcf: A tool to find multi-scale community profiles in biological networks. *Computer methods and programs in biomedicine*, 112(3):665–672, 2013.

[45] Shang Gao, Jia Zeng, Abdallah M ElSheikh, Ghada Naji, Reda Alhajj, Jon Rokne, and Douglas Demetrick. A closer look at social boundary genes reveals knowledge to gene expression profiles. *Current Protein and Peptide Science*, 12(7):602–613, 2011.

[46] Luz García-Alonso, Roberto Alonso, Enrique Vidal, Alicia Amadoz, Alejandro de María, Pablo Minguez, Ignacio Medina, and Joaquín Dopazo. Discovering the hidden sub-network component in a ranked list of genes or proteins derived from genomic experiments. *Nucleic acids research*, 40(20):e158–e158, 2012.

[47] Francis D Gibbons and Frederick P Roth. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome research*, 12(10):1574–1581, 2002.

[48] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[49] Bjorn Goemann, Edgar Wingender, and Anatolij Potapov. An approach to evaluate the topological significance of motifs and other patterns in regulatory networks. *BMC Systems Biology*, 3(1):53+, 2009.

[50] Christopher Greenman, Philip Stephens, Raffaella Smith, Gillian L Dalgliesh, Christopher Hunter, Graham Bignell, Helen Davies, Jon Teague, Adam Butler, Claire Stevens, et al. Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–158, 2007.

[51] Roger Guimera and Luis A Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005.

[52] Roger Guimerà, Marta Sales-Pardo, and Luís A. Nunes Amaral. Module identification in bipartite and directed networks. *Phys. Rev. E*, 76(3):036102, 2007.

[53] Benjamin Haibe-Kains, Catharina Olsen, Amira Djebbari, Gianluca Bontempi, Mick Correll, Christopher Bouton, and John Quackenbush. Predictive networks: a flexible, open source, web application for integration and analysis of human gene networks. *Nucleic acids research*, 40(D1):D866–D875, 2012.

[54] Jing-Dong J Han, Nicolas Bertin, Tong Hao, Debra S Goldberg, Gabriel F Berriz, Lan V Zhang, Denis Dupuy, Albertha JM Walhout, Michael E Cusick, Frederick P Roth, et al. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, 430(6995):88–93, 2004.

[55] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *cell*, 100(1):57–70, 2000.

[56] Carmen Hernando, Mercè Mora, Ignacio M Pelayo, and Carlos Seara. Some structural, metric and convex properties on the boundary of a graph. *Electronic Notes in Discrete Mathematics*, 24:203–209, 2006.

[57] John Hopcroft, Omar Khan, Brian Kulis, and Bart Selman. Tracking evolving communities in large linked networks. *Proceedings of the national academy of sciences of the United States of America*, 101(Suppl 1):5249–5253, 2004.

[58] William Hoppitt, Neeltje J Boogert, and Kevin N Laland. Detecting social transmission in networks. *Journal of Theoretical Biology*, 263(4):544–555, 2010.

[59] Katsuhisa Horimoto and Hiroyuki Toh. Statistical estimation of cluster boundaries in gene expression profile data. *Bioinformatics*, 17(12):1143–1151, 2001.

[60] Steve Horvath and Jun Dong. Geometric interpretation of gene coexpression network analysis. *PLoS computational biology*, 4(8):e1000117, 2008.

[61] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13, 2009.

[62] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9), 2010.

[63] T. Ideker and N. J. Krogan. Differential network biology. *Mol Syst Biol*, 8:565, 2012.

[64] Trey Ideker and Roded Sharan. Protein networks in disease. *Genome research*, 18(4):644–652, 2008.

[65] S. Imoto, S. Kim, T. Goto, S. Miyano, S. Aburatani, K. Tashiro, and S. Kuhara. Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *Journal of bioinformatics and computational biology*, 1(2):231–52, 2003.

[66] Piers J. Ingram, Michael P. Stumpf, and Jaroslav Stark. Network motifs: structure does not determine function. *BMC genomics*, 7(1):108+, 2006.

[67] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

[68] Hawoong Jeong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

[69] Daxin Jiang, Chun Tang, and Aidong Zhang. Cluster analysis for gene expression data: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, 16(11):1370–1386, 2004.

[70] Russell G. Jones and Craig B. Thompson. Tumor suppressors and cell metabolism: a recipe for cancer growth. *Genes & Development*, 23(5):537–548, 2009.

[71] Pall F Jonsson, Tamara Cavanna, Daniel Zicha, and Paul A Bates. Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. *BMC bioinformatics*, 7(1):2, 2006.

[72] A. Kamburov, K. Pentchev, H. Galicka, C. Wierling, H. Lehrach, and R. Herwig. Consensuspathdb: toward a more complete picture of cell biology. *Nucleic acids research*, 39(Database issue):D712–7, 2011.

[73] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[74] Nadav Kashtan and Uri Alon. Spontaneous evolution of modularity and network motifs. *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13773–13778, 2005.

[75] Steven M Kay. Fundamentals of statistical signal processing: detection theory. 1998.

[76] Brian P Kelley, Bingbing Yuan, Fran Lewitter, Roded Sharan, Brent R Stockwell, and Trey Ideker. Pathblast: a tool for alignment of protein interaction networks. *Nucleic acids research*, 32(suppl 2):W83–W88, 2004.

[77] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[78] Samuel Kerrien, Yasmin Alam-Faruque, Bruno Aranda, I Bancarz, Alan Bridge, C Derow, Emily Dimmer, Marc Feuermann, A Friedrichsen, R Huntley, et al. Intactopen source resource for molecular interaction data. *Nucleic acids research*, 35(suppl 1):D561–D565, 2007.

[79] Hiroaki Kitano. Computational systems biology. *Nature*, 420(6912):206–210, 2002.

[80] Hiroaki Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–1664, 2002.

[81] Johannes F. Knabe, Chrystopher L. Nehaniv, and Maria J. Schilstra. Do motifs reflect evolved function?–No convergent evolution of genetic regulatory network subgraph topologies. *Bio Systems*, 94(1-2):68–74, 2008.

[82] Arun Konagurthu and Arthur Lesk. On the origin of distribution patterns of motifs in biological networks. *BMC Systems Biology*, 2(1):73+, 2008.

[83] A. Kraskov, H. Stogbauer, and P. Grassberger. Estimating mutual information. *Phys. Rev. E*, 69(6):066138, 2004.

[84] Kevin N Laland, John Odling-Smee, and Sean Myles. How culture shaped the human genome: bringing genetics and the human sciences together. *Nature Reviews Genetics*, 11(2):137–148, 2010.

[85] Renaud Lambiotte, J-C Delvenne, and Mauricio Barahona. Laplacian dynamics and multiscale modular structure in networks. *arXiv preprint arXiv:0812.1770*, 2008.

[86] G. R. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and W. S. Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–35, 2004.

[87] Matthieu Latapy, Clémence Magnien, and Nathalie Del Vecchio. Basic notions for the analysis of large two-mode networks. *Social Networks*, 30(1):31–48, 2008.

[88] D.-S. Lee, J. Park, K. A. Kay, N. A. Christakis, Z. N. Oltvai, and A.-L. Barabási. The implications of human metabolic network topology for disease comorbidity. *Proc Natl Acad Sci*, 105(29):9880–9885, 2008.

[89] Homin K. Lee, Amy K. Hsu, Jon Sajdak, Jie Qin, and Paul Pavlidis. Coexpression analysis of human genes across many microarray data sets. *Genome Research*, 14(6):1085–1094, 2004.

[90] Marius Leordeanu and Martial Hebert. A spectral technique for correspondence problems using pairwise constraints. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1482–1489. IEEE, 2005.

[91] Ai Li and Steve Horvath. Network neighborhood analysis with the multi-node topological overlap measure. *Bioinformatics*, 23(2):222–231, 2007.

[92] Huai Li, Yu Sun, and Ming Zhan. Exploring pathways from gene co-expression to network dynamics. 541, 2008.

[93] Jie Li, Anne E.G. Lenferink, Yinghai Deng, Catherine Collins, Qinghua Cui, Enrico O. Purisima, Maureen D. O'Connor-McCourt, and Edwin Wang. Identification of high-quality cancer prognostic markers and metastasis network modules. *Nature Communications*, 1:34–, 2010.

[94] Xiaoping Liu, Zhi-Ping Liu, Xing-Ming Zhao, and Luonan Chen. Identifying disease genes and module biomarkers by differential interactions. *Journal of the American Medical Informatics Association*, 19(2):241–248, 2012.

[95] K. Lo, A. E. Raftery, K. M. Dombek, J. Zhu, E. E. Schadt, R. E. Bumgarner, and K. Y. Yeung. Integrating external biological knowledge in the construction of regulatory networks from time-series expression data. *BMC systems biology*, 6:101, 2012.

[96] Sherene Loi, Benjamin Haibe-Kains, Christine Desmedt, Françoise Lallemand, Andrew M Tutt, Cheryl Gillet, Paul Ellis, Adrian Harris, Jonas Bergh, John A Foekens, et al. Definition of clinically distinct molecular subtypes in estrogen receptor–positive breast carcinomas through genomic grade. *Journal of clinical oncology*, 25(10):1239–1246, 2007.

[97] Ying Lu and Jiawei Han. Cancer classification using gene expression data. *Information Systems*, 28(4):243–268, 2003.

[98] Hong W. Ma and An P. Zeng. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19(2):270–7, 2003.

[99] H Craig Mak, Mike Daly, Bianca Gruebel, and Trey Ideker. Cellcircuits: a database of protein network models. *Nucleic acids research*, 35(suppl 1):D538–D545, 2007.

[100] Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, Gustavo Stolovitzky, et al. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804, 2012.

[101] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7 Suppl 1:S7, 2006.

[102] Adam A Margolin, Kai Wang, Wei Keat Lim, Manjunath Kustagi, Ilya Nemenman, and Andrea Califano. Reverse engineering cellular networks. *Nature Protocols*, 1(2):662–671, 2006.

[103] Florian Markowetz and Rainer Spang. Inferring cellular networks–a review. *BMC bioinformatics*, 8(Suppl 6):S5, 2007.

[104] Maria S Massa, Monica Chiogna, and Chiara Romualdi. Gene set analysis exploiting the topology of a pathway. *BMC systems biology*, 4(1):121, 2010.

[105] Gnen Mehmet and Alpaydin Ethem. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12(July):2211–2268, 2011.

[106] M. Meila. Comparing clusterings—an information based distance. *J Multivariate Analysis*, 98(5):873–895, 2007.

[107] C. Men, J. Wang, Y. M. Qin, B. Deng, and X. L. Wei. Characterizing electrical signals evoked by acupuncture through complex network mapping: a new perspective on acupuncture . *Computer Methods and Programs in Biomedicine*, 104(3):498–504, 2011.

[108] Franziska Michor, Yoh Iwasa, and Martin A Nowak. Dynamics of cancer progression. *Nature Reviews Cancer*, 4(3):197–205, 2004.

[109] Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, and Uri Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, 2004.

[110] Melanie Mitchell. Complex systems: Network thinking. *Artificial Intelligence*, 170(18):1194–1212, 2006.

[111] James Moody and Douglas R White. Structural cohesion and embeddedness: A hierarchical concept of social groups. *American Sociological Review*, pages 103–127, 2003.

[112] S. Mostafavi and Q. Morris. Fast integration of heterogeneous data sources for predicting gene function with limited annotation. *Bioinformatics*, 26(14):1759–65, 2010.

[113] Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878, 2010.

[114] I. A. Nepomuceno-Chamorro, J. S. Aguilar-Ruiz, and J. C. Riquelme. Inferring gene regression networks with model trees. *BMC bioinformatics*, 11:517, 2010.

[115] T. Nepusz, H. Yu, and A. Paccanaro. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat Methods*, 9(5):471–2, 2012.

[116] M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2), 2004.

[117] Mark Newman. *Networks: an introduction.* Oxford University Press, 2010.

[118] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[119] Mark EJ Newman and Juyong Park. Why social networks are different from other types of networks. *Physical Review E*, 68(3):036122, 2003.

[120] M. C. Nicoletti, J. R Jr. Bertini, M. M. Tanizaki, T. C. Zangirolami, V. M. Goncalves, A. C. Horta, and R. C. Giordano. On-line prediction of the feeding phase in high-cell density cultivation of rE. coli using constructive neural networks. *Computer Methods and Programs in Biomedicine*, 111(1):228–248, 2013.

[121] Martin A Nowak, Franziska Michor, Natalia L Komarova, and Yoh Iwasa. Evolutionary dynamics of tumor suppressor gene inactivation. *Proceedings of the National Academy of Sciences of the United States of America*, 101(29):10635–10638, 2004.

[122] Alberto Ochoa, Julio Ponce, Rubén Jaramillo, Francisco Ornelas, Alberto Hernandez, Daniel Azpeitia, Arturo Elías, and Arturo Hernández. Analysis of cyber-bullying in a virtual social networking. In *HIS*, pages 229–234, 2011.

[123] Yoshifumi Okada, Takehiko Sahara, Satoru Ohgiya, and Tomomasa Nagashima. Detection of cluster boundary in microarray data by reference to mips functional catalogue database. *GIW2005*, 2005.

[124] Gabriel Östlund, Mats Lindskog, and Erik LL Sonnhammer. Network-based identification of novel cancer genes. *Molecular & Cellular Proteomics*, 9(4):648–655, 2010.

[125] Gergely Palla, Albert-László Barabási, and Tamás Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, 2007.

[126] Juyong Park and Albert-László Barabási. Distribution of node characteristics in complex networks. *Proc Natl Acad Sci*, 104(46):17916–17920, 2007.

[127] Georgios A. Pavlopoulos, Charalampos N Moschopoulos, Sean D. Hooper, Reinhard Schneider, and Sophia Kossida. jclust: A clustering and visualization toolbox. *Bioinformatics*, 2009.

[128] Yudi Pawitan, Judith Bjöhle, Lukas Amler, Anna-Lena Borg, Suzanne Egyhazi, Per Hall, Xia Han, Lars Holmberg, Fei Huang, Sigrid Klaar, et al. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Research*, 7(6):R953, 2005.

[129] Erin D Pleasance, R Keira Cheetham, Philip J Stephens, David J McBride, Sean J Humphray, Chris D Greenman, Ignacio Varela, Meng-Lay Lin, Gonzalo R Ordóñez, Graham R Bignell, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(7278):191–196, 2010.

[130] Kornelia Polyak. Heterogeneity in breast cancer. *The Journal of clinical investigation*, 121(10):3786, 2011.

[131] Robert J Prill, Pablo A Iglesias, and Andre Levchenko. Dynamic properties of network motifs contribute to biological network organization. *PLoS Biol*, 3(11):e343, 2005.

[132] P. Qiu, A. J. Gentles, and S. K. Plevritis. Reducing the computational complexity of information theoretic approaches for reconstructing gene regulatory networks. *Journal of computational biology : a journal of computational molecular cell biology*, 17(2):169–76, 2010.

[133] Helmut Ratschek and JON ROKNE. Exact and optimal convex hulls in 2d. *International Journal of Computational Geometry & Applications*, 10(02):109–129, 2000.

[134] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, 2002.

[135] Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74(1):016110, 2006.

[136] Luke Rendell, Robert Boyd, Daniel Cownden, Marquist Enquist, Kimmo Eriksson, Marc W Feldman, Laurel Fogarty, Stefano Ghirlanda, Timothy Lillicrap, and Kevin N Laland. Why copy others? insights from the social learning strategies tournament. *Science*, 328(5975):208–213, 2010.

[137] Ryan M. Rifkin and Ross A. Lippert. Notes on regularized least-squares. Technical report, MIT, 2007.

[138] Alexander W Rives and Timothy Galitski. Modular organization of cellular networks. *Proceedings of the National Academy of Sciences*, 100(3):1128–1133, 2003.

[139] Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proc Natl Acad Sci*, 105(4):1118–1123, 2008.

[140] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[141] Jianhua Ruan, Angela K Dean, and Weixiong Zhang. A general co-expression network-based approach to gene expression analysis: comparison and applications. *BMC systems biology*, 4(1):8, 2010.

[142] Marta Sales-Pardo, Roger Guimerà, André A. Moreira, and Luís A. Nunes Amaral. Extracting the hierarchical organization of complex systems. *Proc Natl Acad Sci*, 104(39):15224–15229, 2007.

[143] Miquel Salicru, Sergi Vives, and Tian Zheng. Inferential clustering approach for microarray experiments with replicated measurements. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 6(4):594–604, 2009.

[144] Lawrence K. Saul and Sam T. Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *J. Mach. Learn. Res.*, 4(Dec.):119–155, 2003.

[145] Charles L Sawyers. The cancer biomarker problem. *Nature*, 452(7187):548–552, 2008.

[146] Jan Schellenberger, Junyoung Park, Tom Conrad, and Bernhard Palsson. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*, 11(1):213+, 2010.

[147] Thomas Schlitt and Alvis Brazma. Current approaches to gene regulatory network modelling. *BMC bioinformatics*, 8(Suppl 6):S9, 2007.

[148] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature genetics*, 34(2):166–76, 2003.

[149] Eran Segal, Nir Friedman, Daphne Koller, and Aviv Regev. A module map showing conditional activity of expression modules in cancer. *Nature genetics*, 36(10):1090–1098, 2004.

[150] Anurag Sharma and Christian W Omlin. Determining cluster boundaries using particle swarm optimization. *Enformatika*, 15, 2006.

[151] Jianbo Shi and Jitendra Malik. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[152] Victor Spirin, Mikhail S. Gelfand, Andrey A. Mironov, and Leonid A. Mirny. A metabolic network in the evolutionary context: Multiscale structure and modularity. *Proc Natl Acad Sci*, 103(23):8774–8779, 2006.

[153] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. The cancer genome. *Nature*, 458(7239):719–724, 2009.

[154] Alexander Sturn, John Quackenbush, and Zlatko Trajanoski. Genesis: cluster analysis of microarray data. *Bioinformatics*, 18(1):207–208, 2002.

[155] Gang Su, Allan Kuchinsky, John H. Morris, David J. States, and Fan Meng. Glay: community structure analysis of biological networks. *Bioinformatics*, 26(24):3135–3137, 2010.

[156] Chayant Tantipathananandh, Tanya Berger-Wolf, and David Kempe. A framework for community identification in dynamic social networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 717–726. ACM, 2007.

[157] Ian W Taylor, Rune Linding, David Warde-Farley, Yongmei Liu, Catia Pesquita, Daniel Faria, Shelley Bull, Tony Pawson, Quaid Morris, and Jeffrey L Wrana. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature biotechnology*, 27(2):199–204, 2009.

[158] L. P. Thomsen, U. M. Weinreich, D. S. Karbing, Jensen V. G. Helbo, M. Vuust, J. B. Frorkjaer, and S. E. Rees. Can computed tomography classifications of chronic obstructive pulmonary disease be identified using Bayesian networks and clinical data? . *Computer Methods and Programs in Biomedicine*, 110(3):361–368, 2013.

[159] Ali Torkamani and Nicholas J Schork. Identification of rare cancer driver mutations by network reconstruction. *Genome research*, 19(9):1570–1578, 2009.

[160] K. Tsuda, H. Shin, and B. Scholkopf. Fast protein classification with multiple networks. *Bioinformatics*, 21 Suppl 2:ii59–65, 2005.

[161] Alfred Ultsch. *U*-matrix: a tool to visualize clusters in high dimensional data.* Fachbereich Mathematik und Informatik, 2003.

[162] L. J. P. van der Maaten, E. O. Postma, and H. J. van den Herik. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10:1–41, 2009.

[163] P. van Nes, D. Bellomo, M. J. T. Reinders, and D. de Ridder. Stability from Structure: Metabolic Networks Are Unlike Other Biological Networks. *EURASIP Journal on Bioinformatics and Systems Biology*, 2009, 2009.

[164] Laura J van't Veer, Hongyue Dai, Marc J Van De Vijver, Yudong D He, Augustinus AM Hart, Mao Mao, Hans L Peterse, Karin van der Kooy, Matthew J Marton, Anke T Witteveen, et al. Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871):530–536, 2002.

[165] A. Vázquez, R. Dobrin, D. Sergi, J. P. Eckmann, Z. N. Oltvai, and A. L. Barabási. The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *Proc Natl Acad Sci*, 101(52):17940–17945, 2004.

[166] Dennis Vitkup, Peter Kharchenko, and Andreas Wagner. Influence of metabolic network structure and function on enzyme evolution. *Genome Biology*, 7(5):R39+, 2006.

[167] Bert Vogelstein and Kenneth W Kinzler. Cancer genes and the pathways they control. *Nature medicine*, 10(8):789–799, 2004.

[168] F. Wang and C. Zhang. Label propagation through linear neighborhoods. *IEEE Transactions on Knowledge and Data Engineering*, 20(1):55–67, 2008.

[169] X. Wang, N. Gulbahce, and H. Yu. Network-based methods for human disease gene prediction. *Brief Funct Genomics*, 10(5):280–93, 2011.

[170] Yixin Wang, Jan GM Klijn, Yi Zhang, Anieta M Sieuwerts, Maxime P Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E Meijer-van Gelder, Jack Yu,

et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365(9460):671–679, 2005.

[171] Yu-Chao Wang and Bor-Sen Chen. A network-based biomarker approach for molecular investigation and diagnosis of lung cancer. *BMC medical genomics*, 4(1):2, 2011.

[172] Z. Wang, B. Zineddin, J. Liang, N. Zeng, Y. Li, M. Du, J. Cao, and X. Liu. A novel neural network approach to cDNA microarray image segmentation. *Computer Methods and Programs in Biomedicine*, 111(1):189–198, 2013.

[173] D. Warde-Farley, S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, C. T. Lopes, A. Maitland, S. Mostafavi, J. Montojo, Q. Shao, G. Wright, G. D. Bader, and Q. Morris. The genemania prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research*, 38(Web Server issue):W214–20, 2010.

[174] Sebastian Wernicke and Florian Rasche. Fanmod: a tool for fast network motif detection. *Bioinformatics*, 22(9):1152–1153, 2006.

[175] Richard Wooster and Kurtis E Bachman. Catalogue, cause, complexity and cure; the many uses of cancer genome sequence. *Current opinion in genetics & development*, 20(3):336–341, 2010.

[176] Z. Wu, X. Zhao, and L. Chen. Identifying responsive functional modules from protein-protein interaction network. *Mol Cells*, 27(3):271–7, 2009.

[177] Takuji Yamada and Peer Bork. Evolution of biomolecular networks - lessons from metabolic and protein interactions. *Nature Reviews Molecular Cell Biology*, 10(11):791–803, 2009.

[178] Yoshihiro Yamanishi, J-P Vert, and Minoru Kanehisa. Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics*, 20(suppl 1):i363–i370,

2004.

[179] Bo Yang, William K Cheung, and Jiming Liu. Community mining from signed social networks. *Knowledge and Data Engineering, IEEE Transactions on*, 19(10):1333–1348, 2007.

[180] M. K. Yeung, J. Tegner, and J. J. Collins. Reverse engineering gene networks using singular value decomposition and robust regression. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9):6163–8, 2002.

[181] Andy M Yip and Steve Horvath. Gene network interconnectedness and the generalized topological overlap measure. *BMC bioinformatics*, 8(1):22, 2007.

[182] Golan Yona, William Dirks, and Shafquat Rahman. Comparing algorithms for clustering of expression data: how to assess gene clusters. In *Computational Systems Biology*, pages 479–509. Springer, 2009.

[183] Haiyuan Yu, Philip M Kim, Emmett Sprecher, Valery Trifonov, and Mark Gerstein. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS computational biology*, 3(4):e59, 2007.

[184] Mikhail Zaslavskiy, Francis Bach, and Jean-Philippe Vert. Global alignment of protein–protein interaction networks by graph matching methods. *Bioinformatics*, 25(12):i259–1267, 2009.

[185] Bin Zhang, Steve Horvath, et al. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1):1128, 2005.

[186] S. Q. Zhang, W. K. Ching, N. K. Tsing, H. Y. Leung, and D. Guo. A new multiple regression approach for the construction of genetic regulatory networks. *Artificial intelligence in medicine*, 48(2-3):153–60, 2010.

[187] Jing Zhao, Guo H. Ding, Lin Tao, Hong Yu, Zhong H. Yu, Jian H. Luo, Zhi W. Cao, and Yi X. Li. Modular co-evolution of metabolic networks. *BMC Bioinformatics*, 8:311, 2007.

[188] Jing Zhao, Hong Yu, Jian H. Luo, Zhi W. Cao, and Yi X. Li. Hierarchical modularity of nested bow-ties in metabolic networks. *BMC Bioinformatics*, 7:386, 2006.

[189] X. J. Zhou, M. C. Kao, H. Huang, A. Wong, J. Nunez-Iglesias, M. Primig, O. M. Aparicio, C. E. Finch, T. E. Morgan, and W. H. Wong. Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nature biotechnology*, 23(2):238–43, 2005.

[190] Xuefeng Zhou, Ramanjulu Sunkar, Hailing Jin, Jian-Kang Zhu, and Weixiong Zhang. Genome-wide identification and analysis of small rnas originated from natural antisense transcripts in oryza sativa. *Genome Research*, 19(1):70–78, 2009.

[191] Xiaojin Zhu and Andrew B. Goldberg. *Introduction to Semi-Supervised Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool, 2009.