COMPUTATIONAL METHODS FOR DISCOVERING FUNCTIONAL MODULES FROM PROTEIN INTERACTION NETWORKS

by

Phuong Dao

M.Sc., Simon Fraser, 2009

B.C.S., Carleton University, 2006

A Thesis submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

in the
School of Computing Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Summer 2012

APPROVAL

Name: Phuong Dao

Degree: Doctor of Philosophy

Title of Thesis: Computational Methods For Discovering Functional Modules From Protein Interaction Networks

Examining Committee: Dr. Andrei Bulatov
Chair

Dr. Martin Ester, Co-senior Supervisor

Dr. Cenk Sahinalp, Co-senior Supervisor

Dr. Artem Cherkasov, Supervisor

Dr. Colin Collins, Supervisor

Dr. Peter Unrau, SFU Examiner

Dr. Roded Sharan, External Examiner

Date Approved: 9 August 2012

Partial Copyright Licence

Abstract

Recent studies have suggested that molecular interaction networks within cells could be decomposed into different subnetworks of molecules that are involved in common biological processes. Such subnetworks are known as pathways, protein complexes or, in general, as functional modules. Many computational methods have been developed to discover functional modules based on various hypotheses. For example, network motifs are abundant subnetworks in natural networks but not random networks with similar global properties. Networks motifs have been utilized for comparing protein-protein interaction (PPI) networks of various organisms and for assessing the random models in terms of capturing the global and local properties of PPI networks. In another example, subnetwork markers are connected subnetworks from PPI networks in which member gene expressions correlate with labels of the samples. Such subnetwork markers could be used as predictors for phenotype of the samples such as the disease statuses of the patients.

In this dissertation, I first present novel computational methods for discovering network motifs that use the confidence scores from protein interactions. Since there are many false positives and false negatives in the current binary PPI networks, utilizing confidence scores could result in better network motifs. I have used this algorithm to compare PPI networks of prokaryotic unicellular, eukaryotic unicellular and multicellular organisms. Later, I present two efficient and optimal computational approaches for identifying subnetwork markers. The first one utilizes confidence scores from PPIs. And the second one is a randomized algorithm for discovering the subnetworks markers with the best predicting performance. I have applied these algorithms to predict disease statuses of colon cancer and breast cancer patients and treatment outcomes of a combinatory therapy for a breast cancer study.

Acknowledgments

First of all, I am deeply indebted to my senior supervisors Dr. Martin Ester and Dr. Sulleyman Cenk Sahinalp, for their endless support, encouragement, guidance through my research. Dr. Sahinalp is not only a wonderful supervisor but also a good friend. During my study at SFU, they have spent a lot of time, efforts and finance on supervising my works, improving my communication skills and sending me to conferences. I would like to thank my other supervisors Dr. Colin Collins and Dr. Artem Cherkasov. I have had a fruitful collaboration and wonderful time working at Dr. Collins's lab. Dr. Cherkasov is a great and very sincere supervisor who introduced me to Computational Biochemistry world. I would also like to thank Dr. Funda Ergun for her support when I started my study at SFU.

I particularly thank all of my current and previous lab mates: Nazanin Bakhshi, Dr. Hamid Chitsaz, Mohsen Jamali, Faraz Hach, Iman Hajirasouliha, Farhad Hormozdiari, Dr. Fereydoun Hormozdiari, Bo Hu, Dr. Emre Karakoc, Yen-Yi Lin, Andrew Mcpherson, Samaneh Moghaddam, Ibrahim Numanagic, Dr. Raheleh Salari, Dr. Alexander Schoenhuth, Lucas Swanson, Kendric Wang, Deniz Yorukoglu. I really appreciate many thoughtful research discussions and fun hang out times with them. I also thank other friends from the department: Amir Avani, Navid Imani, Dr. Hossein Jowhari, Hossein Maserat, Nhi Nguyen, Dr. Murray Patterson, Shahab Tasharrofi, Carrie Wang, Winona Wu.

Finally and most importantly, I wish to thank my parents and my sister for their unconditional support. Without them, I would haven't got this far.

Contents

Chapter 1

Introduction

1.1 Motivations

In classical molecular biology, the functions of a cell are studied through the functions of its components such as DNAs, RNAs, proteins, metabolites and other organic molecules. However, the functions of a biological system might not be due to the functions of individual components [78]. In fact, most of the functions of a cell usually arise from complex interactions among the consitutent molecules [78]. For example, in yeast there is a signal transduction system that converts the detection of a pheromone into the act of mating. However, no single molecule in the system is known for for amplifying the input signal provided by the pheromone molecule [78]. Thus, many recent studies have focused on understanding the structure and dynamics of the interaction network of molecules.

High-throughput technologies have recently provided comprehensive mappings of cellular molecular interaction networks such as protein-protein interaction (PPI) networks. Such interactomes have helped to understand the physical architecture of the cells . For instance, many studies have derived pairwise protein interactions [187, 223, 197, 20] or protein complexes [23, 57, 112, 183] . Other types of physical interactions that have been mapped systematically include transcriptional protein-DNA interactions [159, 91] and kinase-substrate interactions [153, 125]. Many databases have also been developed to record these experimentally derived protein interactions together with predicted ones: BIND [9], BioGRID [185], DIP [165], IntAct [107], MINT [25], MIPS [144], HPRD [108].

Recent studies have suggested that molecular interaction networks within cells could be decomposed into different subnetworks of molecules that are involved in common biological

processes [78, 14]. Such sets of molecules are known as signalling pathways, protein complexes or in general functional modules. One of the well known protein complex is the ribosome, a large molecule which is responsible for formation of proteins from individual amino acids using messenger RNA.

Many computational tools are developed to extract functional modules from only PPI data such as network motifs. Network motifs are small subgraphs which occurs more frequent in PPI networks than random networks of similar global properties such as degree distribution. Studies on these network motifs have yielded insights into the information processing in biological networks [79, 128].

Other computational methods have been developed for discovering functional modules by integrating network topology data (PPIs) with other genomic profiles. Such tools can identify active functional modules which are subnetworks from a PPI network of which member genes are differentially expressed in a set of conditions/samples. Specific activation of member genes can be detected through gene/protein expression profiles. Recent computational efforts on active functional modules have focused on subnetwork markers of which the member gene activities correlate with the labels of the samples. For example, tumour samples taken from a cancer study usually come with clinical labels such as subtypes of the cancers. Recent studies show that subnetwork markers could be used as predictors for the disease status of the patients and provide a comprehensive view of the molecular mechanisms underlying pathology.

1.2 Contributions

The problems of discovering functional modules such as network motifs and subnetwork markers can be formulated as combinatorial optimization problems, which are typically solved through heuristics, exact/approximation algorithms, or machine learning methods. In this dissertation we will develop a randomized approximation algorithm for discovering network motifs, an exact enumeration algorithm and a randomized algorithm for identifying subnetwork markers. These algorithms not only have efficient running time and space but also have provable performance on real data sets. In what follows, we summarize our contributions on developing these algorithms:

- Previous computational methods for network motif discovery are based on binary PPI networks derived from high throughput experiments. However, there is a high false

discovery rate in these experiments. Recently, many groups for example STRING database [93] have assigned the confidence scores to the protein interactions by integrating many large scale experiments and other genomic evidences such as mRNA coexpression. Utilizing the confidence scored PPI networks can yield more reliable network motifs. Subgraph counting or counting the number of isomorphic subnetworks from a PPI network of a given query subnetwork has been the main algorithmic tool for discovering network motifs. We will present a novel randomized approximation algorithm based on color coding to count the occurrences of a given subgraph in a confidence scored PPI network. In other words, we count the weight of non-induced isomorphic subnetworks of a given tree T with k vertices in a confidence scored PPI network G with n vertices with polynomial running time with n, provided $k = O(logn)$ for a given error probability and an approximation ratio. We also introduce two definitions for the weight of an subnetwork in which is isomorphic to the query tree T.

- Similar to computational tools for discovering network motifs, current computational methods for identifying subnetwork markers have not made use of confidence scored PPI networks. We present a novel computational strategy wDCB for exhaustive enumeration of dense subnetwork biclusters of which the average edge weight is more than a given threshold and member genes are differentially expressed in sufficiently many, but not necessarily all the samples. Conditions on partial differential expression model the fact that samples from cancer patients usually divide into many different subgroups. These subnetwork markers can correspond to dysregulated pathways in many, but not necessarily all samples under consideration.

- Current computational methods for identifying subnetworks do not provide both efficient running time and optimality of subnetwork markers. The authors from the seminal works [31] and others propose heuristic methods which do not guarantee the optimality of the solution for marker selection. Other exact approaches based on branch and bound or exhaustive enumeration as introduced in the above can yield an optimal solution under some fixed set of parameters; however, their worst-case running time can be super-polynomial (and hence intractable). Thus, we introduce a novel and efficient randomized algorithm to compute optimally discriminative subnetworks for classification of samples from different classes. The discriminative score is calculated as the difference between the total distance between samples from different classes and

the total distance between samples from the same class. Our algorithm is based on the color-coding paradigm [6] which allows for identifying the optimally discriminative subnetwork markers for any given error probability. Since the running time of our algorithm is a logarithmic function of the error probability, we can set the error probability to a small value, close to zero, while the running time does not increase much. When the maximum size of a subnetwork is $k = O(logn)$, where n is the size of the network, we have a polynomial time algorithm with a fixed error probability.

1.3 Organization of This Thesis

The dissertation is organized as follows:

- In the second chapter, we look into experiments to derive protein-protein interactions. Then we discuss various computational methods for assigning confidence scores to PPIs. At the end of the chapter, we introduce two databases for PPI networks that our algorithms are implemented on.

- In the third chapter, we review existing computational approaches for discovering functional modules. We first discuss computational tools solely based on PPI network data. Later we discuss other approaches for integrating network topology and other genomic data such as gene expression and genomic variation profiles. We conclude the chapter with discussion about computational methods for discovering subnetwork markers.

- In Chapter 4, we present our randomized algorithm for counting network motifs in a confidence scored PPI network. Later, we show how to apply the algorithms to compare the weighted PPI networks of various species. We show that there are differences between PPI networks of multicellular and prokaryotic organisms in terms of the occurrences of network motifs while global properties of such networks such as degree distribution and clustering coefficient fail to capture.

- In Chapter 5, we present the wDCB algorithm for complete enumeration of densely connected subnetwork markers. We then show how to apply wDCB on two cancer datasets. We show that the predictive performance of WDCB outperform other

heuristics for deriving subnetwork markers and single gene marker approaches which ranked the discrimative scores of genes based on statistical tests such as t-test.

- In Chapter 6, we present the OPTDIS algorithm for identifying optimal discriminative subnetwork markers. We present its application to predicting chemotherapy response of patients from a breast cancer study. The predictive performances of OptDis is better than all other approaches including wDCB. At the end of the chapter, we examine difference sources of performance improvement of OptDis.

- In the last chapter, we first provide a brief summary of the algorithms presented in this thesis. Later, we discuss the limitations of the current algorithms. At last, we conclude the thesis with future works.

Chapter 2

Protein-Protein Interaction Networks

In this chapter, we will discuss *in vivo* and *in vitro* experiment techniques for deriving PPIs to build PPI networks. Then we look at how to combine these PPI networks with other genomic data to build functional protein interaction networks in which edges denote pairs of proteins with similar functions. We will introduce some databases of PPIs and functional protein interaction networks such as HPRD [108] and STRING [181, 195]. Finally, we look at other types of biological network models.

2.1 Experiments for Deriving Protein-Protein Interactions

Protein-protein interaction data have increased dramatically throughout the last few years. For instance, the systematic identification of pairwise protein interactions [187, 223, 197, 20] or protein complexes [23, 57, 112, 183] has been a widely used strategy for understanding the physical architecture of the cell. Other types of physical interactions that are being mapped systematically include transcriptional protein-DNA interactions [159, 91] and kinase-substrate interactions [153, 125].

There are many experimental techniques for detecting PPIs: Bimolecular Fluorescence Complementation [106], Chemical Crosslinking [27], Co-Immunoprecipitation (CoIP) [53], Tandem Affinity Purification [160] and yeast Two-Hybrid (Y2H) [222] . However, the main high throughput methods for detecting protein interactions probably are CoIP, TAP and

Y2H.

2.1.1 Yeast Two-Hybrid Experiments

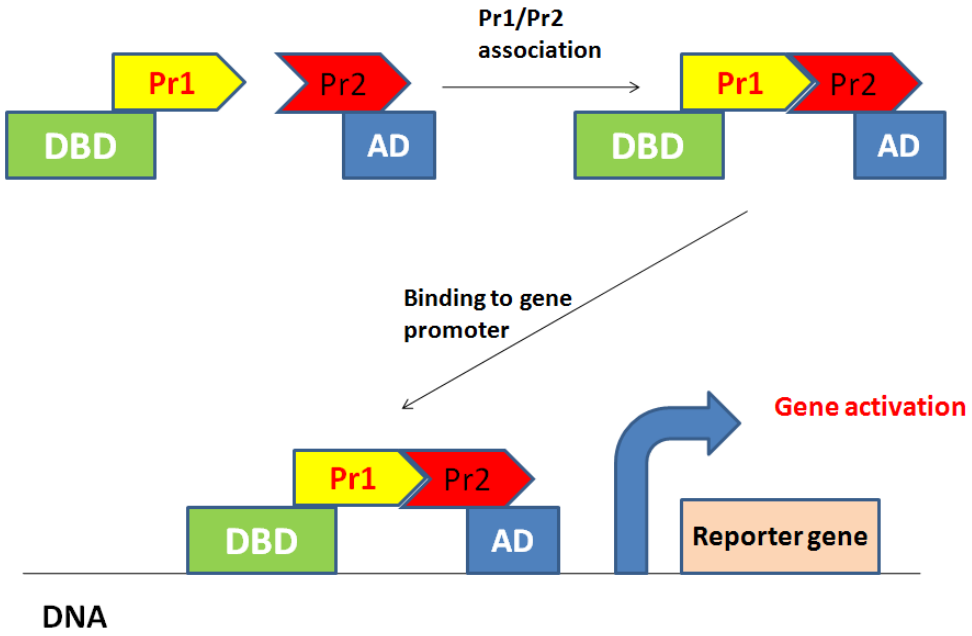


Figure 2.1: A yeast two-hybrid experiment. In this experiment, the presence of interaction between two proteins of interest Pr1 and Pr2 is detected by measuring the mRNA expression of the reporter protein. The physical binding of Pr1 and Pr2 will form a transcription factor that is capable of binding to the promoter region using binding domain (DBD) attached to Pr1 of the reporter gene and activating its transcription using the activating domain (AD) attached to Pr2. Figure is taken from [120].

The Y2H technique is a wet lab technique for detecting physical binding between two proteins [222] or a protein and a DNA molecule [99, 86]. The idea is that the interaction of two considered proteins causes the transcription of a reporter gene. The transcription of a reporter gene can only be activated if a transcription factor is present. This transcription factor consists of two protein domains: a DNA-binding domain (DBD) that is capable of binding to the promoter region of the reporter gene and an activation domain (AD) that is capable of activating the transcription. So if either of these domains is absent then transcription of the reporter gene will be unlikely. Hence, each of two proteins of interests is attached to either the AD or DBD domain. The one attached to DBD in its N-terminus

is called the bait while the one attached to activation domain is called the prey. If there is a physical binding between two proteins of interests, mRNA expression of the reporter gene could be captured or detected using microarray or RNA-Seq technologies. Figure 2.2 illustrates a Y2H experiment. Moreover, the expression level of the reporter gene can be used as a measure of interaction between two protein of interests. Since Y2H experiments are done *in vitro*, they can detect spurious protein interactions which are physical bindings of protein domains outside of cellular environments.

The yeast *S. cerevisiae* is the most common used model organism for high throughput two hybrid experiments. The Gal4 transcription factor domains (DBD and AD) are used to fused to two proteins of interest. The expression of reporter gene LacZ is measured to detect the existence of the fused protein product.

2.1.2 Co-Immunoprecipitation and Tandem Affinity Purification Experiments

CoIP experimental technique is used to detect interactions among a bait protein and other proteins in a cell [118]. First, the bait protein is marked by a tag. The bait protein is put into a cell. Potential interaction partners could bind to the bait protein during this process. Then an antibody which can recognize the tag is used to capture bait protein and precipitate it. Any proteins which already binded to the bait protein are also precipitated. An mass spectrometry experiment is used to detect the presence of the precipitated proteins. Figure 2.2 illustrates a CoIP experiment. Similar to CoIP experiments, the TAP experiments requires two successive steps of protein purification. Recent genome-wide TAP experiments were performed by Krogan [112] and [57] providing updated protein interaction data for yeast *S. cerevisiae*. Since a precipitated protein could bind to another precipitated protein but the bait protein, these experimental techniques can not distinguish direct from indirect PPIs. However, it can be used to detect protein complexes or interactions of multiple proteins. In contrast to Y2H approach, the accuracy of CoIP and TAP are comparable to those of small-scale experiments since the interactions are examined inside the cellular environment.

2.1.3 Scoring Protein Interactions

The above high throughput technologies for discovering PPIs suffer from high false positive rate [40, 216]. It was estimated by the study [203], the number of false interactions returned

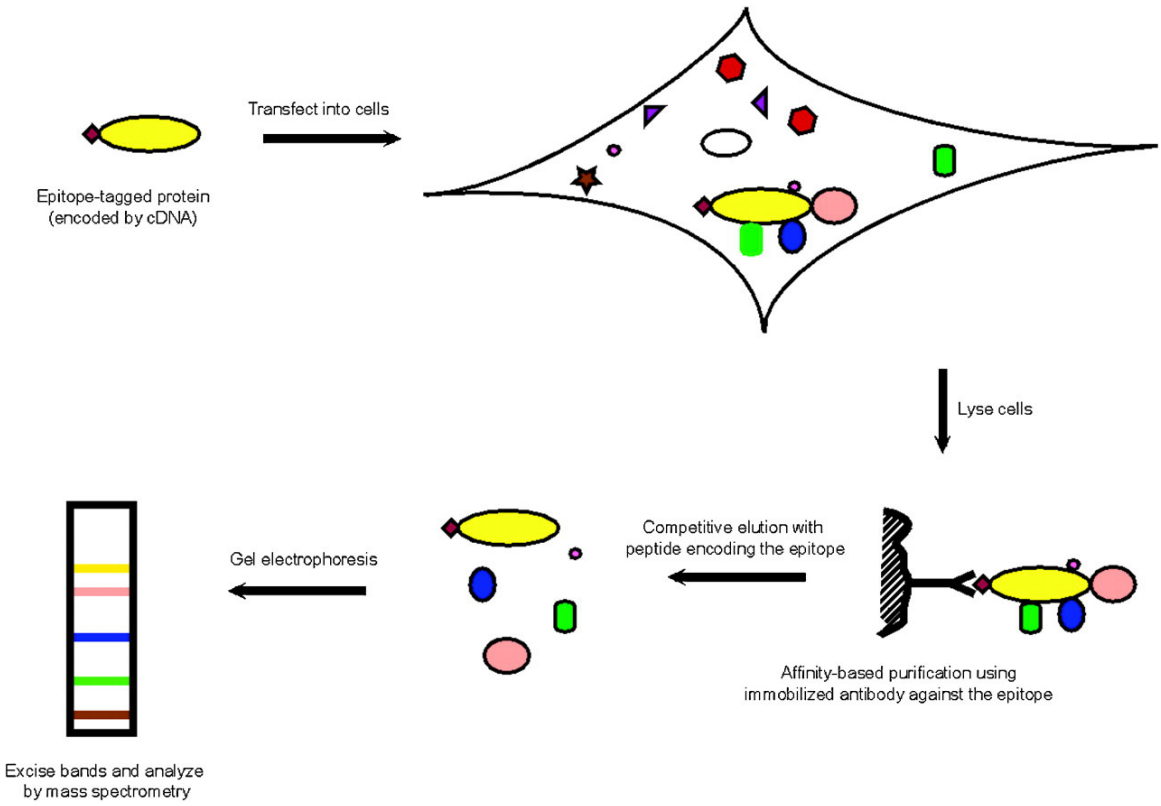


Figure 2.2: A CoIP experiment. A) A protein of interest (the bait) is attached to affinity tag. B) The bait is put into the cells in order to check whether there are bindings between the bait and associated proteins C) The cells are then lysed. D) The lysed protein complex is purified using a pull down assay with an antibody to the tag. E) Protein bands are extracted and digested into small peptide fragments. The peptides are identified using mass spectrometry methods. The protein associated with the bait is determined by comparing its peptide fingerprint against known databases. Figure was taken from [132].

by a Y2H experiment could be up to 50%. The error rate can be reduced by integrating data from various experiments as a false interactions are not be likely to be observed in repeated experiments. The false negative rate of Y2H experiments could be even higher. Reguly et. al. [158] estimated the number of protein interactions in the yeast network is around 30k whereas current Y2H experiments could detect about 20k interactions.

Various computational approaches have been developed to reduce the false positive rates by integrating data from many high throughput experiments. Gavin *et al.* [58] designed a simple approach for deriving the confidence scores for high-throughput interaction data. Suppose we would like to estimate the confidence score for a pair of proteins a and b. Let N be the number of high throughput experiments, $N_{together}$ be the number times both a, b are observed to be pulled out together, N_a (N_b) be the number of times a (b) is pulled out alone. Then the confidence score is estimated as follows:

$$w_{ab} = \log \frac{N_{together} N}{(N_a + 1)(N_b + 1)}$$

A more sophisticated method [194] quantifies the confidence scores for protein-protein interactions using logistic regression as follows. Each gold standard positive or negative interaction is associated to a vector of multiple dimensions that corresponds to experiments from Y2H, TAP, CoIP, large scale and small scale experiments. For each dimension, an interaction is assigned with the confidence score derived from the corresponding experiment. Using logistic regression model on training data, the authors can assign a confidence score for a novel pair of two genes/proteins.

2.2 Functional Protein Association Networks

A functional association network is a generalization of a protein interaction network. There exists an edge between two proteins if they physically interact with each other or they are predicted to interact with each other. Or two proteins are connected by an edge if they share similar functions or are predicted to share similar functions.

Functional associations can be predicted based on the vast and increasing high throughput genomic, transcriptomic and proteomic data. The predictions are possible due to the fact that genes whose products interact physically or in a protein complex often have similar functions [203, 215]. Moreover, genes that exhibit similar patterns of expression [189], synthetic lethality [225], chemical sensitivity [62] often have some similar functions. Previous studies

also reveal shared functions among genes with similar phylogenetic profiles [148] or with similar protein domains [80].

2.3 Databases of PPI and Functional Protein Association Networks

There is an increasing number of databases of protein interaction network with quickly growing number of interactions for various organisms: BIND [9], BioGRID [185], DIP [165], IntAct [107], MINT [25], MIPS [144], HPRD [108]. These databases record experimentally determined protein-protein interactions. Besides, there are many available collections of well studied pathways: CellMap [1], KEGG [101], NCI Pathway Interaction Database [2], Panther [131] and Reactome [36]. Recently, databases of functional association networks such as GeneMANIA [137], Skypainter [219] and STRING [181, 195] have been built and growing at the increasing speed.

In this thesis, we will discuss the Human Protein Interaction Database (HPRD) and the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database in details since they were used in our experiments. Other databases of protein interactions and functional associations of proteins are built based on similar principles.

2.3.0.1 HPRD

As in the latest update [108], there are around 30047 annotated protein sequences. Around 39000 PPIs are documented in HPRD. Experiments for PPIs are divided into three categories namely in vitro, in vivo and yeast two hybrid (Y2H). Sixty percent of PPIs in HPRD are supported by a single experiment whereas around 26% of them are found to have two of the three experimental methods annotated. Overall, in HPRD, 8710 proteins are annotated with at least one PPI, whereas 2015 and 774 proteins have more than 5 or 10 PPIs.

2.3.0.2 STRING

Search Tool for the Retrieval of Interacting Genes (STRING) database [181, 195] has recorded protein functional associations based on various genomic contexts. First, they import protein interactions from high throughput experiments and interactions from co-expression analysis. They also search for genes that are found in close proximity within chromosomes, genes that

jointly encode for a single fusion protein which is a good indicator for functional linkage, gene families which have similar phylogenetic profiles. All of these evidences are good indicators that there are functional linkages among genes/gene products of interests.

Other sources of interactions in STRING are text-mining and interaction transfer among organisms. They parse a large body of scientific texts for example abstracts from PubMed. Using Natural Language Processing, they search for statistically relevant co-occurrences of gene names, and also extract a subset of semantically specified interactions. For the transfer of interactions between organisms, they estimate whether a pair of interacting proteins found conserved in another organism justifies the transfer of the interaction to that other organism.

Each edge in STRING database is assigned with a confidence score. This confidence score is integrated based on confidence scores from various evidences. And it is often higher than the individual sub scores reflecting the increasing confidence when there are various evidences that support the functional association. Suppose that W is final confidence score and W_i's are the confidence scores of the supported evidences. Based on the naive Bayesian framework and the assumption of independence of the evidences, the confidence score W is computed as

$$W = 1 - \prod_i (1 - W_i)$$

The functional interactions stored in STRING are divided into three categories based on their confidence scores: low confidence (scores < 0.4), medium (scores from 0.4 to 0.7) and high (scores > 0.7).

2.4 Other Biological Networks

Gene/mRNA coexpression networks are built from gene expression profiles across many tissue types and experimental conditions. Nodes in the network represent genes and two nodes are connected if the corresponding gene expression profiles are significantly coexpressed. The typical measures for coexpression are Pearson and Spearman correlations.

Gene regulatory networks are built on nodes which are genes, mRNAs and proteins and their interactions include transcription, translation, transcriptional regulation and post-translational reactions. Specifically, the nodes have distinct identities if they correspond to diverse cellular components, and the edges can have two different signs corresponding to

activation and inhibition.

Metabolic networks are designed to capture the metabolism of an organism. That is the biological process that generates essential components such as amino acids, sugars and lipids, and the energy required to synthesize them and to use them in creating proteins and cellular structures. Nodes in metabolic networks are proteins, particles, molecules and metabolites. Edges represent chemical reactions and each edge has labels which are enzymes or genes.

Chapter 3

Functional Module Discovery

Biological networks have been observed to share global properties with other technological and social networks such as the Internet [13, 94, 188]. For example, two nodes in these networks tend to have short distance between them. In many of these networks, there are a few nodes with many more connections than the average node has. Thus, it might indicate that there are some common principles that govern these complex networks which allows the knowledge from well studied non-biological systems to be transferred to characterize complex interaction networks of molecules in the cells.

Recent studies by Milo *et al.* [134] suggested that complex networks could be constructed from subunits (modules) with similar structures or functions. Recent research in biological networks has also suggested that molecular interaction network in a cell could be decomposed into subnetworks with similar functions or in short functional modules [78, 14]. Many hypotheses have been formulated to discover functional modules. Probably, the two most prominent ones are: 1) abundant subnetworks in biological networks but not random networks with similar global properties (network motifs) 2) subnetworks in biological networks in which there many interactions among the component genes (dense subnetworks).

In this chapter, we will discuss related works on computational methods for discovering functional modules from protein interaction networks. In Section 3.2, we will discuss approaches that only take into account of the topology of protein interaction network. In Section 3.3, we will discuss methods that not only take into account of network topology but also other genomic data. At the end of this chapter, we will discuss computational methods for discovering functional modules that exist in a particular condition but not the others and applications in biomarker discovery. Before that, we will discuss graph theory terminologies

that will be discussed throughout this chapter.

3.1 Graph Theory Terminology

Before we delve deeper into each category, we introduce the necessary terminologies. We denote the original network $G = (V, E)$ and for an edge $uv \in E$ $(u \in V, v \in V)$ we denote $0 \leq w(uv) \leq 1$ as the associated confidence score for the interaction. Suppose that $n = |V|$ is the number of vertices in the network. In an unweighted network $w(uv) = 1$ if there is an interaction; 0 otherwise.

Two graphs G and H are said to be **isomorphic** if there is a mapping $f(v)$ from $V(H)$ to $V(G)$ such that $(v, w) \in E(H)$ if and only if $(f(v), f(w)) \in E(G)$. If such a map exists, it is called an **isomorphism** from H to G.

A subgraph H of a graph G is said to be **induced** if, for any pair of vertices x and y of H, xy is an edge of H if and only if xy is an edge of G. In other words, H is an **induced subgraph** of G if it has exactly the edges that appear in G over the same vertex set. If the vertex set of H is the subset S of $V(G)$, then H can be written as $G[S]$ and is said to be **non-induced** by S.

Intuitively, a **tree decomposition** of a graph is the mapping of the graph into a tree where each node of the tree is a subset of vertices of the graph. vertices are adjacent only when the corresponding subtrees intersect. The tree decomposition of a graph is not unique. And tree decompositions are also known as junction trees, clique trees.

Thus, given a graph $G = (V, E)$, a **tree decomposition** is a pair (S, T) where $S = S_1, ..., S_n$ is a family of subsets of V, and T is a tree whose nodes are the subsets S_i, satisfying the following properties:

Each vertex v in V belongs to one of S_i's or the union of all sets S_i equals V.

For each edge (v, w) in the graph, there is a subset S_i such that $v, w \in S_i$.

If S_i and S_j both contain a vertex v, then each S_k of the tree T on the path between S_i and S_j contains v as well.

The width of a tree decomposition (S, T) is the size of its largest set S_i minus one. The **treewidth** of a graph G is the minimum width among all possible tree decompositions of G.

3.2 Using Network Topology

The input for the approaches in this section is a biological network. Edges could be assigned with/without confidence scores. The output of the below algorithms is a list of sets of genes. These sets are usually connected subnetworks in the examined biological networks. Many approaches do not output gene sets which are connected subnetworks, however, their member genes are at short distance from one another based on predefined measurements from network topology.

3.2.1 Network Motifs

The idea to build larger systems by combining smaller subsystems has been applied extensively in many areas of science and engineering. Engineering design of a large system usually is a hierarchical organization of similar subsystems of structures or functions. The subsystems are able to interact and exchange information by adhering to a standard interface. Recently, many research groups have also observed that some subnetworks are abundant in naturally occurring, evolving biological networks [157, 73, 134]. Studying these abundant subnetworks has helped in yielding insights into the information process in biological networks [128, 174].

Milo *et al.* [134] have recently looked at ways in which such networks can be broken down into smaller functional units in order to more easily identify structures within the network. The authors defined a network motif as a subgraph that occurs much more frequently in a network G than one in a random network whose global properties are similar to those of G. Similarly, a subgraph that occurs much less frequently in G in comparison to random networks is called an anti-motif of G. Some of most important motifs are the feed-forward motif and bifan motif (see Figure 3.1). A number of recent studies have proposed mathematical models to study the dynamics of feed-forward motifs [79, 128]. The information processing roles of feed-forward motifs have also been experimentally verified [129, 161, 224].

There are many generalizations of network motifs in various types of biological networks. The authors of [113] suggested reaction motifs in metabolic networks. These reaction motifs are subnetworks within metabolic networks that share similar functional annotations rather than similar topologies. Banks *et al.* [12] proposed network schema which add properties to nodes and edges such that network schema can be wildly applicable to many biological networks. A network schema is a subgraph where genes/proteins can come with functional

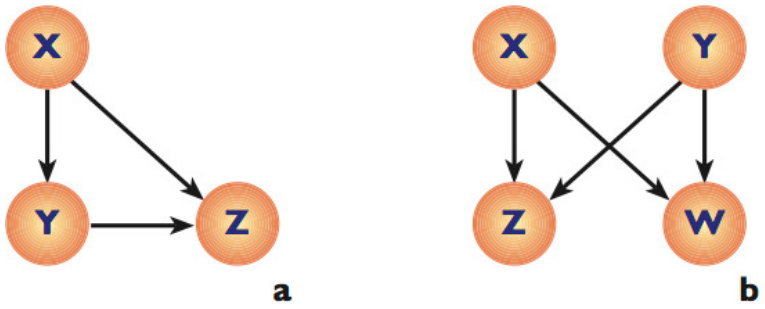


Figure 3.1: Some abundant network motifs in yeast gene regulatory network: a) a feed-forward loop b) a bifan. Figure is taken from [132].

annotations or putative domains, edges could be specified as regulatory relationships for example inhibiting/promoting. And its number of occurrences should be more than expected from a random network. Network motifs could be considered as subgraphs that are present in in a set of biological networks instead of occurring many times within one. Sharan *et al.* [170] generalized the concept of network motifs as subgraphs that are present in PPI networks of many species.

3.2.1.1 Exact Counting

The algorithm in the seminal work [134] is probably the most popular algorithm for counting the occurrences of a particular subgraph S in a network $G = (V, E)$. The algorithm basically uses depth first search to enumerate all the connected non-induced subgraphs of G which have size $|S|$ and are isomorphic to S. There are maximum $O(n^k)$ such subgraphs (when G is a complete graph). The number is then corrected by dividing by the number of automorphisms of S.

In order to assess the significance of the number of occurrences t of S in G, random networks that capture the topological properties of G are generated. For each random network, the number of occurrences of S is computed and we have a distribution of the numbers of occurrences of S in random networks. A statistical test such as t-test is used to compute p-value for the number of occurrences of S in G.

Subgraph counting has been applied to compare the protein interaction networks with networks generated by random models. Such comparison helps to assess the quality of different generative random models that capture the properties of protein interaction network such as degree distribution, betweenness and subgraph distributions. It was argued that

the distribution of subgraphs of up to $k = 5$ vertices in the yeast PPI network is quite different from that of the preferential attachment model [151]. The authors from the same group have also used the distribution of subgraphs of up to $k = 5$ to assess the quality of geometric random models in emulating PPI networks [82, 152]. Hormozdiari *et al.* [84] demonstrated that the subgraph distribution of the preferential attachment model and that of the duplication model for $k \leq 6$ can be substantially different and the seed network of the duplication method could be chosen in a way that its subgraph distribution can be made very similar to that of the available PPI networks including that of yeast.

Grochow and Kellis [67] have improved the running time of this simple counting algorithm significantly using many heuristics. The first one is based on graph or vertex invariants. An invariant is a property that is the same for isomorphic graphs. The number of vertices, the number of edges, and the degree distribution are some common use invariants. If two graphs differ in one of these invariants, they are not isomorphic to each other. The sequence of degrees of the neighbours of a vertex is used. That is when we need to match a vertex v from S to a vertex v' in G, the sequence of degrees of v's neighbours should be a subset of the sequence of degrees of v''s neighbours.

The second heuristic is the use of symmetry breaking to eliminate double counting the automorphisms of an isomorphic subgraph in G. The authors divide the vertices of S into groups. Vertices from the same group could map to each other to create an automorphism. For example, when S is a triangle of three vertices, there is only one group including all the vertices. When S is path of size three, there is two groups, one group includes two vertices of degree one and the other group includes vertex of degree two. The third heuristic is based on the degree distribution of the original network. Since the algorithm discards all the vertices at which all the isomorphic subgraphs are already searched, starting depth first search at a hub or high degree node in the network simply could give more subsets of vertices to look at. The vertices are sorted in the increasing order of their degrees then the degree sequence of their neighbours.

Grochow and Kellis [67] were able to search for all network motifs of size 7 within a couple of hours in yeast PPI network. This a significant improvement over other naive approaches. By combining the counting of smaller network motifs, they could discover a network motif of 29 nodes. This motif consists of two protein complexes that are enriched with chromatin modification and histone acetylation.

Later algorithms for exact counting the copies of a subgraph H in a graph G have

improved the average running time across different subgraphs of the same size. The running time of the algorithms typically depends on a parameter from the subgraph H for example the treewidth, pathwidth and maximum independent set. The authors from [213] compute the number of copies of an H in G in $O^*(2^s n^{k-s+3})$ time where k is the size in terms of vertices in H and s is the size of the maximum independent set of H. Note that the O^* notation omits a $poly(k)$ factor. This algorithm relies on fast algorithms for computing the permanent of a matrix. Another algorithm for exact counting the occurrences of H in G with running time depending on the treewidth of H was proposed by the authors from [54]. This algorithm runs in time and space $\binom{n}{k/2} n^{O(t \log k)}$ where t is the treewidth of H.

3.2.1.2 Approximate Counting

Kashtan *et al.* [103] proposed a sampling procedure to approximate the number of occurrences of a particular subgraph in a network. First the algorithm pick a random edge in a gene regulatory network. From the current sampled subgraph, one of the adjacent edges is picked with a uniform probability. The process is repeated until we obtain a k-node subgraph. However the sampled subgraph is defined as the induced subgraph that contain the same set of vertices i.e. including the sampled edges and together unsampled edges on the same set of vertices. The probabilities of sampling non-induced subgraphs are not equal eventhough they are isomorphic to each other. So the algorithm needs to correct for this unbiased sampling. Finally, the authors calculate the concentration of each possible subgraph that have the same number of vertices k. Suppose that we have K possible subgraphs $H_1, H_2, \ldots, H_K$ of size k and $S_1, S_2, \ldots, S_K$ are the numbers of their occurrences. The concentration of H_i is calculated as:

$$C_i = \frac{S_i}{\sum_i S_i}$$

The authors ran the sampling algorithm on a variety of networks including a WWW network [13] that consists 3.25 x 10^5 nodes, 1.46 x 10^6 edges for subgraphs up to 5 nodes. In all the experiments that they performed, they showed that the results converge toward the real values within 10^5 samples or less.

Arvind and Raman [8] used the color coding approach to count the number of subgraphs in a given graph G which are isomorphic to a *bounded treewidth graph* H. They give a randomized approximate counting algorithm with running time $k^{O(k)} \cdot n^{b+O(1)}$ where n and k are the number of vertices in G and H, respectively, and b is the treewidth of H. The

framework which they use is based on approximate counting via sampling [102]. Even when $k = O(\log n)$, the running time of this algorithm is *super-polynomial* with n, and thus is not practical.

Alon and Gutner [5] combined the color coding technique with a construction of what is called *Balanced Families of Perfect Hash Functions* to obtain a *deterministic* algorithm to count the number of *simple paths or cycles* of size k in an input graph G with running time $2^{O(k \log \log k)} n^{O(1)}$, still *super-polynomial* in n when $k = O(\log n)$.

The above algorithm has been improved by [4], given an additive error ϵ and error probability δ, we present a randomized approximation algorithm that with success probability $1 - 2\delta$ outputs a number within ϵ of the number of non-induced occurrences of a tree T of k vertices in a graph G of n vertices running in time $O(|E| \cdot 2^{O(k)} \cdot \log(1/\delta) \cdot \frac{1}{\epsilon^2})$. Note that if $k = O(\log n)$ and ϵ, δ are fixed, we have a polynomial time algorithm. The idea is to divide detecting an occurrence of T in G into detecting occurrences of subgraphs T and T in G. Assigning colors or labels to vertices in G keeps track of which vertices in T and T have been used so far i.e. to make sure that T and T does not share any vertices. The algorithms consists of many iterations. At each iteration, there are two main steps:

1. **Color coding**. Assign each vertex of input graph G independently and uniformly at random with one of the k colors.

2. **Counting**. Using dynamic programming, count the number of non-induced occurrences of T in which each vertex has a unique color.

For each iteration, the probability that an subgraph H in G isomorphic to T is colorful is $k!/k^k = O(e^{-k})$. Thus, after around $O(e^k)$ iterations, we have high constant probability that we count H one time. By summing up the number of colorful occurrences of T in G and accounting for some overcounting factor, they can obtain a good estimate of the real number of occurrences.

In [4], the authors obtained treelet distributions (distributions of occurrences of trees up to 10 vertices) of available PPI networks of unicellular organisms (*Saccharomyces cerevisiae*, *Escherichia coli* and *Helicobacter Pyloris*), which are all quite similar, and a multicellular organism (*Caenorhabditis elegans*) which is significantly different. Furthermore, the treelet distribution of the unicellular organisms are similar to that obtained by the duplication model but are quite different from that of the preferential attachment model.

Other randomized algorithms have improved the running time for counting subgraphs of smaller size and with specific structures. Gonen *et al.* [65] designed a sublinear time algorithm to approximate the occurrences of a path of length 2 or in general a star of small size. Tsourakakis *et al.* [202] proposed a randomized approximation algorithm for counting triangles in a massive graph.

3.2.2 Dense Subnetworks

Perhaps the second common hypothesis about biological functional modules is that dense subnetworks in biological networks may correspond to functional subunits. Analogous to network motifs where their occurrences are more frequent than expected , a dense subnetwork in a biological network contains more edges among the constituent genes than a random subnetwork of the same size. It is interesting to note that dense subnetworks seem to be more abundant in biological networks than random networks generated to match the topological properties for example degree distribution of the natural networks [34].

Dense subnetworks from protein interaction networks have been observed to correspond to protein complexes or functional units of protein complexes [184, 9]. Densely connected subnetworks that are conserved across protein interaction networks of various organisms and species also correspond to protein complexes or functional modules [171]. The authors from [9, 184] revealed that dense sunetworks that presents in gene coexpression networks also correspond to functional modules. In genetic interaction networks, a functional module may consist of two sets of nodes A and B where there are abundant of interactions among member genes from A and from B and less connections among genes in a same group [105]. A and B could correspond to complete/partial known pathways [105].

In the ideal case, a dense subnetwork S is a clique which contain $|S| * (|S| - 1)/2$ edges. Finding out whether a network contain a clique of at least k vertices is a NP-Complete problem. However, in practice many heuristics, enumeration approaches, and approximation algorithms could yield relative good results in terms of meaningful biological subnetworks.

Computational methods have been developed for extracting dense subnetworks from protein interaction and other biological networks. They can be broadly divided into three categories: minimum/normalized cut based, average clustering coefficient based, and random walk based approaches.

3.2.2.1 Graph Cut Based Approaches

First of all, we give the definitions of a minimum cut and a minimum normalized cut. Suppose that we have two disjoint subsets of vertices $A \subset V$ and $B \subset V$ such that $A \cap B = \emptyset$ and $A \cup B = V$. an edge cut is a set of edges E_c such that $E(G) - E_c$ is disconnected. The set of edges E_c with the minimum number of edges is defined as a minimum cut. We defined a cut $cut(A, B)$ as the weights of edges pass through A and B:

$$cut(A, B) = \sum_{u \in A, v \in B} w(uv)$$

A minimum cut is the cut with the minimum value by the above definition. In order to define a normalized cut, we define $assoc(A, V)$, $assoc(A, V)$ as the total weights of edges connecting vertices A, B to other vertices in V:

$$assoc(A, V) = \sum_{u \in A, t \in V} w(ut)$$

$$assoc(B, V) = \sum_{v \in B, t \in V} w(vt)$$

The value of a normalized cut is defined as

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

A minimum normalized cut [175] is the one with minimum value defined as above.

The computational methods in this category are usually top-down approaches. They start with an original network, divide the network into two subnetworks such that there are less edges among these two. Then, they repeat the same process for each subnetwork.

The CLICK (CLuster Identication via Connectivity Kernels) algorithm [173] is a graph partitioning approach that separates a graph into several subgraphs based on minimum cuts. It first separates the whole network into two subnetworks G_1 and G_2 based on minimum cuts. It then recursively partition G_1 and G_2 into small small graphs. The recursive process ends when the minimum cut is less than a certain threshold. The authors applied the algorithm on mRNA coexpression networks to reveal tighter coexpressed gene clusters and clustered proteins based on their sequence similarity into families. Another graph partitioning algorithm highly-connected subgraph (HCS) [77] is also based on minimum cut. Original network G is rst separated into two subgraphs G_1 and G_2 in which G_1 is a highly connected subgraph and G_2 is not. The algorithm repeats the process on G_2.

The minimum cut criterion usually cuts small sets of nodes from the graph. As verified by the experiments in [85] found HCS often cuts off one node in each iteration. Since it is time consuming to compute a minimum cut, algorithms based on minimum cuts could take a long time to run. Instead of using minimum cut as a criterion for graph partitioning, the CODENSE algorithm [85] is similar to CLICK but based on normalized cuts. The authors built a consensus graph from many mRNA coexpression networks from many yeast microarray datasets. They applied CODENSE on this consensus graph to discover many gene clusters where member genes share a biological process annotation. They also used these gene clusters to annotate genes with unknown functions.

3.2.2.2 Clustering Coefficient Based Approaches

In contrast to the approaches based on minimum cut and minimum normalized cut, the algorithms developed for this category are usually bottom up approaches. They start with each vertex in the original network and grow a subnetwork by adding one vertex at each iteration as long as the density defined by clustering coefficient is greater than some threshold. However, each algorithm has a distinct mathematical definition of density.

One of the most common clustering coefficient is defined as follows. Let $S \in V$ be a subset of vertices. The clustering coefficient of S is

$$density(S) = \frac{2\sum_{u,v \in S} w(uv)}{|S|(|S|-1)}$$

S is α-dense connected subnetwork if $density(S) \geq \alpha$. S is α-dense connected subnetwork if S is α-dense and S forms a connected subnetwork in G.

Many greedy algorithms and heuristics have been developed using the above density notion. MCODE [9] defines a weight for each vertex u as the density of the best subnetwork formed by vertices which are adjacent to u. They greedily grow a cluster adding a node that maximize the density of the best subnetworks formed by adjacent vertices. The authors applied MCODE on yeast *Saccharomyces cerevisiae* protein interaction networks and discovered that gene clusters correspond to many known protein complexes. NetworkBLAST [100, 170] was originally designed for comparing multiple PPI networks but is applicable for finding densely connected subnetworks in a PPI network by growing the subnetworks by adding one node at each iteration. Similar greedy strategies have been used by other other algorithms CFinder [145], DCPLus [7] , SPICI [96] for detecting protein complexes

in protein interaction networks. The authors in [121] formulate the problem of finding the maximum dense subgraphs across many mRNA coexpression networks as a quadratic integer program. Then authors relaxed several constraints and solve the problem based on convex continuous optimization.

Several enumeration algorithms have been developed for the above density definition. These algorithms work based on the observation that α-densely connected network S ($\alpha \geq 0.5$) contains a α-densely connected subnetwork S' where $|S'| = |S| - 1$. The authors in [136, 35] extended the observation for densely connected subnetworks. The observation entails a bottom up search fashion. Starting from each node in a protein interaction network, at the end of k-th iteration, we maintain the list of all the densely subnetworks of size k. At beginning of each iteration, we extend a dense subnetwork in the current list by adding an adjacent vertices as long as the newly formed subnetwork is still dense. Spirin and Mirny [184] proposed an exhaustive enumeration of all cliques in a PPI network. The authors from [61] enumerate all the α-dense but not connected subnetworks ($\alpha \geq 0.5$) in yeast and human PPI networks. Finally the authors from [136, 35] provide algorithms for enumerating all the α-densely connected subnetworks ($\alpha \geq 0.5$) in PPI networks.

Another common density notion is α-quasi-clique. A subset of vertices $S \in V$ is a α-quasi-clique if each vertex v in S has a degree at least $\alpha(|S| - 1)$. In a weighted network, we can generalized the degree of vertex v in S as $d(v, S) = \sum_{u \in S, uv \in E} w(vu)$. The authors from [147, 95] proposed efficient pruning techniques for mining all the frequent quasi-cliques from many gene coexpression networks. Another relaxed version of α-quasi-clique requires the average degree $\sum_{uv \in E} w(vu)/|S|$ is greater a threshold. It is interesting to note that, the densest subgraphs by this definition could be retrieved by algorithms from [63] in polynomial time by computing series of s-t min cuts. There is a 2-approximation to the problem [26] but it does not yield the densest subgraphs but offers much more efficient running time. [164] utilized both versions for annotating the functions of genes from Arabidopsis genome.

3.2.2.3 Random Walk Based Approaches

The Markov clustering (MCL) algorithm was designed specially simple and weighted graphs [50]. The MCL algorithm simulates random walks within a graph by the alternation of expansion and inflation operations. Expansion refers to taking the power of a stochastic matrix using matrix multiplication. Inflation operation changes the probabilities for all these walks in the graph, boosting the probabilities of intra-cluster walks and reducing inter-cluster

walks so that the resulting matrix is again a stochastic matrix i.e. all the entries in the matrix are non negative and for each row or each column, the summation of the entries is 1. The process is repeated until all the entries in the matrix do not change. All the edges in the resulted graph which are less than some threshold are removed. The returned clusters are the connected components.

Enright *et al.* [50] employed the MCL algorithm for the assignment of proteins to families. A protein-protein similarity graph is built where edges within the graph are weighted according to a sequence similarity score obtained from an algorithm such as BLAST. [149] applied MCL to the protein interaction network of Saccharomyces cerevisiae to detect functional modules. Others [112, 154, 76, 56] have applied MCL algorithm to detect protein complexes from the Tandem Affinity Purification and Co-Immunoprecipitation data.

3.2.2.4 Other Approaches

King *et al.* [111] proposed a local search algorithm Restricted Neighborhood Search Clustering Algorithm (RNSC) based on the tabu search. The algorithm begins with an initial random or user-input clustering and denes a cost function. Nodes are then randomly added to or removed from clusters to nd a partition with minimum cost. RNSC removes clusters based on their size, density and functional homogeneity. The disadvantage of RNSC is that it depends on the quality of initial clustering which is random or user defined.

Navlakha *et al.* [139] proposed a novel graph summarization (GS) technique based on graph compression [140] to discover functional modules from a PPI networks. The method defines a biological module as a set of proteins that have similar sets of interaction partners. GS compresses the original PPI graph into a summary graph where the nodes correspond to biological modules. The authors applied GS to predict complex memberships, biological processes, functional annotations.

3.3 Using Network Topology And Other Genomics Data

In the previous section, we have looked at various computational methods for discovering functional modules in terms of network motifs and dense subnetworks from PPI networks. These protein interactions in these PPI networks are recorded in regular lab conditions. Many of these interactions are collected by different labs when the cells are in different states. Thus, all the protein interactions which are reported from a database provide a static

view of the interaction networks in which all the methods in the previous section apply on. However, living cells are dynamics by nature: a functional module or biological process might be active in a condition and off in another condition. Therefore, detecting active functional modules in a condition or a set of conditions could provide valuable insights into dynamic behaviours of the cells.

Recent high-throughput genomic technologies have generated increasingly vast amount of data about cells in various states and under various conditions. These technologies allow for simultaneous genome-wide assaying of the snapshots of genomic variation, gene expression, DNA methylation, microRNA expression of the cells under the conditions or states of interest. Therefore, integrating PPI data with other genomic data could help to discover active biological processes/functional modules under a condition or set of conditions.

Previous works have established the interconnection between multi omics data and protein interactions. And these interconnections have been exploited to discover new hypotheses. For example, the authors in [59, 71] have shown that the encoded proteins of genes with similar expression profiles are more likely to interact. Later, Jansen *et al.* [92] made use of pairwise gene expression similarities to predict protein interactions. In another example, it is well known that cancer is a disease of pathways and it is hypothesized that somatic mutations target genes in some regulatory and signalling pathways [72]. Recent studies on cancer based on sequencing assess whether mutated genes significantly overlap known cancer pathways [180, 218, 41].

In this section, we will explore various computational methods for discovering functional modules by integrating PPI network data and multi-omics profiles. The approaches are presented in the chronological order.

The input of the considered algorithms here is a biological network and expression profiles of genes. Expression profile of a gene is mRNA expression measurements under different conditions or with different samples from a population. Other approaches in this section also consider genomic variations from different samples or individuals. The output like before is a list of gene sets.

Before we go into the descriptions of the prominent approaches, we first describe a couple of notations that we are going to make use of. We denote $g_1, g_2, \ldots, g_n$ as the genes/protein products which have available measurements. We denote the original network $G = (V, E)$ and for each $uv \in E$ $(u \in V, v \in V)$ we denote $0 \leq w(uv) \leq 1$ is the associated confidence score for the interaction. Suppose that $n = |V|$ is the number of vertices in the network.

In an unweighted network $w(uv) = 1$ if there is an interaction; 0 otherwise. In addition to the biological network G, we denote M denote the expression matrix of gene expression measured across m conditions. So M is a $n \times m$ matrix and M_{ij} is the expression of gene i measured at condition j.

3.3.1 Ideker *et al.*

Ideker *et al.* [89] was probably the first to introduce a computational approach for discovering active subnetworks in a set of samples. Each gene g_i in a sample is associated with a $z(g_i)$ score that quantifies its expression change. The z score of a subnetwork S with k vertices is computed as

$$z(S) = \sum_i \frac{z(g_i)}{\sqrt{k}}$$

For each subnetwork, z values are computed for all the conditions. These z scores are then combined into a single z score. The authors designed a heuristic approach to search for subnetworks with the best z score based on Simulated Annealing. Random subnetworks with the same size are sampled and their z scores are computed. For each subnetwork discovered from the original network, a p-value is computed from the distribution of z scores from random subnetworks. The insignificant subnetworks are then discarded. The authors applied their algorithms on a yeast data set.

3.3.2 Segal *et al.*

Segal *et al.* [169] proposed a probabilistic model to search for sets of genes that have similar expression profiles, and have a significant number of protein-protein interactions among them. Suppose that each gene belongs to one of k functional modules $f_1, f_2, \ldots, f_k$. They denote A_{ij} ($1 \leq i \leq n$, $1 \leq i \leq m$, $A_{ij} \geq 0$) as the probability that the gene g_i is in the functional module f_j. Thus, for each gene g_i, we have

$$\sum_j A_{ij} = 1$$

The authors used a naive Bayes model for conditional distribution of the expression of gene g_i given that it belongs to the functional module f_j. Let θ_E be the parameters for this Bayes model and denote this probability $P(M, A|\theta_E)$. Remind the readers that M is the gene

expression matrix. And the authors made use of Markov Random Field model to estimate the probability that a group of genes belong to a functional module given a binary protein interaction network G. Let θ_N be the parameters for this Markov Random Field model and denote this probability as $P(G, A|\theta_N)$. Assuming the independence between two probability models, they could estimate:

$$P(G, M, A|\theta_E, \theta_N) = P(M, A|\theta_E) \times P(G, A|\theta_N)$$

The authors utilized the Expectation Maximization algorithm to learn the parameters θ_B and θ_N. Initial assignment of genes into functional modules is based on the MCL algorithm [50]. The authors ran this algorithms on two yeast gene expression datasets and they could not only discover coherent functional groups but also protein complexes.

3.3.3 Hanisch *et al.*

Hanisch *et al.* [75] proposed a biclustering model for to combine expression data and protein interaction network. For each pair of genes, they calculated a distance score from the underlying biological network and a distance score from their expression profiles. The similarity scores from network and expression data are then combined into a single distance score. Specifically, for each pair of genes g and g', the distance score from the underlying biological network $\lambda_{net}(gg')$ is calculated as the shortest path from gene g to gene g' in the distance graph where a weight for an interaction is equal 1 minus its confidence score. The distance score $\lambda_{exp}(gg')$ is calculated as 1 minus the Pearson correlation from their gene expression profiles. They combined the computed similarity scores as follows:

$$1 - 0.5x(\lambda_{net}(gg') + \lambda_{exp}(gg'))$$

The authors then applied hierarchical average linkage clustering on the similarity matrix. Each cluster starts with a single gene. At each step, the clustering method joins two clusters A and B if the average pairwise distance between gene g in A and gene g' in B is the smallest among any two pairs of clusters. The method stops when the smallest average pairwise distance is greater than some threshold.

The authors applied their algorithm on a yeast time series data set. In this experiment, yeast is inoculated into a sugar rich medium. Gene expression are measured when the sugar is progressively depleted through seven time points.

3.3.4 Ulitsky *et al.*

Ulitsky *et al.* [204] designed MATISSE for discovering functional modules from binary PPI network and gene expression data. The idea is that each functional module corresponds to a connected subgraph in the considered PPI network and their gene expression profiles are very similar. Let $G_{exp} = (V_{exp}, E_{exp})$ be the similarity graph built from calculating pairwise expression profiles where there is an edge $gg' \in E_{exp}$ if the correlation between two gene expression profiles is greater than some threshold. The authors would like to partition the genes into k non-overlapping sets $F_1, F_2, \ldots, F_k$ such that F_i $(1 \leq i \leq k)$ forms a connected subgraph in G and each F_i has a significant number of edges that belong to E_{exp} compared to one of a random connected subnetwork of similar size.

Since the combinatorial optimization is intractable, the authors implemented several heuristics for solving the problem. The heuristic that seems to perform the best on biological data consists of three steps. First, several small seed networks that consist of one node and its neighbors are picked. In the second step, seed networks are extended by adding each node at a time. In the last step, the significant score for each subnetwork is computed and the insignificant ones are filtered.

The authors from the same group later proposed CEZANNE [206], an improved version of MATISSE that could run on a confidence score PPI network. Now instead of requiring each functional module to be a connected subgraph as in the binary PPI network, each functional module should be a densely connected subnetwork. A functional module is dense if the minimum cut of the induced subgraph is greater than some threshold.

3.3.5 Colak *et al.*

Colak *et al.* [35] designed Densely Connected Biclustering (DECOB) to discover functional modules which are densely connected subgraphs from a binary PPI network. To remind the readers the density of a connected subgraph $H = (V_H, E_H)$ is defined as

$$density(H) = \frac{2\sum_{u,v \in V_H} w(uv)}{|V_H|(|V_H| - 1)}$$

Constituent genes from a functional module H are required to have similar expression or homogeneous in at least k columns/conditions in the expression matrix M. Suppose that we discretize the gene expression matrix M and obtain expression matrix B where $B_{ij} = 1/-1$ if gene g_i is up-regulated/down-regulated in the j-th condition; $B_{ij} = 0$ otherwise. Let

$C = \{c_1, c_2, \ldots, c_k\}$ be the set of conditions that member genes of the functional module required to have similar expression. For every pair of genes g_i and g_j in V_H and for any l-th condition ($l \in C$), we have $B_{il} = B_{jl}$.

The problem of discovering densely connected homogeneous modules is intractable. However with a certain density threshold and the sufficient number of conditions k, the number of densely connected homogeneous functional modules is not abundant. Thus, the authors designed an efficient enumeration algorithm to discover all the proposed functional modules. This algorithm is similar to the one discussed in Section 3.2.2. And it is based on the observation that a densely connected homogeneous functional module of size k contains a densely connected homogeneous functional module of size $k-1$. The authors derived a bottom up search approach: starting where each module has only one node, extending module by one node at a time while maintain connectivity, density and homogeneity constraints.

3.3.6 Vandin *et al.*

Vandin *et al.* [210] proposed a computational method to integrate genomic variation profiles and PPI network data. Specifically, the authors made use of somatic mutation data from from Cancer Genome Atlas and lung adenocarcinoma samples from the Tumor Sequencing Project. It is known that two tumors rarely have the same complement of mutations in cancer and most cancer genes are mutated at much lower frequencies. The observed frequency of mutation is an inadequate measure of the importance of a gene. However, most previously pathway analyses are using the frequencies.

The goal is to recover subnetworks that harbour a significant number of mutations compared to random subnetworks from PPI networks. A subnetwork or a subset of genes is mutated in (covers) a patient if the patient has a somatic mutation in one of the member genes. The computational problem is to uncover a connected subnetwork with an upper limit on the number of vertices that cover the most number of patients. The authors could recover known pathways such as p53 pathway that are previously known to be important in these cancer data sets.

3.3.7 Kim *et al.*

Kim *et al.* [110] proposed a computational approach that aims to extract driver mutations and deregulated pathways from Glioblastoma multiforme (GBM) patients. They integrated

multi omics data including copy number, gene expression profiles and protein interaction networks. Assuming that disease-associated gene expression changes are caused by genomic alterations, they uncover potential paths from copy number altered regions to differentially expressed genes through a network of molecular interactions.

Similar to the approach by Ulitsky *et al.* [204], a gene g covers a sample if it is differentially expressed in that sample. At first, the authors look for a set S of minimum number of genes such that each sample is covered with at least a pre-defined number of times. Based on correlation with the expression profiles in S, only genes with significant copy number alteration and high correlation scores remain. For each remaining gene g, they assess whether there is a path in protein interaction network from g to other genes in S. The weights of the edges in the PPI network are calculated as correlation scores of gene expression profiles.

3.3.8 Sharan *et al.*

Sharan *et al.* [172] proposed a computational approach for constructing signalling pathways from phosphoproteomics data. Phosphoproteomics data could provide additional and better information compared to gene expression data. By reporting phosphorylation status of proteins using mass spectrometry, they are could be more reliable as a change in phosphorylation status usually reflects a change in protein activity. The phosphoproteomics data can indicate which proteins might be potential drug targets by using the kinase inhibitors. In this work, the authors tried to construct signalling network models from two datasets for epidermal growth factor receptor (EGFR) signalling and interleukin 1 (IL-1) signalling pathways. In both datasets, the cells were stimulated with different ligands and treated with different inhibitors, the activity (phosphorylation) levels of certain proteins were recorded.

Each vertex (a protein) in G is either active (1) or inactive (0). They also assume that the state of a vertex u is a boolean function $f(u)$ of the states of its direct predecessors $P(u)$. They further assume that $f(U)$ is monotone non-decreasing with respect to its input and regulatory relationships. In more details, given a vertex u and one of its predecessors $v \in P(u)$, the function $f(u)$ is monotone non-decreasing in v ($\bar{v}$) when v promotes/inhibits the activity of u. The input data is a set of experiments in which some genes/vertices are perturbed and the states of some affected vertices are recorded. The computational problem is to find a boolean function $f(u)$ for each vertex u that fits the data the best.

3.3.9 Other Approaches

Most of the earlier approaches on active subnetwork discovery focus on discovering well-characterized pathways which are active in the considered experiments. For example, gene-set enrichment analysis (GSEA) [191] takes a set of genes and check out whether genes from the set are significantly differentially expressed compared to a randomly picked gene set with the same size. Many other approaches have been developed with improved statistical procedures [115, 201, 47]. Recent approaches have been developed to consider regulatory relationships among member genes rather than consider them as a set [44, 198, 212].

Many computational approaches have been improved on the seminal approach by Ideker *et al.* [89]. Dittrich *et al.* [42] designed an exact approach for discovering functional modules. Specifically, the authors formulated an integer linear program to solve the problem, however, there is no guarantee that it can run efficiently for a large PPI network like human. Guo *et al.* [69] generalized the approach proposed by Ideker *et al.* [89] on weighted interaction network. Specifically, they searched for the optimal subnetworks where edges are assigned to weights equal to correlation scores computed from expression data.

Other groups designed computational methods to uncover functional modules in the models of signalling pathways [186, 168, 226]. For example, Steffen *et al.* [186] designed Netsearch to reconstruct signalling pathways which consists of linear paths that start at any membrane protein and ending at any DNA-binding protein. Scores of the linear pathways are computed based on correlation of the member genes. Only top ranked pathways still remain. Other groups designed computational approaches to discovered signalling pathways specifically for quantitative trait loci and gene perturbation data [193, 221]. For example, Yeger-Lotem *et al.* [221] uncovers signalling pathways from perturbed genes to affected/differentially expressed genes in a yeast data set.

3.4 Using Labels of Samples

Finally, we will review computational methods for discovering functional modules that correlate labels of the samples. Unlike the input for the approaches in the previous section, the samples are supplied with meta data that can potentially divide the samples into different groups. For example, tumour samples from taken a cancer study usually come with clinical information such as subtypes of the cancers. Functional modules in which member gene expression correlates with the subtypes of cancers/diseases could provide a comprehensive

view of the molecular mechanisms underlying pathology. Such functional modules could be used as predictors or subnetwork markers for the disease status of the patients.

In recent years, there is an increasing number of prognostic markers of cancers that have been discovered through genome-wide expression data [64, 3, 208, 156, 217]. In these studies, each gene is ranked based on its differential expression between subgroups of samples and top ranked genes are picked. Typical measures for differential expression are t-test and mean-based fold-change. These approaches are known as single gene marker based approaches. In breast cancer, single gene marker based approaches [208, 217] have identified marker sets around 70 genes which are 60-70% accurate for prediction of metastasis of breast cancer.

Despite these successful results, single gene based approaches suffer from multiple weaknesses. First, they may only detect the genes with the strongest differentiation whereas single genes with weaker signals that could be combined to have much better discrimination scores could be missed. Moreover, single gene based approaches could be very sensitive to noise and variations in training data, resulting in marker sets which are not reproducible in another data set. Recent studies have shown that small differences in training data (such as the changes in ratio of subtypes) may produce very different marker sets. In fact, two studies on breast cancer [208, 217] that identified around 70 gene signatures only have 3 overlapping genes. However, it is hypothesized that a reproducible marker set should provide more robust predictive performance which is required for clinical applications [49].

To address these shortcomings, many groups have aimed to identify de novo markers associated with phenotype by integrating gene expression data and network topology. Here each marker is called subnetwork marker and its activity is calculated as a function of the expression of component genes in the subnetwork. In the seminal work, Chuang *et al.* [31] introduced the use of all members of a protein-protein interaction (PPI) subnetwork as a marker for predicting metastasis in breast cancer. Chuang *et al.* [31] demonstrated that subnetwork markers are more robust, i.e. their results tend to provide more reproducible results across different cohorts of patients. Moreover, subnetwork markers seem to provide better insights into pathways involved in tumor progression.

The input of the considered algorithms here is a biological network, expression/genomic variation profiles of genes. Expression/genomic variation profile of a gene is measured for all the samples from a population. The samples here come with meta data that could help us to categorize them into different groups. The output of all the algorithms is again a list of

gene sets. The algorithms in this section should utilize the meta data from the samples in the process of discovering subnetwork markers.

The general strategy for discovering subnetwork markers consists of three major steps: (1) data integration, (2) search for optimal subnetworks, and (3) marker selection. These steps are illustrated in 3.2.

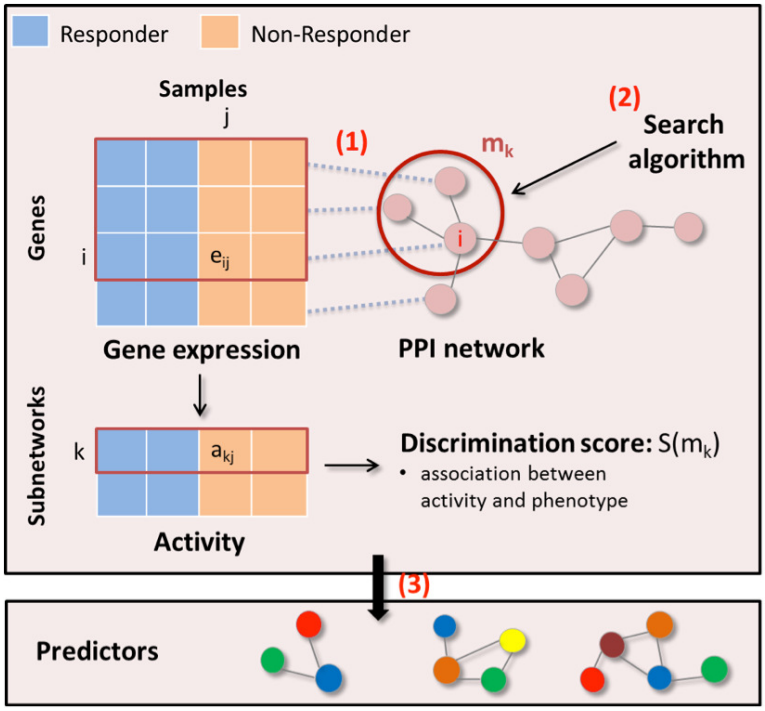


Figure 3.2: A general strategy for identifying subnetwork markers: 1) mapping gene expression profiles to protein products in the PPI network 2) Search for optimal subnetworks whose gene expression profiles correlate with the labels of the samples 3) Retain the best subnetworks as predictors. Here we have two groups of tumor samples: sensitive to a treatment (responder) and resistant to a treatment (non-responder)

In the first step, the gene expression profile and PPI data are integrated by overlaying each gene in the expression profile onto its corresponding protein in the PPI network. In this way, an edge is assigned for a pair of genes if there exists an interaction between their corresponding proteins. Note that only genes with corresponding proteins in the network are used to identify subnetwork markers in the subsequent steps.

In the second step, a search algorithm is employed to identify subnetworks with activities that best correlate with the phenotype (such as response to treatment). A subnetwork is a

set of connected genes extracted from the PPI network, and the activity of a subnetwork in any sample is calculated as a function of the expression levels of constituent genes of the subnetwork in that sample. A typically used function is the aggregate one. This aggregation essentially collapses many gene features into one subnetwork feature that captures the discriminatory potential of multiple gene markers in a single metagene marker. For example, if gene A discriminates drug-response in one set of patients and gene B discriminates drug-response in a second set of patients, then the aggregate activity of these two genes can potentially discriminate response in both sets of patients. To identify candidate subnetwork markers, the search algorithm scans the PPI network for subnetworks with maximal discrimination scores, where the discrimination score is a measure of the association between subnetwork activity and the phenotype. If a simple greedy search algorithm is used, it would search in the following manner. To find the optimal subnetwork that includes a specific gene (seed gene), the algorithm starts by including that gene in the subnetwork. Then, it iteratively adds neighbouring genes from the PPI network into the subnetwork if they improve the discrimination score. If no additional gene can be added to improve the discrimination score, then that subnetwork is considered the optimally discriminative subnetwork containing the seed gene and is subsequently added to the list of candidate subnetwork markers. The search algorithm is applied repeatedly, using each node in the PPI network as the seed gene, in order to return the list of all candidate subnetwork markers.

In the third and final step, all the candidate subnetwork markers returned by the search algorithm are ranked based on their discrimination scores and the top k subnetworks are selected as predictors of phenotype. The activity levels of the selected subnetwork markers are used to train a classifier for predicting on new samples.

3.4.1 Chuang *et al.*

Chuang *et al.* [31] were one of the first that proposed a computational approach to discover connected subnetworks whose member gene expression is correlated with phenotypes. For each gene g_i in a sample, it is associated with a $z(g_i)$ score that quantifies its expression change. For each subnetwork S with k vertices, we define the subnetwork activity as

$$z(S) = \sum_i \frac{z(g_i)}{\sqrt{k}}$$

The goal here is find a connected subnetwork that is the best correlated with the

phenotype variable C. The correlation measure is Mutual Information (MI). Unfortunately, this problem is NP-Hard. The authors designed a greedy algorithm to search for optimal subnetworks by extending one node at a time. The authors assessed the significance of the returned subnetworks by applying the same greedy algorithms on the datasets where the labels of the samples are swapped. For each subnetwork discovered from the original network, a p value is computed from the distribution of MI scores of subnetworks discovered when the sample labels are swapped. The insignificant subnetworks are then discarded. The author applied their algorithm to predict metastasis status of breast cancer patients. The authors shown that the subnetwork markers have better predictive performance than top ranked genes using t-test. The identified subnetworks are significantly more reproducible between different breast cancer cohorts than individual marker genes selected without network information and provide better models of the molecular mechanisms underlying metastasis.

3.4.2 Ulitsky *et al.*

Ulitsky *et al.* [205] proposed a computational method that detects differentially expressed subnetworks from case/control gene expression data and PPI networks. A gene is differentially expressed or *covers* a case sample if it is over/under expressed based on a statistical significant score from the distribution of the gene expression in the control samples. A subnetwork covers a sample if a member gene covers that sample. The computational problem is to find the smallest subnetwork such that it covers all the case samples and each sample is covered with at least a predefined number of times. Even though the problem is NP-Hard, the authors could provide some basic approximation algorithms. The authors applied their algorithms on case/control gene expression data from Huntington disease.

3.4.3 Hwang *et al.*

Hwang *et al.* [87] proposed a novel approach for discovering network based markers by modelling similarity among samples in the training data using hyperedges in a hypergraph. Vertices are samples and an hyperedge is present among some samples if a gene g is either up-regulated or down-regulated in that set of samples. Labelled (unlabelled) samples/vertices are from training (testing) data respectively. The goal is to assign a label for an unlabelled sample/vertex and a weight to each gene/hyperedge to maintain the similarities

among vertices that have edges in the hypergraph and the PPI network. The weight of a gene/hyperedge reflects how informative the gene in predict samples in the training data. The authors formulated a quadratic program to optimize the following criteria: 1) penalizing inconsistent labelling of samples that have a lot of edges in the hypergraph; 2) penalizing inconsistent labelling of samples with known outcomes 3) penalizing inconsistent weighting of pairs of genes that form edges in the PPI network.

The authors applied their algorithm on two breast cancer expression datasets to predict metastasis status of breast tumors. Without looking for connected or densely connected subnetworks, subnetworks formed by their genes are significantly enriched with biological pathways are closely linked with breast cancer. Note that this approach can also make use of samples from the testing set. Thus, it could utilize the interrelations among samples in the testing set through the use of hypergraph. So it probably results in better predictive performance.

3.4.4 Chowhudry *et al.*

Chowhudry *et al.* [30] proposed an interesting approach to discover subnetwork markers from PPI networks based on information theory. For the purpose of illustrations, I present a simplified model of this approach. Gene expression data from the case samples are discretized into two states (1: differentially expressed, 0: otherwise). Each subnetwork H of k vertices is associated with 2^k network states and $States(H)$ be the set of states of H. For each network state S, its probability/frequency $f(S)$ is same as the number of case samples in which member genes exhibit the same state. If the expression of genes from a subnetwork form a frequent state i.e. many case/control samples are associated with it, the subnetwork could be used as a predictor. The higher the frequency is the lower the entropy of a subnetwork is. Formally, the entropy of the subnetwork S is

$$f(H) = \sum_{S \in States(H)} f(S) log(f(S))$$

The goal is to look for a connected subnetwork with minimum entropy. This problem is also NP-Hard, thus, the authors provided a branch and bound approach to search for the subnetwork with the minimum score. The subnetwork activity is calculated as a linear combination of expression levels of its component genes. The weights of the linear combination is learnt through training a neural network model. They applied their algorithms on colorectal cancer data sets to predict the disease progression.

3.4.5 Fortney *et al.*

Fortney *et. al* [55] designed a novel method based on density for discovering subnetwork markers to predict the chronological ages of *C. elegans*. The density definition is quite different from the ones introduced earlier. The density of a subnetwork is inversely proportional to the number edges among the genes from the subnetworks and genes outside of the subnetworks (external weight) and proportional to the number of edges among the member genes (internal weight). Thus, they defined the external weight (w_{ext}) and internal weight (w_{int}) of a subnetwork as follows:

$$\begin{aligned} w_{ext}(H) &= \sum_{g \in H, g' \notin H, gg' \in E} w(gg') \\ w_{int}(H) &= \sum_{g, g' \in H, gg' \in E} w(gg') \end{aligned} \tag{3.1}$$

Thus, the density of H is $w_{int}^2(H)/(1+w_{ext}^2(H))$. The authors made use of the functional association network WormNet [116]. In fact, it is a genetic interaction network that was constructed for around 80% of the known/predicted genes. It is also a weighted network where the weights of the edges are computed using Spearman correlation from gene expression profiles. Subnetwork activity is calculated as the average expression of the member genes. And the correlation between the subnetwork activity and phenotype is also calculated using Spearman correlation. The authors used a simple greedy algorithm to discover their defined subnetwork markers. The authors demonstrated that their subnetwork markers could be a relative good predictors for aging process and the subnetworks are enriched for many genes known longevity genes. The authors also made used of their subnetwork markers to annotate genes without annotations.

3.4.6 Other Approaches

Earlier approaches for discovering subnetwork markers utilized curated pathway databases or groups of genes annotated with same Gene Ontology (GO) terms. For example, Guo *et al.* [68] ranked groups of genes annotated with the same GO terms for their correlation with different subtypes of cancer cell lines. They also examined two possibilities of estimating the activity of a group of genes as the mean or median of the expression levels of constituent genes. Bild *et al.* [18] proposed to take first principle component as pathway/network

activity instead of taking average expression of members like previous approaches. They have shown that pathway activities could be used as predictors for patient subgroups, and sensitivity to therapeutic compounds. Later, Lee *et al.* [114] examined various methods for combing expression profiles of a set of genes into a single value: average or median as in [68], first principle component as in [18], and divided by square root of the number of genes as in [31]. If pathway/subnetwork activity is estimated as in the seminal work by Chuang *et al.* [31], it seems to yield the best predictive performance. Teschendorff *et al.* [200] discovered decomposed known pathways into subnetworks in which member genes have similar expression profiles. Then they ranked the subnetworks based on their correlation (Penalized Cox regression) with the phenotype.

Similar to MATISSE that was discussed earlier, Su *et al.* [190] proposed a greedy algorithm to discover a connected subnetwork whose combined activity is correlated with the phenotype and member genes have high correlations in their expression profiles. Note that in this approach, subnetwork activity is defined as linear combination of expression of member genes. Taylor *et al.* [199] proposed an interesting approach for discovering network based markers based on hubs in the interaction networks. They have shown that the average Pearson correlation of the expression of a hub protein and its interacting partners can be used to reliably predict survival of breast cancer patients. Zhu *et al.* [228] designed a support vector machine approach that incorporates protein interaction information on expression data sets related to the Parkinson's disease and breast cancer. The authors adds an extra penalty term into the objective function of the quadratic program to enforce that genes form edges in PPI networks should be taken/leaved out together.

Recent approaches on subnetwork marker discovery have integrated genomic variation with/without expression profiles with PPI networks. Chen *et al.* [28] proposed an approach that first discovers the frequently altered copy number genes and subnetwork markers rooted at these genes in PPI networks for predicting survival time in ovarian cancer. Vandin *et al.* [209] also integrated somatic mutation and copy number data to discover connected subnetworks that are correlated with the survival time in ovarian cancer.

Chapter 4

Confidence-Scored Network Motif Counting

A current major issue in evolutionary systems biology is to reliably quantify both organismic complexity and evolutionary diversity from a systemic point of view. While currently available biomolecular networks provide a data basis, the assessment of network similarity has remained both biologically and computationally challenging. Since currently available network data is still incomplete, simple edge statistics, for example, do not apply. Moreover, recent research has revealed that many biomolecular networks share global topological features which are robust to missing edges, which rules out many more straightforward approaches to the topic (see e.g. [84] for a related study on global features such as degree distribution, k-hop reachability, betweenness and closeness). On the more sophisticated end of the scale of such approaches would be attempts to perform and appropriately score alignments of the collection of all systemic subunits of two organisms. However, the development of workable scoring schemes in combination with related algorithms comes with a variety of obvious, yet unresolved, both biological and computational issues. Clearly, any such scoring schemes would already establish some form of condensed, systemic evolutionary truth by themselves.

This explains why recent approaches focused on monitoring differences between biomolecular networks in terms of *local structures*, which likely reflect biological arrangements such as functional subunits. A seminal study which reported that statistically overrepresented graphlets, i.e. small subnetworks, are likely to encode pathway fragments and/or similar

functional cellular building blocks [134] sparked more interest in the topic. In the meantime, to discover and to count *biomolecular network motifs* has become a thriving area of research which poses some intriguing algorithmic problems. As is summarized in the comprehensive review [32], such approaches are supported by various arguments.

As discussed in the earlier chapter, most of the previous works focused on determining the number of all possible ''induced'' subgraphs in a PPI network, which already is a very challenging task. Recently developed algorithms improved on this by counting induced subgraphs of size up to $k = 6$ [84] and $k = 7$ [67]. However, the running time of these techniques all increase exponentially with k. To count subgraphs of size $k \geq 8$ required novel algorithmic tools. A substantial advance was subsequently provided in [4] which introduced the ''color coding'' technique for counting non-induced occurrences of subgraph topologies in the form of bounded treewidth subgraphs, which includes trees as the most obvious special case. Counting non-induced occurrences of network motifs is not only challenging but also quite desirable since non-induced patterns are often correlated to induced occurrences of denser patterns which, in turn, often reflect functional cellular building blocks, as is widely established (e.g. [227]).

While these studies successfully revealed differences between PPI networks of uni- and multicellular organisms, a binary edge has remained a notoriously noisy datum. However, none of the studies considered PPI networks with weighted edges where edge weights reflect the confidence that the interactions are of cellular relevance instead of being experimental artifacts. Weighted network data have recently become available and have already been employed for other purposes (see e.g. [206] and the references therein for a list of weighted network data sources). One of the main reasons for the lack of network motif studies on such data might be that to exhaustively mine biomolecular networks with probabilistic edge weights poses novel computational challenges.

In [38] and this chapter, we show how to apply the ''color coding'' technique to networks with arbitrary edge weights and two different scoring schemes for weighted subgraphs. Edge weights are supposed to reflect our confidence in the interactions, as provided for instance in the *STRING* database, and we will apply a scoring scheme which reflects our expectation[1] in entire subgraphs to be present or not. *STRING* is a major resource for assessments of protein interactions and/or associations predicted by large-scale experimental data (in the

[1]Expectation is meant to be in the very sense of probability theory, by interpreting confidence scores as probabilities

broad sense, including e.g. literature data) of various types (see [93] for the latest issue of STRING and the references therein for earlier versions). Here, we focus on physical protein interactions in order to follow up on the recent discussions. Clearly, statistics on weighted PPI networks will establish substantial improvements over studies on binary network data in terms of statistical significance and robustness.

We compute the expected number of non-induced occurrences (E-values) of tree motifs G' (''treelets'') with k vertices in a network G with n vertices in time polynomial in n, provided $k = O(\log n)$. Note that, in binary edge weight graphs, computation of the number of expected occurrences and counting occurrences is equivalent when interpreting an edge to be an interaction occurring with probability 1. This provides the basis on which we can benchmark our results against previous studies. We use our algorithm to obtain normalized treelet distributions, that is the sum of the weights of non-induced occurrences of different tree topologies of size $k = 8, 9$[2] normalized by the total weight of all non-induced trees of size $8, 9$ for weighted PPI networks. We analyze the prokaryotic, unicellular organisms (*E.coli*, *H.pylori*), *B. subtilis* and *T. pallidum*, which are all quite similar, the eukaryotic unicellular organism *S.cerevisiae* (Yeast), and a multicellular organism (*C.elegans*). Beyond the previously reported similarities among the prokaryotic organisms, we were able to also reveal strong differences between Yeast and the prokaryotes. As before, statistics on C.elegans are still different from all other ones. As a last point, we demonstrate that our weighted treelet distributions are *robust* relative to reasonable amounts of network sparsification as suggested by [74].

To summarize, we present a novel randomized approximation algorithm to count the weight of non-induced occurrences of a tree T with k vertices in a weighted-edge network G with n vertices in time polynomial with n, provided $k = O(\log n)$ for a given error probability and an approximation ratio. We prove that resulting weighted treelet distributions are robust and sensitive measures of PPI network similarity. Our experiments then confirm, for the first time on a statistically reliable PPI network data, that uni- and multicellular organisms are different on an elevated systemic cellular level. Moreover, for the first time, we report such differences also between pro- and eukaryotes.

[2]We recall that there are 23 resp. 47 different tree topologies on 8 resp. 9 nodes, see e.g. [142].

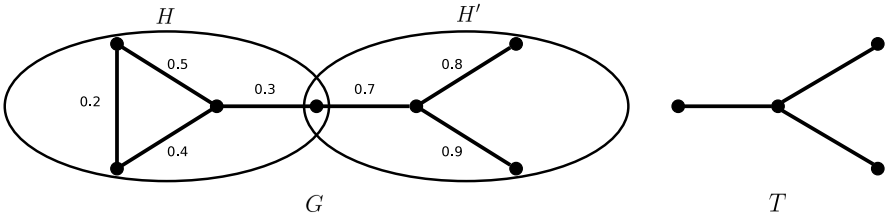


Figure 4.1: An example of counting the total weight of non-induced subgraphs in a given network G which are isomorphic to a query tree T.

4.1 Problem Definition

In the following, let $G = (V, E)$ be a graph on n vertices and $w : E \to \mathbb{R}$ be an edge-weight function. Let T be a tree on k vertices where, in the following, $k = O(\log n)$. We define $\mathcal{S}(G, T)$ to be the set of non-induced subgraphs of G which are isomorphic to T and let $E(H)$ to be the edges of such a subgraph $H \in \mathcal{S}$. We extend w to weight functions on the members of $\mathcal{S}(G, T)$ by either defining

$$w(H) = \prod_{e \in E(H)} w(e) \quad \text{or} \quad w(H) = \sum_{e \in E(H)} w(e) \tag{4.1}$$

Note that if $w(e)$ is interpreted as the probability that e is indeed present in G then, assuming independence between the edges, $w(H)$ of the first case is just the probability that H is present in G. In the following, we will focus on the first case. Proofs for the second choice of $w(H)$ can be easily obtained, mutatis mutandis, after having replaced multiplication by addition in the definition of $w(H)$. Finally, let $w(G, T) = \sum_{H \in \mathcal{S}(G,T)} w(H)$ be the total weight of non-induced occurrences of T in G. We would like to provide reliable estimates $\hat{w}(G, T)$ on $w(G, T)$. Note that $w(G, T)$ is the number of expected occurrences of T in G due to the linearity of expectation.

Consider Fig. 4.1. Here, T is a star-like tree on 4 vertices. There are two subgraphs H and H' in G which are isomorphic to T; therefore, $w(G, T) = w(H) + w(H')$. In the case that the weight of a subgraph in G is calculated as the product of the weights of its edges, we have $w(G, T) = w(H) + w(H') = 0.5 \times 0.4 \times 0.3 + 0.7 \times 0.8 \times 0.9$. In the other case, we have $w(G, T) = w(H) + w(H') = (0.5 + 0.4 + 0.3) + (0.7 + 0.8 + 0.9)$.

4.2 Computational Methods

In the following, in order to estimate $w(G,T)$ by color coding, we will randomly assign k colors to the vertices of G where k is the size of T. Therefore, we introduce the notations $[k] = \{1, ..., k\}$ for the set of k colors and $\mathcal{S}(G, T, [k])$ for the set of all non-induced subgraphs of G which are *colorful* in terms of $[k]$, that is occurrences of T where each vertex has been assigned to a different color.

The following algorithm ApproxWeightedOccur, when given an approximation factor ϵ and an error probability δ, computes an estimate $\hat{w}(G,T)$ of $w(G,T)$ efficiently in n and k, given that $k = O(\log n)$ such that with probability $1 - 2\delta$, $\hat{w}(G,T)$ lies in the range $[(1-\epsilon)w(G,T), (1+\epsilon)w(G,T)]$.

Algorithm 1 ApproxWeightedOccurr (G, T, ϵ, δ)

$G = (V, E)$, $k \leftarrow |V(T)|$, $t \leftarrow \log(1/\delta)$, $p \leftarrow k!/k^k$, $s \leftarrow 4/(\epsilon^2 p)$
for $i = 1$ to t **do**
  $Y_i \leftarrow 0$
  for $j = 1$ to s **do**
    Color each vertex of G independently and uniformly at random with one of k colors
    $X \leftarrow$ total weight of colorful subgraphs of G which are isomorphic to T
    $Y_i \leftarrow Y_i + X$
  end for
  $Y_i \leftarrow Y_i/s$
end for
$Z \leftarrow$ median of $Y_1 \dots Y_t$
Return Z/p as the estimate $\hat{w}(G,T)$ of $w(G,T)$

The following lemmas give rise to a theorem that supports our claims from above.

Lemma 4.1. *The algorithm* ApproxWeightedOccurr *(G, T, ϵ, δ) returns $\hat{w}(G,T)$ such that with probability at least $1 - 2\delta$ we have $(1-\epsilon)w(G,T) \leq \hat{w}(G,T) \leq (1+\epsilon)w(G,T)$*

Proof. Consider we are at starting the j-th iteration of the second for loop. Since each vertex in G is colored independently and uniformly at random with one of k colors, the probability p that the vertices of a subgraph H of size k are assigned to k different colors (H is *colorful*) evaluates as $p = k!/k^k$. Let x_H be the indicator random variable whose value is $w(H)$ if H is colorful in a random coloring of G and 0 otherwise. We define $X = \sum_{H \in \mathcal{S}(G,T)} x_H$, which is the random variable that counts the total weight of colorful subgraphs of G which are

isomorphic to T. The expected value of X is

$$E(X) = E(\sum_{H \in \mathcal{S}(G,T)} x_H) = \sum_{H \in \mathcal{S}(G,T)} E(x_H) = \sum_{H \in \mathcal{S}} w(H)p = w(G,T)p \tag{4.2}$$

In order to get a bound on the variance $Var(X)$ of X, we estimate $Var(x_H)$ and $Cov(x_H, x_{H'})$ for $H, H' \in \mathcal{S}(G,T)$ as follows. We first observe that $Var(x_H) = E(x_H^2) - E^2(x_H) \leq E(x_H^2) = [w(H)]^2 p$. Moreover, the probability that both H and H' are colorful is at most p which implies

$$Cov(x_H, x_{H'}) = E(x_H x_{H'}) - E(x_H)E(x_{H'}) \leq E(x_H x_{H'}) \leq w(H)w(H')p.$$

Therefore, the variance of X satisfies

$$\begin{aligned} Var(X) &= \sum_{H \in \mathcal{S}} Var(x_H) + \sum_{H \neq H' \in \mathcal{S}} Cov(x_H, x_{H'}) \\ &\leq \sum_{H \in \mathcal{S}} w^2(H)p + \sum_{H \neq H' \in \mathcal{S}} w(H)w(H')p = (\sum_{H \in \mathcal{S}} w(H))^2 p = w^2(G,T)p. \end{aligned} \tag{4.3}$$

Since Y_i is the average of s independent copies of random variable X, we have $E(Y) = E(X) = w(G,T)p$ and $Var(Y_i) = Var(X)/s \leq w^2(G,T)p/s$. Therefore, the probability that Y is smaller than or bigger than its expectation by at least $\epsilon w(G,T)p$ due to $s = \frac{4}{\epsilon^2 p}$ is at most

$$P(|Y_i - w(G,T)p| \geq \epsilon w(G,T)p) \leq \frac{w(G,T)^2 p}{\epsilon^2 w(G,T)^2 p^2 s} = \frac{1}{\epsilon^2 ps} = \frac{1}{4}. \tag{4.4}$$

Thus, with constant error probability, Y_i/p is an ϵ-approximation of $w(G,T)$. To obtain error probability $1 - 2\delta$, we compute t independent samples of Y_i (using the first for loop) and replace Y_i/p by Z/p where Z is the median of Y_i's. The probability that Z is less than $(1-\epsilon)w(G,T)p$ is the probability that at least half of the copies of Y_i computed are less than Z, which is at most $\binom{t}{t/2}4^{-t} \leq 2^{-t}$. Similarly we can estimate the probability that Z is bigger than $(1+\epsilon)w(G,T)p$. Therefore, if $t = \log(1/\delta)$ then with probability $1 - 2\delta$ the value of $\hat{o}$ will lie in $[(1-\epsilon)w(G,T), (1+\epsilon)w(G,T)]$. ◇ ◇

We still need to argue that given the graph G where each vertex is colored with one of k colors, we can compute the total weight of all non-induced colorful occurrences $w(G,T,[k])$ of T in G which refers to the variable X in the second for loop efficiently.

Lemma 4.2. *Given a graph G where each vertex has one of k colors, we can estimate $w(G,T,[k])$ in time $O(|E| \cdot 2^{O(k)})$.*

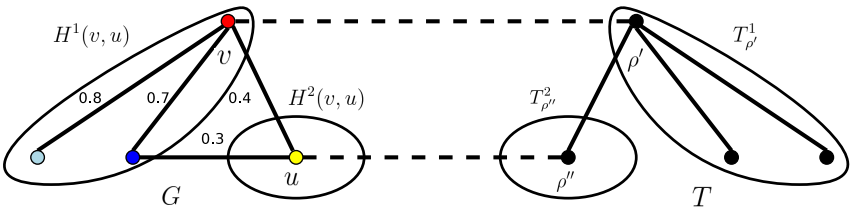


Figure 4.2: Counting the total weight of colorful non-induced occurrences of T in G by counting the total weight of colorful non-induced occurrences of subtrees $T_{\rho'}$ and $T_{\rho''}$ in G.

Proof. We pick a vertex ρ of T and consider T_ρ to be a rooted version of the query tree T with designated root ρ. We will compute $w(G, T_\rho, [k])$ recursively in terms of subtrees of T_ρ; so let $T'_{\rho'}$ be any subtree T' of T with designated root ρ'. Let $C \subset [k]$ and $\mathcal{S}(G, T'_{\rho'}, v, C)$ be the set of all non-induced occurrences of $T'_{\rho'}$ in G which are rooted at v and colorful with colors from C and $w(v, T'_{\rho'}, v, C) = \sum_{H \in \mathcal{S}(G, T'_{\rho'}, v, C)} w(H)$ to be the total weight of all such occurrences. We observe that

$$w(G, T, [k]) = \frac{1}{q} \sum_{v \in G} w(G, T_\rho, v, [k]) \tag{4.5}$$

where q is equal to one plus the number of vertices ϱ in T for which there is an automorphism that ρ is mapped to ϱ. For example, if T in Figure 4.2 is rooted at ρ'', q is equal to 3. The key observation is that we can compute $w(G, T_\rho, v, [k])$ or the total weight of colorful non-induced subtrees rooted at v in G which are isomorphic to T_ρ in terms of total weight of colorful non-induced occurrences of subtrees of T_ρ in G. Let $T'_{\rho'}$ be an arbitrary rooted subtree of T_ρ. We decompose $T'_{\rho'}$ into two smaller subtrees and count the total weight of colorful non-induced occurrences of these subtrees in G as follows. We choose a child ρ'' of ρ' and, by removing the edge between ρ' and ρ'' to decompose $T'_{\rho'}$ into two rooted subtrees $T^1_{\rho'}$ (that does not contain ρ'') and $T^2_{\rho''}$ (that does not contain ρ'); see Figure 4.2 for an example. Analogously, for every neighbor u of v in G, we denote a colorful copy of $T^1_{\rho'}$ rooted at v by $H^1(v, u)$ and a colorful copy of $T^2_{\rho''}$ rooted at u by $H^2(v, u)$. To obtain a copy H of $T'_{\rho'}$ in G by combining $H^1(v, u)$ and $H^2(v, u)$, $H^1(v, u)$ and $H^2(v, u)$ must be colorful for color sets $C_1(v, u), C_2(v, u)$ such that $C_1(v, u) \cap C_2(v, u) = \emptyset, C_1 \cup C_2 = C$ where the cardinality of C is the number of vertices of $T'_{\rho'}$. Finally, independent of the choice of u, we have

$$w(H) = w(H^1(v, u))w(H^2(v, u))w(vu) \tag{4.6}$$

To initialize the base case of single-vertex trees $T'_{\rho'}$, we set $w(G, T'_{\rho'}, v, \{i\}) = 1$ if the color of v is i; otherwise 0. In general, we have

$$
\begin{aligned}
w(G, T'_{\rho'}, v, C) &= \frac{1}{d} \sum_{u \in N(v)} \sum_{\substack{C_1 \cap C_2 = \emptyset \\ C_1 \cup C_2 = C}} \sum_{\substack{H^1 \in \mathcal{S}(G, T^1_{\rho'}, v, C_1) \\ H^2 \in \mathcal{S}(G, T^2_{\rho''}, u, C_2)}} w(H^1(v,u)) w(H^2(v,u)) w(vu) \\
&= \frac{1}{d} \sum_{u \in N(v)} \sum_{\substack{C_1 \cap C_2 = \emptyset \\ C_1 \cup C_2 = C}} w(G, T^1_{\rho'}, v, C_1) w(vu) w(G, T^2_{\rho''}, u, C_2).
\end{aligned} \tag{4.7}
$$

Note that $w(H)$ will, as in the summands, appear exactly d times across the different suitable choices of color sets C_1, C_2. For example, H in Fig. 4.2, rooted at v, is a colorful copy of a star-like rooted tree T with three leaves. There are three different ways by which one can decompose H into a path of length 2, denoted by H_1 and a single node H_2, meaning that $d = 3$ in this case. Now observe that the weight of H $w(H)$ will appear three times as a summand in the above summation scheme, according to the three different decompositions of H. The proof for the case of additive weight schemes proceeds mutatis mutandis, after having replaced multiplication by addition and adjusted the base case ($w(G, T'_{\rho'}, v, \{i\} = 0$ regardless of the color of v). Let $N(v)$ be the set of neighbors of v, we have

$$
\begin{aligned}
&w(G, T'_{\rho'}, v, C) \\
&\qquad = \frac{1}{d} \sum_{u \in N(v)} \sum_{\substack{C_1 \cap C_2 = \emptyset \\ C_1 \cup C_2 = C}} n_2 w(G, T^1_{\rho'}, v, C_1) + n_1 n_2 w(vu) + n_1 w(G, T^2_{\rho''}, u, C_2)
\end{aligned} \tag{4.8}
$$

where $n_1 = |\mathcal{S}(G, T^1_{\rho'}, v, C_1)|, n_2 = |\mathcal{S}(G, T^2_{\rho''}, u, C_2)|$ are the cardinalities of the respective sets of colorful copies of $T^1_{\rho'}$ resp. $T^2_{\rho''}$ rooted at v resp. u which can be computed efficiently [4] and parallely with $w(G, T^1_{\rho'}, v, C_1)$ and $w(G, T^2_{\rho''}, u, C_2)$.

Note that each $w(G, T'_{\rho'}, v, C)$ can be computed in $O(deg(v) \cdot 2^{O(k)})$ time where $deg(v)$ is the degree of v. Thus, the computation of the total weight of colorful non-induced occurrences of T in G is in $O(|E|2^{O(k)})$ time. ◇

Theorem 4.1. *The algorithm* ApproxWeightedOccurr *(G, T, ϵ, δ) estimates the total weight of non-induced occurrences of a tree T in G with additive error ϵ and with probability at least $1 - 2\delta$ and runs in time $O(|E| \cdot 2^{O(k)} \cdot \log(1/\delta) \cdot \frac{1}{\epsilon^2})$ where $|E|$ is the number of edges in the input network.*

Proof. Now we only need to consider its running time. Notice that we need to repeat the color coding step and counting step $s \cdot t$ times and each iteration runs in time $O(|E| \cdot 2^{O(k)})$ where

$|E|$ is the number of edges in the input network. Thus, since $1/p = k^k/k! = O(e^k) = O(2^{O(k)})$, the asymptotic running time of our algorithm evaluates as

$$O(s \cdot t \cdot |E| \cdot 2^{O(k)}) = O(|E| \cdot 2^{O(k)} \cdot \log(1/\delta) \cdot \frac{1}{\epsilon^2 p}) = O(|E| \cdot 2^{O(k)} \log(1/\delta) \cdot \frac{1}{\epsilon^2}). \quad (4.9)$$

$\diamond$

4.3 Experimental Results

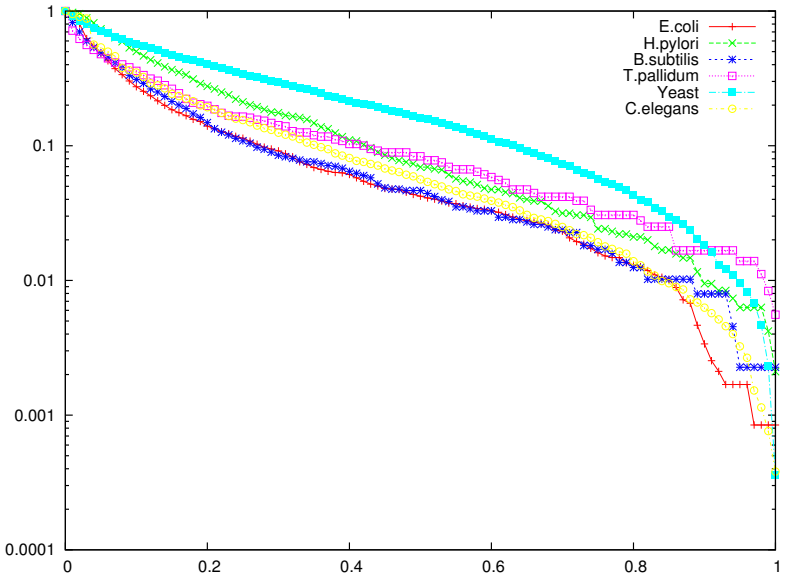
4.3.1 Data and Implementation

Weighted PPI Networks We downloaded PPI networks with confidence scores from the *STRING* database, version 8.0 [93] for the prokaryotic, unicellular organisms *E.coli, H.pylori, B.subtilis, T. pallidum*, the eukaryotic, unicellular organism *S.cerevisiae* (Yeast) and the eukaryotic, multicellular organism *C.elegans*. As mentioned above, edges are assigned to weights reflecting confidence scores which can be interpreted as probabilities that the corresponding interactions are not experimental artifacts. See Table 4.1 for some basic statistics about these networks.

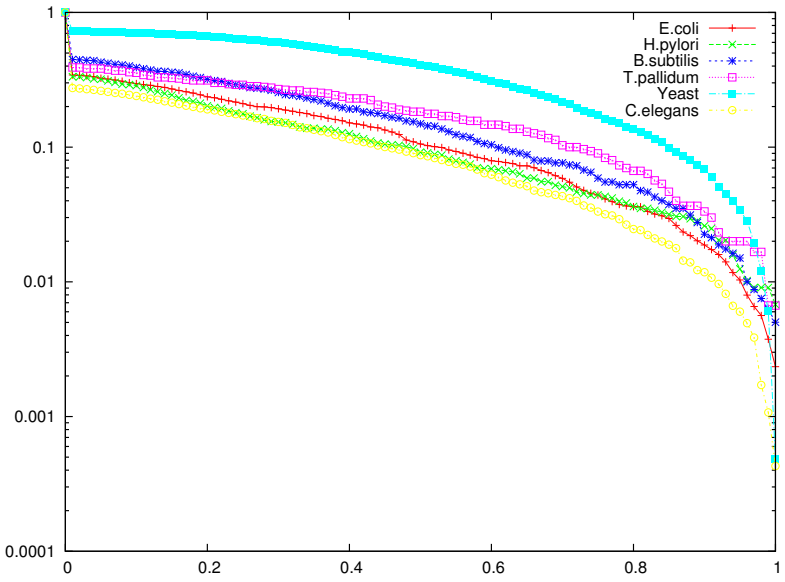
Table 4.1: Number of vertices, edges, in the studied PPI networks.

	E.coli	*H.pylori*	*B. subtilis*	*T. pallidum*	*S.cerevisiae*	*C.elegans*
Vertices	2482	1019	939	398	5913	5353
Edges	22476	9274	9184	4198	207075	43100

Other Global Properties of Confidence-Scored PPI Networks Two of the most commonly studied properties to measure the global structure of PPI networks are degree distributions and clustering coefficient distributions. First we extend the concepts of degree and clustering coefficient distributions to confidence-scored PPI networks; then we compare these *weighted degree and clustering cofficient distributions* of the studied species. For any node v in a confidence-scored PPI network G, let $N(v)$ be the set of vertices that are adjacent to v in G. We denote $deg(v)$ be expected degree of node v i.e. $deg(v) = \sum_{u \in N(v)} w(uv)$ where $w(uv)$ is the confidence score of the edge uv. Let $P(k)$ be the fraction of the number nodes that have expected degree greater than or equal to k. Since we would like to compare PPI networks with differences in the number of nodes and edges, we obtain $P'(k)$ be the fraction



(a) Weighted cumulative degree distributions



(b) Weighted cumulative clustering coefficient distributions

Figure 4.3: Weighted cumulative degree distributions (top row) and weighted cumulative clustering coefficient distributions (bottom row) of the prokaryotes *H.pylori*, *E.coli*, *B.subtilis*, *T. pallidum*, *S.cervisiae* (Yeast) and *C.elegans* PPI networks (top row)

of the number nodes that have expected degree greater than or equal to k times maximum degree in the network where $0 \leq k \leq 1$. Thus, the weighted cumulative distribution of a PPI network G is the distribution of $P'(k)$ for $0 \leq k \leq 1$. To obtain a robust weighted cumulative degree distribution, we first remove top one percent of vertices with highest degrees.

Similarly for any node v with $|N(v)| \geq 2$ in a PPI network G, let $C(v) = \{us | u, s \in N(v), us \in E(G)\}$. We denote $c(v)$ be expected clustering coefficient of v:

$$c(v) = \sum_{e \in C(v)} \frac{2 * w(e)}{|N(v)| * (|N(v)| - 1)}$$

In the case that $|N(v)| \leq 1$, we set $c(v) = 0$. Let $H(k)$ be the fraction of the number nodes that have expected clustering coefficient greater than or equal to k where $0 \leq k \leq 1$. Thus, the weighted cumulative clustering coefficient distribution of a PPI network G is the distribution of $H(k)$ for $0 \leq k \leq 1$.

To compare global network properties of unicellular prokaryotic organisms *E.coli, H.pylori, B.subtilis, T. pallidum*, the unicellular eukaryotic organism *S.cerevisiae* (Yeast) and the multicellular eukaryotic organism *C.elegans*, we obtained the weighted cumulative degree distributions and clustering coefficient distributions Figure 4.3 of these species.

As seen from the figures, the weighted cumulative distributions of degree and clustering coefficient of different species especially unicellular prokaryotic organisms and multicellular eukaryotic organism *C.elegans* are quite similar. Eventhough the distributions of Yeast are quite different from ones of unicellular prokaryotic organisms, they are pretty much similar in their shapes. These motivate us to study local network properties of these species in order to compare them.

Query Tree Topologies There are 23 and 47 possible tree topologies with 8 and 9 nodes respectively. We obtained the list of treelets from the Combinatorial Object Server [142]. Here we give the lists of all tree topologies with 8 vertices in Figure 4.4 and with 9 vertices in Figure 4.5. Note that the treelets are enumerated from 1 in these figures while they are enumerated from 0 in the normalized weighted treelet distributions.

Implementation / Choice of Parameters We implemented our algorithm ApproxWeightOccur with the multiplicative weight scheme such that the weight of a query tree can be interpreted as its probability to be present in the network, as aforementioned. We

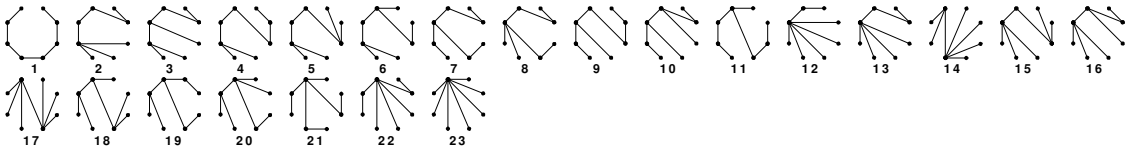


Figure 4.4: List of treelets with 8 vertices

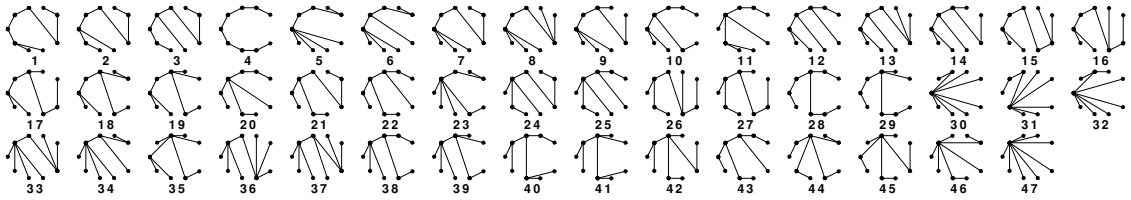


Figure 4.5: List of treelets with 9 vertices

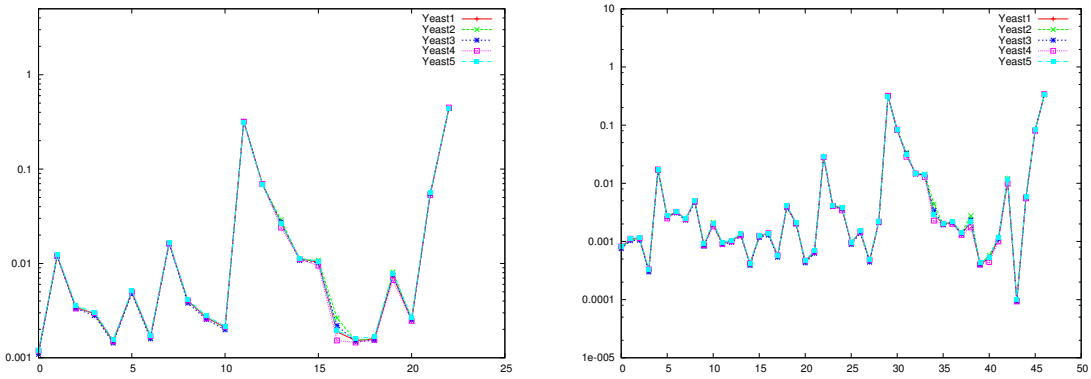
set $\epsilon = 0.01$ as the approximation ratio and $\delta = 0.001$ as the error probability. Then we computed expected numbers of occurrences of all query trees of size 8 and 9 for each of the networks described above. By normalizing the occurrences of the different query trees of size 8 resp.9 over the 23 resp. 47 different query trees, we obtained size 8 resp. size 9 treelet distributions which we refer to as *normalized weighted treelet distributions*. The idea behind normalizing expected occurrences is for comparing PPI networks with different number of nodes and edges and to increase robustness with respect to missing data which still is a considerable issue in PPI network studies. We will demonstrate the robustness by an approved series of experiments [74] in the subsequent subsection 4.3.2. Experiments were performed on a Sun Fire X4600 Server with 64GB RAM and 8 dual AMD Opteron CPUs with 2.6 Ghz speed each.

4.3.2 Robustness Analysis

In order to assess the reliability of the normalized weighted treelet distributions as a measure of weighted PPI network similarity one needs to ensure that they are robust w.r.t. small alterations to the network. This is motivated by the fact that currently available PPI data is still rather noisy, containing significant amounts of both false positive and false negative edges. In this section, we evaluate the robustness of normalized weighted treelet distributions meaning that minor changes in the weighted PPI networks do not result in drastic changes in their normalized weighted treelet distributions.

Therefore, we used the random sparsification method which was proposed in [74] and was

applied in earlier studies [4]. The method iteratively sparsifies networks by removing vertices and edges in a sampling procedure. This is based on two parameters, the bait sampling probability α_b and the edge sampling probability α_e which refer to sampling vertices and edges. As in [4], we set $\alpha_b = 0.7$ and $\alpha_e = 0.7$ and shrank the weighted PPI network of Yeast to five smaller networks accordingly with approximately the same number of nodes and edges. A comparison of the normalized weighted treelet distributions of the shrunken networks is displayed in Figure 4.6. As can be seen, the normalized weighted treelet distributions are very similar to one another which confirms the robustness of the normalized weighted treelet distributions.



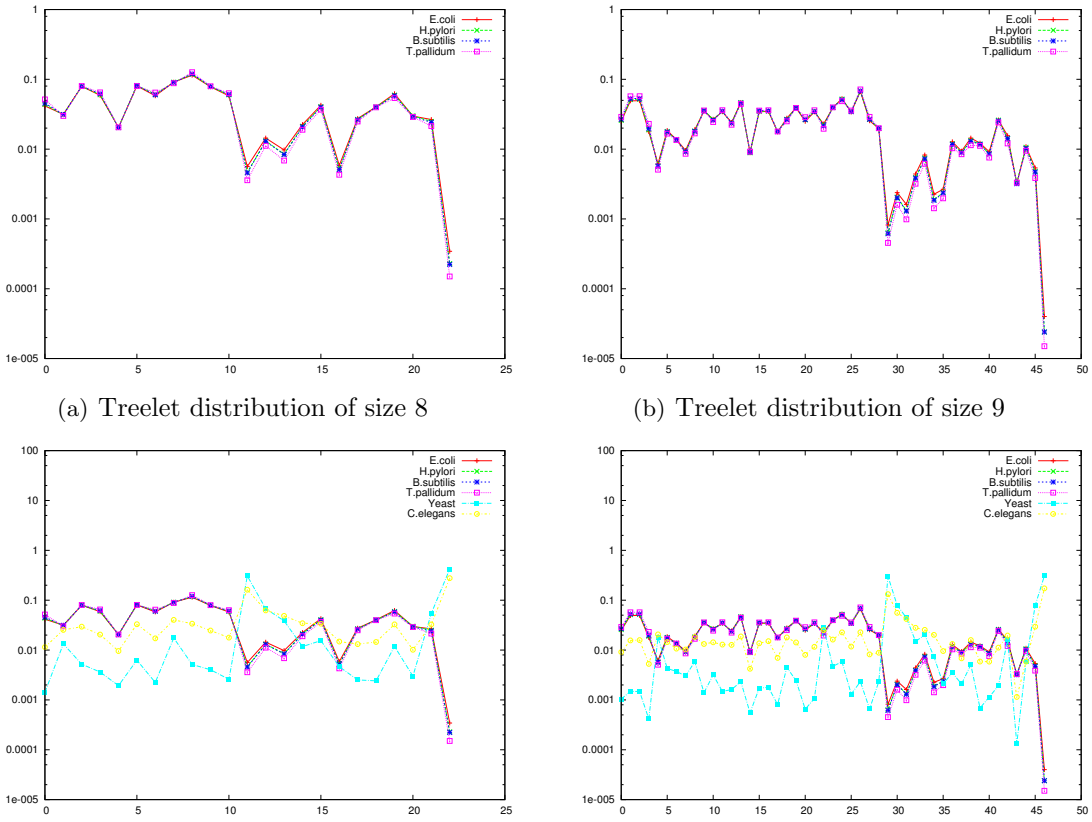
(a) Treelet distribution of size 8 (b) Treelet distribution of size 9

Figure 4.6: Normalized weighted treelet distributions of of five networks (a) size 8, (b) size 9 generated from the *S.cervisiae* Yeast PPI network with both the bait and edge sampling probability equal to 0.7.

4.3.3 Comparison of PPI Networks

In order to be able to appropriately benchmark our results against previous findings we considered the same organisms that were examined in [4]. We also considered the two prokaryotic organisms *B.subtilis* (a Gram-negative bacterium commonly found in soil) and *T.pallidum* (a Gram-negative pathogen giving rise to congenital syphilis). The corresponding weighted treelet distributions are displayed in Figure 4.7. The upper row of figures shows that the treelet distributions of the prokaryotic organisms are all similar. This is quite amazing since the weighted PPI networks have been determined in experiments which were independent of one another and without the integration of cross-species associations [93].

As can be seen in the bottom row of Figure 4.7, the treelet distributions of the Yeast PPI network is quite different from the ones of the prokaryotic organisms, which had not been observed in the boolean networks used in [4]. Still, there are obvious differences between the unicellular organisms and *C.elegans*, the multicellular model organism under consideration. It might be interesting to note that the greatest differences occur for the expected numbers of occurrences of tree topologies 23 resp. 47, which are the stars with 8 resp. 9 nodes.



(a) Treelet distribution of size 8

(b) Treelet distribution of size 9

(c) Treelet distribution of size 8

(d) Treelet distribution of size 9

Figure 4.7: Normalized weighted treelet distributions of the prokaryotes *H.pylori*, *E.coli*, *B.subtilis*, *T. pallidum* PPI networks (top row) and of the prokaryotes *H.pylori*, *E.coli*, *B.subtilis*, *T. pallidum*, *S.cervisiae* (Yeast) and *C.elegans* PPI networks (bottom row)

Chapter 5

Subnetwork Marker Discovery by Density-Constrained Biclustering

Inference of subnetwork markers comes with most demanding computational and combinatorial challenges due to the tremendous number of plausible subnetwork patterns to be examined. Here we aim at solving a combinatorial problem which particularly addresses that the same cancer can come in many different subtypes and stages of progression which cannot be necessarily distinguished by visual inspection (e.g. [162]). Namely, we address that pathways which are dysregulated in cancer can show in many, but not all cancer patients. This reflects that cancer is a most diverse disease which, nonetheless, can be classified---there are phenomena which are common to many (but not necessarily all) different specimens.

In [37] and this chapter, we present a computational strategy to solve this combinatorial search problem and show that applying it results in exhaustive enumeration of *subnetwork biclusters* that is combinations of gene and sample clusters where participating genes form dense, connected subgraphs in a PPI network. Hence our markers can be taken as (fractions of) pathways which are dysregulated in sufficiently many, but not necessarily all cancer (subtype) samples. To serve the purposes of a fair benchmarking competition we *first* perform cross-platform classification on colon cancer datasets as described in the state-of-the-art approach of [30] and outperform the prior approaches partly by raising accuracy by a relative increase of nearly 50%. *Second*, we perform cross-validation (within the same platform) experiments on breast cancer as described in [133] and outperform all approaches which yield universal, platform-independent markers. In both cases, we analyze the subnetworks

associated with our top-ranked markers.

Three things are substantially different:

1. The PPI networks we employ are confidence-scored [93].

2. Our subnetworks need not only be connected, but also need to contain a sufficient amount of edge weight (= confidence scores).

3. In our case, *all* genes in a subnetwork need to be dysregulated in a subset of patients of size at least L, but *not necessarily in all* patients. In other words, the genes of the subnetwork and the L cancer samples in which the subnetwork, as a whole, is dysregulated form a *bicluster*. See Figure 5.1 for two examples of subnetwork markers and the Problem Definition section for precise definitions.

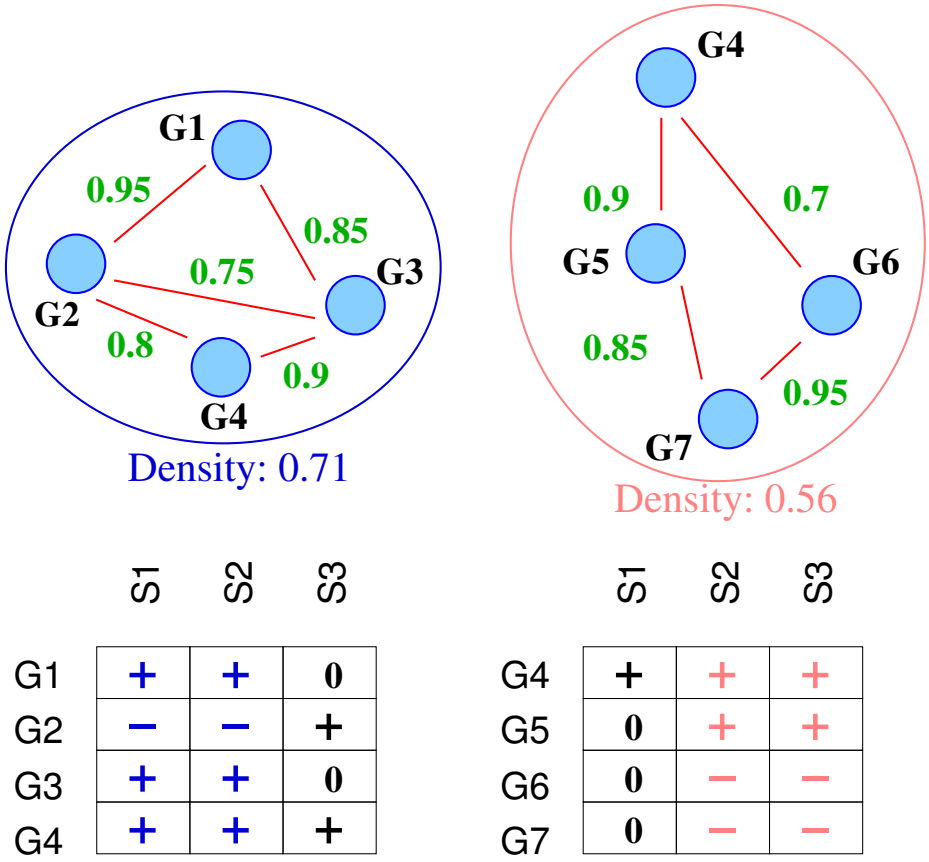
The advantage of confidence-scored physical PPI networks is that each detected physical interaction is rated by the likelihood that the interaction does play a cellular role and is not merely an experimental artifact. As a consequence, dense connectivity can be interpreted as that the genes in the subnetwork establish a cellular functional element through physically interacting with each other which comes from accumulating high confidence scores within the subnetwork [93]. In fact, many markers we compute are enriched with Gene Ontology terms whereas this is not as obvious for the previous approaches (see the section Experimental Results). The third point finally reflects the discussion from above: unlike in the previous approaches, we would like to have markers apply *as an entity* for a *sufficient percentage but not necessarily all* cancer samples.

5.1 Problem Definition

Let $G = (V, E)$ be a network where the set of nodes V is identified with the genes resp. their associated proteins and an edge $e = uv$ indicates a potential physical protein-protein interaction between the proteins associated with $u, v \in V$. We also have a weight function on the edges

$$w : E \to [0, 1]$$

where $w(e)$ is the confidence score associated with edge $e \in E$. We recall that $w(e)$ reflects our degree of belief that the physical protein-protein interaction associated with e plays a functional cellular role. In order to have gene expression experiments included in our considerations we have a *differential expression label function*



	S1	S2	S3
G1	+	+	0
G2	−	−	+
G3	+	+	0
G4	+	+	+

	S1	S2	S3
G4	+	+	+
G5	0	+	+
G6	0	−	−
G7	0	−	−

Figure 5.1: Two density constrained biclusters [see the below for the definition of density] where genes are differentially, either consistently over- (+) or under- (−) expressed in a subset of size at least 2 of cancer samples. 0 is for no differential expression.

$$D : V \to \{+, -, 0\}^K$$

which assigns a K-dimensional vector $D(v)$ with entries + (overexpressed in cancer sample), − (underexpressed in cancer sample) and 0 (not differentially expressed in cancer sample) to each of the nodes where K is the number of cancer samples in the dataset. We denote the i-th entry of $D(v)$ by $D(v)_i$ such that, for example, $D(v)_i = +$ means that gene v is overexpressed in cancer sample i. We then define:

- The *density* $\theta(G')$ of a subnetwork $G' = (V', E')$ of G is

$$\theta(G') := \frac{\sum_{e \in E'} w(e)}{\binom{|V'|}{2}} = \frac{2 \cdot \sum_{e \in E'} w(e)}{|V'|(|V'| - 1)}$$

  where $\binom{|V'|}{2}$ is the number of possible edges in G'.

- G' is called α-*dense* if

$$\theta(G') \geq \alpha$$

  where $\alpha \in [0, 1]$.

- An α-dense, connected subnetwork G' is called α-*densely connected*.

- A subset of genes $V' \subset V$ is called a *differential L-bicluster* if there is a subset $\{i_1, ..., i_L\} \subset \{1, ..., K\}$ such that

$$D(v)_{i_l} = ... = D(v)_{i_L} \in \{+, -\}$$

  for all $v \in V'$. That is each gene needs to be consistently differentially either over- or underexpressed in a subset of samples of size at least L. As an example, see Figure 5.1. There, genes $G1, G2, G3, G4$ resp. $G4, G5, G6, G7$ form a differential bicluster with respect to the samples $S1, S2$ resp. $S2, S3$.

- An α-densely connected subnetwork $G' = (V', E')$ where V' forms a differential L-bicluster is called a α-density constrained L-bicluster.

5.2 Computational Methods

We would like to devise a strategy by which to tractably mine all α-density constrained L-biclusters. To outline our strategy we define:

- A graph property is called *strong antimonotone* if in each graph of size n with the property every induced subgraph of size $n - 1$ has the property.

- A graph property is called *loose antimonotone* if in each graph of size n with the property there is an induced subgraph of size $n - 1$ with the property.

Strong antimonotonicity implies loose antimonotonicity. As a simple example consider graphs where nodes are labeled by either red or blue color. Clearly, the property to have only red nodes is strong antimonotone: all subgraphs of a red graph are red. As a simple example for loose antimonotonicity consider paths: clearly, removing either the start or the end node results in another, shorter path. However, not every node can be removed---removing internal nodes splits the path. Another loose antimonotone property on red-blue graphs is that at least half of the nodes are red. Removing blue nodes works while removing red nodes does not necessarily result in a predominantly red colored graph. We make a few observations:

- Combining a strong antimonotone with a loose antimonotone results in a loose antimonotone property. For example, in red-blue colored graphs, to be a red path is a loose antimonotone property.
- Combining a loose antimonotone with a loose antimonotone property does not necessarily result in a loose antimonotone property. Consider the property (on red-blue colored graphs) to be a path with at least half of the nodes being red. To see that this is not loose antimonotone take a path of length 4 where both start and end node are colored red whereas the two internal nodes are colored blue. Removal of none of the nodes results in a predominantly red colored path.

In the following, we will show that α-density is loose antimonotone. Before proving this, we need a lemma. And in what follows, we denote the necessary notations for proving the lemma.

We denote $G - v$ as the graph which results from removing the node v from G together with all incident edges. The diameter $\operatorname{diam}(G)$ of a graph $G = (V, E, w)$ is defined as

$$\operatorname{diam}(G) := \max_{u,v \in V} d(u, v) \tag{5.1}$$

where $d(u, v) \in \mathbb{N}$ is the length of the shortest path between u and v in terms of number of edges to be traveled (independent of edge weight). We also refer to a node v for which there is u such that $d(u, v) = \operatorname{diam}(G)$ as a *diameter node*.

The following, straightforward lemma is crucial for our main result. Note that an analogous, more general statement was established before ([136]). The diameter argument shown here, however, is crucial for the version of the proof displayed here. Therefore, we display the corresponding alternative proof.

Lemma 5.1. *Let G be a boolean connected graph and v be a diameter node. Then $G - v$ is connected.*

Proof. Let u be such that $\mathrm{diam}\,(G) = d(u, v)$. Consider an arbitrary node $u' \neq v$ and the shortest path between u and u'. The assumption that this shortest path leads through v would yield

$$d(u, u') \;=\; d(u, v) \,+\, d(v, u') \;\geq\; d(u, v) \,+\, 1 \;=\; \mathrm{diam}\,(G) \,+\, 1 \;>\; \mathrm{diam}\,(G). \quad (5.2)$$

which is a contradiction w.r.t. the definition of the diameter of a graph. Hence in $G - v$ every node u' is connected to u which yields that $G - v$ is connected. ◇

As stated above, the ideas of the lemma are related to the ideas of the proof of the subsequent theorem. Note that in this theorem the observation that one has to distinguish between graphs with different diameters establishes the technical novelty in this class of results. As abovementioned, a full compendium of observations necessary for related theorems for boolean edge-weight graphs can be looked up in [136].

We are aware that there are also several other alternative proofs, sometimes seemingly simpler, but as compensation based on strong theorems from graph theory (such as a version making use of the block tree of a connected graph). We opted to display the following proof which does not need any stronger results from graph theory. Clearly, this is a matter of taste.

Theorem 5.1. *In every connected, weighted-edge graph $G = (V, E, w)$ where $\theta(G) = \alpha \geq 1/2$ there is a node $v \in V$ such that also $G - v$ is connected and $\theta(G - v) \geq \alpha$.*

Proof.

$$\theta(G) = \frac{w(V)}{n(n-1)/2} \geq \alpha \geq 1/2 \quad (5.3)$$

translates to

$$w(V) \geq \alpha n(n-1)/2 \geq n(n-1)/4. \quad (5.4)$$

Assuming that there is a node $v \in V$ for which

$$w(v) \leq \alpha(n-1) \quad (5.5)$$

yields

$$\begin{aligned}\theta(G-v) = \frac{w(V)-w(v)}{(n-1)(n-2)/2} &\overset{(5.5)}{\geq} \frac{w(V)-\alpha(n-1)}{(n-1)(n-2)/2} \\ &\overset{(5.4)}{\geq} \frac{\alpha(n/2-1)}{(n-2)/2} = \alpha \cdot \frac{n/2-1}{n/2-1} = \alpha.\end{aligned} \tag{5.6}$$

Therefore, it suffices to show that there exists a node v in G such that $G-v$ both is connected and $w(v) \leq \alpha(n-1)$. We do this by distinguishing between graphs G with different diameters.

diam(G) > 2: Let u and u' be such that $\text{diam}\,(G) = d(u,u')$. Thanks to lemma 5.1 both $G-u$ and $G-u'$ are connected subgraphs. Therefore, it remains to show that $w(u) \leq \alpha(n-1)$ or $w(u') \leq \alpha(n-1)$.

We denote by $N(u), N(u')$ the sets of neighbors of u, u' in the boolean version G_{bool}. The choice of u, u' ($d(u,u') = \text{diam}\,(G) > 2$) yields that u and u' cannot share a neighbor. W.l.o.g. let u be the node with less neighbors which implies (G was supposed to be loop-free)

$$|N(u)| \leq \frac{|V - \{u,u'\}|}{2} = \frac{1}{2}(|V|-2). \tag{5.7}$$

Therefore,

$$w(u) \;=\; \sum_{v \in N(u)} w(uv) \;\overset{w(uv)\leq 1}{\leq}\; |N(u)| \;\overset{(5.7)}{\leq}\; \frac{|V|-2}{2} \;\leq\; \alpha(n-2) \;<\; \alpha(n-1) \tag{5.8}$$

which yields that $G-u$ is of the desired quality.

diam(G) $= 2$: Here, every node u which is not connected to every other node is a diameter node hence, due to lemma 5.1, $G-u$ is connected for any such u. To find such a node with $w(u) \leq \alpha(n-1)$ would deliver a good candidate to be removed. Therefore, it remains to treat the case where

$$w(u) > \alpha(n-1) > \frac{n-1}{2} \tag{5.9}$$

for every diameter node u. However, (5.9) translates to that every diameter node u is connected to more than half of the other nodes. Let u' be a complementary diameter node for u (i.e. $d(u,u') = 2$) and, w.l.o.g. let u be the node with less neighbors. The assumption that u and u' share at most one neighbor would then yield

$$w(u) \leq |N(u)| \leq \frac{n-2}{2} \tag{5.10}$$

which is a contradiction to (5.9). Hence every two diameter nodes share at least two neighbors. As a consequence, since every non-diameter node is connected with every other node, every pair of nodes shares at least two neighbors. [A non-diameter node shares all neighbors with other non-diameter nodes and at least $(n-1)/2+1$ with the diameter nodes, due to (5.9).] Therefore, $G-u$ is connected for all u. It remains to show that there is a node u such that $w(u) \leq \alpha(n-1)$.

However, assuming the contrary yields the contradictory

$$\theta(G) \;=\; \frac{2w(G)}{n(n-1)} \;=\; \frac{\sum_{u\in V} w(u)}{n(n-1)} \overset{(5.9)}{>} \frac{\sum_{u\in V} \alpha(n-1)}{n(n-1)} \overset{|V|=n}{=} \frac{\alpha(n-1)n}{n(n-1)} \;=\; \alpha. \quad (5.11)$$

diam(G) $= 1$: In this case, G is a clique, which again translates to that one can remove all nodes u without that $G-u$ is disconnected. Therefore, the proof proceeds completely analogously to the last part of the case diam $(G) = 2$. ◇

In our setting, we obtain the following results where $G-v$ is the subgraph of G which results from removing v and all edges incident to v:

Theorem 5.2.

1. *Every subgraph of a differential bicluster of degree at least L is a differential bicluster of degree at least L.*

2. *In every connected, weighted-edge graph $G=(V,E,w)$ where $\theta(G)=\alpha \geq 1/2$ there is a node $v \in V$ such that $G-v$ is connected and $\theta(G-v) \geq \alpha$.*

3. *In every (α, L)-density constrained bicluster $G=(V,E)$ where $0.5 \leq \alpha \leq 1.0$ there is a node $v \in V$ such that $G-v$ is a α-density constrained L-bicluster.*

In other words, theorem 5.2 establishes that to be a differential L-bicluster is strong antimonotone whereas to be an α-densely connected graph or to be a density constrained bicluster are both loose antimonotone.

Proof. Strong antimonoticity of differential biclusters is easy. If G is differentially expressed in L samples then so is any subgraph of G. And we know that dense connectivity is loose antimonotone. Since (see above) combining a strong antimonotone with a loose antimonotone property results in a loose antimonotone property, (3) follows immediately from (1) and (2).

An Enumeration Algorithm Theorem 5.2 supports a search strategy which is based on the loose antimonotonicity of density constrained biclusters and is completely analogous to that of [136] for the case $0.5 \leq \alpha \leq 1.0$ which was also employed in[35]. This strategy will yield all α-density constrained L-biclusters U for $\alpha \in [0.5, 1.0]$ which are maximal in the sense that there is no proper α-density constrained L-bicluster which contains U as an induced subgraph. This strategy applies for all loose antimonotone properties and therefore applies when mining α-density constrained L-biclusters. Subnetworks are screened in a breadth-first fashion by starting with subnetworks of size 2 and subsequently neglecting subnetworks of size $n \geq 3$ which do not contain any density constrained bicluster of size $n-1$. Loose antimonotonicity guarantees that subnetworks of size n cannot be density constrained biclusters if not containing a density constrained bicluster of size $n-1$. As was also demonstrated in [136, 35] this results in a tractable strategy when combining PPI network with gene expression data. Here, all maximal density constrained biclusters were computed in runtimes of at most $2-3$ minutes on an ordinary personal computer.

Ranking Procedure The resulting set of all density-constrained biclusters is ranked with respect to statistical significance. We randomly sampled 10^5 connected subnetworks and determined the p-value of a density-constrained L-bicluster G' as the fraction of randomly sampled subnetworks H with $\theta(H) \geq \theta(G')$ and H being consistently dysregulated in at least L samples. We select markers top-down from this p-value based ranking list while discarding biclusters where more than half of the genes are already contained in previously selected markers. We furthermore re-ranked our 50 most significantly dense modules by applying the information-theoretic criteria as described for the approaches which were employed for benchmarking. See Description of Benchmarking Partners for details.

Classification Process Classification is performed by a support vector machine approach implementing a linear kernel using Matlab's *svmclassify*. For colon cancer vs. healthy classification, the training data is identical with that used for marker computation (i.e. either GSE8671 or GSE10950). For colon cancer with vs. without liver metastasis, markers are computed using GSE8671 or GSE10950 and classification is performed by leave-one-out cross-validation in GSE6988. This coincides with the procedures described in [30]. For breast cancer TP53 wildtype vs. mutant markers are computed using GSE3494 and classification is performed by leave-one-out cross-validation in the same dataset. The breast cancer classification scheme is the *only non-cross-platform* experiment. In colon cancer data used

for marker computation and classification test data come from two different platforms. For *feature space construction*, we choose the best K markers to obtain a feature space of dimension K. Each sample j is transformed into a K-dimensional vector $A(j) \in \mathbb{R}^K$ where the entries $A(j)_k$ for each marker k are $A(j)_k := \sum_v E(v, j)/K$ where v ranges over all genes v contained in the subnetwork associated with marker k. In other words, each sample j becomes a point $A(j)$ in the K-dimensional marker feature space $\mathbb{R}^K$.

5.3 Experimental Results

5.3.1 Network Data and Cancer Datasets

Network Data We downloaded the licensed protein-protein interaction network from the STRING database, version 8.1 [93]. In this version, STRING network consists of 9927 proteins and 62539 edges. Edges have a positive confidence score in case that there is evidence that the two proteins in question *physically* interact within a cellular context. We opted to exclusively treat physical interactions since comparison partners only considered (ordinary) physical protein-protein interaction networks. Note that their methods do not allow to make use of edge weights. For these methods unweighted PPI network data as described in [30] was used.

Colon Cancer Gene Expression Data In analogy to [30]'s study we treated the microarray datasets with the accession numbers GSE8671, GSE10950 and GSE6988 from the Gene Expression Omnibus [60] database. GSE8671 contains 8987 gene expression profiles across 32 prospectively collected adenomas with those of normal mucosa from the same individuals [163]. GSE10950 contains 18171 gene expression profiles across normal and tumor pairs [97]. GSE6988 contains 17104 gene expression profiles for 25 normal colorectal mucosa, 27 primary colorectal tumors, 13 normal liver, 27 liver metastasis and 20 primary colorectal tumors without liver metastasis [109].

Breast Cancer Gene Expression Data We considered the gene expression dataset GSE3494 treated in [133] along with all available additional information. Experiments performed in [133] aim at predicting TP53 mutation status, tumor grade and survival time. Therefore, they first identify platform-specific (Affymetrix U133 A and B) probes as being correlated with TP53 mutation and estrogen receptor status as well as tumor grade,

using multivariate linear regression from their own data. Subsequently, they select 32 such platform-specific probes as being the features which yield best accuracy when performing cross-validation on their own data. This means that accuracy values cannot be taken as unbiased results since feature selection is based on the outcome of the cross-validation.

Differential Expression For the colon cancer datasets GSE8671 and GSE10950 we determine differential expression as follows. We first normalize expression values for each gene v individually. Let $E(v, j)$ be the resulting normalized expression value for gene v in sample j. We then determine the top 10% of the values $E(v, j)$ in each sample j and declare them ''overexpressed''. In both datasets samples come in pairs cancer vs. healthy. Let j_1 be the cancer and j_2 be healthy sample for one patient l. We then put $D(v)_l = +$ resp. $D(v)_l = -$ if v is overexpressed in j_1, but not in j_2 resp. the other way round.

In the breast cancer dataset GSE3494 (see below) we determine a normal distribution for all values and normalize the entire data accordingly. For an arbitrary sample l let $E(v, l)$ be the corresponding normalized value. Subsequently, $D(v)_l = +$ for a sample l if $E(v, l)$ is among the top 5% resp. $D(v)_l = -$ if $E(v, l)$ is among the lowest 5%.

Description of Benchmarking Partners

The idea which is common to the majority of prior approaches is to aim at inferring genes g whose gene expression profiles $E(g) \in \mathbb{R}^K$ (where K is the number of samples) share large *mutual information* with the phenotype profile $P = (1, ..., 1, 2, ..., 2) \in \{1, 2\}^K$ where $P_k = 1$ if k is a cancer sample and 2 if not (dimensions k are ordered such that cancer tissue samples come before healthy tissue samples). Mutual information is an information-theoretic concept which here can be taken as a measure for the similarity of a K-dimensional vector with P. For groups of genes $g_1, ..., g_N$ mutual information is determined by using the average gene expression profile $\sum_{i=1}^{N} E(g_i)/\sqrt{N}$. Correspondingly, *single gene markers* are computed as genes g where $E(g)$ achieves maximal mutual information with P regardless of any network considerations. *[31]*, as the first subnetwork marker approach, greedily collect genes g which form a connected subnetwork in the PPI network such that the average expression profile has high mutual information with P. *[30]* aim at finding groups of genes which have low *network distance* where network distance equal to 1 translates to a connected subnetwork. The algorithm described (NetCover) then finds minimal groups of genes having bounded network distance such that the group *covers* the samples. This translates that for each

sample, one of the genes in the group is dysregulated. A theorem then points out that such groups achieve good mutual information.

In all of these approaches subnetwork markers are ranked according to the average mutual information they achieve. Note that the subnetworks produced by [30]'s method are not necessarily connected. None of the approaches from above follows the idea that certain subnetworks might be dysregulated in some, but not all samples. In fact, expression levels of genes in one subnetwork can vary across the patients, individually for each gene while trying to cover as many patients as possible. Our approach specifically prevents this.

5.3.2 Analysis of Colon Cancer Dataset

Marker Computation We computed and subsequently ranked subnetwork markers as described in the Computational Methods section both using GSE8671 (parameter choices: $\alpha = 0.5, L = 3$) and GSE10950 ($\alpha = 0.5, L = 2$). Parameters were chosen as non-restrictive as possible such that the total number of subnetwork markers did not exceed 1000. Throughout this section, our method is referred to as wDCB. Since GSE6988 does not contain paired cancer/control samples one cannot compute markers as described in [30] We also computed and ranked subnetwork markers as described in [31] (GMI), single gene markers (SGM) as described in [30] and were provided with subnetwork markers by [30] extracted from GSE8671, accordingly ranked (NETCOVER=NC). However, we were neither provided with the subnetwork markers from GSE10950 nor the implementation of the NC algorithm. In the following, values for [30] referring to subnetwork markers extracted from GSE10950 are adopted from their paper.

Classification Performance We was performed as described in the Computational Methods section, using support vector machines for both GMI and NC as was evaluated as yielding maximal predictive power in both cases [30]. Predictions refer to predicting cancer vs. healthy tissue resp. with vs. without liver metastasis (henceforth referred to as ''Prognosis'') in GSE6988 using the markers from GSE8671 and GSE10950. Note that we cannot display certain values referring to markers from GSE10950 for NC since we were not provided with the corresponding subnetworks nor the software. In the following, positives (=P) and negatives (=N) are cancer and healthy resp. liver metastasis and non-liver-metastasis tissue samples such that true resp. false positives resp. negatives (=TP,FP,TN,FN) are correctly resp. misclassified cancer/metastasis resp. healthy/non-metastasis samples.

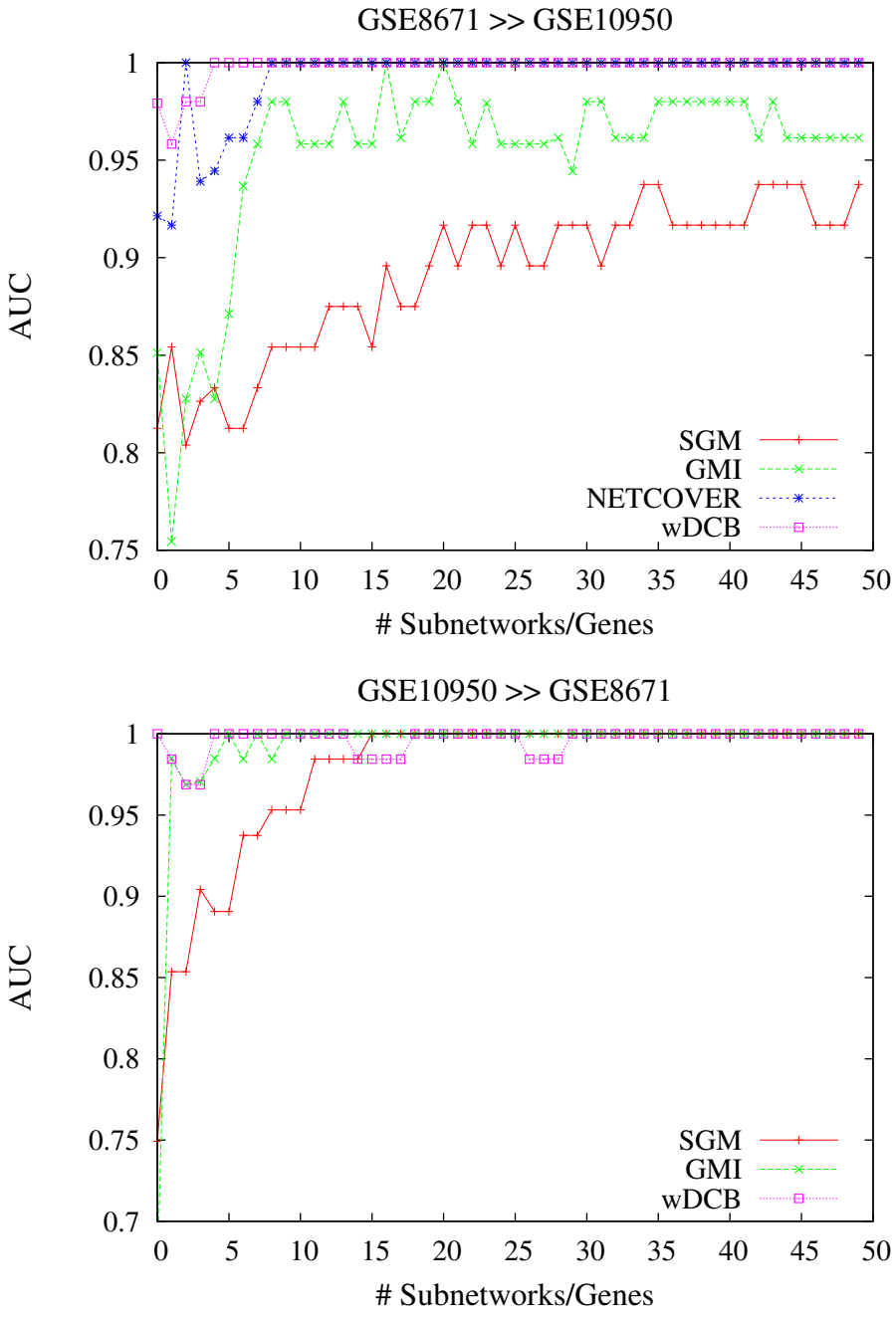


Figure 5.2: AUC for different choices of numbers of subnetwork markers using markers extracted from GSE10950 for cancer vs. non-cancer prediction of GSE8671. Values for NETCOVER were not available.

Figure 5.2 displays AUC (area under the precision-recall curve) which is computed as the arithmetic average of precision ($= TP/(TP + FP)$) and recall ($= TP/(TP + FN)$) for different choices of subnetwork markers and the two prediction tasks where markers are chosen according to the corresponding rankings.

Note that values for NC using markers from GSE10950 are missing due to the above-mentioned reasons. In [30] an average AUC of 0.86 is reported for prediction of GSE6988 (wDCB: 0.91, see Table 5.2 for more information).

See also Figure 5.3 and 5.4 for plots referring to making predictions in GSE8671 from GSE10950 and vice versa; the corresponding results are rather negligible---every competitor achieves AUC / Accuracy close to 100%. We recall that GSE6988 is the most difficult dataset, due to size (123 samples) and comprehensiveness in terms of subtypes and stages of progression.

K	SGM	GMI	NC	wDCB	SGM	GMI	NC	wDCB
	8671→6988				10950→6988			
1	0.56	**0.84**	0.72	**0.84**	0.63	0.37	N/A	**0.77**
5	0.73	0.72	0.72	**0.82**	0.82	0.68	N/A	**0.86**
10	0.76	0.76	0.83	**0.85**	0.82	0.81	N/A	**0.88**
20	0.80	0.84	0.86	**0.89**	0.84	0.83	N/A	**0.89**
30	0.80	0.83	0.84	**0.91**	0.83	**0.85**	N/A	**0.85**
40	0.85	0.85	0.87	**0.90**	0.84	0.84	N/A	**0.89**
50	0.85	0.84	0.85	**0.93**	0.81	0.82	N/A	**0.89**
	8671→6988, Prognosis				10950→6988, Prognosis			
1	**0.57**	**0.57**	0.51	0.56	0.57	**0.68**	N/A	0.47
5	**0.74**	0.62	**0.74**	0.6	0.63	**0.81**	N/A	0.68
10	0.76	0.77	0.74	**0.88**	0.57	**0.77**	N/A	0.74
20	0.72	0.62	0.77	**0.83**	0.61	0.79	N/A	**0.85**
30	0.65	0.74	0.83	**0.88**	0.63	0.81	N/A	**0.85**
40	0.67	0.79	0.83	**0.90**	0.78	0.85	N/A	**0.89**
50	0.74	0.77	0.81	**0.92**	0.76	0.85	N/A	**0.91**

Table 5.1: Accuracy for varying numbers K of markers relating to experiments on colon cancer. NC=NETCOVER. Boldface: top score. NC 10950 subnetworks are not available. See Supplement for sensitivity and specificity values

We also display Accuracy ($= (TP + TN)/(P + N)$) values in table 5.1 for predictions using GSE8671 markers, which are available for all competitors. See also Tables 3 and 4 in the Supplement for more values on cancer vs. non-cancer including sensitivity (recall) and specificity ($= TN/N$) which translate to fractions of correctly predicted cancer resp. healthy samples. We do think that sensitivity/specificity/accuracy statistics make most sense.

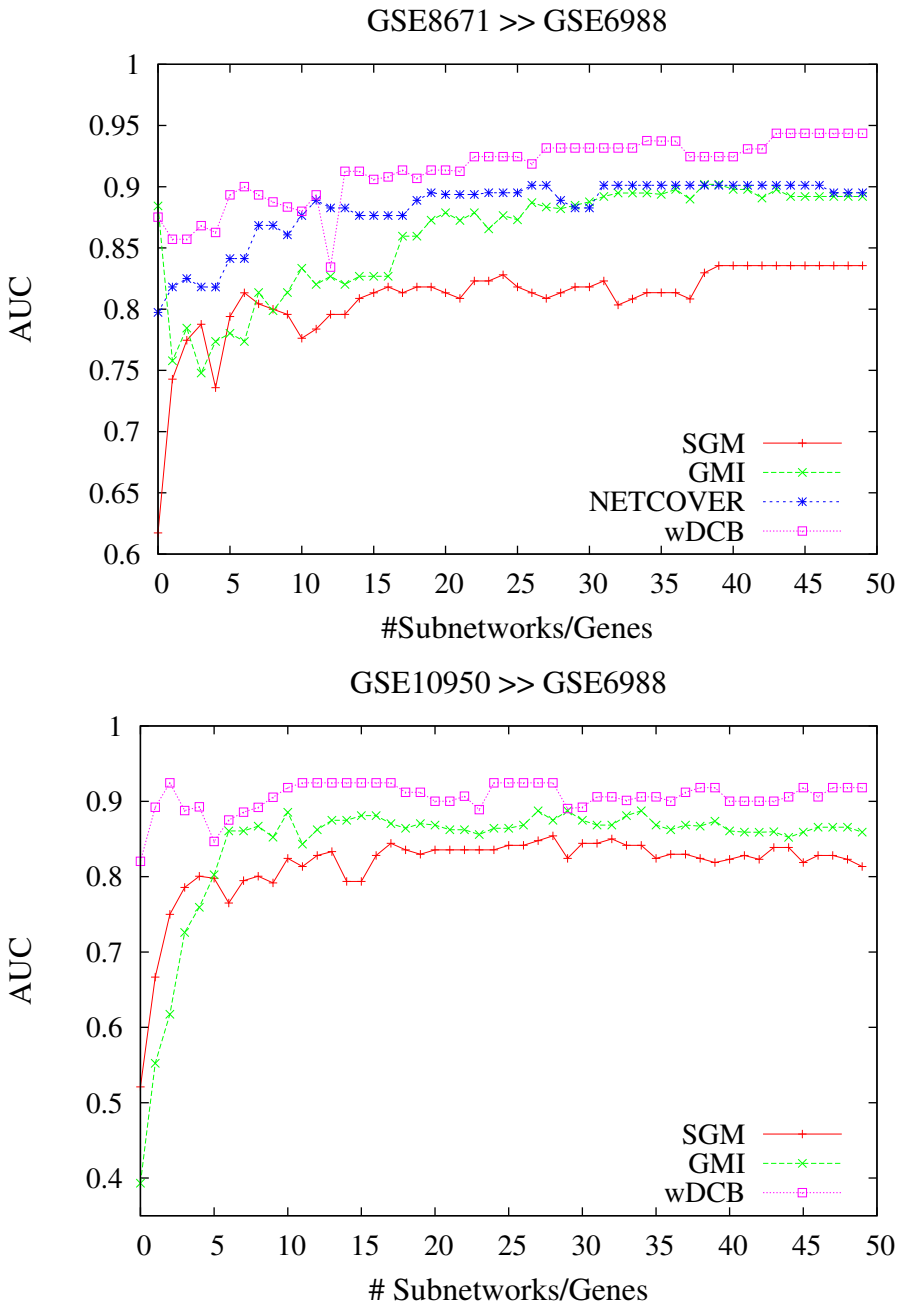


Figure 5.3: Colon cancer: AUC vs. numbers of subnetwork markers using markers extracted from GSE8671 and GSE10950 for cancer vs. non-cancer prediction in GSE6988.

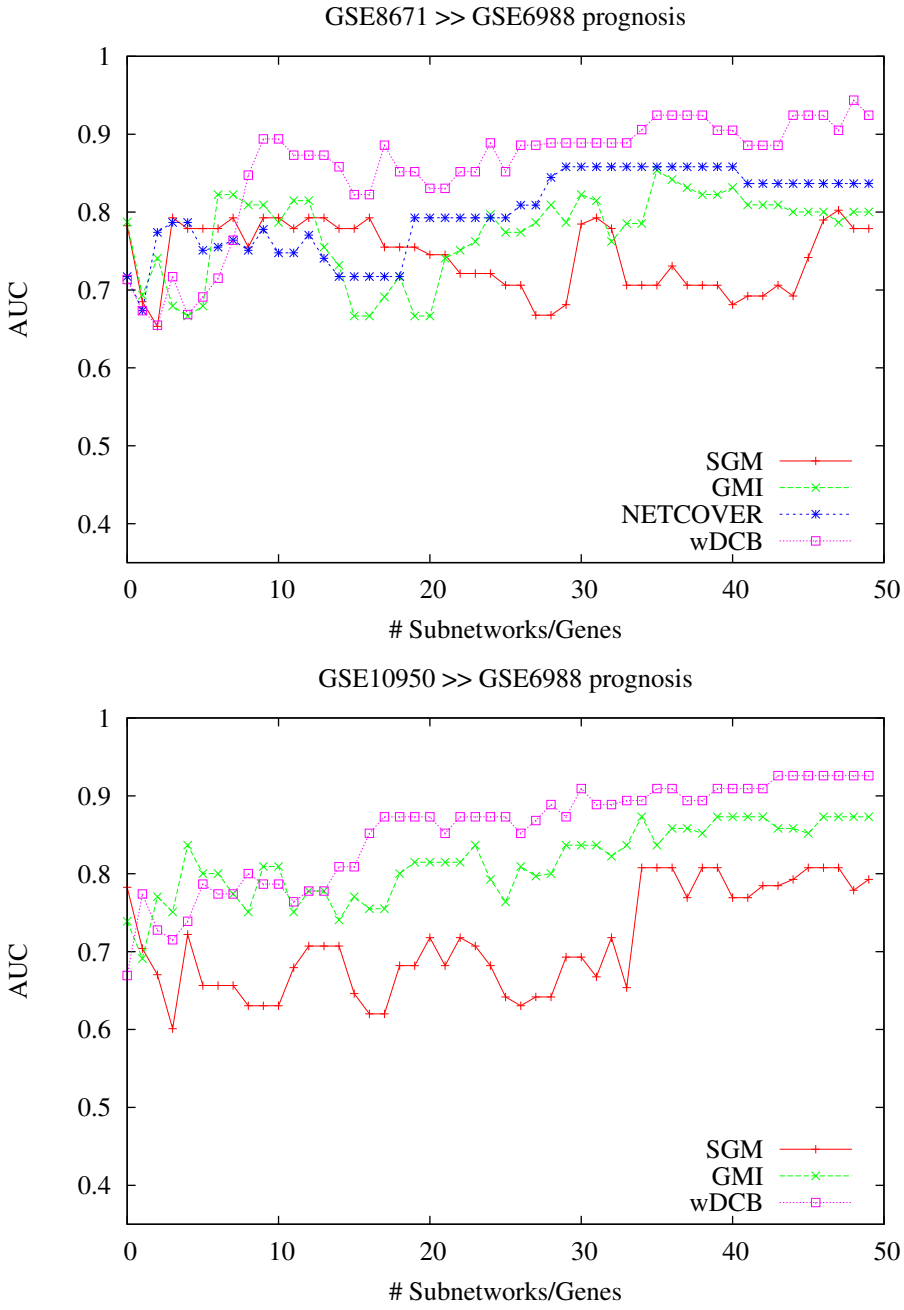


Figure 5.4: Colon cancer: AUC vs. numbers of subnetwork markers using markers extracted from GSE8671 and GSE10950 for liver metastasis vs. non-metastasis prediction in GSE6988.

However, for fairness reasons, we followed the workflow scheme of [30] which is based on AUC. Overall, our method outperforms all competitors both when predicting cancer vs. non-cancer and metastasis vs. non-metatastis. In the latter case, when using subnetworks from GSE8671, the increase in accuracy from 83%, the best value obtained by the competitors, to 92%, obtained by our method wDCB is quite remarkable. Note that this is a relative increase of more than 50% (9% out of possible 17%) translating to more than 50% less misclassified samples. In conclusion, our method proves best on a difficult colon cancer dataset in all categories tested, raising accuracy beyond 90% as the only method in three test cases.

	8671 Subnetworks				10950 Subnetworks			
	#	ER-50	6988	10950	#	ER-50	6988	8671
GMI	806	0.38	0.86	0.95	755	0.34	0.84	0.99
NC	923	0.12	0.87	0.99	N/A	N/A	0.86	N/A
wDCB	282	0.76	0.91	1.00	216	0.74	0.91	1.00

Table 5.2: Head-to-head statistics for subnetwork marker approaches: [#]: total number of subnetwork markers computed, [ER-50]: Gene Ontology Enrichment of the top 50 subnetwork markers, [6988], [10950] and [8671]: Average AUC when classifying GSE6988, GSE10950 and GSE8671 with the top 50 markers

Enrichment Analysis In table 5.2 we also display statistics for the subnetwork marker methods on the total number of networks and the Gene Ontology (GO) term enrichment of the top 50 markers used for prediction. Average AUC refers to averaging values in the plots of Figure 5.2. Values for NETCOVER on markers extracted from GSE10950 have been adopted from the corresponding paper [30]. Enrichment is based on statistical significance ($p = 10^{-3}$) relative to the hypergeometric probability distribution, Bonferroni corrected for multiple testing, computed by making use of Gene Ontologizer [15]. The obviously superior enrichment rate of wDCB subnetwork markers (76% resp. 74%) certainly is an explanation for their high quality and covenient interpretability in terms of cellular contexts. Note that NETCOVER subnetworks need not be connected which might explain the relatively inferior enrichment. Personal communication revealed that the authors expect the connected components to be more enriched. In conclusion, there is good evidence that our subnetwork markers are biologically more meaningful.

Top Subnetwork Markers GO enrichment analysis of the 186 genes identified in the top subnetworks from GSE8671 revealed a significant role for genes involved in the biological

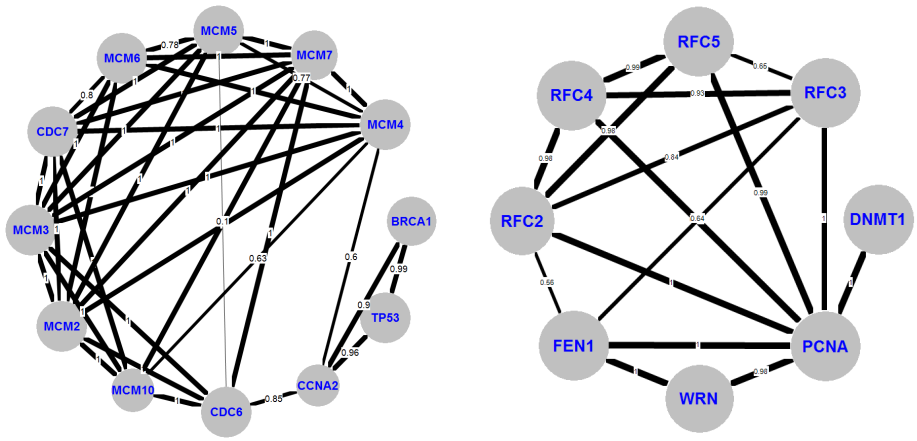


Figure 5.5: Top-ranked subnetwork markers both extracted from GSE8671 (top) and GSE10950 (bottom). Edge weights indicate confidence scores.

Dataset	Ranking	Genes in Subnetworks	Enriched GO Terms
GSE8671	1	BRCA1 CCNA2 CDC6 CDC7 MCM10 MCM2 MCM3 MCM4 MCM5 MCM6 MCM7 TP53	DNA replication initiation (p=6.39e-14) DNA replication (p=2.13e-13) DNA metabolic process (p=6.15e-12)
	2	BRCA1 CHEK1 EXO1 FEN1 MLH1 MRE11A MSH2 MSH6 PCNA PRKDC RAD51 TP53	DNA repair (p=6.44e-18) Cellular response to DNA damage stimulus (p=5.29e-17) Response to DNA damage stimulus, (p=1.905371e-16)
	3	CD2 CD247 CD28 FYN ITKLCK LCP2 PTPN22 PTPN6 PTPRC ZAP70	Leukocyte activation (p=2.70e-8) Cell receptor signaling pathway (p=3.08e-8) T cell activation (p=5.53e-8)
GSE10950	1	DNMT1 FEN1 PCNA RFC2 RFC3 RFC4 RFC5 WRN	nucleotide-excision repair (p=5.01e-11)) Protein-DNA loading ATPase activity (p=3.08e-10) DNA clamp loader activity (p=3.08e-10)
	2	CDC6 DBF4 GMNN MCM10 MCM4 MCM6 MCM7	DNA replication (p=1.61e-10) DNA metabolic process (p=4.96e-8) DNA replication initiation (p=1.42e-7)
	3	CCNA2 CDC2 CDC25C CKS1B MYC RBL1 SKP2	cell cycle (p=2.94e-6) Cell cycle process (p=1.57e-3) Cell division (p=4.57e-3)

Table 5.3: Analysis of gene and gene ontology term content of the three top-ranked subnetworks both extracted from GSE8671 and GSE10950

processes of DNA replication, DNA metabolic process, DNA repair and cell cycle progression (Bonferonni corrected, $p < 1e - 20$). In particular tumor suppressor genes such as TP53, BRCA1 and mis-match repair genes MLH1, MSH2 and MSH6 all well characterized genes known to be involved in colon cancer tumorigenesis [52] are featured in the top ranked subnetworks. The top ranked subnetwork contains TP53 and most of the minichromosome maintenance (MCM) complex components, which are essential for replication of DNA during cell division. In particular MCM2 and MCM5, have been shown to be early markers for CRC [22] and overall almost all CRC display dysregulation of the TP53 pathway through mutations or other means of functional inactivation.

See Figure 5.5 and Table 5.3 for pictures and additional statistics on our colon cancer top markers. See also Table 5.2 for a comparative enrichment analysis of all subnetwork marker approaches which reveals that $\approx 75\%$ of our colon cancer top markers are enriched with GO terms which substantially differs from other subnetwork approaches (at most 38% of the top markers are enriched).

5.3.3 Analysis of Breast Cancer Dataset

Here, we use [133] as a guideline. We focus on TP53 mutation status and predict wildtype (wt) vs. mutant (mt), a binary classification task. We first compute markers from GSE3494 and subsequently employ the suggested leave-one-out cross-validation scheme in the same dataset. As has been recently pointed out [49, 48, 31] non-cross-platform evaluations (marker computation and classification are performed in the same dataset) come with two issues: first, they are biased towards markers which do not have to rely on mapping probes to well-established gene identifiers and second, single gene markers traditionally ‘‘overperform’’, i.e. when using them for classification on other platforms their predictive power tends to significantly decrease. We recall that cross-platform stability is a major source of motivation for subnetwork marker approaches. In the following we will distinguish between Single Probe Markers (SPM) that is a single gene marker approach making use of all probe data available in GSE3494 even if probes cannot be mapped (possibly reflecting non-coding RNA etc.) [1] SGM (single gene markers) which is the equivalent of SPM using only mappable gene probes, GMI [31] and our approach wDCB which both rely on mapping probes onto nodes in PPI

[1]The signature genes reported in [133] are 32 such probes chose such as to achieve maximum training accuracy in the cross-validation scheme.

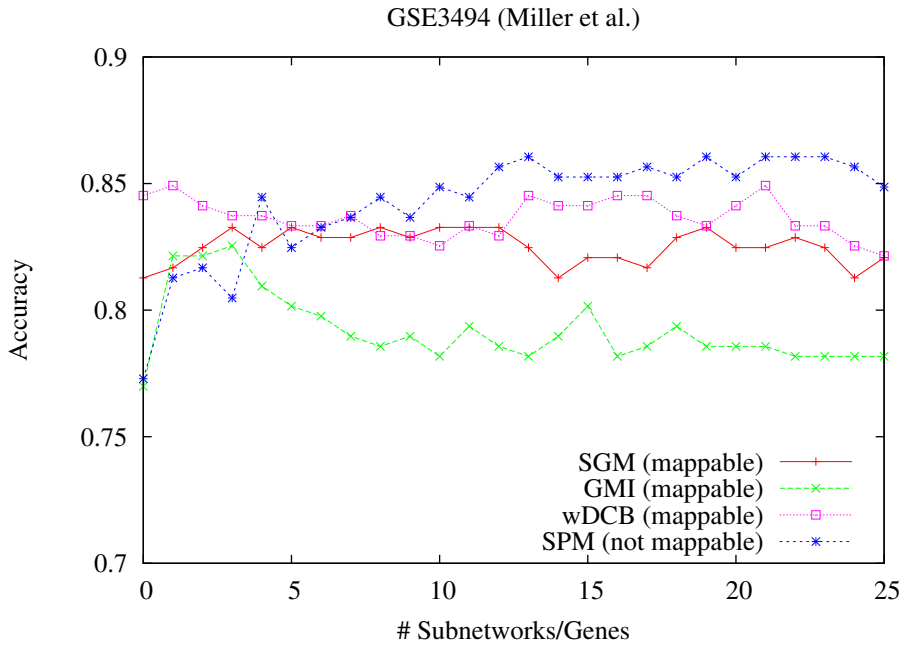


Figure 5.6: Breast cancer: Accuracy vs. numbers of subnetwork markers using markers extracted from GSE3494 for predicting TP53 mutation status (wildtype vs. mutant) in GSE3494 (leave-one-out cross-validation).

networks.

Marker Computation We computed markers for SPM, SGM, GMI and wDCB as described in the Computational Methods section. For wDCB, we used parameters $\alpha = 0.5, L = 5$ again chosen as being most non-restrictive while keeping the computed numbers of subnetworks below 1000.

Classification Performance We plotted accuracy vs. different numbers of markers (see Figure 5.6) and observed that for more than 25 markers none of the methods achieved further improvements. The non-mappable SPM achieve maximum accuracy for numbers of markers between 5 and 25 whereas wDCB achieves best values for choosing only up to 5 top-ranked markers. Among the approaches generating universally mappable marker sets, wDCB performs best. Note that, as was reported in previous studies, it is reasonable to assume that the mappable single gene marker set SGM will suffer from decreased performance

rates in cross-platform evaluations [48, 31] whereas such effects have not been reported for subnetwork marker approaches. We conclude that our approach wDCB comes is of substantial value also in breast cancer subtyping.

Top Subnetwork Markers Here we focused on the role of TP53, whose expression signature was used previously to classify prognostic classes in two breast cancer and one liver cancer cohorts with known TP53 status [133]. We found similar enrichment of GO terms such as DNA replication, DNA metabolic process and cell cycle progression (Bonferonni corrected, $p < 1e - 20$) in the 174 genes identified in our top subnetworks used for classification of TP53 mutational status. Furthermore, the subnetworks identified for known TP53 status in breast cancer were comprised of many of the same genes identified in the colon cancer subnetwork analysis (65 genes in total, $\approx 35\%$ overlap). Given the well characterized role of dysregulated TP53 signaling (e.g. caused by TP53 mutations) in both colon and breast cancers, these findings suggest that in addition to its utility for developing multivariate classifiers, DCB may also have additional functionality for extracting biologically relevant networks of genes.

Chapter 6

Optimal Subnetwork Markers Predict Drug Response

In the treatment of cancers, patients presenting tumors with similar clinical characteristics will often respond differently to the same chemotherapy [211]. In fact, for many types of cancer, only a minority of treated patients will observe regression of tumour growth. This is the case for both conventional chemotherapeutic agents and newer targeted therapies that affect specific molecules. To achieve an effective cancer treatment, it is critical to identify the underlying mechanisms that confer chemoresistance in some tumours but not others.

The advent of genome-wide expression profiling technologies has allowed the discovery of novel biomarkers for cancer diagnosis, prognosis and treatment [211]. While some progress has been made towards identifying reliable prognostic markers for breast and other cancers, development of molecular markers predictive of response to chemotherapy has proved to be far more difficult [211].

In recent years, a number of studies have used genome-wide expression profiling to identify genes that could be used as predictors of drug response in breast cancer [81, 33]. In these studies, single gene marker methods were used, where each gene is individually ranked for differential expression and the top genes were selected as predictors known as single gene markers. Other studies [117, 123] required single gene markers not only to be differentially expressed but also to have similar coexpression between the training and test cohorts. While some of these predictive markers have shown promising results in a limited number of patient cohorts, many of these signatures have failed to achieve similar

performance in additional validation studies [19]. In addition, single gene markers developed from different cohorts have been shown to have very little overlap [48]. A further limitation of single gene markers is that they provide relatively limited insight into the biological mechanisms underlying response to drug response. Thus, predictive markers with robust performance, greater reproducibility and improved insights into drug action - which are critical for clinical application - still remains elusive.

Motivated by the limitations in predicting drug response using single gene markers and the better performance promised by subnetwork markers, this chapter aims to identify subnetwork markers to predict chemotherapeutic response.

Available approaches for subnetwork marker discovery have a number of disadvantages. Network based approaches such as [31, 55, 29] are heuristic methods and thus do not guarantee the optimality of the solution for marker selection - an optimal solution would presumably provide a better prediction performance. The branch and bound approach [30] or exhaustive enumeration using biclustering in the previous chapter [37] can yield an optimal solution under some fixed set of parameters; however their worst-case running time can be super-polynomial (and hence intractable). Therefore, there is a keen need for designing efficient algorithms to retrieve the optimal subnetwork markers that could successfully distinguish samples from different classes.

In [39] and this chapter, we introduce a novel and efficient randomized algorithm to compute "optimally discriminative" subnetworks for classification of samples from different classes. The discriminative score is calculated as the difference between the total distance between samples from different classes and the total distance between samples from the same class. Our algorithm is based on the color coding paradigm [6], which allows for identifying the optimally discriminative subnetwork markers for any given error probability. Since the running time of our algorithm is a logarithmic function of the error probability, we can set the error probability to a small value - close to zero - while the running time does not increase much. When the maximum size of a subnetwork is $k = O(\log n)$ where n is the size of the network, we have a polynomial time algorithm with a fixed error probability. Since the discriminative score is additive, we can easily adapt our method to retrieve subnetwork markers to distinguish samples from more than two classes. This is very helpful in particular when there are more than three categories for responses to treatment: complete, partial and non-response.

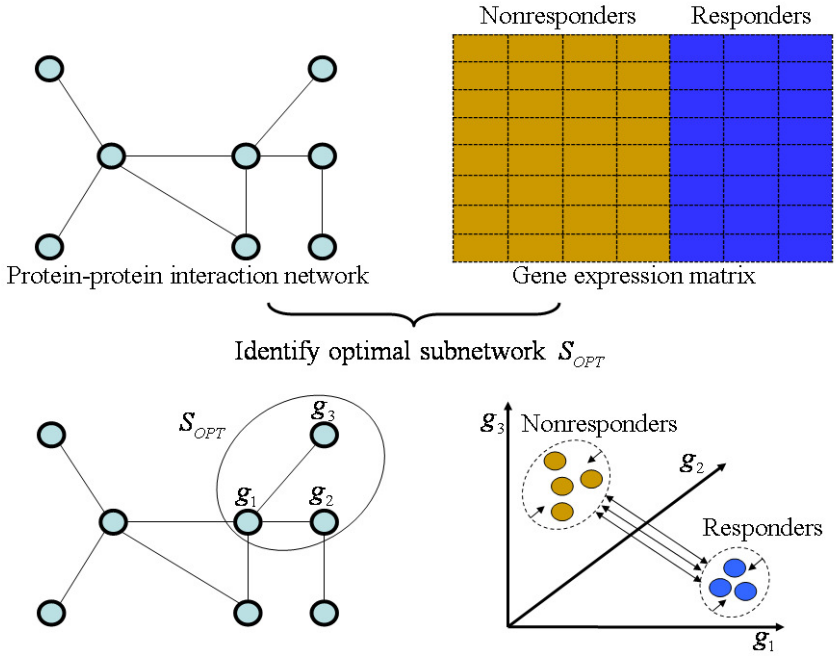


Figure 6.1: The main idea behind our approach: samples (denoted as points in a high dimensional space) are projected into k-dimensional space while ensuring that samples from the same class are clustered together, while samples from different classes stay separated. These k dimensions/genes have to form a connected subnetwork in a PPI network. The main difference between our approach and earlier ones is that we can identify the optimal subnetwork S_{OPT} in polynomial time when $k = O(\log n)$; here n is the size of the network. This is done by minimizing the total distance of samples from same class while maximizing the total distance of samples from different classes.

6.1 Problem Definition and Its Complexity

In our methodology, each patient sample is represented as a point in high dimensional space where each dimension represents one gene. We perform dimensionality reduction by projecting samples (points) into a subspace of at most k dimensions such that samples from different classes are well separated. The separation criteria is defined based on minimizing the distances of samples from the same class while maximizing the distances of samples from different classes. Figure 6.1 sketches the idea behind our approach.

We formalize our problem as the Optimal Discriminating k-Subnetwork (ODkS) problem

below. We then assess the complexity of the problem and finally give a randomized algorithm to solve it for any given error probability.

Before formally defining ODkS problem, we would like to introduce the notations used. Without loss of generality, we assume that we have only two classes of samples: positive and negative. Note that it is easy to generalize our approach for more than two classes. Let A and A' denote the expression matrices for positive and negative samples respectively. For each gene g_i, let A_i and A'_i respectively denote the expression profiles of gene g_i in positive class and negative class. For expression matrix A (A'), let $A_i(j)$ ($A'_i(j)$) denote the expression of g_i in sample j.

Given n genes, let a and a' denote the number of samples in positive class and negative class respectively. We denote the PPI network by $G = (V, E)$, where $|V| = n$ and $|E| = m$.

We define the weight function *score* on subnetwork S as the difference between the total of distance between samples from different classes and the total of distance between samples from the same class - under L_1 distance:

$$\begin{aligned} score(S) &= \sum_{j=1}^{a}\sum_{j'=1}^{a'}\sum_{\forall i:g_i\in S}\frac{|A_i(j)-A'_i(j')|}{aa'} \\ &\quad -\sum_{j=1}^{a}\sum_{j'=1}^{a}\sum_{\forall i:g_i\in S}\frac{|A_i(j)-A_i(j')|}{aa} \\ &\quad -\sum_{j=1}^{a'}\sum_{j'=1}^{a'}\sum_{\forall i:g_i\in S}\frac{|A'_i(j)-A'_i(j')|}{a'a'} \end{aligned}$$

The ODkS problem asks to compute the connected subnetwork S_{OPT} ($|S_{OPT}| \leq k$) from G such that S_{OPT} ''distinguishes'' samples from different classes ''optimally'', i.e. $score(S_{OPT})$ is the maximum among $score(S)$'s for any connected subnetwork S. We call S_{OPT} the optimally discriminative subnetwork.

For any connected subnetwork S, $score(S)$ could be rewritten as:

$$\begin{aligned} score(S) = \sum_{\forall i:g_i\in S}\Bigg(&\sum_{j=1}^{a}\sum_{j'=1}^{a'}\frac{|A_i(j)-A'_i(j')|}{aa'} \\ &-\sum_{j=1}^{a}\sum_{j'=1}^{a}\frac{|A_i(j)-A_i(j')|}{aa}-\sum_{j=1}^{a'}\sum_{j'=1}^{a'}\frac{|A'_i(j)-A'_i(j')|}{a'a'}\Bigg) \end{aligned}$$

We will extend the discriminative score function *score* so that it can apply on a single gene.

We assign each gene g_i to a weight $score(g_i)$:

$$score(g_i) = \sum_{j=1}^{a}\sum_{j'=1}^{a'} \frac{|A_i(j) - A'_i(j')|}{aa'}$$
$$- \sum_{j=1}^{a}\sum_{j'=1}^{a} \frac{|A_i(j) - A_i(j')|}{aa} - \sum_{j=1}^{a'}\sum_{j'=1}^{a'} \frac{|A'_i(j) - A'_i(j')|}{a'a'}$$

Now we can rewrite the discriminative score of a connected subnetwork S as:

$$score(S) = \sum_{\forall i: g_i \in S} score(g_i)$$

Thus, identifying the optimally discriminative connected subnetwork S_{OPT} ($|S| \leq k$) is equivalent to finding the connected subnetwork for which the total weight of the vertices is maximum possible. A variant of this problem without any restriction on the size of the subnetworks ($k \leq n$) was defined to extract dysregulated pathways in different cancer types by two independent studies ([42, 155]). Both studies provided integer linear programming formulations but rather than solving the IP formulation, [155] solved a relaxed version of the program, thus, didn't give the optimal solution, and [42] tried to solve the integer linear program using a cutting plane method - however, this approach doesn't guarantee a worst-case running time.

Another variant of the ODkS problem, the Connected k-Subgraph problem (where the weights of vertices are either 0 or 1), is proved to be NP-hard by [83]. Here we prove that ODkS problem is also NP-hard:

Theorem 6.1. *The ODkS problem is NP-hard even when we have one sample for each class.*

Proof. The reduction is done from Connected k-Subgraph problem defined by [83]. We are given an instance of Connected k-Subgraph problem where we have a graph $G = (V, E)$, a weight function $h : V \longrightarrow \{0, 1\}$ and positive integers k and l. For a subnetwork S, let $g(S)$ be the number of vertices with weight 1. The Connected k-Subgraph problem asks whether there exists a subgraph S in G with at most k vertices such that $g(S) \geq l$.

We build an instance of the ODkS problem as follows. The network G' for the instance of ODkS problem is the same as the given graph G i.e. $V' = V$ and $E' = E$. We only have one sample $a = 1$ for the positive class and another sample for the negative class $a' = 1$. For

each gene g_i corresponding to a vertex v_i in G, set $A_i(1) = 0$ and $A'_i(1) = h(v)$. Now for every vertex v_i such that $h(v_i) = 1$, we have the discriminative score $score(v_i) = 1$. By the construction, the discriminative score of any connected subnetwork S' from G' is equivalent to the number of vertices with weight 1 of the corresponding subgraph S in G. Thus, G has a subgraph S with at most k vertices and $g(S) \geq l$ if and only if the network G' has a subnetwork S' also with at most k vertices and $score(S) \geq l$. $\diamond$

6.2 Computational Methods

A Randomized Algorithm In this section, we give a randomized algorithm to solve the ODkS problem for any given error probability. This randomized algorithm is based on color coding technique [6].

Color coding is an algorithmic technique that was first introduced by [6] to detect a simple path or a cycle of length k in a given graph. The algorithm consists of a predefined number of iterations. In each iteration, there are two main steps: assign each vertex uniformly at random with one of k colors and detect whether there is a ''colorful'' path or cycle of length k in the given graph. A path or cycle is colorful if it is not the case that two vertices u, v in the path or cycle have the same color.

The idea behind the algorithm is the clever use of colors to reduce the number of paths that need to consider in the detecting step. In the naive algorithm, we need to keep track of every vertices visited so far which uses $O(n^k)$ time and space. Now we only keep track of all possible sets of vertices of distinct colors which only take $O(n2^k)$ time and space.

Color coding is widely applicable in the context of retrieving ''homologous'' subnetworks from a PPI network given a particular query pathway or protein complex [168, 179, 43, 21]. Color coding has also been successfully applied to retrieve network motifs (subnetworks which are recurrent more than expected in a PPI network) and comparing PPI networks of different species [4, 38].

Similar to color coding technique, our algorithm consists of a predefine number of iterations n_i (we will show how to determine n_i later). Each iteration consists of two main steps:

1. Assign a vertex uniformly at random with one of k colors.

2. Identifying the colorful connected subnetwork S'_{OPT} ($|S'_{OPT}| \leq k$) with the maximum

discriminative score $score(S'_{OPT})$.

We remind the readers that S_{OPT} is the optimally discriminative connected subnetwork while S'_{OPT} is the colorful optimally discriminative subnetwork after each iteration. After n_i iterations, we return S'_{OPT} of some iteration that has the the maximum $score(S'_{OPT})$. We will prove that we return S_{OPT} with the given error probability δ by determining the number iterations n_i and identifying the colorful optimally discriminative subnetwork S'_{OPT} in the second step efficiently.

In the following, we describe how to estimate the number iterations n_i. For each iteration, the probability that we could retrieve S_{OPT} is the same as the probability that S_{OPT} is colorful which is $k!/k^k \geq e^{-k}$. In order to boost the success probability to at least $1-\delta$ for a given error probability δ, we need

$$n_i \leq ln(1/\delta)e^k$$

iterations to yield the S_{OPT}.

In what follows, we describe an efficient dynamic programming algorithm to retrieve the S'_{OPT}. At each iteration, for any vertex $v \in V$ let $color(v)$ denote the color of v. By extending the notation of the discriminating function w defined earlier, we let $score(u, T)$ denote the colorful connected subnetwork S' such that S' contains u, the color set of vertices in S' is T and S' has the maximum discriminative score compared to ones of the colorful connected subnetwork S'''s that contain u. For the base case, for each vertex u, we have:

$$score(u, \{c\}) = \begin{cases} score(u) & \text{if } c = color(v) \\ -\infty & \text{otherwise.} \end{cases}$$

In the general case we can compute $score(u, T)$ as follows:

$$score(u, T) = \max_{\forall v: uv \in E} \big\{ \max_{\forall P,Q: P \cap Q = \emptyset, P \cup Q = T} \{score(u, P) + score(v, Q)\} \big\}$$

Here we assume that the addition of $-\infty$ and any real number or $-\infty$ is $-\infty$. We first compute $score(v, T_1)$ for each vertex v and each set T_1 of one color and so on. In the final step, we compute $score(v, T_k)$ for each vertex v and each set T_k of k colors. Now we compute $score(S'_{OPT})$ as follows:

$$score(S'_{OPT}) = \max_{\forall v: v \in V} \big\{ \max_{\forall T: T \neq \emptyset, |T| \leq k} \{score(v, T)\} \big\}$$

Now we estimate the running time complexity of this randomized algorithm. Let $deg(u)$ be the degree of vertex u. For any vertex u and a set of colors T, in order to compute each $score(u, T)$, it takes $O(deg(u)2^{|T|})$ time. To retrieve S'_{OPT} at each iteration, it takes $O(mk4^k)$ time. Thus, the worst-case running time to retrieve S_{OPT} is $O(mk \ln 1/\delta(4e)^k)$. For our interests in subgraphs of small size $k = O(\log n)$ and for a fixed probability of error, it takes polynomial time to find the optimally discriminative subnetwork S_{OPT}.

Ranking Subnetwork Markers From now on, we fix the error probability $\delta = 0.001$ and the maximum size of a subnetwork $k = 7$ of for any experiment performed later. For each vertex $v \in V$ and for each size n' from 4 to k, we compute the optimal discriminative subnetwork that contain v with n' vertices. In total, we have at most kn subnetworks.

For each subnetwork S, we aggregate the expression profiles of genes in S into a metagene s:

$$\begin{aligned} A_s(j) &= \sum_{g_i \in S} A_i(j)/|S| \ (1 \le j \le a) \\ A'_s(j') &= \sum_{g_i \in S} A'_i(j')/|S| \ (1 \le j' \le a') \end{aligned}$$

Now the normalized discriminative score of a subnetwork S is calculated in the same way as we calculate the discriminative score $score(g)$ for any gene g in Section 2.1. We rank all the extracted subnetworks by their normalized discriminative score. Then we select subnetwork markers from the top to the bottom of the list as follows. Suppose L is the number of genes in the selected subnetworks so far and S is the current considered subnetwork. S is only selected if we have at least $|S|/2$ genes that are not from L. We finish the selection process with 50 subnetworks.

Classification Process and Performance Assessment We always consider top 50 subnetworks for our method for any experiment performed after this point. For any l $(1 \le l \le 50)$, we represent a sample using top l subnetworks $(S_1, ..., S_l)$ as follows. Each sample j is transformed into a l-dimensional vector $V(j) \in \mathbb{R}^l$ where the entries $V(j)_l$ for each marker l are

$$V(j)_l := \sum_{v \in S_l} E(v, j)/|S_l|$$

where v ranges over all genes v contained in the subnetwork marker S_l and $E(v, j)$ is the expression of gene v in sample j. In other words, each sample j becomes a point $V(j)$ in the

l-dimensional feature space $\mathbb{R}^l$. Now, all the classification experiments were performed using 3-nearest neighbour classifier under L_1 distance.

Since the tested datasets have an imbalanced ratio between number of samples in positive and negative class, accuracy is not a good measure for classification performance. We utilize Matthews Coefficient Correlation (MCC) as a measure to compare different classifiers [24]. MCC is essentially the Pearson correlation between the vectors of predicted labels and true labels of a testing set. Suppose that TP is the number of true positives, TN the number of true negatives, FP the number of false positives and FN the number of false negatives. The MCC can be also calculated as follows:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

If one of the four sums in the denominator is zero, the denominator set to one. This results in a Matthews correlation coefficient of zero. MCC value of 1 indicates a perfect prediction, -1 an inverse prediction, and 0 a completely random prediction.

MCC is a recommended measure when compared with other measures for classification performance [10]. We have chosen MCC over area under ROC curve (AUC) to facilitate comparison to competing models from the MAQC-II study [176].

6.3 Experimental Results

Dataset and Network We retrieved the human protein-protein interaction data from the Human Protein Reference Database (HPRD) version April 2010 [108]. By including both binary interactions and considering each protein complex as a clique of proteins, we obtained 46,370 protein interactions involving 9,617 proteins.

We assessed the performance of our method on a human breast cancer dataset contributed by the University of Texas M.D. Anderson Cancer Center (MDACC, Houston, TX, USA). The gene expression profiles were retrieved from NCBI Gene Expression Omnibus (GEO) with accession number GSE20194. Gene expression data from 230 stage I-III breast cancers were generated from fine needle aspiration specimens of newly diagnosed breast cancers before any therapy. Patients received 6 months of neoadjuvant chemotherapy that comprising paclitaxel (T), 5-fluorouracil(F), doxorubicin (A) and cyclophosphamide (C) (and denoted as TFAC) followed by surgical resection of the cancer. Responders to chemotherapy was categorized as a pathological complete response i.e. no residual invasive cancer in the breast

or lymph nodes or residual invasive cancer. RNA extraction and gene expression profiling were performed in multiple batches over time using Affymetrix U133A microarrays. This dataset was split into two different cohorts according to the time of collection. One cohort consists of 130 samples while the other one consists of 100 samples. The expression profiles were normalized with Robust-chip Median Average (RMA) algorithm [90] and adjusted for batch effect using ComBat [98]. Prior to model generation, the expression values of the two cohorts were normalized but not standardized.

Classification Performance We evaluated the performance of our method (we denote as OptDis) against both single gene marker models and other subnetwork-based methods following the workflow presented by the MicroArray Quality Control (MAQC)-II studies [176, 150]. In those studies, the MAQC project assessed the performance and limitations of various data analysis methods in developing and validating microarray-based predictive models with the ultimate goal of discovering best practices. Thirty-six groups participated in the project to develop classifiers for 13 large datasets, including the one used in our study. MAQC models (denoted as MAQC) were constructed by these groups using different methods for data processing (i.e. normalization), feature selection, and classification.

To assess the predictive performance, we performed two analyses. In the forward cross-dataset (FXD) analysis, we treated the 130 patient cohort as the training set used for deriving markers, and validated their performance on the 100 patient cohort. We also performed the complementary backward cross-dataset (BXD) analysis and swapped the cohorts used in training and validation. In Figure 6.2, we compare the performance of OptDis against single gene marker models. The single gene marker classifier constructed using t-test is denoted by SGM and includes only genes that map to the PPI network. For each mappable gene, the corresponding probe with the lowest p value was used in the model. We also compared the performance of our method OptDis against implementations of existing subnetwork-based methods, one based on mutual information (GreedyMI) [31], and another based on dense subnetworks (we denote as Dense) using the STRING functional network [37]. The density threshold to extract all dense subnetworks is set at 0.7 as implemented in [37]. Note that, top 50 subnetworks for GreedyMI and Dense are ranked based on their mutual information scores. Starting from around 20 features, the performance of OptDis is better than competing methods. While the maximum MCC value is not that high, it is still significant compared to the random classifier which has a MCC value of 0. Moreover,

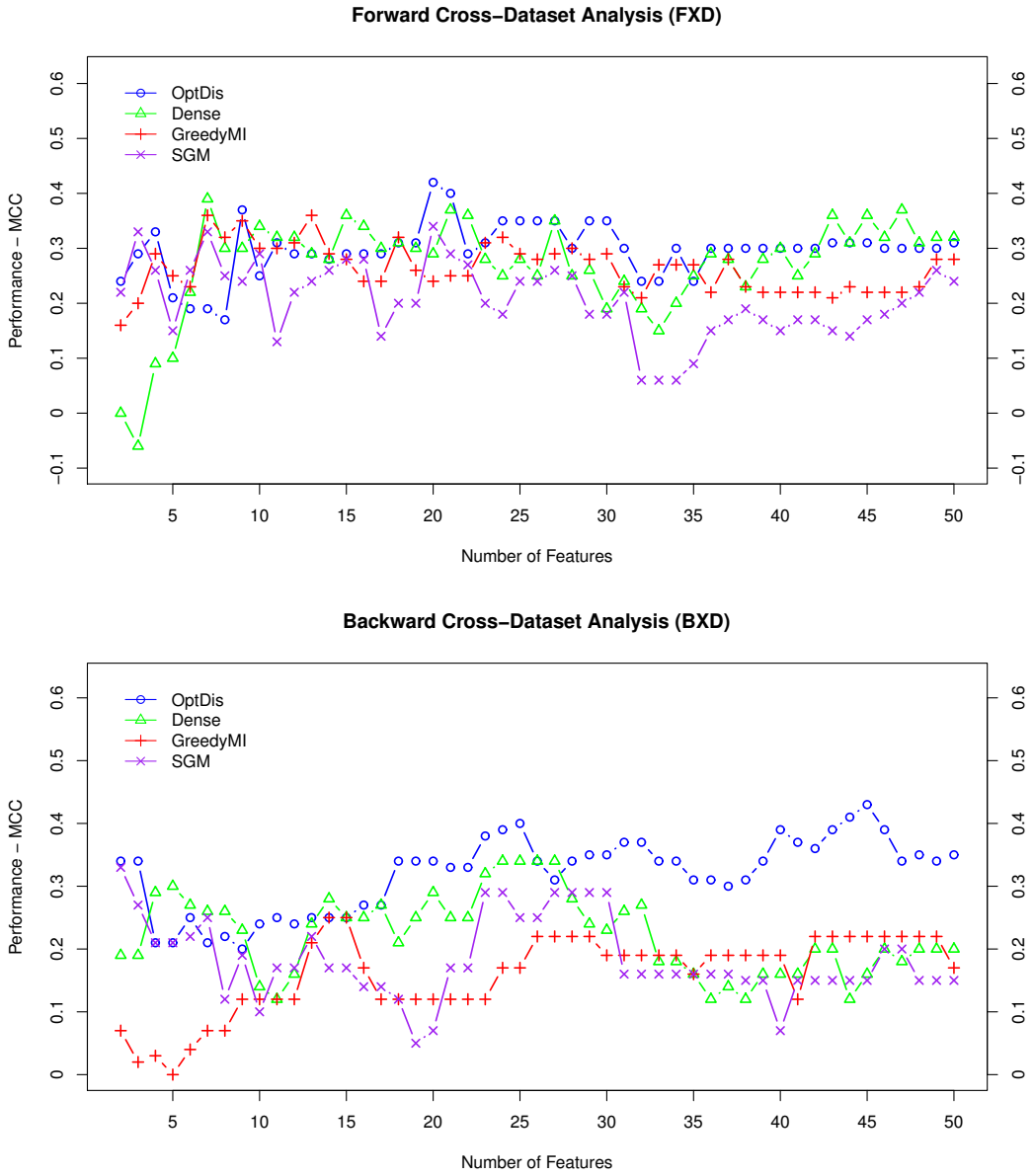


Figure 6.2: Line graphs show the MCCs for different predictive models using the top 1 to 50 features. The compared approaches are single gene marker model based on t-test (SGM) and subnetwork marker models include [31] (GreedyMI), dense subgraphs from STRING functional network by [37] (Dense) and our approach (OptDis).

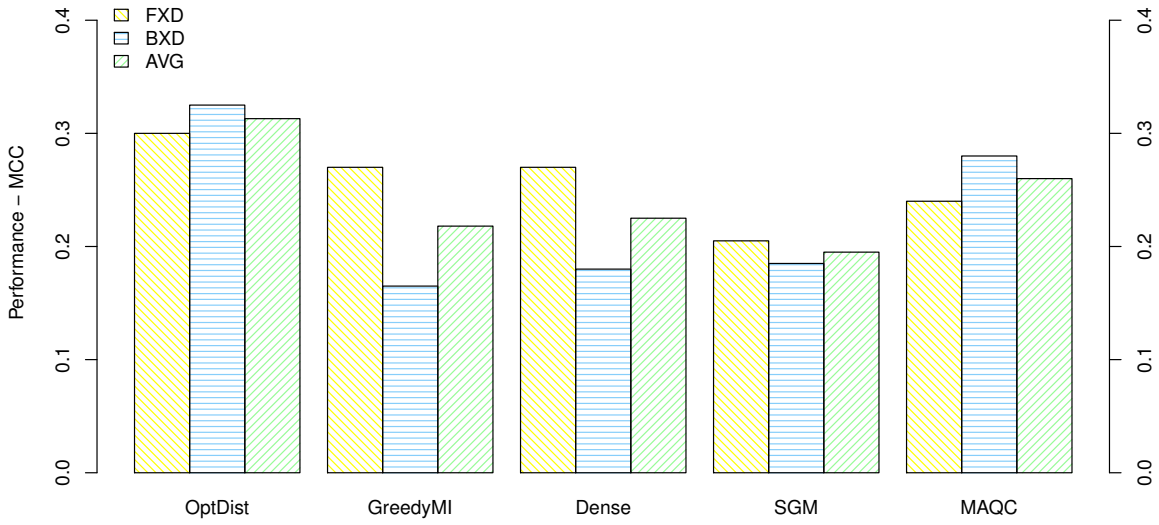


Figure 6.3: Bar charts show the average MCCs of different predictive models. Single gene marker models include based on t-test (SGM) and models from MAQC project (MAQC). Subnetwork marker models include [31] (GreedyMI), [37] (Dense), and our method (OptDis). The white bars and blue bars show the classification performance in FXD and BXD analyses respectively. The green bars show the average of the white and blue bars.

predicting response to chemotherapy has been shown as a difficult endpoint to predict in the recent MAQC publications [176]. The difficulties might be due to the known heterogeneity within tumours of the same cancer type, subtype-specific response, differences in drug metabolism between individuals, and variations in chemotherapy schedules between patients [150]. Figure 6.3 shows the average performance of models in cross-dataset validation of FXD and BXD analyses. Here, the average performance for a model is the average MCC of 50 models generated using the top 1 to 50 features. The MAQC performance was derived from the average of top model from each participating group. As shown in Figure 6.3, OptDis outperforms all other competitors on the average classification performance in FXD and BXD analyses.

For further analyses, we compared the average best performance of different classifiers in

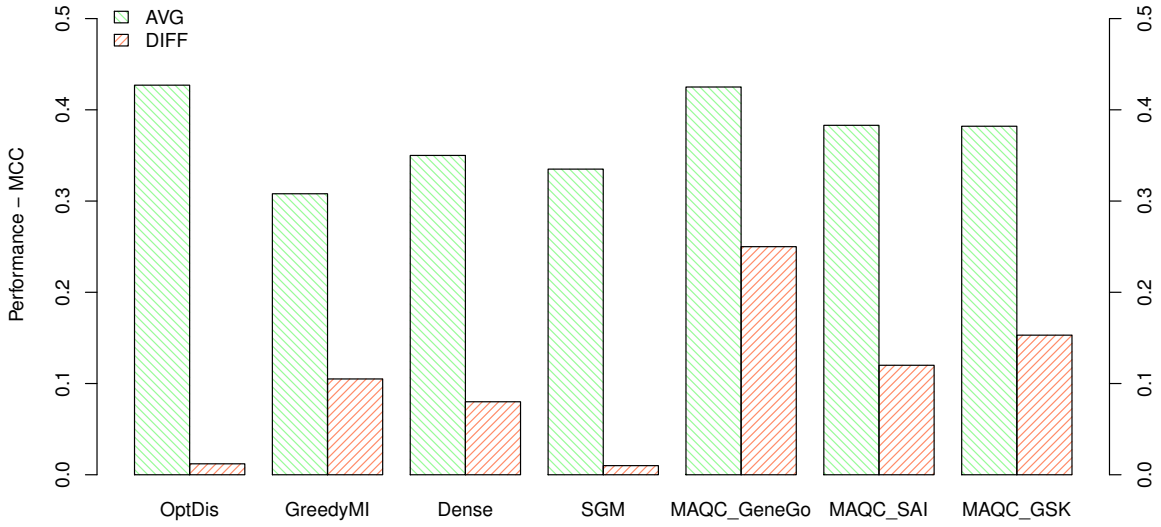


Figure 6.4: The bar charts show the average best MCCs of different classifiers. The average best performance of a classifier is the average of its best model from FXD analysis and its best model from BXD analysis. Single gene marker models include simple one based on t-test (SGM) and top 3 models from MAQC project (MAQC_ GeneGo, MAQC_ SAI, MAQC_ GSK). Subnetwork marker models include [31] (GreedyMI), [37](Dense), and our approach (OptDis). The green bars show the average best MCCs. A red bar of a classifier shows the difference in terms of MCC between its best model from FXD analyses and its best one from BXD analyses.

Figure 6.4. The average best performance of a classifier is the average of its best model from FXD analysis and its best one from BXD analysis. Here we compare against the top three MAQC models. Figure 6.4 shows that our top OptDis model has consistent performance in cross-dataset validation experiments. In contrast, the top three MAQC models show discrepancy in performance when the datasets used for training and test were swapped - especially in the case of the MAQC_ GeneGo model, which has the largest difference in performance (0.25) between the FXD and BXD analysis. The second and third best MAQC models also show similar discrepancy in performance.

Figure 6.5 shows the performance of OptDis against one of the predictive model constructed, where the constituent genes were taken from the top x ($1 \leq x \leq 50$) OptDis subnetworks (we denote as SGM_OptDis). We also compare our method against another single gene marker model that ranks all genes by t-test and matches the number of genes in the top x ($1 \leq x \leq 50$) subnetworks from OptDis (SGM_M). OptDis is consistently better than SGM_OptDis across different number of features. This suggests the importance of treating genes as functional modules. Moreover, on the average constituent genes taken from OptDis subnetworks tend to perform better than genes from simple predictive model using t-test. Hence, OptDis subnetworks might capture genes more informative to predicting chemotherapy response.

In summary, our subnetwork markers have the best combination of relatively high performance and greater stability between different cohorts of patients and thus could be more clinically applicable to other independent cohorts of patients.

Reproducibility of Predictive Markers We compared the reproducibility of subnetwork markers identified by OptDis against gene markers by deriving top markers from the two different cohorts of breast patients and calculating the number of overlapping genes. Since each subnetwork marker may comprise multiple genes, we compared subnetwork markers to an equivalent number of gene markers equal to the number of genes in those subnetworks (i.e. 1 subnetwork marker = k gene markers). The degree of gene overlap across a range of top markers is shown in Figure 6.6. With ten subnetwork markers, OptDis markers already have 25% reproducibility, which is much higher than the 8% reproducibility for an equivalent number of top gene markers. Although the percentage of overlap for gene markers increases as more genes are considered, it remains consistently lower than the reproducibility of subnetwork markers. The greater reproducibility of OptDis markers may contribute to its more robust performance in cross-dataset validation experiments.

Role of Predictive Markers in Drug Response

Gene Function Analysis

We hypothesized that the set of 39 genes (O39) common between the two T50 SN signatures trained on different cohorts may be important to the activity of TFAC therapy. Some of their biological functions are listed in Table 6.1. About half are implicated in apoptosis, suggesting

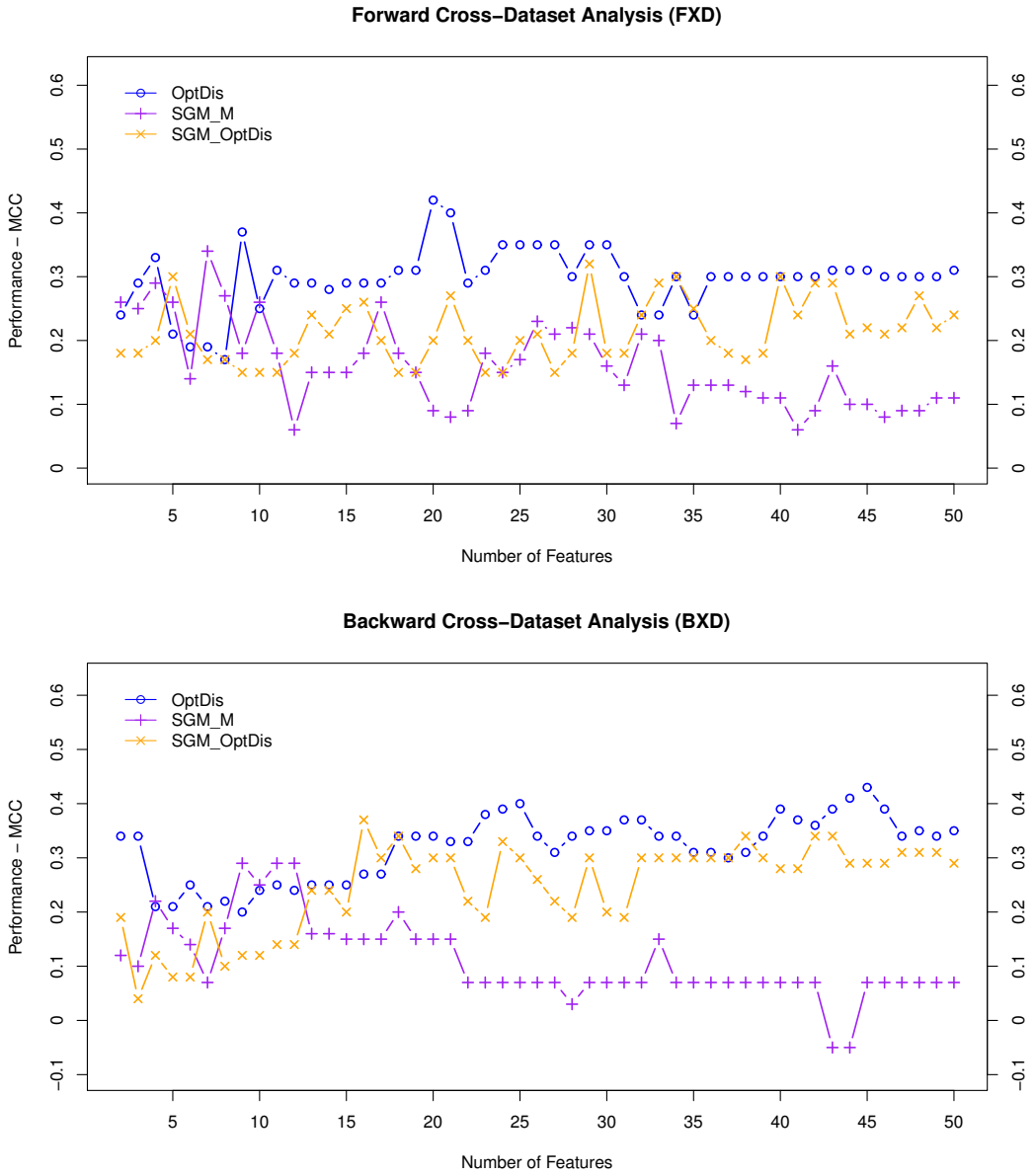


Figure 6.5: Line graphs show the MCCs for different predictive models: our approach (OptDis), the model that constitutes genes from top x ($1 \leq x \leq 50$) subnetworks from OptDis (SGM_OptDis) and another model that ranks genes by t-test and matches the number of genes from top x ($1 \leq x \leq 50$) subnetworks from OptDis (SGM_M).

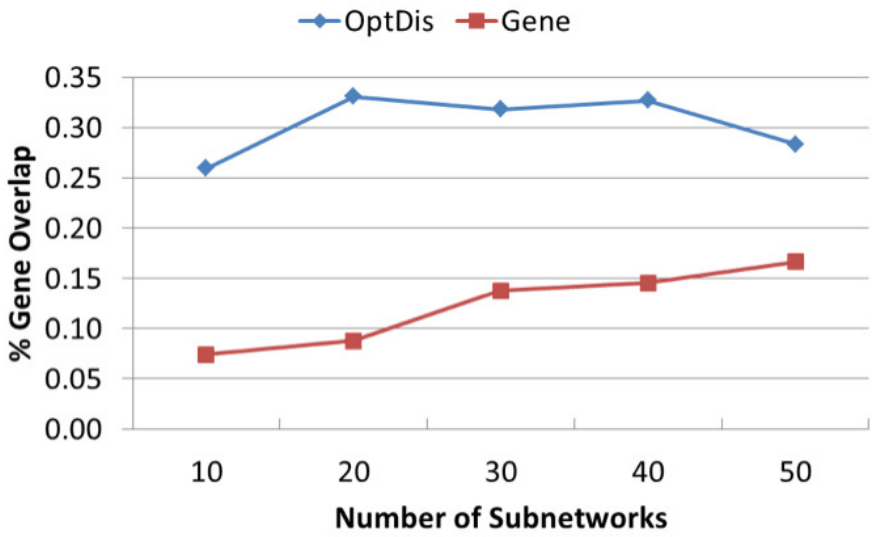


Figure 6.6: Reproducibility is quantified as the degree of gene overlap between top markers identified from different datasets. This overlap is calculated for 10, 20, 30, 40, and 50 OptDis subnetworks and the equivalent number of genes derived from t-test.

that changes in strengths of pro-apoptotic and anti-apoptotic signals can induce resistance to chemotherapy. There are also genes involved in DNA repair, which is expected given many of the anticancer drugs within TFAC therapy induce DNA damage (i.e. cyclophosphamide by cross linking DNA strands).

Some of the 39 genes have specific functions related to mechanism of individual TFAC drugs. Paclitaxel is a mitotic-inhibitor that stabilizes microtubule activity during mitosis and induces cell death. While paclitaxel is known to act on beta-tubulin, some studies [104] have also shown association between the actin and tubulin cytoskeleton in drug response, and suggest that regulation of actin cytoskeleton can induce sensitivity to mitotic-inhibitors. From our O39 list, the EVL, RET, and CST3 genes have regulatory roles in organization and assembly of actin filaments.

Fluouracil's primary anti-cancer activity blocks DNA replication by suppressing thymidylate synthetase activity and depleting thymidine [127]. *In vitro* studies have shown that AR and IGF2, from our O39 list, can increase incorporation of thymidine, which acts in antagonist to thymidylate synthetase suppression, to allow DNA synthesis through the actions

of thymidine kinase [220, 146].

Doxorubicin is an anthracycline antibiotic that intercalates with DNA and causes double-stranded breaks induce to cell apoptosis or disruption in mitosis [135, 138]. SMAD3 from our list has been observed to affect BRCA1-dependent double-stranded DNA break repair in breast cancer cell lines and thus potentially may contribute to differential response to doxorubicin [45].

Signalling Pathway Analysis

Finally, we also compared subnetwork and single gene markers based on their insights into the mechanisms underlying drug response. We derived the T50 SN, T50 SG, and Tx SG from the combined cohort of 230 patients and used the Ingenuity Pathway Analysis software (IPA; Ingenuity © Systems, www.ingenuity.com) to identify significant pathway associations. Interestingly, several signalling pathways associated with chemotherapy response were identified for SN markers, whereas no significantly enriched pathways were found for the T50 and T111 SG markers (Figure 6.7). A closer examination of the top associated pathways suggests response to TFAC treatment is affected by the crosstalk between tumour subtype-specific mechanisms and pathways regulating apoptosis. Chemotherapy response in breast cancer have been observed to be subtype-specific [182], with ER+ tumours exhibiting much higher response rates to taxane-based therapies than ER- tumours [124, 51, 150]. Therefore, it was expected to find that the predictive subnetwork signature was strongly enriched for genes activating the estrogen receptor (ER) signalling pathway. For the same reason, we also observe an enrichment for the androgen receptor (AR) signalling pathway. With nearly all ER+ tumors and few ER- tumours showing AR expression [141], it is likely that AR-based

Enriched terms	Gene symbols	p-value
Apoptosis	AR, EP300, ESR1, GADD45G, IGF2, IGF1R, IGFBP4, IL6ST, MAPK3, MDM2, MED1, NCOA3, PRKACA, RARA, RET, SHC1, SMAD3, SRC, TSC2	1.27E-06
DNA Synthesis	AR, ESR1, IGF2, IGFBP4, IL6ST, MDM2, SHC1, SRC	1.74E-06
Actin Filament Organization	EVL, CST3, RET, SRC, TSC2	7.16E-03
DNA Repair	GADD45G, MDM2, RARA, SMAD3	1.89E-02

Table 6.1: Table of enriched molecular and cellular functions related to drug response of overlapping gene set. The p-values are adjusted using Benjamini-Hochberg method.

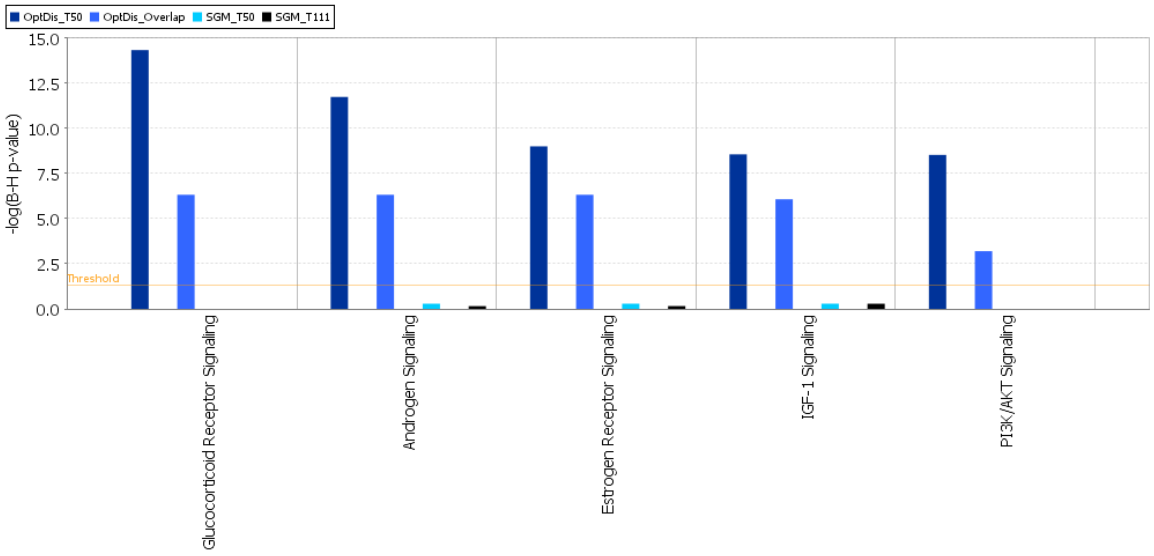


Figure 6.7: Signaling pathways associated with TFAC response ranked by enrichment in T50 SN from our method OptDis. We also show the comparison among the enrichments of genes from T50 SN (dark blue), C39 (light blue), T50 SG (cyan), and T111 SG markers (black). Significantly enriched pathways have Benjamini-Hochberg corrected p-values above threshold of 0.05 (dotted line).

subnetworks serve as good predictive markers of TFAC treatment based on their association with ER status. Based on the enriched IPA pathways associated with response, we speculate that the differential response between subtypes may be attributed to differential regulation of apoptosis. Experimental studies have shown that expression of ERα selectively inhibits paclitaxel-induced apoptosis through modulation of glucocorticoid receptor activity [192].

Other response-associated pathways may also contribute to differential response to TFAC treatment. For example, signalling of insulin-like growth factor (IGF-1) has known functions in cancer proliferation and inhibition of apoptosis, and has been experimentally implicated in chemotherapy resistance [46, 66, 17]. The PI3K/AKT pathway can also increase resistance to taxane-based therapies through downstream anti-apoptotic effectors BCL-2 and BCL-XL [130]. Experiments have shown that tumours with increased phosphorylated BCL-2 expression have increased sensitivity to paclitaxel compared to tumours with reduced expression [177].

We measured the reproducibility of these pathway enrichments by performing IPA pathway analysis on both C39 genes and the T50 SNs derived from the pooled 230 patients

using another SN method (GreedyMI). Figure 6.7 shows both predictive SN signatures were significantly enriched with the same pathways, which may implicate a strong role for these pathways in response to TFAC treatment.

6.4 Source of Performance Improvement

As known in the previous section, OptDis has better performance over single gene markers ranked by t-test . There could be two primary factors that may contribute to the improvement of classification performance of OptDis:

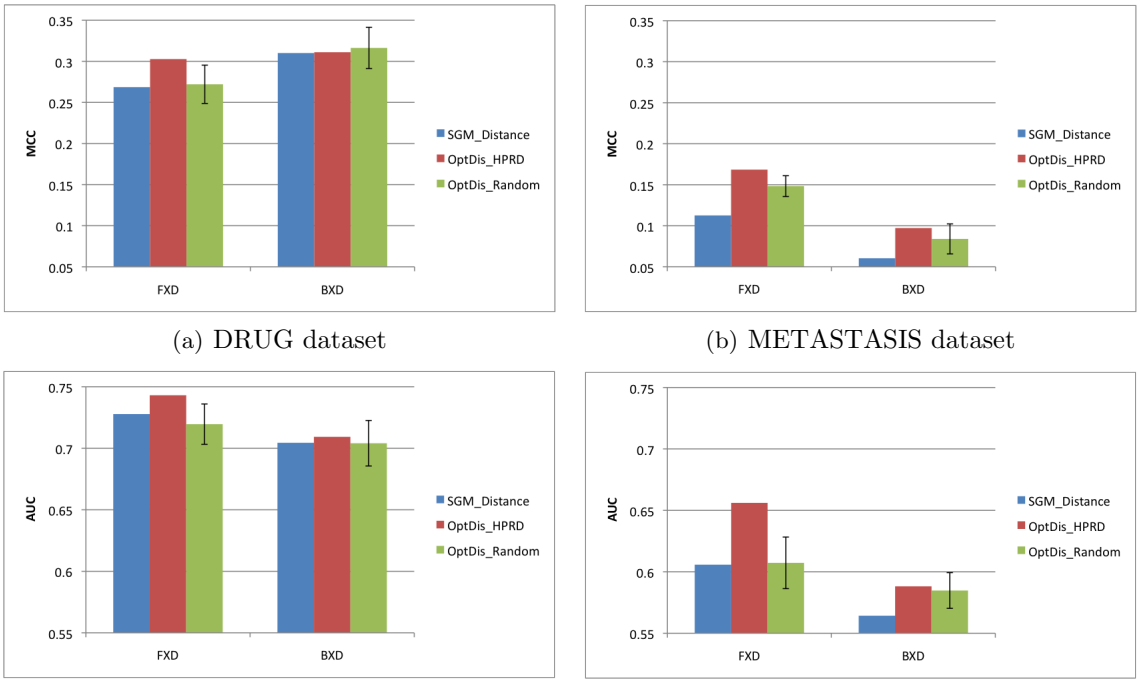
1. Distance-based discrimination score function
2. Use of prior knowledge from PPI networks

Approaches deriving single gene marker rank genes based on how well they distinguish samples from different classes. In the previous section, single genes are ranked by on t-test while subnetwork markers are ranked based on a distance function. Thus the distanced based function could be the source of the improvement in performance of subnetwork markers.

Each subnetwork marker consists of genes which form a connected subnetwork or are in proximity with each other in PPI networks. This is different from approaches deriving single gene markers which do not enforce genes must be near each other on the PPI networks. Thus prior knowledge from PPI network could contribute the improvement in the performance of subnetwork markers.

In this section, we focus on assessing the value of these two aspects. We evaluated the worth of the distance-based discrimination score by comparing the average cross-dataset validation performance of the top gene markers identified by the distance score (SGM_Distance) and the top subnetwork markers from OptDis (OptDis_HPRD). Similar to previous experiments, average performance is calculated as the average of the 50 classifiers built across the range of top markers. We evaluated the value of using prior knowledge, in the form of known protein-protein interactions from PPI networks, to guide marker discovery. To investigate, we re-ran OptDis on randomized networks generated using the Erdős-Rényi model. This model produces a random network with the same average degree as the original PPI network in other words with the same number of edges. Note that this model might not preserve the degree distribution of the original network which is known to be similar among all the PPI networks.

The classification performance of OptDis markers identified using the true PPI network



(a) DRUG dataset

(b) METASTASIS dataset

(c) DRUG dataset

(d) METASTASIS dataset

Figure 6.8: Average classification performances of single gene markers ranked by distance function, OptDis on the natural PPI network HPRD, OptDis on randomized networks on two datasets DRUG and METASTASIS as measured in MCC (top row) and in AUC (bottom row). The average performance for each predictive model is estimated using the top 1 to 50 features. The error bars for the performance of OptDis on randomized networks denote the standard variation on 20 randomized networks. Note that single gene markers ranked based on distance functions matches the number of genes from top x ($1 \leq x \leq 50$) subnetworks from OptDis on HPRD network.

(OptDis_HPRD) against the random networks (OptDis_Random) and the top gene markers identified by the distance score (SGM_Distance) are shown for two datasets. The first dataset is the one that was examined in the previous section and we denote this one as DRUG.

The second dataset is a combination of two datasets on human breast cancer previously reported by van de Vijver *et al.* [207] and Wang *et al.* [217]. Expression profiles from both datasets were obtained from primary breast tumors but hybridized to two different microarray platforms (Agilent oligonucleotide Hu25K microarrays and Affymetrix HGU133a GeneChips). For 78 patients in [207] and 106 in [217], metastasis had been detected during follow-up visits within 5 years of surgery. Profiles for these patients were assigned to the

class metastatic, whereas profiles for the remaining 217 and 180 patients were labeled non-metastatic. We denote this dataset as METASTASIS. Similarly, we denote the classification process by training on [217] and testing on [207] as FXD and the classification process by training on [207] and testing on [217] as BXD.

The classification performance of all the experiments is based on MCC and AUC as shown in Figure 6.8. In order to compute more accurate AUC values for all the experiments, we made use of nearest neighbour classifier and the number of nearest neighbors k is equal to 5.

If interaction knowledge from PPI networks is useful, then the performance of OptDis should decrease when it runs on the random network. However, from the BXD analyses of two datasets, the performance of OptDis on random networks is comparable to the original network. To explain this, we note that the performance of subnetwork markers using the random networks appears to be similar with the performance of gene markers using the distance-based scoring function in DRUG dataset. Thus the distance-based function appears to contribute to the entire the performance improvements demonstrated by OptDis in the DRUG dataset. This suggests that OptDis does not find edge information from the random networks to help in marker discovery in the DRUG dataset.

From FXD analyses of both datasets, we observe a significant drop in performance between OptDis_HPRD and OptDis_Random. Moreover, the average performance of OptDis on the original network is ranked first among all the average performances of OptDis in the randomized networks of both datasets. Thus, results from the FXD analyses from both datasets suggest that knowledge of protein-protein interactions improves marker discovery and classification performance.

Chapter 7

Conclusion

Functional modules which are groups of molecules in the cells together with the interactions among them exhibit a particular function under a biological process. Discovering functional modules is an essential task towards understanding molecular biology of the cells. In this thesis, we present novel computational methods for discovering functional modules in terms of network motifs and subnetwork markers. Our algorithms not only have efficient running time but also have good accuracy on real world biological datasets.

7.1 Summary

In the first part of this dissertation, we have presented background on PPI networks and computational approaches for discovering functional modules from PPI networks. In Chapter 2, we have discussed various wet lab techniques for deriving PPI networks and computational methods for assigning confidence scores for PPIs.

In Chapter 3, we have reviewed existing computational methods for identifying functional modules in static conditions or under a condition or a set of conditions of interest. We have only focused on approaches for detecting functional modules from PPI and functional protein association networks. We also discussed computational approaches for discovering functional modules in which the multi-omics profiles of member genes correlate with the phenotype.

To quantify organismic complexity and evolutionary diversity from a systemic point of view poses challenging biological and computational problems. In Chapter 4, we have investigated *normalized weighted treelet distributions*, based on the exploration of PPI network whose edges are assigned to confidence scores, which can be retrieved from the

STRING database. As a theoretical novelty, we have extended the color coding technique to weighted networks. As a novelty in terms of applications, we applied it to confidence-scored PPI networks. As a result, we were able to reveal differences between uni- and multicellular as well as pro- and eukaryotic organisms. Systemic differences based on local features in PPI networks between pro- and eukaryotes had not been reported before. In sum, our study reveals novel systemic differences and confirms previously reported ones on substantially more reliable PPI network data. Our study also confirms that confidence-scored PPI networks can capture the biological reality more accurately than currently available boolean PPI network data.

Recent studies have strongly confirmed that cancer comes in a great variety of phenotypes as well as multiple evolutionary stages. In Chapter 5, we have explicitly addressed this when searching for systemic subnetwork markers: we employ a biclustering approach (wDCB) which allows that our markers may apply to several but not all cancer samples under examination. As a result, we have outperformed the state-of-the-art approaches, achieving relative increases in prediction accuracy of about 50% in the most demanding cross-platform instances. Our top-ranked markers contained, for example, well-known dysregulated genes involved in TP53 signaling.

In Chapter 6, we have described a novel network-based classification algorithm (OptDis) using the color coding technique to identify optimally discriminative subnetwork markers. Focusing on PPI networks, we have applied our algorithm to drug response studies: we have evaluated our algorithm using published cohorts of breast cancer patients treated with combination chemotherapy. We have shown that our OptDis method improves over previously published subnetwork methods and provides better and more stable performance compared with other subnetwork including wDCB and single gene methods. We have also shown that our subnetwork method produces predictive markers that are more reproducible across independent cohorts and offer valuable insight into biological processes underlying response to therapy. At the end of this chapter, we assessed the main sources of the predicting performance improvement of OptDis.

7.2 Limitations

Most of the computational methods for discovering functional modules including the ones introduced in this thesis have utilized molecular interaction networks which are constructed

mainly from PPI networks. As mentioned in Chapter 2, PPI network data are incomplete and noisy. Some interactions have not been discovered yet. As a result, existing computational methods that search for connected subnetworks from PPI networks could miss many functional modules. Some interactions exist only in some particular conditions. As a consequence, extracted functional modules may not exist under a given condition, thus, are false discoveries. To tackle the incompleteness problem of PPI networks, as we have learnt from Chapter 3, many groups have integrated other genomic evidences to build protein functional association networks. For example, the STRING database has utilized mRNA coexpression and comparative genomics to transfer PPIs from related species to the considered species. To provide a local view of PPI networks, recent efforts have tried to build molecular interaction networks under specific conditions. Thus, future PPI network data could not only be more complete but also provide a local view under the condition of interest.

Many existing algorithms for functional module discovery have integrated gene expression into PPI network data. However, regulation of gene expression is complex and controlled by many factors. In addition, gene expression does not necessarily correspond to protein expression under the condition of interest. Therefore, high expression level of a pair of genes which form an edge in a PPI network may not imply an interaction between their protein products. However, the identification of interactions in PPI networks could be improved with the progress of wet lab techniques to measure protein expression such as mass spectrometry.

7.3 Future Work

Although we have made good progress with algorithms for discovering network motifs and subnetwork markers, there are several interesting directions that need to be explored, in particular the following ones:

- First, it can been seen from previous works in Chapter 3 and our approach in Chapter 4 that currently available approaches for discovering network motifs only take into account of network topology. Since there is a high false positive rate in PPI networks, using network topology may result in false positive discovery. On the other hand, a set of genes which are highly coexpressed and connected from a network might correspond to a functional module. This suggests that integrating PPI network data and gene expression profiles could yield more reliable network motifs. We could redefine

network motifs as connected subgraphs which are more abundant in PPI networks than in random networks with the same global properties and all member genes are coexpressed.

- It is well known that many genes/proteins are known to be expressed only in specific tissues and thus the biological networks in various tissues could be very different. Moreover, the majority of protein interactions in online databases are collected from various tissue types and are examined under standard lab conditions. Thus, current available PPI networks can only offer a static view of molecular interaction networks in cells. Recent efforts have aimed to examine molecular interaction networks under specific tissue types [178] or under specific conditions [11]. As condition-specific interaction data will become abundant in the near future, our algorithms in this thesis in general computational approaches for functional module discovery could be re-examined and could potentially result in better identification. For example, tissue specific data could be utilized to detect tissue specific functional modules.

- Recent high throughput genomic technologies have produced various measurements that capture activities of many molecules DNAs, mRNAs, proteins, metabolites, miRNAs and etc... However, integrating network topology with data from single genomic technology is usually done. Consider all of these multi-omic profiles under a systems biology view could provide more complete and reliable view of molecular interactions of the cells. Thus, novel computational methods for integrating all multi-omic data (for example integrate genomic variations, gene expression, miRNA expression and network topology) will be highly in demand. In addition, future computational tools for integrating multi-omic profiles and network topology should not only discover a functional module as a set of genes but should also recover the regulatory relationships (for example promoting/inhibiting) among its members.

Bibliography

[1] Cellmap. `http://cancer.cellmap.org`.

[2] Nci pathway interaction database. `http://pid.nci.nih.gov`.

[3] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503--511, Feb 2000.

[4] N. Alon, P. Dao, I. Hajirasouliha, F. Hormozdiari, and S. C. Sahinalp. Biomolecular network motif counting and discovery by color coding. *Bioinformatics*, 24:i241--249, Jul 2008.

[5] Noga Alon and Shai Gutner. Balanced families of perfect hash functions and their applications. *ACM Transactions on Algorithms*, 6(3), 2010.

[6] Noga Alon, Raphael Yuster, and Uri Zwick. Color-coding. *J. ACM*, 42(4):844--856, 1995.

[7] M. Altaf-Ul-Amin, Y. Shinbo, K. Mihara, K. Kurokawa, and S. Kanaya. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics*, 7:207, 2006.

[8] Vikraman Arvind and Venkatesh Raman. Approximation algorithms for some parameterized counting problems. In *ISAAC*, pages 453--464, 2002.

[9] G. D. Bader, D. Betel, and C. W. Hogue. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, 31:248--250, Jan 2003.

[10] Pierre et al. Baldi. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412--424, 2000.

[11] S. Bandyopadhyay, M. Mehta, D. Kuo, M. K. Sung, R. Chuang, E. J. Jaehnig, B. Bodenmiller, K. Licon, W. Copeland, M. Shales, D. Fiedler, J. Dutkowski, A. Guenole, H. van Attikum, K. M. Shokat, R. D. Kolodner, W. K. Huh, R. Aebersold, M. C. Keogh, N. J. Krogan, and T. Ideker. Rewiring of genetic networks in response to DNA damage. *Science*, 330(6009):1385--1389, Dec 2010.

[12] E. Banks, E. Nabieva, R. Peterson, and M. Singh. NetGrep: fast network schema searches in interactomes. *Genome Biol.*, 9:R138, 2008.

[13] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509--512, Oct 1999.

[14] A.-L. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101--113, 2004.

[15] S. Bauer, S. Grossmann, M. Vingron, and P.N. Robinson. Ontologizer 2.0 - a multifunctional tool for go term enrichment analysis and data exploration. *Bioinformatics*, 24:1650--1, 2008.

[16] P. Beltrao, J. C. Trinidad, D. Fiedler, A. Roguev, W. A. Lim, K. M. Shokat, A. L. Burlingame, and N. J. Krogan. Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. *PLoS Biol.*, 7:e1000134, Jun 2009.

[17] S. et al. Benini. Inhibition of insulin-like growth factor I receptor increases the antitumor activity of doxorubicin and vincristine against Ewing's sarcoma cells. *Clin Cancer Res.*, 7:1790--1797, Oct 2001.

[18] A. H. Bild, G. Yao, J. T. Chang, Q. Wang, A. Potti, D. Chasse, M. B. Joshi, D. Harpole, J. M. Lancaster, A. Berchuck, J. A. Olson, J. R. Marks, H. K. Dressman, M. West, and J. R. Nevins. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, 439:353--357, Jan 2006.

[19] H. Bonnefoi, C. Underhill, R. Iggo, and D. Cameron. Predictive signatures for chemotherapy sensitivity in breast cancer: are they ready for use in the clinic? *Eur. J. Cancer*, 45:1733--1743, Jul 2009.

[20] P. Braun, A. R. Carvunis, B. Charloteaux, M. Dreze, J. R. Ecker, D. E. Hill, F. P. Roth, M. Vidal, M. Galli, P. Balumuri, V. Bautista, J. D. Chesnut, R. C. Kim, C. de los Reyes, P. Gilles, C. J. Kim, U. Matrubutham, J. Mirchandani, E. Olivares, S. Patnaik, R. Quan, G. Ramaswamy, P. Shinn, G. M. Swamilingiah, S. Wu, J. R. Ecker, M. Dreze, D. Byrdsong, A. Dricot, M. Duarte, F. Gebreab, B. J. Gutierrez, A. MacWilliams, D. Monachello, M. S. Mukhtar, M. M. Poulin, P. Reichert, V. Romero, S. Tam, S. Waaijers, E. M. Weiner, M. Vidal, D. E. Hill, P. Braun, M. Galli, A. R. Carvunis, M. E. Cusick, M. Dreze, V. Romero, F. P. Roth, M. Tasan, J. Yazaki, P. Braun, J. R. Ecker, A. R. Carvunis, Y. Y. Ahn, A. L. Barabasi, B. Charloteaux, H. Chen, M. E. Cusick, J. L. Dangl, M. Dreze, J. R. Ecker, C. Fan, L. Gai, M. Galli, G. Ghoshal,

T. Hao, D. E. Hill, C. Lurin, T. Milenkovic, J. Moore, M. S. Mukhtar, S. J. Pevzner, N. Przulj, S. Rabello, E. A. Rietman, T. Rolland, F. P. Roth, B. Santhanam, R. J. Schmitz, W. Spooner, J. Stein, M. Tasan, J. Vandenhaute, D. Ware, P. Braun, and M. Vidal. Evidence for network evolution in an Arabidopsis interactome map. *Science*, 333:601--607, Jul 2011.

[21] Sharon Bruckner, Falk Hüffner, Richard M. Karp, Ron Shamir, and Roded Sharan. Topology-free querying of protein interaction networks. In *RECOMB*, pages 74--89, 2009.

[22] M. Burger. Mcm2 and mcm5 as prognostic markers in colon cancer: A worthwhile approach. *Digestive Diseases and Sciences*, 54(2):197--198, 2008.

[23] G. Butland, J. M. Peregrin-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan, M. Davey, J. Parkinson, J. Greenblatt, and A. Emili. Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature*, 433:531--537, Feb 2005.

[24] Matthews BW. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim Biophys Acta.*, 405(2):442--51, 1975.

[25] A. Ceol, A. Chatr Aryamontri, L. Licata, D. Peluso, L. Briganti, L. Perfetto, L. Castagnoli, and G. Cesareni. MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.*, 38:D532--539, Jan 2010.

[26] Moses Charikar. Greedy approximation algorithms for finding dense components in a graph. In *APPROX*, pages 84--95, 2000.

[27] C. S. Chen and H. Zhu. Protein microarrays. *BioTechniques*, 40:423, 425, 427 passim, Apr 2006.

[28] L. Chen, J. Xuan, J. Gu, Y. Wang, Z. Zhang, T. L. Wang, and I. e. M. Shih. Integrative network analysis to identify aberrant pathway networks in ovarian cancer. *Pac Symp Biocomput*, pages 31--42, 2012.

[29] Salim A. Chowdhury and Mehmet Koyutürk. Identification of coordinately dysregulated subnetworks in complex phenotypes. In *Pacific Symposium on Biocomputing*, pages 133--144, 2010.

[30] Salim A. Chowdhury, Rod K. Nibbe, Mark R. Chance, and Mehmet Koyutürk. Subnetwork state functions define dysregulated subnetworks in cancer. In *RECOMB*, pages 80--95, 2010.

[31] H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, 3:140, 2007.

[32] G. Ciriello and C. Guerra. A review on models and algorithms for motif discovery in protein-protein interaction networks. *Brief Funct Genomic Proteomic*, 7(2):147--156, Mar 2008.

[33] S. Cleator, A. Tsimelzon, A. Ashworth, M. Dowsett, T. Dexter, T. Powles, S. Hilsenbeck, H. Wong, C. K. Osborne, P. O'Connell, and J. C. Chang. Gene expression patterns for doxorubicin (Adriamycin) and cyclophosphamide (cytoxan) (AC) response and resistance. *Breast Cancer Res. Treat.*, 95:229--233, Feb 2006.

[34] R. Colak, F. Hormozdiari, F. Moser, A. Schonhuth, J. Holman, M. Ester, and S. C. Sahinalp. Dense graphlet statistics of protein interaction and random networks. *Pac Symp Biocomput*, pages 178--189, 2009.

[35] R. Colak, F. Moser, J. S. Chu, A. Schonhuth, N. Chen, and M. Ester. Module discovery by exhaustive search for densely connected, co-expressed regions in biomolecular interaction networks. *PLoS ONE*, 5:e13348, 2010.

[36] D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, S. Jupe, I. Kalatskaya, S. Mahajan, B. May, N. Ndegwa, E. Schmidt, V. Shamovsky, C. Yung, E. Birney, H. Hermjakob, P. D'Eustachio, and L. Stein. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, 39:D691--697, Jan 2011.

[37] Phuong Dao, Recep Colak, Raheleh Salari, Flavia Moser, Elai Davicioni, Alexander Schönhuth, and Martin Ester. Inferring cancer subnetwork markers using density-constrained biclustering. *Bioinformatics*, 26(18), 2010.

[38] Phuong Dao, Alexander Schönhuth, Fereydoun Hormozdiari, Iman Hajirasouliha, Süleyman Cenk Sahinalp, and Martin Ester. Quantifying systemic evolutionary changes by color coding confidence-scored ppi networks. In *WABI*, pages 37--48, 2009.

[39] Phuong Dao, Kendric Wang, Colin Collins, Martin Ester, Anna Lapuk, and Süleyman Cenk Sahinalp. Optimally discriminative subnetwork markers predict response to chemotherapy. *Bioinformatics*, 27(13):205--213, 2011.

[40] C. M. Deane, ?. Salwi?ski, I. Xenarios, and D. Eisenberg. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell Proteomics*, 1:349--356, May 2002.

[41] L. Ding, G. Getz, D. A. Wheeler, E. R. Mardis, M. D. McLellan, K. Cibulskis, C. Sougnez, H. Greulich, D. M. Muzny, M. B. Morgan, L. Fulton, R. S. Fulton, Q. Zhang, M. C. Wendl, M. S. Lawrence, D. E. Larson, K. Chen, D. J. Dooling, A. Sabo, A. C. Hawes, H. Shen, S. N. Jhangiani, L. R. Lewis, O. Hall, Y. Zhu, T. Mathew, Y. Ren, J. Yao, S. E. Scherer, K. Clerc, G. A. Metcalf, B. Ng, A. Milosavljevic, M. L. Gonzalez-Garay, J. R. Osborne, R. Meyer, X. Shi, Y. Tang, D. C. Koboldt, L. Lin,

R. Abbott, T. L. Miner, C. Pohl, G. Fewell, C. Haipek, H. Schmidt, B. H. Dunford-Shore, A. Kraja, S. D. Crosby, C. S. Sawyer, T. Vickery, S. Sander, J. Robinson, W. Winckler, J. Baldwin, L. R. Chirieac, A. Dutt, T. Fennell, M. Hanna, B. E. Johnson, R. C. Onofrio, R. K. Thomas, G. Tonon, B. A. Weir, X. Zhao, L. Ziaugra, M. C. Zody, T. Giordano, M. B. Orringer, J. A. Roth, M. R. Spitz, I. I. Wistuba, B. Ozenberger, P. J. Good, A. C. Chang, D. G. Beer, M. A. Watson, M. Ladanyi, S. Broderick, A. Yoshizawa, W. D. Travis, W. Pao, M. A. Province, G. M. Weinstock, H. E. Varmus, S. B. Gabriel, E. S. Lander, R. A. Gibbs, M. Meyerson, and R. K. Wilson. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, 455:1069--1075, Oct 2008.

[42] Marcus T. Dittrich, Gunnar W. Klau, Andreas Rosenwald, Thomas Dandekar, and Tobias Müller. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. In *ISMB*, pages 223--231, 2008.

[43] Banu Dost, Tomer Shlomi, Nitin Gupta 0002, Eytan Ruppin, Vineet Bafna, and Roded Sharan. Qnet: A tool for querying protein interaction networks. In *RECOMB*, pages 1--15, 2007.

[44] S. Draghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichita, C. Georgescu, and R. Romero. A systems biology approach for pathway level analysis. *Genome Res.*, 17:1537--1545, Oct 2007.

[45] A. et al. Dubrovska. TGFbeta1/Smad3 counteracts BRCA1-dependent repair of DNA damage. *Oncogene*, 24:2289--2297, Mar 2005.

[46] S. E. Dunn, R. A. Hardman, F. W. Kari, and J. C. Barrett. Insulin-like growth factor 1 (IGF-1) alters drug sensitivity of HBL100 human breast cancer cells by inhibition of apoptosis induced by diverse anticancer drugs. *Cancer Res.*, 57:2687--2693, Jul 1997.

[47] Bradley Efron and Robert Tibshirani. On testing the significance of sets of genes. *Annals of Applied Statistics*, pages 107--129, 2007.

[48] L. Ein-Dor, O. Zuk, and E. Domany. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci. U.S.A.*, 103:5923--5928, Apr 2006.

[49] Liat Ein-Dor, Itai Kela, Gad Getz, David Givol, and Eytan Domany. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171--178, 2005.

[50] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, 30:1575--1584, Apr 2002.

[51] P. Farmer, H. Bonnefoi, P. Anderle, D. Cameron, P. Wirapati, P. Wirapati, V. Becette, S. Andre, M. Piccart, M. Campone, E. Brain, G. Macgrogan, T. Petit, J. Jassem,

F. Bibeau, E. Blot, J. Bogaerts, M. Aguet, J. Bergh, R. Iggo, and M. Delorenzi. A stroma-related gene signature predicts resistance to neoadjuvant chemotherapy in breast cancer. *Nat. Med.*, 15(1):68--74, Jan 2009.

[52] E.R. Fearon and B. Vogelstein. A genetic model for colorectal tumorigenesis. *Cell*, 61:759--767, 1990.

[53] S. Fields. High-throughput two-hybrid analysis. The promise and the peril. *FEBS J.*, 272:5391--5399, Nov 2005.

[54] Fedor V. Fomin, Daniel Lokshtanov, Venkatesh Raman, Saket Saurabh, and B. V. Raghavendra Rao. Faster algorithms for finding and counting subgraphs. *J. Comput. Syst. Sci.*, 78(3):698--706, 2012.

[55] K. Fortney, M. Kotlyar, and I. Jurisica. Inferring the functions of longevity genes with modular subnetwork biomarkers of Caenorhabditis elegans aging. *Genome Biol.*, 11:R13, 2010.

[56] Caroline C. Friedel, Jan Krumsiek, and Ralf Zimmer. Bootstrapping the interactome: Unsupervised identification of protein complexes in yeast. In *RECOMB*, pages 3--16, 2008.

[57] A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dumpelfeld, A. Edelmann, M. A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A. M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell, and G. Superti-Furga. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440:631--636, Mar 2006.

[58] A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415:141--147, Jan 2002.

[59] H. Ge, Z. Liu, G. M. Church, and M. Vidal. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat. Genet.*, 29:482--486, Dec 2001.

[60] Gene expression omnibus. http://www.ncbi.nlm.nih.gov/geo/, 2010.

[61] E. Georgii, S. Dietmann, T. Uno, P. Pagel, and K. Tsuda. Enumeration of condition-dependent dense modules in protein interaction networks. *Bioinformatics*, 25:933--940, Apr 2009.

[62] G. Giaever, D. D. Shoemaker, T. W. Jones, H. Liang, E. A. Winzeler, A. Astromoff, and R. W. Davis. Genomic profiling of drug sensitivities via induced haploinsufficiency. *Nat. Genet.*, 21:278--283, Mar 1999.

[63] A. V. Goldberg. Finding a maximum density subgraph. Technical report, Technical report, 1984.

[64] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531--537, Oct 1999.

[65] Mira Gonen, Dana Ron, and Yuval Shavitt. Counting stars and other small subgraphs in sublinear time. In *SODA*, pages 99--116, 2010.

[66] J. L. Gooch, C. L. Van Den Berg, and D. Yee. Insulin-like growth factor (IGF)-I rescues breast cancer cells from chemotherapy-induced cell death--proliferative and anti-apoptotic effects. *Breast Cancer Res. Treat.*, 56:1--10, Jul 1999.

[67] Joshua A. Grochow and Manolis Kellis. Network motif discovery using subgraph enumeration and symmetry-breaking. In *RECOMB*, pages 92--106, 2007.

[68] Z. Guo, T. Zhang, X. Li, Q. Wang, J. Xu, H. Yu, J. Zhu, H. Wang, C. Wang, E. J. Topol, Q. Wang, and S. Rao. Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinformatics*, 6:58, 2005.

[69] Zheng Guo, Yongjin Li, Xue Gong, Chen Yao, Wencai Ma, Dong Wang, Yanhui Li, Jing Zhu, Min Zhang, Da Yang, and Jing Wang. Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. *Bioinformatics*, 23(16):2121--2128, 2007.

[70] S. P. Gygi, B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb, and R. Aebersold. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.*, 17:994--999, Oct 1999.

[71] A. Hahn, J. Rahnenfuhrer, P. Talwar, and T. Lengauer. Confirmation of human protein interaction data by human expression data. *BMC Bioinformatics*, 6:112, 2005.

[72] W. C. Hahn and R. A. Weinberg. Modelling the molecular circuitry of cancer. *Nat. Rev. Cancer*, 2:331--341, May 2002.

[73] J. D. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430:88--93, Jul 2004.

[74] J. D. Han, D. Dupuy, N. Bertin, M. E. Cusick, and M. Vidal. Effect of sampling on topology predictions of protein-protein interaction networks. *Nat. Biotechnol.*, 23(7):839--844, Jul 2005.

[75] D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer. Co-clustering of biological networks and gene expression data. *Bioinformatics*, 18 Suppl 1:S145--154, 2002.

[76] G. T. Hart, I. Lee, and E. R. Marcotte. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics*, 8:236, 2007.

[77] Erez Hartuv and Ron Shamir. A clustering algorithm based on graph connectivity. *Inf. Process. Lett.*, 76(4-6):175--181, 2000.

[78] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402:47--52, Dec 1999.

[79] F. Hayot and C. Jayaprakash. A feedforward loop motif in transcriptional regulation: induction and repression. *J. Theor. Biol.*, 234:133--143, May 2005.

[80] H. Hegyi and M. Gerstein. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.*, 288:147--164, Apr 1999.

[81] K. R. et al. Hess. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J. Clin. Oncol.*, 24:4236--4244, Sep 2006.

[82] Desmond J. Higham, Marija Rasajski, and Natasa Przulj. Fitting a geometric graph to a protein-protein interaction network. *Bioinformatics*, 24(8):1093--1099, 2008.

[83] D. S. Hochbaum and A. Pathria. Node-optimal connected k-subgraphs. *Manuscript, UC Berkeley*, 1994.

[84] F. Hormozdiari, P. Berenbrink, N. Przulj, and S. C. Sahinalp. Not all scale-free networks are born equal: the role of the seed graph in PPI network evolution. *PLoS Comput. Biol.*, 3:e118, Jul 2007.

[85] H. Hu, X. Yan, Y. Huang, J. Han, and X. J. Zhou. Mining coherent dense subgraphs across massive biological networks for functional discovery. *Bioinformatics*, 21 Suppl 1:i213--221, Jun 2005.

[86] J. A. Hurt, S. A. Thibodeau, A. S. Hirsh, C. O. Pabo, and J. K. Joung. Highly specific zinc finger proteins obtained by directed domain shuffling and cell-based selection. *Proc. Natl. Acad. Sci. U.S.A.*, 100:12271--12276, Oct 2003.

[87] TaeHyun Hwang, Ze Tian, Rui Kuang, and Jean-Pierre Kocher. Learning on weighted hypergraphs to integrate protein interactions and gene expressions for cancer outcome prediction. In *ICDM*, pages 293--302, 2008.

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

[89] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18 Suppl 1:S233--240, 2002.

[90] Rafael. A. Irizarry, Benjamin M. Bolstad, Francois Collin, and Leslie M. Cope. Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, 31, 2003.

[91] V. R. Iyer, C. E. Horak, C. S. Scafe, D. Botstein, M. Snyder, and P. O. Brown. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, 409:533--538, Jan 2001.

[92] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302:449--453, Oct 2003.

[93] L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. von Mering. STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, 37(Database issue):D412--416, Jan 2009.

[94] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407:651--654, Oct 2000.

[95] Daxin Jiang and Jian Pei. Mining frequent cross-graph quasi-cliques. *TKDD*, 2(4), 2009.

[96] P. Jiang and M. Singh. SPICi: a fast clustering algorithm for large biological networks. *Bioinformatics*, 26:1105--1111, Apr 2010.

[97] X. Jiang, J. Tan, J. Li, S. Kivime, X. Yang, L. Zhuang, P.L. Lee, M.T. Chan, L.W. Stanton, E.T. Liu, B.N. Cheyette, and Q. Yu. Dact3 is an epigenetic regulator of wnt/beta-catenin signaling in colorectal cancer and is a therapeutic target of histone modifications. *Cancer Cell*, 13:529--41, 2008.

[98] W. E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8:118--127, Jan 2007.

[99] J. K. Joung, E. I. Ramm, and C. O. Pabo. A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions. *Proc. Natl. Acad. Sci. U.S.A.*, 97:7382--7387, Jun 2000.

[100] M. Kalaev, M. Smoot, T. Ideker, and R. Sharan. NetworkBLAST: comparative analysis of protein networks. *Bioinformatics*, 24:594--596, Feb 2008.

[101] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, 32:D277--280, Jan 2004.

[102] Richard M. Karp and Michael Luby. Monte-carlo algorithms for enumeration and reliability problems. In *FOCS*, pages 56--64, 1983.

[103] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20:1746--1758, Jul 2004.

[104] M. Kavallaris. Microtubules and resistance to tubulin-binding agents. *Nat. Rev. Cancer*, 10:194--204, Mar 2010.

[105] R. Kelley and T. Ideker. Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.*, 23(5):561--566, May 2005.

[106] T. K. Kerppola. Design and implementation of bimolecular fluorescence complementation (BiFC) assays for the visualization of protein interactions in living cells. *Nat Protoc*, 1:1278--1286, 2006.

[107] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, C. Jandrasits, R. C. Jimenez, J. Khadake, U. Mahadevan, P. Masson, I. Pedruzzi, E. Pfeiffenberger, P. Porras, A. Raghunath, B. Roechert, S. Orchard, and H. Hermjakob. The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, 40:D841--846, Jan 2012.

[108] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey. Human Protein Reference Database--2009 update. *Nucleic Acids Res.*, 37:D767--772, Jan 2009.

[109] D.H. Ki, H.C. Jeung, C.H. Park, S.H. Kang, G.Y. Lee, W.S. Lee, N.K. Kim, H.C. Chung, and S.Y. Rha. Whole genome analysis for liver metastasis gene signatures in colorectal cancer. *International Journal of Cancer*, 121, 2007.

[110] Y. A. Kim, S. Wuchty, and T. M. Przytycka. Identifying causal genes and dysregulated pathways in complex diseases. *PLoS Comput. Biol.*, 7(3):e1001095, Mar 2011.

[111] A. D. King, N. Przulj, and I. Jurisica. Protein complex prediction via cost-based clustering. *Bioinformatics*, 20:3013--3020, Nov 2004.

[112] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrin-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. St Onge, S. Ghanny, M. H. Lam, G. Butland, A. M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili, and J. F. Greenblatt. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, 440:637--643, Mar 2006.

[113] Vincent Lacroix, Cristina G. Fernandes, and Marie-France Sagot. Reaction motifs in metabolic networks. In *WABI*, pages 178--191, 2005.

[114] E. Lee, H. Y. Chuang, J. W. Kim, T. Ideker, and D. Lee. Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.*, 4:e1000217, Nov 2008.

[115] H. K. Lee, W. Braynen, K. Keshav, and P. Pavlidis. ErmineJ: tool for functional analysis of gene expression data sets. *BMC Bioinformatics*, 6:269, 2005.

[116] I. Lee, B. Lehner, C. Crombie, W. Wong, A. G. Fraser, and E. M. Marcotte. A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. *Nat. Genet.*, 40:181--188, Feb 2008.

[117] J. K. Lee, D. M. Havaleshko, H. Cho, J. N. Weinstein, E. P. Kaldjian, J. Karpovich, A. Grimshaw, and D. Theodorescu. A strategy for predicting the chemosensitivity of human cancers and its application to drug discovery. *Proc. Natl. Acad. Sci. U.S.A.*, 104:13086--13091, Aug 2007.

[118] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298:799--804, Oct 2002.

[119] B. Levin. *"Genes IX"*. "Jones and Bartlett", 2007.

[120] Georges J. J. Lhoest. Protein-Protein, Protein-Drug Interactions and MS. `http://www.specmetcrime.com/noncovalent_complexes_in_mass_s.htm`.

[121] W. Li, C. C. Liu, T. Zhang, H. Li, M. S. Waterman, and X. J. Zhou. Integrative analysis of many weighted co-expression networks using tensor computation. *PLoS Comput. Biol.*, 7:e1001106, Jun 2011.

[122] P. Liang and A. B. Pardee. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, 257:967--971, Aug 1992.

[123] C. Liedtke, J. Wang, A. Tordai, W. F. Symmans, G. N. Hortobagyi, L. Kiesel, K. Hess, K. A. Baggerly, K. R. Coombes, and L. Pusztai. Clinical evaluation of chemotherapy response predictors developed from breast cancer cell lines. *Breast Cancer Res. Treat.*, 121:301--309, Jun 2010.

[124] C. et al. Liedtke. Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. *J. Clin. Oncol.*, 26:1275--1281, Mar 2008.

[125] R. Linding, L. J. Jensen, G. J. Ostheimer, M. A. van Vugt, C. J?rgensen, I. M. Miron, F. Diella, K. Colwill, L. Taylor, K. Elder, P. Metalnikov, V. Nguyen, A. Pasculescu, J. Jin, J. G. Park, L. D. Samson, J. R. Woodgett, R. B. Russell, P. Bork, M. B. Yaffe, and T. Pawson. Systematic discovery of in vivo phosphorylation networks. *Cell*, 129:1415--1426, Jun 2007.

[126] Y. Loewenstein, E. Portugaly, M. Fromer, and M. Linial. Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space. *Bioinformatics*, 24:i41--49, Jul 2008.

[127] D. B. Longley, D. P. Harkin, and P. G. Johnston. 5-fluorouracil: mechanisms of action and clinical strategies. *Nat. Rev. Cancer*, 3:330--338, May 2003.

[128] S. Mangan and U. Alon. Structure and function of the feed-forward loop network motif. *Proc. Natl. Acad. Sci. U.S.A.*, 100:11980--11985, Oct 2003.

[129] S. Mangan, A. Zaslaver, and U. Alon. The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *J. Mol. Biol.*, 334:197--204, Nov 2003.

[130] B. T. McGrogan, B. Gilmartin, D. N. Carney, and A. McCann. Taxanes, microtubules and chemoresistant breast cancer. *Biochim. Biophys. Acta*, 1785:96--132, Apr 2008.

[131] H. Mi, N. Guo, A. Kejariwal, and P. D. Thomas. PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic Acids Res.*, 35:D247--252, Jan 2007.

[132] Roger L. Miesfeld. Amg lecture 23.

[133] L.D. Miller, J. Smeds, J. George, V.B. Vega, L. Vergara, A. Ploner, Y. Pawitan, P. Hall, S. Klaar, E.T. Liu, and J. Bergh. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc. Natl. Acad. Sci. USA*, 102(38):13550--13555, 2005.

[134] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298:824--827, Oct 2002.

[135] G. Minotti, P. Menna, E. Salvatorelli, G. Cairo, and L. Gianni. Anthracyclines: molecular advances and pharmacologic developments in antitumor activity and cardiotoxicity. *Pharmacol. Rev.*, 56:185--229, Jun 2004.

[136] F. Moser, R. Colak, A. Rafiey, and M. Ester. Mining cohesive patterns from graphs with feature vectors. In *SDM*, pages 593--604, 2009.

[137] S. Mostafavi, D. Ray, D. Warde-Farley, C. Grouios, and Q. Morris. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol.*, 9 Suppl 1:S4, 2008.

[138] A. F. Munro, D. A. Cameron, and J. M. Bartlett. Targeting anthracyclines in early breast cancer: new candidate predictive biomarkers emerge. *Oncogene*, 29:5231--5240, Sep 2010.

[139] S. Navlakha, M. C. Schatz, and C. Kingsford. Revealing biological modules via graph summarization. *J. Comput. Biol.*, 16:253--264, Feb 2009.

[140] Saket Navlakha, Rajeev Rastogi, and Nisheeth Shrivastava. Graph summarization with bounded error. In *SIGMOD Conference*, pages 419--432, 2008.

[141] L. A. Niemeier, D. J. Dabbs, S. Beriwal, J. M. Striebel, and R. Bhargava. Androgen receptor in breast cancer: expression in estrogen receptor-positive tumors and in estrogen receptor-negative tumors with apocrine differentiation. *Mod. Pathol.*, 23:205--212, Feb 2010.

[142] University of Victoria. Combinatorial object server. *http://www.theory.csc.uvic.ca/ cos*, since 1995.

[143] S. E. Ong, B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey, and M. Mann. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell Proteomics*, 1:376--386, May 2002.

[144] P. Pagel, S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, P. Mark, V. Stumpflen, H. W. Mewes, A. Ruepp, and D. Frishman. The MIPS mammalian protein-protein interaction database. *Bioinformatics*, 21:832--834, Mar 2005.

[145] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814--818, Jun 2005.

[146] A. Pedram, M. Razandi, R. C. Sainson, J. K. Kim, C. C. Hughes, and E. R. Levin. A conserved mechanism for steroid receptor translocation to the plasma membrane. *J. Biol. Chem.*, 282:22278--22288, Aug 2007.

[147] Jian Pei, Daxin Jiang, and Aidong Zhang. Mining cross-graph quasi-cliques in gene expression and protein interaction data. In *ICDE*, pages 353--354, 2005.

[148] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U.S.A.*, 96:4285--4288, Apr 1999.

[149] J. B. Pereira-Leal, A. J. Enright, and C. A. Ouzounis. Detection of functional modules from protein interaction networks. *Proteins*, 54:49--57, Jan 2004.

[150] V. et al. Popovici. Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. *Breast Cancer Res.*, 12:R5, 2010.

[151] N. Przulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23:e177--183, Jan 2007.

[152] Natasa Przulj, Oleksii Kuchaiev, Aleksandar Stevanovic, and Wayne Hayes. Geometric evolutionary dynamics of protein interaction networks. In *Pacific Symposium on Biocomputing*, pages 178--189, 2010.

[153] J. Ptacek, G. Devgan, G. Michaud, H. Zhu, X. Zhu, J. Fasolo, H. Guo, G. Jona, A. Breitkreutz, R. Sopko, R. R. McCartney, M. C. Schmidt, N. Rachidi, S. J. Lee, A. S. Mah, L. Meng, M. J. Stark, D. F. Stern, C. De Virgilio, M. Tyers, B. Andrews, M. Gerstein, B. Schweitzer, P. F. Predki, and M. Snyder. Global analysis of protein phosphorylation in yeast. *Nature*, 438:679--684, Dec 2005.

[154] S. Pu, J. Vlasblom, A. Emili, J. Greenblatt, and S. J. Wodak. Identifying functional modules in the physical interactome of Saccharomyces cerevisiae. *Proteomics*, 7:944--960, Mar 2007.

[155] Y. Qiu, S. Zhang, X-S. Zhang, and L. Chen. Identifying differentially expressed pathways via a mixed integer linear programming model. 3:475--486, 2009.

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

[157] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551--1555, Aug 2002.

[158] T. Reguly, A. Breitkreutz, L. Boucher, B. J. Breitkreutz, G. C. Hon, C. L. Myers, A. Parsons, H. Friesen, R. Oughtred, A. Tong, C. Stark, Y. Ho, D. Botstein, B. Andrews, C. Boone, O. G. Troyanskya, T. Ideker, K. Dolinski, N. N. Batada, and M. Tyers. Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *J. Biol.*, 5(4):11, 2006.

[159] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. Genome-wide location and function of DNA binding proteins. *Science*, 290:2306--2309, Dec 2000.

[160] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Seraphin. A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.*, 17:1030--1032, Oct 1999.

[161] M. Ronen, R. Rosenberg, B. I. Shraiman, and U. Alon. Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl. Acad. Sci. U.S.A.*, 99:10555--10560, Aug 2002.

[162] A. Rosenwald et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large b-cell lymphoma. *The New England Journal of Medicine*, 346(25):1937--1947, 2002.

[163] J. Sabates-Bellver et al. Transcriptome profile of human colorectal adenomas. *Molecular Cancer Research*, 5:1263--1275, 2007.

[164] Barna Saha, Allison Hoch, Samir Khuller, Louiqa Raschid, and Xiao-Ning Zhang. Dense subgraphs with restrictions and applications to gene annotation graphs. In *RECOMB*, pages 456--472, 2010.

[165] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, 32:D449--451, Jan 2004.

[166] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467--470, Oct 1995.

[167] B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[168] J. Scott, T. Ideker, R. M. Karp, and R. Sharan. Efficient algorithms for detecting signaling pathways in protein interaction networks. *J. Comput. Biol.*, 13:133--144, Mar 2006.

[169] E. Segal, H. Wang, and D. Koller. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, 19 Suppl 1:i264--271, 2003.

[170] R. Sharan, S. Suthram, R. M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. M. Karp, and T. Ideker. Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. U.S.A.*, 102:1974--1979, Feb 2005.

[171] R. Sharan, I. Ulitsky, and R. Shamir. Network-based prediction of protein function. *Mol. Syst. Biol.*, 3:88, 2007.

[172] Roded Sharan and Richard Karp. Reconstructing boolean models of signaling. In *RECOMB*, 2012.

[173] Roded Sharan and Ron Shamir. Center click: A clustering algorithm with applications to gene expression analysis. In *ISMB*, pages 307--316, 2000.

[174] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat. Genet.*, 31:64--68, May 2002.

[175] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888--905, 2000.

[176] L. et al. Shi. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.*, 28:827--838, Aug 2010.

[177] M. Shitashige, M. Toi, T. Yano, M. Shibata, Y. Matsuo, and F. Shibasaki. Dissociation of Bax from a Bcl-2/Bax heterodimer triggered by phosphorylation of serine 70 of Bcl-2. *J. Biochem.*, 130:741--748, Dec 2001.

[178] T. Shlomi, M. N. Cabili, M. J. Herrgard, B. ?. Palsson, and E. Ruppin. Network-based prediction of human tissue-specific metabolism. *Nat. Biotechnol.*, 26(9):1003--1010, Sep 2008.

[179] Tomer Shlomi, Daniel Segal, Eytan Ruppin, and Roded Sharan. Qpath: a method for querying pathways in a protein-protein interaction network. *BMC Bioinformatics*, 7:199, 2006.

[180] T. Sjoblom, S. Jones, L. D. Wood, D. W. Parsons, J. Lin, T. D. Barber, D. Mandelker, R. J. Leary, J. Ptak, N. Silliman, S. Szabo, P. Buckhaults, C. Farrell, P. Meeh, S. D. Markowitz, J. Willis, D. Dawson, J. K. Willson, A. F. Gazdar, J. Hartigan, L. Wu, C. Liu, G. Parmigiani, B. H. Park, K. E. Bachman, N. Papadopoulos, B. Vogelstein, K. W. Kinzler, and V. E. Velculescu. The consensus coding sequences of human breast and colorectal cancers. *Science*, 314:268--274, Oct 2006.

[181] B. Snel, G. Lehmann, P. Bork, and M. A. Huynen. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.*, 28:3442--3444, Sep 2000.

[182] T. Sorlie, Y. Wang, C. Xiao, H. Johnsen, B. Naume, R. R. Samaha, and A. L. B?rresen-Dale. Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across three different platforms. *BMC Genomics*, 7:127, 2006.

[183] M. E. Sowa, E. J. Bennett, S. P. Gygi, and J. W. Harper. Defining the human deubiquitinating enzyme interaction landscape. *Cell*, 138:389--403, Jul 2009.

[184] V. Spirin and L. A. Mirny. Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. U.S.A.*, 100:12123--12128, Oct 2003.

[185] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, 34:D535--539, Jan 2006.

[186] M. Steffen, A. Petti, J. Aach, P. D'haeseleer, and G. Church. Automated modelling of signal transduction networks. *BMC Bioinformatics*, 3:34, Nov 2002.

[187] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksoz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E. E. Wanker. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122:957--968, Sep 2005.

[188] S. H. Strogatz. Exploring complex networks. *Nature*, 410:268--276, Mar 2001.

[189] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302:249--255, Oct 2003.

[190] J. Su, B. J. Yoon, and E. R. Dougherty. Identification of diagnostic subnetwork markers for cancer in human protein-protein interaction network. *BMC Bioinformatics*, 11 Suppl 6:S8, 2010.

[191] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, 102:15545--15550, Oct 2005.

[192] M. Sui, Y. Huang, B. H. Park, N. E. Davidson, and W. Fan. Estrogen receptor alpha mediates breast cancer cell resistance to paclitaxel through inhibition of apoptotic cell death. *Cancer Res.*, 67:5337--5344, Jun 2007.

[193] S. Suthram, A. Beyer, R. M. Karp, Y. Eldar, and T. Ideker. eQED: an efficient method for interpreting eQTL associations using protein networks. *Mol. Syst. Biol.*, 4:162, 2008.

[194] S. Suthram, T. Shlomi, E. Ruppin, R. Sharan, and T. Ideker. A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics*, 7:360, 2006.

[195] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. von Mering. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, 39:D561--568, Jan 2011.

[196] C. S. Tan, B. Bodenmiller, A. Pasculescu, M. Jovanovic, M. O. Hengartner, C. J?rgensen, G. D. Bader, R. Aebersold, T. Pawson, and R. Linding. Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. *Sci Signal*, 2:ra39, 2009.

[197] K. Tarassov, V. Messier, C. R. Landry, S. Radinovic, M. M. Serna Molina, I. Shames, Y. Malitskaya, J. Vogel, H. Bussey, and S. W. Michnick. An in vivo map of the yeast protein interactome. *Science*, 320:1465--1470, Jun 2008.

[198] Adi L. Tarca, Sorin Draghici, Purvesh Khatri, Sonia S. Hassan, Pooja Mittal, Jung sun Kim, Chong Jai Kim, Juan Pedro Kusanovic, and Roberto Romero. A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75--82, 2009.

[199] I. W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, and J. L. Wrana. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.*, 27:199--204, Feb 2009.

[200] A. E. Teschendorff, S. Gomez, A. Arenas, D. El-Ashry, M. Schmidt, M. Gehrmann, and C. Caldas. Improved prognostic classification of breast cancer defined by antagonistic activation patterns of immune response pathway modules. *BMC Cancer*, 10:604, 2010.

[201] L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, and P. J. Park. Discovering statistically significant pathways in expression profiling studies. *Proc. Natl. Acad. Sci. U.S.A.*, 102:13544--13549, Sep 2005.

[202] Charalampos E. Tsourakakis, U. Kang, Gary L. Miller, and Christos Faloutsos. Doulion: counting triangles in massive graphs with a coin. In *KDD*, pages 837--846, 2009.

[203] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403:623--627, Feb 2000.

[204] I. Ulitsky and R. Shamir. Identification of functional modules using network topology and high-throughput data. *BMC Syst Biol*, 1:8, 2007.

[205] Igor Ulitsky, Richard M. Karp, and Ron Shamir. Detecting disease-specific dysregulated pathways via analysis of clinical expression profiles. In *RECOMB*, pages 347--359, 2008.

[206] Igor Ulitsky and Ron Shamir. Identifying functional modules using expression profiles and confidence-scored protein interactions. *Bioinformatics*, 25(9):1158--1164, 2009.

[207] M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards. A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, 347(25):1999--2009, Dec 2002.

[208] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530--536, Jan 2002.

[209] F. Vandin, P. Clay, E. Upfal, and B. J. Raphael. Discovery of mutated subnetworks associated with clinical data in cancer. *Pac Symp Biocomput*, pages 55--66, 2012.

[210] Fabio Vandin, Eli Upfal, and Benjamin J. Raphael. *De Novo* discovery of mutated driver pathways in cancer. In *RECOMB*, pages 499--500, 2011.

[211] L. J. van't Veer and R. Bernards. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*, 452:564--570, Apr 2008.

[212] Charles J. Vaske, Stephen C. Benz, J. Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua M. Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics [ISMB]*, 26(12):237--245, 2010.

[213] Virginia Vassilevska and Ryan Williams. Finding, minimizing, and counting weighted subgraphs. In *STOC*, pages 455--464, 2009.

[214] M. Vidal. A biological atlas of functional maps. *Cell*, 104:333--339, Feb 2001.

[215] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417:399--403, May 2002.

[216] C. von Mering, E. M. Zdobnov, S. Tsoka, F. D. Ciccarelli, J. B. Pereira-Leal, C. A. Ouzounis, and P. Bork. Genome evolution reveals biochemical networks and functional modules. *Proc. Natl. Acad. Sci. U.S.A.*, 100:15428--15433, Dec 2003.

[217] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, T. Jatkoe, E. M. Berns, D. Atkins, and J. A. Foekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365:671--679, 2005.

[218] L. D. Wood, D. W. Parsons, S. Jones, J. Lin, T. Sjoblom, R. J. Leary, D. Shen, S. M. Boca, T. Barber, J. Ptak, N. Silliman, S. Szabo, Z. Dezso, V. Ustyanksky, T. Nikolskaya, Y. Nikolsky, R. Karchin, P. A. Wilson, J. S. Kaminker, Z. Zhang, R. Croshaw, J. Willis, D. Dawson, M. Shipitsin, J. K. Willson, S. Sukumar, K. Polyak, B. H. Park, C. L. Pethiyagoda, P. V. Pant, D. G. Ballinger, A. B. Sparks, J. Hartigan, D. R. Smith, E. Suh, N. Papadopoulos, P. Buckhaults, S. D. Markowitz, G. Parmigiani, K. W. Kinzler, V. E. Velculescu, and B. Vogelstein. The genomic landscapes of human breast and colorectal cancers. *Science*, 318(5853):1108--1113, Nov 2007.

[219] G. Wu, X. Feng, and L. Stein. A human functional protein interaction network and its application to cancer data analysis. *Genome Biol.*, 11:R53, 2010.

[220] C. Q. Yang, X. Zhan, X. Hu, A. Kondepudi, and J. F. Perdue. The expression and characterization of human recombinant proinsulin-like growth factor II and a mutant

that is defective in the O-glycosylation of its E domain. *Endocrinology*, 137:2766--2773, Jul 1996.

[221] E. Yeger-Lotem, L. Riva, L. J. Su, A. D. Gitler, A. G. Cashikar, O. D. King, P. K. Auluck, M. L. Geddie, J. S. Valastyan, D. R. Karger, S. Lindquist, and E. Fraenkel. Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat. Genet.*, 41:316--323, Mar 2009.

[222] K. H. Young. Yeast two-hybrid: so many interactions, (in) so little time.. *Biol. Reprod.*, 58:302--311, Feb 1998.

[223] H. Yu, P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J. F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A. S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A. L. Barabasi, J. Tavernier, D. E. Hill, and M. Vidal. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322:104--110, Oct 2008.

[224] A. Zaslaver, A. E. Mayo, R. Rosenberg, P. Bashkin, H. Sberro, M. Tsalyuk, M. G. Surette, and U. Alon. Just-in-time transcription program in metabolic pathways. *Nat. Genet.*, 36:486--491, May 2004.

[225] L. V. Zhang, O. D. King, S. L. Wong, D. S. Goldberg, A. H. Tong, G. Lesage, B. Andrews, H. Bussey, C. Boone, and F. P. Roth. Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network. *J. Biol.*, 4:6, 2005.

[226] X. M. Zhao, R. S. Wang, L. Chen, and K. Aihara. Uncovering signal transduction networks from high-throughput data by integer linear programming. *Nucleic Acids Res.*, 36:e48, May 2008.

[227] X. Zhu, M. Gerstein, and M. Snyder. Getting connected: analysis and principles of biological networks. *Genes Dev.*, 21(9):1010--1024, May 2007.

[228] Y. Zhu, X. Shen, and W. Pan. Network-based support vector machine for classification of microarray samples. *BMC Bioinformatics*, 10 Suppl 1:S21, 2009.