Exploiting structure information for network dissimilarity characterization – Application to disease network analysis and treatment prediction

by

Serene W. H. Wong

A Dissertation submitted to the Faculty of Graduate Studies
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Graduate Program in Computer Science and Engineering
York University
Toronto, Ontario
January, 2014

Abstract

Exploiting structure information for network dissimilarity characterization –
Application to disease network analysis and treatment prediction

Serene W. H. Wong

Doctor of Philosophy

Graduate Department of Computer Science and Engineering

York University

2014

Most cancers lack effective early disease markers, prognostic and predictive signatures, primarily due to tumor heterogeneity. As a result, we fail treating cancer heterogeneity due to multiple ways cancer initiates and develops treatment resistance. Models that represent these differences and the underlying molecular mechanism in cancer enhance the possibility in characterizing and in turn treating cancer successfully.

We introduce novel graph-based methods for exploiting network structure information in the comparison between any graphs, and validate them on non-small cell lung cancer datasets. We generate normal and tumor graphs using normal and tumor samples from gene expression datasets, where vertices are genes, and edges connect co-expressed genes. In the first part of this dissertation, we propose a systems approach with an aim to revert disease conditions to healthy ones through treatments. In order to achieve the objective, we propose three novel methods to 1) systematically identify network structure differences between normal and tumor graphs, 2) identify and prioritize drug combinations based on extracted network structure differences, and 3) computationally estimate the potential of the proposed drug combination to "repair" deregulated subgraphs. Biological validation of the predictions highlights that our systems approach is a promising method to provide treatment options to non-small cell lung cancer through the rewiring of disease networks. In the second part of this dissertation, we introduce the notion of differential

graphlet community to detect deregulated subgraphs between any graphs such that the network structure information is exploited. We observed a trend that the shortest path lengths are shorter for tumor graphs than for normal graphs between genes that are in differential graphlet communities, suggesting that cancer creates shortcuts between biological processes that may not be present in normal conditions. In the third part of this dissertation, we propose a heuristic, the differential correlation graph approach, that identifies areas that are different between the normal and tumor graph, and perform graphlet enumeration on the identified areas. Results showed that our approach achieves accurate estimation in the difference between normal and tumor states by performing network comparisons in important areas only.

Acknowledgements

I am thankful to my supervisors, Prof. Nick Cercone and Prof. Igor Jurisica. I would like to thank Prof. Cercone for his generous financial support, his kindness, his understanding, his guidance and support. I would like to thank Prof. Jurisica for introducing me to this exciting field. I am really thankful for his invaluable guidance, passion, and inspiration.

I would like to thank, past and present, members of the Jurisica lab. Special thanks to: Chiara, for performing biological validations, and answering biology questions; Marc, for carrying out wet lab experiments; Max, for tremendous help all through my Ph.D. studies; Kristen, for helping me to get started in this field; Christian, for maintaining the NX environment, and for introducing me to beaches nearby; Dan, for pointing me to the right data; Ali and Dave, for Navigator help; Richard, for keyboard holder installation; Sara, for going through the Ph.D. experience together. I would not be able to acknowledge every member by name, but thank you all for the excellent learning and collaborative environment.

I would like to thank Hai-Yun for her artistic and technical help. Special thanks to Paul, the computer operations coordinator at York, for his indescribable helpfulness throughout my studies.

Above all, I thank God for His presence, guidance and providence.

Contents

List of Tables

List of Figures

Publications related to the dissertation

Chapter 3 - Network rewiring approach to drug repositioning. Novel non-small cell lung cancer treatment option.

This chapter in part is published in: Serene Wong, Max Kotlyar, Dan Strumpf, Nick Cercone, Frances A Shepherd, Ming-Sound Tsao, Igor Jurisica. Systematic, comparative network analysis on non-small cell lung cancer [poster]. *Proceedings of the American Association for Cancer Research*, Volume 53, March 2012. Abstract no: 4912 [116].

The paper for this chapter is written, and it's under revisions.

Chapter 4 - Comparative network analysis via differential graphlet communities

Serene W. H. Wong, Nick Cercone and Igor Jurisica. *BMC Systems Biology*, under review.

Chapter 5 - A heuristic for finding graphlets that are different between normal and tumor graphs

This chapter in part is based on: Serene Wong, Nick Cercone, Igor Jurisica. Characterizing healthy and disease states by systematically comparing differential correlation networks in lung. *Advances in Health Informatics Conference*, 2012 [115]. Best student paper award in the Advances in Health Informatics Conference, 2012.

Chapter 1

Introduction

The study of large networks for network structure analysis has continued to grow and is an active research area in systems biology (e.g., [58], [89], [80], [40]). Large networks are used to model many real-world phenomena; social, biological and technological networks are a few examples. These real-world phenomena are modeled with nodes and edges of a network where nodes are components and edges are relationship between two components. In this dissertation, we focus on biological networks.

For over a century, reductionism has dominated biological research [10]. Individual cellular components and their functions are studied. Despite of the success of focusing on individual components, it is increasingly clear that rarely an individual molecule is responsible for a biological function. Instead, most biological functions are due to interactions between different cellular constituents. Systems biology is a new dimension to traditional approaches. Systems biology uses a holistic, rather than a reduction view to understand complex biological phenomena [4]. Not only are individual constituents studied, the interactions between constituents in a network are studied as well. For example, how different constituents function together in a network.

One area of biological network research is the analysis of network structures. Analysis of biological network structures can enhance the understanding to biological functions of

cellular components, underlining mechanisms of disease, and drug discoveries. A challenging problem (in the post-genomic era) is to identify relationship between network topology and disease, and network comparisons can obtain such relationship. By comparing networks representing different states, differences of network structure corresponding to different states can be extracted. These extracted network structure differences can be used to gain insights to the underlying mechanisms and treatments for complex disease. Comparing networks with different conditions is extremely useful, for example, comparing networks with different stages or subtypes in cancer, comparing networks with different drug treatments, comparing networks with disease development in different time points.

Graph theory has been an important tool to compare networks, to identify structural properties, and to provide insights to the underlying mechanisms of disease by linking network structures to different types of biological data such as gene expressions, gene signatures, known and novel drugs, and protein-protein interactions. In this dissertation, graph theory will be used for network comparisons. Methods in comparing networks can be applied to other real-world networks, but we confine our comparisons on biological networks. In particular, we compared normal versus non-small cell lung cancer co-expression networks.

A *co-expression network* is an undirected graph such that individual vertices represent genes, and an edge represents the two genes co-expressed. Gene expression studies enable us to understand the mechanism in the molecular level as responses to stimuli are reflected in gene expression levels [24]. Gene expression studies allow the revolutionising of molecular medicine such as the potential to classify, predict diagnosis and prognosis of disease. Thus, we focus on comparing co-expression networks in this dissertation.

According to the American Cancer Society for 2013, in the United States, cancer is the second leading cause of death, and it is estimated that lung cancer accounts for 27% of all deaths from cancer [3]. Lung cancer has three main types, and non-small cell

lung cancer (NSCLC) is the most common type, accounting for 85% of lung cancers [3]. Hence, we are interested in gaining insights in NSCLC.

While methods in comparing networks can be applied to other networks, we confine our network comparisons on normal (usually generated from pathologically normal tissue adjacent to tumor or from healthy controls) versus non-small cell lung cancer co-expression networks.

1.1 Objectives

The goal of this dissertation is to design algorithms to compare network structures between different states of networks systematically. Particularly, graphs that are compared are normal versus disease networks. The ultimate objective in extracting network structure differences is to gain insights to the underlying mechanisms and treatments for complex disease. For example, relating the identified network structures to deregulated genes in signatures, to known and novel drugs, to protein-protein interactions. Yet the main contribution of this dissertation remains to translate insights to the underlying mechanisms or treatments for clinical use. We focus on NSCLC in this dissertation, but the algorithms can be applied generally.

Comparing network structures between graphs provides useful insights, but large graph comparison is computationally intensive as it involves the subgraph isomorphism problem which is NP-complete [45]. Thus, heuristics for network comparison have arisen [92]. In this dissertation, we developed a heuristic that reduce search space by identifying relevant areas for network comparison.

1.2 Contributions

In this dissertation, we address the problem of comparing graphs, and extract network structure differences between them. We focus on comparing normal and disease graphs

in this dissertation, but the algorithms can be applied generally. Most previous network comparisons between healthy and disease networks used 1) simple gene connectivity or its variations; 2) edge or the mean edge weight between groups; 3) membership of cliques to compare the graphs. Furthermore, after differential networks are obtained, often there is no systematic network analysis on them. Although comparing network structures will provide important information for the understanding of disease mechanism, it has not yet been used to its full potential. In this dissertation, we developed novel methods that use network structure information to compare graphs. Based on the extracted network structure differences, we analyzed and designed methods in order to gain insights to the underlying mechanisms and treatments for diseases. While these approaches are generic, we evaluated our approaches on NSCLC datasets.

We proposed a systems approach with an aim to revert disease conditions to healthy ones through treatments. First, we developed a systematic approach to extract network structure differences between normal and NSCLC graphs. Second, based on the network structure differences, a computational method is designed to identify drug combinations in order to "repair" the wiring of the identified subgraphs in tumor samples; i.e., to make the tumor graph more similar to the normal graph. Third, a novel, systematic approach that provides insights on both mechanistic impacts and therapeutic effects of drug treatments on networks is introduced. Validations of drug combination predictions, both mechanistically - measuring whether the graphs are altered as predicted, and functionally - whether the cells show positive effect of the treatment, showed promising results.

We developed a novel method to detect deregulated subgraphs between any graphs such that the network structure information is exploited. Deregulated subgraphs refer to subgraphs that are present in the tumor state, but are not present in the normal state. This approach circumvents the exponential growth of computation required as the deregulated subgraph size increases, and enables the systematically exploring of protein communities with larger size, which provide stronger biological context. Importantly,

this approach has the ability to include a gene into more than one deregulated subgraph. The ability for overlapping deregulated subgraphs is important because genes can have multiple functions under different biological contexts. The analysis led to intriguing results; the difference in network topology between normal and tumor graphs provides insights to the underlying molecular mechanism in NSCLC. In particular, a trend that the shortest path lengths are shorter for tumor graphs than for normal graphs in deregulated subgraphs is observed, suggesting that tumor cells can create shortcuts between biological processes that may not be present in normal conditions.

Comparing network structures between graphs is useful, but large graph comparison is computationally intensive. However, not all areas of the graphs are needed to perform comparisons. We designed a heuristic that identifies areas that are different between the normal and tumor graph, and perform graphlet enumeration on the identified areas. Results showed that our method achieves accurate estimation in the difference between normal and tumor states by performing network comparisons in important areas only.

1.3 Organization of the thesis

The structure of the dissertation is as follow. Chapter 2 defines graph theoretic terminologies and biological terminologies that are used throughout the dissertation, and presents background work that are closely related to our research. Chapter 3 introduces a systems approach with an aim to revert disease conditions to healthy ones through treatments. Chapter 4 presents the notion of differential graphlet community to detect deregulated subgraphs between any graphs such that the network structure information is exploited. Chapter 5 describes a heuristic, the differential correlation graph approach, that reduces search space by identifying relevant areas for graphlet enumeration. Chapter 6 summarizes the contributions of the dissertation and discusses future work. Appendix A presents the prognostic signatures that are used throughout the dissertation. Appendix

B presents the drug concentrations that were used, and the drug validation results for the impact on the deregulated subgraph in Chapter 3. Appendix C presents information regarding the overlapping of genes in differential graphlet communities with pathways and GO biological processes in Chapter 4.

Chapter 2

Background

The main focus of the dissertation is on comparative biological network analysis, thus, we provide a brief overview of the research area in biological network analysis. We begin by first introducing graph theoretic and biological terminologies. Then, we discuss the background work on biological network analysis.

2.1 Graph theoretic terminology

A graph is composed of vertices and edges [113]. Let $G(V, E)$ denote a graph where V is the set of vertices, and E, $E \subseteq V$ x V, is the set of edges in G. $|V|$ denotes the number of vertices in G, and $|E|$ denotes the number of edges in G. $V(G)$ denotes the set of vertices in G, and $E(G)$ denotes the set of edges in G. A graph can be *undirected* or *directed*. A directed graph consists of $V(G), E(G)$ and a function that assigns an ordered pair vertices to an edge. Thus, an edge in a directed graph is an ordered pair. The first vertex of the ordered pair is the tail, and the second vertex of the ordered pair is the head. An edge goes from its tail to its head. A graph is *weighted* if its edges or vertices are associated with a numerical label.

A graph is *complete* if there exists an edge between all pairs of vertices. A complete graph is also known as a *clique*. Let x and y be vertices from G. y is *adjacent* to x if

there is an edge between x and y, and y is a *neighbour* of x. Let $N_n(x)$ denote the set of vertices that are adjacent to x, and $N_n(x)$ is the *neighbourhood* of x.

A *degree* of a vertex x, $d(x)$, is the number of incident edges to x. A *path* in a graph that contains no loop contains vertices that can be ordered such that 2 vertices are adjacent if and only if they are consecutive in the ordering. The *diameter* of a graph is the maximum shortest path length between any pair of vertices. A *cycle* is a graph having the same number of vertices and edges, and its vertices are ordered along a circle such that 2 vertices are adjacent if and only if they are consecutive in the ordering.

A *connected graph*, G, is a graph such that $\forall u, v \in V(G)$, there is a path between u and v; otherwise, G is a *disconnected graph*. A *forest* is a graph with no cycle. A *tree* is a connected graph with no cycle.

A *subgraph* H of G is a graph such that $V(H) \subseteq V(G), E(H) \subseteq E(G)$ and H has the same assignment of vertices to edges as in G. An *induced subgraph*, H, is a subgraph such that $E(H)$ consists of all edges that are connected to $V(H)$ in G. The maximal connected subgraphs of G are called the *components* of G.

2.2 Biological terminology

Proteins are very important components to living organisms, and they are responsible for most functions in a cell [56]. The roles of proteins include providing structural support and infrastructure of living things, they are enzymes that make necessary chemical reactions for life, they are sensors and detectors, and they control the on and off states of genes. All proteins are formed by linear sequences of basic units called the amino acids. Proteins can be as long as 4500 amino acids. Proteins can also fold to form three dimensional structures, which provide specific chemical functionalities.

If proteins are the work horses, then nucleic acids are the drivers that control actions in the biochemical world [56]. Genetic information are all stored in deoxyribonucleic

acid (DNA) which are sequences of nucleic acid units. Each nucleic acid unit is called a nucleotide. There exists four nucleotides in DNA. The encoding for the primary structure of proteins are in nucleic acids, which is the primary role of nucleic acids. The *genome* of an organism refers to all genetic information as a whole there is in an organism.

The basic processes of protein synthesis include transcription, splicing and translation. Transcription is the process to make a messenger ribonucleic acid (mRNA) molecule from a portion of the DNA molecule by using the DNA as a template to make a complementary strand of ribonucleic acid (RNA). This resulted RNA contains both exons and introns. Exons are segments that contains protein coding, and introns are segments that do not contain protein coding. Splicing is the process to take out the introns, and splice the exons together, and the product is then used as the blueprint to make proteins, known as translation [56]. The *central dogma* of molecular biology refers to the process in which information transfer among DNA, mRNA and protein [24].

In general, each cell (with few exceptions) in the body has the same DNA. There are different type of cells in the body, and a lot of difference is in the subset of genes that a cell expressed. Besides the fact that different types of cells can express different subsets of genes, different responses to stimuli can also lead to expressing different subsets of genes [24]. Thus, different cell types or responses to stimuli are reflected in gene expression levels. Gene expression studies enable us to understand the mechanism in the molecular level [24]. Gene expression studies allow the revolutionising of molecular medicine such as the potential to classify, predict diagnosis and prognosis of disease. Studies (e.g., [23]) have shown that genes that are involved in common processes are often *co-expressed*. *Gene expression profile* or *signature* describes a cell's molecular state in a specific condition [118], [24], and can be used to infer cellular phenotypes.

Despite of the success of focusing on individual components, it is increasingly clear that most biological functions are due to interactions between different cellular constituents. Thus, various networks have emerged including *protein-protein interaction*

networks and *co-expression networks*. A *protein-protein interaction (PPI) network* is generally represented as an undirected network that represents physical interactions between proteins. However, in general, protein-protein interactions are directed, we just do not have the directionality information for most physical protein-protein interactions. One of the challenges for computational biology is to predict directionality, strength and time of interactions. A *co-expression network* is an undirected graph such that individual vertices represent genes, and an edge represents the two genes co-expressed. Besides networks, *biological pathways* are important in research in biology. A *biological pathway* is the combination of actions in series among molecules to accomplish tasks such as triggering the assembling of new molecules, turning genes on and off, and can cause other changes in a cell. Some common types of biological pathways involved metabolism, gene regulation and signal transduction [2].

There is a major bioinformatics initiative project, the *Gene ontology* project, that aims to standardize representations of attributes on genes and gene products across databases [8]. *Gene ontology* is often used in the area of biological network research.

2.3 Network properties

In order to compare and characterize different complex networks, some network measures are needed. There are two main categories of network properties used to compare biological networks, global network properties and local network properties. Global network properties study the overall network, while local network properties focus on local structures or patterns of the network [92].

2.3.1 Global network properties

Some global network properties that have been studied extensively include diameter, degree distribution, clustering coefficient and network centrality measures. As stated in

Section 2.1, the *diameter* of a graph is the maximum shortest path length between any pair of nodes according to classical graph theory. However, very often, the *diameter* of a graph is the average shortest path length between all pairs of nodes in the context of analysis of large networks [91]. The *degree distribution* of G, commonly denoted as $P(k)$, is the probability in which any randomly selected node has degree k [90]. The clustering coefficient measures the average probability of two neighbours of any node being adjacent, and centrality measures identify important vertices in complex graphs. The clustering coefficient and centrality measures are to be discussed in turn.

Clustering coefficient

The *clustering coefficient*, C, of a network measures the average probability of two neighbours of any node being adjacent [110]. More formally, let C_x denote the clustering coefficient for node x. Then C_x is defined to be $\frac{2E_x}{n_x(n_x-1)}$ where E_x is the number of edges between all the neighbours of x, and n_x is the number of neighbours of x. C is the average of C_x for all x in the network.

Degree centrality

The idea of *degree centrality* is that a vertex is important if it is involved in many interactions. The *degree centrality* [77] of a vertex u measures the number of edges that are incident to it. *Degree centrality* of u is defined as

$$C_d(u) = d(u).$$

Closeness centrality

The idea of the *closeness centrality* is that a vertex is important if it is "close" to other vertices in the graph. *Closeness centralities*, the *center* and the *median* are defined in

turn. Let $d(x, y)$ be the shortest path length between vertices x and y. The *center* [77], [113] of G , $Cen(G)$ is the set

$$Cen(G) = \{x \in V | e(x) = r(G)\}$$

where $e(x)$ is the *excentricity* of $x \in V$ defined as

$$e(x) = \max_{y \in V} d(x, y)$$

and *radius* $r(G)$ is defined as

$$r(G) = \min_{x \in V} e(x).$$

The *median* [77] of G is the set

$$Med(G) = \{x \in V | s(x) = \sigma(G)\}$$

where $s(x)$ is the status of $x \in V$ defined as

$$s(x) = \sum_{y \in V} d(x, y)$$

and $\sigma(G)$ is defined as

$$\sigma(G) = \min_{x \in V} s(x).$$

Betweenness centrality

The *betweenness centrality* [41], [77] takes into account the global connectivities of G, and it is defined as follow. $\{u, v, w \in V | u \neq v,\ v \neq w\}$,

$$BC(w) = \sum_{u,v \in V} \frac{S_{uv}(w)}{S_{uv}}$$

where S_{uv} is the number of shortest paths between u and v, and $S_{uv}(w)$ is the number of shortest paths between u and v that traverse through w. The idea of *betweenness centrality* is that a vertex is important if it is involved in a high proportion of paths between other vertices.

2.3.2 Local network properties

Global network properties examine network properties of entire networks, but for some applications, for example, the characterization of biological networks, more detailed network properties are needed. In this section, we discuss local network properties, motifs and graphlets.

Network motifs [81] are small subgraphs in a network whose patterns appear significantly more than in randomized networks. Randomized networks in [81] are generated by preserving the number of incoming and outgoing edges for each node. Motifs can be found in different complex networks, from biochemistry, neurobiology, and engineering. Different motifs are found for different types of networks. For example, two transcription-regulatory networks, one from yeast Saccharomyces and another from Escherichia coli, both have a 3-node motif known as the feed-forward loop. However, in food webs networks, this 3-node feed-forward loop is under-represented [81]. Thus, motifs can be used to characterize broad classes of networks. The network motifs approach, however, would miss those subgraphs that are functionally significant but not statistically significant [81].

Another approach for measuring local network properties is the use of *graphlets* [90]. *Graphlets* are all non-isomorphic connected induced graphs on a certain number of vertices. In Figure 2.1, all 5 node graphlets are shown. Our dissertation closely relates to graphlets, thus, we discuss graphlets in detail in Section 2.4.

There are differences between graphlets and motifs. Motifs depend on the randomization scheme, but graphlets do not because graphlets do not have to be over-represented when compared to randomized networks [95]. There are many different models of ran-

dom networks which we would not diverge to in this document, but motifs are dependent upon the model that is chosen. Graphlets, on the other hand, are able to identify all structures, not only the over-represented ones, and are independent of random network models.

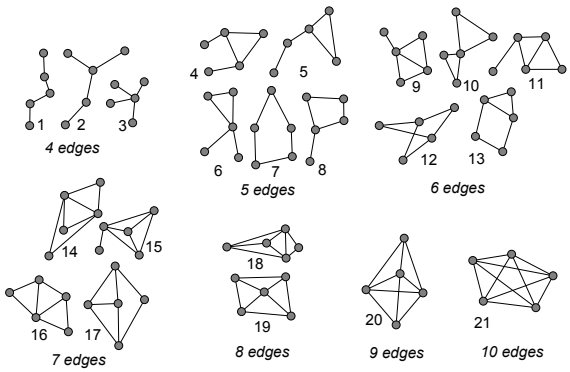


Figure 2.1: All twenty-one 5-node graphlets, all non-isomorphic, connected, induced graphs on 5 vertices.

2.4 Graphlets

Graphlets are all non-isomorphic connected induced graphs on a specific number of vertices [90]. By definition, they have the ability to capture all the local structures on a certain number of vertices. Several graphlet-based network properties have been developed, *relative graphlet frequency distance (RGF-distance)* [94], *graphlet degree distribution agreement (GDD agreement)* [92] and *Graphlet degree signature* [80]. Each of these graphlet-based network properties will be discussed in this section. Pržulj *et al.* used the RGF-distance to determine the random graph model that is the most accurate representation of PPI networks [94]. Furthermore, Pržulj showed that geometric random graphs modeled eukaryotic PPI networks well using the GDD-agreement [92]. A random geometric graph $G(n, r)$ is a graph with n vertices distributed at random independently and uniformly in a metric space with radius r such that $u, v \in V$, $E = \{\{u, v\} | 0 < \|u - v\| \leq r\}$ and $\|\cdot\|$ is any distance norm [94, 86]. Milenković *et al.* demonstrated that proteins

in PPI that are grouped using the graphlet degree signature are in the same protein complexes and the biological functions that they carry out are the same [80].

2.4.1 Relative graphlet frequency distance

A network measure, *graphlet frequency*, is introduced which is the count of the number of graphlets of each category (graphlets 1 to 29 in Figure 2.2) a network contains [94].

The *relative frequency of graphlets* is defined to be $\frac{N_i(G)}{T(G)}$, where $N_i(G)$ is the number of graphlets of category i, $i \in [1, ..., 29]$ in graph G, and $T(G) = \sum_{i=1}^{29} N_i(G)$, that is the total number of graphlets in graph G. In this measure, the similarity of two graphs does not depend on the number of nodes and edges the two graphs have, but instead, should reflect the relative frequency of graphlets [94].

The RGF-distance between graphs G and H is denoted by $D(G, H)$.

$$D(G, H) = \sum_{i=1}^{29} |F_i(G) - F_i(H)|,$$

where $F_i(G) = -log\frac{N_i(G)}{T(G)}$. Logarithm is used due to the difference of frequency in graphlets, at times, the difference can be of order of several magnitude. The use of logarithm would prevent the RGF-distance measure to be totally dominated by the most frequently appeared graphlets [94].

2.4.2 Graphlet degree distribution

The network similarity measure *graphlet degree distribution (GDD)* [92] is based on graphlets, and it's a direct generalization of degree distribution.

Let there be two graphs G and H, and let f be an *isomorphism* from G to H. f is a bijection from $V(G)$ to $V(H)$, and $ab \in E(G)$ if and only if $f(a)f(b) \in E(H)$. An *automorphism* is an *isomorphism* in which a graph maps onto itself. Let $a, b \in V(G)$ and

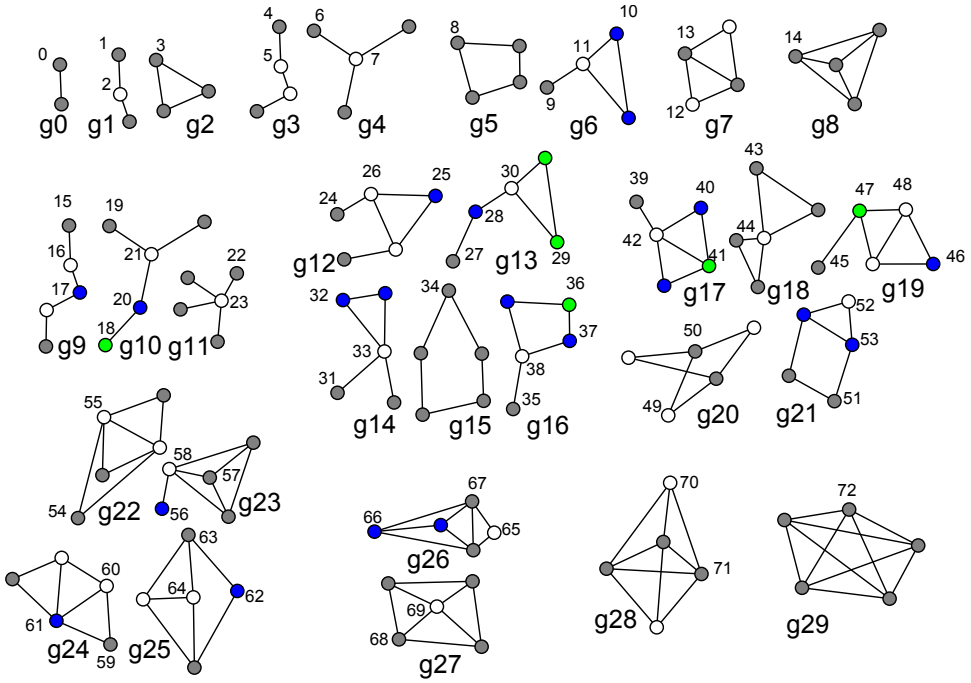


Figure 2.2: 2-5 node graphlets with automorphism orbits 0 .. 72.

$f \in Aut(G)$, where $Aut(G)$ denotes the automorphism group of G. The *automorphism orbit* of a is $\{b \mid b = f(a)\}$.

From graphlets $g0$ to $g29$, there are 73 automorphism orbits, refer to Figure 2.2. There is a graphlet degree distribution for each automorphism orbit, and thus there are 73 graphlet degree distributions.

Based on GDDs, Pržulj [92, 93] developed the *GDD agreement* measure to compare network similarity. The idea is to reduce the 73 GDDs into a scalar agreement between $[0, 1]$ where 0 means that the networks are at a distance, and 1 means that the distributions of the two graphs are identical.

Let $d_G^j(k)$ denote the number of nodes that touch automorphism orbit j in G k times. Pržulj [92] states that most of the information is in the lower degrees of the distribution, and that the information in the extreme high degrees of the distribution is due to noise. Thus, it is desired to scale $d_G^j(k)$ as follow:

$$S_G^j(k) = \frac{d_G^j(k)}{k} \tag{2.1}$$

$S_G^j(k)$ is normalized with respect to the total area, T_G^j, and is denoted as $N_G^j(k)$:

$$T_G^j = \sum_{k=1}^{\infty} S_G^j(k) \tag{2.2}$$

$$N_G^j(k) = \frac{S_G^j(k)}{T_G^j}. \tag{2.3}$$

The distance of the automorphism orbit j between two graphs, G and H is defined to be:

$$D^j(G,H) = \frac{1}{\sqrt{2}}(\sum_{k=1}^{\infty}[N_G^j(k) - N_H^j(k)]^2)^{1/2}. \tag{2.4}$$

The jth GDD *agreement* is defined to be:

$$A^j(G,H) = 1 - D^j(G,H),\ for\ j\ \in\ \{0,1,...,72\}. \tag{2.5}$$

The *agreement* for graph G and H can be defined as the arithmetic mean over $A^j(G,H)$ for all j:

$$A_{arith}(G,H) = \frac{1}{73}\sum_{j=0}^{72} A^j(G,H) \tag{2.6}$$

or the geometric mean over $A^j(G,H)$ for all j:

$$A_{geo}(G,H) = (\prod_{j=0}^{72} A^j(G,H))^{1/73}. \tag{2.7}$$

2.4.3 Graphlet degree signature

A vertex similarity measure is introduced based on the *Graphlet degree signature* for each node [80]. The *Graphlet degree signature* for each node, n, is a vector that has 73

coordinates corresponding to automorphism orbits 0..72 for $2-5$ node graphlets (refer to Figure 2.2), and it counts the number of orbits that n touches. n_j denotes the j^{th} coordinate of n. For example, if n touches orbit 0 once, then n_0 is 1.

To compute node signature distances, in addition to the graphlet degree signature vector, a weight vector, W is also used. W is a vector with 73 coordinates, w_j, $j \in \{0..72\}$, corresponding to the 73 automorphism orbits. Different orbits are assigned different weight because the count of some orbits depend on other orbits. Thus, orbits that are not affected by many other orbits are assigned higher weights. Let o_j denote the number of orbits that affect orbit j, $j \in \{0..72\}$. $w_j = 1 - \frac{log(o_j)}{log(73)}$.

The distance of nodes m, n of orbit j is defined as:

$$D_j(m,n) = w_j \times \frac{|log(m_j+1) - log(n_j+1)|}{log(max\{m_j, n_j\}+2)}$$

The distance between m, n is:

$$D(m,n) = \frac{\sum_{j=0}^{72} D_j}{\sum_{j=0}^{72} w_j}$$

2.5 Biological network comparisons

Several approaches to compare co-expression networks constructed from two conditions, for example, healthy and disease samples, have been developed.

Choi *et al.* compared a normal network with a tumor network by mapping edges in the normal and tumor network to functional interactions using Gene Ontology terms [27]. If *gene* 1 belongs to *category* 1 according to GO annotations, and *gene* 2 belongs to *category* 2, then the edge *gene* $1-$ *gene* 2 is mapped to the category pair *category* $1-$ *category* 2. Measurements *normal coexpression score* (NCS) and *tumor coexpression score* (TCS) are proposed to detect the change in strengths of functional interactions in normal versus in tumor conditions. If there are many co-expressed gene pairs mapped to a particular category pair in normal or in tumor, then the category pair will have a high

NCS or *TCS* respectively. Choi *et al.* identified function interactions that are inactivated or enhanced in cancer when a normal network and a tumor network is compared; however, the comparison considered only edge by edge differences between the two conditions. Fuller *et al.* performed differential network analysis on weighted co-expression networks for lean and obese mice [42]. The connectivity of gene i is defined to be $k_i = \sum_{u \neq i} a_{iu}$ where a_{iu} is the correlation for genes i and u. $k_1(i)$ and $k_2(i)$ are the connectivity for gene i in network k_1 and k_2 respectively. Let $K_1(i) = \frac{k_1(i)}{max(k_1)}$ and $K_2(i) = \frac{k_2(i)}{max(k_2)}$, the differential connectivity is defined to be $DiffK(i) = K_1(i) - K_2(i)$. The identification of genes that are differentially expressed and differentially connected is the aim of the differential network analysis. Fuller *et al.* plotted *DiffK* against student t-test statistic, and obtained eight sectors that have either high absolute values for *DiffK* or t-statistic values, or both. The comparison is on a variation of connectivity of each gene between the two conditions. Watson *et al.* used hierarchical clustering to identify co-expressed gene groups, and used a resampling method to find gene groups that are co-expressed in one condition and not another [109]. The mean pairwise correlations for genes in differentially co-expressed groups are compared between two conditions. Pairwise correlations that have the most changes between the two conditions are examined. The comparison is on each edge weight in the two conditions, and on the mean edge weight between the differentially co-expressed gene groups. Voy *et al.* compared co-expression networks for irradiated and sham-irradiated mice using cliques and edges [108]. The comparison is on the difference of clique memberships on genes or subsets of genes between the two conditions; or on each edge weight in the two conditions.

Other approaches use dependency networks to compare healthy and disease networks. Qiu *et al.* constructed dependence networks for normal and cancer samples [96]. Edges do not represent correlations, but dependency determined by eigenvalue patterns. Dependence-network-based biomarkers are identified by the norm of all columns or rows of the matrix $D_{normal} - D_{cancer}$, where D_{normal} is the adjacency matrix for normal, and

D_{cancer} is the adjacency matrix for cancer. Then, the approach used the change of connections in neighboring nodes of each node to identify biomarkers. Zhang *et al.* detected topological changes between two conditions in transcriptional networks using a differential dependency network analysis [120]. The approach used a local dependency model such that for each node given its parents, there can be more than one conditional probability distribution. Differential dependency networks were extracted; however, there is no systematic analysis on network structure information in the differential dependency networks.

The most straightforward way for such network comparison is to use the connectivity of each gene in the healthy and disease network [33]. Previous methods used diverse approaches to compare two networks: 1) simple gene connectivity or its variations; 2) edge or the mean edge weight between groups; 3) membership of cliques. Furthermore, after differential networks are obtained, there is no systematic network analysis on them. Although comparing network structures will provide important information for the understanding of disease mechanism, it has not yet been used to its full potential. We propose novel methods that use network structure information to compare any graphs.

2.6 Computational challenges

Comparing all aspects of large networks is a challenging problem as it involves the subgraph isomorphism problem which is NP-complete [45]. The subgraph isomorphism problem is defined to be the following: given two graphs G and H as input, determines if there exists a subgraph in G such that it is isomorphic to H.

A version of the subgraph isomorphism problem that is most relevant for our purposes is the enumeration induced subgraph isomorphism problem. The enumeration induced subgraph isomorphism problem is defined such that given G and some small fixed graph H, the solution will include an enumeration of all sets of nodes from G that are isomorphic

induced subgraph to H [91]. In this section, we briefly give an overview of approaches that directly relate to the dissertation, giving more details on approaches that are most related to the focus.

Since the subgraph isomorphism problem is NP-complete, one reasonable solution is to restrict the subgraph size for searching. The limitation is dependent upon the availability of computing power and the algorithm that is used [47]. Algorithms for exact counting of small subgraphs have been developed. For example, Batagelj *et al.* developed an algorithm that returns all triangles in a graph [11], and Marcus *et al.* presented a graphlet enumerator for all graphlets of size up to four [76]. Milo *et al.* presented an algorithm that exhaustively enumerate all subgraphs for a given size, n, in the network [81]. The algorithm scales at least as the size of the network, for it scales with the number of n-subgraph in the network [64].

Wernicke [111] proposed an algorithm, EnumerateSubgraphs (ESU), to enumerate all size-n subgraphs. All vertices in the input graph are labeled in an increasing order. Each vertex of the input graph is a starting point for the algorithm. The algorithm keeps 2 sets of vertices, $V_{Extension}$ and $V_{Subgraph}$. $V_{Subgraph}$ contains the partial list of a subgraph, and $V_{Extension}$ is the working set. The basic idea of the algorithm is as follow. Start with a vertex v, only vertices with the following 2 properties are added into $V_{Extension}$: 1) their labels are greater than that of v and 2) they are neighbors of the newly added vertex, but not the neighbors to vertices already in $V_{Subgraph}$. An arbitrary vertex from $V_{Extension}$ is moved to $V_{Subgraph}$. $V_{Extension}$ is updated according to the above 2 properties. When $|V_{Subgraph}| = n$, then the algorithm outputs the graph corresponding to vertices in $V_{Subgraph}$, and it is a size-n subgraph. ESU is a recursive algorithm, and it is able to enumerate all size-n subgraphs. ESU is implemented in Fanmod, a fast tool to detect network motifs [112].

In addition to restricting the subgraph size for searching, methods that approximate subgraph frequencies instead of counting the exact subgraph frequencies have developed.

Kashtan *et al.* [64] developed a sampling method to estimate the concentration of subgraphs and to detect motifs. Kashtan *et al.*'s sampling method has a major drawback as the method needs to correct the non-uniform sampling problem which is computationally expensive. Wernicke developed an uniform sampling method of size-n subgraphs based on the algorithm ESU that overcomes the above drawback [111].

Pržulj *et al.* developed two heuristics to efficiently estimate graphlet frequency distribution for high confidence PPI networks and geometric random networks: *Neighborhood Local search (NLS)* and *Targeted Node Processing (TNP)* [88].

In NLS, the idea is that a random seed node is chosen from the network, and a specific graphlet is searched in the seed's neighborhood [88]. For any specific graphlet, with n nodes and m edges, the following is performed. A node, v, is randomly selected. v and its neighbors are put in a set. From this set, a subset of n connected nodes are randomly selected. If the subset of n connected nodes contains the induced graphlet G_s with m edges, then the algorithm returns. Otherwise, the neighborhood of G_s is searched for a n-node subgraph with number of edges closer to m than G_s.

In TNP, the idea of the heuristic is that an exhaustive search for graphlets in a small part of the network is performed, and use the graphlet frequency distribution obtained to estimate the graphlet frequency distribution of the entire network [88]. The heuristic is based on the observation that geometric random networks have a sparser boundary, and the rest of the network has a uniform structure. Thus, the hypothesis is that in the sparser boundary of the network, exhaustive search for graphlets can be done quickly; due to the uniformity structure of the rest of the network, the graphlet frequency obtained from the boundary can be used to estimate the graphlet frequency of the network.

Chapter 3

Network rewiring approach to drug repositioning. Novel non-small cell lung cancer treatment option.

3.1 Introduction

Most cancers lack effective early disease markers, prognostic and predictive signatures, primarily due to tumor heterogeneity. As a result, we fail treating cancer heterogeneity due to multiple ways cancer initiates and develops treatment resistance. Models that represent these differences and the underlying molecular mechanism in cancer enhance the possibility in characterizing and in turn treating cancer successfully.

Traditionally, systems approaches that aim to understand and target diseases have been lacking. Most studies focus on individual targets and specific mutations, but not on the impact on signaling or molecular networks [38]. However, knowing how a network responses to a drug is essential for targeting it best. If the way tumor cells are rewired and entered into a new state is known, then it will be easier to force tumor cells out of that state [38].

We use gene expressions from tumor and normal samples to create normal and tumor graphs. Based on coexpression differences, we predict drug combinations that would make the two graphs more similar, with the hypothesis that making tumor graphs more similar to normal graphs will be beneficial. We then validate prediction both mechanistically - measuring whether the graphs are altered as predicted, and functionally - whether the cells show positive effect of the treatment. The goal of our systems approach is to rewire disease networks, i.e., making the disease graph more similar to the normal graph through drug treatments. In order to achieve the objective, there are several problems that need to be addressed. First, we need to systematically identify network structure differences between normal and tumor graphs. Second, we need to identify and prioritize drug combinations based on the extracted network structure differences. Third, we need to systematically quantitate the potential of the proposed drug combinations to "repair" deregulated subgraphs. We use methods based on graph theory to solve these challenges. While the proposed methods are generic, we evaluated them on NSCLC datasets.

While identifying differences between normal and tumor samples using gene groups (e.g., [59]) proved useful, increasing evidence shows that network-based approaches provide substantial benefits (e.g., [58], [28], [40]). For example, Ideker *et al.* showed that top-scoring subnetworks overlap well with known regulatory mechanisms [58]. Chuang *et al.* showed that identified subgraphs were more reproducible, and better predict breast cancer metastasis than individual genes [28]. Fortney *et al.* showed that subnetworks are effective biomarkers in the prediction of aging [40]. Network structure provides evidence for protein function (e.g., [89], [80]). For example, Pržulj *et al.* observed that proteins from different functional classes have different network properties. Milenković *et al.* showed that topologically similar proteins are in same protein complexes, and perform same biological functions. Thus, identifying network structure differences between normal and tumor samples may provide useful clues about carcinogenesis (the process in which cancer cells are transformed from normal cells). In turn, we may be able to

computationally predict what drug combinations will likely "rewire" tumor graphs to resemble normal graphs.

A key drug treatment to cancer involves cytotoxic chemotherapy that kills both cancer and normal cells that divide rapidly [5]. The cytotoxic chemotherapy approach is a one-size-fits-all approach. We are moving towards a new era of personalized molecular medicine, a medical model that customizes treatments to individual patients. The personalized molecular medicine approach is a genetically targeted approach [5]. The recognition of outstanding drug combinations is needed for targeted cancer therapy to be at its full potential as a key challenge in drug combinations in cancer is to overcome genetic heterogeneity and drug resistance [5]. In order to identify desirable drug combinations, new computational methods including network biology approaches are needed [5]. Given the large number of possible drug combinations, it would be infeasible to evaluate all of them in biological experiments due to cost. Thus, computational approaches are desirable to prioritize potential combinations.

As mentioned previously, our systems approach first extracts network structure differences between normal and tumor graphs, then base on the deregulated subgraphs, we identify drug combinations. In order to compare and characterize different complex networks, some network measures are needed. There are two kinds of network measures, global network properties and local network properties. Refer to Section 2.3 for more detail on network properties; we briefly remind the reader about them here. Global network properties examine the overall network, while local network properties focus on local structures or patterns of the network [92]. Some common global network properties used are degree distribution, diameter and clustering coefficient; however, these measures do not contain the detail needed to capture the structural characteristics of biological networks [95]. Thus, more sensitive local structure measurements have emerged. *Graphlets* [90], by definition, have the ability to capture all the local structures on a certain number of vertices. *Graphlets* are all non-isomorphic connected induced graphs on a specific

number of vertices. We propose a graphlet approach to systematically identify network structure differences between any graphs, in this dissertation, between normal and tumor graphs.

The identification of disease pathways and modules in networks greatly contributes to drug development [67]. Given knowledge on mechanism of diseases, prioritization of potential drug targets can be enhanced with networks [57]. The network structure differences obtained through the graphlet approach provide us with potential knowledge of disease mechanism. Exploiting the detected disease modules, we propose a graph-based computational method to prioritize potential drug combinations with a goal to rewire tumor graphs. Importantly, our approach identifies not only individual drugs, but also drug combinations.

We propose a novel, systematic evaluation method in order to determine if the proposed drug combinations are indeed able to "repair" the wiring of deregulated subgraphs in tumor samples. The proposed approach systematically determines the mechanistic impact of drug treatments on: i) the wiring of the edges, ii) individual nodes and iii) the deregulated subgraph. Furthermore, the evaluation provides therapeutic effects on NSCLC. We validated three identified drug combinations on three NSCLC cell lines.

3.2 Methods

Sections 3.2.1 - 3.2.4 describe the datasets, prognostic signatures, notation and implementation. Sections 3.2.5 - 3.2.6 present the graphlet approach. Section 3.2.7 presents the graph-based approach for prioritizing drug combinations. Section 3.2.8 describes the systematic evaluation of mechanistic and therapeutic impact of drug treatments.

3.2.1 Datasets

Datasets were obtained from Gene Expression Omnibus database [36]. We applied our approach to 3 NSCLC microarray gene expression datasets [55, 104, 70]; referred to as Hou, Su, and Landi respectively in this chapter. Datasets were chosen based on the number of normal and tumor samples they contain. Refer to Table 3.1 for more details.

Authors	GSE #	Title	Description
J. Hou *et al. (PLoS One, 2010)*	19188	Expression data for early stage NSCLC	91 patients, 91 tumor and 65 adjacent normal lung tissue samples
L. Su *et al. (BMC Genomics, 2007)*	7670	Expression data from lung cancer	Pairwise tumor-normal samples from 27 patients
M. T. Landi *et al. (PLoS One, 2008)*	10072	Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival	107 lung adenocarcinoma and normal lung samples, 58 tumor and 49 non-tumor tissues

Table 3.1: 3 non-small cell lung cancer datasets are used [55, 104, 70].

3.2.2 Prognostic signatures

Eighteen prognostic NSCLC signatures [44, 12, 14, 114, 17, 37, 106, 26, 50, 75, 87, 97, 15, 72, 71, 73, 105] were used in our approach. Refer to Table A.1 in Appendix A for a list of genes that were used.

3.2.3 Notation

Let $e1$ and $e2$ be edges from G. $e1$ is adjacent to $e2$ if $e1$ and $e2$ share a common vertex. Let $N(e)$ denote the neighbourhood edges of e, i.e., the set of edges that are adjacent to $e \in E$ but not including e.

The set of dataset is denoted as $D = \{Hou, Su, Landi\}$. For $i \in D$, T_i, N_i denote the tumor and normal graph for dataset i respectively. Let S be any subgraph identified to be different between normal and tumor graphs. $T\_all \subseteq E(S)$ is a set of edges such that $t \in T\_all$ is in some T_i but not in any N_i. $abscorr_{T_i}(e)$ represents the absolute correlation value of e in the tumor graph for dataset i. $abscorr_{N_i}(e)$ represents the absolute correlation value of e in the normal graph for dataset i. Let $N_{T_i}(e)$ denote the neighbourhood edges of e in the tumor graph of dataset i.

3.2.4 Implementation

GO enrichment analysis was performed using the GOstats package [39] in R. The t test and Mann Whitney test were performed in R 2.15.0. Node signature distance was computed using graphcrunch 2 [68]. The enumeration of all graphlets was executed using Fanmod [112]. Graph visualization was from NAViGaTOR version 2.3 - Network Analysis, Visualization, & Graphing TORonto [22].

3.2.5 Graphlet approach

Recall that *Graphlets* are all non-isomorphic connected induced graphs on a specific number of vertices [90]. By definition, they have the ability to capture all the local structures on a certain number of vertices. We propose a graphlet approach to systematically identify network structure differences between any graphs, in this dissertation, between normal and tumor graphs [116].

Relative graphlet frequency distance (RGF-distance) [94] and Graphlet degree distribution agreement (GDD-agreement) [92] have been developed as local network structure measures between two graphs using graphlet frequencies and graphlet degree distributions respectively. Both of them return a scalar for the difference between two graphs; thus, throwing away important information that would help further characterization of network structure differences. Previous graphlet based measures are useful for comparing graphs efficiently, since only scalars need to be evaluated. However, our aim is to make the most of graphlet information, and use it to further characterize network structure differences between any graphs. We propose a novel method that make full use of the enumeration of n-node graphlets in graphs A and B. Our method detects deregulated subgraphs that differ between the two graphs, and corresponding network structures from compared graphs will be returned.

We propose a graphlet approach to systematically extract network structure differences between any graphs, in this dissertation, between normal and NSCLC graphs [116]. Normal and NSCLC graphs are generated using coexpression values from normal and NSCLC samples respectively. The construction of graphs is described below. In our approach, we enumerate all n-node graphlets in normal graphs and NSCLC graphs. We then separate the n-node graphlets into different categories, and focus on graphlets that are tumor-specific. As with all exhaustive search algorithms, the graphlet approach is a simple way to solve the problem, and our approach guarantees to extract all n-node graphlets that are different between the normal and NSCLC graph for any specified n. The graphlet approach is systematic because all n-node graphlets from the normal and NSCLC graphs are enumerated, and no subgraph of size n will be missed. Furthermore, by the definition of graphlets, the graphlet approach has the ability to capture local structures of biological networks.

The graphlet approach involves the subgraph isomorphism problem, which is NP-complete. As n increases, the number of different types of subgraphs increases exponen-

tially [94], and the time as well as the memory needed to determine isomorphic subgraphs increases exponentially as well [84]. On the other hand, the number of genes that function together is often more than a few. In previous graphlet-based measures [94, 92], n ranges from 2 to 5. Since exploring protein communities with larger size provides stronger biological context, the largest feasible graphlet size with respect to previous graphlet-based measures is chosen; that is, $n = 5$. Figure 2.1 shows all 5-node graphlets.

Construction of co-expression graphs

While the approach is generic, we evaluated it on three NSCLC gene expression datasets. Three NSCLC gene expression datasets (Section 3.2.1) and eighteen prognostic NSCLC signatures (Section 3.2.2) are the input to the method. The union of genes from all eighteen prognostic gene signatures is denoted as PGS. For each gene expression dataset, $i \in D$, genes in i are intersected with PGS, and the resulting gene set is denoted as gi.

Two co-expression graphs for each dataset, a normal and a tumor graph, are generated using normal and tumor samples, respectively. The co-expression graphs are generated using the following approach, for both normal and tumor samples:

- calculate pairwise Pearson correlations for all gene pairs;
- rank edges according to their absolute correlation values;
- select gene pairs with the top 1% of the absolute correlation values.

The construction of co-expression graphs is highlighted in Figure 3.1.

Enumeration of graphlets

For each dataset, given a normal and a tumor graph, all 5-node graphlets are enumerated. We separate the enumeration of 5-node graphlets into 3 categories.

1. NORMAL: graphlets that are only in the normal graph.

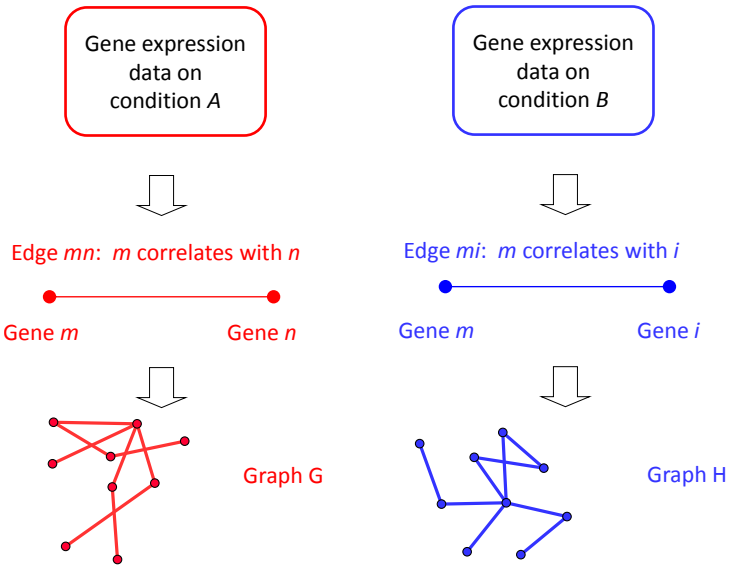


Figure 3.1: The construction of co-expression graphs. Graph G represents condition A, and Graph H represents condition B.

2. BOTH: graphlets that are in the normal and tumor graphs, but with structural differences.

3. TUMOR: graphlets that are only in the tumor graph.

We focus on graphlets that are in the tumor category, and those that have the same membership across all 3 datasets. The graphlet approach identifies interactions between proteins that are deregulated in tumors. Deregulations are seen from the difference in network structures between the normal and tumor graph. The graphlet approach is highlighted in Figure 3.2.

3.2.6 Biological meaning on identified network structures

In order to interpret possible biological meaning of the network structures identified as different between normal and tumor samples, we performed several analyses. We performed GO enrichment analysis using Gene ontology [8] to determine if the identified network structure differences are involved in any biological processes. Let F denote the

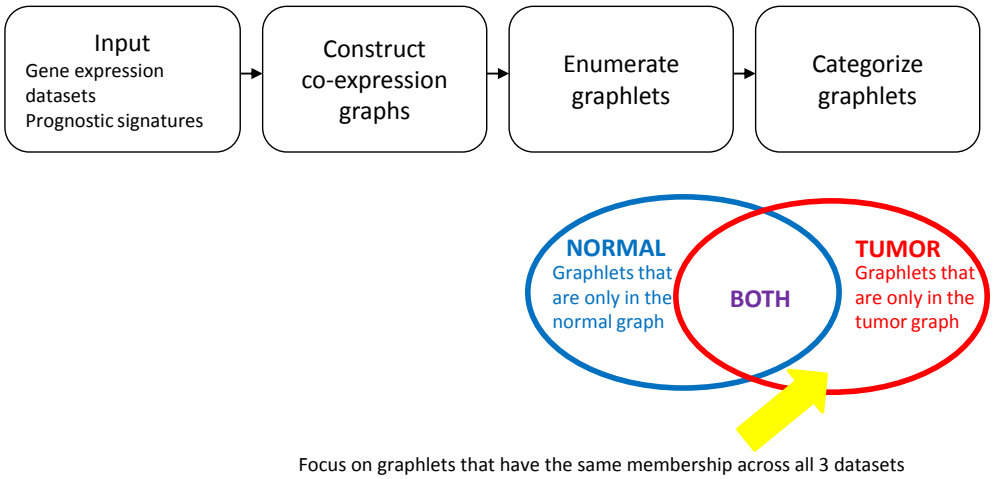


Figure 3.2: The graphlet approach.

union of all subgraphs that are significantly enriched in some GO terms (there was only one significant GO term). We further annotate F with information from literature and PPI databases. In particular, we determine if genes in F are therapeutic targets, and if edges in F are PPIs.

GO enrichment analysis

5-node graphlets in the tumor category that have the same membership across all 3 datasets are compared with the background network. The set of background genes is the intersection of $gi\ \forall i \in D$. GO enrichment analysis is performed using GOstats [39]. A conditional hypergeometric method [39] from GOstats is used. Significant terms are controlled for multiple testing using FDR [13] with a cut-off of 0.05.

PPI analysis

In order to determine if co-expressed edges in identified network structures have PPI evidence, PPI enrichment analysis with the hypergeometric test is performed. Experimentally detected PPIs are obtained from the Interologous Interaction Database (I2D) version 1.95 [21], and computationally predicted PPIs are obtained from FpClass [66]. The set of background genes is the intersection of $gi\ \forall i \in D$, denoted as Bg. The set

of background interactions used in the hypergeometric test is $\binom{|Bg|}{2}$, representing the number of possible interactions among Bg.

3.2.7 Graph-based approach for prioritizing drug combinations

We define a computational method to prioritize drug combinations using identified differences in graphlets. The main goal is to predict drug combinations that have potential to rewire tumor graphs and make them more similar to normal graphs. In order to systematically evaluate and prioritize possible drug combinations, we define an objective function - impact weight function - which maximizes impact while minimizes the number of drugs needed. We first discuss data sources that were used, then we define the impact weight function.

Data source

We have data sources for vertices, edges and drugs.

There are two data sources for vertices, gene expression datasets that are described in Section 3.2.1 and gene expression lung datasets in Cancer Data Integration Portal (CDIP) `http://ophid.utoronto.ca/cdip` (2011). $\forall i \in D$, we determine which genes in F are significantly *up-regulated* or *down-regulated*. A gene is significantly *up-regulated* (or *down-regulated*) if the mean expression value of its tumor (normal) samples is greater than the mean expression value of its normal (tumor) samples, and $p < 0.05$ (t-test, controlled for multiple testing using FDR [13]). Let $a \in \{significantly\ up–\ or\ down–regulated\ genes\ in\ F\}$. $m(a)$ is the number of datasets that have node a as significantly up- or down-regulated in D, and $cdip(a)$ is the number of lung datasets in CDIP that have the direction of node a as datasets in D. $m(a)$ is normalized to $[0, 1]$ by dividing each $m(a)$ by $max\{m(a)\} \forall a$, and $cdip(a)$ is normalized to $[0, 1]$ by dividing each $cdip(a)$ by $max\{cdip(a)\} \forall a$.

The set of genes in F that is significantly *up-regulated* or *down-regulated* plus LCK is denoted as Sg. LCK was included because out of all genes that are not significantly up or down in F, LCK is the most highly connected in F. Our analysis is not only on nodes but on edges as well, thus, it is desirable for LCK to be included.

There are three data sources for edges, gene expression datasets that are described in Section 3.2.1, experimentally detected PPI from I2D [21] and predicted PPI from FpClass [66]. Edge $e \in E(F)$ comprises vertices $\{a, b\}$, $a, b \in V(F)$. A PPI score, $PS(a, b) \in [0, 1]$, is defined for each edge using I2D [21] and FpClass [66]. $PS(a, b)$ is the prediction score from FpClass [66] if (a, b) is a predicted PPI, and 1 if (a, b) is an experimentally detected PPI in I2D [21]. Three scoring functions are defined using the gene expression datasets (see below). Thus, each edge is associated with three scoring functions and a PPI score. The PPI score is used as evidence on edges indicating that edges not only represent co-expression, but also evidence about physical PPI.

The main data source we use for drugs is the Comparative Toxicogenomics Database (CTD) [32]. The chemical-gene interaction table in CTD provides us with drug-vertex pair information. For each gene in Sg, putative compounds which reverse the gene expression direction are obtained from the chemical-gene interaction table in CTD, giving us drug-vertex pairs. The set of putative compounds obtained from CTD is denoted as PC. Additional data sources are used to further prioritize PC. We annotate PC with information from the U.S. Food and Drug Administration (FDA), clinical trials, GI50 values from NCI-60 (we use studies from National Cancer Institute to determine whether drugs are able to cause 50% growth inhibition of the given concentrations on cell lines) and literature (including American Society of Clinical Oncology (ASCO), American Association for Cancer Research (AACR), International Association for the Study of Lung Cancer (IASLC) meetings). Compounds that do not have at least one of the above additional evidence are filtered out, and the set of compound that remains is denoted as PC_f.

PubMed IDs are provided for chemical-gene interactions in CTD. If CTD provides a family of compound for any drug in PC_f, we select a specific compound from that family according to the PubMed reference. For example, polyphenol is in PC_f, and polyphenols are compounds that contain more than one phenol group [32]. A PubMed ID is referenced for our drug-vertex pair that involves polyphenols. The specific compound in the class of polyphenols that CTD references for is Epicatechin. Thus, we use Epicatechin for polyphenols in our drug validation experiment.

These data sources are used to define the impact weight function (see below), and they provide input to the method. Putting all the pieces of information together, the intuition of the approach is as follow. Given 1) an impact weight on vertices (define below) and 2) drug-vertex pairs, the approach searches for best combinations of drugs to "rewire" tumor samples. The method searches for vertices with top impact weights, and from the drug-vertex pairs, chooses drugs that associate with these vertices.

Impact weight

The purpose of the impact weight is to estimate the impact of a drug on a given subgraph. The impact weight comprises two components: edge weight and node weight. Edge $e \in E(F)$ comprises vertices $\{a, b\}$, $a, b \in V(F)$.

The node weight, $wn(a)$, for node a is defined as:

$$wn(a) = m(a) + cdip(a).$$

The edge weight incorporates biological as well as topological information, and it comprises of three scoring functions, $I(e)$, $C(e)$, $S(e)$ and a PPI score (defined above).

1. $I(e) \in [0, 1]$ is the average absolute correlation value of e itself among datasets that contain e in their tumor graphs, but not in their normal graphs. Refer to Algorithm 1.

2. $C(e) \in [0,1]$ is the average absolute correlation difference of $N(e)$ between the tumor and normal graph among datasets that contain e in their tumor graphs. Refer to Algorithm 2. $C_i(e)$ denotes the absolute correlation difference of $N(e)$ between the tumor and normal graph for dataset i.

3. $S(e) \in [0,1]$ is the average node signature distance (described in Section 2.4.3) between the tumor and normal graph for a and b among datasets that contain e in their tumor graphs. Refer to Algorithm 3 for the algorithm of $S(e)$. $S_i(e)$ denotes the node signature distance score for e in dataset i. $D(a\_T_i, a\_N_i)$ denotes the node signature distance (described in Section 2.4.3) between node a in the tumor and normal graph for dataset i. $S_i(e) \in [0,2]$. In order for $S(e)$ to be in [0,1], the average value of $S_i(e)$ is multiplied by $\frac{1}{2}$. It is desirable for $S(e)$ to be in $[0,1]$ as $C(e)$ and $I(e)$ are in $[0,1]$.

Input: An edge e, Graphs T_i, N_i, $\forall i \in D$
Output: $I(e) \in [0,1]$, the average absolute correlation value of e
$counter = 0$;
foreach $i \in D$ **do**
  `// e is not in tumor`
  if $e \notin E(T_i)$ **then**
    go to next dataset;
  `// e in tumor, not in normal`
  else if $e \in E(T_i) \wedge e \notin E(N_i)$ **then**
    $I(e) = I(e) + abscorr_{T_i}(e)$;
    increment $counter$;
  `// Done if normal contains e`
  else if $e \in E(T_i) \wedge e \in E(N_i)$ **then**
    return $I(e) = 0$;
end
return $I(e) = \frac{1}{counter} I(e)$;

Algorithm 1: Compute $I(e)$, the average absolute correlation value of e.

$we(a,b)$ denotes the edge weight for edge e, and is defined as:

Input: An edge e, Graphs T_i, N_i, $\forall i \in D$
Output: $C(e) \in [0, 1]$, the average absolute correlation difference of $N(e)$ between tumor and normal graphs

$counter = 0$;
foreach $i \in D$ **do**
  // e is not in tumor
  if $e \notin E(T_i)$ **then**
    $C_i(e) = 0$;
    go to next dataset;
  else
    increment *counter*;
    // loop through the neighborhood edges of e in tumor
    foreach $l \in N_{T_i}(e)$ **do**
      // assign max weight if l is in tumor, but not in normal
      if $l \in E(T_i) \wedge l \notin E(N_i)$ **then**
        $C_i(e) = C_i(e) + 1$;
      // Take absolute correlation difference if l is in tumor and normal
      else if $l \in E(T_i) \wedge l \in E(N_i)$ **then**
        $C_i(e) = C_i(e) + |abscorr_{N_i}(l) - abscorr_{T_i}(l)|$;
    end
    $C_i(e) = \frac{1}{|N_{T_i}(e)|} C_i(e)$;
end
return $C(e) = \frac{1}{counter} \sum_{i \in D} C_i(e)$;

Algorithm 2: Compute $C(e)$, the average absolute correlation difference of $N(e)$ between tumor and normal graphs.

Input: An edge e comprises of vertices a and b, Graphs T_i, N_i, $\forall i \in D$
Output: $S(e) \in [0, 1]$, the average node signature distance between tumor and normal graphs for a and b

$counter = 0$;
foreach $i \in D$ **do**
  // e is not in tumor
  if $e \notin E(T_i)$ **then**
    $S_i(e) = 0$;
    go to next dataset;
  else
    increment *counter*;
    // calculate node signature distance for a, assign max value if a is in tumor only
    if $a \in V(T_i) \wedge a \notin V(N_i)$ **then**
      $D(a\_T_i, a\_N_i) = 1$;
    else if $a \in V(T_i) \wedge a \in V(N_i)$ **then**
      compute $D(a\_T_i, a\_N_i)$; // using GraphCrunch 2
    // calculate node signature distance for b, assign max value if b is in tumor only
    if $b \in V(T_i) \wedge b \notin V(N_i)$ **then**
      $D(b\_T_i, b\_N_i) = 1$;
    else if $b \in V(T_i) \wedge b \in V(N_i)$ **then**
      compute $D(b\_T_i, b\_N_i)$; // using GraphCrunch 2
    $S_i(e) = D(a\_T_i, a\_N_i) + D(b\_T_i, b\_N_i)$;
end
return $S(e) = \frac{1}{2}(\frac{1}{counter} \sum_{i \in D} S_i(e))$;

Algorithm 3: Compute $S(e)$, the average node signature distance between tumor and normal graphs for a and b.

$$we(a,b) = \begin{cases} I(a,b) + C(a,b) + S(a,b) & \text{if } num_N = 0 \\ 0 & \text{if } num_N \geq 1. \end{cases}$$

where $I(a,b) = I(e) * \frac{1}{3} num_T$, $C(a,b) = C(e) * \frac{1}{3} num_T$, $S(a,b) = S(e) * \frac{1}{3} num_T$, num_T is the number of datasets that have (a,b) in the tumor graphs, and num_N is the number of datasets that have (a,b) in the normal graphs. $I(a,b)$, $C(a,b)$, $S(a,b) \in [0,1]$. $I(e)$, $C(e)$, $S(e)$ are independent of the number of datasets that have e in their tumor graphs. $I(a,b)$, $C(a,b)$, $S(a,b)$ take into account the number of datasets that have (a,b) in their tumor graphs. An edge, e, should have higher weight if all datasets have e in their tumor graphs.

The impact weight of $x \in V(F)$, $impactWeight(x)$, is defined as:

$$impactWeight(x) = wn(x) + \sum_{i \in N_n(x)} wn(i) + \sum_{(x,i) \in E(F)} [we(x,i) + PS(x,i)] \tag{3.1}$$

An impact weight is calculated for each gene in Sg.

Computational method to determine drug combinations

Using the impact weight, the drug-vertex pairs, drug combinations are identified from PC_f using the following criteria:

1. select the gene with maximized impact weight.

2. maximize intersection with genes in Sg.

3. minimize the number of drug used.

We imposed some restrictions on the computation of drug combinations. We limit the maximum number of drug to be 2 in the computation. Suppose that we have drug-vertex pairs $d1 - v1, d2 - v2$ where $d1, d2 \in PC_f$, $v1, v2 \in Sg$, and $impactWeight(v1) >$

$impactWeight(v2)$. Let's also suppose that *drug combination* $\leftarrow \{d1\}$, and $[(v1 \cup N_n(v1)) \cap Sg] \neq Sg$. Then $d2$ can only be added to *drug combination* if $v2 \notin N_n(v1)$. The reason for the latter restriction is that if $v2 \in N_n(v1)$, we infer that $d1$ covers $v2$ already.

3.2.8 Systematic evaluation of mechanistic and therapeutic impact of drug treatments

We propose a systematic evaluation to quantitate the potential of the proposed drug combinations to "repair" deregulated subgraphs. The goal is to validate the proposed treatment options by 1) functional changes of cells, and 2) mechanistic changes of network structures as predicted from the graphlet approach through changes in genes. Functional changes of cells are measured by sulforhodamine B (SRB) assay. SRB assays are used to measure cytotoxicity and cell proliferation caused by the application of drugs [107]. More details on SRB can be found in [107]. We used SRB assay as a surrogate measurement of NSCLC cell viability. Mechanistic changes are measured by quantitative polymerase chain reaction (qPCR) on genes. Biological experiments were performed by Dr. Chiara Pastrello and Marc Angeli in Dr. Igor Jurisica's lab.

We used three NSCLC cell lines (a cell line is a population of cells having the same genetic make-up as the cells are derived from a single cell) for in vitro (experiments conducted in components of an organism that are isolated from an organism's biological environment) validation on our treatment options. A549, H460 and H1975 were used. These cell lines were used because 1) they are NSCLC cell lines, and the graphlet approach is applied on NSCLC gene expression datasets, and 2) genes $\in Sg$ in these cell lines expressing the same direction as in datasets $\in D$ overlap well. A549 and H1975 are cell lines derived from adenocarcinoma, and H460 is from large cell carcinoma. A549, H460 and H1975 cells were cultured, and cell lines were treated with the identified drug

combinations and their individual drugs. Drug concentrations that were used are in Appendix Table B.1.

SRB and qPCR

For each treatment, for each cell line, there are 12 replicates of SRB values. SRB values are averaged for each treatment, for each cell line with outliers removed. For each gene in F, there are 6 replicates of qPCR expression values for each treatment and for each cell line, qPCR expression values are averaged with outliers removed. Outliers are removed when there is an experimental failure (e.g., missing values, values at the extremes of the scale, values lower than the background, values lower than the negative control) or an evident technical variability (e.g., a replicate that is very different from the other replicates).

Rewiring of deregulated edges

The intuition of rewiring deregulated tumor edges is that T_all should "disappear" from F after treatment. Considering that changes are not "black and white", we take edges in T_all that also rank among the top in non-treated cell lines in terms of absolute correlation values, and hypothesize that those edges should rank low after treatment.

For each gene in F, there are 3 averaged real time PCR expression values, one value corresponding to each cell line. For each treatment:

- calculate pairwise Pearson correlations for all gene pairs in F;
- rank edges according to their absolute correlation values.

Edges in T_all that also rank among the top in non-treated cell lines in terms of absolute correlation values are determined. The histogram of absolute correlation values of T_all in non-treated is plotted, and the histogram shows a natural split in absolute

correlation value at 0.8, which is thus used as a threshold to determine which edges rank among the top in non-treated, and this set of edges is denoted as T_11.

Let $R(T\_11)_{NT}$ denote the set of rank for T_11 in non-treated. The set of rank for T_11 in non-treated is obtained as follow. For each edge in T_11, determine where it ranks in the absolute correlation value ranking in non-treated. Let $R(T\_11)_{treatment}$ denotes the set of rank for T_11 in treatment. If we can show that the median of $R(T\_11)_{treatment}$ is larger than that of $R(T\_11)_{NT}$, it would provide evidence that the change is concordant with prediction.

Impact on individual nodes

For each gene in Sg, qPCR expression values are used to determine if the gene expression increased or decreased after treatment with respect to non-treated. Specifically, for each treatment, for each gene in Sg, the averaged qPCR value after treatment is divided by the averaged qPCR value for non-treated (this ratio is also called the *fold change*); if the answer is greater than 1, the direction after treatment for the gene is up, and if the answer is less than 1, the direction after treatment is down. The significance of directions is determined by the two-sided mann-whitney test.

Recall that we have drug-vertex pairs in our prediction, let $d - v$ be a drug-vertex pair where $d \in PC_f$ and $v \in Sg$. The prediction is as follow: if v is up-regulated (or down-regulated), the direction of v will be down (up) after d is applied. Before d is applied, the direction of v is computed as described in Section 3.2.7 using datasets in D. After d is applied, the direction of v is determined as in the aforementioned computation using qPCR data. We compute how many drug-vertex pairs indeed have this effect.

Impact on the deregulated subgraph

For all $v \in Sg$, there is a direction calculated for v. The direction of v is computed as described in Section 3.2.7. For each drug treatment applied, we compute how many

$v \in Sg$ would have its direction reversed. Reverse means that if v is up (or down) before a treatment, then v is down (up) after the treatment is applied. The direction of v after a treatment is applied is determined as in the aforementioned computation using qPCR data.

3.3 Results and Discussion

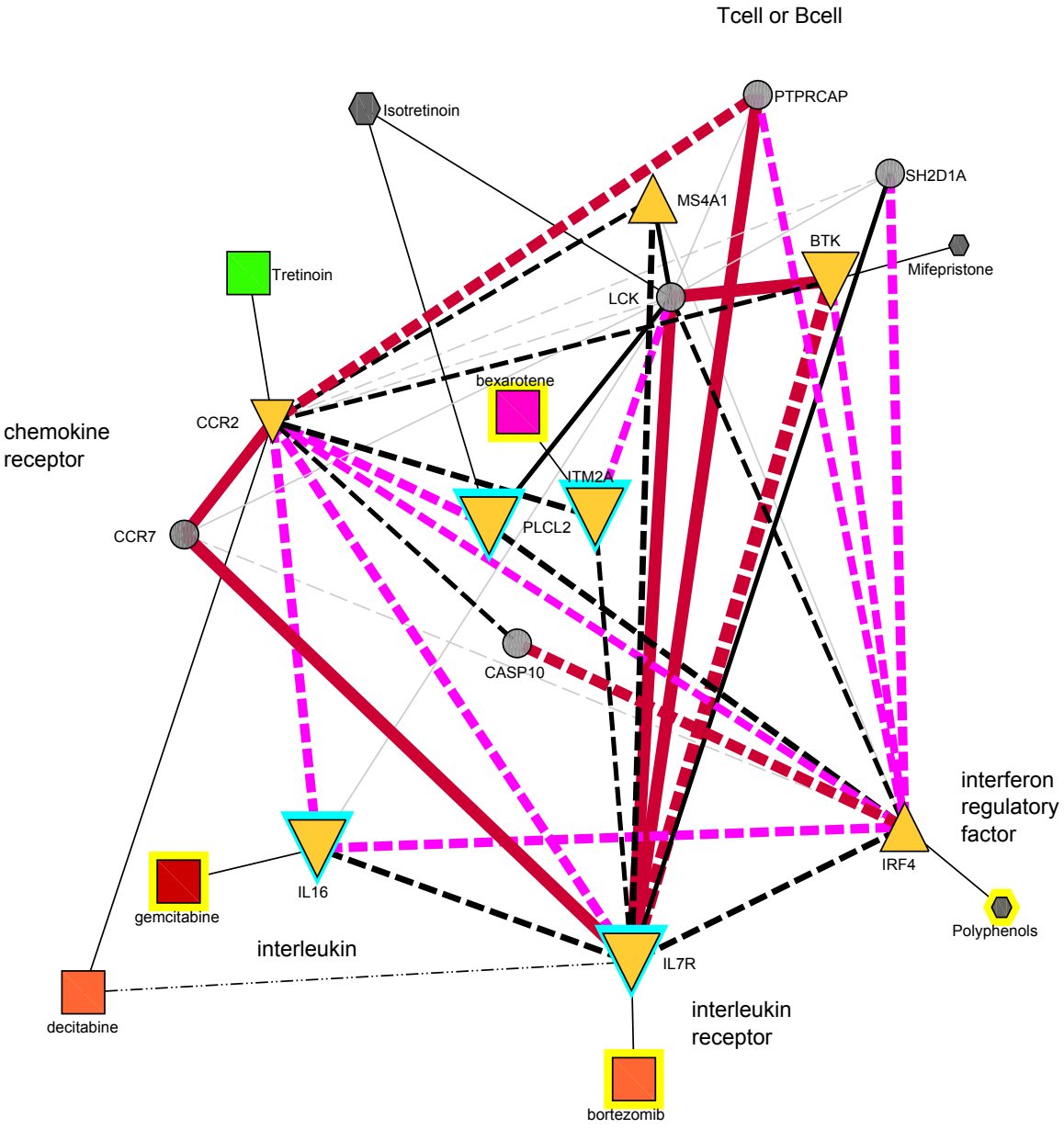
Section 3.3.1 presents the results for the graphlet approach, Section 3.3.2 presents results for the graph-based approach for prioritization of drug combinations, and Section 3.3.3 presents results for the systematic evaluation of mechanistic and therapeutic impact of drug treatments.

3.3.1 Results for the graphlet approach

Gene enrichment analysis

Gene enrichment analysis resulted in 9 subgraphs that are significantly enriched. All 9 subgraphs are enriched in the term "regulation of lymphocyte activation" ($p < 0.05$), and genes are related to chemokine receptors (CCR2, CCR7), interleukin (IL16), interleukin receptor (IL7R), interferon regulatory factor (IRF4), and T cells or B cells (PTPRCAP, SH2D1A, LCK, BTK, MS4A1). Table 3.2 provides the name of the genes. Notably, evading immune destruction is an emerging hallmark of cancer [51].

More importantly, the graphlet approach identified not only gene groups that are different between normal and tumor samples, but also the interactions between genes that are deregulated in tumors. Figure 3.3 shows the resulting subgraph, F, from the union of nodes and edges in the 9 aforementioned subgraphs. Deregulated interactions in tumor that are present in all 3, 2, 1 datasets are in red, pink and black respectively.



Legend for:
Node: gene
Edge: gene-gene

line thickness	width increases with score
dash line	with no PPI evidence
solid line	with PPI evidence
red	3 datasets with tumor
pink	2 datasets with tumor
black	1 dataset with tumor
grey	in normal also
up-triangle	differentially expressed, up-regulated size is w.r.t the # of datasets
down-triangle	differentially expressed, down-regulated size is w.r.t the # of datasets
highlight	*cdip(a)* > 0.33, *a* is differentially expressed

Legend for:
Node: drug
Edge: gene-drug

solid line	drug affects the gene expression in the opposite direction
dash dot-dot line	drug affects the gene expression
square	with recent lung and other cancer trials
polygon	with recent other cancer trials, size is w.r.t. the # of clincial trials for other cancer
red	FDA approved for NSCLC & other cancer; GI50 positive
orange	FDA approved for other cancer; GI50 positive
pink	FDA approved for other cancer
green	GI50 positive
grey	No FDA/GI50 evidence
highlight	with ASCO/IASLC/AACR NSCLC abstracts

Figure 3.3: The union of nodes and edges in the 9 subgraphs in the tumor category that are significantly enriched in the term "regulation of lymphocyte activation" ($p < 0.05$). The drug-vertex pairs are also shown.

Gene Symbol	Name	Related to
CCR2	chemokine (C-C motif) receptor 2	Chemokine receptor
CCR7	chemokine (C-C motif) receptor 7	
IL16	interleukin 16	Interleukin
IL7R	interleukin 7 receptor	Interleukin receptor
IRF4	interferon regulatory factor 4	Interferon regulatory factor
PTPRCAP	protein tyrosine phosphatase, receptor type, C-associated protein	T cells or B cells
SH2D1A	SH2 domain containing 1A	
LCK	lymphocyte-specific protein tyrosine kinase	
BTK	bruton agammaglobulinemia tyrosine kinase	
MS4A1	membrane-spanning 4-domains, subfamily A, member 1	

Table 3.2: Genes identified that are related to the emerging hallmark: evading immune destruction.

PPI analysis

The PPI analysis resulted in 13/38 edges that have known or predicted interaction evidence. This overlap with PPIs is significant $P(X \geq 13) = 4.440892e-16$, as determined using the hypergeometric test. Interactions with PPI evidence are shown as solid lines in Figure 3.3.

Literature evidence

Some genes in F are found to be promising therapeutic targets in other cancers, e.g., Bruton agammaglobulinemia tyrosine kinase (BTK): "Bruton tyrosine kinase represents a promising therapeutic target for treatment of chronic lymphocytic leukemia and is effectively targeted by PCI-32765" [53] and "The Bruton tyrosine kinase inhibitor PCI-32765 blocks B-cell activation and is efficacious in models of autoimmune disease and B-cell malignancy" [54].

Interferon regulatory factor 4 (IRF4) may also be a promising therapeutic target, "IRF4 silencing inhibits Hodgkin lymphoma cell proliferation, survival and CCL5 secretion" [6].

3.3.2 Results for the graph-based approach for prioritization of drug combinations

The purpose of the impact weight is to estimate the impact of a drug in F. Table 3.3 displays the impact weights using formula 3.1.

Node	Neighbors	Impact weight
IL7R	BTK, CCR2, CCR7, IL16, IRF4, ITM2A, LCK, MS4A1, PTPRCAP, SH2D1A	27.63
LCK	BTK, CCR2, CCR7, IL16, IL7R, IRF4, ITM2A, MS4A1, PLCL2, PTPRCAP, SH2D1A	25.62
CCR2	BTK, CASP10, CCR7, IL16, IL7R, IRF4, ITM2A, LCK, MS4A1, PLCL2, PTPRCAP, SH2D1A	23.99
IRF4	BTK, CASP10, CCR2, CCR7, IL16, IL7R, LCK, MS4A1, PLCL2, PTPRCAP, SH2D1A	20.36
BTK	CCR2, IL7R, IRF4, LCK	12.32
IL16	CCR2, IL7R, IRF4, LCK	9.47
MS4A1	CCR2, IL7R, IRF4, LCK	8.24
ITM2A	CCR2, IL7R, LCK	7.49
PLCL2	CCR2, IRF4, LCK	6.65

Table 3.3: The list of impact weights for Sg in the identified subgraph.

Drug combinations are identified using the impact weights, limiting the maximum number of drug used to be 2. For example, IL7R has the highest impact weight (Table 3.3), and its neighbors include all genes in Sg except PLCL2. Thus, IL7R affects all

genes except PLCL2. Therefore, maximum impact with minimum number of drugs should target IL7R and PLCL2. The resulting drug combination comprises Bortezomib + Isotretinoin (see Figure 3.3). Similarly, LCK has the second highest impact weight, and its neighbors include all genes. Thus, the drug in this case is Isotretinoin.

We identified 4 drug combinations that cover all 9 genes (see Table 3.4). We also identified 2 drug combinations that cover 6/9 genes. Combination 5 maximizes the impact weight for drug combinations that cover 6/9 genes. Combination 6 maximizes the impact weight and minimize the overlapping neighbors for drug combinations that cover 6/9 genes.

Combination No.	Drug combinations	Nodes	Impact weight
1	bortezomib	IL7R	27.63
	Isotretinoin	PLCL2	6.65
2	Isotretinoin	LCK	25.62
3	Tretinoin	CCR2	23.99
4	Polyphenols	IRF4	20.36
	Bexarotene	ITM2A	7.49
5	Mifepristone	BTK	12.32
	Gemcitabine	IL16	9.47
6	Mifepristone	BTK	12.32
	Bexarotene	ITM2A	7.49

Table 3.4: This table displays the identified drug combinations. Combinations $1-4$ cover all 9 genes. Combination 5 maximizes the impact weight for drug combinations that cover 6/9 genes. Combination 6 maximizes the impact weight and minimizes the overlapping neighbors.

After prioritizing drug combinations for validation, we also considered combinations which contain drugs currently used, in clinical trials, or reported with potential clinical use. Gemcitabine is an FDA-approved drug for NSCLC, and Bexarotene + Erlotinib's clinical activity is encouraging in NSCLC [35]. We considered combinations that contain Gemcitabine, Bexarotene or Erlotinib as these drugs can be used as a positive control. Thus, our final combinations for biological validation include:

1. Mifepristone + Gemcitabine

2. Polyphenols + Bexarotene + Erlotinib (i.e., Epicatechin + Bexarotene + Erlotinib, see Section 3.2.7)

3. Mifepristone + Bexarotene + Erlotinib.

3.3.3 Results for the systematic evaluation of mechanistic and therapeutic impact of drug treatments

Therapeutic impact of drug treatments

SRB assay was used as a surrogate measurement of NSCLC cell viability. Absorbance was measured after a drug or a drug combination was applied for 48 hours. For each drug combination, we compared the drug combination with the individual drugs that form the drug combination. For Mifepristone + Gemcitabine, we compared Mifepristone + Gemcitabine with Mifepristone individually, and with Gemcitabine individually. For Epicatechin + Bexarotene + Erlotinib, we compared Epicatechin + Bexarotene + Erlotinib with Epicatechin individually, with Bexarotene individually, and with Erlotinib individually. Furthermore, since Bexarotene + Erlotinib has encouraging clinical activity in NSCLC [35], we compared Epicatechin + Bexarotene + Erlotinib with Bexarotene + Erlotinib as well. For Mifepristone + Bexarotene + Erlotinib, we compared Mifepristone + Bexarotene + Erlotinib with Mifepristone individually, with Bexarotene individually, with Erlotinib individually, and with Bexarotene + Erlotinib for the aforementioned reason. Figures 3.4, 3.5, and 3.6 show promising results: for all three cell lines, for all three drug combinations, the cell viability (absorbance percentage with respect to Dimethyl sulfoxide (DMSO)) for the predicted drug combinations are lowest. Importantly, the predicted drug combinations have lower cell viability than Gemcitabine, Erlotinib and Bexarotene + Erlotinib in their respective comparisons. The reader is reminded that Gemcitabine, Erlotinib and Bexarotene + Erlotinib can be used as a positive control as

the former two are FDA approved NSCLC drugs, and the latter has encouraging clinical activity in NSCLC [35].

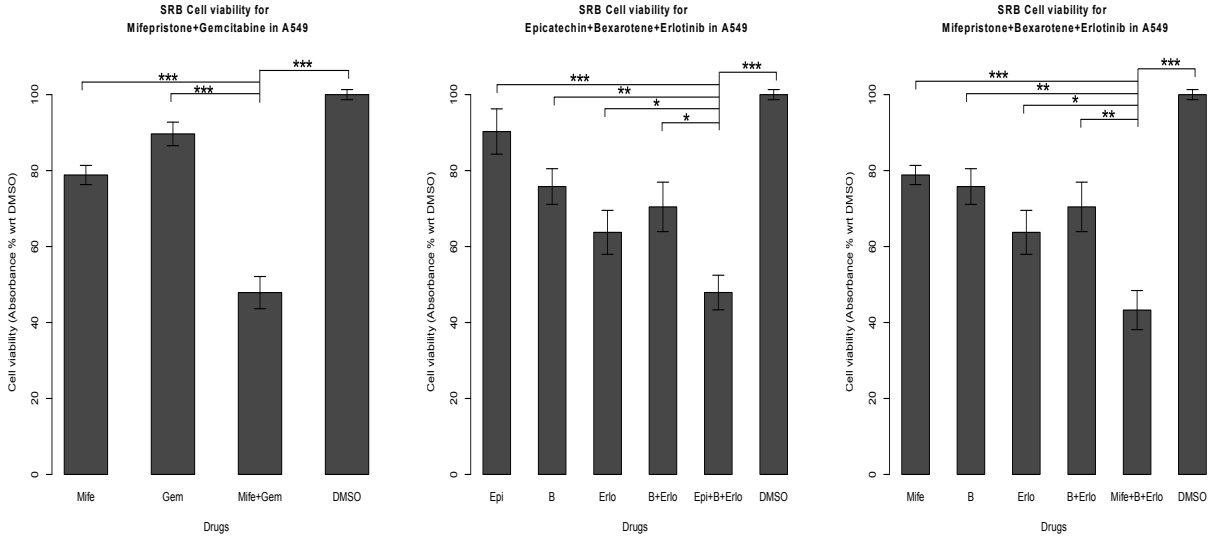


Figure 3.4: The cell viability for all three predicted drug combinations are significantly the lowest in A549. The mean of absorbance percentage with respect to DMSO are shown in the graphs. Error bars represent standard errors. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; unpaired one-sided Mann-Whitney test. B is Bexarotene, Erlo is Erlotinib, Epi is Epicatechin, Mife is Mifepristone and Gem is Gemcitabine.

Results for the rewiring of deregulated edges

Recall that if we can show that the median of $R(T\_11)_{treatment}$ is larger than that of $R(T\_11)_{NT}$, it would provide evidence that the change is concordant with prediction. Table 3.5 shows that for all treatments, the median of $R(T\_11)_{treatment}$ is larger than that of $R(T\_11)_{NT}$. Furthermore, all treatments except for Epicatechin, the median of $R(T\_11)_{treatment}$ is significantly larger than that of $R(T\_11)_{NT}$ (adjusted $p < 0.05$; one-sided Mann Whitney test; adjusted with FDR). The adjusted p value for Epicatechin is slightly above the significance threshold, $p = 0.050142$.

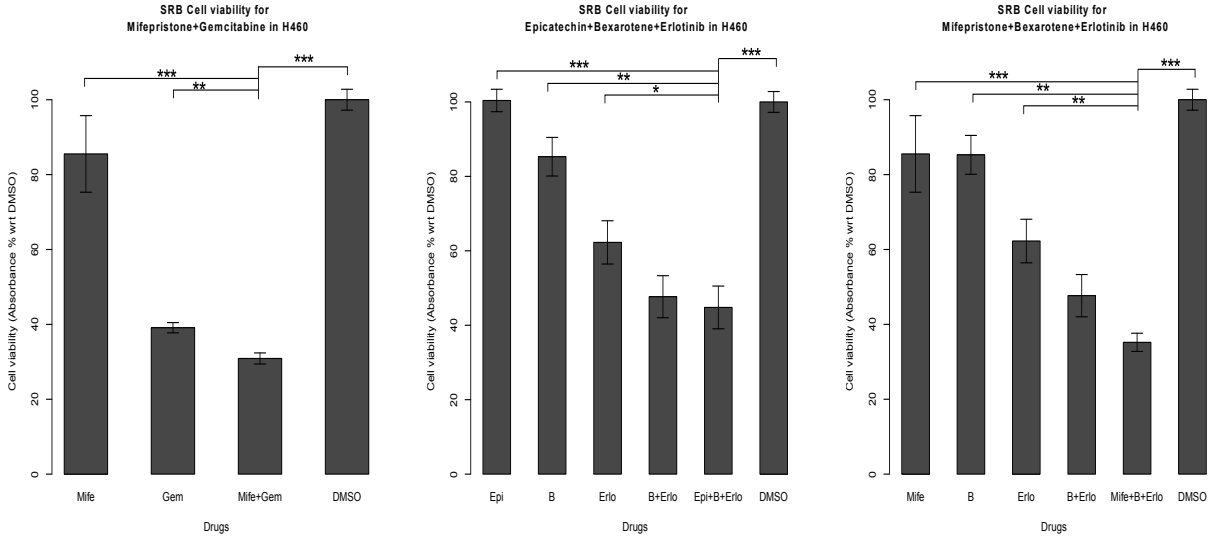


Figure 3.5: The cell viability for all three predicted drug combinations are lowest in H460. The mean of absorbance percentage with respect to DMSO are shown in the graphs. Error bars represent standard errors. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; unpaired one-sided Mann-Whitney test. B is Bexarotene, Erlo is Erlotinib, Epi is Epicatechin, Mife is Mifepristone and Gem is Gemcitabine.

Results for the impact on individual nodes

There are four drugs $\in PC_f$ from the drug combinations for biological validation; thus, we have four drug-vertex pairs to be validated for the impact on individual nodes. Table 3.6 displays the results. The predictions for BTK, ITM2A and IL16 are confirmed in all three cell lines; out of 9 predictions, 6 fold changes are ≥ 1.78. The prediction for IRF4 is confirmed in H460. The effect of Epicatechin on IRF4 may be dependent on histology as the reader is reminded that A549 and H1975 are cell lines derived from adenocarcinoma, but H460 is from large cell carcinoma.

Results for the impact on the deregulated subgraph

Not only do we want to determine the effect of putative compound on individual nodes, we want to further determine the cascade effect on the deregulated network when a drug

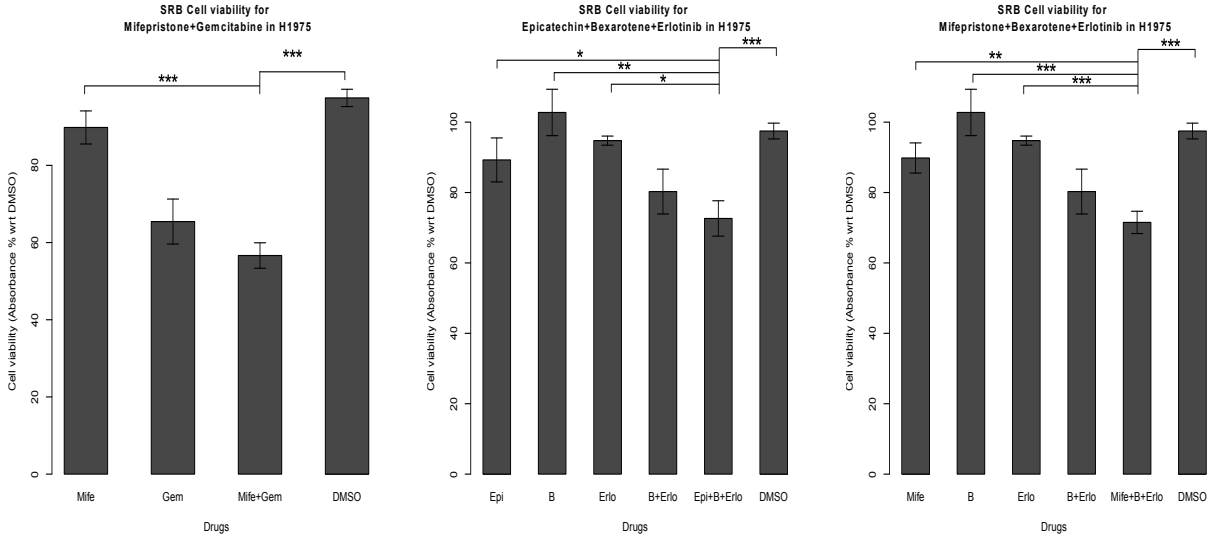


Figure 3.6: The cell viability for all three predicted drug combinations are lowest in H1975. The mean of absorbance percentage with respect to DMSO are shown in the graphs. Error bars represent standard errors. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; unpaired one-sided Mann-Whitney test. B is Bexarotene, Erlo is Erlotinib, Epi is Epicatechin, Mife is Mifepristone and Gem is Gemcitabine.

is applied. All comparisons between predicted and validated directions are in Tables 3.7 - 3.9, all fold changes are in Appendix Tables B.2 - B.4. We highlight several results in the figures and text of this section.

In H460, the Mifepristone + Gemcitabine combination performs the best, better than Mifepristone or Gemcitabine alone; and reverses 8/9 genes in the deregulated subnetwork (Figure 3.7). Recall that Gemcitabine is an FDA-approved NSCLC drug; thus, our predicted combination provides better in vitro results. Epicatechin + Bexarotene + Erlotinib performs well in A549 and H1975, adenocarcinoma cell lines, but not in H460, a large cell carcinoma cell line. In A549, Epicatechin + Bexarotene + Erlotinib reverses 7/9 genes, and the combination performs better than Bexarotene, Erlotinib and Bexarotene + Erlotinib. Note that the predicted combination performs better than Erlotinib, an FDA approved NSCLC drug, which reverses 3/9 genes. Mifepristone + Bexarotene + Erlotinib also performs well in A549, see Figure 3.8.

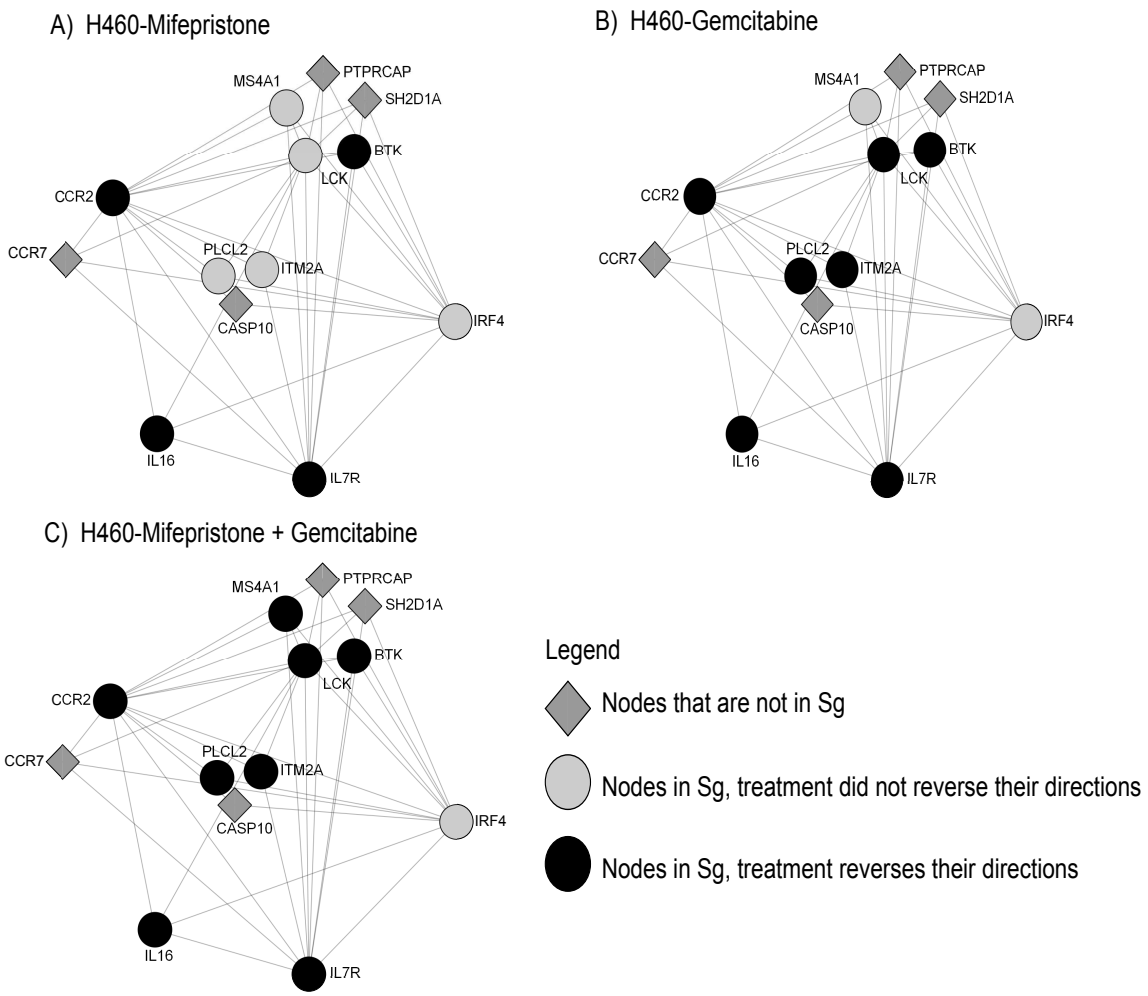


Figure 3.7: In H460, the proposed combination Mifepristone + Gemcitabine (C) performs the best, and reverses 8/9 genes. Mifepristone + Gemcitabine performs better than Mifepristone (A) or Gemcitabine (B) alone, and Gemcitabine is an FDA approved drug for NSCLC.

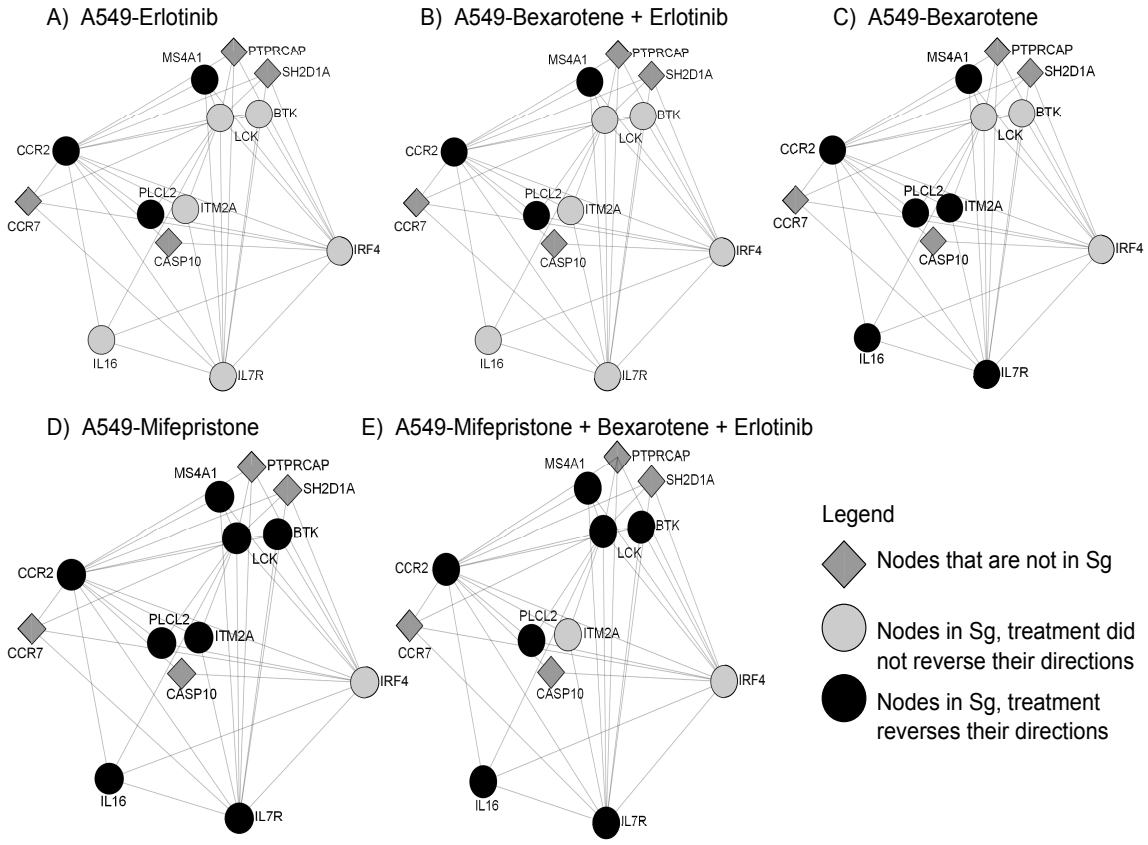


Figure 3.8: In A549, the proposed combination Mifepristone + Bexarotene + Erlotinib (E) performs better than Erlotinib (A), Bexarotene (C) and Bexarotene + Erlotinib (B). Mifepristone + Bexarotene + Erlotinib performs better than Erlotinib (A) and Bexarotene + Erlotinib (B), which is an FDA approved NSCLC drug and a drug combination whose clinical activity is encouraging in NSCLC [35] respectively. Mifepristone + Bexarotene + Erlotinib reverses 7/9 genes while Erlotinib reverses only 3/9 genes, and Bexarotene + Erlotinib reverses 3/9 genes.

Drug	Median of $R(T\_11)_{NT}$	Median of $R(T\_11)_{treatment}$	P value
Bexarotene	16	48	0.02
Bexarotene+Erlotinib	16	40	0.01
DMSO	16	40	0.02
Epicatechin+Bexarotene+Erlotinib	16	46	0.01
Epicatechin	16	57	0.05
Erlotinib	16	38	0.01
Gemcitabine	16	39	0.01
Mifepristone+Bexarotene+Erlotinib	16	38	0.02
Mifepristone	16	37	0.03
Mifepristone+Gemcitabine	16	30	0.02

Table 3.5: Rewiring of deregulated tumor edges. The median of $R(T\_11)_{treatment}$ is larger than that of $R(T\_11)_{NT}$. P values are adjusted using FDR.

Drug - Vertex pair	Prediction	Validation for A549	Validation for H1975	Validation for H460
Mifepristone - BTK	up	up	up	up
Bexarotene - ITM2A	up	up	up	up
Gemcitabine - IL16	up	up	up	up
Epicatechin - IRF4	down	up	up	down

Table 3.6: The first column is the drug-vertex pair. Suppose that we have a drug-vertex pair, $d - v$. Prediction is the predicted direction of v after d is applied. Validation for A549 is the validated direction of v after A549 cells are treated with d. Validation for H1975 is the validated direction of v after H1975 cells are treated with d. Validation for H460 is the validated direction of v after H460 cells are treated with d.

A549: Mifepristone+Gemcitabine										
name	CCR2-dir	BTK-dir	IRF4-dir	PLCL2-dir	IL16-dir	MS4A1-dir	IL7R-dir	ITM2A-dir	LCK-dir	Match
Mife	up* =	up* =	up*	up* =	up =	down =	up =	up* =	up =	8
Gem	up* =	up* =	up*	up* =	up =	up*	up* =	up* =	up =	7
Mife+Gem	up =	up =	up*	up* =	up* =	down =	up =	up* =	up =	8
H1975: Mifepristone+Gemcitabine										
name	CCR2-dir	BTK-dir	IRF4-dir	PLCL2-dir	IL16-dir	MS4A1-dir	IL7R-dir	ITM2A-dir	LCK-dir	Match
Mife	up* =	up =	up*	up* =	up =	up	down*	up =	up =	6
Gem	up* =	up =	down =	up =	up =	up	down	up =	up* =	7
Mife+Gem	up* =	up* =	up	up =	up =	up	down	up =	up* =	6
H460: Mifepristone+Gemcitabine										
name	CCR2-dir	BTK-dir	IRF4-dir	PLCL2-dir	IL16-dir	MS4A1-dir	IL7R-dir	ITM2A-dir	LCK-dir	Match
Mife	up =	up =	up	down*	up =	up	up* =	down	down	4
Gem	up* =	up* =	up*	up =	up =	up	up* =	up* =	up =	7
Mife+Gem	up* =	up =	up	up =	up* =	down =	up* =	up* =	up =	8

Table 3.7: The directions for $v \in Sg$ after A549, H1975 and H460 cells are treated with Mifepristone, Gemcitabine, and Mifepristone+Gemcitabine. * indicates $p < 0.05$ (Two-sided Mann-Whitney test); = indicates the validated direction matches the predicted direction. The column - Match presents the number of genes in Sg such that their validated directions match the predicted directions. Mife is Mifepristone and Gem is Gemcitabine.

A549: Epicatechin+Bexarotene+Erlotinib										
name	CCR2-dir	BTK-dir	IRF4-dir	PLCL2-dir	IL16-dir	MS4A1-dir	IL7R-dir	ITM2A-dir	LCK-dir	Match
Epi	up* =	up* =	up*	up* =	up* =	down =	up =	up* =	up =	8
B	up =	down	up	up =	up =	down =	up =	up =	down	6
Erlo	up* =	down	up*	up* =	down	down =	down	down	down	3
B+Erlo	up =	down*	up	up =	down	down =	down*	down	down	3
Epi+B+Erlo	up* =	up* =	up*	up* =	down	down =	up =	up =	up =	7
H1975: Epicatechin+Bexarotene+Erlotinib										
name	CCR2-dir	BTK-dir	IRF4-dir	PLCL2-dir	IL16-dir	MS4A1-dir	IL7R-dir	ITM2A-dir	LCK-dir	Match
Epi	up =	down	up*	up* =	down	up	up =	up* =	down	4
B	down	down*	up	up =	down*	down =	down*	up =	down	3
Erlo	up* =	up =	up	up* =	down	down =	down*	up* =	up =	6
B+Erlo	up* =	down	up	up* =	down	up	down	up* =	down	3
Epi+B+Erlo	up =	up =	up*	up =	up =	down =	down	up =	down	6
H460: Epicatechin+Bexarotene+Erlotinib										
name	CCR2-dir	BTK-dir	IRF4-dir	PLCL2-dir	IL16-dir	MS4A1-dir	IL7R-dir	ITM2A-dir	LCK-dir	Match
Epi	down	down*	down =	down*	down	down =	up =	down	up =	4
B	up* =	up =	up	up =	up =	up	up =	up =	up =	7
Erlo	up* =	down	down =	down	down	up	up =	up* =	up =	5
B+Erlo	up* =	up* =	down =	down	up* =	down =	up* =	up =	up =	8
Epi+B+Erlo	down*	down	down =	down	down	up	down*	down	up =	2

Table 3.8: The directions for $v \in Sg$ after A549, H1975 and H460 cells are treated with Epicatechin, Bexarotene, Erlotinib, Bexarotene + Erlotinib, and Epicatechin + Bexarotene + Erlotinib. * indicates $p < 0.05$ (Two-sided Mann-Whitney test); = indicates the validated direction matches the predicted direction. The column - Match presents the number of genes in Sg such that their validated directions match the predicted directions. B is Bexarotene, Erlo is Erlotinib, Epi is Epicatechin.

A549: Mifepristone+Bexarotene+Erlotinib										
name	CCR2-dir	BTK-dir	IRF4-dir	PLCL2-dir	IL16-dir	MS4A1-dir	IL7R-dir	ITM2A-dir	LCK-dir	Match
B	up =	down	up	up =	up =	down =	up =	up =	down	6
Erlo	up* =	down	up*	up* =	down	down =	down	down	down	3
B+Erlo	up =	down*	up	up =	down	down =	down*	down	down	3
Mife	up* =	up* =	up*	up* =	up =	down =	up =	up* =	up =	8
Mife+B+Erlo	up* =	up* =	up	up* =	up* =	down =	up =	down	up =	7
H1975: Mifepristone+Bexarotene+Erlotinib										
name	CCR2-dir	BTK-dir	IRF4-dir	PLCL2-dir	IL16-dir	MS4A1-dir	IL7R-dir	ITM2A-dir	LCK-dir	Match
B	down	down*	up	up =	down*	down =	down*	up =	down	3
Erlo	up* =	up =	up	up* =	down	down =	down*	up* =	up =	6
B+Erlo	up* =	down	up	up* =	down	up	down	up* =	down	3
Mife	up* =	up =	up*	up* =	up =	up	down*	up =	up =	6
Mife+B+Erlo	up =	up =	up*	up =	down	down =	down	up =	down	5
H460: Mifepristone+Bexarotene+Erlotinib										
name	CCR2-dir	BTK-dir	IRF4-dir	PLCL2-dir	IL16-dir	MS4A1-dir	IL7R-dir	ITM2A-dir	LCK-dir	Match
B	up* =	up =	up	up =	up =	up	up =	up =	up =	7
Erlo	up* =	down	down =	down	down	up	up =	up* =	up =	5
B+Erlo	up* =	up* =	down =	down	up* =	down =	up* =	up =	up =	8
Mife	up =	up =	up	down*	up =	up	up* =	down	down	4
Mife+B+Erlo	down	down	down =	down*	up =	down =	down*	down*	down	3

Table 3.9: The directions for $v \in Sg$ after A549, H1975 and H460 cells are treated with Mifepristone, Bexarotene, Erlotinib, Bexarotene + Erlotinib and Mifepristone + Bexarotene + Erlotinib. * indicates $p < 0.05$ (Two-sided Mann-Whitney test); = indicates the validated direction matches the predicted direction. The column - Match presents the number of genes in Sg such that their validated directions match the predicted directions. B is Bexarotene, Erlo is Erlotinib, and Mife is Mifepristone.

3.4 Conclusion

We developed a network rewiring approach that provides treatment options to NSCLC. The goal to our systems approach is to rewire disease networks, i.e., making the disease graph more similar to the normal graph through drug combination treatments. In order to achieve the objective, we proposed three novel methods to 1) systematically identify network structure differences between normal and tumor graphs, 2) identify and prioritize drug combinations based on detected deregulated graphs, and 3) computationally estimate the potential of the proposed drug combination to "repair" deregulated subgraphs, making disease graphs more similar to normal graphs.

The systematic graphlet approach resulted in 9 subgraphs significantly enriched in the GO term "regulation of lymphocyte activation" ($p < 0.05$), and evading immune destruction is an emerging hallmark of cancer. Furthermore, the deregulated subgraph is enriched in PPIs, and contains genes that are found in literature to be promising therapeutic targets in other cancers.

Exploiting the identified disease module, we proposed a graph-based computational method to prioritize potential drug combinations with a goal to rewire tumor graphs, making them more similar to normal graphs. Importantly, our approach identifies not only individual drugs, but also drug combinations. We computationally identified 6 drug combinations to rewire the deregulated subgraph.

We performed a systematic evaluation on 3 drug combinations on 3 NSCLC cell lines in order to determine if the predicted drug combinations are indeed able to "repair" the wiring of the deregulated subgraph in tumor samples. The evaluation provides therapeutic effects of the drug combinations on NSCLC as well as the mechanistic impact of drug treatments on: i. the wiring of the edges, ii. individual nodes and iii. the deregulated subgraph.

SRB assay was used as a surrogate measurement of NSCLC cell viability. For each drug combination, we compared the drug combination with the individual drugs that

form the drug combination. For all three cell lines, for all three drug combinations, the cell viability is lowest for the predicted drug combinations. Importantly, the predicted drug combinations have lower cell viability than the tested FDA approved NSCLC drugs in their respective comparisons.

The mechanistic impact of drug treatments is promising as well. For the rewiring of edges, we showed that for all tested drug combinations, the median rank of deregulated tumor edges after treatment is significantly larger than that of non-treated (adjusted $p < 0.05$, one-sided Mann Whitney test, adjusted using FDR), resulting in rewired disease graphs that are more similar to normal graphs. Furthermore, the mechanistic impact of drug treatments on individual nodes are encouraging. The predictions for BTK, ITM2A and IL16 are confirmed in all three cell lines, and the prediction for IRF4 is confirmed in the large cell carcinoma cell line. The results for the cascade treatment effect on the deregulated network is also reassuring. For example, the Mifepristone + Gemcitabine combination in H460 performs extremely well; not only did the predicted combination performs better than Mifepristone or Gemcitabine alone, but it is able to reverse 8/9 genes in the deregulated subnetwork.

Results have shown that our systems approach is a promising method to provide treatment options to NSCLC through the rewiring of disease networks, i.e., making the disease graph more similar to the normal graph through drug combination treatments.

Chapter 4

Comparative network analysis via differential graphlet communities

4.1 Background

The identification of differences between healthy and affected tissues is important for the understanding of disease. Differential expression studies that compare gene expression levels between healthy and affected tissues have been developed [33]. Differential expression studies usually involve detecting statistical significance changes to the mean expressions of individual genes [30]. Some studies associated changes in mean expression levels in gene groups or pathways with disease phenotypes [33]. However, useful prognostic signatures are not necessarily the most differentially expressed genes [19]. Differential co-expression approaches that compare co-expression patterns between healthy and diseased samples have been developed (e.g., [65]). Studies have identified several highly differentially co-expressed transcriptional regulators involved in cancer, but their mean expressions did not change much [33].

Identification of differences between healthy and diseased tissues is important, but the difference should not be limited to gene groups. Difference in network structure

is essential as studies have shown that systematically studying structural properties of biological networks can bring forth important insights, for example, determining the relationship between network topology and protein functions, or network topology and the underlying disease mechanism. Jeong *et al.* suggested that the most highly connected proteins are those that are the most important to survival [60]. Pržulj *et al.* observed that lethal proteins are not only highly connected, but they are articulation points [89]. Jonsson *et al.* provided insight of global network properties of cancer proteins, and found that cancer proteins, on average, had twice as many interacting partners as non-cancer proteins [61]. These results have to be interpreted carefully as trends can be due to literature bias; however, they suggest that there is a relationship between structures and functions in networks that needs to be explored further.

Furthermore, network-based approaches have been successful in identifying subnetworks for classification (e.g., [28]), for recovering of known and uncovering of novel biological functions (e.g., [58]). For example, Ideker *et al.* showed that top-scoring subnetworks overlap well with known regulatory mechanisms [58]. Chuang *et al.* showed that identified subgraphs were more reproducible, and better predict breast cancer metastasis than individual genes [28]. Subnetworks have also been shown to be effective biomarkers in the prediction of aging [40]. Thus, identification of differences between healthy and diseased tissues should include differences in network structures.

Several approaches to compare co-expression networks constructed from healthy and disease samples have been developed, Section 2.5 provides details on them.

In order to compare and characterize different complex networks, we can use global or local network properties. Refer to Section 2.3 for more detail on network properties; we briefly remind the reader about them here. Global network properties examine the overall network, while local network properties focus on local structures or patterns of the network [92]. One approach for measuring local network properties is the use of graphlets. *Graphlets* are all non-isomorphic connected induced graphs on a specific

number of vertices [90]. By definition, they have the ability to capture all the local structures on a certain number of vertices.

In Chapter 3, we propose a novel method that make full use of the enumeration of n-node graphlets in graphs A and B [116]. The graphlet approach detects deregulated subgraphs that differ between the two graphs, and corresponding network structures from compared graphs are returned. In this chapter, we propose another novel graphlet-based method to identify network structure differences between any graphs. Our approach circumvents the exponential growth of computation required as the graphlet size increases, and enables the systematically exploring of protein communities with larger size, which provide stronger biological context. The size of our detected deregulated communities can be much larger than the size of individual graphlets.

We introduce the notion of *differential graphlet community* to detect deregulated subgraphs between any graphs such that the network structure information is exploited. The differential graphlet community approach overcomes a limitation of some existing approaches (e.g., [42, 109]); importantly, it has the ability to include a gene into more than one deregulated subgraph. The ability for overlapping differential graphlet communities is important because genes can have multiple functions under different biological contexts. While the differential graphlet community approach is generic, we evaluated our approach on three NSCLC datasets. The approach led to intriguing results; the difference in network topology between normal and tumor graphs provides insights to the underlying molecular mechanism in NSCLC. In particular, a trend that the shortest path lengths are shorter for tumor graphs than for normal graphs in differential graphlet communities is observed, suggesting that tumor cells can create shortcuts between biological processes that may not be present in normal conditions. Examples of shortcuts that are observed, and are in agreement with known mechanism in literature include the crosstalk between the Jak-STAT and NF-kappaB pathways or STAT3 signaling enabling crosstalk among tumor and immune cells, resulting in an immunosuppressive network.

4.2 Methods

4.2.1 Graphlet approach

We proposed a graphlet approach in Chapter 3 to systematically extract network structure differences between normal and NSCLC graphs [116]. We briefly review the graphlet approach in this section. The graphlet approach enumerates all n-node graphlets in normal graphs and NSCLC graphs. This method involves the subgraph isomorphism problem, which is NP-complete [45]. As n increases, the number of different types of subgraphs increases exponentially [94], and the time and memory needed to determine isomorphic subgraphs increases exponentially as well [84]. The use of differential graphlet communities can help circumvent this exponential growth of computation and space required. Importantly, the number of genes that function together is often more than a few. Previous approaches considered $2-5$ node graphlets [94, 92]. Since exploring protein communities with larger size provides stronger biological context, the largest feasible graphlet size with respect to previous graphlet-based measures is chosen; that is, n is 5. The graphlet approach is systematic because all 5-node graphlets from the normal and NSCLC graphs are enumerated, and no subgraph of size 5 will be missed.

4.2.2 Differential graphlet community

Enumerating 5-node graphlets ensures that all non-isomorphic connected induced graphs on 5 nodes will be considered. However, the number of genes that function together is often more than 5. Furthermore, any 2 graphlets, A and B can potentially have 4 nodes that overlap. Thus, we extend the approach to consider graphlet communities with a goal to identify the difference in the properties of networks between different graphs – in this dissertation, between normal and tumor graphs.

Palla *et al.* [85] defines a community as the union of all k-cliques such that one can reach to another by a chain of adjacent k-cliques. A k-clique is a complete graph with k

vertices. Adjacent k-cliques are k-cliques that share $k-1$ nodes. A differential graphlet community is defined as the union of all k-graphlets such that one can reach to another by a chain of adjacent k-graphlets. Adjacent k-graphlets are graphlets that share $k-1$ nodes. Since all 5-node graphlets are enumerated, k is 5 for the purpose of this chapter. Figure 4.3 illustrates the notion of differential graphlet community.

The differential graphlet community approach detects deregulated subgraphs that differ between two graphs. There are several advantages to the differential graphlet community approach. First, the proposed approach has the ability to include a gene into more than one deregulated subgraph. The ability for overlapping differential graphlet communities is important because genes can have multiple functions in biological systems. Second, the differential graphlet community approach circumvents the exponential growth of computation required as the graphlet size increases, and enables the systematically exploring of protein communities with larger size which provide stronger biological context. Thus, although the size of each graphlet is 5, the sizes of differential graphlet communities can be much larger. Third, no predetermined size or number of deregulated subgraphs are required as input to the method, size and the number of communities are determined automatically.

The construction of co-expression graphs is discussed in Section 3.2.5. For each dataset, given a normal and a tumor graph, all 5-node graphlets are enumerated. We separate the enumeration of 5-node graphlets into 3 categories.

1. NORMAL: graphlets that are only in the normal graph.
2. BOTH: graphlets that are in the normal and tumor graphs, but with structural differences.
3. TUMOR: graphlets that are only in the tumor graph.

We focus on graphlets that are in the tumor category, and those that have the same membership across all 3 datasets. Differential graphlet communities are then computed

for the extracted graphlets. The differential graphlet community analysis identifies interactions between proteins that are deregulated in tumors. Deregulations are seen from the difference in network structures between the normal and tumor graph.

4.2.3 Datasets

We applied our approach to 3 NSCLC microarray datasets [55, 104, 70], referred to as Hou, Su, and Landi. The same datasets are used in Chapter 3. Refer to Section 3.2.1 for more detail. Datasets have been selected based on the number of normal and tumor samples they contain, and were downloaded from Gene Expression Omnibus database [36].

We used four independent NSCLC microarray gene expression datasets [74, 99, 82, 48] to validate our results (referred to as Lu, Sanchez, Okayama and Girard, respectively). Table 4.1 provides more detail on the 4 datasets.

4.2.4 Notation

Let Hou_N, Su_N, $Landi_N$ denote the normal graphs for Hou, Su, and Landi respectively. Similarly, let Hou_T, Su_T, $Landi_T$ denote the tumor graphs for Hou, Su, and Landi respectively.

Let g_{T-Hou}, g_{T-Su}, $g_{T-Landi}$ denote the set of graphlets that are in the tumor category for datasets Hou, Su, and Landi respectively. Let M_{TALL} denote the set containing sets of 5 vertices such that $V(h) = V(s) = V(l)$ for some $h \in g_{T-Hou},\ s \in g_{T-Su},\ l \in g_{T-Landi}$. $|M_{TALL}|$ is the number of graphlets that have the same membership across all 3 datasets in the tumor category.

Differential graphlet communities are then computed on g_{T-Hou} for all $h \in g_{T-Hou}$, g_{T-Su} for all $s \in g_{T-Su}$, $g_{T-Landi}$ for all $l \in g_{T-Landi}$ such that $V(h), V(s), V(l) \in M_{TALL}$.

Authors	GSE #	Title	Description
T. P. Lu *et al.*	19804	Genome-wide screening of transcriptional modulation in non-smoking female lung cancer in Taiwan	120 samples: 60 normal samples, 60 tumor samples
A. Sanchez-Palencia *et al.*	18842	Gene expression analysis of human lung cancer and control samples	91 samples: 45 controls, 46 tumor samples
H. Okayama *et al.*	31210	Gene expression data for pathological stage I-II lung adenocarcinomas	246 samples: 20 normal samples, 226 tumor samples
L. Girard *et al.*	31547	MSKCC-A Primary Lung Cancer Specimens	50 samples: 20 adjacent normal lung controls, 30 tumor samples

Table 4.1: 4 other independent non-small cell lung cancer gene expression datasets [74, 99, 82, 48].

We have identified three differential graphlet communities for each dataset, referred to as: $dGC_{Hou}i,\ i \in \{1,2,3\}$ for Hou, $dGC_{Su}i,\ i \in \{1,2,3\}$ for Su and $dGC_{Landi}i,\ i \in \{1,2,3\}$ for Landi. Importantly, note that $V(dGC_{Hou}i) = V(dGC_{Su}i) = V(dGC_{Landi}i),\ i \in \{1,2,3\}$ respectively, and thus the computation returns the same number of differential graphlet communities for each dataset.

All shortest paths are computed between all vertex pairs in $V(dGC_{Hou}i),\ i \in \{1,2,3\}$ for Hou_N and for Hou_T. All shortest paths are computed between all vertex pairs in $V(dGC_{Su}i),\ i \in \{1,2,3\}$ for Su_N and for Su_T. Finally, all shortest paths are computed between all vertex pairs in $V(dGC_{Landi}i),\ i \in \{1,2,3\}$ for $Landi_N$ and for $Landi_T$.

Let $dGCsp_{HouN}i,\ i \in \{1,2,3\}$ denote the shortest path graph for differential graphlet community i for dataset Hou in Hou's normal graph. $dGCsp_{HouN}i,\ i \in \{1,2,3\}$ con-

tains all shortest paths in Hou_N between all vertex pairs in $V(dGC_{Hou}i),\ i \in \{1,2,3\}$. Let $dGCsp_{HouT}i,\ i \in \{1,2,3\}$ denote the shortest path graph for differential graphlet community i for dataset Hou in Hou's tumor graph.

4.2.5 Shortest path distribution

After obtaining deregulated subgraphs, comparing network structures is important for the understanding of disease mechanisms. In order to better utilize network structure information obtained from the deregulated subgraphs, we computed shortest path distributions on differential graphlet communities.

Visualization of differential graphlet communities in NAViGaTOR [22] shows that there are fewer vertex pairs xy such that x is adjacent to y among vertices in $V(dGC_{Hou}i),\ i \in \{1,2,3\}$ for Hou_N than in $dGC_{Hou}i,\ i \in \{1,2,3\}$ respectively. Similar results are observed for Su and Landi datasets. To quantify these observations, we performed a systematic shortest path distribution analysis.

Shortest path distributions are computed for:

- $dGCsp_{HouN}i,\ i \in \{1,2,3\}$ and $dGCsp_{HouT}i,\ i \in \{1,2,3\}$;
- $dGCsp_{SuN}i,\ i \in \{1,2,3\}$ and $dGCsp_{SuT}i,\ i \in \{1,2,3\}$;
- $dGCsp_{LandiN}i,\ i \in \{1,2,3\}$ and $dGCsp_{LandiT}i,\ i \in \{1,2,3\}$.

Significance of shortest path distribution differences between normal and tumor graphs is determined by the Mann-Whitney test. A constant C is used to replace infinity distance (i.e., non-reachable vertices). By the nature of the Mann-Whitney test, results from different Cs will be the same if C is greater than all non-infinity lengths in the compared shortest path distributions. Thus, without loss of generality, C is set to be 100 as the maximum shortest path length is 12.

4.2.6 Pathway and GO analysis

In order to gain biological insights from network structures of the differential graphlet communities, and to test whether edges in differential graphlet communities are within a pathway or across pathways, nodes are overlapped with pathways and Gene Ontology (GO). Pathway databases used include Encyclopedia of Homo Sapiens Genes and Metabolism (HumanCyc) [98], Kyoto Encyclopedia of Genes and Genomes (KEGG) [63], National Cancer Institute - Pathway Interaction Database [100], Reactome [78] and The Cancer Cell Map [9]. KEGG was downloaded on Feb 2011; remaining databases were downloaded from Pathway Commons [25] on Aug, 2012. Annotations for GO ontology - biological process were downloaded from Quick GO from European Bioinformatics Institute [16] on Aug, 2012.

The intersection of $dGCsp_{HouT}i$, $dGCsp_{SuT}i$ and $dGCsp_{LandiT}i$ is taken for $i \in \{1, 2, 3\}$, and is denoted as $dGCsp_{ALL}i,\ i \in \{1, 2, 3\}$. $V(dGCsp_{ALL}i),\ i \in \{1, 2, 3\}$ are intersected with individual pathways and GO biological processes.

4.2.7 Implementation

The shortest path distribution analysis and the differential graphlet community analysis were written using the igraph package [29] version 0.5.5.2 in R. The differential graphlet community analysis adapted the implementation of the clique percolation algorithm in the wiki website of igraph [1]. The Mann-Whitney test was performed in R 2.15.0. The enumeration of all 5-node graphlets was executed using Fanmod [112]. Fanmod is a fast tool to detect network motifs, and contained an algorithm, EnumerateSubgraphs (ESU), by Wernicke [111], to enumerate all size-n subgraphs. Graph visualization was from NAViGaTOR version 2.3 - Network Analysis, Visualization, & Graphing TORonto [22].

4.3 Results and discussion

We identified three differential graphlet communities for each dataset; for all 3 differential graphlet communities, for all 7 datasets, we observed a trend that the shortest path lengths are shorter for tumor graphs compared to normal graphs. Differential graphlet communities $dGC_{Hou}i$, $dGC_{Su}i$ and $dGC_{Landi}i$, $i \in \{1, 2, 3\}$ are presented in Figures 4.1-4.3. Note that the difference in wiring in individual datasets could be due to the difference in disease stage as well as the difference in histology.

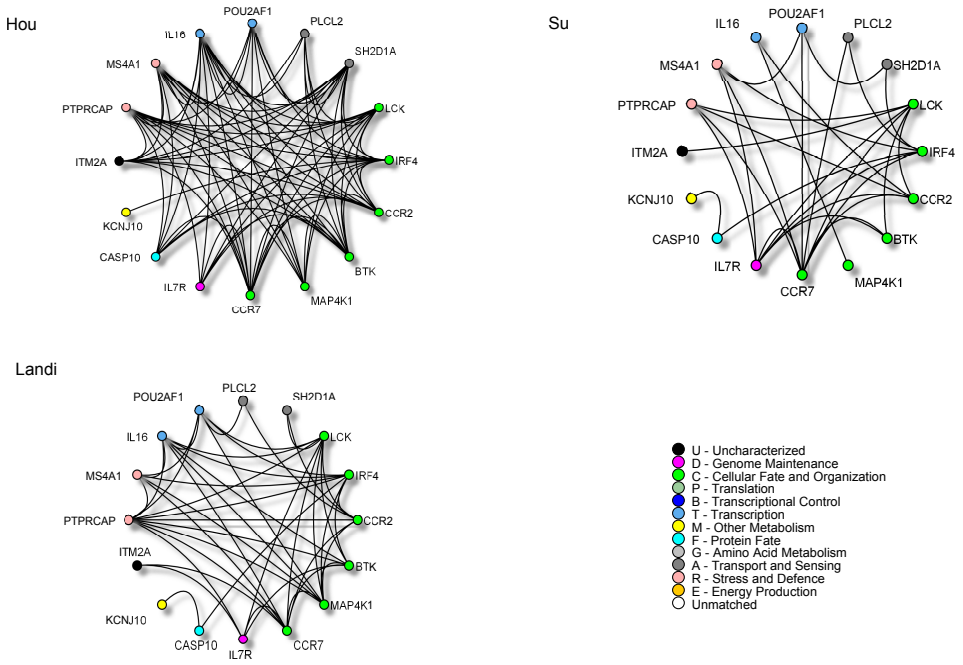


Figure 4.1: $dGC_{Hou}1$, $dGC_{Su}1$ and $dGC_{Landi}1$ are shown. Edges connect co-expressed genes. Nodes are sorted and colored based on GO biological function.

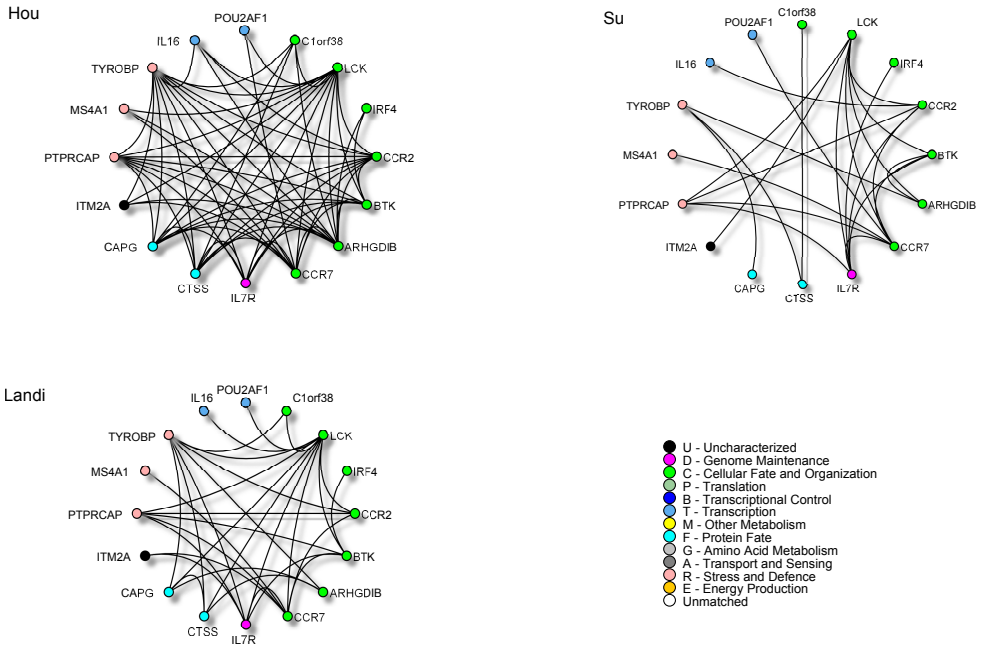


Figure 4.2: $dGC_{Hou}2$, $dGC_{Su}2$ and $dGC_{Landi}2$ are shown. Edges connect co-expressed genes. Nodes are sorted and colored based on GO biological function.

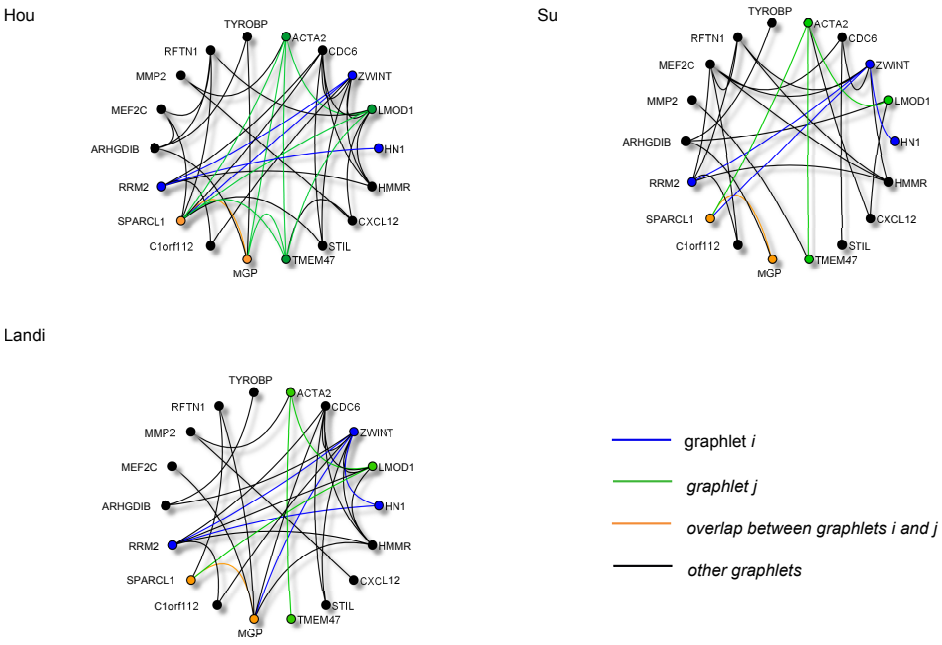


Figure 4.3: $dGC_{Hou}3$, $dGC_{Su}3$ and $dGC_{Landi}3$ are shown. Edges connect co-expressed genes. Differential graphlet communities are formed by graphlets; graphlet i is in blue, and graphlet j is in green for some i, j that form $dGC3$. Other graphlets that form $dGC3$ are in black (other graphlets that overlap with graphlets i, j are not shown).

We also present the comparisons of shortest path distributions for:

- $dGCsp_{HouN}i$ versus $dGCsp_{HouT}i$ for $i \in \{1, 2, 3\}$;
- $dGCsp_{SuN}i$ versus $dGCsp_{SuT}i$ for $i \in \{1, 2, 3\}$;
- $dGCsp_{LandiN}i$ versus $dGCsp_{LandiT}i$ for $i \in \{1, 2, 3\}$.

For readability, simpler terms are used in the Figures. For example, shortest path distribution for Landi for $dGC1$ refers to the comparison of the shortest path distributions between $dGCsp_{LandiN}1$ and $dGCsp_{LandiT}1$.

Figures 4.4 - 4.6 show that for all 3 datasets, for all 3 differential graphlet communities, tumor graphs have shorter shortest paths than normal graphs; the median of shortest path lengths in normal is significantly larger compared to tumor graphs (adjusted p values $\leq 1.13E-20$; one-sided Mann-Whitney test). This suggests that tumor cells can cause crosstalk between biological processes that usually does not exist under normal conditions.

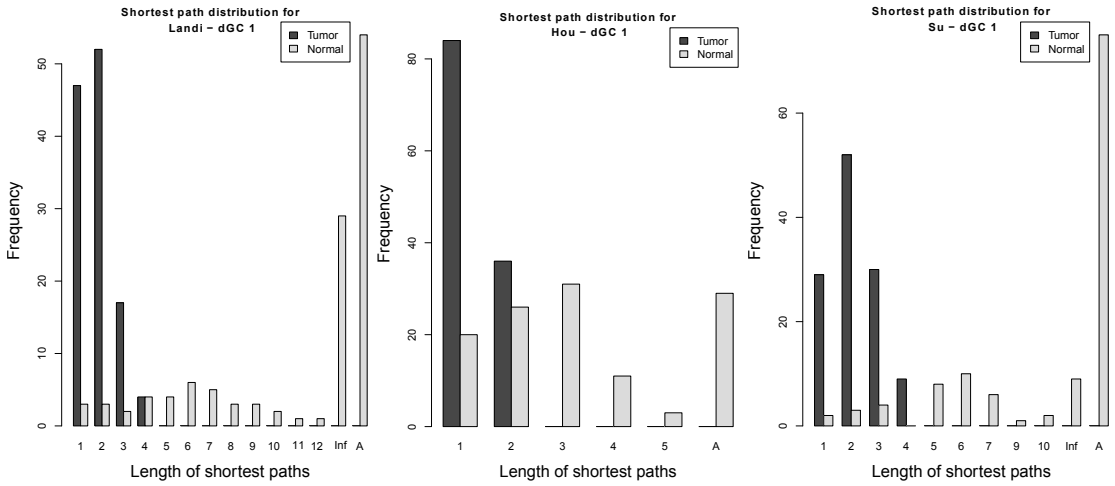


Figure 4.4: Shortest path distributions for $dGC1$ for Landi, Hou and Su datasets. Inf represents shortest path between unreachable nodes. A is the number of node pairs that have infinity as the distance due to the absence of nodes in the graph.

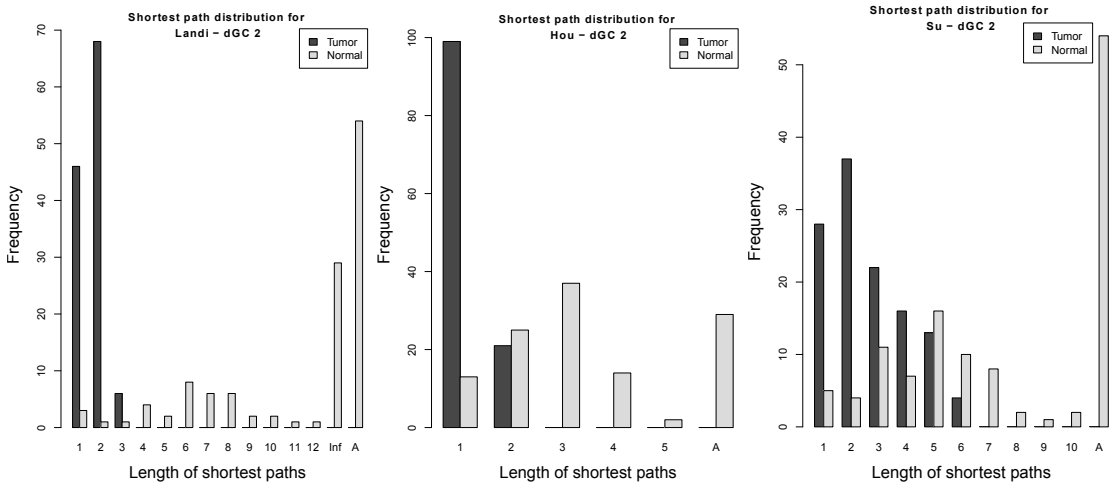


Figure 4.5: Shortest path distributions for $dGC2$ for Landi, Hou and Su datasets. Inf represents shortest path between unreachable nodes. A is the number of node pairs that have infinity as the distance due to the absence of nodes in the graph.

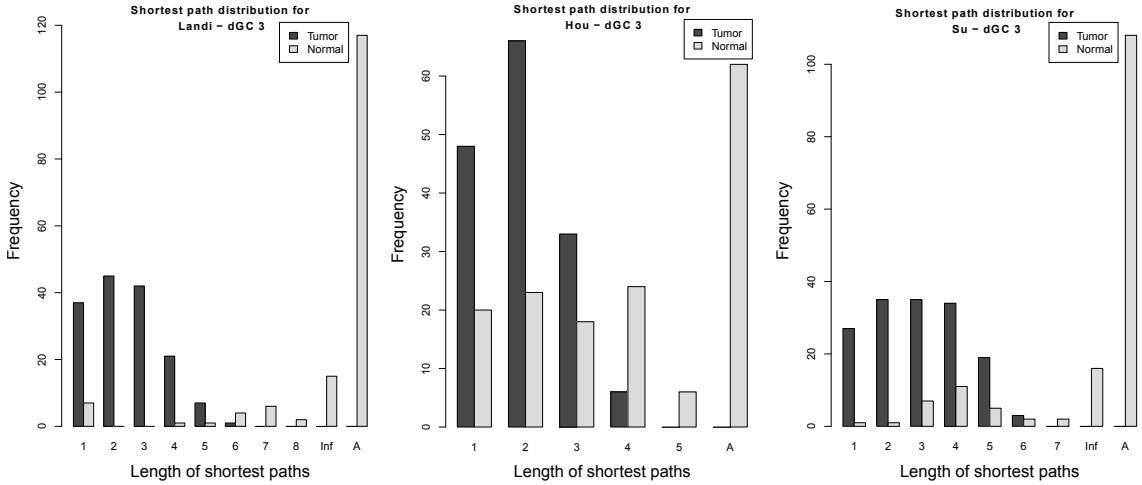


Figure 4.6: Shortest path distributions for $dGC3$ for Landi, Hou and Su datasets. Inf represents shortest path between unreachable nodes. A is the number of node pairs that have infinity as the distance due to the absence of nodes in the graph.

To further validate the observed trend, we used four independent NSCLC datasets – Lu, Sanchez, Okayama and Girard [48, 74, 82, 99]. In all 4 datasets, for all 3 differential graphlet communities, the observed trend is confirmed: tumor graphs have shorter shortest paths compared to normal graphs; the median of shortest path lengths in normal is

significantly larger than tumor graphs (adjusted p values $\leq 2.61E-13$; one-sided Mann-Whitney test). Figures 4.7, 4.8 and 4.9 show the observed trend for different datasets for differential graphlet community 1, 2 and 3, respectively.

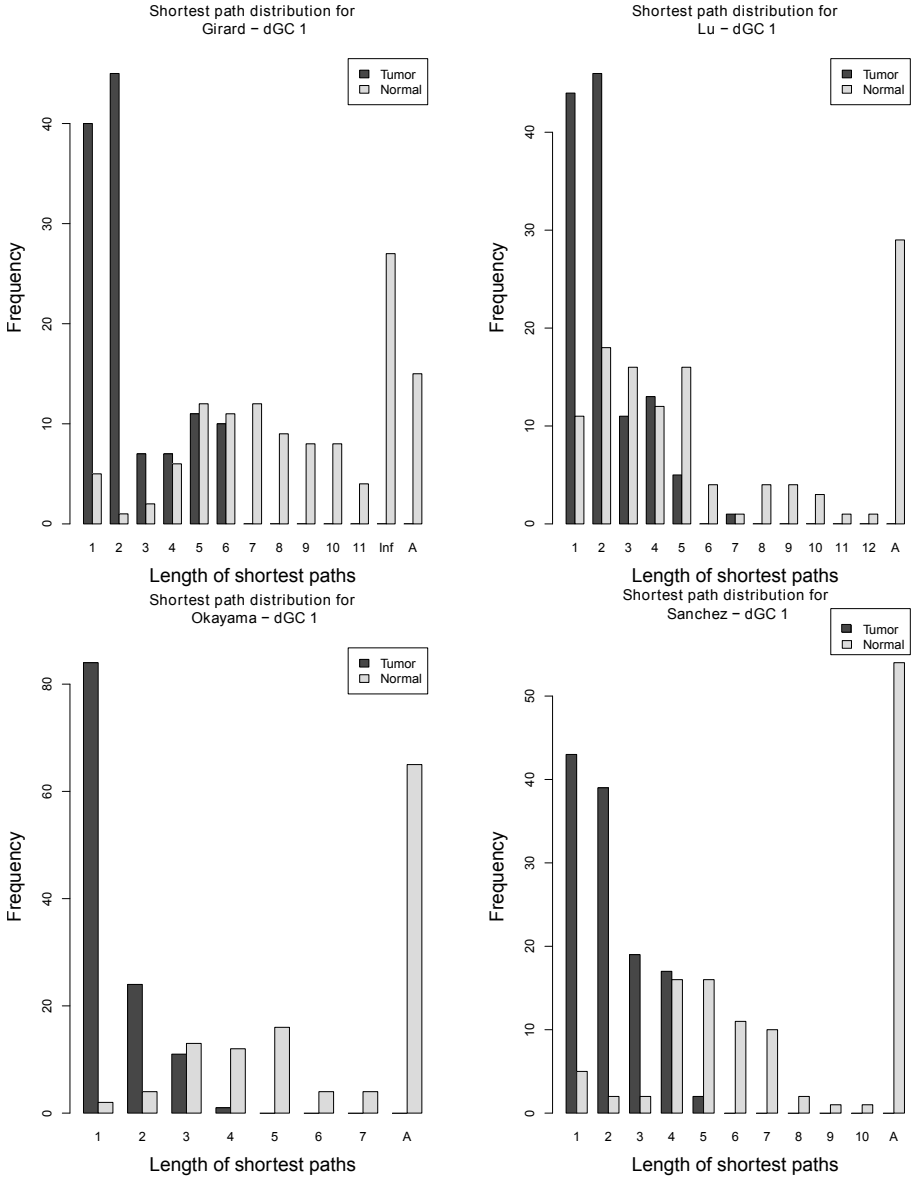


Figure 4.7: Shortest path distribution for $dGC1$ for Girard and Lu datasets are shown at the top. Shortest path distribution for $dGC1$ for Okayama and Sanchez datasets are shown in the bottom. Inf represents shortest path between unreachable nodes. A is the number of node pairs that have infinity as the distance due to the absence of nodes in the graph.

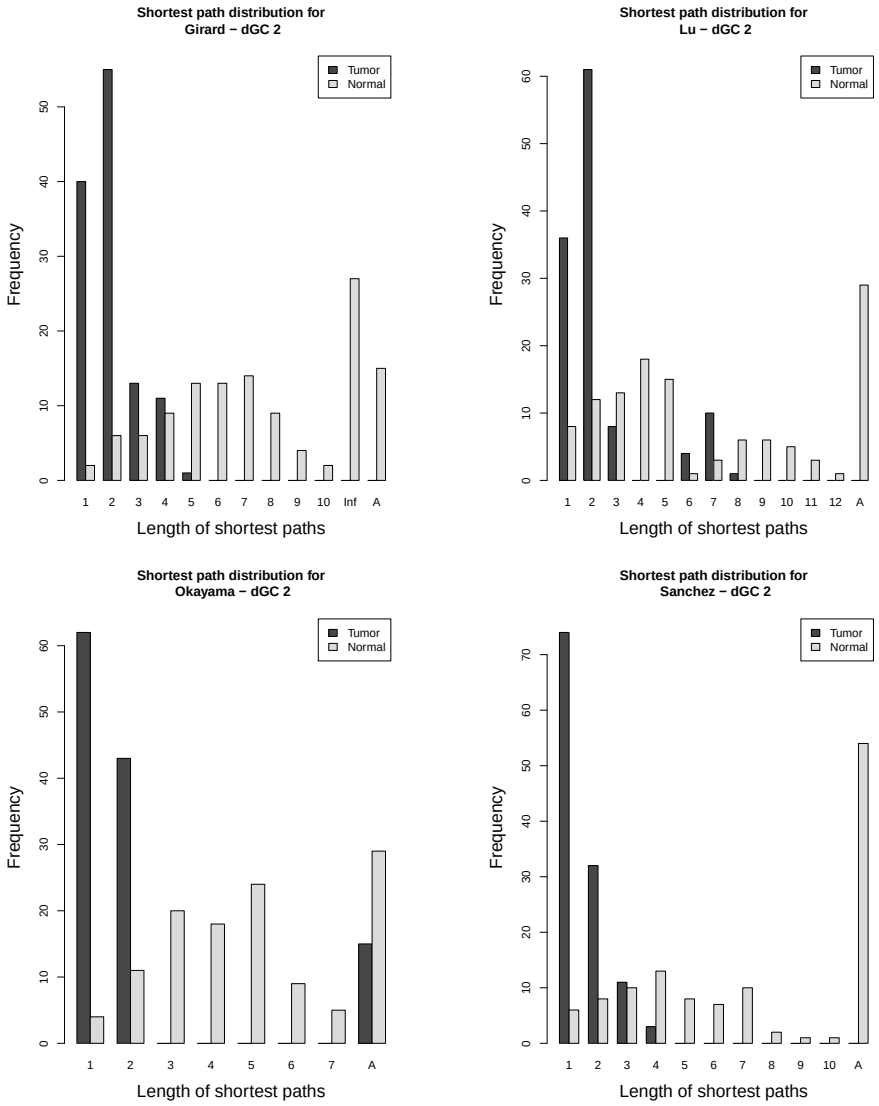


Figure 4.8: Shortest path distribution for $dGC2$ for Girard and Lu datasets are shown at the top. Shortest path distribution for $dGC2$ for Okayama and Sanchez datasets are shown in the bottom. Inf represents shortest path between unreachable nodes. A is the number of node pairs that have infinity as the distance due to the absence of nodes in the graph.

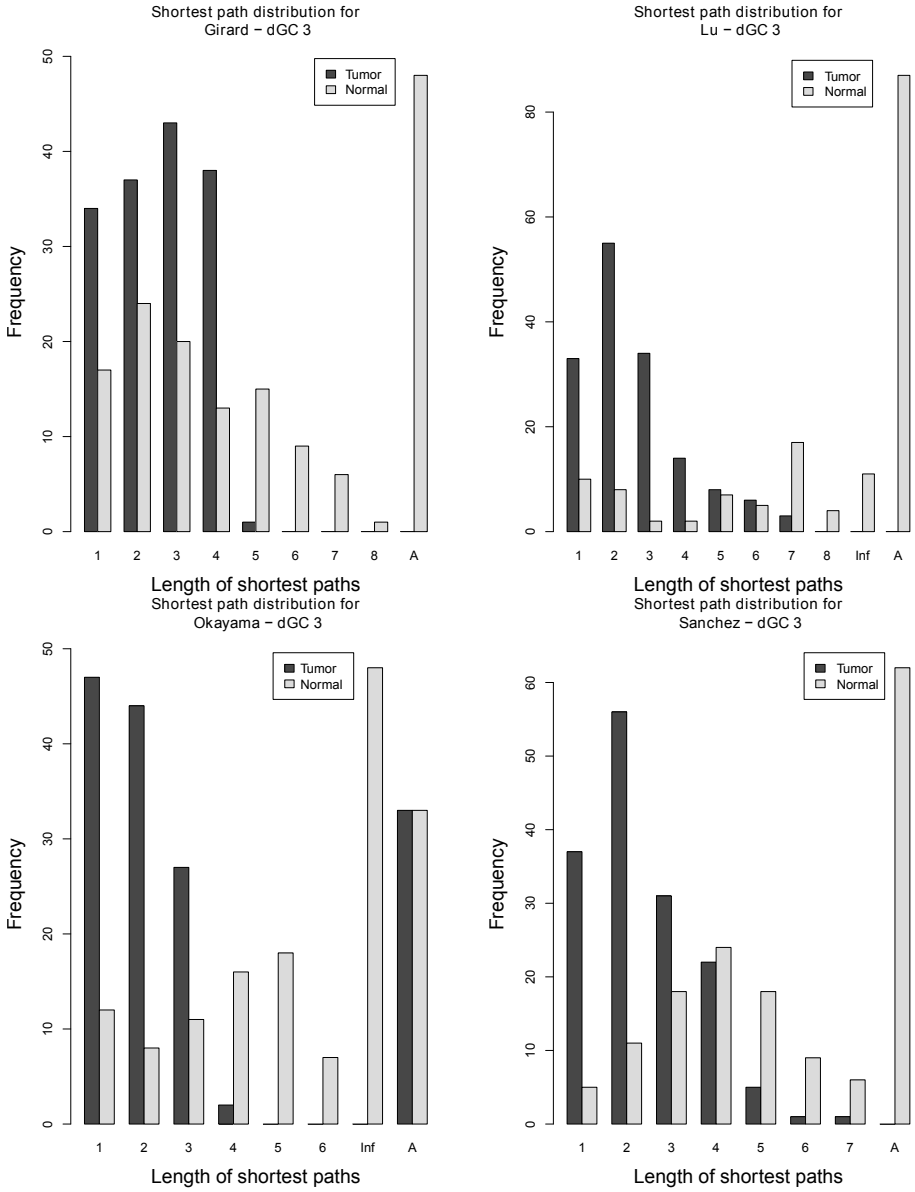


Figure 4.9: Shortest path distribution for *dGC*3 for Girard and Lu datasets are shown at the top. Shortest path distribution for *dGC*3 for Okayama and Sanchez datasets are shown in the bottom. Inf represents shortest path between unreachable nodes. A is the number of node pairs that have infinity as the distance due to the absence of nodes in the graph.

Thus, for all 7 datasets, for all 3 differential graphlet communities, we observed a trend that the shortest path lengths are shorter for tumor graphs compared to normal graphs; the median of shortest path lengths in normal is larger than that of tumor graphs, as determined using the one-sided Mann-Whitney test (adjusted p values $\leq 2.61E-13$).

4.3.1 Biological meaning of differential graphlet communities

From the shortest path distributions across all 7 datasets and all 3 differential graphlet communities, we observed a trend that the shortest path lengths are shorter for tumor graphs than for normal graphs. The observed trend suggests that tumor cells can create shortcuts between biological processes that are usually not connected under normal conditions.

In order to test whether edges in differential graphlet communities are within a pathway or across pathways, nodes in differential graphlet communities are overlapped with pathways and GO biological processes, and are presented in Tables C.1 - C.9 in Appendix C.

A proof-of-concept.

We use an example from $dGCsp_{ALL}2$ as a proof-of-concept to demonstrate that the differential graphlet community approach provides insights into the underlying mechanism, and potential novel treatments for NSCLC. Figure 4.10 presents $dGCsp_{ALL}2$ labeled with pathway information, and it shows that many edges in $dGCsp_{ALL}2$ are across different pathways suggesting crosstalk between them.

In $dGCsp_{ALL}2$, there are many edges crossing between members of the chemokine signaling pathway, Jak-STAT signaling pathway, Canonical NF-kappaB pathway and the B cell receptor signaling pathway. It has been reported that Jak-STAT signaling pathway and Canonical NF-kappaB pathway have *STAT3* and *NF-kappaB* "collaborating" in cancer [49]. The activation of *STAT3* and *NF-kappaB* as well as the interaction between

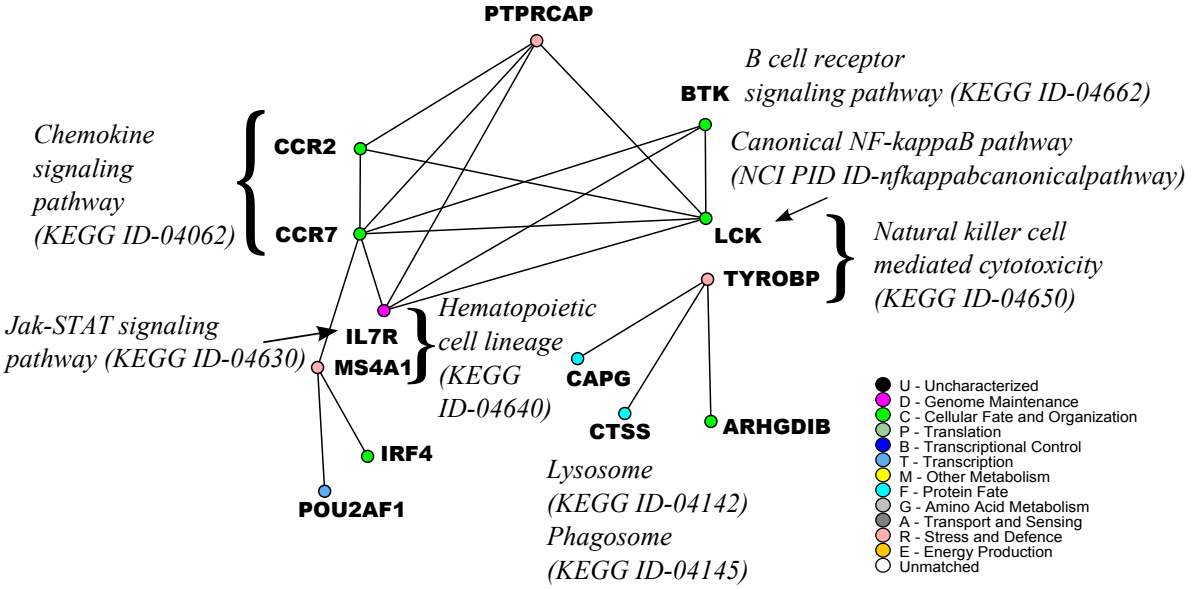


Figure 4.10: An example from $dGCsp_{ALL}2$. Edges link co-expressed genes. Nodes are colored based on GO biological function. *IL7R* belongs to the Jak-STAT signaling pathway and the hematopoietic cell lineage. *LCK* belongs to the canonical NF-kappaB pathway and the natural killer cell mediated cytotoxicity.

them are important for controlling the communication between a malignant cell and its microenvironment. Often, *STAT3* and *NF-kB* are basally active in neoplastic cells. A global profiling of mouse lung cells showed that *STAT3* controlled the expression of a large number of genes, and some *NF-kappaB* target genes were among them [31]. Genes that are controlled by *STAT3* and *NF-kappaB* include chemokines, *PAI-1*, *Bcl3*, *Bcl2*, *GADD45β* and *SOCS3*. This suggests that STAT3 and NF-kappaB pathways have to work together for the induction of specific groups of genes [49].

CCR2 and *CCR7* are chemokine receptors in the chemokine signaling pathway identified in $dGCsp_{ALL}2$. Genes that encode chemokines are among targets for *STAT3* and *NF-kappaB* [49]. Chemoattractants are crucial for recruiting and renewing various cells in the tumor microenvironment. In particular, *CCL2*, a *CCR2* ligand, controls the enrollment of myeloid cells, which induce tumor-associated macrophage (TAM) or myeloid-derived suppressor cells (MDSC) [49]. In the tumor microenvironment, TAM can promote

tumor and MDSC can suppress T cells [20]. Another chemokine receptor in $dGCsp_{ALL}2$ is *CCR7*. *CCL19*/*CCL21*/*CCR7* play a role in attracting immunosuppressive T regulatory cells [18]. Therefore, *STAT3* and *NF-kappaB*, through the regulation of chemokine synthesis, can determine which groups of immune cells are active in the tumor microenvironment.

Not only is *STAT3* observed to have crosstalk with *NF-kappaB*, STAT3 signaling also enables crosstalk among tumor and immune cells, resulting in an immunosuppressive network [119]. This crosstalk via STAT3 signaling involves hematopoietic progenitor cells, and hematopoietic cell lineage is also present in $dGCsp_{ALL}2$ (*IL7R*, *MS4A1*). Furthermore, pathways related to immune cells are also present in $dGCsp_{ALL}2$. Increase in *STAT3* activity in hematopoietic progenitor cells encourages the production of immature myeloid cells, and increases the amount of plasmacytoid dendritic cells. The amount of immature dendritic cell is also increased. Both immature dendritic cells and plasmacytoid dendritic cells encourage and accumulate regulatory T cells in the tumor microenvironment. *STAT3* activity prevents immature dendritic cells from maturing. However, mature dendritic cells are able to stimulate $CD8^+$ T cell's and natural killer cell's anti-tumor effects. *IL7R* and *MS4A1* belongs to the lymphoid stem cell branch, and the lymphoid stem cell branch is responsible for the maturing of T and B cell, as seen from the hematopoietic cell lineage in KEGG [63]. From the primary immunodeficiency pathway in KEGG, *LCK* can affect the maturing of T cell, and *BTK* can affect the maturing of B cell. Although *IL7R* and *MS4A1* are involved in the lymphoid stem cell branch, and not the myeloid stem cell, other crosstalk among tumor and immune cells is possible. Note that the plasmacytoid dendritic cells also belong to the lymphoid stem cell branch.

BTK also has edges across different pathways. *BTK* can relate to the crosstalk between *STAT3* and *NF-kappaB*, as *BTK* is crucial in the survival of B cell as well as the activation of *NF-kappaB* [101]. *BTK* can also relate to the crosstalk among tumor

and immune cells involving hematopoietic progenitor cells since *BTK* plays an important role in the maturation of B cell as mentioned above.

PTPRCAP, protein tyrosine phosphatase receptor type C-associated protein, is another vertex that has edges across different pathways. Several protein tyrosine phosphotases, PTPs, have been associated with the regulation of JAKs [102], and the JAK-STAT pathway is important for controlling immune responses [102]. Furthermore, T-cell protein tyrosine phosphatase is identified to be a crucial regulator in the signaling of immune cells [34]. *PTPRCAP* is particularly associated with CD45, an important controller of B and T lymphocyte activation [46]. In $dGCsp_{ALL}2$, edges are present between *PTPRCAP* and the chemokine receptors, as well as between *PTPRCAP* and the Jak-STAT signaling pathway.

The example from $dGCsp_{ALL}2$ highlights different crosstalk between pathways or among tumor and immune cells. There can be other crosstalk and interpretations to $dGCsp_{ALL}2$, yet this proof-of-concept demonstrates that the differential graphlet community approach provides insights to the underlying mechanism and potential treatments for NSCLC. Importantly, the differential graphlet community approach does not only return gene groups, but the edges between them as well. Systematically comparing network structure enables the identification and characterization of differences between tumor and normal samples, and enables the formalization of functional hypotheses and prioritization of biological experiments.

4.4 Conclusions

We have developed a graph-based approach that systematically characterizes network structure differences between any graphs, and used it for identifying lung cancer-specific differences between normal and tumor graphs. We proposed using differential graphlet communities for detecting deregulated subgraphs between any graphs. The differential

graphlet community approach reveals gene group and wiring differences between compared graphs – in this dissertation, between normal lung and lung cancer. Going beyond using connectivity of each gene or each edge to compare the identified deregulated subgraphs, we used shortest path distributions on differential graphlet communities in order to exploit network structure information on identified deregulated subgraphs. Importantly, the differential graphlet community approach enables a gene to participate in more than one deregulated subgraph. The ability for overlapping differential graphlet communities is important because genes can have multiple functions in different context. Interestingly, this approach identified difference in network topology between normal and tumor graphs which provides insights to the underlying molecular mechanism in NSCLC. In particular, across all 3 NSCLC datasets and all 3 identified differential graphlet communities, a trend that the shortest path lengths are shorter for tumor graphs than for normal graphs is observed; the median of shortest path lengths in normal is significantly larger compared to tumor graphs (adjusted p values $\leq 1.13E-20$; one-sided Mann-Whitney test). The results suggest that tumor cells can create shortcuts between biological processes that may not be present under normal conditions. We have further validated these results on 4 independent NSCLC datasets. As a proof-of-concept to demonstrate that the differential graphlet community approach provides insights to the underlying mechanism for NSCLC, we highlighted crosstalk between pathways and among tumor and immune cells that are revealed through the systematic graph-based analysis. Examples of crosstalk that are observed include the crosstalk between the Jak-STAT and NF-kappaB pathways or STAT3 signaling enabling crosstalk among tumor and immune cells, resulting in an immunosuppressive network. The systematic network structure comparison enables the identification of network structure differences between tumor and normal samples. Ultimately, new therapies and drug discoveries will benefit from identifying such information.

Chapter 5

A heuristic for finding graphlets that are different between normal and tumor graphs

5.1 Introduction

In Chapter 3 and Chapter 4, we have shown that comparing network structures that characterize healthy and disease state provides insights to the underlying mechanisms and treatment options for complex disease like cancer. With technological advancement, biological data will continue to grow, likely resulting in networks with millions of nodes and edges. While graphlet analysis provides valuable information in comparing networks, the subgraph isomorphism problem is NP-complete [45]; thus, we need heuristics to obtain this information efficiently. There are two groups of approaches: 1) develop approximate but efficient graphlet counting heuristics; and 2) reduce the search space by identifying relevant areas for graphlet enumeration (similar to [88]). We have discussed the computational challenges in graphlets counting in Section 2.6. In this chapter, we introduce a method for reducing search space. We develop the *differential correlation graph* approach

to identify areas that are likely to be different between the normal and tumor graph, and perform graphlet enumeration on the identified areas only. We introduce the notion of *backbone* to explain why the *differential correlation graph* approach works well.

Section 5.2 describes the *differential correlation graph* approach. Section 5.3 introduces the notion of *backbone*. Section 5.4 presents the benchmark for evaluation, Section 5.5 provides the results, and Section 5.6 gives some concluding remarks.

5.2 The differential correlation graph (DCG) approach

Differential expression studies have been developed to compare gene expression levels between healthy and affected tissues [33]. Differential expression studies usually involve detecting statistical significance changes to the mean expressions of individual genes [30]. Some studies associated changes in mean expression levels in gene groups or pathways with disease phenotypes [33]. However, useful prognostic signatures are not necessarily the most differentially expressed genes[19]. Differential co-expression approaches that compare co-expression patterns between healthy and diseased samples have developed (e.g., [65]). We use DCG to identify differences in network structures between normal and tumor graphs because 1) DCG captures the difference in gene expression correlation values between normal and tumor samples, and 2) DCG has relatively few edges among its vertices; DCG spans to different areas instead of having too many edges between the same vertices (we formalize the notion of backbone to describe this property in Section 5.3). We obtain network structure differences by using neighborhoods of depth 4 of DCGs. We consider areas of distance 4 around each node in DCGs because we use 5-node graphlets in previous chapters; thus, in order to compare our heuristic with approaches described in previous chapters, we consider neighborhoods of depth 4 of DCGs.

We construct normal and tumor co-expression graphs as previously described (see Section 3.2.5). Let N, T denote the normal and tumor graph respectively. In the DCG

method, two correlation matrices for each dataset, a normal and a tumor correlation matrix, are generated using normal and tumor samples, by calculating pairwise Pearson correlations for all gene pairs. The normal correlation matrix is denoted as MN, and the tumor correlation matrix is denoted as MT. Let MN_{ij}, MT_{ij} represent the correlation value from the i^{th} row and j^{th} column in MN and MT respectively. The differential correlation matrix MD is defined as $MD_{ij} = |MN_{ij} - MT_{ij}|$. MD_{ij} is the absolute differential correlation value between genes i and j. All gene pairs are ranked according to their absolute differential correlation values. Let DCG_x denote the DCG constructed by taking the top $x\%$ of gene pairs in MD. For each $v \in V(DCG_x)$, enumerate all 5-node graphlets involving v. Figure 2.1 shows all 5-node graphlets. The design of the heuristic is to reduce the search space by identifying relevant areas for graphlet enumeration; thus, in order to compare our heuristic with the graphlet approach (described in Chapter 3), we use 5-node graphlets as previous chapters also use 5-node graphlets. The reasons for using 5-node graphlets are described in Section 3.2.5. The DCG approach is highlighted in Figure 5.1.

5.3 Backbones

The intuition of a backbone is to have relatively few edges among its nodes. For DCG_x, as x increases, the number of edges increases. Suppose that we have DCG_x and DCG_y where $x < y$, and $e \in E(DCG_y)$, $e \notin E(DCG_x)$. We want e to be between two vertices that are not already in $V(DCG_x)$, or at least one of them is not already in $V(DCG_x)$. Given a fix number of edges, we desire DCGs to span over more vertices, instead of being dense graphs that span over few vertices.

Section 2.1 defines graph theoretic terminologies used in this chapter. The reader is reminded that a *tree* is a connected graph with no cycle, and a *forest* is a graph with

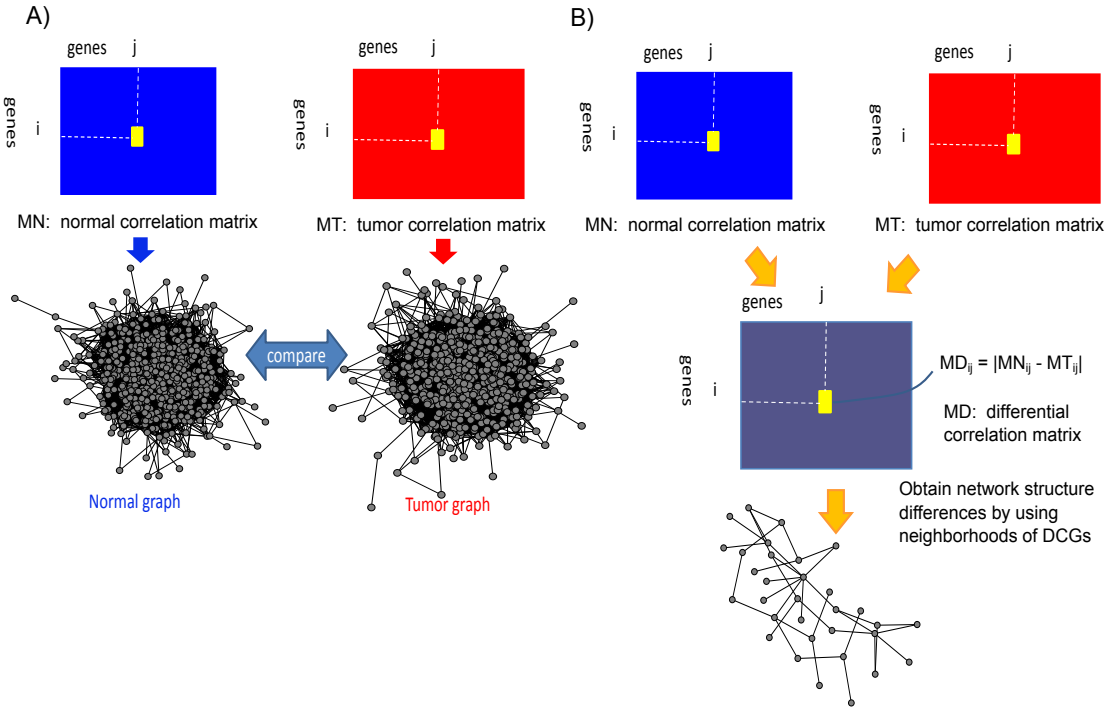


Figure 5.1: Instead of comparing between the entire normal and tumor graph (A), the *DCG* approach (B) obtains network structure differences by using neighborhoods of *DCG*s.

no cycle [113]. The maximal connected subgraphs of G are called the *components* of G [113].

A *pseudotree* is a connected graph such that $|V| = |E|$, that is, a tree with an extra edge that forms a cycle [43]. A *pseudoforest* is a graph such that each of its connected component is either a tree or a pseudotree [43]. We define a *n-node graphlet backbone* and a *n-backbone* with intuitions similar to that of a pseudotree and a pseudoforest respectively, i.e., having relatively few edges among their nodes. We define a *n-node graphlet backbone* to be a n-node graphlet with m edges, where $m \leq n + 1$. We define a *n-backbone* to be a graph such that each of its component, C, is a tree if $|C| < n$ or all n-node graphlets are n-node graphlet backbones in C if $|C| \geq n$. Note that when $|C| < n$, it is not possible to enumerate n-node graphlet in C.

5.4 Benchmark for evaluation

We applied the *DCG* approach on the same datasets, *Hou*, *Landi*, and *Su* as previous chapters (see Section 3.2.1), as we have shown that biologically meaningful deregulated subgraphs are obtained.

In Chapter 3, we separate the enumeration of 5-node graphlets from normal and tumor graphs into 3 categories: normal, tumor and both. We remind the reader that normal category contains graphlets that are only in the normal graph, and tumor category contains graphlets that are only in the tumor graph. Since we want to identify network structure differences between normal and tumor graphs, the benchmark for evaluation is focused on the normal and tumor category.

In previous chapters, we focused on graphlets that are in the tumor category, and those that have the same membership across all 3 datasets. The set of graphlets in the tumor category having the same membership across all 3 datasets is denoted as $all3$, and $|all3| = 323$. $all3$ represents deregulated graphlets that are most important biologically speaking as all three datasets captured these graphlets. Thus, we use $all3$ as our benchmark for evaluation as well.

Benchmark 1 - the normal and tumor category

Let $ApproxNC_{DCG}$, $ApproxTC_{DCG}$ be the number of graphlets obtained for the normal and tumor category using the *DCG* method respectively. Let $EnumNC$, $EnumTC$ denote the number of graphlets obtained for the normal and tumor category through enumeration respectively. Let the accuracy obtained for the *DCG* method be denoted as $AccNC_{DCG}$ for the normal category, and $AccTC_{DCG}$ for the tumor category.

$$AccNC_{DCG} = \frac{ApproxNC_{DCG}}{EnumNC} * 100$$

$$AccTC_{DCG} = \frac{ApproxTC_{DCG}}{EnumTC} * 100$$

We generated 9 DCGs using the top 0.1%, 0.2% and 0.3% of gene pairs in MD for each of the three datasets. Higher percentages are not chosen as heuristics would be meaningless if they need to process more nodes than exhaustive searches. $hou001$ denotes the top 0.1% of DCG for dataset Hou, $hou002$ denotes the top 0.2% of DCG for dataset Hou, and $hou003$ denotes the top 0.3% of DCG for dataset Hou. Similar representations are used for datasets $Landi$ and Su.

Benchmark 2 - the all3 category

Let $Approxall3_{DCG}$ be the number of graphlets obtained for $all3$ using the DCG heuristic. Let the accuracy obtained for the DCG method be denoted as $Accall3_{DCG}$.

$$Accall3_{DCG} = \frac{Approxall3_{DCG}}{|all3|} * 100$$

5.5 Results and Discussion

We present the results for the DCG approach in this section.

Table 5.1 presents the results for the DCG approach for the normal category. The result for the top 0.1% for all three datasets already performs very well; achieving greater than 87% accuracy, processing less than or equal to 57.01% of nodes. The top 0.2% returns high accuracy, with the lowest accuracy being 95.8%. Of course, the top 0.3% gives the best accuracy. Trivially, there is a trade-off between accuracy and computational demand. Note that $V(N)$, $V(T)$ and $V(DCG)$ are not the same in general; thus, it is possible for the percentage of node processed with respect to N (or T) to be greater than 100%.

DCG	$\|V(DCG)\|$	% Node wrt N	$EnumNC$	$ApproxNC_{DCG}$	$AccNC_{DCG}$
hou001	179	38.66	4182593	3666267	87.66
hou002	265	57.24	4182593	4007103	95.80
hou003	328	70.84	4182593	4133480	98.83
landi001	183	57.01	15748654	13801528	87.64
landi002	276	85.98	15748654	15546347	98.72
landi003	350	109.03	15748654	15714772	99.78
su001	186	43.97	6137000	5800794	94.52
su002	300	70.92	6137000	6106397	99.50
su003	379	89.60	6137000	6134931	99.97

Table 5.1: Results for the normal category for the DCG approach.

Table 5.2 shows the results for the DCG approach for the tumor category. Results for the top 0.1% for Hou and Su are extremely good, achieving 99.33% and 92.54% accuracy respectively, but not for $Landi$, achieving 50.69% accuracy. We investigated the reason as to why $landi001$ performs so poorly. Let $notInNT$ denote the set of nodes, $v \in V(DCG)$, $v \notin V(N)$ and $v \notin V(T)$. Nodes in $notInNT$ will not help to identify graphlets in the normal or tumor category as they are not in the normal or the tumor graph. Let inN denote the set of nodes, $v \in V(DCG)$, $v \in V(N)$ and $v \notin V(T)$. Let inT denote the set of nodes, $v \in V(DCG)$, $v \notin V(N)$ and $v \in V(T)$. Let $inNT$ denote the set of nodes, $v \in V(DCG)$, $v \in V(N)$ and $v \in V(T)$. We did a breakdown of nodes in DCG for the top 0.1% for all three datasets, refer to Table 5.3. We found that 16.39% of nodes in $landi001$ belong to $notInNT$ compared to 2.79% and 2.15% for $hou001$ and $su001$ respectively. Thus, $landi001$ does not perform well because it contains a large percentage of nodes that would not help to identify graphlets in the normal and tumor category.

DCG	$\|V(DCG)\|$	% Node wrt T	$EnumTC$	$ApproxTC_{DCG}$	$AccTC_{DCG}$
hou001	179	43.03	8577395	8520169	99.33
hou002	265	63.70	8577395	8574195	99.96
hou003	328	78.85	8577395	8576125	99.99
landi001	183	49.46	12180103	6174288	50.69
landi002	276	74.59	12180103	10164804	83.45
landi003	350	94.59	12180103	11560964	94.92
su001	186	42.96	9081990	8404783	92.54
su002	300	69.28	9081990	9056966	99.72
su003	379	87.53	9081990	9081225	99.99

Table 5.2: Results for the tumor category for the DCG approach.

DCG	# of Nodes in $notInNT$	# of Nodes in inT	# of Nodes in inN	# of Nodes in $inNT$
hou001	5	18	19	137
landi001	30	32	38	83
su001	4	32	17	133
hou002	7	26	40	192
landi002	48	52	55	121
su002	14	56	35	195
hou003	10	34	57	227
landi003	66	72	68	144
su003	23	68	54	234

Table 5.3: Node breakdown for DCGs.

From both the normal and tumor category, we see that in general, unless the application requires close to perfect accuracy, there is no need to compute DCGs with higher percentages than top 0.1% since the top 0.1% of DCGs already perform so well. The DCG approach works very well even at the top 0.1% if the DCG does not contain too many nodes in $notInNT$. Thus, one can perform a quick check on the number of nodes in $notInNT$ to determine if accurate results can be achieved.

The $all3$ category contains the most important graphlets that differed between the normal and tumor graph because all three datasets picked up these graphlets. Thus, even if the heuristic cannot achieve high accuracy in the normal or tumor category, it is important for the heuristic to obtain graphlets in the $all3$ category. The DCG approach performs extremely well, refer to Table 5.4. For all three datasets at the top 0.1%, the

performance for the *all*3 category is 100%, 99.07% and 95.98%. Importantly, although *landi*001 did not perform well in the tumor category, *landi*001 performs extremely well in this key category.

DCG	$\|V(DCG)\|$	% Node wrt N	% Node wrt T	$Approxall3_{DCG}$	$Accall3_{DCG}$
hou001	179	38.66	43.03	323	100
hou002	265	57.24	63.70	323	100
hou003	328	70.84	78.85	323	100
landi001	183	57.01	49.46	320	99.07
landi002	276	85.98	74.59	322	99.69
landi003	350	109.03	94.59	322	99.69
su001	186	43.97	42.96	310	95.98
su002	300	70.92	69.28	323	100.00
su003	379	89.60	87.53	323	100.00

Table 5.4: Results for the all3 category for the DCG approach.

Since the top 0.1% DCGs already perform so well, we do not need to go to the top 0.2% or 0.3% DCGs. Why do the top 0.1% DCGs perform so well? There are two reasons. First, differential co-expression values capture co-expression values that differed between the normal and tumor condition. The co-expression is calculated from gene expressions, which provide much information about the normal and tumor states. Second, the top 0.1% DCGs for Hou, Su and $Landi$ are 5-node backbones. 5-node backbones allow edges in DCGs to span to different areas instead of having too many edges between the same vertices. Since in this chapter as well as in previous chapters, 5-node graphlets are used; thus, using 5-node backbones is a natural choice. 5-node graphlets are used for the aforementioned reasons, and we will not repeat them here. By enumerating graphlets on nodes in DCGs, the neighborhoods of depth 4 of DCGs are used for obtaining graphlets in the normal, tumor and *all*3 category.

Tables 5.5 and 5.8 show that the top 0.1% DCG for Hou is a 5-node backbone. All components in the top 0.1% DCG for Hou that have less than 5 nodes, they are trees (having $|C| - 1$ edges, where C denotes a component). All components in the top 0.1% DCG for Hou that have greater than or equal to 5 nodes, all 5-node graphlets in them

are 5-node graphlet backbones. Similarly, Tables 5.6 and 5.8 show that the top 0.1% *DCG* for *Landi* is a 5-node backbone; Tables 5.7 and 5.8 show that the top 0.1% *DCG* for Su is a 5-node backbone.

Component	Nodes	Edges
1	2	1
2	2	1
3	2	1
4	2	1
5	2	1
6	2	1
7	2	1
8	2	1
9	2	1
10	2	1
11	2	1
12	3	2
13	3	2
14	3	2
15	4	3
16	5	4
17	139	172

Table 5.5: All components in hou001. Nodes are the number of nodes in the component, and edges are the number of edges in the component.

Table 5.8 also compares the 5-node graphlet distributions between Ns, Ts and the top 0.1% *DCG*s for Hou, Su and *Landi*. There is no graphlet for graphlet numbers that are greater than or equal to 13 for all of the three top 0.1% *DCG*s, which is not the case for Ns and Ts. Graphlet numbers increase monotonically with the number of edges between the 5 vertices. Note that the computation for graphlet distributions are not needed for the *DCG* approach; graphlet distributions are only used to illustrate that the top 0.1% *DCG*s for Hou, Su and *Landi* are backbones. Note also that we are not claiming that all *DCG*s that are n-node backbones will perform well as there are other factors to be considered such as the aforementioned $|notInNT|$. We are using 5-node backbone to explain why the top 0.1% *DCG*s for Hou, Su and *Landi* perform so well.

Component	Nodes	Edges
1	2	1
2	2	1
3	2	1
4	2	1
5	2	1
6	2	1
7	2	1
8	2	1
9	2	1
10	2	1
11	2	1
12	3	2
13	3	2
14	3	2
15	3	2
16	5	4
17	6	5
18	138	148

Table 5.6: All components in landi001. Nodes are the number of nodes in the component, and edges are the number of edges in the component.

5.6 Concluding remarks

We described a heuristic, the DCG approach, that performs well in obtaining graphlets that differed between normal and tumor graphs by identifying relevant areas for graphlet enumeration. From the three NSCLC datasets, we showed that if only a low percentage of vertices in DCGs are in $notInNT$, it is sufficient to achieve accurate estimation in the difference between normal and tumor states using only the top 0.1% DCGs. For example, we obtained a 99.33% accuracy from the top 0.1% DCG for Hou in the tumor category. Furthermore, we showed that the top 0.1% DCGs are able to achieve excellent accuracy in the $all3$ category, achieving accuracies as high as 100%. Recall that the $all3$ category represents deregulated graphlets that are most important biologically speaking as all three datasets captured these graphlets.

Component	Nodes	Edges
1	2	1
2	2	1
3	2	1
4	2	1
5	2	1
6	2	1
7	2	1
8	2	1
9	2	1
10	2	1
11	2	1
12	2	1
13	2	1
14	2	1
15	3	2
16	3	2
17	3	2
18	3	2
19	4	3
20	4	3
21	4	3
22	4	3
23	18	18
24	112	124

Table 5.7: All components in su001. Nodes are the number of nodes in the component, and edges are the number of edges in the component.

In order to explain why the top 0.1% *DCG*s for *Hou*, *Su* and *Landi* perform so well, we introduce the notion of backbone. Intuitively, a backbone is a graph that has relatively few edges among its vertices; allowing it to span to different areas instead of having many edges between the same vertices.

While the *DCG* approach is generic, we applied it to three NSCLC datasets. A future work is to evaluate how well the *DCG* approach works on other cancer and other disease datasets. In this chapter, the benchmark of evaluation is on the normal and tumor graphs that are used throughout the dissertation. Another future work is to evaluate the approach on different sizes of graphs.

No.	Hou_N	Hou_T	hou 001	$Landi_N$	$Landi_T$	$landi$ 001	Su_N	Su_T	su 001
1	1106313	1315759	2165	2212733	1580462	1095	1268630	1180453	855
2	1325614	1735266	3928	3236163	2684372	1649	1567616	1951757	1589
3	149038	228778	1224	539277	534023	506	204360	332008	329
4	427382	970202	0	2066653	1583524	0	716688	1160982	0
5	427986	947766	0	1735082	1220161	0	681149	930447	0
6	251277	619477	0	1388068	1217524	0	452958	812807	0
7	5416	7186	2	30462	15755	1	9078	10618	0
8	61536	70682	262	304551	193711	127	94913	130146	158
9	137608	595457	0	1288541	1051188	0	332769	702967	0
10	35637	118965	0	301030	213192	0	81666	162023	0
11	131491	541341	0	1073163	782939	0	297820	593723	0
12	609	529	23	6796	3318	4	1201	1754	5
13	17890	30024	0	157245	90810	0	36382	62438	0
14	8038	87056	0	119890	107858	0	26554	69210	0
15	43732	527789	0	658272	551003	0	157465	376583	0
16	35088	199449	0	514164	368467	0	107944	264932	0
17	2991	7048	0	42660	23841	0	8090	15429	0
18	17448	324251	0	408137	332110	0	80248	228167	0
19	1929	8437	0	42027	26701	0	7011	17684	0
20	4016	156303	0	145409	121768	0	26735	82351	0
21	718	94794	0	40343	39388	0	7506	25294	0

Table 5.8: 5-node graphlet distributions. Refer to Figure 2.1 for all 5-node graphlets. No. refers to graphlet numbers.

Chapter 6

Conclusions and future work

6.1 Conclusions

We proposed novel approaches to compare graphs, and to extract network structure differences between them. We focused on comparing normal and disease graphs in this dissertation, but the algorithms can be applied generally. Base on the extracted network structure differences, we analyzed and designed methods in order to gain insights to the underlying mechanisms and treatments for diseases.

In Chapter 3, we demonstrated how graphlets facilitate network structure comparison, and in turn how this information can be used to predict novel treatment options. We proposed a systems approach with an aim to revert disease conditions to healthy ones through treatments. In order to achieve the objective, we proposed three novel methods to 1) systematically identify network structure differences between normal and tumor graphs, 2) identify and prioritize drug combinations based on extracted network structure differences, and 3) computationally estimate the potential of the proposed drug combination to "repair" deregulated subgraphs, making disease graphs more similar to normal graphs. Validations of drug combination predictions, both mechanistically and functionally are performed. Results have shown that our systems approach is a promis-

ing method to provide treatment options to NSCLC through the rewiring of disease networks, i.e., making the disease graph more similar to the normal graph through drug combination treatments.

In Chapter 4, we introduced the notion of differential graphlet community to detect deregulated subgraphs between any graphs such that the network structure information is exploited. The differential graphlet community approach systematically captures network structure differences between any graphs. This approach circumvents the exponential growth of computation required as the deregulated subgraph size increases, and enables the systematically exploring of protein communities with larger size, which provide stronger biological context. Importantly, this approach has the ability to include a gene into more than one deregulated subgraph. The differential graphlet community approach led to exciting results, providing insights to the underlying molecular mechanism in NSCLC. In particular, across all three NSCLC datasets, we observed a trend that the shortest path lengths are shorter for tumor graphs than for normal graphs between genes that are in differential graphlet communities, suggesting that cancer creates shortcuts between biological processes that may not be present in normal conditions. Importantly, we have validated these intriguing results on four independent datasets. Examples of shortcuts that are observed, and are in agreement with known mechanism in literature include the crosstalk between the Jak-STAT and NF-kappaB pathways or STAT3 signaling enabling crosstalk among tumor and immune cells, resulting in an immunosuppressive network.

Comparing network structures between graphs is useful, but large graph comparison is computationally intensive. However, not all areas of the graphs are needed to perform comparisons. In Chapter 5, we proposed a heuristic, the differential correlation graph approach, that identifies areas that are different between the normal and tumor graph, and perform graphlet enumeration on the identified areas. Results showed that our approach achieves accurate estimation in the difference between normal and tumor states

by performing network comparisons in important areas only. We also introduce the notion of network backbone to explain why the differential correlation graph approach works well.

Our study increases the current exploitation of network structures in the comparison between networks. We introduce two novel graphlet-based methods, and an efficient heuristic for exploiting network structure information in the comparison between any graphs, and we validate them on comparing graphs generated from NSCLC datasets. Going beyond identifying network structure differences, our approaches resulted in insights to the underlying molecular mechanism in NSCLC, as well as treatment options to NSCLC through the rewiring of disease networks.

We have demonstrated that the potential is enormous in going from comparative network analysis to the understanding of the underlying mechanisms of disease to treatment options. Although we have proposed several novel methods to discover these immense treasures, we are still far from fully understanding them.

6.2 Future work

We propose some future work that are closely related to our dissertation.

6.2.1 Pseudo dominating set of differential correlation graph (PDS)

In Chapter 5, we used the differential correlation graph to reduce search space by identifying relevant areas for graphlet enumeration. We can use dominating sets of differential correlation graphs to further reduce search space.

A dominating set of G is defined to be a subset S, $S \subseteq V(G)$ such that $\forall v \in V(G)$, either $v \in S$ or $v \in N_n(s), s \in S$ [52]. The domination number for G, $\gamma(G)$, is the number of vertices that is present in the smallest dominating set for G. Given a graph G and an

input k, the dominating set problem is to determine if $\gamma(G) \leq k$ is true. The dominating set problem is a NP-complete decision problem [45]. Thus, many heuristics have been developed (e.g., [69]).

The notion of dominating sets is used in the field of designing routing protocols for wireless networks, e.g., [103, 117]. Dominating sets are used to locate central nodes in wireless networks for efficient routing, e.g., [103, 117]. The dominating set in a wireless network can be viewed as a skeleton of the network where data can be efficiently routed through it because each node in the network is at most one hop away from the skeleton. We are inspired to apply the notion of dominating sets to identify skeletons to correlation difference graphs where graphlets that are differed between normal and tumor graphs can be identified. In the biological network context, Milenković *et al.* suggested that biologically essential proteins can be obtained through dominating sets of PPI networks [79], and Molnár Jr. *et al.* studied how the size of minimum dominating set in scale-free networks scales [62].

Algorithm-PDS

The algorithm aims to identify a skeleton for the given differential correlation graph. The objective of the algorithm is to design a heuristic to identify as many deregulated subgraphs as possible without an exhaustive search on the normal and tumor graph. The objective is not to design a heuristic to obtain a minimum dominating set; rather, it is designed to identify as many deregulated subgraphs as possible.

Let DCG, N, T denote the differential correlation graph, the normal graph, and the tumor graph respectively. The inputs to the algorithm are DCG, $V(N)$ and $V(T)$. The output to the algorithm is the pseudo dominating set for DCG. The idea of the algorithm is as follow. At any time, every node in $V(DCG)$ belongs to 1 of 3 sets: S denotes the set of nodes that are in the pseudo dominating set; $Grey$ denotes the set of nodes that are either the neighbors of nodes in S or nodes that would not be considered as candidates

for S; $White$ denotes the set of nodes that have not been processed yet. The algorithm halts when there is no more nodes in $White$. $corr(a, b)$ denotes the absolute correlation value for $\{a, b\}$, $a, b \in V(DCG)$. We define the correlation weight, $scorr(v)$, as:

$$scorr(v) = \begin{cases} 1 + \sum_{(v,i) \in E(DCG)} corr(v, i), i \in N_n(v) \wedge i \in White & \text{if } v \in White \\ 0 & \text{if } v \in S \, or \, v \in Grey. \end{cases}$$

In each step, we select the node with the highest correlation weight to S.

In order to enhance the identification of deregulated subgraphs, the heuristics has a pre-process step. Let $notInNT$ denote the set of nodes, $v \in V(DCG)$, $v \notin V(N)$ and $v \notin V(T)$. All nodes in $notInNT$ are put into $Grey$ because these nodes will not help to identify deregulated subgraphs if they are not in the normal or the tumor graph. Let inN denote the set of nodes, $v \in V(DCG)$, $v \in V(N)$ and $v \notin V(T)$. Let inT denote the set of nodes, $v \in V(DCG)$, $v \notin V(N)$ and $v \in V(T)$. All nodes in inN and all nodes in inT are put into S. Nodes that are in the normal graph only, and nodes that are in the tumor graph only will help to identify graphlets that are different between the two graphs. The neighbors of nodes in S are put into $Grey$. The algorithm for PDS is shown in Algorithm 4.

The reason for the algorithm to return a pseudo dominating set for DCG, and not a dominating set is because of the pre-process step in the heuristic. When all nodes in $notInNT$ are put into $Grey$, the output of the heuristic may not satisfy the definition of a dominating set, which is fine for our purposes. Putting nodes in $notInNT$ to $Grey$ informs us that those are not important areas to look for deregulated subgraphs.

6.2.2 Size of graphlets

In Chapter 3, we proposed the graphlet approach to identify network structure differences between any graphs. We used 5-node graphlets in the graphlet approach for reasons

Input: DCG, $V(N)$, $V(T)$
Output: Pseudo dominating set for DCG
// Initialization
$S \leftarrow \emptyset$;
$Grey \leftarrow \emptyset$;
$White \leftarrow V(DCG)$;
// Pre-process step
$Grey \leftarrow Grey \cup \{v\}, \forall v \in notInNT$;
$S \leftarrow S \cup \{v\}, \forall v \in inN$;
$S \leftarrow S \cup \{v\}, \forall v \in inT$;
$Grey \leftarrow Grey \cup \{v\}, \forall v \in N_n(s), s \in S$;
$White \leftarrow White \setminus \{v\}, \forall v \in S$;
$White \leftarrow White \setminus \{v\}, \forall v \in Grey$;
// Main step
while $White \neq \emptyset$ **do**
  choose $m \in \{v | scorr(v) = max_{u \in V(DCG)}\{scorr(u)\}\}$; // m will be from *White* as otherwise, the correlation weight is 0
  $S \leftarrow S \cup \{m\}$;
  $Grey \leftarrow Grey \cup \{v\}, \forall v \in N_n(m)$;
  $White \leftarrow White \setminus \{m\}$;
  $White \leftarrow White \setminus \{v\}, \forall v \in N_n(m)$;
end

Algorithm 4: Algorithm for the pseudo dominating set of differential correlation graph.

described in Section 3.2.5. The graphlet approach using 5-node graphlets obtained biological meaningful network structure differences (Section 3.3.1). Nevertheless, one can examine network structure differences from the graphlet approach using graphlets of different sizes.

In Chapter 4, we proposed the differential graphlet community approach. Recall that a differential graphlet community is formed by n-node graphlets. The approach circumvents the exponential growth of computation required as the graphlet size increases, and enables the systematically exploring of protein communities with larger size which provide stronger biological context. We used 5-node graphlets in the differential graphlet community approach for reasons described in Section 4.2.1. What about if different n's are used for the n-node graphlets? How different would the differential graphlet communities be? The reader is reminded that a differential graphlet community is defined as the union of all n-graphlets such that one can reach to another by a chain of adjacent n-graphlets. Adjacent n-graphlets are graphlets that share $n-1$ nodes. Trivially, it would not be very meaningful if $n \leq 3$. 2-node graphlets are edges, and 3-node graphlets will likely form large differential graphlet communities as well because of the differential graphlet community definition. On the other hand, it would also not be meaningful if n is too large. One of the main advantage of the differential graphlet community approach is that it circumvents the exponential growth of computation required as the graphlet size increases and still results in protein communities with larger size which provide stronger biological context. Thus, having large n's would defeat the design of our approach. Nevertheless, one can investigate the network properties on differential graphlet communities when different n's are used for the n-node graphlets.

6.2.3 Applications to other networks

In Chapter 1, we discussed that comparing networks with different conditions is extremely useful, for example, comparing networks with different stages or subtypes in cancer,

comparing networks with different drug treatments, comparing networks with disease development in different time points. We focus on comparing normal and NSCLC graphs in this dissertation, but the algorithms can be applied generally. Applying our algorithms on the comparisons between normal and other diseases graphs, or other aforementioned different conditions would be interesting. Furthermore, besides biological networks, many real-world phenomena are modeled with large networks, for example, social, technological and information networks. Applying our algorithms in these networks can also be of interest. Of course, tuning of the algorithms will likely be needed to tailor for any specific fields or conditions.

6.2.4 Applications of other biological techniques

With technological advancement, new biological techniques continue to arise. As data from new techniques becomes available, we can apply our approaches on them. For example, we used microarray datasets for gene expressions in the dissertation; we can apply our approaches on RNA-Seq (RNA Sequencing) data.

In addition to applying data from new techniques, there are other existing techniques whereby we can validate our drug combinations. We used cell lines in the dissertation; we can validate our drug combinations on xenograft models.

Appendices

Appendix A

Prognostic gene signatures

A.1 Prognostic signatures

Eighteen prognostic NSCLC signatures were used [44, 12, 14, 114, 17, 37, 106, 26, 50, 75, 87, 97, 15, 72, 71, 73, 105] throughout the dissertation. Refer to Table A.1 for a list of genes that were used.

Genes (Entrez gene ID)						
3964	3156	7385	396	5754	2800	3276
10857	2517	1066	397	638	284611	26469
6512	1476	6188	3267	4356	163486	7443
7379	5621	3579	6773	10320	79627	51442
5193	2064	324	6781	9412	9473	79650
5191	383	1909	1315	8899	51104	116138
7480	3623	498	5428	545	54665	83604
11113	5054	4504	10321	29896	8532	58495
6399	3383	1520	641	6642	144193	11055
1106	6181	5725	1822	7162	23231	11235
9536	6175	4485	8028	864	23347	80303
1787	6741	1271	4673	4666	25836	79083
8573	3249	4140	1006	5168	256949	79053
2069	3643	1803	3169	2966	256227	55006
23037	3932	4134	5316	5150	728340	29775
375449	5836	5295	9520	8621	730394	8284
753	7054	5122	637	3603	79818	60468
11079	7056	2048	10718	1298	9652	8539
8473	1514	25802	54020	3184	23253	84084

Continued on next page

Table A.1 – *Continued from previous page*

Genes (Entrez gene ID)						
10622	3912	1381	6281	23650	133121	57556
27324	4313	5037	10096	9898	129285	1798
1317	3145	894	746	8581	80108	92552
10257	4256	7465	6138	1974	23389	55204
10057	5788	624	390	2535	7846	79101
3638	308	4836	7114	2886	2011	80157
9422	3868	1992	8766	10406	1355	54462
27429	5452	6541	51646	8404	54808	79762
5889	8273	11151	59	6596	159090	54414
9452	307	6241	6234	4175	90233	57161
10488	2736	7072	2782	10795	79694	64816
7305	3576	3301	6434	8243	114784	9200
23036	6036	6282	100291837	9790	10444	57665
4976	2548	887	6227	23180	11011	28316
26025	369	1236	7329	5790	57862	57110
8828	596	1880	6047	841	54919	22931
8029	251	6546	8480	5349	84440	56344
10492	1191	3445	10265	1108	84864	55734
7975	6513	4363	3766	10948	25901	56943
1500	5834	5217	6868	22949	94134	55505
4068	6440	5245	80742	4784	10523	55900
8996	1559	133	3430	27332	10346	56948
1838	931	6291	23635	3837	221154	11179
4037	4953	4488	3020	4978	91526	3890
261734	650	2201	3021	9975	196403	55732
3161	2547	3667	7436	23395	56983	54014
11156	5916	7078	3339	5031	8495	51761
4094	3852	157	57198	5050	27127	10137
23451	634	1119	4693	5339	23	55752
4246	966	403	7073	3662	25983	54498
8676	3082	5268	1869	5094	55578	54952
4308	3660	3080	271	1266	84548	55863
9252	4811	6337	1671	5794	80321	23163
10434	7184	1650	8140	6491	84236	11163
5495	6628	822	5214	2220	2595	440270
11169	1359	7066	5460	23531	51099	55888
5064	8061	10124	5342	8717	6574	4139
7743	3248	1811	5343	1740	23522	8911
10263	3543	846	3398	6307	5279	55691
7204	1832	8744	1824	6453	9896	100130086
10845	4582	729230	2692	1874	9806	51402
23683	5618	6772	9590	4916	1794	6683

Continued on next page

Table A.1 – *Continued from previous page*

Genes (Entrez gene ID)						
9860	1363	2475	2122	3131	1902	51454
22943	3575	4110	5970	1432	2125	6400
23052	7554	5780	51035	814	843	7750
22872	5578	3802	3083	6804	9138	11012
11130	2697	7080	6742	5450	5976	10157
11328	5566	546	4486	2981	11184	1616
25796	2596	6157	3382	3091	535	26083
10527	7494	6387	695	1847	113178	51566
19	655	5993	4208	4225	114907	26249
8876	378	372	3823	11099	57655	11232
11062	6651	6511	7033	1848	89970	8697
27107	522	5294	1591	152485	92856	51155
9453	2780	223	55697	348654	30001	23305
24139	2057	6164	92070	80315	79731	267
10648	2921	7203	2589	64285	83737	7224
86	5870	23438	1628	339318	26018	57476
1528	4437	837	2529	388969	84502	51440
216	1644	56252	547	80206	89782	4854
4860	825	10965	3778	3077	57554	23228
4609	3589	10972	7786	11259	92304	23265
1950	2525	7424	7866	605	9971	11238
1442	5764	5663	3609	132332	84955	117178
1443	4763	5498	7074	9540	64421	22890
796	2357	429	64714	4957	2304	23589
3512	2263	8087	7710	84075	54738	10809
2243	2065	9364	1846	84986	1316	11187
2244	6192	2719	10949	8852	4258	23151
714	7317	6356	5058	2778	990	1837
5196	6699	3899	7965	79605	398	51400
2512	5567	113	4250	10807	8904	56105
29923	25825	10564	27445			

Table A.1: Genes in the 18 prognostic signatures (Entrez gene ID)

Appendix B

Drug validation

B.1 Drug concentration

Effective concentrations for Bexarotene, Epicatechin, Erlotinib, Gemcitabine and Mifepristone were evaluated using a serial dilution curve of 5 points. Cells were treated with the half maximal inhibitory concentrations (IC50), refer to Appendix Table B.1.

Drug	Concentration used
Bexarotene	4uM
Epicatechin	100uM
Erlotinib	2uM
Gemcitabine	30uM
Mifepristone	10uM
Bexarotene + Erlotinib	4uM + 2uM
Bexarotene + Erlotinib + Epicatechin	4uM + 2uM + 100uM
Bexarotene + Erlotinib + Mifepristone	4uM + 2uM + 10uM
Mifepristone + Gemcitabine	10uM + 30uM

Table B.1: Drug concentrations that were used.

B.2 Drug validation results for the impact on the deregulated subgraph

We present the results for the impact on the deregulated subgraph for Chapter 3. For all $v \in Sg$, for all treatment, for all cell lines, fold change information is presented in Appendix Tables B.2 - B.4.

A549: Mifepristone+Gemcitabine									
name	CCR2-fc	BTK-fc	IRF4-fc	PLCL2-fc	IL16-fc	MS4A1-fc	IL7R-fc	ITM2A-fc	LCK-fc
Mife	47.41	15.84	4.2	10.77	10.51	0.7	12.55	2.83	5.34
Gem	87.56	9.87	13.09	27.78	10.08	17.28	41.07	6.48	19.7
Mife+Gem	7.51	1.47	2.25	23.82	9.26	0.12	11.7	4.03	2.02
H1975: Mifepristone+Gemcitabine									
name	CCR2-fc	BTK-fc	IRF4-fc	PLCL2-fc	IL16-fc	MS4A1-fc	IL7R-fc	ITM2A-fc	LCK-fc
Mife	3.3	1.06	13.2	15.02	1.61	32.05	0.13	36.84	1.16
Gem	19.94	5.42	0.07	26.02	4.81	18.85	0.71	148.79	7.64
Mife+Gem	62.12	38.4	11.82	408.31	9.48	11.51	0.64	1237.98	12.53
H460: Mifepristone+Gemcitabine									
name	CCR2-fc	BTK-fc	IRF4-fc	PLCL2-fc	IL16-fc	MS4A1-fc	IL7R-fc	ITM2A-fc	LCK-fc
Mife	1.42	1.78	2.24	0.18	2.84	2.98	2.79	0.61	0.92
Gem	32.6	13.57	42.49	7.84	37.13	36.49	18.64	49.02	35.51
Mife+Gem	6.37	2.15	15.5	2.51	6.75	0.77	5.68	10.99	5.56

Table B.2: The fold changes (fc) for $v \in Sg$ after A549, H1975 and H460 cells are treated with Mifepristone, Gemcitabine, and Mifepristone+Gemcitabine. Mife is Mifepristone and Gem is Gemcitabine.

A549: Epicatechin+Bexarotene+Erlotinib									
name	CCR2-fc	BTK-fc	IRF4-fc	PLCL2-fc	IL16-fc	MS4A1-fc	IL7R-fc	ITM2A-fc	LCK-fc
Epi	20.47	10.87	2.58	7.36	10.18	0.23	14.93	3.62	5.27
B	1.35	0.74	1.93	2.14	1.27	0.18	1.55	1.05	0.58
Erlo	3.04	0.81	5.09	3.92	1	0.02	0.56	0.9	0.76
B+Erlo	1.87	0.33	1.79	2.29	0.9	0.1	0.23	0.72	0.25
Epi+B+Erlo	16.01	3.07	5.64	5.09	0.14	0.52	10.8	3.18	6.35
H1975: Epicatechin+Bexarotene+Erlotinib									
name	CCR2-fc	BTK-fc	IRF4-fc	PLCL2-fc	IL16-fc	MS4A1-fc	IL7R-fc	ITM2A-fc	LCK-fc
Epi	1.24	0.74	9.55	9.37	0.76	1.12	3.08	15.87	0.79
B	0.65	0.22	5.51	5.96	0.51	0.05	0.1	23.13	0.31
Erlo	7.78	1.32	16.8	6.99	0.72	0.2	0.07	9.76	2.31
B+Erlo	1.87	0.71	91.08	9.59	0.54	1.29	0.6	7.3	0.16
Epi+B+Erlo	2.1	1.82	52.6	9.97	1.55	0.95	0.43	9.54	0.62
H460: Epicatechin+Bexarotene+Erlotinib									
name	CCR2-fc	BTK-fc	IRF4-fc	PLCL2-fc	IL16-fc	MS4A1-fc	IL7R-fc	ITM2A-fc	LCK-fc
Epi	0.56	0.18	0	0.17	0.97	0.35	1.07	0.44	2.03
B	12	1.61	1.24	1.09	2.82	28.08	1.43	1.08	1.47
Erlo	1.87	1	0.37	0.23	0.77	14.94	1.31	2.25	1.64
B+Erlo	3.96	1.89	0.42	0.42	4.78	0.14	6.79	1.58	1.86
Epi+B+Erlo	0.3	0.28	0.06	0.19	0.82	1.61	0.07	0.49	2.16

Table B.3: The fold changes (fc) for $v \in Sg$ after A549, H1975 and H460 cells are treated with Epicatechin, Bexarotene, Erlotinib, Bexarotene + Erlotinib, and Epicatechin + Bexarotene + Erlotinib. B is Bexarotene, Erlo is Erlotinib, Epi is Epicatechin.

A549: Mifepristone+Bexarotene+Erlotinib									
name	CCR2-fc	BTK-fc	IRF4-fc	PLCL2-fc	IL16-fc	MS4A1-fc	IL7R-fc	ITM2A-fc	LCK-fc
B	1.35	0.74	1.93	2.14	1.27	0.18	1.55	1.05	0.58
Erlo	3.04	0.81	5.09	3.92	1	0.02	0.56	0.9	0.76
B+Erlo	1.87	0.33	1.79	2.29	0.9	0.1	0.23	0.72	0.25
Mife	47.41	15.84	4.2	10.77	10.51	0.7	12.55	2.83	5.34
Mife+B+Erlo	24.72	2.82	1.79	4.9	2.22	0.22	1.78	0.47	1.79
H1975: Mifepristone+Bexarotene+Erlotinib									
name	CCR2-fc	BTK-fc	IRF4-fc	PLCL2-fc	IL16-fc	MS4A1-fc	IL7R-fc	ITM2A-fc	LCK-fc
B	0.65	0.22	5.51	5.96	0.51	0.05	0.1	23.13	0.31
Erlo	7.78	1.32	16.8	6.99	0.72	0.2	0.07	9.76	2.31
B+Erlo	1.87	0.71	91.08	9.59	0.54	1.29	0.6	7.3	0.16
Mife	3.3	1.06	13.2	15.02	1.61	32.05	0.13	36.84	1.16
Mife+B+Erlo	1.29	1.33	20.68	58.8	0.98	0.13	0.51	12.02	0.96
H460: Mifepristone+Bexarotene+Erlotinib									
name	CCR2-fc	BTK-fc	IRF4-fc	PLCL2-fc	IL16-fc	MS4A1-fc	IL7R-fc	ITM2A-fc	LCK-fc
B	12	1.61	1.24	1.09	2.82	28.08	1.43	1.08	1.47
Erlo	1.87	1	0.37	0.23	0.77	14.94	1.31	2.25	1.64
B+Erlo	3.96	1.89	0.42	0.42	4.78	0.14	6.79	1.58	1.86
Mife	1.42	1.78	2.24	0.18	2.84	2.98	2.79	0.61	0.92
Mife+B+Erlo	0.69	0.28	0.03	0.26	1.12	0.9	0.18	0.24	0.46

Table B.4: The fold changes (fc) for $v \in Sg$ after A549, H1975 and H460 cells are treated with Mifepristone, Bexarotene, Erlotinib, Bexarotene + Erlotinib and Mifepristone + Bexarotene + Erlotinib. B is Bexarotene, Erlo is Erlotinib, and Mife is Mifepristone.

Appendix C

Pathway and GO information

Tables containing information regarding the overlapping of genes in differential graphlet communities with pathways and GO biological processes (used in Chapter 4).

GO ID	GO Term	Gene in biological process (Entrez gene ID)
GO:0006810	transport	3766
GO:0006811	ion transport	3766
GO:0006813	potassium ion transport	3766
GO:0006355	regulation of transcription, DNA-dependent	5450,3662
GO:0006351	transcription, DNA-dependent	5450,3662
GO:0007165	signal transduction	729230,1236,23228,3575
GO:0019221	cytokine-mediated signaling pathway	3662,729230,3575
GO:0002606	positive regulation of dendritic cell antigen processing and presentation	1236
GO:0002885	positive regulation of hypersensitivity	1236
GO:0002922	positive regulation of humoral immune response	1236
GO:0006935	chemotaxis	729230,1236
GO:0006955	immune response	729230,1236,931,3575
GO:0007186	G-protein coupled receptor signaling pathway	729230,1236
GO:0032496	response to lipopolysaccharide	1236
GO:0032649	regulation of interferon-gamma production	1236
GO:0032735	positive regulation of interleukin-12 production	1236
GO:0045060	negative thymic T cell selection	1236

Continued on next page

Table C.1 – *Continued from previous page*

GO ID	GO Term	Gene in biological process (Entrez gene ID)
GO:0050706	regulation of interleukin-1 beta secretion	1236
GO:0050862	positive regulation of T cell receptor signaling pathway	3932,1236
GO:0070098	chemokine-mediated signaling pathway	729230,1236
GO:0072610	interleukin-12 secretion	1236
GO:0090023	positive regulation of neutrophil chemotaxis	1236
GO:0097029	mature dendritic cell differentiation	1236
GO:2000510	positive regulation of dendritic cell chemotaxis	1236
GO:2000522	positive regulation of immunological synapse formation	1236
GO:2000525	positive regulation of T cell costimulation	1236
GO:2000526	positive regulation of glycoprotein biosynthetic process involved in immunological synapse formation	1236
GO:0006954	inflammatory response	729230,3766
GO:0006468	protein phosphorylation	695,11184,3932
GO:0050729	positive regulation of inflammatory response	729230
GO:0016310	phosphorylation	695,11184,3932
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	3662
GO:0006915	apoptotic process	843
GO:0045893	positive regulation of transcription, DNA-dependent	3662
GO:0042981	regulation of apoptotic process	843
GO:0045087	innate immune response	729230,843
GO:0006508	proteolysis	843
GO:0035556	intracellular signal transduction	695,23228
GO:0006629	lipid metabolic process	23228
GO:0007601	visual perception	3766
GO:0007628	adult walking behavior	3766
GO:0007596	blood coagulation	3932
GO:0001974	blood vessel remodeling	729230
GO:0006874	cellular calcium ion homeostasis	729230
GO:0007268	synaptic transmission	3766
GO:0000165	MAPK cascade	11184
GO:0051384	response to glucocorticoid stimulus	3766

Continued on next page

Table C.1 – *Continued from previous page*

GO ID	GO Term	Gene in biological process (Entrez gene ID)
GO:0016032	viral reproduction	3932
GO:0042493	response to drug	3932
GO:0007243	intracellular protein kinase cascade	11184
GO:0006950	response to stress	11184
GO:0007166	cell surface receptor signaling pathway	3575
GO:0006366	transcription from RNA polymerase II promoter	5450
GO:0051289	protein homotetramerization	3766
GO:0008624	induction of apoptosis by extracellular signals	843
GO:0031295	T cell costimulation	3932
GO:0042535	positive regulation of tumor necrosis factor biosynthetic process	729230
GO:0022010	central nervous system myelination	3766
GO:0007267	cell-cell signaling	4068
GO:0043123	positive regulation of I-kappaB kinase/NF-kappaB cascade	843
GO:0034765	regulation of ion transmembrane transport	3766
GO:0071805	potassium ion transmembrane transport	3766
GO:0006952	defense response	5790
GO:0018105	peptidyl-serine phosphorylation	11184
GO:0007204	elevation of cytosolic calcium ion concentration	729230
GO:0006959	humoral immune response	4068,5450
GO:0007259	JAK-STAT cascade	729230
GO:0030168	platelet activation	3932
GO:0042110	T cell activation	3662
GO:0050870	positive regulation of T cell activation	729230,3932
GO:0016525	negative regulation of angiogenesis	729230
GO:0010820	positive regulation of T cell chemotaxis	729230
GO:0032729	positive regulation of interferon-gamma production	729230
GO:0043388	positive regulation of DNA binding	3662
GO:0051930	regulation of sensory perception of pain	3766
GO:0030217	T cell differentiation	3932
GO:0042391	regulation of membrane potential	3766
GO:0006917	induction of apoptosis	3932,843
GO:0048169	regulation of long-term neuronal synaptic plasticity	3766

Continued on next page

Table C.1 – *Continued from previous page*

GO ID	GO Term	Gene in biological process (Entrez gene ID)
GO:0006919	activation of cysteine-type endopeptidase activity involved in apoptotic process	3932
GO:0007194	negative regulation of adenylate cyclase activity	729230
GO:0009611	response to wounding	729230
GO:0030097	hemopoiesis	3932
GO:0019048	virus-host interaction	729230,3932
GO:0050852	T cell receptor signaling pathway	3932
GO:0050900	leukocyte migration	3932
GO:0002407	dendritic cell chemotaxis	729230
GO:0051209	release of sequestered calcium ion into cytosol	3932
GO:0007257	activation of JUN kinase activity	11184
GO:0045954	positive regulation of natural killer cell mediated cytotoxicity	4068
GO:0043011	myeloid dendritic cell differentiation	3662
GO:0032743	positive regulation of interleukin-2 production	729230
GO:0043966	histone H3 acetylation	3662
GO:0002827	positive regulation of T-helper 1 type immune response	729230
GO:0046641	positive regulation of alpha-beta T cell proliferation	729230
GO:0050690	regulation of defense response to virus by virus	3932
GO:0006968	cellular defense response	4068,729230
GO:0090026	positive regulation of monocyte chemotaxis	729230
GO:0060333	interferon-gamma-mediated signaling pathway	3662
GO:0010107	potassium ion import	3766
GO:0042113	B cell activation	931
GO:0045404	positive regulation of interleukin-4 biosynthetic process	3662
GO:0000185	activation of MAPKKK activity	11184
GO:0055075	potassium ion homeostasis	3766
GO:0014003	oligodendrocyte development	3766
GO:0043967	histone H4 acetylation	3662
GO:0060337	type I interferon-mediated signaling pathway	3662

Continued on next page

Table C.1 – *Continued from previous page*

GO ID	GO Term	Gene in biological process (Entrez gene ID)
GO:0051385	response to mineralocorticoid stimulus	3766
GO:0045086	positive regulation of interleukin-2 biosynthetic process	3662
GO:0006882	cellular zinc ion homeostasis	3932
GO:0051249	regulation of lymphocyte activation	3932
GO:2000473	positive regulation of hematopoietic stem cell migration	729230
GO:0051935	glutamate uptake involved in synaptic transmission	3766
GO:0010574	regulation of vascular endothelial growth factor production	729230
GO:0019725	cellular homeostasis	729230
GO:0090265	positive regulation of immune complex clearance by monocytes and macrophages	729230
GO:0000018	regulation of DNA recombination	3575
GO:0038111	interleukin-7-mediated signaling pathway	3575
GO:0060075	regulation of resting membrane potential	3766
GO:0060081	membrane hyperpolarization	3766
GO:0002829	negative regulation of type 2 immune response	729230
GO:0035705	T-helper 17 cell chemotaxis	729230
GO:0043310	negative regulation of eosinophil degranulation	729230
GO:2000439	positive regulation of monocyte extravasation	729230
GO:2000451	positive regulation of CD8-positive, alpha-beta T cell extravasation	729230
GO:2000464	positive regulation of astrocyte chemotaxis	729230
GO:0051938	L-glutamate import	3766
GO:0009637	response to blue light	3766
GO:0021554	optic nerve development	3766
GO:0034122	negative regulation of toll-like receptor signaling pathway	3662
GO:0045082	positive regulation of interleukin-10 biosynthetic process	3662
GO:0045368	positive regulation of interleukin-13 biosynthetic process	3662

Continued on next page

Table C.1 – *Continued from previous page*

GO ID	GO Term	Gene in biological process (Entrez gene ID)
GO:0045622	regulation of T-helper cell differentiation	3662

Table C.1: Intersection of genes with GO biological processes for differential graphlet community 1

GO ID	GO Term	Gene in biological process (Entrez gene ID)
GO:0006355	regulation of transcription, DNA-dependent	3662,5450
GO:0006351	transcription, DNA-dependent	3662,5450
GO:0007165	signal transduction	729230,7305,1236,3575
GO:0007275	multicellular organismal development	397
GO:0007411	axon guidance	7305
GO:0019221	cytokine-mediated signaling pathway	729230,3662,3575
GO:0002606	positive regulation of dendritic cell antigen processing and presentation	1236
GO:0002885	positive regulation of hypersensitivity	1236
GO:0002922	positive regulation of humoral immune response	1236
GO:0006935	chemotaxis	729230,1236
GO:0006955	immune response	729230,1520,1236,397,3575,931
GO:0007186	G-protein coupled receptor signaling pathway	729230,1236
GO:0032496	response to lipopolysaccharide	1236
GO:0032649	regulation of interferon-gamma production	1236
GO:0032735	positive regulation of interleukin-12 production	1236
GO:0045060	negative thymic T cell selection	1236
GO:0050706	regulation of interleukin-1 beta secretion	1236
GO:0050862	positive regulation of T cell receptor signaling pathway	3932,1236
GO:0070098	chemokine-mediated signaling pathway	729230,1236
GO:0072610	interleukin-12 secretion	1236
GO:0090023	positive regulation of neutrophil chemotaxis	1236
GO:0097029	mature dendritic cell differentiation	1236
GO:2000510	positive regulation of dendritic cell chemotaxis	1236

Continued on next page

Table C.2 – *Continued from previous page*

GO ID	GO Term	Gene in biological process (Entrez gene ID)
GO:2000522	positive regulation of immunological synapse formation	1236
GO:2000525	positive regulation of T cell costimulation	1236
GO:2000526	positive regulation of glycoprotein biosynthetic process involved in immunological synapse formation	1236
GO:0006954	inflammatory response	729230
GO:0006468	protein phosphorylation	695,3932
GO:0050729	positive regulation of inflammatory response	729230
GO:0016310	phosphorylation	695,3932
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	3662
GO:0045893	positive regulation of transcription, DNA-dependent	3662
GO:0007264	small GTPase mediated signal transduction	397
GO:0051056	regulation of small GTPase mediated signal transduction	397
GO:0045087	innate immune response	729230,1520
GO:0006508	proteolysis	1520
GO:0035556	intracellular signal transduction	7305,695
GO:0007596	blood coagulation	3932
GO:0001974	blood vessel remodeling	729230
GO:0006874	cellular calcium ion homeostasis	729230
GO:0030036	actin cytoskeleton organization	397
GO:0016032	viral reproduction	3932
GO:0043547	positive regulation of GTPase activity	397
GO:0042493	response to drug	3932
GO:0007166	cell surface receptor signaling pathway	3575
GO:0006366	transcription from RNA polymerase II promoter	5450
GO:0007162	negative regulation of cell adhesion	397
GO:0031295	T cell costimulation	3932
GO:0042535	positive regulation of tumor necrosis factor biosynthetic process	729230
GO:0050776	regulation of immune response	7305
GO:0007229	integrin-mediated signaling pathway	7305
GO:0006952	defense response	5790

Continued on next page

Table C.2 – *Continued from previous page*

GO ID	GO Term	Gene in biological process (Entrez gene ID)
GO:0007204	elevation of cytosolic calcium ion concentration	729230
GO:0006959	humoral immune response	5450
GO:0007259	JAK-STAT cascade	729230
GO:0030168	platelet activation	3932
GO:0042110	T cell activation	3662
GO:0050870	positive regulation of T cell activation	729230,3932
GO:0016525	negative regulation of angiogenesis	729230
GO:0010820	positive regulation of T cell chemotaxis	729230
GO:0032729	positive regulation of interferon-gamma production	729230
GO:0043388	positive regulation of DNA binding	3662
GO:0030217	T cell differentiation	3932
GO:0019882	antigen processing and presentation	1520
GO:0002474	antigen processing and presentation of peptide antigen via MHC class I	1520
GO:0006917	induction of apoptosis	3932
GO:0006928	cellular component movement	397
GO:0007266	Rho protein signal transduction	397
GO:0006919	activation of cysteine-type endopeptidase activity involved in apoptotic process	3932
GO:0007194	negative regulation of adenylate cyclase activity	729230
GO:0042590	antigen processing and presentation of exogenous peptide antigen via MHC class I	1520
GO:0009611	response to wounding	729230
GO:0030097	hemopoiesis	3932
GO:0019048	virus-host interaction	729230,3932
GO:0050852	T cell receptor signaling pathway	3932
GO:0050900	leukocyte migration	3932
GO:0002407	dendritic cell chemotaxis	729230
GO:0051209	release of sequestered calcium ion into cytosol	3932
GO:0043011	myeloid dendritic cell differentiation	3662
GO:0032743	positive regulation of interleukin-2 production	729230
GO:0043966	histone H3 acetylation	3662
GO:0002250	adaptive immune response	1520

Continued on next page

Table C.2 – *Continued from previous page*

GO ID	GO Term	Gene in biological process (Entrez gene ID)
GO:0002827	positive regulation of T-helper 1 type immune response	729230
GO:0046641	positive regulation of alpha-beta T cell proliferation	729230
GO:0050690	regulation of defense response to virus by virus	3932
GO:0006968	cellular defense response	729230,7305
GO:0090026	positive regulation of monocyte chemotaxis	729230
GO:0002281	macrophage activation involved in immune response	7305
GO:0002283	neutrophil activation involved in immune response	7305
GO:0060333	interferon-gamma-mediated signaling pathway	3662
GO:0042113	B cell activation	931
GO:0045404	positive regulation of interleukin-4 biosynthetic process	3662
GO:0043967	histone H4 acetylation	3662
GO:0060337	type I interferon-mediated signaling pathway	3662
GO:0045086	positive regulation of interleukin-2 biosynthetic process	3662
GO:0006882	cellular zinc ion homeostasis	3932
GO:0051249	regulation of lymphocyte activation	3932
GO:0097067	cellular response to thyroid hormone stimulus	1520
GO:2000473	positive regulation of hematopoietic stem cell migration	729230
GO:0010574	regulation of vascular endothelial growth factor production	729230
GO:0019725	cellular homeostasis	729230
GO:0090265	positive regulation of immune complex clearance by monocytes and macrophages	729230
GO:0000018	regulation of DNA recombination	3575
GO:0038111	interleukin-7-mediated signaling pathway	3575
GO:0002480	antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-independent	1520

Continued on next page

Table C.2 – *Continued from previous page*

GO ID	GO Term	Gene in biological process (Entrez gene ID)
GO:0002829	negative regulation of type 2 immune response	729230
GO:0035705	T-helper 17 cell chemotaxis	729230
GO:0043310	negative regulation of eosinophil degranulation	729230
GO:2000439	positive regulation of monocyte extravasation	729230
GO:2000451	positive regulation of CD8-positive, alpha-beta T cell extravasation	729230
GO:2000464	positive regulation of astrocyte chemotaxis	729230
GO:0034122	negative regulation of toll-like receptor signaling pathway	3662
GO:0045082	positive regulation of interleukin-10 biosynthetic process	3662
GO:0045368	positive regulation of interleukin-13 biosynthetic process	3662
GO:0045622	regulation of T-helper cell differentiation	3662

Table C.2: Intersection of genes with GO biological processes for differential graphlet community 2

GO ID	GO Term	Gene in biological process (Entrez gene ID)
GO:0006355	regulation of transcription, DNA-dependent	4208
GO:0006351	transcription, DNA-dependent	4208
GO:0007049	cell cycle	990
GO:0008283	cell proliferation	6491
GO:0007165	signal transduction	7305,8404,6387
GO:0007275	multicellular organismal development	6491,397,4256
GO:0007411	axon guidance	7305
GO:0006935	chemotaxis	6387
GO:0006955	immune response	397,6387
GO:0007186	G-protein coupled receptor signaling pathway	6387
GO:0070098	chemokine-mediated signaling pathway	6387
GO:0001764	neuron migration	6387
GO:0008285	negative regulation of cell proliferation	990
GO:0008152	metabolic process	4313

Continued on next page

Table C.3 – *Continued from previous page*

GO ID	GO Term	Gene in biological process (Entrez gene ID)
GO:0055114	oxidation-reduction process	6241
GO:0030154	cell differentiation	4256
GO:0030900	forebrain development	6491
GO:0045666	positive regulation of neuron differentiation	6387
GO:0007264	small GTPase mediated signal transduction	397
GO:0043066	negative regulation of apoptotic process	6491,6387
GO:0051056	regulation of small GTPase mediated signal transduction	397
GO:0030198	extracellular matrix organization	4313
GO:0001701	in utero embryonic development	6491
GO:0030324	lung development	4256
GO:0051301	cell division	990
GO:0007155	cell adhesion	6387
GO:0044281	small molecule metabolic process	6241
GO:0006508	proteolysis	4313
GO:0035556	intracellular signal transduction	7305
GO:0022617	extracellular matrix disassembly	4313
GO:0007420	brain development	6387
GO:0035264	multicellular organism growth	6491
GO:0008284	positive regulation of cell proliferation	6387
GO:0006874	cellular calcium ion homeostasis	6387
GO:0030036	actin cytoskeleton organization	397
GO:0009615	response to virus	6387
GO:0001938	positive regulation of endothelial cell proliferation	6387
GO:0008217	regulation of blood pressure	59
GO:0051216	cartilage development	4256
GO:0006260	DNA replication	6241,990
GO:0008156	negative regulation of DNA replication	990
GO:0030334	regulation of cell migration	6387
GO:0007067	mitosis	990
GO:0030335	positive regulation of cell migration	6387
GO:0006936	muscle contraction	25802
GO:0043547	positive regulation of GTPase activity	397
GO:0001843	neural tube closure	6491
GO:0007368	determination of left/right symmetry	6491
GO:0001666	response to hypoxia	6387
GO:0001502	cartilage condensation	4256

Continued on next page

Table C.3 – *Continued from previous page*

GO ID	GO Term	Gene in biological process (Entrez gene ID)
GO:0030500	regulation of bone mineralization	4256
GO:0007224	smoothened signaling pathway	6491
GO:0000082	G1/S transition of mitotic cell cycle	6241,990
GO:0043434	response to peptide hormone stimulus	6387
GO:0031100	organ regeneration	6387
GO:0000278	mitotic cell cycle	6241,990
GO:0007162	negative regulation of cell adhesion	397
GO:0050776	regulation of immune response	7305
GO:0007229	integrin-mediated signaling pathway	7305
GO:0030903	notochord development	6491
GO:0007281	germ cell development	6387
GO:0009725	response to hormone stimulus	4256
GO:0014829	vascular smooth muscle contraction	59
GO:0001569	patterning of blood vessels	6387
GO:0021915	neural tube development	6491
GO:0042221	response to chemical stimulus	4256
GO:0009186	deoxyribonucleoside diphosphate metabolic process	6241
GO:0008045	motor axon guidance	6387
GO:0000079	regulation of cyclin-dependent protein kinase activity	990
GO:0009314	response to radiation	6387
GO:0009612	response to mechanical stimulus	4256,6387
GO:0007584	response to nutrient	4256
GO:0051592	response to calcium ion	4256
GO:0001525	angiogenesis	4313
GO:0006928	cellular component movement	397
GO:0007266	Rho protein signal transduction	397
GO:0000075	cell cycle checkpoint	990
GO:0000084	S phase of mitotic cell cycle	990
GO:0000216	M/G1 transition of mitotic cell cycle	990
GO:0001503	ossification	4256
GO:0051259	protein oligomerization	6241
GO:0008344	adult locomotory behavior	6387
GO:0001947	heart looping	6491
GO:2000107	negative regulation of leukocyte apoptotic process	6387
GO:0015949	nucleobase-containing small molecule interconversion	6241
GO:0055086	nucleobase-containing small molecule metabolic process	6241

Continued on next page

Table C.3 – *Continued from previous page*

GO ID	GO Term	Gene in biological process (Entrez gene ID)
GO:0000578	embryonic axis specification	6491
GO:0008354	germ cell migration	6387
GO:0008064	regulation of actin polymerization or depolymerization	6387
GO:0009408	response to heat	6387
GO:0048754	branching morphogenesis of a tube	4256
GO:0050930	induction of positive chemotaxis	6387
GO:0042098	T cell proliferation	6387
GO:0006968	cellular defense response	7305
GO:0090026	positive regulation of monocyte chemotaxis	6387
GO:0002281	macrophage activation involved in immune response	7305
GO:0002283	neutrophil activation involved in immune response	7305
GO:0048842	positive regulation of axon extension involved in axon guidance	6387
GO:0000076	DNA replication checkpoint	990
GO:0000083	regulation of transcription involved in G1/S phase of mitotic cell cycle	6241,990
GO:0022029	telencephalon cell migration	6387
GO:0030574	collagen catabolic process	4313
GO:0008015	blood circulation	6387
GO:0051290	protein heterotetramerization	6241
GO:0051929	positive regulation of calcium ion transport via voltage-gated calcium channel activity	6387
GO:0033603	positive regulation of dopamine secretion	6387
GO:0009263	deoxyribonucleotide biosynthetic process	6241
GO:0007089	traversing start control point of mitotic cell cycle	990
GO:0009262	deoxyribonucleotide metabolic process	6241

Continued on next page

Table C.3 – *Continued from previous page*

GO ID	GO Term	Gene in biological process (Entrez gene ID)
GO:0001667	ameboidal cell migration	6387
GO:0090007	regulation of mitotic anaphase	990
GO:0071777	positive regulation of cell cycle cytokinesis	990
GO:0033504	floor plate development	6491
GO:0051984	positive regulation of chromosome segregation	990

Table C.3: Intersection of genes with GO biological processes for differential graphlet community 3

KEGG ID	KEGG pathway	Gene in pathway (Entrez gene ID)
04010	MAPK signaling pathway	11184
04060	Cytokine-cytokine receptor interaction	729230,1236,3575
04062	Chemokine signaling pathway	729230,1236
04210	Apoptosis	843
04380	Osteoclast differentiation	695,3932
04622	RIG-I-like receptor signaling pathway	843
04630	Jak-STAT signaling pathway	3575
04640	Hematopoietic cell lineage	931,3575
04650	Natural killer cell mediated cytotoxicity	4068,3932
04660	T cell receptor signaling pathway	3932
04662	B cell receptor signaling pathway	695
04664	Fc epsilon RI signaling pathway	695
04971	Gastric acid secretion	3766
05340	Primary immunodeficiency	695,3932,3575

Table C.4: Intersection of genes with Kegg pathways for differential graphlet community 1

KEGG ID	KEGG pathway	Gene in pathway (Entrez gene ID)
04060	Cytokine-cytokine receptor interaction	729230,1236,3575
04062	Chemokine signaling pathway	729230,1236
04142	Lysosome	1520
04145	Phagosome	1520
04380	Osteoclast differentiation	7305,695,3932
04612	Antigen processing and presentation	1520
04630	Jak-STAT signaling pathway	3575
04640	Hematopoietic cell lineage	3575,931
04650	Natural killer cell mediated cytotoxicity	7305,3932
04660	T cell receptor signaling pathway	3932
04662	B cell receptor signaling pathway	695
04664	Fc epsilon RI signaling pathway	695
04722	Neurotrophin signaling pathway	397
04962	Vasopressin-regulated water reabsorption	397
05340	Primary immunodeficiency	695,3932,3575

Table C.5: Intersection of genes with Kegg pathways for differential graphlet community 2

KEGG ID	KEGG pathway	Gene in pathway (Entrez gene ID)
00230	Purine metabolism	6241
00240	Pyrimidine metabolism	6241
00480	Glutathione metabolism	6241
01100	Metabolic pathways	6241
04010	MAPK signaling pathway	4208
04060	Cytokine-cytokine receptor interaction	6387
04062	Chemokine signaling pathway	6387
04110	Cell cycle	990
04115	p53 signaling pathway	6241
04270	Vascular smooth muscle contraction	59
04360	Axon guidance	6387
04380	Osteoclast differentiation	7305
04512	ECM-receptor interaction	3161
04650	Natural killer cell mediated cytotoxicity	7305
04670	Leukocyte transendothelial migration	4313,6387
04672	Intestinal immune network for IgA production	6387
04722	Neurotrophin signaling pathway	397
04912	GnRH signaling pathway	4313
04962	Vasopressin-regulated water reabsorption	397
05200	Pathways in cancer	4313
05219	Bladder cancer	4313

Table C.6: Intersection of genes with Kegg pathways for differential graphlet community 3

Pathway	Source	Gene in pathway (Entrez gene ID)
The role of Nef in HIV-1 replication and disease pathogenesis	REACTOME	3932
Nef Mediated CD4 Down-regulation	REACTOME	3932
Nef-mediates down modulation of cell surface receptors by recruiting them to clathrin adapters	REACTOME	3932
Nef and signal transduction	REACTOME	3932
Platelet activation, signaling and aggregation	REACTOME	3932
Hemostasis	REACTOME	3932
GPVI-mediated activation cascade	REACTOME	3932
Host Interactions of HIV factors	REACTOME	3932
HIV Infection	REACTOME	3932
Signal Transduction	REACTOME	1236
Chemokine receptors bind chemokines	REACTOME	1236

Continued on next page

Table C.7 – *Continued from previous page*

Pathway	Source	Gene in pathway (Entrez gene ID)
Signaling by GPCR	REACTOME	1236
GPCR ligand binding	REACTOME	1236
Class A/1 (Rhodopsin-like receptors)	REACTOME	1236
Peptide ligand-binding receptors	REACTOME	1236
Adaptive Immune System	REACTOME	3932
TCR signaling	REACTOME	3932
Phosphorylation of CD3 and TCR zeta chains	REACTOME	3932
Translocation of ZAP-70 to Immunological synapse	REACTOME	3932
Immune System	REACTOME	3662,3932,843
PD-1 signaling	REACTOME	3932
CD28 co-stimulation	REACTOME	3932
CD28 dependent PI3K/Akt signaling	REACTOME	3932
CD28 dependent Vav1 pathway	REACTOME	3932
Generation of second messenger molecules	REACTOME	3932
Downstream TCR signaling	REACTOME	3932
Costimulation by the CD28 family	REACTOME	3932
TRAIL signaling	REACTOME	843
FasL/ CD95L signaling	REACTOME	843
Death Receptor Signalling	REACTOME	843
Extrinsic Pathway for Apoptosis	REACTOME	843
Apoptosis	REACTOME	843
Innate Immune System	REACTOME	843
RIG-I/MDA5 mediated induction of IFN-alpha/beta pathways	REACTOME	843
NF-kB activation through FADD/RIP-1 pathway mediated by caspase-8 and -10	REACTOME	843
Cytokine Signaling in Immune system	REACTOME	3662
Interferon alpha/beta signaling	REACTOME	3662
Interferon Signaling	REACTOME	3662
Interferon gamma signaling	REACTOME	3662
Regulation of cytoplasmic and nuclear SMAD2/3 signaling	NCI_NATURE	3662,3932
IL2 signaling events mediated by STAT5	NCI_NATURE	3932
EphrinA-EPHA pathway	NCI_NATURE	3932
Canonical NF-kappaB pathway	NCI_NATURE	3932
EPHA forward signaling	NCI_NATURE	3932

Continued on next page

Table C.7 – *Continued from previous page*

Pathway	Source	Gene in pathway (Entrez gene ID)
Syndecan-1-mediated signaling events	NCI_NATURE	695,4068,11184,3662,3932,843
Regulation of CDC42 activity	NCI_NATURE	3662,3932
Glypican pathway	NCI_NATURE	695,4068,11184,3662,3932,843
GMCSF-mediated signaling events	NCI_NATURE	695,4068,11184,3662,3932,843
Insulin Pathway	NCI_NATURE	695,4068,11184,3662,3932,843
Nectin adhesion pathway	NCI_NATURE	695,4068,11184,3662,3932,843
CD40/CD40L signaling	NCI_NATURE	3932
TRAIL signaling pathway	NCI_NATURE	695,4068,11184,3662,3932,843
LPA receptor mediated events	NCI_NATURE	3932
IGF1 pathway	NCI_NATURE	695,4068,11184,3662,3932,843
PLK1 signaling events	NCI_NATURE	3932
CDC42 signaling events	NCI_NATURE	3662,3932
Signaling events mediated by Hepatocyte Growth Factor Receptor (c-Met)	NCI_NATURE	695,4068,11184,3662,3932,843
Glypican 1 network	NCI_NATURE	695,4068,11184,3662,3932,843
Fc-epsilon receptor I signaling in mast cells	NCI_NATURE	695
PDGF receptor signaling network	NCI_NATURE	695,4068,11184,3662,3932,843
EphrinB-EPHB pathway	NCI_NATURE	3932
Integrin family cell surface interactions	NCI_NATURE	695,4068,11184,3662,3932,843
IL1-mediated signaling events	NCI_NATURE	3662,3932
Caspase cascade in apoptosis	NCI_NATURE	843
Calcineurin-regulated NFAT-dependent transcription in lymphocytes	NCI_NATURE	3662
Internalization of ErbB1	NCI_NATURE	695,4068,11184,3662,3932,843
TGF-beta receptor signaling	NCI_NATURE	3662,3932
BCR signaling pathway	NCI_NATURE	695,11184
Signaling events mediated by TCPTP	NCI_NATURE	3662
Signaling events mediated by VEGFR1 and VEGFR2	NCI_NATURE	695,4068,11184,3662,3932,843
Beta1 integrin cell surface interactions	NCI_NATURE	695,4068,11184,3662,3932,843
Thromboxane A2 receptor signaling	NCI_NATURE	3932
Urokinase-type plasminogen activator (uPA) and uPAR-mediated signaling	NCI_NATURE	695,4068,11184,3662,3932,843
TCR signaling in naïve CD8+ T cells	NCI_NATURE	3932
Signaling by Aurora kinases	NCI_NATURE	3932
Polo-like kinase signaling events in the cell cycle	NCI_NATURE	3932
IFN-gamma pathway	NCI_NATURE	695,4068,11184,3662,3932,843

Continued on next page

Table C.7 – *Continued from previous page*

Pathway	Source	Gene in pathway (Entrez gene ID)
PAR1-mediated thrombin signaling events	NCI_NATURE	695,4068,11184,3662,3932,843
Regulation of p38-alpha and p38-beta	NCI_NATURE	3662,3932
PDGFR-beta signaling pathway	NCI_NATURE	695,4068,11184,3662,3932,843
Integrin-linked kinase signaling	NCI_NATURE	3662,3932
Ephrin B reverse signaling	NCI_NATURE	3932
p38 MAPK signaling pathway	NCI_NATURE	3662,3932
EGF receptor (ErbB1) signaling pathway	NCI_NATURE	695,4068,11184,3662,3932,843
p75(NTR)-mediated signaling	NCI_NATURE	3932
Thrombin/protease-activated receptor (PAR) pathway	NCI_NATURE	695,4068,11184,3662,3932,843
Class I PI3K signaling events	NCI_NATURE	695,4068,11184,3662,3932,843
Arf6 signaling events	NCI_NATURE	695,4068,11184,3662,3932,843
EPO signaling pathway	NCI_NATURE	695
Plasma membrane estrogen receptor signaling	NCI_NATURE	695,4068,11184,3662,3932,843
IL2-mediated signaling events	NCI_NATURE	3932
FAS (CD95) signaling pathway	NCI_NATURE	695,3932,843
IL3-mediated signaling events	NCI_NATURE	695,4068,11184,3662,3932,843
AP-1 transcription factor network	NCI_NATURE	3662,3932
Signaling events regulated by Ret tyrosine kinase	NCI_NATURE	3932
amb2 Integrin signaling	NCI_NATURE	3932
mTOR signaling pathway	NCI_NATURE	695,4068,11184,3662,3932,843
ErbB receptor signaling network	NCI_NATURE	695,4068,11184,3662,3932,843
Signaling events mediated by focal adhesion kinase	NCI_NATURE	695,4068,11184,3662,3932,843
VEGF and VEGFR signaling network	NCI_NATURE	695,4068,11184,3662,3932,843
Alpha9 beta1 integrin signaling events	NCI_NATURE	695,4068,11184,3662,3932,843
Signaling events mediated by Stem cell factor receptor (c-Kit)	NCI_NATURE	11184
ALK1 pathway	NCI_NATURE	3662,3932
Arf6 downstream pathway	NCI_NATURE	695,4068,11184,3662,3932,843
JNK signaling in the CD4+ TCR pathway	NCI_NATURE	11184
Role of Calcineurin-dependent NFAT signaling in lymphocytes	NCI_NATURE	3662
IL4-mediated signaling events	NCI_NATURE	3662
Signaling events mediated by PTP1B	NCI_NATURE	3932

Continued on next page

Table C.7 – *Continued from previous page*

Pathway	Source	Gene in pathway (Entrez gene ID)
Validated transcriptional targets of AP1 family members Fra1 and Fra2	NCI_NATURE	3662
CXCR4-mediated signaling events	NCI_NATURE	11184,3932
S1P1 pathway	NCI_NATURE	695,4068,11184,3662,3932,843
Alpha-synuclein signaling	NCI_NATURE	3932
ATM pathway	NCI_NATURE	4068,843
TNF receptor signaling pathway	NCI_NATURE	3662,3932,843
p53 pathway	NCI_NATURE	4068,843
ErbB1 downstream signaling	NCI_NATURE	695,4068,11184,3662,3932,843
EGFR-dependent Endothelin signaling events	NCI_NATURE	695,4068,11184,3662,3932,843
ATR signaling pathway	NCI_NATURE	4068,843
Regulation of nuclear SMAD2/3 signaling	NCI_NATURE	3662,3932
ALK1 signaling events	NCI_NATURE	3662,3932
Arf6 trafficking events	NCI_NATURE	695,4068,11184,3662,3932,843
Endothelins	NCI_NATURE	695,4068,11184,3662,3932,843
TCR signaling in naïve CD4+ T cells	NCI_NATURE	11184,3932
IL5-mediated signaling events	NCI_NATURE	695,4068,11184,3662,3932,843
IL23-mediated signaling events	NCI_NATURE	3932
Direct p53 effectors	NCI_NATURE	4068,843
IL12-mediated signaling events	NCI_NATURE	3932
IL2 signaling events mediated by PI3K	NCI_NATURE	3932
Sphingosine 1-phosphate (S1P) pathway	NCI_NATURE	695,4068,11184,3662,3932,843
Proteoglycan syndecan-mediated signaling events	NCI_NATURE	695,4068,11184,3662,3932,843
Atypical NF-kappaB pathway	NCI_NATURE	3932
BMP receptor signaling	NCI_NATURE	3662,3932
Endogenous TLR signaling	NCI_NATURE	3932
Aurora A signaling	NCI_NATURE	3932
Class I PI3K signaling events mediated by Akt	NCI_NATURE	695,4068,11184,3662,3932,843
LKB1 signaling events	NCI_NATURE	695,4068,11184,3662,3932,843
TNF alpha/NF-kB	CELL_MAP	843

Table C.7: Intersection of genes with pathways in Pathway Commons for differential graphlet community 1

Pathway	Source	Gene in pathway (Entrez gene ID)
The role of Nef in HIV-1 replication and disease pathogenesis	REACTOME	3932
Nef Mediated CD4 Down-regulation	REACTOME	3932
Nef-mediates down modulation of cell surface receptors by recruiting them to clathrin adapters	REACTOME	3932
Nef and signal transduction	REACTOME	3932
Developmental Biology	REACTOME	7305
Axon guidance	REACTOME	7305
Semaphorin interactions	REACTOME	7305
Other semaphorin interactions	REACTOME	7305
Platelet activation, signaling and aggregation	REACTOME	3932
Hemostasis	REACTOME	3932
GPVI-mediated activation cascade	REACTOME	3932
Host Interactions of HIV factors	REACTOME	3932
HIV Infection	REACTOME	3932
Signal Transduction	REACTOME	1236,397
Signaling by Rho GTPases	REACTOME	397
Rho GTPase cycle	REACTOME	397
Chemokine receptors bind chemokines	REACTOME	1236
Signaling by GPCR	REACTOME	1236
GPCR ligand binding	REACTOME	1236
Class A/1 (Rhodopsin-like receptors)	REACTOME	1236
Peptide ligand-binding receptors	REACTOME	1236
Adaptive Immune System	REACTOME	7305,3932,1520
TCR signaling	REACTOME	3932
Phosphorylation of CD3 and TCR zeta chains	REACTOME	3932
Translocation of ZAP-70 to Immunological synapse	REACTOME	3932
Immune System	REACTOME	7305,3932,1520,3662
Cell-Cell communication	REACTOME	7305
Signal regulatory protein (SIRP) family interactions	REACTOME	7305
PD-1 signaling	REACTOME	3932
Class I MHC mediated antigen processing & presentation	REACTOME	1520
CD28 co-stimulation	REACTOME	3932
CD28 dependent PI3K/Akt signaling	REACTOME	3932
CD28 dependent Vav1 pathway	REACTOME	3932

Continued on next page

Table C.8 – *Continued from previous page*

Pathway	Source	Gene in pathway (Entrez gene ID)
Generation of second messenger molecules	REACTOME	3932
Downstream TCR signaling	REACTOME	3932
Costimulation by the CD28 family	REACTOME	3932
Endosomal/Vacuolar pathway	REACTOME	1520
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	REACTOME	7305
Antigen processing-Cross presentation	REACTOME	1520
Cytokine Signaling in Immune system	REACTOME	3662
Interferon alpha/beta signaling	REACTOME	3662
Interferon Signaling	REACTOME	3662
Interferon gamma signaling	REACTOME	3662
Regulation of cytoplasmic and nuclear SMAD2/3 signaling	NCI_NATURE	3932,3662
IL2 signaling events mediated by STAT5	NCI_NATURE	3932
EphrinA-EPHA pathway	NCI_NATURE	3932
Canonical NF-kappaB pathway	NCI_NATURE	3932
EPHA forward signaling	NCI_NATURE	3932
Syndecan-1-mediated signaling events	NCI_NATURE	695,3932,3662
Regulation of CDC42 activity	NCI_NATURE	3932,3662,397
Glypican pathway	NCI_NATURE	695,3932,3662
GMCSF-mediated signaling events	NCI_NATURE	695,3932,3662
Insulin Pathway	NCI_NATURE	695,3932,3662
Nectin adhesion pathway	NCI_NATURE	695,3932,3662
CD40/CD40L signaling	NCI_NATURE	3932
TRAIL signaling pathway	NCI_NATURE	695,3932,3662,397
LPA receptor mediated events	NCI_NATURE	3932
IGF1 pathway	NCI_NATURE	695,3932,3662
PLK1 signaling events	NCI_NATURE	3932
CDC42 signaling events	NCI_NATURE	3932,3662,397
Signaling events mediated by Hepatocyte Growth Factor Receptor (c-Met)	NCI_NATURE	695,3932,3662
Glypican 1 network	NCI_NATURE	695,3932,3662
Fc-epsilon receptor I signaling in mast cells	NCI_NATURE	695
PDGF receptor signaling network	NCI_NATURE	695,3932,3662
EphrinB-EPHB pathway	NCI_NATURE	3932
Integrin family cell surface interactions	NCI_NATURE	695,3932,3662
IL1-mediated signaling events	NCI_NATURE	3932,3662

Continued on next page

Table C.8 – *Continued from previous page*

Pathway	Source	Gene in pathway (Entrez gene ID)
Caspase cascade in apoptosis	NCI_NATURE	397
Calcineurin-regulated NFAT-dependent transcription in lymphocytes	NCI_NATURE	3662
Internalization of ErbB1	NCI_NATURE	695,3932,3662
TGF-beta receptor signaling	NCI_NATURE	3932,3662
BCR signaling pathway	NCI_NATURE	695
Signaling events mediated by TCPTP	NCI_NATURE	3662
Signaling events mediated by VEGFR1 and VEGFR2	NCI_NATURE	695,3932,3662
Beta1 integrin cell surface interactions	NCI_NATURE	695,3932,3662
Thromboxane A2 receptor signaling	NCI_NATURE	3932
Urokinase-type plasminogen activator (uPA) and uPAR-mediated signaling	NCI_NATURE	695,3932,3662
TCR signaling in naïve CD8+ T cells	NCI_NATURE	3932
Signaling by Aurora kinases	NCI_NATURE	3932
Polo-like kinase signaling events in the cell cycle	NCI_NATURE	3932
IFN-gamma pathway	NCI_NATURE	695,3932,3662
Regulation of RAC1 activity	NCI_NATURE	397
PAR1-mediated thrombin signaling events	NCI_NATURE	695,3932,3662
Regulation of p38-alpha and p38-beta	NCI_NATURE	3932,3662
PDGFR-beta signaling pathway	NCI_NATURE	695,3932,3662
Integrin-linked kinase signaling	NCI_NATURE	3932,3662
Ephrin B reverse signaling	NCI_NATURE	3932
p38 MAPK signaling pathway	NCI_NATURE	3932,3662
EGF receptor (ErbB1) signaling pathway	NCI_NATURE	695,3932,3662
p75(NTR)-mediated signaling	NCI_NATURE	3932
Thrombin/protease-activated receptor (PAR) pathway	NCI_NATURE	695,3932,3662
Class I PI3K signaling events	NCI_NATURE	695,3932,3662
Arf6 signaling events	NCI_NATURE	695,3932,3662
EPO signaling pathway	NCI_NATURE	695
Plasma membrane estrogen receptor signaling	NCI_NATURE	695,3932,3662
IL2-mediated signaling events	NCI_NATURE	3932
FAS (CD95) signaling pathway	NCI_NATURE	695,3932,397
IL3-mediated signaling events	NCI_NATURE	695,3932,3662
AP-1 transcription factor network	NCI_NATURE	3932,3662

Continued on next page

Table C.8 – *Continued from previous page*

Pathway	Source	Gene in pathway (Entrez gene ID)
Signaling events regulated by Ret tyrosine kinase	NCI_NATURE	3932
amb2 Integrin signaling	NCI_NATURE	3932
mTOR signaling pathway	NCI_NATURE	695,3932,3662
RAC1 signaling pathway	NCI_NATURE	397
ErbB receptor signaling network	NCI_NATURE	695,3932,3662
Signaling events mediated by focal adhesion kinase	NCI_NATURE	695,3932,3662
VEGF and VEGFR signaling network	NCI_NATURE	695,3932,3662
Alpha9 beta1 integrin signaling events	NCI_NATURE	695,3932,3662
ALK1 pathway	NCI_NATURE	3932,3662
Arf6 downstream pathway	NCI_NATURE	695,3932,3662
Role of Calcineurin-dependent NFAT signaling in lymphocytes	NCI_NATURE	3662
IL4-mediated signaling events	NCI_NATURE	3662
Signaling events mediated by PTP1B	NCI_NATURE	3932
Validated transcriptional targets of AP1 family members Fra1 and Fra2	NCI_NATURE	3662
CXCR4-mediated signaling events	NCI_NATURE	3932
S1P1 pathway	NCI_NATURE	695,3932,3662
Alpha-synuclein signaling	NCI_NATURE	3932
TNF receptor signaling pathway	NCI_NATURE	3932,3662,397
ErbB1 downstream signaling	NCI_NATURE	695,3932,3662
EGFR-dependent Endothelin signaling events	NCI_NATURE	695,3932,3662
Regulation of nuclear SMAD2/3 signaling	NCI_NATURE	3932,3662
ALK1 signaling events	NCI_NATURE	3932,3662
Arf6 trafficking events	NCI_NATURE	695,3932,3662
Endothelins	NCI_NATURE	695,3932,3662
TCR signaling in naïve CD4+ T cells	NCI_NATURE	3932
IL5-mediated signaling events	NCI_NATURE	695,3932,3662
IL23-mediated signaling events	NCI_NATURE	3932
IL12-mediated signaling events	NCI_NATURE	3932
IL2 signaling events mediated by PI3K	NCI_NATURE	3932
Sphingosine 1-phosphate (S1P) pathway	NCI_NATURE	695,3932,3662
Proteoglycan syndecan-mediated signaling events	NCI_NATURE	695,3932,3662
Atypical NF-kappaB pathway	NCI_NATURE	3932
BMP receptor signaling	NCI_NATURE	3932,3662

Continued on next page

Table C.8 – *Continued from previous page*

Pathway	Source	Gene in pathway (Entrez gene ID)
Endogenous TLR signaling	NCI_NATURE	3932
Aurora A signaling	NCI_NATURE	3932
RhoA signaling pathway	NCI_NATURE	397
Class I PI3K signaling events mediated by Akt	NCI_NATURE	695,3932,3662
LKB1 signaling events	NCI_NATURE	695,3932,3662
Regulation of RhoA activity	NCI_NATURE	397

Table C.8: Intersection of genes with pathways in Pathway Commons for differential graphlet community 2

Pathway	Source	Gene in pathway (Entrez gene ID)
Developmental Biology	REACTOME	7305,4208
Axon guidance	REACTOME	7305
Semaphorin interactions	REACTOME	7305
Other semaphorin interactions	REACTOME	7305
CDO in myogenesis	REACTOME	4208
Myogenesis	REACTOME	4208
Mitotic Prometaphase	REACTOME	11130
Mitotic M-M/G1 phases	REACTOME	11130,990
M Phase	REACTOME	11130
DNA Replication	REACTOME	11130,990
Smooth Muscle Contraction	REACTOME	25802,59
Muscle contraction	REACTOME	25802,59
Nuclear Events (kinase and transcription factor activation)	REACTOME	4208
ERK/MAPK targets	REACTOME	4208
Signalling by NGF	REACTOME	4208
NGF signalling via TRKA from the plasma membrane	REACTOME	4208
Signal Transduction	REACTOME	397,4208,6387
CDC6 association with the ORC:origin complex	REACTOME	990
Assembly of the pre-replicative complex	REACTOME	990
DNA Replication Pre-Initiation	REACTOME	990
M/G1 Transition	REACTOME	990
E2F mediated regulation of DNA replication	REACTOME	6241,990
G1/S-Specific Transcription	REACTOME	6241,990

Continued on next page

Table C.9 – *Continued from previous page*

Pathway	Source	Gene in pathway (Entrez gene ID)
G1/S Transition	REACTOME	6241,990
S Phase	REACTOME	990
Cyclin A:Cdk2-associated events at S phase entry	REACTOME	990
G0 and Early G1	REACTOME	990
Mitotic G1-G1/S phases	REACTOME	6241,990
Cell Cycle, Mitotic	REACTOME	6241,11130,990
Regulation of Insulin-like Growth Factor (IGF) Activity by Insulin-like Growth Factor Binding Proteins (IGFBPs)	REACTOME	4313
Diabetes pathways	REACTOME	4313
Removal of licensing factors from origins	REACTOME	990
Regulation of DNA replication	REACTOME	990
Association of licensing factors with the pre-replicative complex	REACTOME	990
CDT1 association with the CDC6:ORC:origin complex	REACTOME	990
Activation of the pre-replicative complex	REACTOME	990
Synthesis of DNA	REACTOME	990
Switching of origins to a post-replicative state	REACTOME	990
Orc1 removal from chromatin	REACTOME	990
CDK-mediated phosphorylation and removal of Cdc6	REACTOME	990
Signaling by Rho GTPases	REACTOME	397
Rho GTPase cycle	REACTOME	397
Chemokine receptors bind chemokines	REACTOME	6387
Signaling by GPCR	REACTOME	6387
GPCR ligand binding	REACTOME	6387
Class A/1 (Rhodopsin-like receptors)	REACTOME	6387
Peptide ligand-binding receptors	REACTOME	6387
Adaptive Immune System	REACTOME	7305
Immune System	REACTOME	7305,4208
Synthesis and interconversion of nucleotide di- and triphosphates	REACTOME	6241
Metabolism	REACTOME	6241
Metabolism of nucleotides	REACTOME	6241
Cell Cycle Checkpoints	REACTOME	990

Continued on next page

Table C.9 – *Continued from previous page*

Pathway	Source	Gene in pathway (Entrez gene ID)
G2/M DNA damage checkpoint	REACTOME	990
G2/M Checkpoints	REACTOME	990
Activation of ATR in response to replication stress	REACTOME	990
Cell-Cell communication	REACTOME	7305
Signal regulatory protein (SIRP) family interactions	REACTOME	7305
TRIF mediated TLR3 signaling	REACTOME	4208
Toll Like Receptor 3 (TLR3) Cascade	REACTOME	4208
TRAF6 Mediated Induction of proinflammatory cytokines	REACTOME	4208
Toll Like Receptor 5 (TLR5) Cascade	REACTOME	4208
Toll Like Receptor 7/8 (TLR7/8) Cascade	REACTOME	4208
MyD88 dependent cascade initiated on endosome	REACTOME	4208
TRAF6 mediated induction of NFkB and MAP kinases upon TLR7/8 or 9 activation	REACTOME	4208
Toll Like Receptor 9 (TLR9) Cascade	REACTOME	4208
Activated TLR4 signalling	REACTOME	4208
MyD88:Mal cascade initiated on plasma membrane	REACTOME	4208
Toll Like Receptor 4 (TLR4) Cascade	REACTOME	4208
NFkB and MAP kinases activation mediated by TLR4 signaling repertoire	REACTOME	4208
MyD88-independent cascade initiated on plasma membrane	REACTOME	4208
Innate Immune System	REACTOME	4208
Toll Receptor Cascades	REACTOME	4208
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	REACTOME	7305
Toll Like Receptor 10 (TLR10) Cascade	REACTOME	4208
MAPK targets/ Nuclear events mediated by MAP kinases	REACTOME	4208
MAP kinase activation in TLR cascade	REACTOME	4208
MyD88 cascade initiated on plasma membrane	REACTOME	4208
Toll Like Receptor 2 (TLR2) Cascade	REACTOME	4208

Continued on next page

Table C.9 – *Continued from previous page*

Pathway	Source	Gene in pathway (Entrez gene ID)
Toll Like Receptor TLR1:TLR2 Cascade	REACTOME	4208
Toll Like Receptor TLR6:TLR2 Cascade	REACTOME	4208
guanosine nucleotides de novo biosynthesis	HUMANCYC	6241
pyrimidine deoxyribonucleotides de novo biosynthesis	HUMANCYC	6241
Wnt signaling network	NCI_NATURE	4313
Osteopontin-mediated events	NCI_NATURE	4313
Regulation of cytoplasmic and nuclear SMAD2/3 signaling	NCI_NATURE	4208
Signaling events mediated by the Hedgehog family	NCI_NATURE	6491
Signaling events mediated by HDAC Class I	NCI_NATURE	4208
Integrins in angiogenesis	NCI_NATURE	4313
Syndecan-1-mediated signaling events	NCI_NATURE	4313,4208,59,4256,6387
Regulation of CDC42 activity	NCI_NATURE	397,4313,4208,4256,6387
Glypican pathway	NCI_NATURE	4313,4208,59,4256,6387
GMCSF-mediated signaling events	NCI_NATURE	4313,4208,59,4256,6387
Insulin Pathway	NCI_NATURE	4313,4208,59,4256,6387
Signaling events mediated by HDAC Class II	NCI_NATURE	4208
Stabilization and expansion of the E-cadherin adherens junction	NCI_NATURE	4313
Glypican 3 network	NCI_NATURE	4313
Nectin adhesion pathway	NCI_NATURE	4313,4208,59,4256,6387
Neurotrophic factor-mediated Trk receptor signaling	NCI_NATURE	4208
TRAIL signaling pathway	NCI_NATURE	397,4313,4208,59,4256,6387
LPA receptor mediated events	NCI_NATURE	4313
IGF1 pathway	NCI_NATURE	4313,4208,59,4256,6387
ATF-2 transcription factor network	NCI_NATURE	4313
HIF-1-alpha transcription factor network	NCI_NATURE	6387
CDC42 signaling events	NCI_NATURE	397,4313,4208,4256,6387
Signaling events mediated by Hepatocyte Growth Factor Receptor (c-Met)	NCI_NATURE	4313,4208,59,4256,6387
Glypican 1 network	NCI_NATURE	4313,4208,59,4256,6387
N-cadherin signaling events	NCI_NATURE	4313

Continued on next page

Table C.9 – *Continued from previous page*

Pathway	Source	Gene in pathway (Entrez gene ID)
PDGF receptor signaling network	NCI_NATURE	4313,4208,59,4256,6387
Integrin family cell surface interactions	NCI_NATURE	4313,4208,59,4256,6387
IL1-mediated signaling events	NCI_NATURE	4208
Caspase cascade in apoptosis	NCI_NATURE	397
Internalization of ErbB1	NCI_NATURE	4313,4208,59,4256,6387
TGF-beta receptor signaling	NCI_NATURE	4208
Posttranslational regulation of adherens junction stability and dissassembly	NCI_NATURE	4313
Signaling events mediated by VEGFR1 and VEGFR2	NCI_NATURE	4313,4208,59,4256,6387
Beta1 integrin cell surface interactions	NCI_NATURE	4313,4208,59,4256,6387
Urokinase-type plasminogen activator (uPA) and uPAR-mediated signaling	NCI_NATURE	4313,4208,59,4256,6387
Signaling mediated by p38-alpha and p38-beta	NCI_NATURE	4208
IFN-gamma pathway	NCI_NATURE	4313,4208,59,4256,6387
Regulation of RAC1 activity	NCI_NATURE	397,4313
PAR1-mediated thrombin signaling events	NCI_NATURE	4313,4208,59,4256,6387
Noncanonical Wnt signaling pathway	NCI_NATURE	4313
Regulation of p38-alpha and p38-beta	NCI_NATURE	4208
PDGFR-beta signaling pathway	NCI_NATURE	4313,4208,59,4256,6387
Canonical Wnt signaling pathway	NCI_NATURE	4313
E-cadherin signaling in the nascent adherens junction	NCI_NATURE	4313
Integrin-linked kinase signaling	NCI_NATURE	4313,4208,4256,6387
Regulation of nuclear beta catenin signaling and target gene transcription	NCI_NATURE	4313
p38 MAPK signaling pathway	NCI_NATURE	4208
EGF receptor (ErbB1) signaling pathway	NCI_NATURE	4313,4208,59,4256,6387
p75(NTR)-mediated signaling	NCI_NATURE	4208
Thrombin/protease-activated receptor (PAR) pathway	NCI_NATURE	4313,4208,59,4256,6387
Class I PI3K signaling events	NCI_NATURE	4313,4208,59,4256,6387
Arf6 signaling events	NCI_NATURE	4313,4208,59,4256,6387
Plasma membrane estrogen receptor signaling	NCI_NATURE	4313,4208,59,4256,6387
FAS (CD95) signaling pathway	NCI_NATURE	397
FOXM1 transcription factor network	NCI_NATURE	4313

Continued on next page

Table C.9 – *Continued from previous page*

Pathway	Source	Gene in pathway (Entrez gene ID)
IL3-mediated signaling events	NCI_NATURE	4313,4208,59,4256,6387
AP-1 transcription factor network	NCI_NATURE	4313,4208,4256,6387
Hypoxic and oxygen homeostasis regulation of HIF-1-alpha	NCI_NATURE	6387
amb2 Integrin signaling	NCI_NATURE	4313
mTOR signaling pathway	NCI_NATURE	4313,4208,59,4256,6387
RAC1 signaling pathway	NCI_NATURE	397,4313
ErbB receptor signaling network	NCI_NATURE	4313,4208,59,4256,6387
Signaling events mediated by focal adhesion kinase	NCI_NATURE	4313,4208,59,4256,6387
VEGF and VEGFR signaling network	NCI_NATURE	4313,4208,59,4256,6387
Alpha9 beta1 integrin signaling events	NCI_NATURE	4313,4208,59,4256,6387
Syndecan-2-mediated signaling events	NCI_NATURE	4313
ALK1 pathway	NCI_NATURE	4208
Arf6 downstream pathway	NCI_NATURE	4313,4208,59,4256,6387
Validated transcriptional targets of AP1 family members Fra1 and Fra2	NCI_NATURE	4313,4256
CXCR4-mediated signaling events	NCI_NATURE	6387
S1P1 pathway	NCI_NATURE	4313,4208,59,4256,6387
ATM pathway	NCI_NATURE	6241,990,4313
TNF receptor signaling pathway	NCI_NATURE	397,4208
p53 pathway	NCI_NATURE	4313
ErbB1 downstream signaling	NCI_NATURE	4313,4208,59,4256,6387
EGFR-dependent Endothelin signaling events	NCI_NATURE	4313,4208,59,4256,6387
Syndecan-4-mediated signaling events	NCI_NATURE	4313,6387
ATR signaling pathway	NCI_NATURE	990,4313
Regulation of nuclear SMAD2/3 signaling	NCI_NATURE	4208
ALK1 signaling events	NCI_NATURE	4208
Angiopoietin receptor Tie2-mediated signaling	NCI_NATURE	4313
Arf6 trafficking events	NCI_NATURE	4313,4208,59,4256,6387
E-cadherin signaling events	NCI_NATURE	4313
Endothelins	NCI_NATURE	4313,4208,59,4256,6387
Regulation of retinoblastoma protein	NCI_NATURE	4208
E2F transcription factor network	NCI_NATURE	6241,990
IL5-mediated signaling events	NCI_NATURE	4313,4208,59,4256,6387
Direct p53 effectors	NCI_NATURE	4313
Sphingosine 1-phosphate (S1P) pathway	NCI_NATURE	4313,4208,59,4256,6387

Continued on next page

Table C.9 – *Continued from previous page*

Pathway	Source	Gene in pathway (Entrez gene ID)
Proteoglycan syndecan-mediated signaling events	NCI_NATURE	4313,4208,59,4256,6387
BMP receptor signaling	NCI_NATURE	4208
RhoA signaling pathway	NCI_NATURE	397,4313
Class I PI3K signaling events mediated by Akt	NCI_NATURE	4313,4208,59,4256,6387
LKB1 signaling events	NCI_NATURE	4313,4208,59,4256,6387
Regulation of RhoA activity	NCI_NATURE	397,4313
Trk receptor signaling mediated by the MAPK pathway	NCI_NATURE	4208
Trk receptor signaling mediated by PI3K and PLC-gamma	NCI_NATURE	4208
TGFBR	CELL_MAP	4208

Table C.9: Intersection of genes with pathways in Pathway Commons for differential graphlet community 3

Bibliography

[1] Community Detection In R, 2012. http://igraph.wikidot.com/community-detection-in-r.

[2] National Human Genome Research Institute, November 2013. http://www.genome.gov/27530687.

[3] American Cancer Society, October 2013. http://www.cancer.org/.

[4] Institute for systems biology, October 2013. http://www.systemsbiology.org/about-systems-biology.

[5] Bissan Al-Lazikani, Udai Banerji, and Paul Workman. Combinatorial drug therapy for cancer in the post-genomic era. *Nat Biotechnol.*, 30(7):679–692, 2012.

[6] Donatella Aldinucci, Marta Celegato, Cinzia Borghese, Alfonso Colombatti, and Antonino Carbone. IRF4 silencing inhibits Hodgkin lymphoma cell proliferation, survival and CCL5 secretion. *British journal of haematology*, 152(2):182–90, 2011.

[7] Manit Arya, Iqbal S. Shergill, Magali Williamson, Lyndon Gommersall, Neehar Arya, and Hitendra R. H. Patel. Basic principles of real-time quantitative PCR. *Expert review of molecular diagnostics*, 5(2):209–219, 2005.

[8] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew

Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25:25–29, 2000.

[9] Gary Bader, Ethan Cerami, Benjamin Gross, and Chris Sander. The Cancer Cell Map. The Computational Biology Center at Memorial Sloan-Kettering Cancer Center and the Institute of Bioinformatics. http://cancer.cellmap.org/cellmap/home.do.

[10] Albert-László Barabási and Zoltán N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5:101–113, February 2004.

[11] Vladimir Batagelj and Andrej Mrvar. A subquadratic triad census algorithm for large sparse networks with small maximum degree. *Social networks*, 23(3):237–243, 2001.

[12] David G. Beer, Sharon L.R. Kardia, Chiang-Ching Huang, Thomas J. Giordano, Albert M. Levin, David E. Misek, Lin Lin, Guoan Chen, Tarek G. Gharib, Dafydd G. Thomas, Michelle L. Lizyness, Rork Kuick, Satoru Hayasaka, Jeremy M.G. Taylor, Mark D. Iannettoni, Mark B. Orringer, and Samir Hanash. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med.*, 8(8):816–24, 2002.

[13] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57(1):289–300, 1995.

[14] Arindam Bhattacharjee, William G. Richards, Jane Staunton, Cheng Li, Stefano Monti, Priya Vasa, Christine Ladd, Javad Beheshti, Raphael Bueno, Michael Gillette, Massimo Loda, Griffin Weber, Eugene J. Mark, Eric S. Lander, Wing Wong, Bruce E. Johnson, Todd R. Golub, David J. Sugarbaker, and Matthew

Meyerson. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci USA*, 98(24):13790–5, 2001.

[15] F. Bianchi, P. Nuciforo, M. Vecchi, L. Bernard, L. Tizzoni, A. Marchetti, F. Buttitta, L. Felicioni, F. Nicassio, and P. P. Di Fiore. Survival prediction of stage I lung adenocarcinomas by expression of 10 genes. *J. Clin. Invest.*, 117(11):3436–3444, Nov 2007.

[16] David Binns, Emily Dimmer, Rachael Huntley, Daniel Barrell, Claire O'Donovan, and Rolf Apweiler. QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics*, 25(22):3045–6, 2009.

[17] F. H. Blackhall, D. A. Wigle, I. Jurisica, M. Pintilie, N. Liu, G. Darling, M. R. Johnston, S. Keshavjee, T. Waddell, T. Winton, F. A. Shepherd, and M. S. Tsao. Validating the prognostic value of marker genes derived from a non-small cell lung cancer microarray study. *Lung Cancer*, 46(2):197–204, Nov 2004.

[18] R. Bonecchi, E. Galliera, E. M. Borroni, M. M. Corsi, M. Locati, and A. Mantovani. Chemokines and chemokine receptors: an overview. *Front Biosci*, 14:540–51, 2009.

[19] Paul C. Boutros, Suzanne K. Lau, Melania Pintilie, Ni Liu, Frances A. Shepherd, Sandy D. Der, Ming-Sound Tsao, Linda Z. Penn, and Igor Jurisica. Prognostic gene signatures for non-small-cell lung cancer. *Proc Natl Acad Sci U S A*, 106(8):2824–2828, 2009.

[20] R. M. Bremnes, K. Al-Shibli, T. Donnem, R. Sirera, S. Al-Saad, S. Andersen, H. Stenvold, C. Camps, and L. T. Busund. The role of tumor-infiltrating immune cells and chronic inflammation at the tumor site on cancer development, progression, and prognosis: emphasis on non-small cell lung cancer. *J Thorac Oncol*, 6(4):824–33, 2011.

[21] Kevin R. Brown and Igor Jurisica. Online predicted human interaction database. *Bioinformatics*, 21(9):2076–2082, 2005.

[22] Kevin R. Brown, David Otasek, Muhammad Ali, Michael J. McGuffin, Wing Xie, Baiju Devani, Ian Lawson van Toch, and Igor Jurisica. NAViGaTOR: Network Analysis, Visualization and Graphing Toronto. *Bioinformatics*, 25(24):3327–3329, 2009.

[23] P. O. Brown and D. Botstein. Exploring the new world of the genome with DNA microarrays. *Nature Genetics Supplement*, 21:33–37, 1999.

[24] Helen C. Causton, John Quackenbush, and Alvis Brazma. *Microarray Gene Expression Data Analysis*, chapter Introduction. Blackwell Publishing, 2003.

[25] Ethan G. Cerami, Benjamin E. Gross, Emek Demir, Igor Rodchenkov, Özgün Babur, Nadia Anwar, Nikolaus Schultz, Gary D. Bader, and Chris Sander. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res*, 39:D685–90, 2011.

[26] H. Y. Chen, S. L. Yu, C. H. Chen, G. C. Chang, C. Y. Chen, A. Yuan, C. L. Cheng, C. H. Wang, H. J. Terng, S. F. Kao, W. K. Chan, H. N. Li, C. C. Liu, S. Singh, W. J. Chen, J. J. Chen, and P. C. Yang. A five-gene signature and clinical outcome in non-small-cell lung cancer. *N. Engl. J. Med.*, 356(1):11–20, Jan 2007.

[27] J. K. Choi, U. Yu, O. J. Yoo, and S. Kim. Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics*, 21(24):4348–4355, 2005.

[28] Han-Yu Chuang, Eunjung Lee, Yu-Tsueng Liu, Doheon Lee, and Trey Ideker. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3(140), 2007.

[29] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695, 2006.

[30] Xiangqin Cui and Gary A Churchill. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.*, 4(4):210, 2003.

[31] Daniel J Dauer, Bernadette Ferraro, Lanxi Song, Bin Yu, Linda Mora, Ralf Buettner, Steve Enkemann, Richard Jove, and Eric B Haura. Stat3 regulates genes common to both wound healing and cancer. *Oncogene*, 24:3397–3408, 2005.

[32] A. P. Davis, B. L. King, S. Mockus, C. G. Murphy, C. Saraceni-Richards, M. Rosenstein, T. Wiegers, and C. J. Mattingly. The Comparative Toxicogenomics Database: update 2011. *Nucleic Acids Res*, 39 (Database issue):D1067–72, January 2011.

[33] Alberto de la Fuente. From 'differential expression' to 'differential networking'– identification of dysfunctional regulatory networks in diseases. *Trends in genetics*, 26(7):326–333, 2010.

[34] Karen M. Doody, Annie Bourdeau, and Michel L. Tremblay. T-cell protein tyrosine phosphatase is a key regulator in immune cell signaling: lessons from the knockout mouse model and implications in human disease. *Immunol Rev*, 228(1):325–41, 2009.

[35] Konstantin H. Dragnev, Tian Ma, Jobin Cyrus, Fabrizio Galimberti, Vincent Memoli, Alexander M. Busch, Gregory J. Tsongalis, Marc A. Seltzer, David Johnstone, Cherie P. Erkmen, William Nugent, James R. Rigas, Xi Liu, Sarah J. Freemantle, Jonathan M. Kurie, Samuel Waxman, and Ethan Dmitrovsky. Combining bexarotene with erlotinib in window of opportunity and phase II trials causes lung cancer responses independent of KRAS mutations. In *Proceedings: AACR 102nd Annual Meeting 2011*, volume 71 of *Supplement 1*. Cancer Research, April 2011. Abstract nr LB-412.

[36] R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207–10, Jan 2002.

[37] H. Endoh, S. Tomida, Y. Yatabe, H. Konishi, H. Osada, K. Tajima, H. Kuwano, T. Takahashi, and T. Mitsudomi. Prognostic model of pulmonary adenocarcinoma by expression profiling of eight genes as determined by quantitative real-time reverse transcriptase polymerase chain reaction. *J. Clin. Oncol.*, 22(5):811–819, Mar 2004.

[38] Janine T. Erler and Rune Linding. Network Medicine Strikes a Blow against Breast Cancer. *Cell*, 149(4):731–733, 2012.

[39] S Falcon and R Gentleman. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257–8, 2007.

[40] Kristen Fortney, Max Kotlyar, and Igor Jurisica. Inferring the functions of longevity genes with modular subnetwork biomarkers of Caenorhabditis elegans aging. *Genome Biology*, 11(2):R13, 2010.

[41] L. C. Freeman. A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1):35–41, 1977.

[42] Tova F. Fuller, Anatole Ghazalpour, Jason E. Aten, Thomas A. Drake, Aldons J. Lusis, and Steve Horvath. Weighted gene coexpression network analysis strategies applied to mouse weight. *Mamm Genome*, 18(6-7):463–472, 2007.

[43] Harold N. Gabow and Robert E. Tarjan. A linear-time algorithm for finding a minimum spanning pseudoforest. *Information Processing Letters*, 27(5):259–263, 1988.

[44] Mitchell E. Garber, Olga G. Troyanskaya, Karsten Schluens, Simone Petersen, Zsuzsanna Thaesler, Manuela Pacyna-Gengelbach, Matt van de Rijn, Glenn D. Rosen, Charles M. Perou, Richard I. Whyte, Russ B. Altman, Patrick O. Brown, David Botstein, and Iver Petersen. Diversity of gene expression in adenocarcinoma of the lung. *Proc Natl Acad Sci USA*, 98(24):13784–9, 2001.

[45] Michael R. Garey and David S. Johnson. *Computers and Intractability-A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, 1979.

[46] Lewis Y. Geer, Aron Marchler-Bauer, Renata C. Geer, Lianyi Han, Jane He, Siqian He, Chunlei Liu, Wenyao Shi, and Stephen H. Bryant. The NCBI BioSystems database. *Nucleic Acids Res*, 38:D492–6, 2010.

[47] Joseph Geraci, Geoffrey Liu, and Igor Jurisica. Algorithms for Systematic Identification of Small Subgraphs. In Jacques van Helden, Ariane Toussaint, and Denis Thieffry, editors, *Bacterial Molecular Networks*, volume 804 of *Methods in Molecular Biology*, pages 219–244. Springer, 2012.

[48] L. Girard, J. D. Minna, W. L. Gerald, P. Saintigny, and L. Zhang. MSKCC-A Primary Lung Cancer Specimens. *Gene Expression Omnibus GSE31547*, 2011.

[49] Sergei Grivennikov and Michael Karin. Dangerous liaisons: STAT3 and NF-kappaB collaboration and crosstalk in cancer. *Cytokine Growth Factor Rev*, 21(1):11–19, February 2010.

[50] L. Guo, Y. Ma, R. Ward, V. Castranova, X. Shi, and Y. Qian. Constructing molecular classifiers for the accurate prognosis of lung adenocarcinoma. *Clin. Cancer Res.*, 12(11 Pt 1):3344–3354, Jun 2006.

[51] Douglas Hanahan and Robert A. Weinberg. Hallmarks of Cancer: The Next Generation. *Cell*, 144(5):646–674, March 2011.

[52] Teresa W. Haynes, Stephen T. Hedetniemi, and Peter J. B. Slater. *Fundamentals of domination in graphs*, volume 208. Marcel Dekker Inc., 1998.

[53] Sarah E. M. Herman, Amber L. Gordon, Erin Hertlein, Asha Ramanunni, Xiaoli Zhang, Samantha Jaglowski, Joseph Flynn, Jeffrey Jones, Kristie A. Blum, Joseph J. Buggy, Ahmed Hamdy, Amy J. Johnson, and John C. Byrd. Bruton tyrosine kinase represents a promising therapeutic target for treatment of chronic lymphocytic leukemia and is effectively targeted by PCI-32765. *Blood*, 117(23):6287–6296, 2011.

[54] Lee A. Honigberg, Ashley M. Smith, Mint Sirisawad, Erik Verner, David Loury, Betty Chang, Shyr Li, Zhengying Pan, Douglas H. Thamm, Richard A. Miller, and Joseph J. Buggy. The Bruton tyrosine kinase inhibitor PCI-32765 blocks B-cell activation and is efficacious in models of autoimmune disease and B-cell malignancy. *PNAS*, 107(29):13075–80, 2010.

[55] J. Hou, J. Aerts, B. den Hamer, W. van Ijcken, M. den Bakker, P. Riegman, C. van der Leest, P. van der Spek, J. A. Foekens, H. C. Hoogsteden, F. Grosveld, and S. Philipsen. Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS One*, 5(4), 2010.

[56] Lawrence Hunter. *Artificial intelligence and molecular biology*, chapter Molecular Biology for Computer Scientists. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1993.

[57] W.C. Hwang, A. Zhang, and M. Ramanathan. Identification of information flow-modulating drug targets: a novel bridging paradigm for drug discovery. *Clin. Pharmacol. Ther.*, 84(5):563–572, 2008.

[58] Trey Ideker, Owen Ozier, Benno Schwikowski, and Andrew F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18(Suppl. 1):S233–S240, 2002.

[59] Rubina S. Ismail, Rae Lynn Baldwin, Junguo Fang, Damaris Browning, Beth Y. Karlan, Judith C. Gasson, and David D. Chang. Differential gene expression between normal and tumor-derived ovarian epithelial cells. *Cancer research*, 60(23):6744–6749, 2000.

[60] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature Brief Communications*, 411:41–42, May 2001.

[61] Pall F. Jonsson and Paul A. Bates. Global topological features of cancer proteins in the human interactome. *Bioinformatics*, 22(18):2291–2297, 2006.

[62] F. Molnár Jr., Sameet Sreenivasan, Boleslaw K. Szymanski, and Gyorgy Korniss. Minimum Dominating Sets in Scale-Free Network Ensembles. *Scientific reports*, 3(1736).

[63] M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, 28:27–30, 2000.

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

[65] Dennis Kostka and Rainer Spang. Finding disease specific alterations in the co-expression of genes. *Bioinformatics*, 20(suppl 1):i194–i199, 2004.

[66] Max Kotlyar. *Prediction of protein-protein interactions and essential genes through data integration*. PhD thesis, University of Toronto, 2011.

[67] Max Kotlyar, Kristen Fortney, and Igor Jurisica. Network-based characterization of drug-regulated genes, drug targets, and toxicity. *Methods*, 57(4):499–507, 2012.

[68] Oleksii Kuchaiev, Aleksandar Stevanović, Wayne Hayes, and Nataša Pržulj. GraphCrunch 2: Software tool for network modeling, alignment and clustering. *BMC Bioinformatics*, 12(24), 2011.

[69] Christian Laforest and Raksmey Phan. Solving the Minimum Independent Domination Set Problem in Graphs by Exact Algorithm and Greedy Heuristic. *RAIRO-Operations Research*, 47(03):199–221, 2013.

[70] M. T. Landi, T. Dracheva, M. Rotunno, J. D. Figueroa, H. Liu, A. Dasgupta, F. E. Mann, J. Fukuoka, M. Hames, A. W. Bergen, S. E. Murphy, P. Yang, A. C. Pesatori, D. Consonni, P. A. Bertazzi, S. Wacholder, J. H. Shih, N. E. Caporaso, and J. Jen. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PLoS One*, 3(2), 2008.

[71] J. E. Larsen, S. J. Pavey, L. H. Passmore, R. Bowman, B. E. Clarke, N. K. Hayward, and K. M. Fong. Expression profiling defines a recurrence signature in lung squamous cell carcinoma. *Carcinogenesis*, 28(3):760–766, Mar 2007.

[72] J. E. Larsen, S. J. Pavey, L. H. Passmore, R. V. Bowman, N. K. Hayward, and K. M. Fong. Gene expression signature predicts recurrence in lung adenocarcinoma. *Clin. Cancer Res.*, 13(10):2946–2954, May 2007.

[73] S. K. Lau, P. C. Boutros, M. Pintilie, F. H. Blackhall, C. Q. Zhu, D. Strumpf, M. R. Johnston, G. Darling, S. Keshavjee, T. K. Waddell, N. Liu, D. Lau, L. Z. Penn, F. A. Shepherd, I. Jurisica, S. D. Der, and M. S. Tsao. Three-gene prognostic classifier for early-stage non small-cell lung cancer. *J. Clin. Oncol.*, 25(35):5562–5569, Dec 2007.

[74] T. P. Lu, M. H. Tsai, J. M. Lee, C.P. Hsu, P. C. Chen, C. W. Lin, J. Y. Shih, P. C. Yang, C. K. Hsiao, L. C. Lai, and E. Y. Chuang. Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women. *Cancer Epidemiol Biomarkers Prev*, 19(10):2590–7, Oct 2010.

[75] Y. Lu, W. Lemon, P. Y. Liu, Y. Yi, C. Morrison, P. Yang, Z. Sun, J. Szoke, W. L. Gerald, M. Watson, R. Govindan, and M. You. A gene expression signature predicts survival of patients with stage I non-small cell lung cancer. *PLoS Med.*, 3(12):e467, Dec 2006.

[76] D. Marcus and Y. Shavitt. {RAGE} – A rapid graphlet enumerator for large networks. *Computer Networks*, 56(2):810–819, 2012.

[77] O. Mason and M. Verwoerd. Graph theory and networks in Biology. *Systems biology, IET*, 1(2):89–119, March 2007.

[78] Lisa Matthews, Gopal Gopinath, Marc Gillespie, Michael Caudy, David Croft, Bernard de Bono, Phani Garapati, Jill Hemish, Henning Hermjakob, Bijay Jassal, Alex Kanapin, Suzanna Lewis, Shahana Mahajan, Bruce May, Esther Schmidt, Imre Vastrik, Guanming Wu, Ewan Birney, Lincoln Stein, and Peter D'Eustachio. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res*, 37(D619-22), 2009.

[79] Tijana Milenković, Vesna Memišević, Anthony Bonato, and Nataša Pržulj. Dominating biological networks. *PloS one*, 6(8):e23016, 2011.

[80] Tijana Milenković and Nataša Pržulj. Uncovering Biological Network Function via Graphlet Degree Signatures. *Cancer Informatics*, 6:257–273, 2008.

[81] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[82] H. Okayama, T. Kohno, Y. Ishii, Y. Shimada, K. Shiraishi, R. Iwakawa, K. Furuta, K. Tsuta, T. Shibata, S. Yamamoto, S. Watanabe, H. Sakamoto, K. Kumamoto, S. Takenoshita, N. Gotoh, H. Mizuno, A. Sarai, S. Kawano, R. Yamaguchi, S. Miyano, and J. Yokota. Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. *Cancer Res*, 72(1):100–11, Jan 2012.

[83] C. N. A. M. Oldenhuis, S. F. Oosting, J. A. Gietema, and E. G. E. de Vries. Prognostic versus predictive value of biomarkers in oncology. *European Journal of Cancer*, 44(7):946–953, 2008.

[84] Saeed Omidi, Falk Schreiber, and Ali Masoudi-Nejad. MODA: An efficient algorithm for network motif discovery in biological networks. *Genes Genet. Syst.*, 84(5):385–395, October 2009.

[85] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814–818, 2005.

[86] Mathew Penrose. *Random geometric graphs*, volume 5. Oxford University Press, 2003.

[87] A. Potti, S. Mukherjee, R. Petersen, H. K. Dressman, A. Bild, J. Koontz, R. Kratzke, M. A. Watson, M. Kelley, G. S. Ginsburg, M. West, D. H. Harpole, and J. R. Nevins. A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. *N. Engl. J. Med.*, 355(6):570–580, Aug 2006.

[88] N. Pržulj, D. G. Corneil, and I. Jurisica. Efficient estimation of graphlet frequency distributions in protein-protein interaction networks. *Bioinformatics*, 22(8):947–980, 2006.

[89] N. Pržulj, D.A. Wigle, and I. Jurisica. Functional topology in a network of protein interactions. *Bioinformatics*, 20(3):340–348, 2004.

[90] Natasa Pržulj. Graph theory analysis of protein-protein interactions. In Igor Jurisica and Dennis Wigle, editors, *Knowledge discovery in proteomics*, volume 8 of *Chapman and Hall/CRC Mathematical biology and medicine series*, pages 73–128. CRC Press Taylor and Francis Group, 2006.

[91] Nataša Pržulj. *Analyzing large biological networks: protein-protein interactions example*. PhD thesis, University of Toronto, 2005.

[92] Nataša Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.

[93] Nataša Pržulj. Erratum to Biological Network Comparison Using Graphlet Degree Distribution. *Bioinformatics*, 26(6):853–854, 2010.

[94] Nataša Pržulj, D. G. Corneil, and I. Jurisica. Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, 2004.

[95] Nataša Pržulj and Tijana Milenković. Computational methods for analyzing and modeling biological networks. In Jake Chen and Stefano Lonardi, editors, *Biological data mining*, pages 397–428. Chapman & Hall/CRC, 2009.

[96] Peng Qiu, Z. Jane Wang, K. J. Ray Liu, Zhang-Zhi Hu, and Cathy H. Wu. Dependence network modeling for biomarker identification. *Bioinformatics*, 23(2):198–206, 2007.

[97] M. Raponi, Y. Zhang, J. Yu, G. Chen, G. Lee, J. M. Taylor, J. Macdonald, D. Thomas, C. Moskaluk, Y. Wang, and D. G. Beer. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Res.*, 66(15):7466–7472, Aug 2006.

[98] Pedro Romero, Jonathan Wagg, Michelle L Green, Dale Kaiser, Markus Krummenacker, and Peter D Karp. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biology*, 6(R2), 2004.

[99] A. Sanchez-Palencia, M. Gomez-Morales, J. A. Gomez-Capilla, V. Pedraza, L. Boyero, R. Rosell, and M. E. Fárez-Vidal. Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *Int J Cancer*, 129(2):355–64, July 2011.

[100] Carl F. Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H. Buetow. PID: The Pathway Interaction Database. *Nucleic Acids Res*, 37:D674–9, 2009.

[101] Nicholas P. Shinners, Gianluca Carlesso, Iris Castro, Kristen L. Hoek, Radiah A. Corn, Robert T. Woodland, Martin L. Scott, Demin Wang, and Wasif N. Khan. Bruton's Tyrosine Kinase Mediates NF-kappaB Activation and B Cell Survival by B Cell-Activating Factor Receptor of the TNF-R Family. *J Immunol*, 179(6):3872–80, 2007.

[102] Ke Shuai and Bin Liu. Regulation of JAK-STAT signalling in the immune system. *Nature Reviews Immunology*, 3:900–911, 2003.

[103] Ivan Stojmenovic, Mahtab Seddigh, and Jovisa Zunic. Dominating sets and neighbor elimination-based broadcasting algorithms in wireless networks. *Parallel and Distributed Systems, IEEE Transactions on*, 13(1):14–25, 2002.

[104] L. Su, C. Chang, Y. Wu, K. Chen, C. Lin, S. Liang, C. Lin, J. Whang-Peng, S.Hsu, C. Chen, and C. F. Huang. Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme. *BMC Genomics*, 8(140), 2007.

[105] Z. Sun, D. A. Wigle, and P. Yang. Non-overlapping and non-cell-type-specific gene expression signatures predict lung cancer survival. *J. Clin. Oncol.*, 26(6):877–883, Feb 2008.

[106] S. Tomida, K. Koshikawa, Y. Yatabe, T. Harano, N. Ogura, T. Mitsudomi, M. Some, K. Yanagisawa, T. Takahashi, H. Osada, and T. Takahashi. Gene expression-based, individualized outcome prediction for surgically treated lung cancer patients. *Oncogene*, 23(31):5360–5370, Jul 2004.

[107] Wieland Voigt. Sulforhodamine B assay and chemosensitivity. In *Chemosensitivity*, volume 110 of *Methods in Molecular Medicine*, pages 39–48. Springer, 2005.

[108] B. H. Voy, J. A. Scharff, A. D. Perkins, A. M. Saxton, B. Borate, E. J. Chesler, L. K. Branstetter, and M. A. Langston. Extracting gene networks for low-dose radiation using graph theoretical algorithms. *PLoS computational biology*, 2(7), 2006.

[109] Michael Watson. CoXpress: differential co-expression in gene expression data. *BMC Bioinformatics*, 7:509, 2006.

[110] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *nature*, 393(6684):440–442, 1998.

[111] Sebastian Wernicke. Efficient Detection of Network Motifs. *IEEE/ACM transactions on computational biology and bioinformatics*, 3(4):347–359, October 2006.

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

[113] D. B. West. *Introduction to graph theory.* Prentice hall, Upper Saddle River, NJ, 2nd edition, 2001.

[114] Dennis A. Wigle, Igor Jurisica, Niki Radulovich, Melania Pintilie, Janet Rossant, Ni Liu, Chao Lu, James Woodgett, Isolde Seiden, Michael Johnston, Shaf Keshavjee, Gail Darling, Timothy Winton, Bobby-Joe Breitkreutz, Paul Jorgenson, Mike Tyers, Frances A. Shepherd, and Ming Sound Tsao. Molecular Profiling of Non-Small Cell Lung Cancer and Correlation with Disease-free Survival. *Cancer Res.*, 62(11):3005–8, 2002.

[115] Serene Wong, Nick Cercone, and Igor Jurisica. Characterizing healthy and disease states by systematically comparing differential correlation networks in lung. In *Advances in Health Informatics Conference*, 2012.

[116] Serene Wong, Max Kotlyar, Dan Strumpf, Nick Cercone, Frances A. Shepherd, Ming-Sound Tsao, and Igor Jurisica. Systematic, comparative network analysis on non-small cell lung cancer [abstract]. In *Proceedings of the 103rd Annual Meeting of the American Association for Cancer Research*, volume 72 of *Cancer Research*, page Abstract nr 4912, Chicago, Mar 31-Apr 4 2012.

[117] Jie Wu and Hailan Li. On calculating connected dominating set for efficient routing in ad hoc wireless networks. In *Proceedings of the 3rd international workshop on Discrete algorithms and methods for mobile computing and communications*, pages 7–14. ACM, 1999.

[118] R. A. Young. Biomedical discovery with DNA arrays. *Cell*, 102:9–15, 2000.

[119] Hua Yu, Marcin Kortylewski, and Drew Pardoll. Crosstalk between cancer and immune cells: role of STAT3 in the tumour microenvironment. *Nature Reviews Immunology*, 7:41–51, 2007.

[120] Bai Zhang, Huai Li, Rebecca B. Riggins, Ming Zhan, Jianhua Xuan, Zhen Zhang, Eric P. Hoffman, Robert Clarke, and Yue Wang. Differential dependency network

analysis to identify condition-specific topological changes in biological networks. *Bioinformatics*, 25(4):526–532, 2009.

Abbreviations

AACR	American Association for Cancer Research
ASCO	American Society of Clinical Oncology
CDIP	Cancer Data Integration Portal
CTD	Comparative Toxicogenomics Database
DMSO	Dimethyl sulfoxide
FDA	U.S. Food and Drug Administration
FDR	False Discovery Rate
GO	Gene Ontology
I2D	Interologous Interaction Database
IASLC	International Association for the Study of Lung Cancer
KEGG	Kyoto Encyclopedia of Genes and Genomes
NSCLC	Non-Small Cell Lung Cancer
PPI	Protein-protein interaction
wrt	with respect to

Glossary

Biological pathway

A *biological pathway* is the combination of actions in series among molecules to accomplish tasks such as triggering the assembling of new molecules, turning genes on and off, and can cause other changes in a cell. Some common types of biological pathways involved metabolism, gene regulation and signal transduction [2].

Carcinogenesis

The process in which cancer cells are transformed from normal cells. (`http://www.cancer.gov/dictionary` (Nov, 2013))

Cytotoxic

"Cell-killing". (`http://www.cancer.gov/dictionary` (Dec, 2013))

Deregulated subgraph

Subgraphs that are present in the tumor state, but are not present in the normal state.

Drug repositioning

Applying known drugs to new uses.

Gene expression profile or signature

A *gene expression profile or signature* describes a cell's molecular state in a specific condition [118].

Gene ontology

Gene ontology is a major bioinformatics initiative project that aims to standardize rep-

resentations of attributes on genes and gene products across databases [8].

Personalized molecular medicine

A medical model that customizes treatments to individual patients.

Predictive signature

In the context of oncology, a *predictive signature* gives information about therapeutic effect, and it can be a therapeutic target [83].

Prognostic signature

In the context of oncology, a *prognostic signature* gives information about the overall cancer outcome of a patient independent of therapy [83].

qPCR

Quantitative polymerase chain reaction (qPCR or real-time quantitative PCR) is a technique to sensitively quantify nucleic acids [7].

SRB assay

Sulforhodamine B (SRB) assay is a test used to measure cytotoxicity and cell proliferation caused by the application of drugs [107].