Computational Methods for Knowledge Integration in the Analysis of Large-scale Biological Networks

by

Archana Ramesh

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2012 by the
Graduate Supervisory Committee

Seungchan Kim, Chair
Patrick Langley
Chitta Baral
Jeffrey Kiefer

ARIZONA STATE UNIVERSITY

August 2012

ABSTRACT

As we migrate into an era of personalized medicine, understanding how bio-molecules interact with one another to form cellular systems is one of the key focus areas of systems biology. Several challenges such as the dynamic nature of cellular systems, uncertainty due to environmental influences, and the heterogeneity between individual patients render this a difficult task. In the last decade, several algorithms have been proposed to elucidate cellular systems from data, resulting in numerous data-driven hypotheses. However, due to the large number of variables involved in the process, many of which are unknown or not measurable, such computational approaches often lead to a high proportion of false positives. This renders interpretation of the data-driven hypotheses extremely difficult. Consequently, a dismal proportion of these hypotheses are subject to further experimental validation, eventually limiting their potential to augment existing biological knowledge.

This dissertation develops a framework of computational methods for the analysis of such data-driven hypotheses leveraging existing biological knowledge. Specifically, I show how biological knowledge can be mapped onto these hypotheses and subsequently augmented through novel hypotheses. Biological hypotheses are learnt in three levels of abstraction -- individual interactions, functional modules and relationships between pathways, corresponding to three complementary aspects of biological systems. The computational methods developed in this dissertation are applied to high throughput cancer data, resulting in novel hypotheses with potentially significant biological impact.

To

Grandpa

ACKNOWEDGEMENTS

Several faculty, friends and family members have helped me complete this dissertation and I would like to express my gratitude to them for their assistance.

I would like to begin by thanking my advisor, Seungchan Kim, for being a supportive advisor through out my graduate education. From coming up with an appropriate topic for my dissertation to helping me prepare for my defense, he has been instrumental in seeing me through completion of this program. I am grateful to him for his guidance, mentoring and most importantly, being available for discussions on an almost daily basis, in spite of his busy schedule. His emphasis on quality and excellence has made a significant impact on my work and I am indebted to him for this.

Special thanks to Pat Langley, member of my dissertation committee. He has worked with me patiently, reviewing my writing and providing me with several painstaking comments. While working on solutions towards a specific application (such as a biological problem) one often forgets the bigger picture. However, in my case, Dr. Langley's constant encouragement and critique helped me build a coherent dissertation with a larger goal in mind.

I would also like to thank the rest of my committee, Chitta Baral and Jeff Kiefer for meeting with me periodically and providing me valuable feedback towards this dissertation.

I am indebted to Sara Nasser for her mentoring and advice during the course of my PhD. She is an incredible researcher and her commitment and dedication never cease to inspire me.

I am also grateful to our lab members - Sungwon Jung, Ina Sen, Michael Verdicchio and Robert Trevino, not only in their contributions to my research but also in creating a friendly lab environment to occasionally vent steam about failed experiments.

Several faculty at ASU deserve credit in helping me through this journey, including Faye Navabi, Mutsumi Nakamura, Subbarao Kambhampati, George Runger, Goran Konjevod.

I would like to thank my friends - Karthik, Muthu, Jyothsana, Rajitha, Shantanu, Preetika, Harsha, Archana, Raghu, Bala, Michelle, Vasundhara, Mridul, Praveen, Aditya, Mike and Lily. Karthik, was a year ahead of me in the program, and has been a great source of advice for several crucial decisions I have had to make.

I would like to thank my family to whom I owe everything. My mother, with her immense strength, pragmatism, and wisdom, and my father, with his endurance, affection, and his sheer brilliance (something I doubt he himself is aware of!), have helped me become the person I am today. I also thank my grandparents for imbibing me with all their core values and my brothers, for keeping me on my toes so I could be someone they would one day look up to. Their unfaltering confidence in my abilities has seen me through many tough days. Finally, I would like to thank my husband for not only being a great friend, but also, one of my strongest advocates. His practical outlook, positivity, and ability to lighten even the most depressing moments has helped me bounce back many times.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Based on cancer incidence and mortality rates until 2007, the American Cancer Society predicted that, in 2012, in the US alone [1], a total of 577,190 patients would die due to cancer and 1,638,910 new occurrences of cancer would be diagnosed. This indicates that although molecular biology has seen progress in the last decade, we are still far from completely understanding the biological processes underlying disease. The key causes for this, include, underestimating the complexity of biological processes, heterogeneity among individual organisms, and the value of the genome being limited by its annotation.

The complexity of biological processes arises from the coordinated activity of biomolecules such as genes, proteins or complexes; heterogeneity, arises from differences across individuals in the activity of these biomolecules; and limitations of the genome arise from their functional diversity. Consequently, studying biological processes, benefits from a systems approach. For example, the PI3K pathway [2], an important pathway in cancer, regulates the signaling of multiple biological processes including apoptosis, cell proliferation, and cell growth. Being up regulated in cancer, this pathway is a promising target for therapy, as it is easier to inhibit activation than suppress tumor function [2]. However, the pathway functions through the interplay of biological interactions among several functionally diverse elements, each of which is regulated differently in distinct individuals. Clearly, a traditional reductionist approach which dissects this pathway into its constituent parts alone, would have difficulty in elucidating how it is regulated [3]. Instead it is important to adopt a systems approach to understand this pathway completely.

Systems biology views biological processes as an integrated network of genes,

proteins, and biomolecular interactions in continuous flux. Computational, statistical and mathematical methods play an important role in this paradigm by facilitating formal representations of complex biological systems. Additionally, they allow for the automated inference of models from experimental data that is frequently too large to manually analyze.

Recently, the popularity of high-throughput technology has led to not only a large amount of experimental data but also the growth of biological knowledge repositories. Such knowledge repositories (for example, Gene Ontology [4, 5] or Pathway Commons [6]) store facts about the functions of biomolecules and their interactions. In most cases, the facts are manually curated and associated with literature citations.

This dissertation develops a framework of computational methods to aid in hypothesis generation in systems biology using such available biological knowledge. Specifically, it focuses on hypotheses generated about one of the core problems of systems biology: how biomolecules interact in normal and diseased cellular systems.

The rest of this chapter is organized in a top-down manner. First, I place this dissertation within the broad context of research in systems biology. Subsequently, I provide an overview of the methods developed in this work. Finally, I outline the rest of the document.

1.1 Knowledge Discovery in Systems Biology

Before describing the framework developed in this dissertation, I first discuss the role of this dissertation in the workflows for scientific knowledge discovery. The two modes of scientific knowledge discovery, hypothesis-driven research and data-driven research, are illustrated in Figure 1.1. The structural components within these workflows are data and knowledge, two distinct yet complementary sources

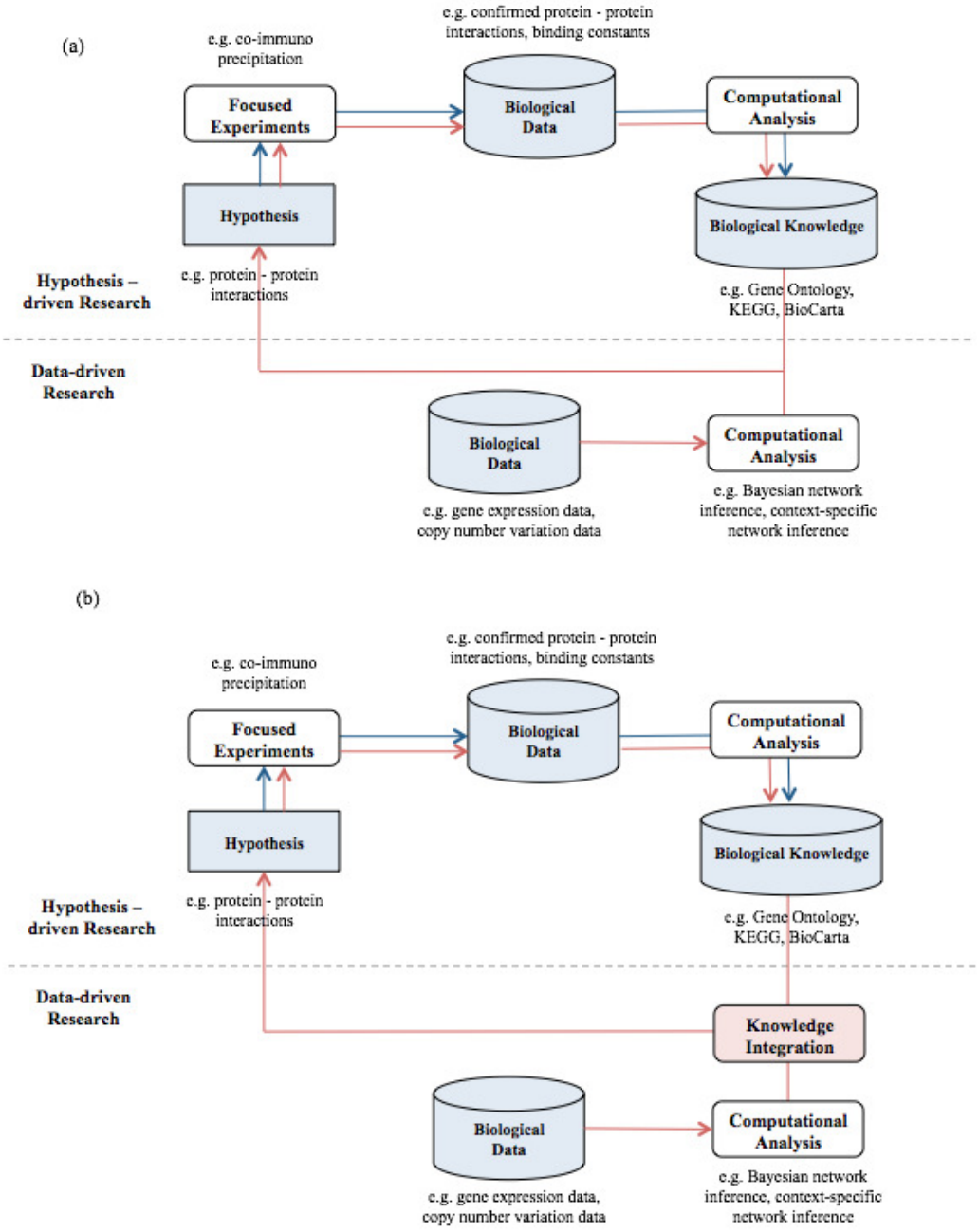

Figure 1.1: Knowledge discovery in systems biology.(a) Workflows for both hypothesis-driven research (indicated by blue arrows) and data-driven research (indicated by red arrows). (b) The role of knowledge integration (the focus of this dissertation) in the workflow for data-driven research.

of information. Data represents the collection of observations or the results of experiments, usually influenced by environmental factors. Data is collected either from focused individual experiments or from high-throughput experiments, shown in Figure 1.1. Examples of biological data include, gene expression profiles, copy number variations, and methylation data. Knowledge, on the other hand, represents validated facts about biological systems. Knowledge is usually manually curated with the help of several experts. Examples of biological knowledge bases include pathway repositories and drug target databases.

Traditionally, biological hypotheses are validated by a hypothesis-driven approach, through focused experiments, studying each hypothesis individually using experimental wet-lab techniques. Such hypotheses, usually formulated by domain experts, could take several forms, such as whether an interaction occurs between a pair of biomolecules or the biological processes that characterize disease sub-types. Validation of the hypothesis involves focused experiments, gathering of biological data and analysis of the data. Findings are usually disseminated through publications and, later, curated into biological knowledge repositories.

In contrast, over the last decade, the widespread availability of large amounts of high-throughput data has led to computational methods for inferring biological hypotheses directly from the data, as shown in Figure 1.1, by the red arrows. Such methods use an underlying mathematical framework, for example, a Bayesian network framework, to learn hypotheses from various kinds of biological data. In contrast to the hypotheses produced by domain experts, such hypotheses tend to be of the order of a few hundred to a few thousand in number. Additionally, since these hypotheses are derived by applying computational methods to high-throughput data (which is often noise-ridden), not all hypotheses are necessarily biologically plausible. Ideally, all data-driven hypotheses would be verified through further

biological experimentation. However, wet-lab experiments are expensive and time-consuming. Consequently, biologists need to manually sift through these hypotheses to determine which ones are plausible for focused experimental validation. Such manual analysis is a cumbersome process due to which only a handful of these hypotheses are validated and the rest are discarded.

In this dissertation, I develop a systematic framework to automatically prioritize such data-driven hypotheses using available biological knowledge. Biological knowledge allows for better interpretation of the data-driven hypotheses by first mapping them to what is currently known about biology and then augmenting what is known with plausible novel hypotheses. I focus on three different levels of biological organization – individual interactions, functional modules, and relationships between functional modules.

1.2 Problem Definition

Now I turn to the computational framework developed in this work (shown in Figure 1.2). This dissertation develops methods for refining data-driven hypotheses using existing biological knowledge at three layers of abstraction: individual interactions (described in Chapter 3), functional modules (described in Chapter 4), and relationships between biological processes (described in Chapter 5). The input and output of the framework are illustrated in Figure 1.2. Such hypotheses can then be validated by a focused hypothesis-driven experiment. The three levels of abstraction correspond to three complementary aspects of understanding biological systems, which, when put together, provide a global understanding of a biological system. In this dissertation, the three levels are treated as independent classes of problems. While this work focuses primarily on hypotheses learned from high throughput data, in theory, the hypotheses could be from any source.

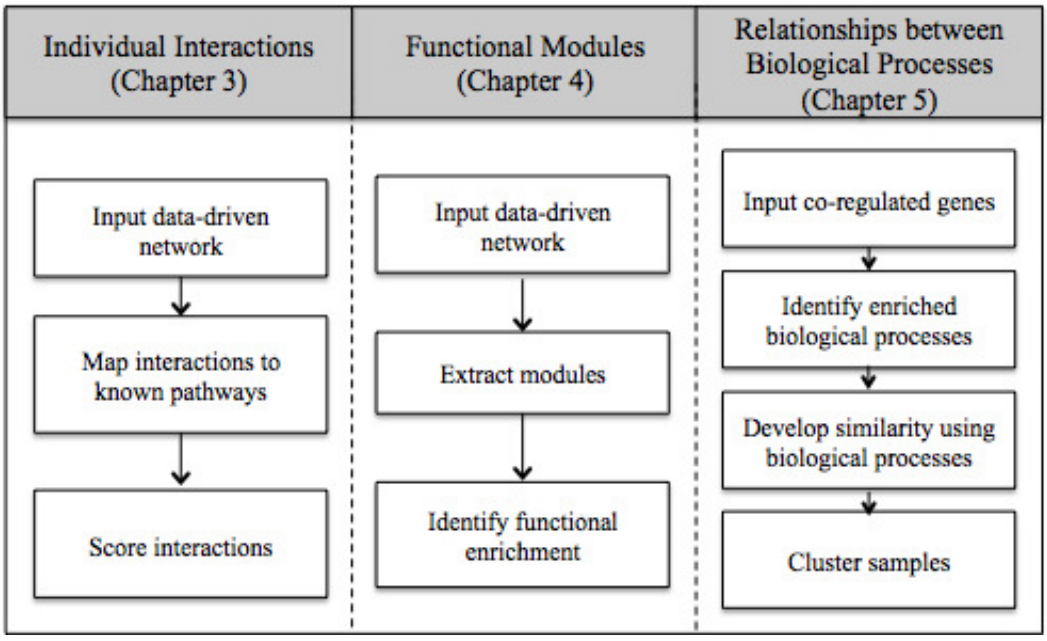

Figure 1.2: Computational framework for knowledge integration (developed in this dissertation).

Individual Interactions

I begin by focusing on hypotheses at the finest level of granularity - individual interactions. Chapter 3 develops methods to score the constituent interactions within a data-driven network to rank the most plausible interactions for further biological validation. This chapter focuses on hypotheses that take the form of data-driven interactions. Using literature-verified biological pathways and additional functional knowledge, the interactions are scored for further validation. High-scoring interactions can be immediately validated as they correspond directly to biological interactions.

Functional Modules

At the second level of granularity, functional modules constitute a set of biomolecular interactions that work towards a common biological purpose [7]. Such modules constitute an important level of biological organization, often corresponding to pathways. In Chapter 4, I focus on identifying functional modules within data-driven networks, using existing biological knowledge. As in the previous chapter, the input hypotheses take the form of data-driven networks of biomolecular interactions. Using sets of genes with pre-defined biological roles, the data-driven networks are partitioned into functional modules. Additionally, this chapter also overlays drug information in the form of therapeutic targets. Functional modules identified here can be used to understand the key biological processes that are active in a given set of samples. Although these functional modules cannot be directly experimentally validated, they provide a mechanism for biologists to focus on a specific set of biological hypotheses, where each hypothesis corresponds to a biomolecular interaction.

Relationships between Biological Processes

In contrast to the previous levels of granularity, at the third level of granularity, the hypotheses take the form of biological processes or functions. In Chapter 5, I use lists of co-regulated genes to identify the biological functions that are active within the data and their relationships across sub-sets of the data. Additionally, knowledge of existing relationships between the biological functions guides the process. While this chapter also provides biologists with specific sets of biological hypotheses to focus on, each hypothesis corresponds to a biological process that is active within the data. Hence, this chapter allows for identifying novel co-occurrence relationships

between biological processes.

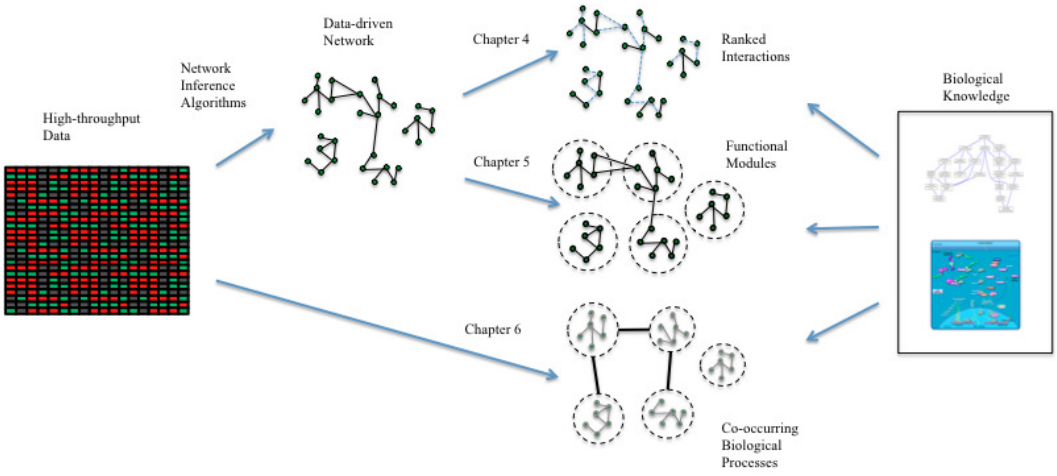

Figure 1.3: Overall inputs and outputs for the knowledge integration framework.

1.3 Dissertation Overview

The rest of this document provide the details of the methods for knowledge integration in the analysis of data-driven networks. Chapter 2 describes related research in this field. Chapter 3 describes methods for scoring individual interactions against a literature-verified database. Chapter 4 describes the methods proposed for the identification of functional modules from data-driven networks using biological gene sets. Chapter 5 focuses upon methods for learning relationships between biological processes active in the high-throughput data. Chapter 6 describes an example focused hypothesis-driven experiment, illustrating how the output of Chapters 3 - 5 could be validated. Finally, Chapter 7 concludes with a summary of the main contributions of the dissertation and directions for future research.

Chapter 2

RELATED RESEARCH

Cellular systems are complex biological processes and identifying the constituent interactions of such systems is one of the core problems of systems biology. As described in Chapter 1, solving this problem could use either a hypothesis-driven or a data-driven approach. However, both approaches rely on the experimental observation of biological processes for either formulation or validation of the hypotheses (in this case, the existence of specific interactions). Technological advances in the last decade allow for diverse experiments which shed light on the different aspects of biological processes. In this chapter, I describe the different kinds of biological data that are currently available. Following this, I outline existing research in methods for generating hypotheses from the data - specifically the inference of gene regulatory networks (used as the input to chapters 3 and 4) from data. Subsequently, I describe the different biological knowledge sources currently available. Finally, I provide an overview of existing work in knowledge integration.

2.1 Biological Data

Most biological experiments rely on the central dogma of molecular biology - the process by which cells control biological processes. Cells store hereditary information within their nucleus, in the form of strands of deoxyribonucleic acid (DNA). Strands of DNA are organized into genes - a molecular unit of heredity that codes for a specific function. While the process by which hereditary instructions are converted into cellular signals is complex, the core of this conversion occurs through two key steps called transcription and translation. In eukaryotes, DNA is converted into messenger ribonucleic acid (or mRNA) through a process called transcription; and mature mRNA is converted into proteins through a process called translation.

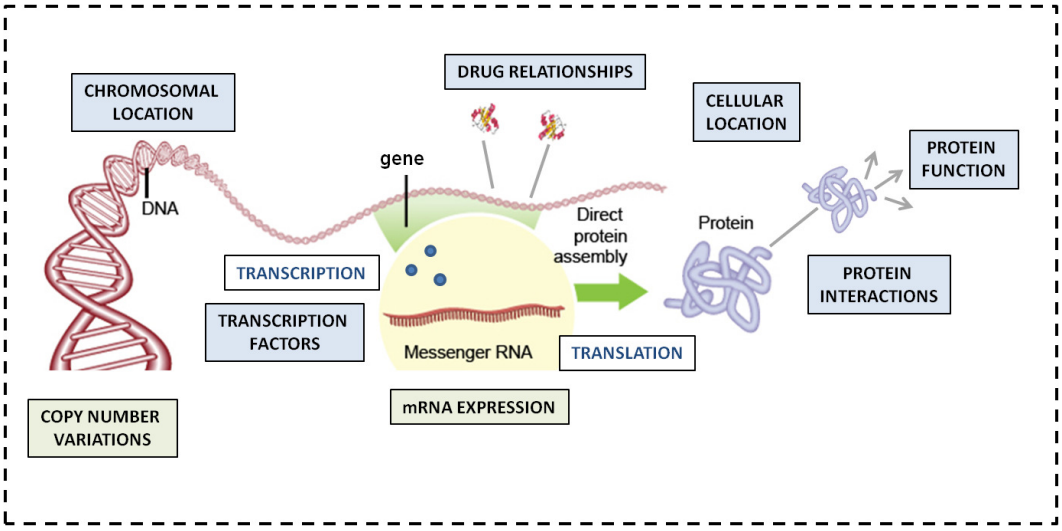

Figure 2.1: Biological data and knowledge in the context of a eukaryotic cell. Green rectangles indicate biological data while blue rectangles indicate biological knowledge.

Proteins then carry out the different functions required for maintaining the cell.

Since a normal individual contains specified amounts of each protein, variations in the amounts of proteins across different individuals could shed light on differences across individuals. As it is relatively easier to gather measurements of mRNA levels over protein levels, mRNA measurements are used in lieu of proteins. They are used to provide information about the role of genes (or proteins) within the cell including their functions, their location, and their relationships with other genes (or proteins) under different phenotypic conditions. Such measurements are called gene expression measurements. Traditionally, gene expression measurements are made by Northern blot experiments. However, recent developments in array-based technology allows the simultaneous hybridization of mRNA to a large number of DNA sequences.

Hybridization is usually measured through fluorescence; hence, a typical microarray experiment results in images which need to be analyzed to identify the arrayed

spots and measure relative fluorescences for each element [8]. Raw gene expression measurements are usually obtained as a ratio of the fluorescences of the target sample to a control. Continuous gene expression measurements can be discretized to fall into three categories – up-regulated (when the expression of the target is higher than the control), down-regulated (when the expression of the target is lower than the control) and neutral (when the expression of the target equals the control).

Abnormal gene expression could be attributed to several reasons and alterations in the DNA copy number is one such reason [9]. Comparative genomic hybridization [10] was the first method developed to measure copy number variations (CNV). Typically, the total genomic DNA is isolated from both a target and a reference population, differentially labelled and hybridized to metaphase chromosomes [9]. Recently, DNA microarrays have been used to achieve comparative genomic hybridization on a genome-wide scale. CNVs represent alterations of fragments of DNA resulting in an abnormal number of copies of the same fragment. With around 12 % of the human genome being susceptible to CNVs [11], CNVs play an important role in the differences between individuals, and thus, also in determining the biological mechanisms behind diseases.

Due to variations in equipments and measurement technologies, high-throughput data is typically pre-processed before any computational analysis. Quackenbush *et. al.* provides a review of the methods popularly used to pre-process high-throughput data [8].

2.2 Inferring Regulatory Networks from High-throughput Data

The large amount of biological data being currently generated has paved the way for several data-driven methods for inferring regulatory networks from this data (reviewed in [12, 13, 14]). These methods differ on the basis of the computational

definition of a biological interaction. The simplest regulatory model - the Boolean network represents genes as discrete ON/OFF switches [15, 16] with regulation modeled as a combination of logical operations. Liang *et. al.* [17] introduced REVEAL, which uses information-theoretic principles to reduce the search space and constructs a large-scale Boolean network from data. Following this, several variations of algorithms [18, 19, 20] have been proposed for inferring Boolean networks from data. Although Boolean networks provide a simple model for regulation, and are useful to understand steady states and network robustness, Boolean networks are known to possess several drawbacks including failing to cope with the dynamics and uncertainty inherent in biological regulation.

Probabilistic Boolean networks (PBN), introduced by Shmulevich *et. al.* [21] incorporates uncertainty and incomplete evidence in the gene regulatory network model by representing each regulatory relationship with several logical functions, each of which is associated with a probability based on data. PBNs have been used in several applications, including constructing a 15 gene sub-network inferred from human glioma expression data [22, 23]. Bayesian networks [24] represent another important class of probabilistic graphical models used in modeling gene regulatory networks and have been extensively applied in genomics [25, 26, 27, 28]. Key strengths of the Bayesian approach include its ability to handle incomplete data, avoid over-fitting, infer causal relationships and encode domain knowledge into the learning framework.

Besides regulatory models, several association-based methods have recently gained popularity in the inference of gene regulatory networks from data. Correlation measures are used to capture the strength of association between two genes, and consequently learn an interaction network [29]. Mutual information is another popular measure extensively used in learning gene interactions [30, 31, 32, 33].

However, while these methods have made progress in providing mechanisms for learning gene regulatory networks from data, they still fail to cope with the heterogeneity inherent in biological data. Several methods have been recently developed to identify sub-type specific modules from heterogeneous data [33, 34, 35, 36, 37]. COALEASCE [35] integrates both gene expression and sequence data to discover regulatory motifs. On the other hand, CONEXIC [36] uses genes with copy number aberrations as putative drivers of gene expression to identify modules. Mukherjee *et. al.* [37] have developes a network clustering approach based on the theory of sparse Gaussian Markov Random Fields, to identify subtypes differing in terms of network phenotype. The concept of a regulatory module (analogous to a bicluster [38] – a sub-set of consistently regulated genes within a sub-set of samples or conditions) was first introduced by Ihmels *et. al.* [39] who has proposed an algorithm which uses a set of seed genes to extract tightly regulated transcription modules from biomedical data. The notion of learning regulatory modules with associated experimental conditions has been introduced by Segal *et. al.* in module networks [40]. A module network [41] is a probabilistic model consisting of modules of co-regulated genes and corresponding regulatory programs. Module networks have been shown to extract regulatory relationships in several biological applications [42, 43, 44]. Similarly, integrative Bayesian network approaches have also been developed in order to learn regulatory networks from data [35] as well as identify driver mutations and the biological processes [36]. Context-specific gene regulatory networks [45, 46] provide a mechanism to learn regulatory relationships between genes using probabilistic measures of consistency.

While these approaches deal with the heterogeneity of the data, as described in chapter 1, the size of these data-driven networks tends to be of the order of several thousand interactions, rendering manual interpretation and biological experimen-

tation for each interaction nearly impossible. Most existing methods focus on the evaluation of a few interactions, resulting in a large number of underutilized hypotheses. Table 2.1 shows some of the popular methods for inferring gene regulatory networks from data and the extent to which the proposed hypotheses (interactions) have been validated. As seen in the table, the field lacks a systematic approach for using biological knowledge in determining plausible biological interactions.

2.3 Biological Knowledge

I now describe the different sources of biological knowledge that could be used in analysis of data-driven networks. Biological knowledge consists of validated facts about biological processes and is typically either manually curated or automatically curated and then manually verified.

Gene Ontology

I begin with one of the most popular knowledge sources - the Gene Ontology (GO). The Gene Ontology(GO) Consortium [4, 5] provides a structured, common, controlled vocabulary for defining the roles of genes and gene products in any organism. The roles of genes and gene products are organized through three independent ontologies : biological process, molecular function and cellular component.

An ontology term is categorized as a biological process when it represents a biological objective to which a gene or gene product contributes (e.g., 'translation'); as a molecular function, when it represents the biochemical activity of a gene product (e.g., 'enzyme') and as a cellular component, when it represents the place in the cell where the gene product is active (e.g., 'ribosome').

The Gene Ontology Annotation (GOA) database [47] provides electronic and manual

Table 2.1: Existing approaches for knowledge integration in interpreting data-driven networks.

Publication	Method	Dataset	Network Scale	Knowledge Integration
Friedman *et. al.* (2000)	Bayesian Networks	Saccharomyces cerevisiae (76 expression measurements; 6177 ORFs)	800 genes; 250 genes used for robustness analysis	Literature validation of few genes selected by a network property.
Ong *et. al.* (2002)	Dynamic Bayesian Networks	E.coli, time series gene expression data (12 data points)	169 genes	Prior knowledge through operons. Literature validation of a few genes.
Hartemik *et. al.* (2002)	Bayesian Networks	Saccharomyces cerevisiae (320 samples)	32 genes + 1 extra variable	Literature validation of a few relevant genes.
Lahdesmaki *et. al.* (2003)	Boolean networks	Saccharomyces cerevisiae gene expression time-series	733 genes; 5 genes studies	Literature validation of a few relevant genes.
Shmulevich *et. al.* (2003)	Probabilistic Boolean networks	Human glioma	15 genes	Literature validation of a few relevant genes.
Lee *et. al.* (2006); Li *et. al.* (2007)	Module networks, Bayesian likelihood score	Saccharomyces cerevisiae (320 samples)	466 candidate regulators, 2355 genes from gene expression,	GO, MIPS, KEGG, TRANSFAC
Kim *et. al.* (2007); Sen et. al (2009)	Context-specific GRN	Target Now Data(17085 probes, 146 samples)	1,790 vertices, 9566 edges	Literature validation of a few relevant genes.

annotations corresponding to each protein in the UniProt Knowledgebase. A GO annotation is a specific association between a GO term identifier and a gene or protein and has a distinct evidence source that supports the association. Each gene product can be annotated to multiple GO terms at different levels of the GO hierarchy.

Pathway Databases

While the Gene Ontology provides information on the functions and locations of biomolecules within the cell, often it is useful to integrate knowledge on their interactions. Biological pathway databases are a popular mechanism for storing information on interactions between biomolecules. Pathguide [48] lists 298 biological pathway resources including protein-protein interactions, metabolic pathways, signaling pathways, pathway diagrams, transcription factors, gene regulatory networks, protein-compound interactions, genetic interaction networks, and protein-sequence focused resources. This dissertation uses Pathway Commons [6] as well the as the the pathway component of the Molecular Signature Database (MSigDB) [49]. Pathway Commons consists of nine different pathway databases. Table 2.3 provides a description of these pathway databases along with the benefits and shortcomings of each resource.

From 2.3, three main aspects are identified that could limit the application of these databases to data-driven networks - species and reliability.

In terms of species, most of the databases consist of well-documented yeast interactions. Human interaction databases mainly consist of Cancer Cell Map, HumanCyc, Human Protein Reference Database and IMID. For the remaining databases the proportion of human interactions within the total database varies between 10% and 50%. Given the relatively low overlap between human and yeast interactions (both orthologs and number of interactions), applicability of yeast interactions to validate

Table 2.2: List of pathway resources in Pathway Commons.

Pathway Resource	Proteins	Interactions	Description	Benefits (+) and Shortcomings (-)
BioGrid [50]	36,196	192,369	protein-protein interactions, genetic interactions	consolidates interactions from different methods both experimental and computational (+); literature sources available for all interactions (+); human interactions constitute 46% of total (-); high-throughput interactions ($\approx$ 60% of total) could be unreliable (-)
Cancer Cell Map [51]	1,245	2,104	protein-protein interactions, signaling pathways	mostly human interactions (+); largely in vivo, in vitro with few yeast-2-hybrid interactions (+); cancer-specific (+); cellular location available (+); small database (-)
HumanCyc [52]	3,999	5,270	metabolic pathways, signaling pathways	human interactions (+); curation level for human interactions is limited (-)
IMID [53]	1,669	1,729	neuronal signaling	human interactions (+); interaction type lacks detail (-); small database (-); specific to neuronal signaling (-)
Human Protein Reference Database [54]	9,871	40,618	protein-protein interactions, complexes, post-translational modifications	in vivo, in vitro and yeast-2-hybrid interactions (+); both direct and complex interactions (+); large number of literature citations (+); interaction type lacks experimental details (-); large overlap with other databases (-)

IntAct [55]	132,806	157,344	protein-protein interactions	detailed experimental methods (+); human interactions only 20 % of total (-)
MINT [56]	92,866	121,824	protein-protein interactions	direct and indirect interactions (+); interactions from several detection methods (+); human interactions only ≈25 % of total (-); high-throughput interactions (≈80 %) could be unreliable (-)
NCI/Nature Pathway Interaction Database [57]	6,186	13,879	signaling pathways, protein-protein interactions	cancer-specific(+); evidence codes (+)
Reactome [58]	9,003	6,139	metabolic pathways, signaling pathways	curates processes with several levels of detail (+); curated manually by peer-review(+); isoform-specific (+); human reactions form only ≈ 10 % of total (-)

interactions from human datasets is limited. Further differences include the total possible interactions in each of the species. For instance, the full yeast protein-protein interaction network contain 37,800-75,500 interactions and the full human network 154,000-369,000, but owing to a high false-positive rate, maps (as of 2006) are roughly only 50% and 10% complete, respectively [59]. Thus, in this dissertation, a sub-set of the database corresponding to human interactions alone was used.

In terms of reliability, interactions within the databases are obtained using several sources - biological experimentation including yeast two-hybrid (Y2H) experiments, tandem affinity purification, as well as computational methods. In general, the interactions which have been verified through manual biological experimentation are known to be more reliable than data-driven interactions. However, even amongst the biological interactions, interactions derived using high-throughput methods such as Y2H screening are characterized by a large number of false positives. For instance, the reliability of Y2H screening mechanisms has been shown to be $\approx$ 50 % [60] using cellular localization and cellular role properties. Amongst the databases, both MINT and BioGrid contain a large proportion of high-throughput interactions (explicitly specified), while Cancer Cell Map and the Human Protein Reference Database contain interactions verified by Y2H screening, which again could be unreliable. However, in this dissertation, interactions were not filtered based on the reliability of the experimental detection method, instead, reliability (through the number of publication counts) is used to weight interactions, since the number of human interactions currently available for use is already limited.

Finally, it is also important to acknowledge the multiple types of interactions within the databases. Methods for the inference of biological interactions from high-throughput data usually derive co-expression based interactions. Such interactions could be attributed to multiple reasons, including protein-protein interactions as well

as transcription factors or regulatory interactions. While this dissertation focuses on protein-protein interactions as a possible explanation, multiple other explanations are possible (including transcription factors) and are beyond the scope of this work.

Chromosomal Location

Genes which are located close to one another tend to share functions making chromosomal location a useful resource for understanding the role of genes within the cell. Chromosomal location information was extracted from the Molecular Signatures Database (MSigDB) [49]. Such information is helpful in identifying effects pertaining to chromosomal amplifications or deletions, and epigenetic silencing. For instance, cytogenetic abnormalities have been attributed to diseases such as leukemia [61]. Similarly, chromosomal location is particularly appropriate in the context of copy number variations.

Drug Databases

Finally, one of the important aspects of understanding the biology behind disease is understanding the role of therapeutic agents. To this end, the publicly available drug database Drugbank [62] is used. Drugbank is a publicly available database consisting of information on the nomenclature, ontology, chemistry, structure, function, action, pharmacology, pharmacokinetics, metabolism and pharmaceutical properties of both small molecule and large molecule drugs, along with target diseases, proteins, genes and organisms on which these drugs act. Layering drug information allows for the ability to identify drug targets from the data.

2.4 Techniques for Knowledge Integration

I now describe the techniques which use these knowledge sources in the analysis of data-driven hypotheses. The methods can be broadly categorized on the basis of the data-driven hypotheses they use as input.

The first category of techniques is devoted to the annotation of interesting gene lists (results from the analysis of a high throughput dataset) with relevant biological processes. The key characteristic of gene set annotation approaches is that knowledge is usually used in the form of gene sets. Such methods [63, 64, 65, 66, 67, 49, 68, 66, 67, 49, 69] tend to use statistical hypothesis testing in the extraction of relevant biological functions. Typically, the output takes the form of a list of overrepresented biological functions. However, such methods have limited application as they can be applied to simply gene lists, where the connectivity between the genes is ignored.

The second category of techniques focuses on data-driven networks and are relatively few in number. In most cases, manual verification is still used, as shown in Table 2.1. Manual verification has been used in validating Bayesian networks learnt from yeast [70], mutual-information based networks derived from human B cells [32], in analyzing relevance networks learnt from cancer cell lines [31] and in validating context-specific networks learnt from both a melanoma dataset [45] and a refractory cancer dataset [46]. However, manually extracting relevant supporting literature for the validation of networks derived from high-throughput data is cumbersome. Further, given the scale of the data-driven networks, only a small number of interactions can be manually validated, even though validating the entire network could yield several novel hypotheses on biological interactions.

In contrast, a few systems have been developed for the automatic validation of data-

driven hypotheses. For example, Bayesian networks derived from yeast have been validated using the Kyoto Encyclopedia of Genes and Genomes (KEGG) [71]. Bader *et. al.* [72] have compared data-driven interaction networks against experimentally-derived networks using a global approach. However, most data-driven networks learnt from high-throughput data require validation at a finer level of granularity as validation is with the view of prioritizing interactions for further experimental verification. To achieve this, Hanalyzer [73] builds knowledge networks based upon multiple sources of interaction evidence (e.g., protein-protein interaction databases) and uses a noisy-OR approach to validate networks derived from experimental data against knowledge networks. Interactions are then scored based on available knowledge and thus, Hanalyzer facilitates the development of novel hypotheses using the combined analysis of knowledge and data, though applications are currently limited to mice.

Techniques devised for yeast and mice are not always applicable to humans as described in the previous section. Additionally, one of the main characteristics of data-driven interactions is that they capture associational relationships – thus the underlying biological interactions that they map onto could be a single interaction or in most cases, a series of interactions.

2.5 Summary

Thus, this chapter provides the background required for the rest of this dissertation. I have described the different kinds of biomedical data used, the methods used to infer regulatory networks from this data, limitations of these methods, biological knowledge sources and approaches for using biological knowledge in the interpretation of data-driven hypotheses. In the next chapter, I move to the approach developed for validating data-driven networks at the individual interaction level of granularity.

Chapter 3

EVALUATING INDIVIDUAL INTERACTIONS

One of the first steps in understanding data-driven interactions is understanding how they map on to known biological pathways. As described in Chapter 2, existing methods for automatically scoring data-driven interactions using biological knowledge are few. Here I discuss the motivation for scoring data-driven interactions, relevant work in this area, the methodology developed in this dissertation and its application to data-driven networks inferred from gene expression data for cancer patients.

3.1 Motivation

Data-driven interactions provide a mechanism to hypothesize regulatory interactions that occur in a specific experimental condition, reflected in the data. However, not all data-driven interactions necessarily correspond to regulatory interactions due to the presence of experimental noise within the data. Instead, data-driven interactions can comprise three types of interactions – known biological interactions, novel biological interactions and false interactions. Consequently, it is necessary to identify these three categories within the data-driven network to effectively utilize these interactions.

Mapping data-driven interactions to known biological interactions allows a biologist to associate the network with a measure of credibility. For instance, when faced with needing to choose between two data-driven networks, the one which has a higher number of edges with reliable explanations would be more likely to augment biological knowledge than the one with fewer edges having less reliable explanations. This measure serves as an indication to how well the network is able to capture

biological interactions which are already known.

However, the true contributions of data-driven networks stem from their ability to hypothesize novel interactions. Hence, a second scoring metric is developed which assigns to each edge the likelihood of that edge being a novel biological interaction. High-scoring interactions subsequently serve as candidate hypotheses for wet-lab experimentation.

3.2 Relevant Work

While Chapter 2 provides an overview of the approaches which motivated this dissertation, here I focus on literature relevant to scoring biological interactions. Most existing literature is devoted to protein-protein interactions and could be categorized into two – methods which score individual protein-protein interactions and methods for extracting pathway structures from protein interactions.

Scoring protein-protein interactions is usually achieved using either biological knowledge or network properties or both. Several kinds of biological knowledge are used for scoring interactions including protein structure information [74], GO annotations [74], expression information [75], sequence homology [74] and the existence of paralog interactions [75]. Methods that use network properties can be categorized based on whether they use local properties, global properties or both. Examples of local properties include proportions of shared neighbors [76, 77] and the extent to which a candidate interaction's neighbors are connected [78]. On the contrary, Saito *et. al.* [79] propose a global approach to rank interactions by classifying interactions into specific categories and using the proportions of interactions in each category to assign reliability scores. Both local and global properties have been combined and used by Liu *et. al.* [80]. Additionally, methods have also used shortest-path approaches [81] as well as Bayesian evidence combination in determining the

reliability of interactions [82].

A few automatic systems have been developed for the the inference of regulatory pathways from protein-protein interaction networks. For instance, Herrgård *et. al.* [83] have evaluated data-driven measurements (not necessarily interactions) in *Escherischia coli* and yeast, using metrics designed to measure the consistency between expression measurements and literature-derived interactions. Similarly, NetSearch [84] has been developed to enumerate and rank linear pathways using gene expression profiles of pathway members. Extending this concept to higher order structures (trees and parallel paths), Scott. *et. al.* [85] have developed a color-coding technique to identify pathways within protein interaction networks. This has been extended to generalized sub-structures by Lu. *et. al.* [86] using a randomized divide-and-conquer scheme. PathFinder [87] uses a combination of knowledge integration and association rule mining to elucidate pathways from protein interactions. Finally, Gitter *et. al.* [88] develop a method to elucidate pathways from protein interaction networks by formulating the problem as an edge orientation problem and focusing on directionality as an important aspect of pathways.

While the approaches developed for protein interaction networks are comparable, such approaches cannot be directly applied to data-driven networks. Protein-protein interactions are the results of direct measurements of protein concentrations, whereas, gene expression is measured at the mRNA level (described in Chapter 2). Hence, data-driven networks inferred from gene expression data tend to represent associational relationships, which could correspond to a direct relationship between the genes (and/or proteins) or, as in most cases, an indirect relationship between the genes (and/or proteins) involved.

In this dissertation, I develop a method to map data-driven interactions onto literature-derived pathways and subsequently score the unmapped interactions using other

biological knowledge sources. The methods developed in this chapter make several contributions to the analysis of data-driven interactions. Firstly, the methods developed here allow for direct interpretation of the data-driven interactions that do map onto biological knowledge. Secondly, the reliability of literature-derived pathways is taken into account while mapping data-driven edges. Thirdly, the scoring method for unmapped interactions allows for the extraction of the best candidate hypotheses from the data-driven interactions which could be validated with wet-lab experiments. Finally, these methods allow for comparisons between different kinds of data-driven networks to determine which one is better at identifying biologically plausible hypotheses.

3.3 Problem Formulation

In this section, I mathematically formulate the problem of assessing data-driven interactions using existing knowledge. Let the data-driven network be represented as $G_d = (V_d, E_d)$ where V_d is the set of genes within the network, E_d is the set of interactions between these genes and associated with each edge is a non-negative edge-weight $w_g : E_d \rightarrow \mathbb{R}^+$ denoting the strength of relationship between the two genes.

Let biological knowledge be represented in the two forms

(a) biological pathways represented as a network $G_b = (V_b, E_b)$ where V_b is the set of genes within the network, E_b is the set of interactions between these genes and associated with each edge (u_b, v_b) is a set of distinct literature citations $C(u_b, v_b) = \{c^1_{uv}, c^2_{uv}, \cdots c^m_{uv}\}$ that report the interaction and,

(b) functional annotations $T(u) = \{t_1, t_2, \cdots t_r\}$, a set of functional terms or pathways associated with the gene u.

The goal of this work is to identify for every edge $(u_d, v_d) \in E_d$,

1. when there is a mapping of the edge to literature-derived pathways, the score $\alpha \rightarrow \mathbb{R}^+$ quantifying the extent to which the edge could be reliably mapped onto existing literature-derived pathways, and

2. when there is no mapping of the edge to literature-derived pathways, $\beta \rightarrow [0,1]$, a score quantifying the extent to which the edge is likely to be a novel biological interaction.

3.4 Identifying Biological Explanations

The first step in assessing data-driven interactions is understanding how they map onto literature-verified pathways [89].

Methodology

Literature-derived pathways are composed of different interactions, each of which could be reliable based on the extent to which it has been studied and experimentally validated. Hence, I first need to define what a reliable interaction is in a literature-derived pathway [89]. As described above, each literature-derived edge (u_b, v_b) is associated with a set of distinct literature citations $C(u_b, v_b) = \{c^1_{uv}, c^2_{uv}, \cdots c^m_{uv}\}$ that report the interaction. The higher the number of citations, the more reliable an interaction is as it implies that several independent groups have been successful in experimentally confirming the interaction.

Definition 1 (**Citation Weight**). The citation weight, $w(u_b, v_b)$ is the reliability associated with a single literature-derived interaction.

Given an edge (u_b, v_b), with distinct literature citations $C(u_b, v_b) = \{c^1_{uv}, c^2_{uv}, \cdots c^m_{uv}\}$, where $m = |C(u_b, v_b)|$ the citation weight is defined as

$$w(u_b, v_b) = \begin{cases} 4 & \text{if } m = 0 \\ 2 & \text{if } m = 1 \\ \frac{1}{log_2(m)} & otherwise \end{cases} \tag{3.1}$$

Using equation 3.1, interactions with lower values of $w(u_b, v_b)$ are more reliable and interactions with higher values of $w(u_b, v_b)$ are less reliable. When m is greater than 1, I take the log-transform to appropriately scale larger number of citations, as after reaching a certain number of citations, more evidence is not necessarily better. When $m = 1$, I assign a weight of 2 to indicate this is equivalent to being half as good as an interaction with 2 citations and similarly assign a weight of 4 for the case when $m = 0$.

After reliability values have been assigned to every edge within the literature-derived network $G_b = (V_b, E_b)$, Djikstra's shortest-path algorithm [90] is applied to every edge within the data-driven network $G_d = (V_d, E_d)$.

Definition 2 (Literature Path Reliability). The literature path reliability, α is the reliability associated with the literature path corresponding to a data-driven interaction (u_d, v_d). This is computed as

$$\alpha = \begin{cases} \sum_{(a,b) \in S} w(a,b) & \text{when } S \neq \emptyset \\ \inf & \text{when } S = \emptyset \end{cases} \tag{3.2}$$

where S is the set of edges in the shortest-path connecting u_d with v_d in G_b. The distribution of α is useful in comparing different data-driven networks and the

following sub-section examines how data-driven networks of a biological data set evaluate against literature-derived pathways.

Application: Glioblastoma Multiforme (TCGA)

In order to demonstrate the application of the methods developed in this chapter, I used the Cancer Genome Atlas (TCGA) - Glioblastoma Multiforme (GBM) data [91]. 301 samples were extracted from the TCGA Portal after screening out samples from cell lines and replicates. 10 normal samples were used as the reference to convert GBM expression values to z-score values by comparing the expression values from GBM samples to the distribution of normal samples. Genes with a low variance across the tumors were filtered out leaving a total of 4166 genes.

While Chapter 2 discusses several existing methods for inferring data-driven networks, in this chapter, I choose three of the algorithms as representatives of three classes of methods – Bayesian networks (directed, probabilistic), ARACNE (undirected, mutual information based) and context-specific networks (directed, incorporates context-specificity).

In the case of the Bayesian networks and the context-specific networks, the expression data was quantized into one of three discretized values $(+1, 0, -1)$ by thresholding the z-score at 1.65 corresponding to 95% significance, while in the case of ARACNE, the continuous values were used as is.

Bayesian Networks : I use the Bayesian networks learnt using the algorithm BANJO, an algorithm which uses a structured iterative learning strategy to find the best possible network given an initial network [92]. The Java implementation on BANJO 2.0 [92] was used to learn the Bayesian network after discretizing the transformed data to three levels. BANJO uses simulated annealing to heuristically search for initial networks. A posterior averaged weighted 'consensus' network was generated.

Context Specific Networks : Context specific gene regulatory networks used here are learnt from gene expression data using the cellular context mining algorithm [46]. Unlike conventional gene regulatory networks, edges in context-specific GRNs represent the interaction conditioned on a subset of samples, i.e. *their biological context*, thus lending adaptability to the model of biological regulation. The parameters for learning the network were set at a maximum crosstalk of 0.3, conditioning of 0.1 and a corrected p-value of 0.05 for extracting context motifs. Subsequently the context-specific gene regulatory network was created using a statistical significance threshold of a corrected p-value of 0.005.

ARACNE : I use the ARACNE algorithm [32] to learn gene regulatory networks. ARACNE learns biological networks from high-throughput data using mutual information between pairs of genes to estimate the strength of the relationship between every pair of genes. ARACNE was also applied to this data using the recommended parameters – a Data Processing Inequality (DPI) threshold of 0.15, a significance p-value of $1 \times e^{-7}$ along with a mutual information threshold of 0.65. Since ARACNE produces an undirected network, the edges yielded by ARACNE were converted to directed edges; for instance an undirected edge (u, v) would correspond to the directed edges (u, v) and (v, u).

Statistical parameters in all three cases were set to obtain a similar number of interactions in all three networks.

Comparing Data-driven Networks Against Literature-derived Pathways

The first set of experiments was to compare how well the data-driven networks constructed from three different algorithms are supported by existing literature-derived pathways. The pathway resources discussed in Chapter 2 were used as a

repository for literature-derived pathways. the shortest-path algorithm was applied to all three networks and the distribution of scores (α) is shown in Table 3.1.

Table 3.1: Evaluating data-driven networks against literature (GBM). The table indicates the number of total annotated edges from the TCGA GBM networks as well as the distribution of the edges across α in comparisons against literature-validated interactions. The total number of annotated edges was computed as the number of edges of the network where both vertices existed in the literature-derived database.

Category	*ARACNE*	*Bayesian*	*Context Mining*
Total Annotated Edges	4154	3051	6157
$\alpha = \text{inf}$	26.6%	28.7%	39.8%
$0 \leq \alpha < 2$	13.7%	5.2%	1.9%
$2 \leq \alpha < 4$	27.5%	24.1%	17.6%
$4 \leq \alpha < 6$	23.5%	30.3%	28.9%
$6 \leq \alpha < 8$	7.7%	10.0%	9.8%
$8 \leq \alpha$	1.0%	1.7%	2.1%

Table 3.1 shows the results obtained from evaluating the three networks using literature-derived pathways. From this table it is interesting to see that, with networks of the order of $\approx$ 3000-6000 genes, all three networks have similar proportions of evaluated interactions. Additionally the median α value, excluding $\alpha = \text{inf}$, was found to be 3.69 in the case of ARACNE, 4.42 in the case of the Bayesian network and 4.7 in the case of the context-specific network. Using these statistics we note that ARACNE has the largest portion of mapped interactions while context-specific networks have the smallest portion of mapped interactions.

Comparing Data-driven Networks Against Random Networks

An interesting question is to understand whether these statistics are random. Specifically, if I had a random network with the same nodes as the data-driven network but a different topology, how would this network evaluate against a biological network

? To achieve this, random networks were generated corresponding to each of the three networks by maintaining the connectivity of each network but re-shuffling the node identifiers (gene names) in each case. Figure 3.4 shows how the three networks compare against random networks in terms of the proportion of edges with literature-derived paths. It is interesting to note that these random networks have similar proportions of literature-validated paths. Figure 3.4 shows the distribution of α across the data-driven networks and the random networks. Again, interestingly, random networks have similar median α scores when compared to data-driven networks.

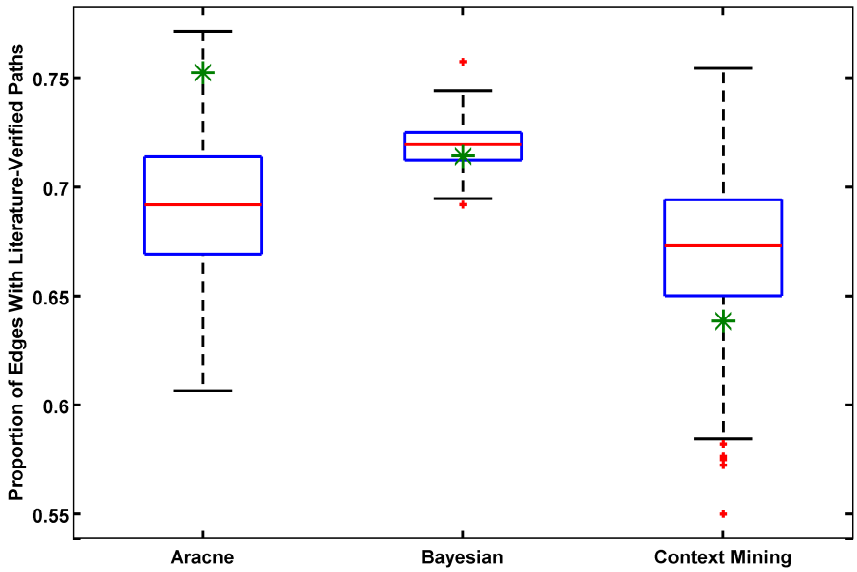

Figure 3.1: Comparing proportions of interactions with literature-verified paths against random networks generated across 1000 permutations. Green dots indicate the proportion of interactions with a literature-verified path.

These performances could be attributed to either the utility of existing data-driven networks, properties of biological networks or, the incompleteness of biological knowledge.

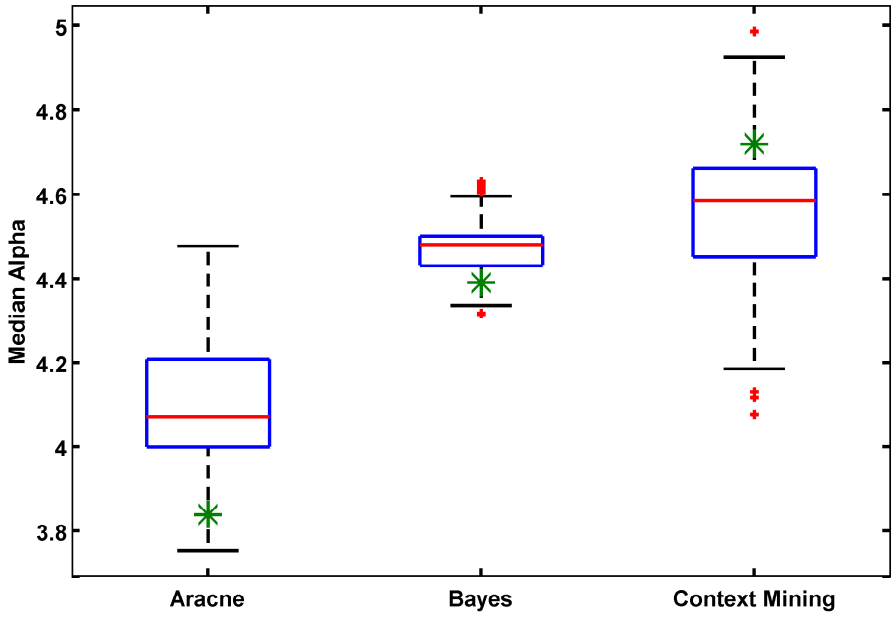

Figure 3.2: Comparing amount of literature evidence (α) for random networks against data-driven networks. Green dots indicate the median values of α for the TCGA data-driven networks.

Number of Knowledge Sources Required for Validation

Finally, I also looked into the number of distinct databases that composed the literature-derived paths corresponding to each data-driven interaction. I observed that, of the edges that were validated, 77.1 % in the case of ARACNE, 84.7 % in the case of the context-specific network and 82.6 % in the case of the Bayesian network required a combination of more than one knowledge source for their evaluation. This indicates that the integration of multiple sources is vital in such validation efforts.

Although evaluation of the data-driven interactions is insightful in determining whether the data-driven interactions are useful, plausible novel hypotheses tend to be those data-driven interactions which cannot be explained by literature-derived paths.

Hence, I now proceed to assess the data-driven interactions using other knowledge sources.

3.5 Assessing Data-driven Interactions

Data-driven interactions which do not map onto existing literature-verified pathways are ideal candidates for possible novel hypotheses. However, prior to wet-lab experimentation it is critical to determine which of the interactions are plausible hypotheses and which ones are likely to be false discoveries.

Methodology

Given a data-driven interaction (u_d, v_d), several factors contribute to whether the interaction is a true biological interaction, including both network properties (such as the connectivity or topology) as well as functional annotations. In this scoring scheme, the likelihood of an interaction between the nodes connected by a data-driven edge, is computed as a combination of several factors. While I specifically address four specific factors, additional evidence could be easily incorporated, if desired.

The first factor I consider is the overlap in neighborhood of the two nodes. The intuition behind this, is that as biological networks tend to be small-world [93], interacting partners with a high overlap in existing neighbors would be more likely to interact than others. This measure takes into account the topology of the data-driven network.

Definition 3 (**Neighbor Overlap**)**.** The neighbor overlap, $O(u_d, v_d)$ represents the combined likelihood of the shared neighbors between the nodes u_d and v_d, to interact with another node.

Given an interaction (u_d, v_d) with N_{u_d} and N_{v_d} neighbors respectively, the neighbor overlap is computed as

$$O(u_d, v_d) = 1 - \prod_{i \in (N_{u_d} \cap N_{v_d})} (1 - a_i) \tag{3.3}$$

where a_i represents the node interaction probability, as described below. The node interaction probability represents a normalized measure of how likely each node is to have another interaction and derives inspiration from the HITS algorithm for identifying authoritative nodes in social networks [94]. The node interaction probabilities are computed by an iterative procedure (Algorithm 1), where in each iteration the node interaction probabilities for each node are updated based on the degree of the node and the node interaction probabilities of its immediate neighbors. Normalizing the node interaction probabilities at each stage ensures convergence. When the algorithm converges, each node is assigned a score which denotes how likely it is to have an interaction, using the logic that nodes with more connections are more likely to have an additional connection relative to nodes with fewer connections. However, this score is computed relative to the nodes in a network and hence depends on the size of the network. To ensure the ability to compare node interaction probabilities across networks of different sizes, the last transformation (lines 18-20) is made using the density of the network assuming an undirected network. As the interaction likelihood of a given node is relative to the rest of the network, the neighbor overlap O thus considers both local properties of the interacting partners as well as global network properties.

Next, I consider the knowledge overlap between the two interacting partners computed using biological annotations of each of the interacting partners. Biological annotations usually tend to represent the functions or biological properties associated

Algorithm 1 Computing the Node Interaction Probabilities

Input : $G = (V, E)$
Output : $A = a_1, a_2, \cdots a_{|V|}$
for all $i \in V$ **do**
    $a_i = 1$
end for
$A' = 0$
while $(A - A')^2 > 0$ **do**
    $A' = A$
    for all $v \in V$ **do**
        for all $u : ((u, v) \in E \parallel (v, u) \in E)$ **do**
            $a_v = a_v + a_u$
        end for
    end for
    for all $v \in V$ **do**
        $a_v = \frac{a_v}{\sqrt{\sum_{i \in V} a_i^2}}$
    end for
end while
for all $v \in V$ **do**
    $a_v = a_v \sqrt{(|E|/(|V|^2)}$
end for

with genes, and are a useful source of evidence as interacting genes usually tend to share common annotations. However, as the scoring method is designed to identify potential novel interactions, it is relatively less likely to find shared biological annotations amongst the two nodes in isolation, and it is important to look at their neighborhoods.

I now define the neighborhood for the nodes participating in a data-driven interaction. For a given data-driven interaction (u_d, v_d), let $(u_d^1, u_d^2, \cdots u_d^j)$ be the set of first-degree neighbors of u_d and $(v_d^1, v_d^2, \cdots v_d^l)$ be the set of first-degree neighbors of v_d. Let the neighborhood of u_d be defined as the set $u_d \cup (u_d^1, u_d^2, \cdots u_d^j)$ and the neighborhood of v_d be the set $v_d \cup (v_d^1, v_d^2, \cdots v_d^l)$. While in theory, it is possible to expand and look at the neighbors of the neighbors and so on, in this dissertation I limit the computation to the immediate neighbors. Each node within this neighborhood is

associated with a set of biological annotations (such as GO terms or pathways).

Definition 4 (**Knowledge Overlap**). The knowledge overlap, $K(u_d, v_d)$ represents the combined enrichment of the shared annotations between the nodes u_d and v_d.

If $T(u_d)$ and $T(v_d)$ are the sets of annotation terms associated with the two neighborhoods respectively, I am interested in the terms which are common to both neighborhoods $T(u_d, v_d) = T(u_d) \cap T(v_d)$. Let $T(u_d, v_d) = \{t_1, t_2, \cdots t_r\}$ be the terms common to both neighborhoods. Also let $\{f_1^u, f_2^u, \cdots f_r^u\}$ and $\{f_1^v, f_2^v, \cdots f_r^v\}$ be the frequencies of the terms within the two neighborhoods.

The knowledge overlap is computed as

$$K(u_d, v_d) = 1 - \prod_{i=1}^{r} (1 - \sqrt{\frac{f_i^u}{(j+1)} \cdot \frac{f_i^v}{(l+1)}}) \tag{3.4}$$

where $(j+1)$ and $(l+1)$ are the sizes of the two neighborhoods of u_d and v_d respectively. The term frequencies for each neighborhood are normalized by the total number of nodes within the neighborhood in order to make the two frequencies comparable. The geometric mean of the two term frequencies is considered as a means to quantify the extent to which the term is present in both nodes. Equation 3.4 was formulated in order to capture the property that having a single overlapping annotation between the neighborhoods is much more important than having no overlapping annotations. However as the number of overlapping annotations increases, additional annotations are important only to the extent in which they are represented within each neighborhoods. Hence, annotations which are over-represented within the neighborhoods would dominate the score.

The knowledge overlap is computed for each biological knowledge source (such as GO annotations, pathways, chromosomal locations) separately.

Definition 5 (**Interaction Likelihood**). The interaction likelihood (β) of a data-driven interaction (u_d, v_d) represents the extent to which the interaction is likely to be a biological interaction.

Given the neighbor overlap ($O(u_d, v_d)$) and knowledge overlap ($K_i(u_d, v_d)$) computed for j different evidence sources, the interaction likelihood of an edge (u_d, v_d) is computed as

$$\beta = 1 - \sqrt[p]{((1 - O(u_d, v_d))^p * \prod_{i=1}^{j} (1 - K_i(u_d, v_d))^p)} \tag{3.5}$$

The formulation is such that it can be easily extended to additional sources of evidence on the interaction between the two. Again, the noisy-OR combination function is used in order to allow scoring components with larger values to dominate the overall sore. In this work, the GO database (of December 2011) along with pathway and chromosomal location information (from the MSigDB database version 3.0) are used as sources of evidence. The GO annotations were filtered based on evidence codes to include only the experimentally verified annotations.

Performance on Gold Standard : Literature-verified Pathways

One of the first challenges in developing a scoring method for distinguishing true data-driven interactions from false positives is in determining if the scoring method is doing what it is supposed to do i.e., assigning high scores to true biological interactions and low scores to false interactions. To achieve this, the first experiment studied the performance of the scoring algorithm on literature-verified interactions from Pathway Commons (illustrated in figure 3.5). Specifically, human interactions with at least two literature citations were used (in order to eliminate unreliable interactions from the gold standard). As negative examples, 10000 random pairs of

nodes (with no path between them in Pathway Commons) were considered (case 1 in figure 3.5). Using this, the interaction likelihood of each interaction was computed. Also, the pathway knowledge source was not used in the simulations (however, it was used in the application study) in order to avoid biases. Three parameters were varied – p, the power to which each term is raised in the Noisy-OR formulation in 3.5, the threshold for the interaction likelihood to determine if an edge is true or not, and the proportion of positive examples within the sampled data set. Precision and recall values were computed at each data point and the results obtained are shown in figure 3.3. As seen in the figure, in all cases, the scoring method developed in this work is able to assign appropriate scores.

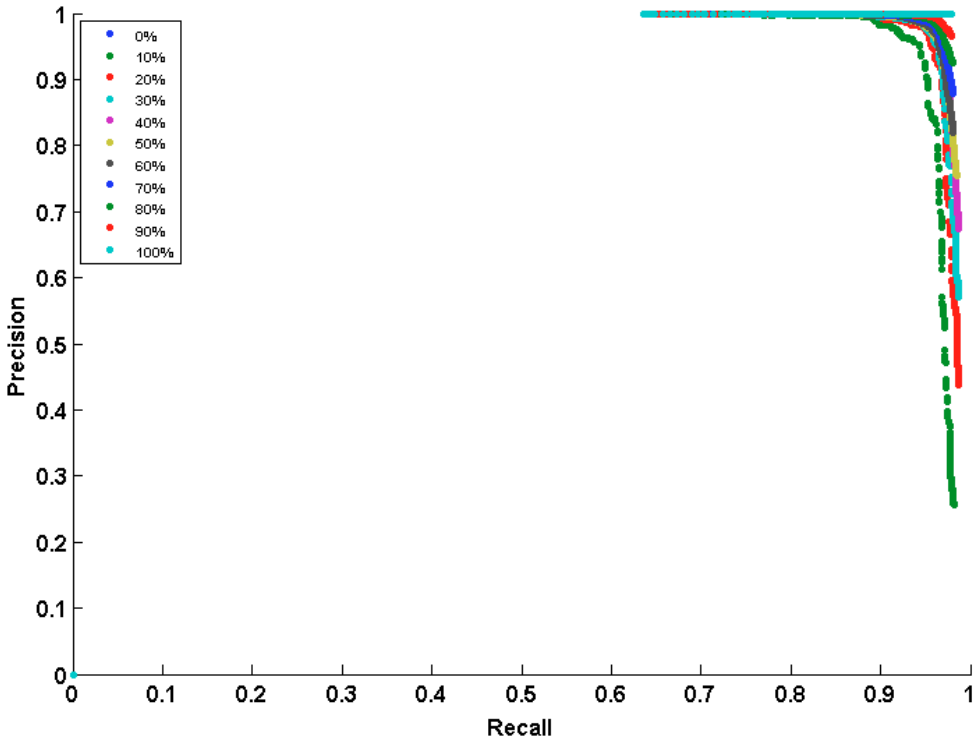

Figure 3.3: Precision-recall curve using literature-derived interactions.

Performance on Simulated Data-driven Interactions

The next challenge was in determining if the scoring algorithm performed well in the context of simulated data-driven interactions. While the ”gold standard” from Pathway Commons was a good way to determine how well the scoring algorithm performed, it still did not capture all properties of data-driven interactions. For instance, data-driven interactions tend to measure associations which could be either a direct biological interaction or a path of multiple biological interactions. Hence, in order to study the performance of the scoring algorithm on data-driven interactions, I simulated interactions as shown in figure 3.5. Specifically, I sampled 10,000 random pairs of nodes from the previously discussed gold standard database (case 2 in figure 3.5). For each such pair, I ensured that the two nodes were connected (by a path of one or more steps) thus simulating a data-driven interaction. The interaction likelihood was applied to these simulated data-driven interactions and the precision-recall values were computed by varying the parameters as before. As seen in figure 3.4, performance degrades with larger proportions of negative examples (interactions with no path). However, as seen in 3.1, this is seldom the case while using data-driven networks.

Performance on Simulated Data-driven Interactions - Incomplete Topology

As discussed in Chapter 2, current estimates of human protein-protein interactions are incomplete. Hence, an important consideration in designing algorithms which utilize such knowledge is to ensure the algorithm copes with incomplete biological knowledge. The next experiment was designed in order to simulate incompleteness in biological knowledge in the form of interactions that are currently not yet discovered. To achieve this, for each of 10,000 random pairs (with a path) considered

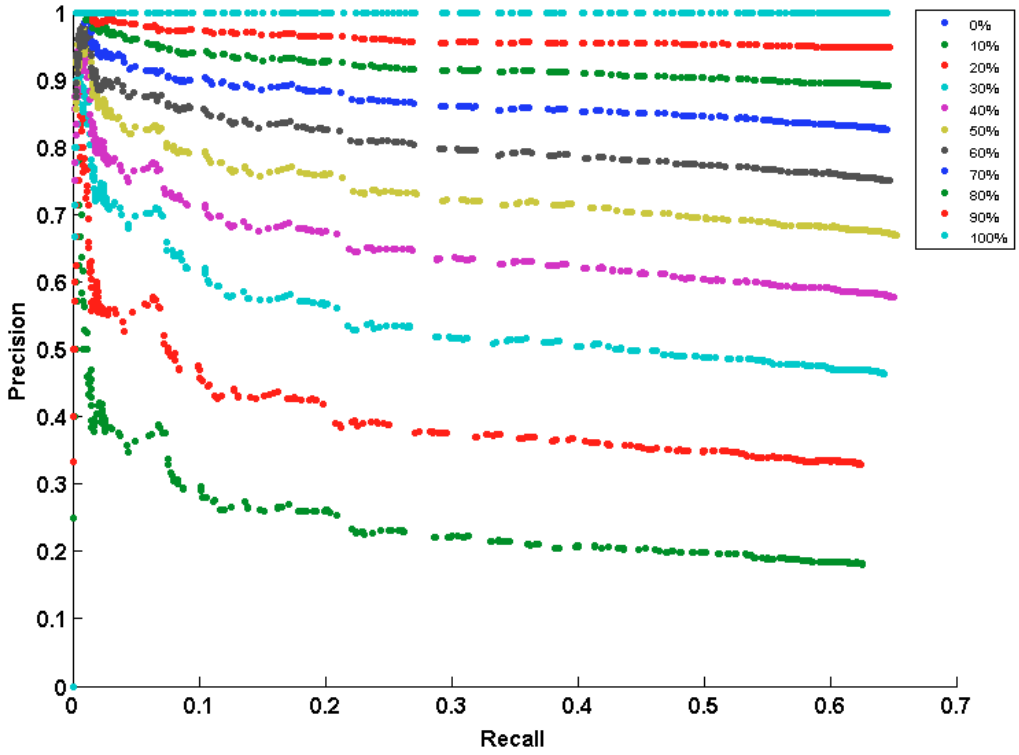

Figure 3.4: Precision-recall curve using simulated data-driven interactions.

in the previous study, the two nodes were forcibly disconnected. The interaction likelihood was computed after disconnection and precision-recall values were computed by varying the parameters as before. As seen in figure **??**, although there is a degradation in performance, it is not very significant.

Performance on Simulated Data-driven Interactions - Incomplete Annotations

In order to further understand how the scoring algorithm performs in light of incomplete biological knowledge, I also studied the effects of incomplete annotations on the interaction likelihood. To achieve this, for each knowledge source, annotations were randomly sampled at 100%, 75%, 50% and 25% as shown in figure 3.7. The interaction likelihood was computed in each case and precision-recall values were computed by varying all parameters as before, except the proportion of positive

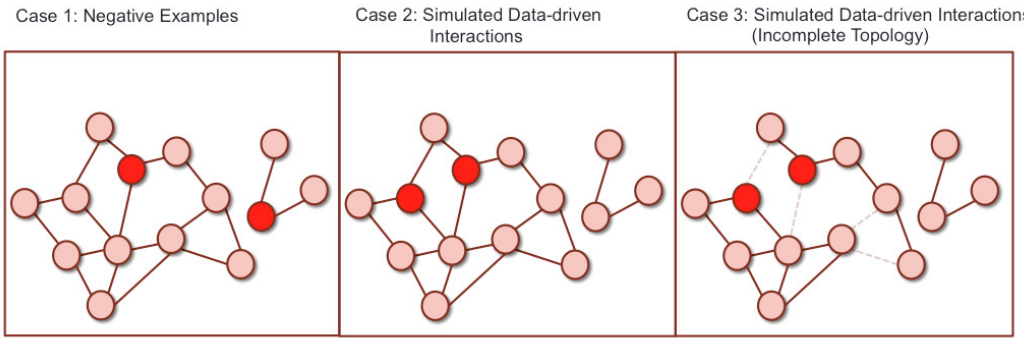

Figure 3.5: Simulating data-driven interactions and incomplete knowledge (topology).

examples which was set to 60% (the lowest observed using data-driven networks). As seen in figures 3.8, 3.9, and 3.10, in all cases there is hardly any degradation in performance, proving that the interaction likelihood copes well with the lack of biological knowledge.

Application to Data-Driven Interactions in GBM

Finally, I examine the interaction likelihoods of the three data-driven networks. Figure 3.11 shows the cumulative distribution of the interaction likelihoods with and without literature-derived paths. Overall, the bayesian network has much lower scores than Aracne or the context-specific network. Further it is interesting to note that although 25 – 40 % of the data-driven interactions did not have literature support, they can be explained through other biological sources. High-scoring interactions could then be extracted and used for further wet-lab verification.

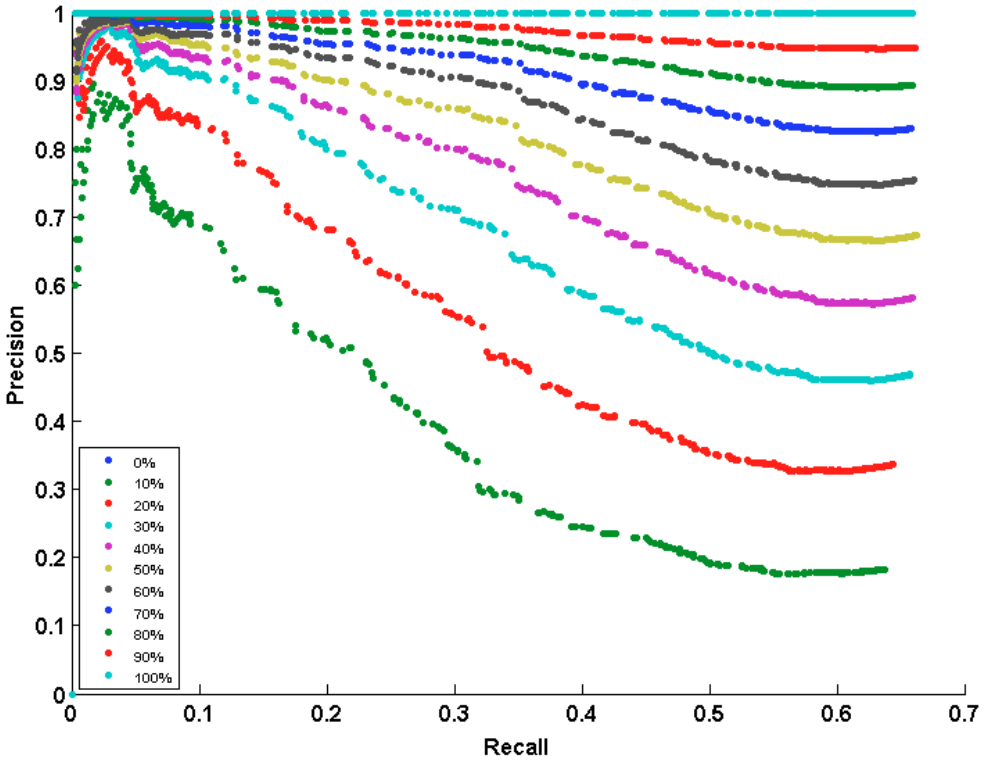

Figure 3.6: Performance of scoring algorithm with incomplete knowledge (topology).

3.6 Summary

In summary, I have proposed a method for both evaluating and scoring data-driven interactions using biological knowledge. I have shown how the evaluation of the data-driven networks could shed light on several interesting properties of biological networks. For a given network size, the proportion of data-driven interactions evaluated against literature-derived pathways is similar across networks learnt using different methods. It is also interesting to note that random networks created using the same set of nodes as the data-driven network results in comparable statistics. Further I also find that the integration of multiple knowledge sources plays an

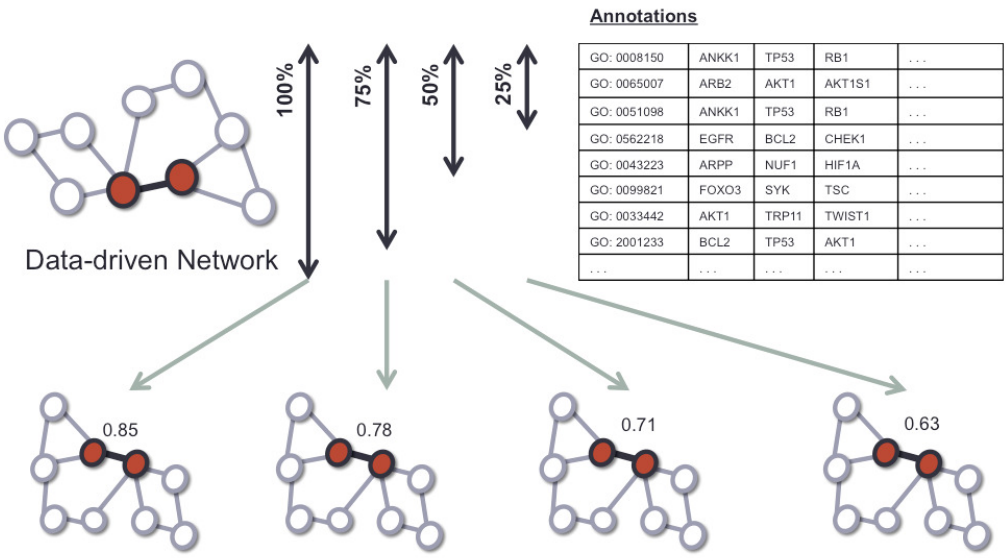

Figure 3.7: Simulating incomplete biological knowledge in the form of missing annotations.

important role in such validation efforts.

I have proposed a scoring method for assessing the likelihood of a data-driven interaction. This metric copes with both incompleteness in biological annotations as well as missing connections in literature-derived pathways. The interaction likelihood has then been applied to score the data-driven networks from a glioma dataset and identify plausible novel hypotheses.

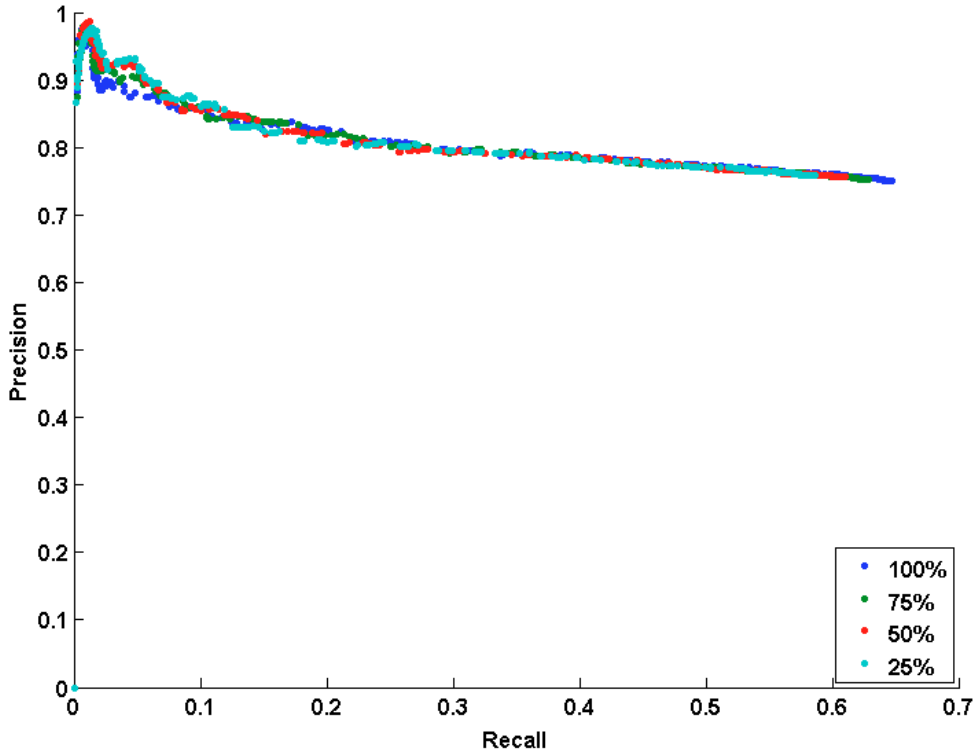

Figure 3.8: Effects of missing GO biological process annotations on scoring performance.

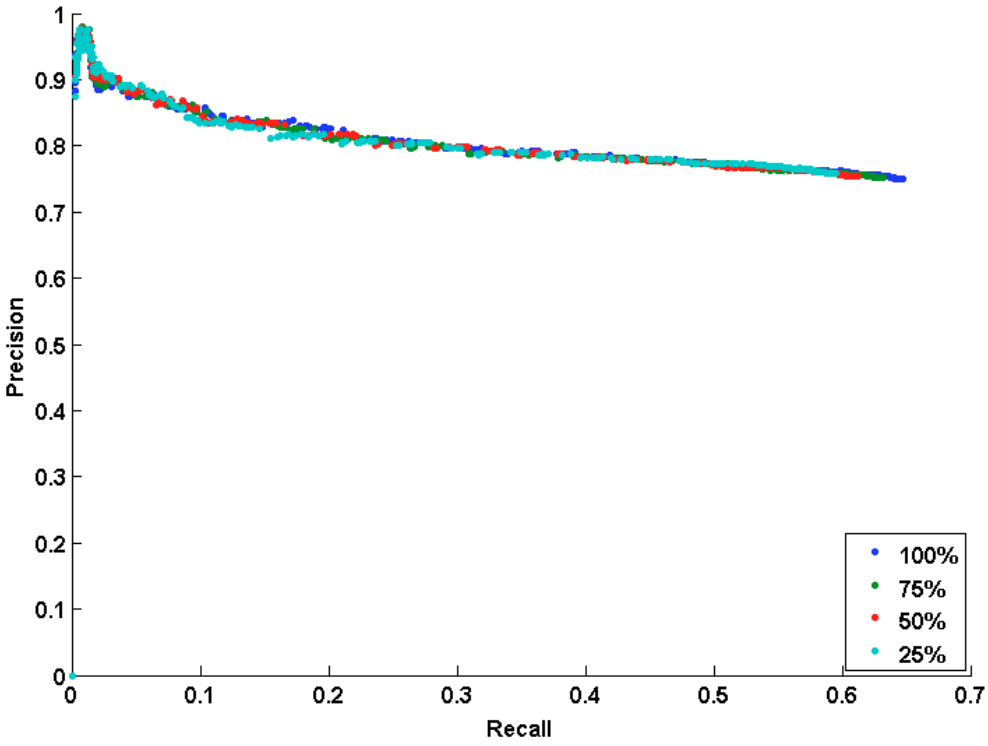

Figure 3.9: Effects of missing GO molecular function annotations on scoring performance.

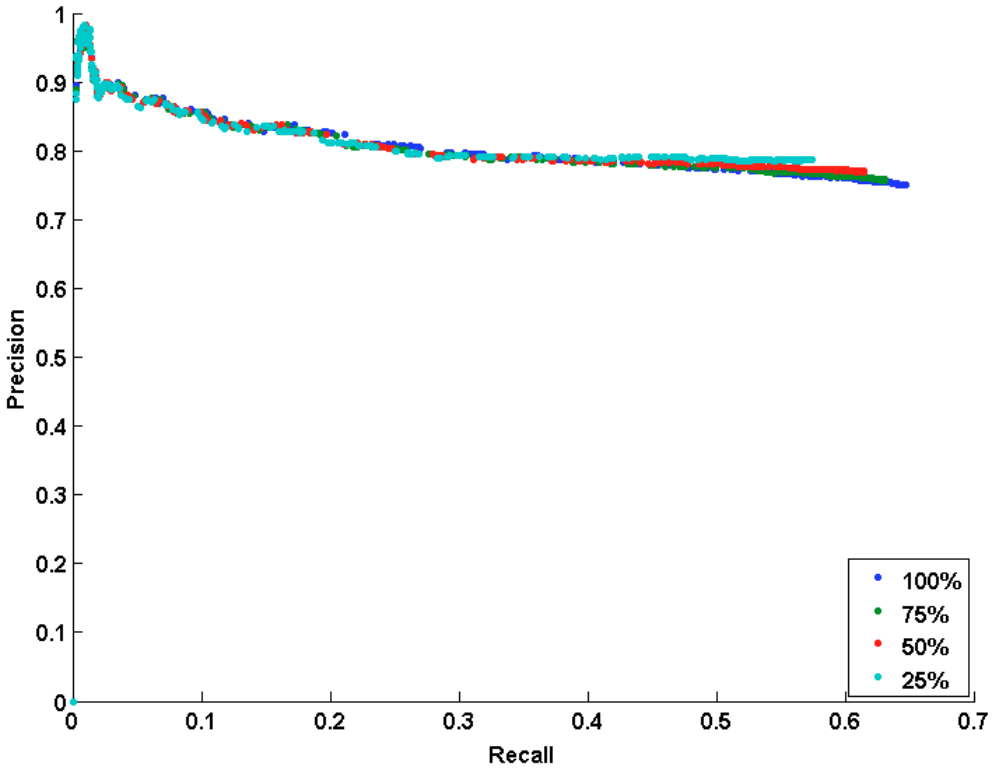

Figure 3.10: Effects of missing chromosomal location annotations on scoring performance.

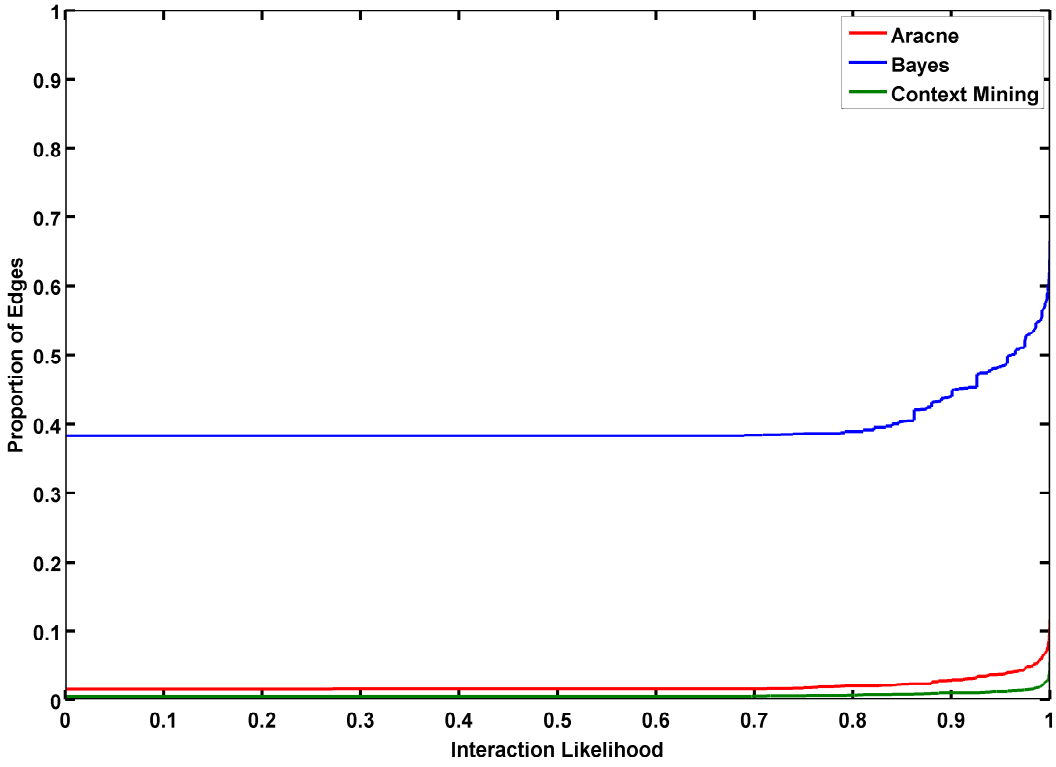

Figure 3.11: Cumulative distributions of interaction likelihood for the data-driven interactions.

Chapter 4

IDENTIFYING FUNCTIONAL MODULES

In the previous chapter, I have proposed a method to evaluate data-driven interactions using existing biological knowledge, along with applications to the data-driven networks from gene expression data derived from cancer patients. However, often biologists are faced with the problem of selecting a sub-set of interactions to focus on. A functional module is a distinct entity consisting of a set of molecular interactions working towards a common biological purpose, separable from the purposes of other modules [7, 95]. Here I discuss the motivation for identifying functional modules within a data-driven network, relevant work in this area, the methodology developed in this dissertation and the functional modules identified in cancer data-driven networks.

4.1 Motivation

Within the cell, functional modules play an important role in facilitating evolution [96, 7, 97]. They represent core biological processes which are robust to change. During evolution, changes to the ways in which functional modules are interconnected with one another occurs, leading to different phenotypes, adapted over time. Hence, identifying functional modules within a data-driven network enables understanding the cell's response to external signals [40] under different phenotypic conditions. For instance, the PI3K/AKT pathway is a functional module regulating cellular growth within normal cells. Within tumor cells however, several genes within the PI3K pathway are over-expressed, above normal levels, leading to abnormal cellular growth. This concerted activity of multiple genes is often difficult to observe at the individual interaction level.

The functional modules extracted from data-driven networks allow biologists to narrow down on the key biological processes that are active within a given set of samples. Among the thousands of interactions within the data-driven network, functional modules provide biologists with a smaller sub-set of interactions to focus on. This dissertation develops an approach to identify functional modules within a data-driven network, associated with both biological processes as well as phenotypic conditions.

4.2 Relevant Work

In this section, I address approaches that have been developed to extract functional modules from biological networks. In chapter 2, I discussed several approaches for identifying sub-type specific modules from data as well extending the modules to module networks. Here I discuss approaches for identifying modules within protein-protein interaction networks.

Biological networks are often represented as graphs (both directed and undirected) with the nodes representing genes (or proteins) and edges representing the presence of an interaction between the two nodes. Consequently, graph theory is often applied to solve problems in biological networks. For instance, Perreira *et. al.* [98] have applied Markov clustering to extract functional modules from networks of yeast protein-protein interactions. Markov clustering allows for graph clustering based on flow between nodes in a network. Subsequently, Perreira *et. al.* extract the functional significance of modules using the consistency of protein classifications within each module. MCODE, proposed by Bader *et. al.* [99], extracts densely connected regions corresponding to molecular complexes from protein-protein interaction networks derived from yeast. Their algorithm uses a vertex weighting scheme which estimates the extent to which a neighborhood forms a clique and uses this measure

to extract complexes from protein-protein networks. Alternately, Palla *et. al.* [100] extract overlapping community structures from protein-protein interaction networks. Their approach relies on the definition of a community as consisting of several complete (fully connected) sub-graphs sharing a large proportion of nodes. Using this definition, communities in yeast protein-protein interaction networks have been analyzed. Navlakha *et. al.* [101] use graph summarization to extract biologically meaningful modules. The approach involves compressing the nodes in the original network into supernodes (composed of a set of nodes) and has been applied to the protein interaction networks derived from yeast. Speed and Performance in Clustering (SPICi) [102] is another fast clustering technique which builds clusters using a greedy approach. The algorithm starts from local seeds with a high weighted degree to add nodes that maintain the density of the clusters and are adjacent to a suitable fraction of nodes within them. SPICi has been applied to extract clusters from yeast protein-protein interaction networks. However, the main shortcoming of the existing methods is in incorporating phenotypic significance while extracting the modules.

In this dissertation, I develop a method to identify functional modules within a data-driven network. The methods developed here identify functional modules with both biological significance and phenotypic significance. Hence, the methods are focused on context-specific gene regulatory networks (GRNs) [46, 103] in order to capture the phenotypic similarities encoded in the interactions. However, these method could be easily applied to other data-driven networks including Bayesian networks [104, 105] and ARACNE [32].

In addition to biological significance, drug target information is also associated with the modules enhancing the therapeutic significance of the modules. The methods developed here are applied to two cancer gene expression datasets yielding insights on

possible associations between tumor-types and several useful clinical implications.

Context-specific Gene Regulatory Networks

Prior to defining the problem of identifying functional modules, I describe context-specific gene regulatory networks as they form the primary input to this work.

I begin by describing the building blocks of context-specific gene regulatory networks – context motifs.

Context Motifs

In its simplest form, a context motif is represented as $M = (G, T)$ where G represents a set of genes and T represents the set of samples under which the genes are expressed consistently within the high throughput dataset. In order to identify context motifs, the set of genes G is divided into driver genes (or drivers) and passenger genes (or drivens), based upon the extent to which the genes exhibit consistent expression (i.e., genes with high coherency in expression are driver genes).

For a single gene, g_i, assuming it is the driver gene for a given context-motif, two probabilistic measures *conditioning* and *crosstalk*, are used, to identify passenger genes that are coherently expressed along with g_i. For simplicity, the expression levels of all genes are assumed to be binary (ON or OFF), although the method could be applied to more than two states. Two cases for g_i are considered, namely when the gene is ON and when the gene is OFF.

In the first case, when g_i is ON, the statistic *conditioning* is used to measure for any other (passenger) gene g_j, the conditional probability of g_j also being ON.

Definition 6 (**Conditioning**). Conditioning (δ_{ij}) is the extent to which contextual

effects diminish the influence of a driver gene g_i on a passenger gene g_j.

$$P(g_j = 1 | g_i = \mathbf{1}) = 1 - \delta_{ij} \tag{4.1}$$

In the second case, when g_i is OFF, the statistic *crosstalk* is used to measure the probability that the state $g_j = 0$ depends on contextual effects alone and not the effects of drivers.

Definition 7 (**Crosstalk**). Crosstalk (η_{ij}) is the extent to which contextual effects outside the driver gene g_i activate a passenger gene g_j.

$$P(g_j = 1 | g_i = \mathbf{0}) = \eta_{ij} \tag{4.2}$$

While the above definitions are provided for context motifs with a single driver genes and binary expression levels, the definition of a context motif can be generalized to a set of driver genes.

With this generalization, a context motif is represented as $M = (D, Y, S, T)$ where D represents a set of driver genes, Y represents the state of the driver genes (e.g., $Y \in \mathscr{Q}^q$ where $\mathscr{Q} = \{0, 1\}$ and $q = |Y_i|$, for a binary quantized dataset), S represents a set of passenger genes and T represents the set of samples under which consistent expression is observed. The hypergeometric test [45] corrected for false discoveries using Benjamini and Hochberg's method [106] is used in order to assess the statistical significance of obtaining a context motif with a given crosstalk and conditioning. A driver gene g_i at activity level y_i is said to regulate a passenger gene x_j when the conditioning δ_{ij} and crosstalk η_{ij} values are lesser than user-specified threshold δ_θ and η_θ, with a statistical significance less than the user-specified threshold p_θ.

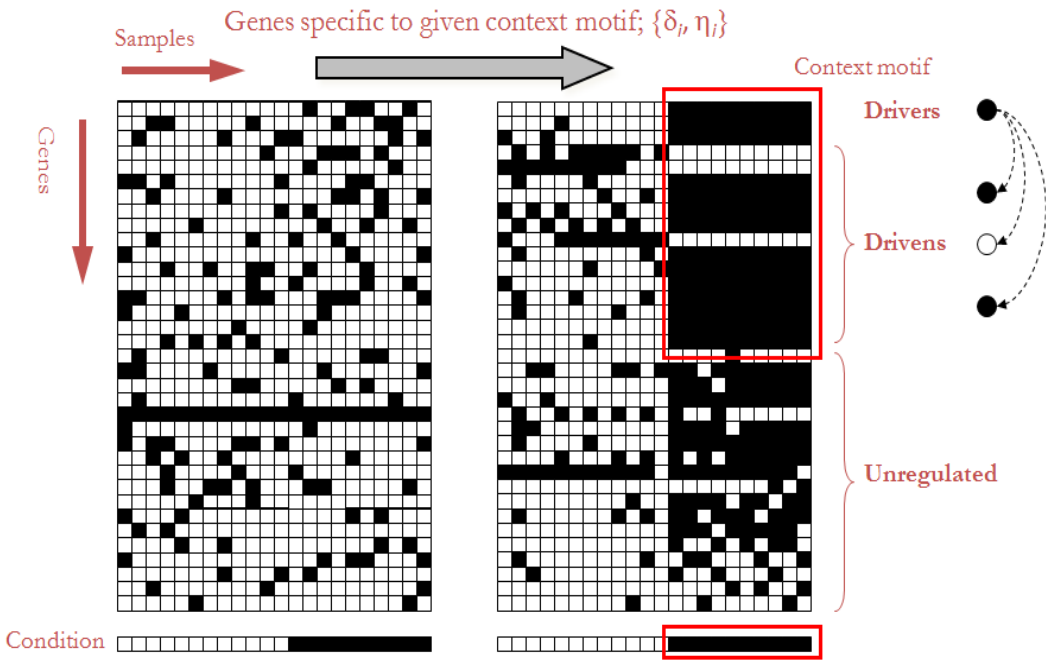

Figure 4.1: Algorithm to learn context motifs from high-throughput data [45].

Context Motif Identification

Using the model described above, all potential context-motifs are identified from data via a process called *in-silico* conditioning [45], a method designed to mimic a biologist's manipulation of the status of a gene in an experiment using techniques such as ectopic expression or gene silencing. Given a gene g_i, the expression of the gene is first set to a certain state, y_i, for example, ON. Then, the samples are divided into two groups, based on the expression of the gene: the first group, T_i with all the samples with $g_i =$ ON, and the second with all the samples with $g_i =$ OFF. The crosstalk and the conditioning for the rest of the genes in the data are estimated to determine which genes show consistent transcriptional activity. Note that a gene can be set to multiple states (ON or OFF if binary), and each gene can be a driver for multiple context motifs. In some cases, different genes can be conditioned across

the same set of samples, leading to multiple driver genes in a context-motif. This process is repeated for all the genes in the data to identify all potential context-motifs. Figure 4.2 illustrates this process.

Additionally, the statistical significance of each context motif is estimated based on its size (number of genes and samples). Given a context-motif $M = (D, Y, S, T)$, the probability of obtaining a context-motif of $l = |D \cup S|$ genes or more by chance is estimated using a permutation based approach. Specifically, the given data set is randomly split into two groups of which one has a sample size of $k = |T|$ and the other has a sample size of $N - k$, where N is the total number of samples in the dataset. The same set of statistics (Eqs. 4.1–4.2) are then used to identify the number of genes filtered by the same thresholds for interference, crosstalk and p-value. By repeating this procedure many times, $\Pr(L \geq l \,|K = k)$ is estimated. The accuracy of the estimation is based on the number of repetitions. 10,000 repetitions was used to build empirical distributions with adequate statistical power, towards the applications described in this dissertation. The estimated probabilities for context-motifs are subsequently corrected for false discoveries using Benjamini and Hochberg's procedure [106]. Using this re-sampling-based approach, the statistical significance of identified context-motifs is found and only those with a significant p-value are considered for further analysis.

Learning Contexts from Context Motifs

Following the identification of context motifs, the learned interactions can be chained and cascaded to understand contextual genomic regulations. A given context motif $M = (D, Y, S, T)$ defines regulatory relationships between the driver genes and the passenger genes, i.e., $g_i \rightarrow g \in S$, specific to T samples with g_i (drivers) conditioned on a specific state $Y = y_i$. A driver g_j in context C_j could be driven by g_i in

another context C_i. When such relationships are added to the implicit driver-driven relationships $g_i \rightarrow g_j$, an interesting graphical structure is obtained representing the relationships between context motifs; this graph is called a *context motif network*. A context motif network differs from other graphical representations in the fact that context motifs connected to one another within the network differ in their sample composition.

The context motif network represents a network of biological interactions where each interaction is specific to a particular sub-set of samples. Typically, after filtering out context-motifs that are statistically insignificant, this network tends to contain a few hundred nodes and several thousand interactions. Further, while functional modules can be seen in this network they are fairly difficult to discern visually, due to the size of the network and the heterogeneity in the dataset, often resulting in interactions conditioned by samples of different sub-types or clinical annotations. Therefore, it is important to extract functional modules from the graph, by promoting regions of strong connectivity – implying considerable overlap in sample composition and removing edges with weaker connectivity between functional modules.

4.3 Problem Formulation

The task of identifying functional modules within data-driven networks poses several challenges. Firstly, data-driven networks have a large number of nodes and connections between the nodes, implying scalability is a desirable property. Secondly, data-driven networks are obtained from high-throughput data which is heterogeneous being drawn from different samples with varying phenotypes. Consequently, it would be relevant to identify modules which not only share a common function but also have phenotypic similarities.

By definition, a functional module consists of set of densely connected nodes, with

loose connections between the modules. It has input and output nodes controlling interactions with the rest of the network. It also possesses internal nodes which do not significantly interact with nodes outside the module [107]. Thus, regulation of the biomolecules within a functional module is tight while the regulation between functional modules tends to be fairly loose. Here the definition of a cut is particularly useful.

Definition 8 (Cut). A cut is a partition of the gene network into two non-empty sets S and $V \setminus S$ denoted by $(S, V \setminus S)$.

Usually a cut is uniquely defined by a set S, and hence any sub-set of V can be called a cut. The *cut-size* is the number of edges that connects vertices in S to those in $V \setminus S$. Given this definition a clustering of the graph G into functional modules $M = (m_1, m_2, \cdots m_k)$ possesses the property that the cut size between modules is lower than the average connectivity within each module.

Definition 9 (Functional Module). A functional module m_i is defined as a triplet (C_i, B_i, T_i) where

- C_i is an induced subgraph of the graph, such that $C_i = (V_c, E_c)$, where $V_c \in V, E_c \in E$; for every edge $(u, v) \in E_c$, $u \in V_c$ and $v \in V_c$
- B_i is a set of biological processes (or pathways) significantly enriched in C_i, and is non-empty.
- T_i is a set of clinical annotations significantly enriched in C_i, and is non-empty.

Based on the above definition, the functional modules identified within the computational gene network are mutually exclusive. However a single functional module could be associated with multiple biological processes as well as a single biological

process could be associated with multiple functional modules, as it is fairly common to find individual genes participating in multiple biological processes (or pathways).

Given such a network, the task of discerning functional modules within the data-driven network is viewed as a two-stage process. The first stage is to partition the network into modules, where the nodes within each module are densely connected and relatively isolated from the rest of the network. The second stage is to associate known biological processes and clinical annotations to each module. The main reason for doing this is in order to allow for flexibility in the biological knowledge and clinical annotations associated with each module.

4.4 Discerning Modules

The first stage in identifying functional modules within the data-driven network is to partition the network into modules.

Methodology

The goal of this stage is to identify modules within the data-driven network in such a way that there are many edges within each module and relatively few edges between modules. Although at a first glance this appears as a graph partitioning problem, this is instead formulated as a clustering problem since balancing the sizes of the clusters is not a key criterion in determining the modules. Further, as opposed to conventional clustering, in this problem, similarity is expressed through whether elements 'share a property' or not (such as a regulatory relationship where genes are co-regulated), rather than the distance between the elements, steering towards graph clustering approaches. The question ofcourse remains, as to whether to consider supervision using a semi-supervised approach.

Semi-supervised clustering aims to organize data points into clusters, using limited amount of information in the form of pair-wise constraints between the data points. Pairwise supervision is usually provided in the form of 'must-link' and 'cannot-link' constraints. A 'must-link' constraint indicates that the two data points in the pair should be placed within the same cluster. A 'cannot-link' constraint indicates that the two data points in the pair should belong to different clusters.

Semi-supervised clustering has successfully applied to several biological problems [108, 109, 110, 111, 112, 113, 114]. Semi-supervised graph clustering [115] has also been shown to be successful in clustering a gene interaction network from yeast (with 216 genes). In the context of data-driven networks however, semi-supervised clustering is not very desirable. Pairwise supervision could be introduced in two dimensions - genes and samples. In the case of genes, pairwise supervision would be by introducing 'must-links' between genes within the same biological pathways or sharing biological functions. However, data-driven interactions learnt from high-throughput data are association networks. Hence, a single computational interaction often maps on to a path of biological interactions (as discussed in Chapter 3). This complicates the assignment of must-link constraints to the genes based on known biological pathways. Directly using biological pathways to enforce must-link constraints would not be applicable to data-driven networks.

Consequently, graph clustering approaches are used to discern modules from data-driven networks . Specifically, two algorithms – Markov clustering and spectral clustering were chosen for identifying functional modules within the data-driven networks. Markov clustering was chosen due to its scalability and ability to automatically determine the number of clusters. Spectral clustering was chosen due to its ability to find an optimal minimum cut while creating well-balanced clusters. In addition, previous successful applications of these algorithms in the bioinformatics

field have yielded promising results [116, 117, 118, 119] indicating these algorithms could be well-suited for this application.

Markov Clustering

The Markov clustering (MCL) algorithm derives its inspiration from the notion of random walks in graphs. If a random walk visits a certain vertex in a cluster, it would be likely to visit several other members of the cluster before leaving the cluster. [120]

The Markov clustering algorithm simulates flow using two (alternating) algebraic operations on matrices. Expansion (identical to matrix multiplication) represents the homogenization of flow across different regions of the graph. Inflation, mathematically equivalent to a Hadamard power followed by diagonal scaling, represents the contraction of flow, making it thicker in regions of higher current and thinner in regions of lower current. Intuitively, expansion corresponds to augmenting the neighbors of a given vertex, and inflation corresponds to promoting those neighbors which have a higher transition probability from a given vertex. The Markov clustering process causes flow to spread out within natural clusters and evaporate in between different clusters [120]. The iteration is continued until a recurrent state or fixpoint is reached.

The exact steps are explained in Algorithm 2. The connected components of the graph induced by the non-zero entries of M provide the required clustering. Proof of concept, mathematical properties and analyses on the complexity and scalability of the algorithm can be found in [121].

Algorithm 2 Markov Clustering

Input: $G = (V, E)$, expansion parameter e, inflation parameter r
while M is not fixpoint **do**
  $M \leftarrow M^e$
  for all $u \in V$ **do**
    for all $v \in V$ **do**
      $M_{uv} \leftarrow M_{uv}^r$
      for all $v \in V$ **do**
        $M_{uv} \leftarrow \frac{M_{uv}}{\sum_{w \in V} M_{uw}}$
      end for
    end for
  end for
end while

Spectral Clustering

Spectral clustering uses the Eigen decomposition of matrix representations of a graph to determine the optimal partitioning of the graph. Although, there has been extensive research in the spectral clustering field, I use the algorithms developed by Shi and Malik [122] – for symmetric clustering, and Meila and Pentney [123] – for asymmetric clustering, because they incorporate information from the edges (in our case, data-driven interactions) in determining the optimal clustering of a graph.

Symmetric Cuts: In graph theory, a cut is defined as

$$cut(A, B) = \sum_{u \in A, v \in B} w_{uv}, \tag{4.3}$$

where A and B are the clusters resulting from the cut between vertices u and v. Finding the minimum cut for Equation 4.3 could result in singletons or clusters with very few nodes, leading to poorly distributed clusters. Thus, there exists a need to balance the clusters. Shi and Malik, have proposed a solution to this problem by normalizing the cuts that create clusters [122]. The cut cost is calculated as a fraction of the weights of the edges in the induced sub-graphs. As finding the exact

solution to the normalized minimum cut problem is considered NP-complete, the authors have found that using the eigenvector corresponding to the second smallest eigenvalue of the Laplacian of an undirected graph (also known as the Fiedler vector) could efficiently provide an approximate discrete solution [122]. The algorithm, referred to as the normalized cut algorithm, recursively splits clusters thresholding the Fiedler vector of the induced sub-graphs until the desired number of clusters are reached.

Asymmetric Cuts: Meila and Pentney [123] provide for the expansion of spectral clustering in multi-way cuts to directed graphs, as the normalized cut is applicable only to undirected graphs. In gene regulation directionality could provide useful information. The weighted cut algorithm, proposed by Meila and Pentney, mathematically transforms a directed graph (with a non-normalized Laplacian matrix, D-A), into a symmetric Hermitian matrix [123] and finds an approximate solution to minimizing a normalized cut. Using the k eigenvectors pertaining to the k smallest eigenvalues of the Hermitian matrix, the weighted cut algorithm applies the k-means algorithm to cluster the graph. In addition, the algorithm allows for user input, balancing parameters T and T' , to normalize the cuts produced by the algorithm. Thus the normalized minimum cut for directed graphs can be expressed as:

$$MNCut(x) = \min_{z_k \in R^n \text{orthon}} \sum_{k=1}^{K} z_k^* H(B) z_k \tag{4.4}$$

where $B = T^{-\frac{1}{2}}(D-A)T^{-\frac{1}{2}}$, K is the number of desired clusters and $H(B)$ is the Hermitian matrix of B.

In this section I describe the comparisons between the algorithms and results obtained from applying the previously described methods to two datasets - a refractory cancer data set and a glioma data set.

Comparing Graph Clustering Algorithms

In our first study, three spectral clustering variants are compared – symmetric spectral clustering with two variants of asymmetric spectral clustering, using different balancing parameters (the average cut and the out-degree cut).

Performance Metrics

I use the metrics coverage and performance [124] to compare the methods.

Definition 10 (**Coverage**)**.** The coverage of a partitioning M is defined as the fraction of intra-cluster edges (q_m) within the complete set of edges (q), i.e

$$\alpha(M) = \frac{q_m}{q} = \frac{q_m}{q_m + \overline{q_m}} \tag{4.5}$$

I choose this metric as it measures the wellness of a cut in a graph by taking the edges within the cluster(s) of a graph as a fraction of all the edges. Thus, the smaller a cut, the better the coverage it would have. Both a graph with no clusters at all and a graph with several disconnected components would have a coverage of 1 due to the absence of inter-cluster edges. Sparsity of the graph would not influence the coverage as long as the intra-connectivity is much higher than the inter-connectivity.

Definition 11 (**Performance**)**.** The performance of a partitioning M is the fraction of intra-cluster edges together with non-adjacent pairs of nodes in different clusters within the set of all pairs of nodes.

$$\beta(M) = \frac{q_m + \sum_{v,w \notin E, v \in m_i, w \in m_j, i \neq j} 1}{\frac{1}{2}n(n-1)} \tag{4.6}$$

The performance of a partitioning M counts the number of 'correctly interpreted pairs of nodes' in a graph. I choose this measure as a means to assess the connectivity within the clusters of the graph. The fewer non-edges (pairs of nodes within the same cluster but lacking an edge between them) there are within a graph, the higher its performance would be. Further, a graph containing several singleton nodes, as well as a fully connected graph with a single giant cluster, would both have a performance of 1, as the number of non-edges would be zero in both cases. The goal is to maximize connectivity within a cluster for better performance and by maximizing intra-connectivity (approaching the number of possible edges of a graph), one can minimize the inter-connectivity. It is notable however, that performance will not do well in sparsely connected large graphs even when there may be substantially fewer edges between clusters.

The above formulae are specific to undirected graphs. Direction when available is associated with each edge e. In the case of directed graphs, the maximum number of edges possible is twice as many as the edges possible in undirected graphs and the formulae are correspondingly modified.

Comparing Symmetric against Asymmetric Spectral Clustering

In the first experiment, symmetric spectral clustering is compared against asymmetric spectral clustering. The average of performance and coverage is used as a measure of the wellness of the clusters, and is plotted against the number of clusters produced, shown in Figure 4.2.

Spectral clustering performed well both on undirected graphs and directed graphs. I notice that the asymmetric algorithms peaked at a higher number of clusters than the symmetric algorithm. This implies that the normalized cut algorithm left intact large, well connected clusters until a certain threshold was reached. I also note that

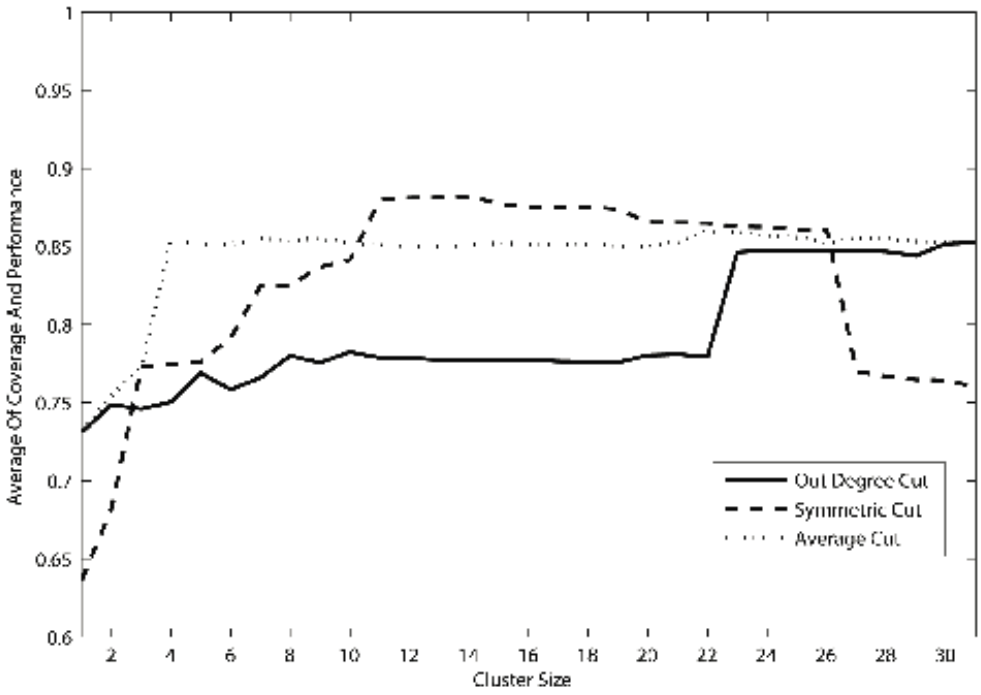

Figure 4.2: Performance and coverage average of spectral clustering

Table 4.1: Performance comparison of Markov and spectral clustering.

		Coverage	Performance
Network 1	Spectral Symmetric	0.97	0.8
(391 nodes	Spectral Asymmetric	0.95	0.6
6200 edges)	MCL	0.94	0.77
Network 2	Spectral Symmetric	0.97	0.93
(1,901 nodes	Spectral Asymmetric	0.71	0.88
33,820 edges)	MCL	0.94	0.89

using the average cut exhibits less fluctuation in performance across different cluster sizes than using the out-degree of the nodes, explained by the fact that the average cut uses the number of nodes as the balancing parameter. However, if in fact a gene regulatory networks follows a scale-free topology then the average cut may not prove to be the most useful in identifying biologically significant clusters because it does not take into account the interactions within a cluster.

Comparing Spectral Clustering against Markov Clustering

In our second study, spectral clustering (symmetric and asymmetric) is compared with Markov clustering. As seen in Table 4.1, in terms of coverage, spectral clustering performed well over both directed and undirected graphs. In terms of performance, the asymmetric case shows a lower performance value than the other two. This indicates that incorporating directionality does not correspond with a significant impact on the clustering, in terms of the performance metrics.

While both MCL and symmetric spectral clustering performed well on the much larger dataset, exhibiting good scalability. However, the spectral clustering techniques required the number of clusters to be pre-specified and this is often not available in biological applications. Hence, for the biological study MCL was applied.

4.5 Associating Biological Knowledge with Modules

Biological knowledge in the form of gene sets and the clinical categories of the samples is used to associate functional significance to each extracted module.

Methodology

Here I describe the methods developed to identify enriched functional and clinical annotations for each functional module.

Enriched Functional Annotations

To understand the biological relevance of the functional modules, I investigate the functional enrichment of each module using pre-defined biologically relevant gene

sets. The hypergeometric test is used to measure the significance of the enrichment and the p-values are corrected for False Discovery Rate (FDR) using Benjamini and Hochberg's method [106]. The Molecular Signatures Database (MSigDB version 3.0) [49] and Pathway Commons (as of September 2010) [6] are used as reference knowledge sources (described in Chapter 2).

Enriched Clinical Annotations

Context-specific gene regulatory networks [46] allow for associating samples with each node (in this case context motifs) in the network. Based on the available clinical annotations, I associate modules with enriched clinical associations.

Since every node has a set of samples associated with it, it is crucial to first associate samples to the modular level based on the frequency of occurrence of samples across the nodes within a module. The sample association score as described in [103] is used for this association.

Definition 12 (**Sample Association Score**). Sample Association Score ($L\left(s_j \in m_i\right)$) is the likelihood that a sample s_j belongs to a module m_i.

Let p be the total number of samples, s_i be the number of samples in a context motif M_i; let m_i be a module made of context motifs $\{M_1, M_2, \ldots, M_m\}$ and the sample s_j is included only in a subset of m_i, $\mathscr{M}^{(j)} \subset \mathscr{C}$. Then, the sample association score is defined as:

$$L\left(s_j \in \mathscr{C}\right) = \frac{\sum_{s_j \in M_i} w(M_i)}{\sum_i w(M_i)}.$$

where

$$w(M_i) = \sqrt[K]{\left(1 - \left(\frac{k_i}{N}\right)^K\right)}, \quad 1 \leq K.$$

Based on this definition I note that $0 \leq L\left(s_j \in \mathscr{C}\right) \leq 1$, where $L\left(s_j \in \mathscr{C}\right) = 0$ indicates no appearance of the sample in any context motif, while $L\left(s_j \in \mathscr{C}\right) = 1$ indicates the presence of the sample in every context motif. The parameter K controls the context-specificity of sample membership to a given module – the higher the K, the more context specific the sample membership. For this application I set $K = 2$. Only samples that had a sample association score > 0.9 were considered to be part of a functional module.

Similarly, the probability that the sample s_j belongs to the module $\mathscr{C}$ can be computed as:

$$Pr\left(s_j \in \mathscr{C}\right) = 1 - \sum_{M_j \in C} p_j^{I\left(s_j\right)} \left(1 - p_j\right)^{1 - I\left(s_j\right)}$$

where $p_j = n_j/N$ and $I\left(s_j\right) = 1$ if $s_j \in m_j$, 0 otherwise. In this study, the probability $Pr\left(s_j \in \mathscr{C}\right)$ was set to 0.05.

Only samples that had a sample association score > 0.8 were considered to be part of a module. Following this, the modules were analyzed for enrichment of specific clinical categories using the hypergeometric test. False discovery rate correction was applied using Benjamini and Hochberg's correction method [106].

Survival Analysis

When survival information was available for the dataset, Kaplan-Meier survival analysis [125] was performed on the samples belonging to each of the functional module with respect to all other samples within the dataset. The chi-squared test was used to identify significant differences in survival.

Drug Annotations

Additionally, drug associations were also studied using drug target information from Drugbank [62]. Drug targets were extracted from the Integrated Druggable list (from Sophic Alliance [1]) consisting of an integrated list of genes determined druggable. The list contains information from four sources, of which two sources were used - the list published by Hopkins & Groom [126] and by Wishart [62] as they were the most reliable.

In summary, Figure 4.3 illustrates the procedure for identification of functional modules from a data-driven network and the nature of the inputs and outputs at each stage.

Application: Refractory Cancer (Target Now)

I now demonstrate the biological application of the methods developed in this dissertation. I use the Target Now (TN) [127] cancer data set. The Target Now study was conducted on refractory cancer patients who did not benefit from standard types of treatment. Late stage cancer is frequently de-differentiated, having lost a great deal of the specialized functions present in the tissue from which it arose. The TN study aims to determine if the patients could derive benefit from therapy with a drug not normally used for their particular form of cancer [127].

The Target Now gene expression profiling experiments were conducted using the Agilent 011521 Human 1A Microarray G4110A platform. Table 4.2 shows the number of samples corresponding to each tumor type. The study consists of 146 patients, spanning 35 different types of tumor. The dataset was filtered based on the

[1] http://www.sophicalliance.com/

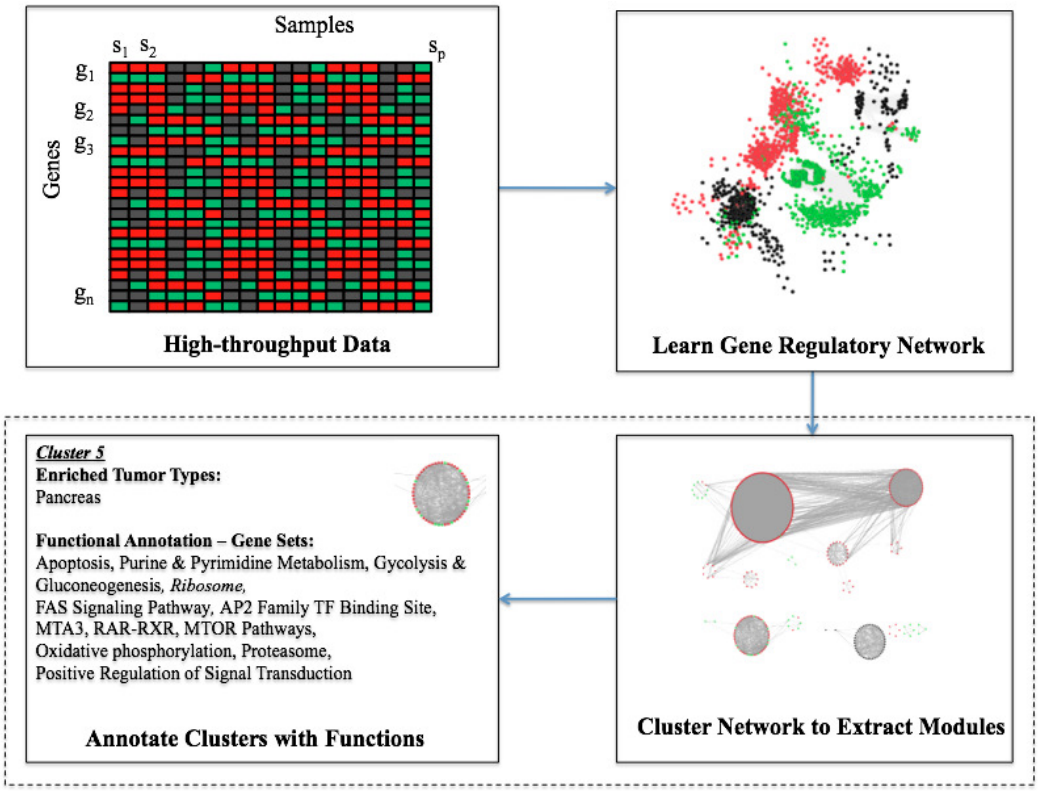

Figure 4.3: Flowchart for the identification of functional modules from a data-driven network. The region enclosed within a dashed rectangle indicates the contributions of this work.

transcription activity of each gene across samples, and reduced to 3,851 genes by eliminating genes with a low variance across samples.

Table 4.2: Distribution of samples in the Target Now data set.

Pancreas	20	Colon	7	Brain	4	Cervical	3	Esophagus	2
Ovarian	19	Kidney	6	Lung	4	Gallbladder	3	Skin	2
Melanoma	18	Salivary	6	Adipose	3	Rectal	3	T Cell	2
Breast	16	Adrenal	5	Bladder	3	Stomach	3	Thyroid	2
Single Sample: Appendix, Cartilage, Chondrosarcoma, Prostate, Testicular, Glioma, Gastric, Ileum, Lymphoma, Monocytes, Eccrine Adenocarcinoma, Rhabdomyosarcoma, Synovial Cell Sarcoma, Skeletal Muscle, Uterus									

The context mining algorithm was applied using a crosstalk < 0.3, conditioning < 0.1 and statistical significance < 0.05. Further, for each context motif (with x genes) the probability of obtaining a context-motif of x genes or more by chance, was computed, and context-motifs with a statistical significance greater than 0.01 were filtered out. Subsequently Markov clustering was applied using an inflation of 1.4 and a total of 28 functional modules were obtained.

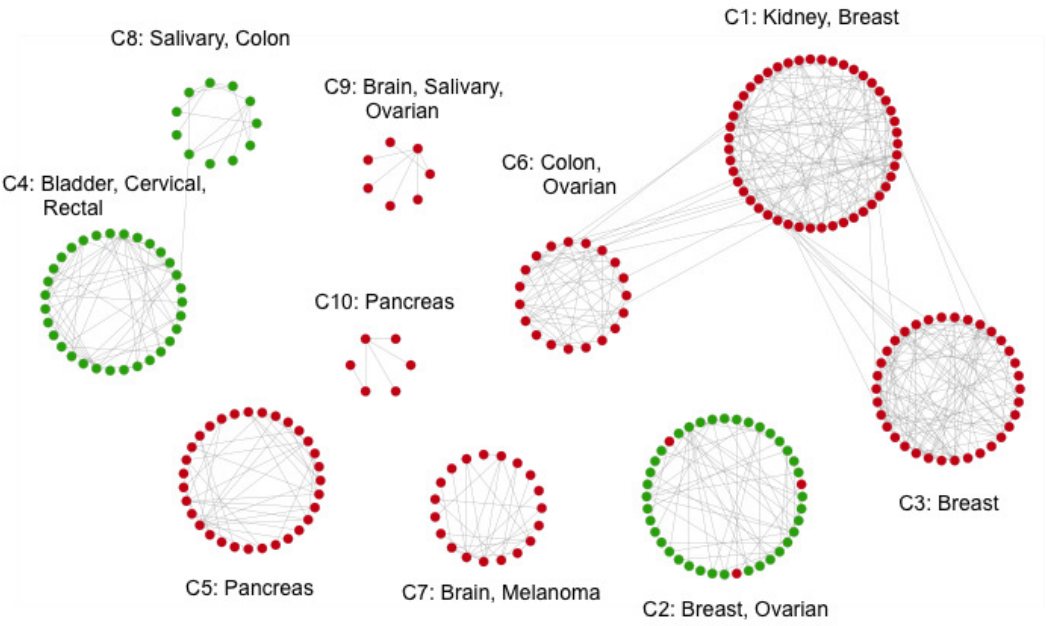

Figure 4.4: Functional modules in refractory cancer (TN).

Enriched tumor types and functional annotations were identified corresponding to these functional modules. Single sample tumors were omitted from the phenotypic enrichment analyses and a statistical significance threshold of 0.05 was used for all functional enrichment. Of these, functional modules with fewer than 10 genes and fewer than 10 samples were eliminated, and the resulting functional modules are shown in Table 4.3. As seen in the table, several interesting tumor type associations were identified.

Starting with module 1, enriched with kidney and breast tumors, it is heartening to see the Wnt/Ca2 signaling pathway enriched in this functional module, confirming the importance of the role of this pathway in multiple tumor types. Amongst the genes, DHFR was identified as a potential drug target, currently being targeted by the drug methotrexate [128, 129] in multiple tumors including renal and breast tumor. Within this functional module, I also identified four other genes to be druggable (shown in Table 4.3). Of particular interest is the gene, PDE4C, which is targeted by the drugs Ketotifen and Iloprost. The randomized phase II trial (NCT00084409) studied how well Iloprost works in preventing lung cancer in patients who are at high risk for this disease.

Module 5, enriched with pancreatic tumor contains, in addition to RRM2, the druggable gene TYMS, a well-known target for several drugs (Raltitrexed, Gemcitabine, Fluorouracil, Pemetrexed, Capecitabine) currently being used to treat the disease. This module also contains the druggable gene PSMB1 (targeted by Bortezomib).

Within module 7, apart from the genes C1QA, C1QB and FCGR3A, currently being targeted by several anti-neoplastic drugs, it is interesting to note the presence of the gene ITGB2 currently being targeted by Simvastatin [130].

Finally, in module 10, of interest are the genes C1R, C1S and PDGFRA all being currently targeted by several anti-neoplastic drugs. Additionally, it is interesting to note the gene PROS1, targeted by the drugs, Drotrecogin alfa (currently being investigated for treating sepsis associated with cancer [131]) and Menadione (currently being used to treat the disease [132]).

Table 4.3: Functional modules in refractory cancer (TN). Table shows the functional modules extracted from the Target Now dataset, along with their enriched clinical annotations and enriched pathways. Functional modules with at least 10 genes and 10 samples are listed here. Tumor types in italics indicate tumor types enriched with an adjusted p-value < 0.05. Acronyms indicate sources of pathway annotations N: NCI Nature Pathways, R: Reactome, GO: Gene Ontology, ST: Signaling Transduction KE. Genes with anti-neoplastic drugs are indicated by a $*$.

Module	Enriched Tumor Types	Enriched Functions	Drivers with Existing Drugs	Druggable Genes
C1	Kidney, Breast	WNT Ca2 Cyclic GMP Pathway [ST]	DHFR [Up]	*PDE4C**, ANKK1, PLCD3, PDE6A
C2	Breast, *Ovarian*	ERK1 ERK2 MAPK Pathway[ST], B Cell Antigen Receptor [ST], Programmed Cell Death [SA], Glucose Catabolic Process [GO], Cellular Carbohydrate Catabolic Process [GO].		CAPNS1, *GSS**, NAGK, *COMT**, EML4, *LIG1**, ALDOA, SARS, *MTR**
C3	Breast	Detection of Stimulus Involved in Sensory Perception [GO], Phototransduction [GO]	STMN4 [Up]	CYP2B6, RHO, *F7**, PLCD3, PPID
C4	Bladder, Cervical, *Rectal*	Regulation of Actin Cytoskeleton by RHO GTPases [Sig], Contractile Fiber[GO], Cell Maturation[GO], Tissue Remodeling [GO], Myofibril [GO], Actin Cytoskeleton Organization and Biogenesis [GO], Regulation of Multicellular Organismal Process [GO]		CLEC3B, MYH11, AOC3, TNXB, CTSG, COL12A1, MFAP4
C5	Pancreas	G2/M Checkpoints[R], Mitotic Prometaphase[R], Lagging & Leading Strand Synthesis[R], Assembly of the pre-replicative complex[R], Regulation of mitotic cell cycle[R], Homologous recombination repair of replication-independent double-strand breaks[R],		

Module	Enriched Tumor Types	Enriched Functions	Existing Drug Targets	Druggable Genes
C5 (continued)		G1 & S Phase[R], Homologous Recombination Repair[R], Chromosome Maintenance[R], G1/S DNA Damage Checkpoints[R], Inactivation of APC/C via direct inhibition of the APC/C complex[R], APC/C:Cdc20 mediated degradation of mitotic proteins & Securin[R], FOXM1 transcription factor network[NCI], Cyclin A/B1 associated events during G2/M transition E2F mediated regulation of DNA replication Mitotic G1-G1/S phases Cdc20:Phospho-APC/C mediated degradation of Cyclin A[R], Chk1/Chk2(Cds1) mediated inactivation of Cyclin B:Cdk1 complex[R], Cyclin D associated events in G1[R], Initiation of checkpoint signal from defective kinetochores[R], Activation of the pre-replicative complex[R], Regulation of APC/C activators between G1/S and early anaphase[R], Mitotic M-M/G1 phases[R], Inhibition of the proteolytic activity of APC/C required for the onset of anaphase by mitotic spindle checkpoint components[R], Repair synthesis of patch 27-30 bases long by DNA polymerase[R], ATM mediated response to DNA double-strand break[R], Ubiquitin Mediated Degradation of Phosphorylated Cdc25A[R], M/G1 Transition[R], Mitotic Spindle Checkpoint[R], Polo-like kinase mediated events[R], Gap-filling DNA repair synthesis and ligation in GG-NER[R], G2/M DNA replication checkpoint[R], G1/S Transition[R], DNA Replication & Pre-Initiation[R], Amplification of signal from the kinetochores[R], Double-Strand Break Repair[R], p53-Independent DNA Damage Response[R], Activation of APC/C and APC/C:Cdc20 mediated degradation of mitotic proteins[R], G1/S-Specific Transcription[R], Amplification of signal from unattached kinetochores via a MAD2 inhibitory signal[R], Mitotic G2-G2/M phases[R], p53-Independent G1/S DNA damage checkpoint[R].	RRM2 [Up]	*TYMS*, GGH, CKS1B, ATP2C1, *PSMB1**, RARS

Module	Enriched Tumor Types	Enriched Functions	Existing Drug Targets	Druggable Genes
C6	Colon, Ovarian		KCNH6 [Up]	ILDR1, ANKK1
C7	Brain, Melanoma	Lysosome [GO], Regulation of Peptidyl Tyrosine Phosphorylation [GO], Defense Response [GO], Positive Regulation of Cellular Protein Metabolic Process [GO], Regulation of Protein Amino Acid Phosphorylation[GO].	C1QB[Up]	CD163, *HCK**, ITGB2, CTSS, LTB, TNFSF13B, *C1QA**, *FCGR3A**
C8	Salivary, Colon			LYZ, LTF
C9	Brain, Salivary, *Ovarian*			TPR, PPIG
C10	Pancreas	Serine Type Endopeptidase Activity [GO], Response to Xenobiotic Stimulus [GO], Cell Recognition [GO].	C1R [Up]	ADH1B, *C1S**, ADH1C, ADH1A, *PROS1**, COL12A1, *PDGFRA**, DCN

Application: Glioblastoma Multiforme (TCGA)

Next, I study the ability of the methods developed in this chapter to identify functional modules within brain tumor. The TCGA GBM data set (described in chapter 3) was used for this purpose. The context mining algorithm was applied using a crosstalk < 0.3, conditioning < 0.1 and statistical significance < 0.01. Further, for each context motif (with x genes) the probability of obtaining a context-motif of x genes or more by chance, was computed, and context-motifs with a statistical significance greater than 0.01 were filtered out. Markov clustering was applied using an inflation of 1.4 and extracted a total of 31 functional modules were obtained.

Enriched tumor types and functional annotations were identified corresponding to these functional modules with a statistical significance threshold of 0.05. Of these, functional modules with fewer than 10 genes and fewer than 10 samples were eliminated, and the resulting functional modules are shown in Table 4.5.

Table 4.5 shows the context-specific gene regulations extracted from this data along with enriched clinical associations, enriched functional annotations. Drug associations for each functional module are shown in Table 4.5, showing specifically the anti-neoplastic drugs currently used to target the genes. It is interesting to note that C1 and C2 are enriched with NF1 and PTEN mutations respectively, previously reported to be characteristic of the Mesenchymal subtype [133]. Additionally, C7 enriched with the Proneural subtype is also enriched with the TP53 mutation.

Further, survival analysis was performed on each of the functional modules against the samples in the rest of the data and three functional modules were identified to show significant survival differences (shown in Figures 4.6, 4.7 and, 4.8). Functional module C7, enriched with the proneural subtype, shows significantly longer survival

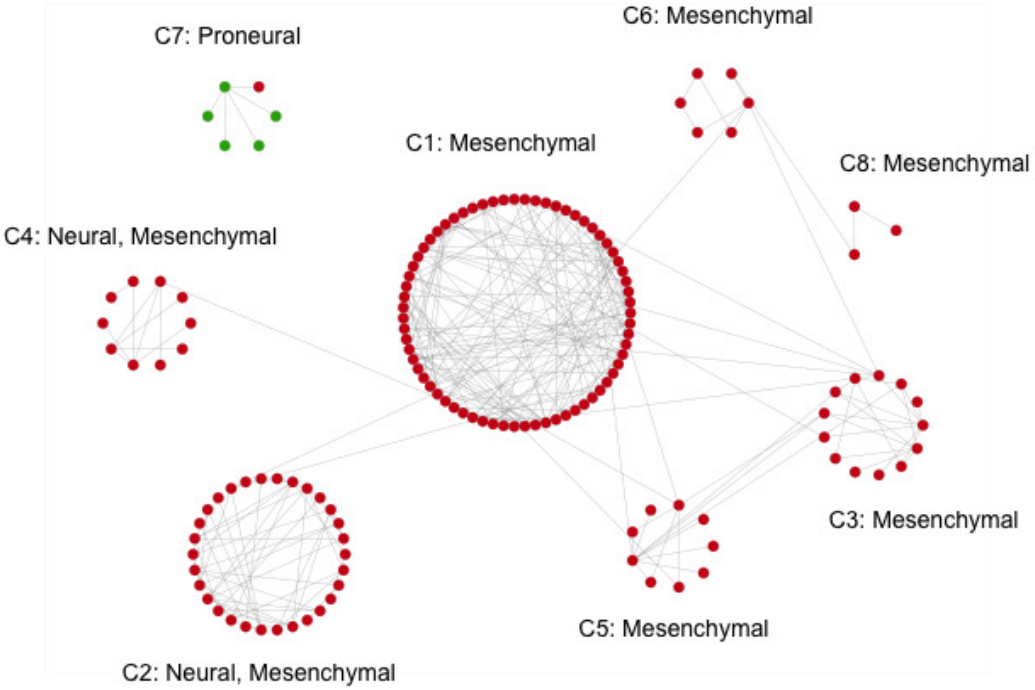

Figure 4.5: Functional modules in glioblastoma multiforme (TCGA).

than the rest of the patients, while the functional modules C2 and C4, enriched with the mesenchymal subtype show significantly shorter survival differences than the rest of the patients.

4.6 Summary

In summary, a method for the identification of functional modules in data-driven networks has been proposed. In comparison with other clustering algorithms, Markov clustering was found to be the most promising in identifying functional modules. Interestingly, the direction of the interactions did not play a role in the clustering.

Functional modules were extracted from both a refractory cancer dataset and the TCGA glioma dataset. Interesting tumor type associations and therapeutic targets were identified for each functional module within each dataset.

Table 4.4: Functional modules in glioblastoma multiforme (TCGA). Table shows the functional modules extracted from the TCGA GBM dataset, along with their enriched clinical annotations and enriched pathways. Functional modules with at least 10 genes and 15 samples are listed here. Acronyms indicate sources of pathway annotations N: NCI Nature Pathways, R: Reactome, GO: Gene Ontology, ST: Signaling Transduction KE.

Module	Mutation	Subtype	Enriched Annotations
C1	CHEK2, NF1	Mesenchymal	Pip3 Signaling In B Lymphocytes [Sig], B Cell Antigen Receptor [ST], B Cell Receptor Complexes [SA], BCR Signaling Pathway [Sig], GA12 Pathway[ST], Cytokine Secretion [GO], G Protein Signaling Adenylate Cyclase Activating Pathway [GO], Protein Secretion [GO], Protein Amino Acid N Linked Glycosylation [GO], Defense Response to Bacterium [GO], Cytokine and Chemokine Mediated Signaling Pathway [GO], Maintenance of Localization [GO], Inositol or Phosphatidylinositol Phosphodiesterase Activity [GO], Leukocyte Chemotaxis [GO], Interleukin 1 Secretion [GO], Adenylate Cyclase Activation [GO], Positive Regulation of Cytokine Biosynthetic Process [GO], Phospholipase C Activity [GO], Mesoderm Development [GO].
C2	PTEN	Neural, Mesenchymal	IL13 Pathway [ST], Leukocyte Chemotaxis [GO], RNA Activity & Catabolic Process [GO], Regulation of Viral Reproduction [GO].
C3	PIK3R1		MMP Cytokine Connection [SA], Protein Phosphatase Binding [GO], Inositol or Phosphatidylinositol Phosphodiesterase Activity [GO], Phospholipase C Activity [GO]
C4	CHEK2, RB1	Neural, Mesenchymal	Positive Regulation of Cytokine Secretion [GO], Kinase Inhibitor Activity [GO], Interleukin 1 Secretion [GO]
C5		Mesenchymal	Negative Regulation of Multicellular Organismal Process[GO], CDC42 Protein Signal Transduction [GO], Response to Steroid Hormone Stimulus [GO], Regulation of Endothelial Cell Proliferation & Synapse Structure and Activity[GO], Axon [GO], Negative Regulation of Cytokine Biosynthetic Process [GO], Hormone Secretion [GO], Pattern Recognition Receptor Activity [GO], Endothelial Cell proliferation [GO], G Protein Signaling Adenylate Cyclase Inhibiting Pathway [GO], Kinase Activator Activity [GO], Regulation of RAS GTPase Activity [GO], Regulation of RHO GTPase Activity [GO].

Module	Mutation	Subtype	Enriched Annotations
C6	CHEK2, PTEN, RB1	Mesenchymal	LPS transferred from LBP carrier to CD14 [R], Pattern Recognition Receptor Activity [GO], Creation of C4 and C2 activators [R], Classical antibody-mediated complement activation [R], Initial triggering of complement [R], LPS transferred from LBP carrier to CD14 [R], Signaling in Immune system [R], Innate Immunity Signaling [R], Complement cascade [R].
C7	A2M, ADAM12, ADM, AIFM1, ALK, ANXA1, ASPM, ATM, ATR, AVIL, BAMBI, BCL11A, BMPR1A, BRCA1, BRCA2, C22orf24, CD46, CDKN2A, CHI3L2, CHL1, CSNK1E, CTNNB1, CYLD, DGKD, DHTKD1, DMBT1, DOCK8, DST, EP300, EPHA2, FBXW7, FN1, FURIN, GATA3, GCLC, GRN, GYPC, ID3, ILK, INHBE, ITGB2, KLF4, KLF6, KLK8, LAX1, LTF, MAP3K6, MAPK7, MAPK9, MDM4, MEOX2, MET, MLL4, MN1, MSH6, MYCN, NEK10, NMBR, NOS3, PI15, PMS2, PRKD2, PRKDC, PROX1, SERPINE1, SLC2A2, SMAD2, SMAD4, SNF1LK2, SOCS1, SPARC, STAT3, STK32B, STK36, TAS1R1, TBK1, TGFBR2, TNFRSF11B, TP53, TRIM33, TSC1, VAV2, ZEB1, ZNF384	Proneural	
C8	RB1	Mesenchymal	Positive Regulation of Cytokine Secretion [GO], Interleukin 1 Secretion [GO]

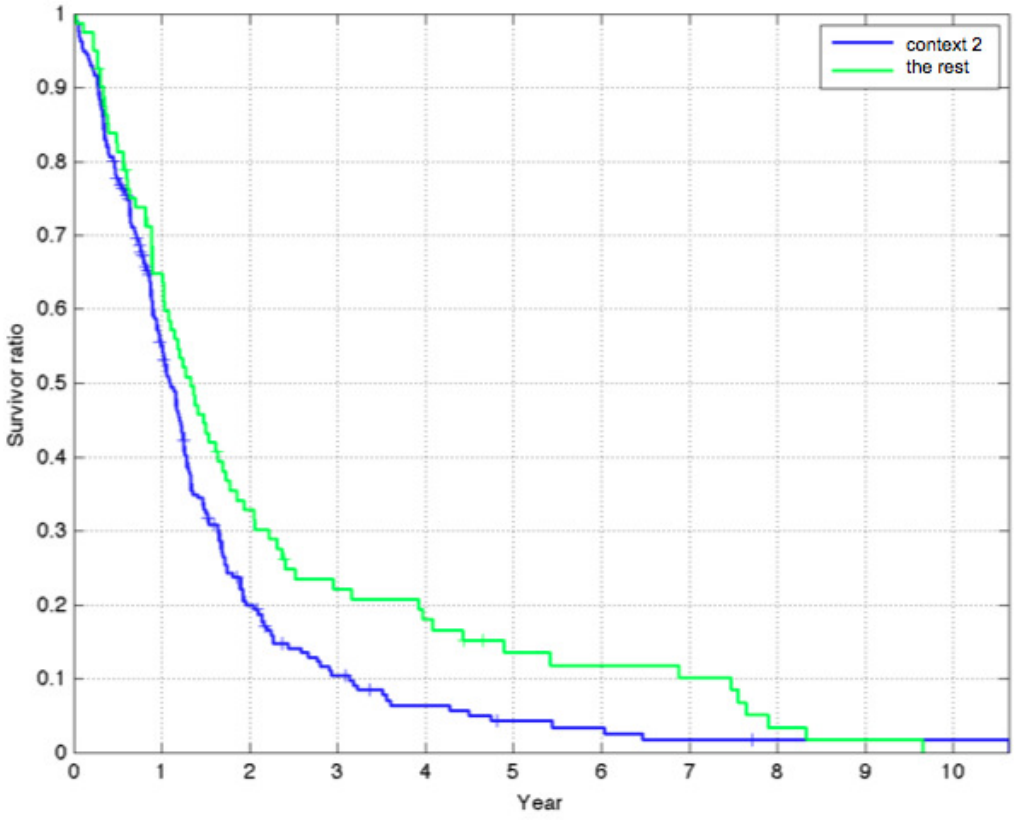

Figure 4.6: Kaplan Meier curve showing survival of C2 against the rest.

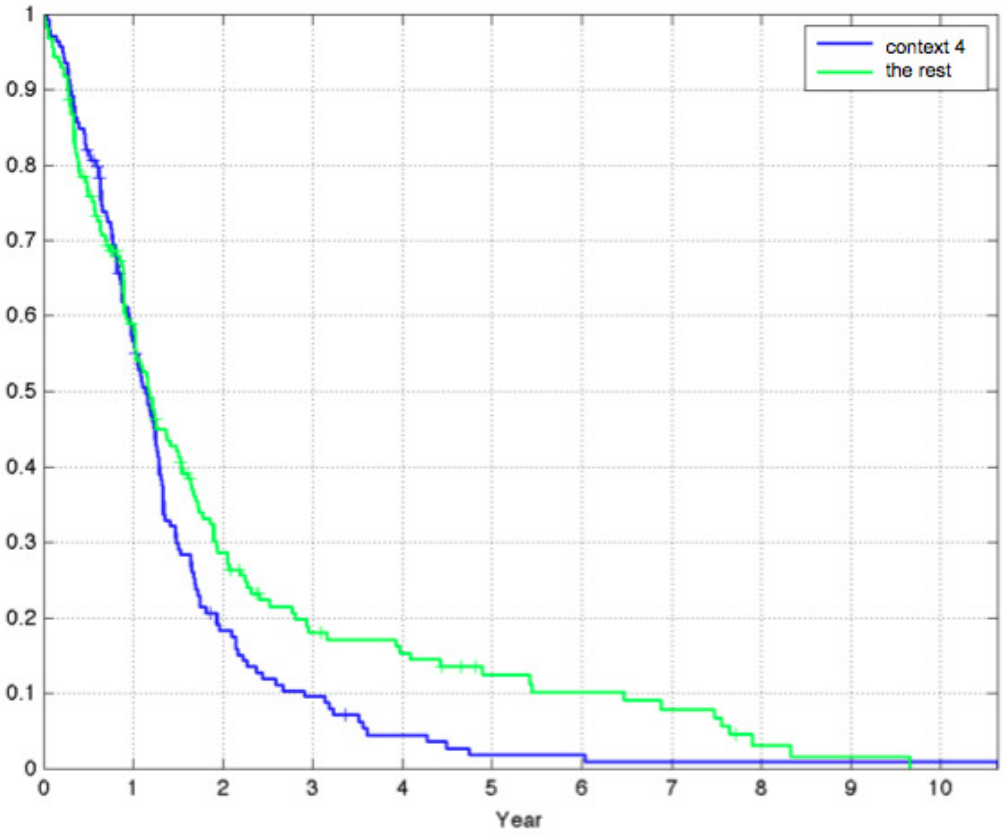

Figure 4.7: Kaplan Meier curve showing survival of C4 against the rest.

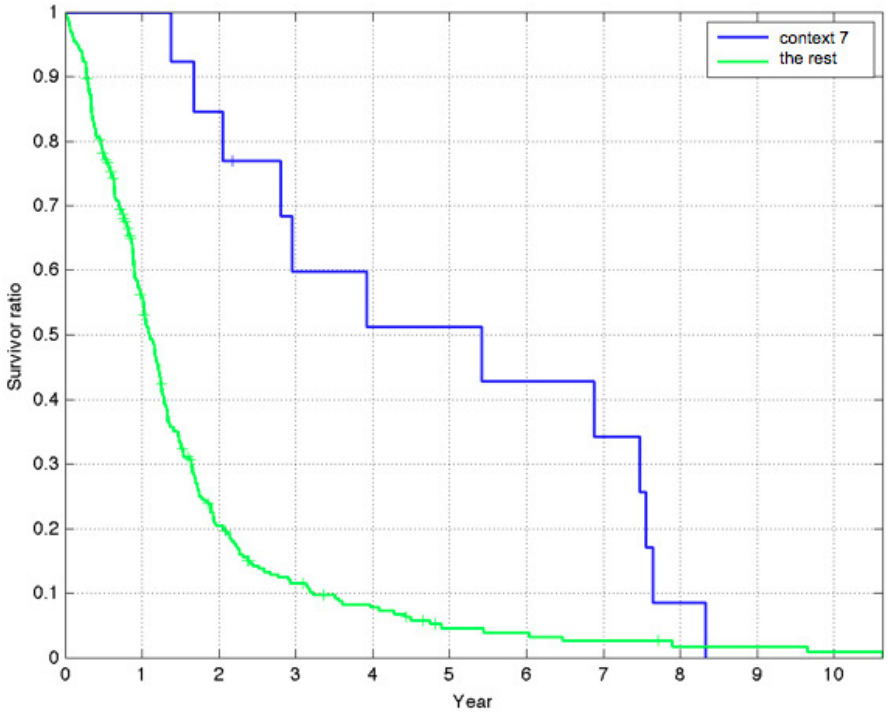

Figure 4.8: Kaplan Meier curve showing survival of C7 against the rest.

Table 4.5: Drug targets for the functional modules (TCGA).

Module	Drug Targets	Drugs
C1	CD33 [Up]	Gemtuzumab ozogamicin;
	SYK [Up]	Sti-571;
	ALOX5 [Up]	Leflunomide;
C2	FCGR2A [Up]	Cetuximab; Gemtuzumab ozogamicin; Trastuzumab; Rituximab; Ibritumomab; Tositumomab; Alemtuzumab; Bevacizumab;
C6	C1QC [Up]	Cetuximab; Gemtuzumab ozogamicin; Trastuzumab; Rituximab; Ibritumomab; Tositumomab; Alemtuzumab; Bevacizumab;
	FCGR3A [Up]	Cetuximab; Gemtuzumab ozogamicin; Trastuzumab; Rituximab; Ibritumomab; Tositumomab; Alemtuzumab; Bevacizumab;
	C1QA [Up]	Cetuximab; Gemtuzumab ozogamicin; Trastuzumab; Rituximab; Ibritumomab; Tositumomab; Alemtuzumab; Bevacizumab;

Chapter 5

LEARNING RELATIONSHIPS BETWEEN BIOLOGICAL PROCESSES

While the previous chapter dealt with identifying functional modules within a high-throughput data set, often we are interested in a higher level of abstraction – determining relationships between the biological processes that are active within a group of patients. In this chapter, I discuss the motivation for studying relationships between biological processes, relevant work in this area, the methodology developed in this dissertation and the relationships between biological processes, identified within a glioma data set.

5.1 Motivation

Biological processes, analogous to functional modules (described in the previous chapter) represent groups of genes or proteins with a common purpose. Understanding relationships between biological processes, specifically, the co-occurence of biological processes allows biologists to focus on the broad-level pathways that are active within a given phenotype. Additionally, it also permits discovery of potentially novel relationships between biological processes.

This chapter develops a method to identify the biological processes that are active within high-throughput data and understand the variations in activity across different sub-sets of patients. While the previous chapter resulted in functional modules for biologists to focus on, here, I develop a mechanism which allows biologists to identify a set of co-occurring biological processes that they are interested in. Further, the method developed here uses not only data-driven information on the activity of a biological process but also existing semantic relationships between biological processes.

5.2 Relevant Work

In this section, I outline methods which use relationships between biological processes or pathways. A few approaches have been developed for scoring pathways over-represented in data. Such methods are similar to annotation approaches and additionally incorporate pair-wise relationships between the genes within the pathways in scoring the pathways. For example, Rahnenfuhrer *et. al.* [134] propose a method to rank pathways within an expression dataset using several similarity measures (correlation, covariance, dot product and cosine similarity) applied between pathways. Gene expression data has also been used to rank pathways significantly represented within the data [135, 136]. However, such methods identify pathways by observing the activity of a pathway across all samples. Biological processes are heterogeneous, implying that the same pathway could be activated in completely different ways in different samples. Hence, it is essential to understand pathway activity on a per-sample basis.

Pathway activity in gene expression has been studied on a per-sample basis using signal transduction pathways [137], however the approach allows viewing the expression of individual pathways one at a time, rather than learning relationships between pathways.

Chagoyen *et. al.* [138] quantify the functional coherence of Gene Ontology biological process terms using the Gene Ontology annotations of a literature-derived protein-protein interaction network for yeast. Statistically significant functional connections are then extracted. While this approach provides a mechanism to extract novel functional relationships as well as analyze a literature-driven network, this approach cannot be applied to identify relationships from high-throughput data.

A few methods have been developed in order to study the co-occurrence of functional terms. Del pozo *et. al.* [139] propose a method to derive functional distances between Gene Ontology terms using the simultaneous occurrence of terms within gene annotations using the cosine similarity metric. Similarly the QuickGO browser [1] allows for viewing similar terms based on other terms that a protein is annotated with. However, the key limitations of such approaches is the lack of integration of biomedical data in learning relationships between the biological processes. Consequently, such methods only allow for learning the relationships between existing biological processes. Learning novel associations between biological processes (or pathways) is difficult to achieve.

This chapter focuses on developing a method to identify co-occurrence relationships between biological processes using both existing relationships as well as the over-representation of the biological processes within high-throughput data. Several contributions are made in this chapter. Firstly, relationships between biological processes are captured by exploiting the biological processes active in each sample individually, accounting for heterogeneity within the data. Secondly, existing knowledge of relationships between biological processes is used to guide the process.

5.3 Problem Formulation

In this section, I mathematically formulate the problem of learning relationships between biological processes. Given an $n \times p$ high throughput dataset $D = (B, S)$ where

- $B = (b_1, b_2 ... b_n)$ is the set of n genes (cellular/biological entities).
- $S = (s_1, s_2 ... s_p)$ is the set of p heterogeneous samples (or patients).

[1] $http://www.ebi.ac.uk/QuickGO$

- $D(g_i, s_j) \in \mathbb{R}$ represents the activity of g_i in sample s_j .

I assume biological knowledge is in the form of a set of terms $P = \{t_1, t_2, \cdots, t_k\}$ where each term t is an element in G, a *directed* graph. In this chapter, Gene Ontology (GO) is considered as a biological knowledge source.

Using this, I am interested in identifying one or more sets of co-occurring biological processes $P_s = \left\{t_1^{(s)}, t_2^{(s)}, \cdots, t_k^{(s)}\right\}$ within D. Similar to functional modules (described in chapter 4), such sets could be associated with clinical annotations. However, unlike functional modules, these co-occurring biological processes are not identified within a data-driven network but instead directly from the high-throughput data.

5.4 Identifying Patient-specific Biological Processes

I now turn to the methodology used to identify co-occurring biological processes. However, first it is essential to identify the biological processes active in a single sample.

Given a set of x co-regulated genes for a specific sample (both up- and down-regulated), out of a total set of X genes, I am interested in the ratio of co-regulated genes that are annotated by the GO term(y) to the total number of genes annotated by the GO term(Y). The ratio y/Y is a measure of the extent to which the genes annotated by a GO term are present within the co-regulated genes and is termed the enrichment ratio.

The statistical significance of this ratio is assessed using the hypergeometric test. This test is used as I sample from the data without replacement. The probability of

randomly obtaining y or more genes enriched within a gene list is computed as

$$p = \sum_{i=y}^{x} \frac{\binom{X-y}{x-i}\binom{y}{i}}{\binom{X}{x}} \tag{5.1}$$

For a given sample, a set of terms P_1 is found by extracting the statistically significant terms (p-value lesser than 0.05). The set P_a comprises a set of GO terms $\left\{t_1^{(a)}, t_2^{(a)}, \cdots, t_q^{(a)}\right\}$ and enrichment ratios $\left\{e_1^{(a)}, e_2^{(a)}, \cdots, e_q^{(a)}\right\}$ for the GO terms.

A naive approach would be to build a matrix of the activity of the GO terms across the samples and then cluster this matrix. However, this approach fails to take into account the existing relationships between GO terms. Hence, I move towards developing a similarity metric between samples based on the GO terms activated in each sample.

5.5 Quantifying the Similarity between Patients

Let $P_1 = \left\{t_1^{(1)}, t_2^{(1)}, \cdots, t_q^{(1)}\right\}$ and $P_2 = \left\{t_1^{(2)}, t_2^{(2)}, \cdots, t_m^{(2)}\right\}$ denote sets of GO terms where each term t is an element in G, a *directed* graph.

The similarity of the two sets P_1 and P_2 is defined as

$$\eta(P_1, P_2) = \kappa * S_K(P_1, P_2) + (1 - \kappa) * S_D(P_1, P_2) \tag{5.2}$$

where $S_K(P_1, P_2)$ is the knowledge-driven similarity between P_1 and P_2 obtained from the Gene Ontology, $S_D(P_1, P_2)$ is the data-driven similarity between P_1 and P_2 obtained from the enrichments of the terms contained in P_1 and P_2 within the high-throughput data and κ is a parameter that controls the relative influence of data and knowledge.

Data-driven Similarity

The data-driven similarity between the two samples can be computed using the correlation of the expression of the genes in the dataset across the two samples. Ideally, the high-throughput data set could be filtered to focus on the most variable genes.

Knowledge-driven Similarity

The knowledge-driven similarity between the two terms could be computed in multiple ways – either by using the enrichment ratios of the GO terms that the two samples are annotated with, or through the semantic similarity between the GO terms that the two samples are annotated with.

Enrichment Similarity

One way of expressing the knowledge-driven similarity is to capture the over-representation of the terms within the high-throughput data. For a pair of samples, two sets of terms P_1 and P_2 are found by extracting the statistically significant terms (p-value lesser than 0.05) using the method previously described. Each set P_a is associated with a set of GO terms $\left\{t_1^{(a)}, t_2^{(a)}, \cdots, t_q^{(a)}\right\}$ and enrichment ratios $\left\{e_1^{(a)}, e_2^{(a)}, \cdots, e_q^{(a)}\right\}$ for the GO terms.

Definition 13 (**Enrichment Similarity**)**.** The enrichment similarity, $E(P_1, P_2)$ represents the extent to which the two sets of terms are both over-represented within a high-throughput data set.

The enrichment similarity is found by

$$E(P_1, P_2) = 1 - [\prod_{i=1}^{m}(1 - \sqrt{e_1^{(i)} * e_1^{(i)}})] \tag{5.3}$$

where $m = |P_1 \cup P_2|$. The enrichment similarity is formulated similar to the knowledge overlap used in chapter 4.

Semantic Similarity

On the contrary, the knowledge-driven similarity could also be expressed through the semantic similarity. The semantic similarity between two sets of terms represents the extent to which the two sets of terms are referring to the same biological concepts. Several methods have been developed for studying the semantic similarity between terms and extended to sets of terms. Such methods are broadly categorized into graph-based similarity measures and information-content based methods. Graph-based methods use the directed acyclic graph encoding of an ontology in order to compare terms. The semantic value of a given term is computed based on the aggregate contribution of all the terms within the DAG, such that terms that are closer to a term t contribute more to its semantics [140]. Alternately, information-theoretic methods of semantic similarity have been addressed in several studies [141, 142]. The idea behind such methods is to utilize the usage of terms within the corpus. Comparisons of the two classes of methods [143, 144] have shown that information-theoretic methods, specifically Resnik's method correlates with gene sequence similarities and gene expression profile better. Hence, in this work, the semantic similarity is derived from information theory principles.

Prior to computing the semantic similarity between two sets of terms, it is first necessary to quantify the information contained within a single term t.

Definition 14 (Information Content). The information content ($i(t)$)) for a Gene

Ontology term quantifies the semantic content within the term.

For a term t in G, this is defined as

$$i(t) = 1 - \frac{log|A_t|}{log|\cup_{k=1}^{n} A_k|} \tag{5.4}$$

where A_t is the set of genes annotated with the term t and n is the total number of terms in G. The idea behind this equation is that a term which annotates a high proportion of genes would be a fairly common term and hence its semantic content would be low [142]. On the contrary, a term which annotates few genes would be one which is much less common and hence more meaningful.

Table 5.1: List of the annotation weights for the GO evidence codes.

Category	Weight
Experimental :	
EXP, IDA, IPI, IMP, IGI, IEP, TAS	1
Computational:	
ISS, ISO, ISA, ISM, IGC, IBA, IBD, IKR, IRD, RCA	0.75
Author Statement / Automatically Assigned :	
NAS, IC, ND, IEA	0.50

Gene annotations are usually curated using different methods, each with varying reliabilities. GO uses evidence codes to assign reliabilities to each annotation and I incorporate this in the information content. First numeric weights are assigned for each evidence code as shown in Table 5.1. Subsequently, given $A_t = (b_1, b_2, \cdots b_p)$ a set of p genes annotated with the term t and the reliability of each annotation $(r_1, r_2, \cdots r_p)$, the size of the set A_t could then be modified to be computed as the sum of the reliabilities of the terms within the set.

$$|A_t| = \sum_{i=1}^{p} r_i \tag{5.5}$$

After this, the semantic similarity between two terms is defined by tracing the path of each term to the root of G [141].

Definition 15 (**Semantic Similarity**)**.** The semantic similarity, $d(t_1,t_2)$ for the Gene Ontology terms t_1 and t_2 is the extent to which the two terms refer to the same biological concept.

$$d(t_1,t_2) = \frac{2 * \max_{t \in \gamma(t_1,t_2)} i(t)}{i(t_1) + i(t_2)} \tag{5.6}$$

where γ is the set of common subsumers to both terms t_1 and t_2. In the case of GO, I focus on the tree corresponding to biological processes. Amongst the path common to the two nodes, the information content of most specific term is used as the common information content between the two nodes. Additionally, in order to incorporate the depth of the two nodes, the information content of the two nodes themselves are also used. The intuition behind this measure is to capture what is common between the two terms. The common ancestor represents the information content that is shared by the nodes.

Finally, the semantic similarity between the two sets $S(P_1,P_2)$ is computed by taking the average of the top h proportion of all pair-wise semantic similarity values. A reasonable starting point for h would be 10 %. The intuition behind this is to capture the most specific terms, and these are represented by the terms with a high semantic similarity.

The similarity metric developed here (η) could then be applied to cluster data by taking into account not only the similarity in expression and but also relationships between the biological processes.

I now turn towards the results obtained while applying these methods to a high-throughput cancer data. GO annotation and ontology files as of February 2012

were used in this work. The ontology was filtered to remove all leaf nodes with no annotated genes or proteins.

Clustering Glioblastoma Multiforme (GBM) Patients

The methods developed here were applied to the glioblastoma multiforme data set described in [145]. This dataset consists of 181 WHO grade IV astrocytoma and 14 non-neoplastic samples from autopsy specimens of cerebral cortex from donors with no history of brain tumor or neurological disorders obtained from the National Neurological Research Brain Bank (Los Angeles, CA). The data was centered using the median of all samples and quantized using a fold change of 1. Co-regulated genes (both up- and down-regulated) were then extracted for each sample. Following this, enriched biological processes were found for each sample using the hypergeometric test and enriched terms were identified using a corrected p-value threshold of 0.05. Of the 195 samples, only 160 samples had any term enrichments and these were used in all subsequent analyses.

Clustering Samples

The first experiment involved clustering the samples using the three variants of the similarity metric, described previously – pure data-driven similarity and knowledge-driven similarity through the enrichment similarity as well as the semantic similarity. Consensus k-means clustering [146] was applied for 100 iterations, varying the number of clusters from 3 to 8. Table 5.2 shows the proportion of samples with a positive silhouette score. As seen in the table, the best performance is achieved when the semantic similarity is used to cluster the samples, at k = 3. Additionally, it is also interesting to see that a purely data-driven metric performs much worse than a knowledge-driven metric.

Table 5.2: Proportion of samples with a positive silhouette score across varying number of clusters (k).

Proportion of Samples With Positive Silhouette Score	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8
Co-expression	69%	72%	72%	69%	70%	75%
Enrichment Similarity	81%	77%	71%	67%	41%	61%
Semantic Similarity	93%	74%	69%	66%	67%	69%

Focusing on the semantic similarity as a metric for clustering, Figure 5.5 shows the consensus heatmaps for number of clusters varying from 3 – 6. The heat map shows the proportion of times that the two samples occupy the same cluster. From the heatmaps, it is clear that setting k at 3 results in clear clusters. It is also interesting to note the presence of additional sub-groups although not very clear, indicating that the data does contain additional sub-groups, however, these sub-groups could be due to a smaller sub set of the biological processes.

Silhouette width values were computed for each sample [147] and only samples with a positive silhouette width were used in further analysis. Silhouette scores reflect whether the assignment of a sample to a cluster is appropriate. Positive scores indicate that a sample is more similar to its own cluster than neighboring clusters. Negative scores indicate samples are assigned to the wrong cluster [147]. Figure 5.5 shows the silhouette plots obtained for k = 3–6. As seen in the figure and in Table 5.2, setting k at 3, resulted in the fewest number of poorly clustered samples. Hence, for all further analyses, k was set to 3 and samples with a negative silhouette score were omitted, leaving a total of 149 samples.

Biological Significance

The obtained clusters were analyzed for enriched clinical subtypes using the chi-squared test, after applying a p-value filter of 0.05. Additionally, survival analyses

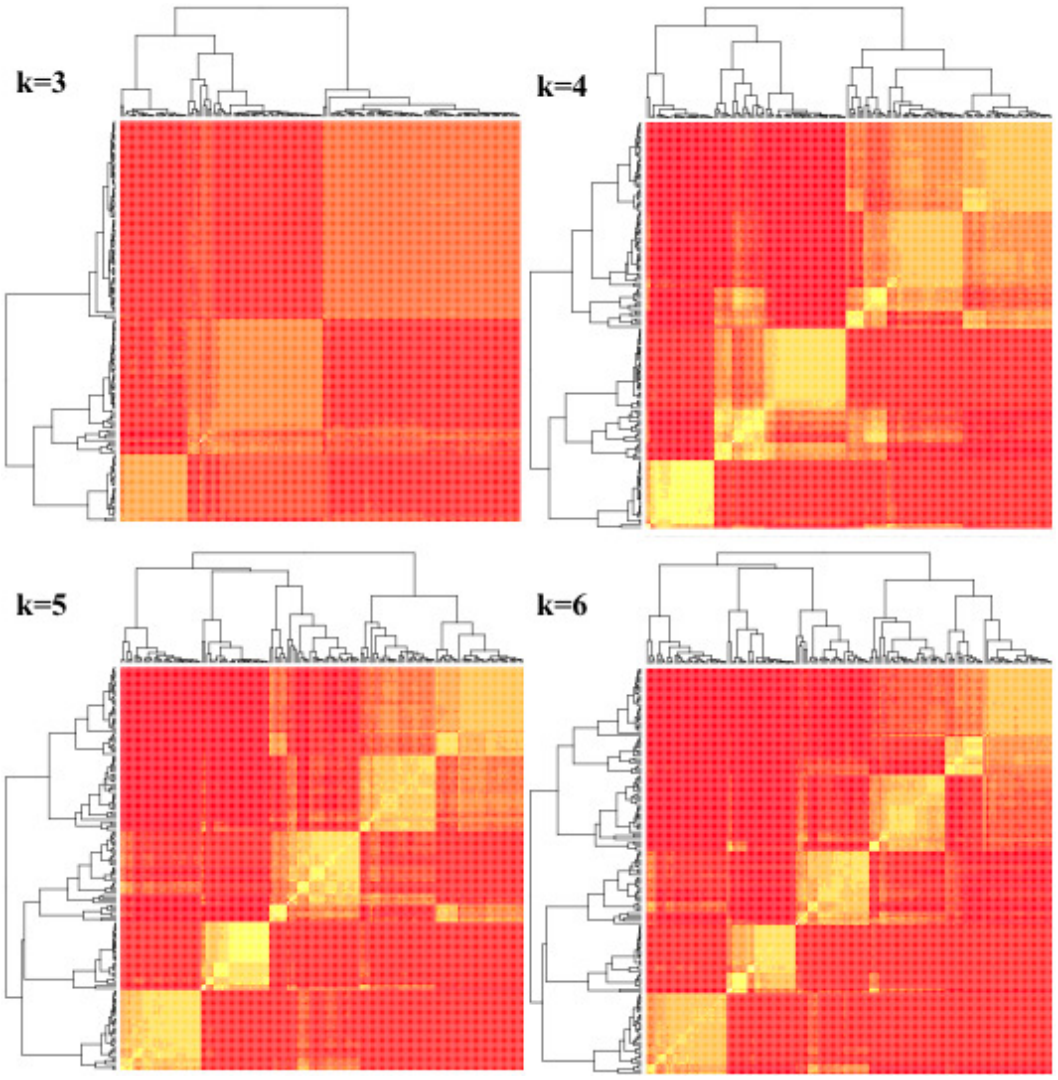

Figure 5.1: Consensus heat maps showing patient clusters in glioblastoma multiforme (GBM). Yellow indicates that the two samples have a high consensus index while red indicates a low consensus index.

was performed on the clusters and the co-occurring biological processes were identified in each cluster by extracting the GO terms which occurred most frequently across the samples within a cluster. Specifically, I extracted terms which occurred in atleast 50 % of the samples within a cluster. Table 5.3 shows a summary of the clusters. Figure 5.5 shows the Kaplan-Meier survival plots for the three clusters.

It is interesting to see that the biological significance of the clusters matches pre-

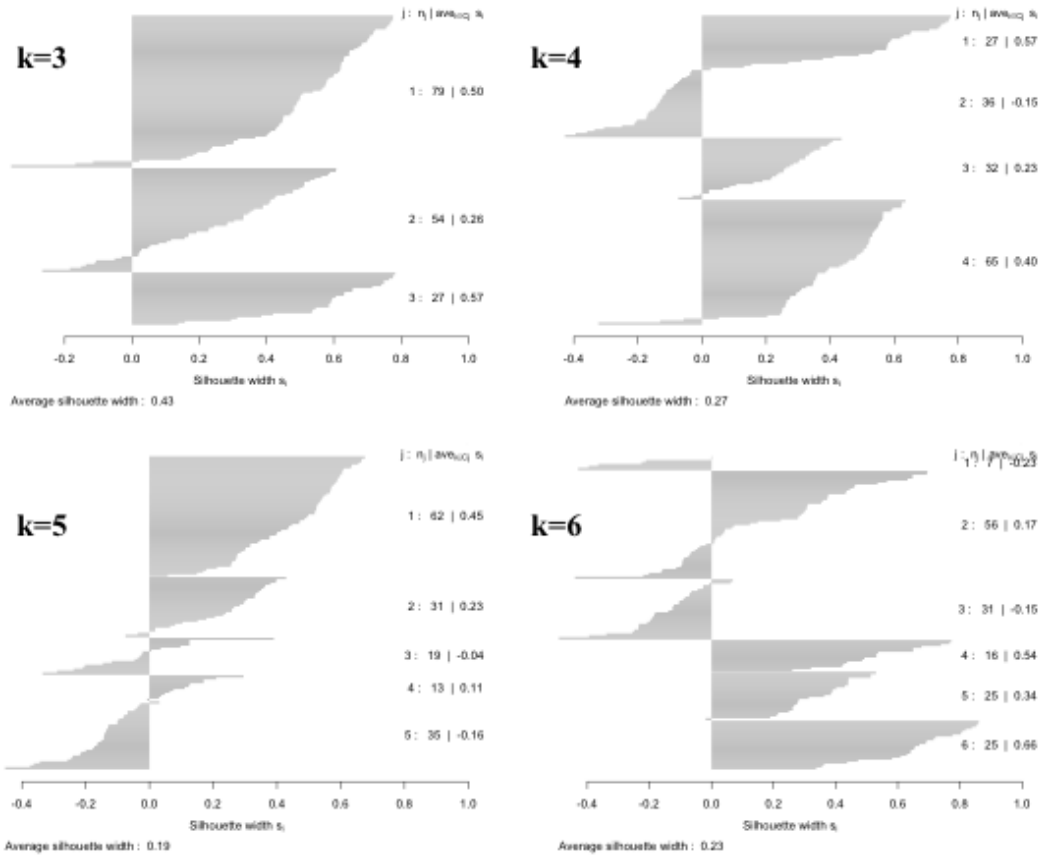

Figure 5.2: Silhouette plots for k = 3-6 showing the 'true' samples of each cluster.

viously reported findings. Survival curves were generated using the Kaplan-Meier analysis and are shown in figure 5.5. Cluster 1 enriched with the proneural subtype was found to have better prognosis, while cluster 3, enriched with the proliferative sub-type was found to exhibit poor prognosis. Contrary to previous findings however, cluster 2, although enriched with the mesenchymal sub-type, had a median survival of 1.17 years, due to the presence of a few long surviving proneural and mesenchymal samples within the cluster.

Additionally, clusters 1 and 3 showed significant survival differences with a p-value of 0.036, and clusters 2 and 3 also showed significant survival differences with a p-value of 0.008. The findings also confirm the associations between prognosis and

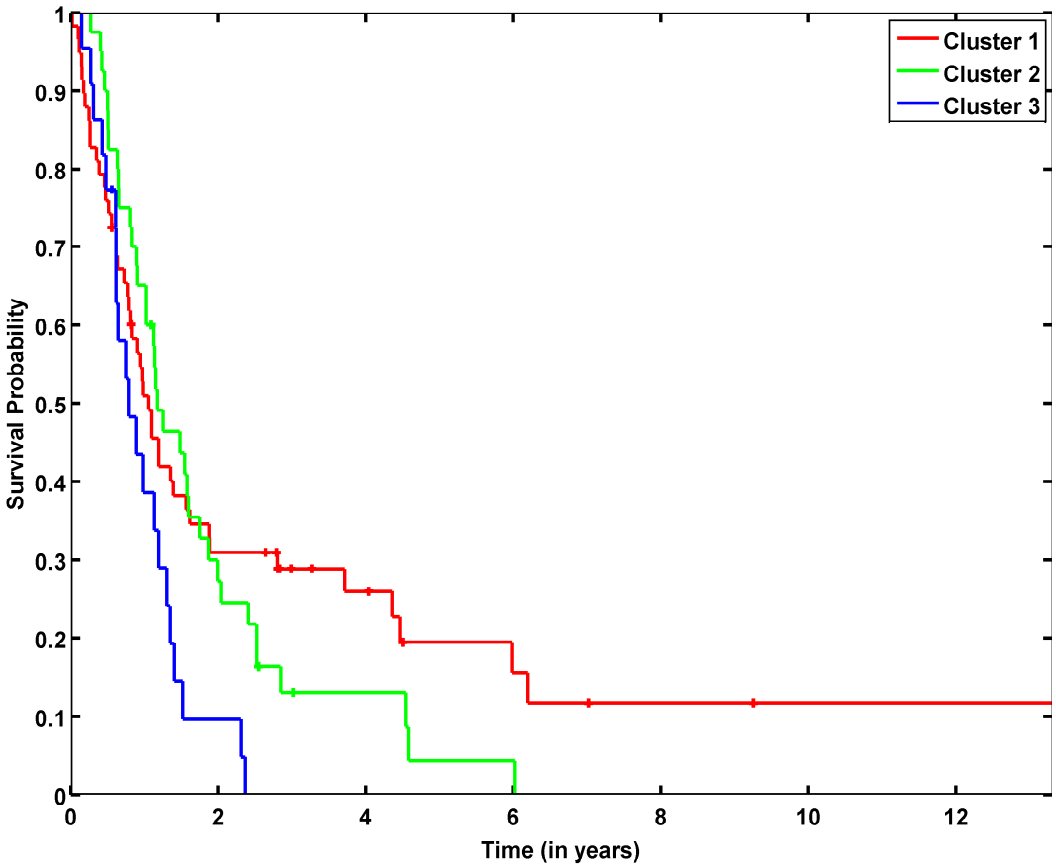

Figure 5.3: Kaplan-Meier survival curves for clusters of biological processes identified in GBM.

the different stages in neurogenesis. The better prognosis subgroup (Cluster 1) is enriched with neurogenesis biomarkers while the poor prognosis group (Cluster 3) was enriched with cell cycle biological processes indicating tumor proliferation. An interesting find is that the normal samples cluster with the proneural sub-type. Further the presence of signaling biological processes within cluster 1, associated with the proneural subtype in addition to the previously reported angiogenesis, could provide insights onto novel characteristics of the proneural sub-type.

Table 5.3: Patient cluster characteristics in glioblastoma multiforme (GBM195).

Cluster	Samples with Survival	Survival P Value (Against Rest)	Median Survival (Years)	Phillips *et. al*. Subtype Enrichment	GO Terms
C1	58	0.13	1.06	Proneural, Normal	nervous system development, neurogenesis, generation of neurons, neuron differentiation, neuron projection morphogenesis, neuron development, cell morphogenesis involved in neuron differentiation, neuron projection development, cell morphogenesis involved in differentiation, axonogenesis, synaptic transmission, transmission of nerve impulse, multicellular organismal signaling, regulation of neurotransmitter levels, neurotransmitter transport, neurotransmitter secretion
C2	40	0.86	1.17	Mesenchymal	defense response, antigen processing and presentation of peptide or polysaccharide antigen via MHC class II, response to wounding, immune system process, immune response, response to biotic stimulus, response to interferon-gamma, response to other organism, response to cytokine stimulus, cellular response to interferon-gamma, inflammatory response, innate immune response, adaptive immune response, adaptive immune response based on somatic recombination of immune receptors built from, immunoglobulin superfamily domains, immunoglobulin mediated immune response, B cell mediated immunity, interferon-gamma-mediated signaling pathway
C3	22	0.01	0.79	Proliferative	mitotic prometaphase, mitotic cell cycle, M phase of mitotic cell cycle, nuclear division, mitosis, cell cycle phase, organelle fission, M phase, cell cycle process, chromosome segregation, cell cycle checkpoint, regulation of cell cycle process, cell division, cell cycle, regulation of cell cycle arrest, sister chromatid segregation,

Cluster	Samples with Survival	Survival P Value (Against Rest)	Median Survival (Years)	Phillips *et. al*. Subtype Enrichment	GO Terms
C3 (continued)					DNA replication, DNA replication-independent nucleosome assembly, spindle organization, chromatin remodeling at centromere, CenH3-containing nucleosome assembly at centromere, DNA replication-independent nucleosome organization, histone exchange, DNA packaging, interphase, interphase of mitotic cell cycle, regulation of cell cycle, mitotic sister chromatid segregation, spindle checkpoint, chromosome organization, DNA-dependent DNA replication, DNA conformation change

Effects of Relative Proportion of Knowledge and Data on Clustering

Finally, an experiment was performed to study how varying the amount of knowledge (against data) would affect the clusters. Equation 5.2 was applied by varying κ between 0 and 1. When $\kappa = 0$, this represented the case when only the correlation of expression values was used for clustering. When $\kappa = 1$, this represented the case when only the semantic similarity was used for clustering. Again, the proportion of samples with a positive silhouette score was used to evaluate the clusterings and the results are shown in Table 5.4. Interestingly, as the proportion of knowledge is increased, the clusterings with k = 3 and 4 (previously reported by [145, 133]) are identified with greater accuracy.

Table 5.4: Proportion of samples with positive silhouette score across varying amounts of knowledge.

Proportion of Knowledge	k=3	k=4	k=5	k=6	k=7	k=8
0%	68%	72%	69%	68%	71%	70%
10%	71%	75%	68%	68%	66%	69%
20%	68%	79%	68%	68%	64%	60%
30%	71%	71%	68%	69%	71%	78%
40%	81%	83%	78%	77%	73%	72%
50%	94%	84%	78%	79%	77%	59%
60%	82%	88%	49%	77%	71%	75%
70%	83%	91%	76%	77%	69%	78%
80%	93%	94%	81%	73%	69%	61%
90%	94%	94%	83%	79%	64%	65%
100%	95%	74%	66%	72%	57%	59%

5.6 Summary

This chapter has dealt with a method to identify co-occurrence relationships between biological processes using both known relationships between biological processes and their over-representation within a data set. The methods developed here have

been applied to a high-throughput glioma data set. Co-occurring biological processes were extracted in the form of three clusters, enriched with previously identified biological sub-types. Further statistically significant survival differences were obtained. Additionally, potentially novel relationships between biological processes were also identified.

Chapter 6

CONCLUSION

Identifying plausible data-driven hypotheses using existing knowledge is an important problem in computational systems biology. In this dissertation, I have developed methods for addressing this problem based on hypotheses which fall into three classes – individual interactions, functional modules and relationships between biological processes. Here I summarize the key contributions of this dissertation and directions for future research.

6.1 Contributions

Individual Interactions

In Chapter 3, I have developed a method to evaluate data-driven interactions against literature-derived pathways. Using the data-driven networks from three sources (ARACNE, Bayesian networks and context-specific networks), I have shown how $\approx$ 60-70% of the interactions within these networks map onto literature-derived paths (refer Table 3.1). Further I have shown how the evaluation of the data-driven networks could shed light on several interesting properties of biological networks. For a given network size, the proportion of data-driven interactions evaluated against literature-derived pathways is similar across networks learnt using different methods. It is also interesting to note that random networks created using the same set of nodes as the data-driven network results in comparable statistics (shown in Figure 3.1 and 3.2). Additionally, I have also learnt that the integration of multiple knowledge sources plays an important role in such validation efforts, with more than 75 % data-driven edges, in all three data-driven networks, requiring a combination of two or more sources for their validation.

Since a significant proportion of the data-driven interactions do not correspond to literature-derived paths, I have developed a scoring method to use additional knowledge sources in determining the likelihood that each of these interactions are plausible novel biological interactions. The scoring method for unmapped interactions allows for the extraction of the best candidate hypotheses from the data-driven interactions which could be validated with wet-lab experiments. Using a simulated data set, I have shown that the scoring metric is able to distinguish between true and false data-driven interactions (difference in distributions is statistically significant with a p-value of $4.61 \times e^{-24}$). This metric copes with both incompleteness in biological annotations as well as missing connections in literature-derived pathways. Finally, these methods are applied to score the data-driven interactions in a glioma dataset and identify plausible novel hypotheses.

Functional Modules

Moving to a higher level of **abstraction**, in Chapter 4, I develop a method to identify functional modules from data-driven networks using graph clustering approaches. Amongst the graph clustering algorithms, I observe that Markov clustering has a tendency to extract biologically significant functional modules, even when the directionality of the data-driven interactions is ignored.

I show how the methods could be applied to a refractory cancer data set, resulting in several interesting tumor-type associations (shown in Figure 4.3 and Table 4.3) between Kidney and Breast tumor; Bladder, Cervical and Rectal; and Brain and Melanoma. Several interesting drug associations were found for the functional modules and many of the identified drug targets were shown to be associated with a drug currently treating some type of cancer, indicating that it could be used towards the tumor in question.

The methods were also applied to a the TCGA glioma data set. Functional modules were found to be enriched with previously discovered sub-types (shown in Table 4.4). Additionally, the modules enriched with the Mesenchymal sub-types were characterized by poor prognosis while the module enriched with the Proneural sub-type was characterized by better prognosis (survival curves are shown in Figures 4.6 – 4.8).

Relationships between Biological Processes

At the level of the relationships between biological processes, I have developed a method to identify co-occurrence relationships among biological processes within a high-throughput data set, capturing the variations in pathway activity across the samples within a given data set. Subsequently, I have developed an approach to evaluate the similarity between samples using both data-derived metrics as well as knowledge-derived metrics.

Interestingly, using knowledge-derived metrics for similarity proved to be more effective than data-derived metrics in clustering biological processes. The approach has been applied to the TCGA glioma data set to identify three sub-groups within the data, confirming previous findings (shown in Table 5.3). The sub-groups corresponded to the different stages in neurogenesis, with the sub-group for better prognosis (Cluster 1) enriched with neurogenesis biomarkers, while the sub-group for poor prognosis (Cluster 3) was enriched with cell cycle biological processes (shown in Figure 5.3).

The relative influences of knowledge and data on the clustering results was also studied (seen in Table 5.4), showing that overall, an increase in knowledge corresponded with better clusters, in the cases of k = 3 and 4 (previously identified sub-groups). This demonstrates that prior knowledge is an important aspect which could be used

to guide the identification of co-occurring biological processes.

6.2 Future Work

While this dissertation has developed methods for identifying plausible biological hypotheses from data-driven hypotheses, there remain several directions for further research.

This dissertation develops methods to refine hypotheses. Currently the approaches are implemented as a suite of tools. An interesting question is the large-scale application of these approaches to identify novel hypotheses from high-throughput data. Specifically, can we create a repository of data-driven hypotheses for community to both utilize and augment?

Secondly, the method developed in this dissertation have been developed as a set of independent tools that could be applied to data-driven hypotheses. As a next step, it would be interesting to study the relationships between these tools. Specifically, could the functional modules identified in chapter 4 be used as input to methods for scoring interactions in chapter 3 ? If so, could we compare the evaluation of data-driven interactions across different functional modules ?

Finally, this dissertation has used existing knowledge sources such as Gene Ontology and pathways. An interesting aspect would be extension of this dissertation to use additional sources such as transcription factors, microRNA information to understand the extent to which these knowledge sources could allow for identifying plausible biological hypotheses.

In summary, this dissertation opens up several possibilities for the effective utilization of data-driven hypotheses using existing biological knowledge, making a significant contribution to the field of knowledge discovery in systems biology.

BIBLIOGRAPHY

[1] "Atlanta: American Cancer Society, Cancer Facts and Figures 2011," 2011.

[2] B. Hennessy, D. Smith, P. Ram, Y. Lu, and G. Mills, "Exploiting the pi3k/akt pathway for cancer drug discovery," *Nature Reviews Drug Discovery*, vol. 4, no. 12, pp. 988–1004, 2005.

[3] M. Van Regenmortel, "Reductionism and complexity in molecular biology," *EMBO reports*, vol. 5, no. 11, p. 1016, 2004.

[4] M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[5] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler, "The Gene Ontology annotation (GOA) database: sharing knowledge in Uniprot with Gene Ontology," *Nucleic acids research*, vol. 32, no. suppl 1, p. D262, 2004.

[6] E. Cerami, B. Gross, E. Demir, I. Rodchenkov, Ö. Babur, N. Anwar, N. Schultz, G. Bader, and C. Sander, "Pathway commons, a web resource for biological pathway data," *Nucleic acids research*, vol. 39, no. suppl 1, p. D685, 2011.

[7] L. Hartwell, J. Hopfield, S. Leibler, A. Murray *et al.*, "From molecular to modular cell biology," *Nature*, vol. 402, no. 6761, p. 47, 1999.

[8] J. Quackenbush *et al.*, "Microarray data normalization and transformation," *nature genetics*, vol. 32, no. supp, pp. 496–501, 2002.

[9] D. Pinkel and D. Albertson, "Array comparative genomic hybridization and its applications in cancer," *Nature Genetics*, vol. 37, pp. S11–S17, 2005.

[10] A. Kallioniemi, O. Kallioniemi, D. Sudar, D. Rutovitz, J. Gray, F. Waldman, and D. Pinkel, "Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors," *Science*, vol. 258, no. 5083, pp. 818–821, 1992.

[11] R. Redon, S. Ishikawa, K. Fitch, L. Feuk, G. Perry, T. Andrews, H. Fiegler, M. Shapero, A. Carson, W. Chen *et al.*, "Global variation in copy number in the human genome," *Nature*, vol. 444, no. 7118, pp. 444–454, 2006.

[12] H. De Jong, "Modeling and simulation of genetic regulatory systems: a literature review," *Journal of computational biology*, vol. 9, no. 1, pp. 67–103, 2002.

[13] G. Karlebach and R. Shamir, "Modelling and analysis of gene regulatory networks," *Nature Reviews Molecular Cell Biology*, vol. 9, no. 10, pp. 770–780, 2008.

[14] N. Soranzo, G. Bianconi, and C. Altafini, "Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data," *Bioinformatics*, vol. 23, no. 13, p. 1640, 2007.

[15] L. Glass and S. Kauffman, "The logical analysis of continuous, non-linear biochemical control networks," *Journal of Theoretical Biology*, vol. 39, no. 1, pp. 103–129, 1973.

[16] S. Kauffman, C. Peterson, B. Samuelsson, and C. Troein, "Random Boolean network models and the yeast transcriptional network," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 25, p. 14796, 2003.

[17] S. Liang, S. Fuhrman, R. Somogyi *et al.*, "REVEAL, a general reverse engineering algorithm for inference of genetic network architectures," in *Pacific Symposium on Biocomputing*, vol. 3, 1998, p. 22.

[18] T. Akutsu, S. Miyano, and S. Kuhara, "Identification of genetic networks from a small number of gene expression patterns under the Boolean network model," in *Pacific Symposium on Biocomputing*, vol. 4, 1999, pp. 17–28.

[19] T. Ideker, V. Thorsson, and R. Karp, "Discovery of regulatory interactions through perturbation: inference and experimental design," in *Pacific Symposium on Biocomputing*, vol. 5, 2000, pp. 302–313.

[20] H. Lahdesmaki, I. Shmulevich, and O. Yli-Harja, "On learning gene regulatory networks under the Boolean network model," *Machine Learning*, vol. 52, no. 1, pp. 147–167, 2003.

[21] I. Shmulevich, E. Dougherty, S. Kim, and W. Zhang, "Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks," pp. 261–274, 2002.

[22] I. Shmulevich, I. Gluhovsky, R. Hashimoto, E. Dougherty, and W. Zhang, "Steady-state analysis of genetic regulatory networks modelled by probabilistic Boolean networks," *Comparative and Functional Genomics*, vol. 4, no. 6, pp. 601–608, 2003.

[23] R. Hashimoto, S. Kim, I. Shmulevich, W. Zhang, M. Bittner, and E. Dougherty, "Growing genetic regulatory networks from seed genes," *Bioinformatics*, vol. 20, no. 8, p. 1241, 2004.

[24] N. Friedman, I. Nachman, and D. Peer, "Learning Bayesian network structure from massive datasets: The sparse candidate algorithm," in *Proceedings of the Fifth Conference on Uncertainty in Artificial Intelligence*, 1999, pp. 206–215.

[25] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian Networks to Analyze Expression Data," *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 601–620, 2000.

[26] J. Yu, V. Smith, P. Wang, A. Hartemink, and E. Jarvis, "Advances to Bayesian network inference for generating causal networks from observational biological data," *Bioinformatics-Oxford*, vol. 20, no. 18, pp. 3594–3603, 2004.

[27] J. Pena, J. Bjorkegren, and J. Tegnér, "Growing Bayesian network models of gene networks from seed genes," *BIOINFORMATICS-OXFORD-*, vol. 21, no. 2, 2005.

[28] A. Hartemink, "Bayesian networks and informative priors: Transcriptional regulatory network models," *Bayesian inference for gene expression and proteomics*, pp. 401–424, 2006.

[29] N. NINDS, "Mining the gene expression matrix: Inferring gene relationships from large scale gene expression data," in *Information processing in cells and tissues*. Plenum Pub Corp, 1998, p. 203.

[30] A. Butte and I. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," in *Pac Symp Biocomput*, vol. 5, 2000, pp. 418–429.

[31] A. Butte, P. Tamayo, D. Slonim, T. Golub, and I. Kohane, "Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 22, p. 12182, 2000.

[32] A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Favera, and A. Califano, "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC bioinformatics*, vol. 7, no. Suppl 1, p. S7, 2006.

[33] J. Faith, B. Hayete, J. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. Collins, and T. Gardner, "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles," *PLoS Biol*, vol. 5, no. 1, p. e8, 2007.

[34] A. Joshi, R. De Smet, K. Marchal, Y. Van de Peer, and T. Michoel, "Module networks revisited: computational assessment and prioritization of model predictions," *Bioinformatics*, vol. 25, no. 4, p. 490, 2009.

[35] C. Huttenhower, K. Mutungu, N. Indik, W. Yang, M. Schroeder, J. Forman, O. Troyanskaya, and H. Coller, "Detailing regulatory networks through large scale data integration," *Bioinformatics*, vol. 25, no. 24, p. 3267, 2009.

[36] U. Akavia, O. Litvin, J. Kim, F. Sanchez-Garcia, D. Kotliar, H. Causton, P. Pochanard, E. Mozes, L. Garraway, and D. Pe'er, "An integrated approach to uncover drivers of cancer," *Cell*, 2010.

[37] S. Mukherjee and S. Hill, "Network clustering: probing biological heterogeneity by sparse graphical models," *Bioinformatics*, 2011.

[38] Y. Cheng and G. Church, "Biclustering of expression data," in *Proceedings of the International Conference on Intelligent Systems for Molecular Biology; ISMB. International Conference on Intelligent Systems for Molecular Biology*, vol. 8, 2000, p. 93.

[39] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai, "Revealing modular organization in the yeast transcriptional network," *Nature genetics*, vol. 31, no. 4, pp. 370–377, 2002.

[40] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman, "Module networks: identifying regulatory modules and their condition-

specific regulators from gene expression data," *Nature genetics*, vol. 34, no. 2, pp. 166–176, 2003.

[41] N. Friedman, "Inferring cellular networks using probabilistic graphical models," *Science*, vol. 303, no. 5659, p. 799, 2004.

[42] S. Lee, D. Pe'Er, A. Dudley, G. Church, and D. Koller, "Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification," *Proceedings of the National Academy of Sciences*, vol. 103, no. 38, p. 14062, 2006.

[43] J. Li, Z. Liu, Y. Pan, Q. Liu, X. Fu, N. Cooper, Y. Li, M. Qiu, and T. Shi, "Regulatory module network of basic/helix-loop-helix transcription factors in mouse brain," *Genome Biology*, vol. 8, no. 11, p. R244, 2007.

[44] N. Novershtern, Z. Itzhaki, O. Manor, N. Friedman, and N. Kaminski, "A functional and regulatory map of asthma," *American journal of respiratory cell and molecular biology*, vol. 38, no. 3, p. 324, 2008.

[45] S. Kim, I. Sen, and M. Bittner, "Mining molecular contexts of cancer via in-silico conditioning," in *Computational systems bioinformatics: CSB2007 Conference proceedings, volume 6, University of California, San Diego, 13-17 August 2007*. Imperial College Pr, 2007, p. 169.

[46] I. Sen, M. Verdicchio, S. Jung, R. Trevino, M. Bittner, and S. Kim, "Context-specific gene regulations in cancer gene expression data," in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*. NIH Public Access, 2009, p. 75.

[47] D. Barrell, E. Dimmer, R. Huntley, D. Binns, C. O'Donovan, and R. Apweiler, "The GOA database in 2009–an integrated Gene Ontology Annotation resource," *Nucleic acids research*, vol. 37, no. Database issue, p. D396, 2009.

[48] G. Bader, M. Cary, and C. Sander, "Pathguide: a pathway resource list," *Nucleic Acids Research*, vol. 34, no. Database Issue, p. D504, 2006.

[49] A. Subramanian, P. Tamayo, V. Mootha, S. Mukherjee, B. Ebert, M. Gillette, A. Paulovich, S. Pomeroy, T. Golub, E. Lander *et al.*, "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15 545–15 550, 2005.

[50] C. Stark, B. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic acids research*, vol. 34, no. suppl 1, p. D535, 2006.

[51] M. S.-K. C. Center, "The cancer cell map," http://cancer.cellmap.org/cellmap/home.do.

[52] P. Romero, J. Wagg, M. Green, D. Kaiser, M. Krummenacker, and P. Karp, "Computational prediction of human metabolic pathways from the complete human genome," *Genome biology*, vol. 6, no. 1, p. R2, 2004.

[53] "Imid," http://www.sbcny.org.

[54] T. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal *et al.*, "Human Protein Reference Database–2009 update," *Nucleic acids research*, vol. 37, no. Database issue, p. D767, 2009.

[55] B. Aranda, P. Achuthan, Y. Alam-Faruque, I. Armean, A. Bridge, C. Derow, M. Feuermann, A. Ghanbarian, S. Kerrien, J. Khadake *et al.*, "The IntAct molecular interaction database in 2010," *Nucleic Acids Research*, vol. 38, no. suppl 1, p. D525, 2010.

[56] A. Chatr-aryamontri, A. Ceol, L. Palazzi, G. Nardelli, M. Schneider, L. Castagnoli, and G. Cesareni, "MINT: the Molecular INTeraction database," *Nucleic acids research*, vol. 35, no. Database issue, p. D572, 2007.

[57] C. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. Buetow, "PID: the pathway interaction database," *Nucleic acids research*, vol. 37, no. Database issue, p. D674, 2009.

[58] I. Vastrik, P. D'Eustachio, E. Schmidt, G. Joshi-Tope, G. Gopinath, D. Croft, B. De Bono, M. Gillespie, B. Jassal, S. Lewis *et al.*, "Reactome: a knowledge base of biologic pathways and processes," *Genome biology*, vol. 8, no. 3, p. R39, 2007.

[59] G. Hart, A. Ramani, and E. Marcotte, "How complete are current yeast and human protein-interaction networks," *Genome Biol*, vol. 7, no. 11, p. 120, 2006.

[60] E. Sprinzak, S. Sattath, and H. Margalit, "How reliable are experimental protein-protein interaction data?" *Journal of molecular biology*, vol. 327, no. 5, pp. 919–923, 2003.

[61] K. Mrozek, N. Heerema, and C. Bloomfield, "Cytogenetics in acute leukemia," *Blood reviews*, vol. 18, no. 2, pp. 115–136, 2004.

[62] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu *et al.*, "Drugbank 3.0: a comprehensive resource for omics research on drugs," *Nucleic Acids Research*, vol. 39, no. suppl 1, p. D1035, 2011.

[63] B. Zeeberg, W. Feng, G. Wang, M. Wang, A. Fojo, M. Sunshine, S. Narasimhan, D. Kane, W. Reinhold, S. Lababidi *et al.*, "GoMiner: a resource for biological interpretation of genomic and proteomic data," *Genome Biol*, vol. 4, no. 4, p. R28, 2003.

[64] F. Al-Shahrour, R. Díaz-Uriarte, and J. Dopazo, "FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes," *Bioinformatics*, vol. 20, no. 4, p. 578, 2004.

[65] M. Robinson, J. Grigull, N. Mohammad, and T. Hughes, "FunSpec: a web-based cluster interpreter for yeast," *BMC bioinformatics*, vol. 3, no. 1, p. 35, 2002.

[66] S. Doniger, N. Salomonis, K. Dahlquist, K. Vranizan, S. Lawlor, B. Conklin *et al.*, "MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data," *Genome biol*, vol. 4, no. 1, p. R7, 2003.

[67] G. Dennis Jr, B. Sherman, D. Hosack, J. Yang, W. Gao, H. Lane, and R. Lempicki, "DAVID: database for annotation, visualization, and integrated discovery," *Genome Biol*, vol. 4, no. 5, p. P3, 2003.

[68] K. Dahlquist, N. Salomonis, K. Vranizan, S. Lawlor, and B. Conklin, "GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways," *Nature genetics*, vol. 31, no. 1, pp. 19–20, 2002.

[69] J. Chang and J. Nevins, "GATHER: a systems approach to interpreting genomic signatures," *Bioinformatics*, vol. 22, no. 23, p. 2926, 2006.

[70] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *Journal of computational biology*, vol. 7, no. 3-4, pp. 601–620, 2000.

[71] N. Nariai, S. Kim, S. Imoto, and S. Miyano, "Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks," in *Pacific Symposium on Biocomputing 2004: Hawaii, USA, 6-10 January 2004*. World Scientific Pub Co Inc, 2003, p. 336.

[72] J. Bader, A. Chaudhuri, J. Rothberg, and J. Chant, "Gaining confidence in high-throughput protein interaction networks," *Nature biotechnology*, vol. 22, no. 1, pp. 78–85, 2003.

[73] S. Leach, H. Tipney, W. Feng, W. Baumgartner, P. Kasliwal, R. Schuyler, T. Williams, R. Spritz, and L. Hunter, "Biomedical discovery acceleration, with applications to craniofacial development," *PLoS Comput Biol*, vol. 5, no. 3, p. e1000215, 2009.

[74] A. Patil and H. Nakamura, "Filtering high-throughput protein-protein interaction data using a combination of genomic features," *BMC bioinformatics*, vol. 6, no. 1, p. 100, 2005.

[75] C. Deane, Ł. Salwiński, I. Xenarios, and D. Eisenberg, "Protein interactions," *Molecular & Cellular Proteomics*, vol. 1, no. 5, pp. 349–356, 2002.

[76] M. Samanta and S. Liang, "Predicting protein functions from redundancies in large-scale protein interaction networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 22, p. 12579, 2003.

[77] D. Goldberg and F. Roth, "Assessing experimentally derived interactions in a small world," *Proceedings of the National Academy of Sciences*, vol. 100, no. 8, p. 4372, 2003.

[78] R. Saito, H. Suzuki, and Y. Hayashizaki, "Interaction generality, a measurement to assess the reliability of a protein–protein interaction," *Nucleic acids research*, vol. 30, no. 5, pp. 1163–1168, 2002.

[79] R. Saito, H. Suzuki, and Y. Hayashizaki, "Construction of reliable protein–protein interaction networks with a new interaction generality measure," *Bioinformatics*, vol. 19, no. 6, pp. 756–763, 2003.

[80] G. Liu, J. Li, and L. Wong, "Assessing and predicting protein interactions using both local and global network topological metrics," *Genome Informatics*, vol. 22, pp. 138–149, 2008.

[81] J. Chen, W. Hsu, M. Lee, and S. Ng, "Increasing confidence of protein interactomes using network topological metrics," *Bioinformatics*, vol. 22, no. 16, pp. 1998–2004, 2006.

[82] D. Li, W. Liu, Z. Liu, J. Wang, Q. Liu, Y. Zhu, and F. He, "Princess, a protein interaction confidence evaluation system with multiple data sources," *Molecular & Cellular Proteomics*, vol. 7, no. 6, pp. 1043–1052, 2008.

[83] M. Herrgård, M. Covert *et al.*, "Reconciling gene expression data with known genome-scale regulatory network structures," *Genome Research*, vol. 13, no. 11, pp. 2423–2434, 2003.

[84] M. Steffen, A. Petti, J. Aach, P. D'haeseleer, and G. Church, "Automated modelling of signal transduction networks," *BMC bioinformatics*, vol. 3, no. 1, p. 34, 2002.

[85] J. Scott, T. Ideker, R. Karp, and R. Sharan, "Efficient algorithms for detecting signaling pathways in protein interaction networks," *Journal of Computational Biology*, vol. 13, no. 2, pp. 133–144, 2006.

[86] S. Lu, F. Zhang, J. Chen, and S. Sze, "Finding pathway structures in protein interaction networks," *Algorithmica*, vol. 48, no. 4, pp. 363–374, 2007.

[87] G. Bebek and J. Yang, "Pathfinder: mining signal transduction pathway segments from protein-protein interaction networks," *BMC bioinformatics*, vol. 8, no. 1, p. 335, 2007.

[88] A. Gitter, J. Klein-Seetharaman, A. Gupta, and Z. Bar-Joseph, "Discovering pathways by orienting edges in protein interaction networks," *Nucleic acids research*, vol. 39, no. 4, pp. e22–e22, 2011.

[89] A. Ramesh, S. Nasser, and S. Kim, "Systematic Validation of Computationally Predicted Interaction Networks with a Literature-Derived Interaction Database," in *Workshop on Computational Systems Biology (Accepted)*, 2011.

[90] E. Dijkstra, "A note on two problems in connexion with graphs," *Numerische mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[91] R. McLendon, A. Friedman, D. Bigner, E. Van Meir, D. Brat, G. Mastrogianakis, J. Olson, T. Mikkelsen, N. Lehman, K. Aldape *et al.*, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, no. 7216, pp. 1061–1068, 2008.

[92] J. Yu, V. Smith, P. Wang, A. Hartemink, and E. Jarvis, "Using bayesian network inference algorithms to recover molecular genetic regulatory networks." International Conference on Systems Biology (ICSB02), December 2002.

[93] A. Barabási and Z. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, 2004.

[94] J. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

[95] E. Ravasz, A. Somera, D. Mongru, Z. Oltvai, and A. Barabási, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, no. 5586, p. 1551, 2002.

[96] F. Jacob, "Evolution and tinkering," *Science*, vol. 196, no. 4295, pp. 1161–1166, 1977.

[97] H. Sauro, "Modularity defined," *Molecular systems biology*, vol. 4, no. 1, 2008.

[98] J. Pereira-Leal, A. Enright, and C. Ouzounis, "Detection of functional modules from protein interaction networks," *Nucleic Acids Res*, vol. 29, pp. 242–245, 2001.

[99] G. Bader and C. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC bioinformatics*, vol. 4, no. 1, p. 2, 2003.

[100] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[101] S. Navlakha, M. Schatz, and C. Kingsford, "Revealing biological modules via graph summarization," *Journal of Computational Biology*, vol. 16, no. 2, pp. 253–264, 2009.

[102] P. Jiang and M. Singh, "SPICi: a fast clustering algorithm for large biological networks," *Bioinformatics*, vol. 26, no. 8, p. 1105, 2010.

[103] A. Ramesh, R. Trevino, D. Von Hoff, and S. Kim, "Clustering Context-specific Gene Regulatory Networks," in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 2010, p. 444.

[104] A. Bernard and A. Hartemink, "Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data," in *Pac Symp Biocomput*, vol. 10, 2005, pp. 459–470.

[105] A. Hartemink *et al.*, "Banjo: Bayesian network inference with java objects," *web site: http://www. cs. duke. edu/amink/software/banjo*, 2005.

[106] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300, 1995.

[107] U. Alon, "Biological networks: the tinkerer as an engineer," *Science*, vol. 301, no. 5641, p. 1866, 2003.

[108] J. Cheng, M. Cline, J. Martin, D. Finkelstein, T. Awad, D. Kulp, and M. Siani-Rose, "A knowledge-based clustering algorithm driven by Gene Ontology," *Journal of biopharmaceutical statistics*, vol. 14, no. 3, pp. 687–700, 2004.

[109] Z. Fang, J. Yang, Y. Li, Q. Luo, and L. Liu, "Knowledge guided analysis of microarray data," *Journal of Biomedical Informatics*, vol. 39, no. 4, pp. 401–411, 2006.

[110] D. Huang and W. Pan, "Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data," *Bioinformatics*, vol. 22, no. 10, p. 1259, 2006.

[111] D. Huang, P. Wei, and W. Pan, "Combining gene annotations and gene expression data in model-based clustering: weighted method," *OMICS: A Journal of Integrative Biology*, vol. 10, no. 1, p. 28, 2006.

[112] M. Brameier and C. Wiuf, "Co-clustering and visualization of gene expression data and gene ontology terms for Saccharomyces cerevisiae using self-organizing maps," *Journal of biomedical informatics*, vol. 40, no. 2, pp. 160–173, 2007.

[113] L. Tari, C. Baral, and S. Kim, "Fuzzy c-means clustering with prior biological knowledge," *Journal of biomedical informatics*, vol. 42, no. 1, pp. 74–81, 2009.

[114] D. Hanisch, A. Zien, R. Zimmer, T. Lengauer *et al.*, "Co-clustering of biological networks and gene expression data," *BIOINFORMATICS-OXFORD-*, vol. 18, pp. 145–154, 2002.

[115] B. Kulis, S. Basu, I. Dhillon, and R. Mooney, "Semi-supervised graph clustering: a kernel approach," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 457–464.

[116] T. Freeman, L. Goldovsky, M. Brosch, S. van Dongen, P. Mazière, R. Grocock, S. Freilich, J. Thornton, and A. Enright, "Construction, Visualisation, and Clustering of Transcription Networks From Microarray Expression Data," *PLoS Comput Biol*, vol. 3, no. 10, pp. 2032–2042, 2007.

[117] B. Samuel Lattimore, S. van Dongen, and M. Crabbe, "GeneMCL in Microarray Analysis," *Computational Biology and Chemistry*, vol. 29, no. 5, pp. 354–359, 2005.

[118] D. Higham, G. Kalna, and M. Kibble, "Spectral Clustering and Its Use in Bioinformatics," *Journal of computational and applied mathematics*, vol. 204, no. 1, pp. 25–37, 2007.

[119] D. Tritchler, S. Fallah, and J. Beyene, "A Spectral Clustering Method for Microarray Data," *Computational Statistics and Data Analysis*, vol. 49, no. 1, pp. 63–76, 2005.

[120] S. van Dongen, "Graph Clustering by Flow Simulation," *University of Utrecht*, 2000.

[121] S. van Dongen, "Technical Report INS-R0010: A Cluster Algorithm for Graphs," *National Research Institute for Mathematics and Computer Science*, 2000.

[122] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 888–905, 2000.

[123] M. Meila and W. Pentney, "Clustering by Weighted Cuts in Directed Graphs," in *Proceedings of the 2007 SIAM International Conference on Data Mining*, 2007.

[124] U. Brandes, M. Gaertler, and D. Wagner, "Experiments on graph clustering algorithms," *Algorithms-ESA 2003*, pp. 568–579, 2003.

[125] E. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, pp. 457–481, 1958.

[126] A. Hopkins, C. Groom *et al.*, "The druggable genome," *Nature Reviews Drug Discovery*, vol. 1, no. 9, pp. 727–730, 2002.

[127] D. Von Hoff, R. Penny, S. Shack, E. Campbell, D. Taverna, M. Borad, D. Love, J. Trent, and M. Bittner, "Frequency of Potential Therapeutic Targets Identified by Immunohistochemistry (IHC) and DNA Microarray (DMA) in Tumors From Patients Who Have Progressed On Multiple Therapeutic Agents," *Journal of Clinical Oncology*, vol. 24, no. 18_suppl, p. 3071, 2006.

[128] H. Bleich, E. Boro, E. Frei III, N. Jaffe, M. Tattersall, S. Pitman, and L. Parker, "New approaches to cancer chemotherapy with methotrexate," *New England Journal of Medicine*, vol. 292, no. 16, pp. 846–851, 1975.

[129] B. Rini, S. Halabi, R. Barrier, K. Margolin, D. Avigan, T. Logan, W. Stadler, P. McCarthy, C. Linker, E. Small *et al.*, "Adoptive immunotherapy by allogeneic stem cell transplantation for metastatic renal cell carcinoma: a calgb intergroup phase ii study," *Biology of Blood and Marrow Transplantation*, vol. 12, no. 7, pp. 778–785, 2006.

[130] Z. Jin, D. Dicker, W. El-Deiry *et al.*, "Enhanced sensitivity of g1 arrested human cancer cells suggests a novel therapeutic strategy using a combination of simvastatin and trail," *Cell Cycle*, vol. 1, no. 1, pp. 82–89, 2002.

[131] J. Friedrich, N. Adhikari, and M. Meade, "Drotrecogin alfa (activated): does current evidence support treatment for any patients with severe sepsis?" *Critical Care*, vol. 10, no. 3, p. 145, 2006.

[132] S. Osada, H. Tomita, Y. Tanaka, Y. Tokuyama, H. Tanaka, F. Sakashita, and T. Takahashi, "The utility of vitamin k3 (menadione) against pancreatic cancer," *Anticancer research*, vol. 28, no. 1A, pp. 45–50, 2008.

[133] R. Verhaak, K. Hoadley, E. Purdom, V. Wang, Y. Qi, M. Wilkerson, C. Miller, L. Ding, T. Golub, J. Mesirov *et al.*, "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1," *Cancer Cell*, vol. 17, no. 1, pp. 98–110, 2010.

[134] J. Rahnenfuhrer, F. Domingues, J. Maydt, and T. Lengauer, "Calculating the statistical significance of changes in pathway activity from gene expression data," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, p. 1055, 2004.

[135] S. Draghici, P. Khatri, A. Tarca, K. Amin, A. Done, C. Voichita, C. Georgescu, and R. Romero, "A systems biology approach for pathway level analysis," *Genome research*, vol. 17, no. 10, p. 1537, 2007.

[136] A. Tarca, S. Draghici, P. Khatri, S. Hassan, P. Mittal, J. Kim, C. Kim, J. Kusanovic, and R. Romero, "A novel signaling pathway impact analysis," *Bioinformatics*, vol. 25, no. 1, p. 75, 2009.

[137] T. Breslin, M. Krogh, C. Peterson, and C. Troein, "Signal transduction pathway profiling of individual tumor samples," *BMC bioinformatics*, vol. 6, no. 1, p. 163, 2005.

[138] M. Chagoyen and F. Pazos, "Quantifying the biological significance of gene ontology biological processesimplications for the analysis of systems-wide data," *Bioinformatics*, vol. 26, no. 3, pp. 378–384, 2010.

[139] A. Del Pozo, F. Pazos, and A. Valencia, "Defining functional distances over gene ontology," *BMC bioinformatics*, vol. 9, no. 1, p. 50, 2008.

[140] J. Wang, Z. Du, R. Payattakool, P. Yu, and C. Chen, "A new method to measure the semantic similarity of GO terms," *Bioinformatics*, vol. 23, no. 10, p. 1274, 2007.

[141] D. Lin, "An information-theoretic definition of similarity," in *Proceedings of the 15th International Conference on Machine Learning*, vol. 1, 1998, pp. 296–304.

[142] P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," *Journal of artificial intelligence research*, vol. 11, no. 4, pp. 95–130, 1999.

[143] X. Guo, R. Liu, C. Shriver, H. Hu, and M. Liebman, "Assessing semantic similarity measures for the characterization of human regulatory pathways," *Bioinformatics*, vol. 22, no. 8, p. 967, 2006.

[144] H. Wang, F. Azuaje, O. Bodenreider, and J. Dopazo, "Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships," in *Computational Intelligence in Bioinformatics and Computational Biology, 2004. CIBCB'04. Proceedings of the 2004 IEEE Symposium on*, 2004, pp. 25–31.

[145] H. Phillips, S. Kharbanda, R. Chen, W. Forrest, R. Soriano, T. Wu, A. Misra, J. Nigro, H. Colman, L. Soroceanu *et al.*, "Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis," *Cancer cell*, vol. 9, no. 3, pp. 157–173, 2006.

[146] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," *Machine learning*, vol. 52, no. 1, pp. 91–118, 2003.

[147] P. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.