A Computational Framework to Model and Learn Context-Specific Gene Regulatory Networks from Multi-Source Data

by

Ina Sen

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

Approved July 2011 by the
Graduate Supervisory Committee:

Seungchan Kim, Chair
Chitta Baral
Michael Bittner
Goran Konjevod

ARIZONA STATE UNIVERSITY

August 2011

ABSTRACT

Reverse engineering gene regulatory networks (GRNs) is an important problem in the domain of Systems Biology. Learning GRNs is challenging due to the inherent complexity of the real regulatory networks and the heterogeneity of samples in available biomedical data. Real world biological data are commonly collected from broad surveys (profiling studies) and aggregate highly heterogeneous biological samples. Popular methods to learn GRNs simplistically assume a single universal regulatory network corresponding to available data. They neglect regulatory network adaptation due to change in underlying conditions and cellular phenotype or both.

This dissertation presents a novel computational framework to learn common regulatory interactions and networks underlying the different sets of relatively homogeneous samples from real world biological data. The characteristic set of samples/conditions and corresponding regulatory interactions defines the cellular context (context). Context, in this dissertation, represents the deterministic transcriptional activity within the specific cellular regulatory mechanism.

The major contributions of this framework include - modeling and learning context specific GRNs; associating enriched samples with contexts to interpret contextual interactions using biological knowledge; pruning extraneous edges from the context-specific GRN to improve the precision of the final GRNs; integrating multi-source data to learn inter and intra domain interactions and increase confidence in obtained GRNs; and finally, learning combinatorial conditioning factors from the data to identify regulatory cofactors.

The framework, Expattern, was applied to both real world and synthetic data. Interesting insights were obtained into mechanism of action of drugs on analysis of NCI60 drug activity and gene expression data. Application to refractory cancer data

and Glioblastoma multiforme yield GRNs that were readily annotated with context-specific phenotypic information. Refractory cancer GRNs also displayed associations between distinct cancers, not observed through only clustering. Performance comparisons on multi-context synthetic data show the framework Expattern performs better than other comparable methods.

To my parents, who have always encouraged and inspired me.

Thank you for your love and blessings.

To my husband, who has supported me in this journey.

Thank you for being there for me.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Seungchan Kim, for his guidance on my research through the years. Dr. Kim has always led by example, highlighting the importance of professionalism and maintaining high standards of quality in research. He encourages discussion which often resulted in suggestions and ideas to improve this dissertation work. I would also like to thank Prof. Chitta Baral, Dr. Goran Konjevod and Dr. Michael Bittner who kindly accepted to serve on my examining committee. Their useful comments and suggestions certainly enhanced this work.

I am indebted to Dr. Gil Speyer from the High Performance Computing Initiative group at ASU for his time and discussions when I undertook to extend and parallelize the code using OpenMPI. I would also like to express my gratitude to the lab members, particularly the post-doctorate researchers - Dr. Sara Nasser and Dr. Sungwon Jung for their advice and contribution in this research, as well as Michael Verdicchio and Archana Ramesh for their effort in this research. I am fortunate to be surrounded by lots of good friends, especially Kahkashan Shaukat, Dr. Pavel Ghosh, Dr. Luis Tari, Dr. Suchismita Tarafdar and Dr. Sucharita Sengupta who have supported me through this journey. I offer my regards to all those who supported me in any way in preparing this dissertation.

My warmest thanks and heartfelt gratitude to my local guardians and my family for their constant support and encouragement. I am especially grateful to my husband for his love, support, understanding and patience through this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

The completion of genome sequencing projects such as The Human Genome Project has precipitated the question of how the basic units of biology, such as genes, act and interact selectively to enable life. Understandably, traditional reductionist approaches to study each biological entity (e.g., DNA, RNA, proteins) in isolation is insufficient to predict the synergistic outcome of the overall system. While an understanding of genes and proteins continue to be important, the focus has shifted to understanding a system's structure and dynamics [1].

Systems biology is the systematic study of interactions in biological systems, viewing the system as an outcome of collaborative interactions between components rather than studying each component individually. The rapid advancement in molecular biology and development of high-throughput platforms provide us with a rich repository of data for study. However, human minds are incapable of inferring the emergent properties of a system from thousands of data points. Thus computational tools play a vital role in the formulation of detailed graphical or mathematical models, refined by hypothesis-driven, iterative systems perturbations and data integration. Computational systems biology provides us with such tools for systems-level understanding of the aggregate outcome of cooperative and complementary interactions in biological systems, such as, our bodies. The main purpose of computational systems biology is to assign biological functions to genes, group of genes and particular gene interactions, and to understand how genes in a cell contribute to specialized function. This dissertation work develops one such computational systems biology tool, a statistical learning framework – Expattern (Extract Pattern), to aid in hypothesis formulation in, and understanding of, systems biology.

Viewing the system at cellular level, we understand that cell type differences

arise because of synthesis and accumulation of different sets of mRNA and protein molecules, i.e., through the expression of different set of genes. The expression level of a particular gene is influenced by the expression level of other genes. Genes interact with each other in order to control and regulate their expression levels. The resulting web of gene/protein interactions forms a Gene Regulatory Network (GRN). GRNs are an abstraction of dynamic interactions in a biological system. Usually the nodes are genes and edges represent direct or indirect interactions between genes or gene products. Ultimately, we wish to know the gene regulatory systems underlying the biological processes.

1.1 Context Specific Gene Regulatory Networks

A normal cell has a repertoire of mechanisms that ensure its proper growth, survival and development of the cell. The cell experiences various situations during its lifetime such as different mutations to the DNA code, change in morphology of neighboring cells, fluctuations in the nutrient supply or even virulent attacks of pathogens. Through all these different conditions and possibilities the cell either adapts or perishes (undergoes apoptosis). Adaptation to the new situation compels certain changes in the mechanisms for regulation of vital genes in the cell, differentiating it from the normal. The existence of different diseases such as cancer, which proliferate, suppressing apoptosis, provides an example of such an adapted system. This prompts us to develop a framework to learn the characteristic set of interactions and conditions that distinguish between normal and disease modulated regulatory mechanisms.

In our framework [2], we define cellular context as the characteristic set of interactions and conditions that represent the deterministic transcriptional activity within the specific cellular regulatory mechanism. It is assumed that when a cell maintains a specific cellular context, for example, a phenotype, it tightly regulates a battery of genes, which would show rather deterministic transcriptional activities.

When the cell moves away from this cellular context or changes to a different cellular state, the behavior of the same set of genes will not appear as deterministic since they now behave without control signals (intrinsic stochastic behavior) or each gene comes under the control of various other external controls. This change in gene regulatory behavior under different cellular contexts results in a different regulatory network for each context, i.e., context-specific gene regulatory network.

The task of learning the contexts from biological data is complicated by the fact that cellular processes are robust, redundant and involve multi-strata interactions (between DNA, RNA, Proteins, miRNA, siRNA). In order to correctly infer the active pathway and associated components, biologists have to invest in costly and time-consuming experiments. Thus, the use of computational approaches to learn the underlying functional connections between the genomic entities from available high throughput data becomes a more viable alternative both in terms of resource cost and time.

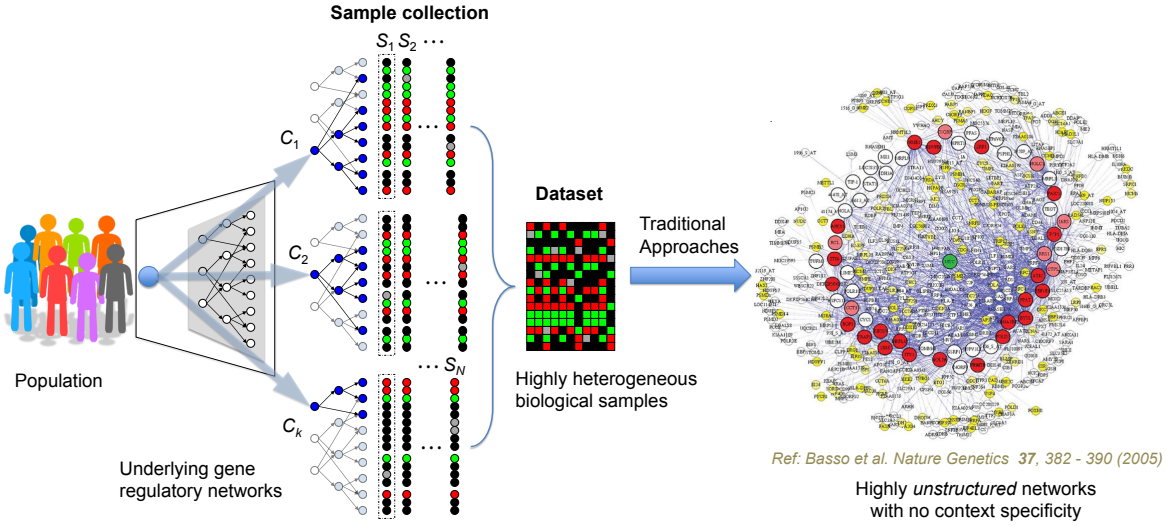


Figure 1.1: Current computational methods to discover underlying regulatory mechanisms assume homogeneity of samples. Use of available heterogeneous biological samples by such methods yield highly unstructured GRNs without any context-specificity.

Except in rare circumstances, biological data are increasingly collected not

from tightly defined and controlled experiments, but from broad surveys, i.e., profiling studies, which invariably lead to highly heterogeneous biological samples. Current computational methods to discover underlying regulatory mechanisms typically assume certain level of homogeneity of samples and are not adequate in dealing with recent explosion of highly heterogeneous high throughput measurements. Figure 1.1 depicts the application of current computational methods to heterogeneous datasets in order to yield GRNs.

Understandably, alterations to the cell's control circuitry would produce diversity in the molecular mechanisms operating in diseased cells. On applying inference methods to such a diverse dataset, the heterogeneity in either the network regulatory connections or operating rules would blur the relationships, reducing the ability to accurately determine consequential regulatory interactions. Cellular contexts account for heterogeneity in the data and identify condition specific regulatory interactions. Thus knowledge of cellular contexts would be highly applicable to fields such as predictive medicine, biomarker discovery and identification of targets for therapeutic intervention.

Briefly, predictive medicine entails predicting disease and instituting preventive measures. Using prevalent cellular contexts with observable symptoms for subtyping of diseases (to further classify patients) could possibly predict deleterious effects of treatment on the patients' health. Thus, knowledge of cellular contexts associated with a particular disease could be applied to diagnose the disease, institute preventive measures and/or prescribe appropriate treatments.

Interestingly, as cellular contexts capture the biological state both in terms of multi-entity interactions and the prevalent conditions, further analysis of relevant cellular contexts would possibly yield a set of biological indicators (instead of a single biological indicator) constituting more robust disease related biomarkers. Finally, the cellular contexts capture the mechanisms that regulate cellular activity.

By comparing cellular contexts across different disease stages, it might be possible to identify pivotal genes or entities in the pathway which influence the trajectory of the cellular development (for instance, towards a preferred phenotype) as strong candidates for therapeutic intervention.

It is imperative for any context specific regulatory network model to also account for switching from one cellular context to another, such as from a healthy context to a diseased one, which would concordantly change the observable set of interactions in the cell. The thesis work models this context-switching system and develops a computational framework, designed to learn the context specific network structure that adequately reflects the relationships in the observed evidence from different data types, i.e., multiple sources. This framework will be applied to problems (disease related) in the biological domain in order to demonstrate the applicability and robustness of the framework.

1.2 Problem Definition

This section formulates the biological problem of identifying cellular contexts to an equivalent computational problem in order to develop relevant computational tools. The study of gene regulatory networks is an important, complex problem in the biological domain. The problem of identifying genetic regulation from high throughput data is an ill-posed problem, considering the relatively high number of genes (biological entities) when compared to the number of samples (experimental conditions/patients). The problem is further confounded by the different levels of heterogeneity of the samples present in high throughput data, an aspect usually neglected by correlation based network learning methods, which treat all samples as instances of the same gene regulatory network. Here, we develop a framework that is able to distinguish between important structural changes in the regulatory relationships from system realizations of different cellular contexts.

Formally, the problem can be stated as given high throughput data ma-

trix D = $[d_{11}d_{12}\ldots d_{1m};\ldots;d_{n1}d_{n2}\ldots d_{nm}]$, containing the expression values (or biomedical data), corresponding to the activity of genes G in samples T, where $G = \{g_1, g_2, ..., g_n\}$ is the set of n genes (cellular/biological entities), and $T = \{t_1, t_2, ..., t_m\}$ is the set of m heterogeneous samples (patients/ subtypes of diseases), then find

- *context motifs* $C_i = \{G_i, Y_i, S_i, T_i\}$ where the set of genes G_i in states specified by a vector expression value Y_i tightly interact[1] with the set of genes S_i within a set of samples T_i.
- *contexts*, network of context motifs with associated common conditions.

Challenges

These are the challenges we will have to meet in order to solve the above problem:

1. Modeling and identification of context motifs:
    a) We need to define a measure of consistent behavior. Using this measure, we need to identify sets of genes that display consistent behavior only within characteristic sets of samples and not outside. Also, when considering consistent activity, we need to allow possible internal biological noise within and external control effects outside the identified set of samples.
    b) Once the context motif model is in place, for context motif identification, we need to estimate the model parameters from the high throughput heterogeneous data. As the parameter estimations are calculated solely from the data, we might obtain false positives as statistical artifacts. We would have to develop a method to minimize false positives from parameter estimation of the model from high throughput data.

[1] Each interaction within the context motif is an edge of the contextual regulatory network. Biological interaction and regulation have been used interchangeably in this dissertation.

c) We require a method to integrate multiple sources of data in the framework to increase the reliability of the results and to obtain a holistic view of system interactions of entities across different sources or domains.

d) We also need to identify combinatorial conditioning factors which complement each other's activity as observed in biology, for example, transcription factors and co-factors necessary for the functioning of the transcription factors.

2. Learning contexts: We need to define similarity between context motifs in terms of sample overlap. This would be required to develop a method to identify the set of context motifs sharing most conditions (set of samples). The conditions associated with the network of interactions from the context motif set would ultimately yield the contexts. Importantly, on combining the context motifs, redundant or extraneous interactions need to be identified and removed.

3. Validation of the approach: We need to develop methods for comparison and validation of the framework. For that, we need to consider different measures of computational accuracy and also methods for biological validation.

1.3 Overview of Framework

Here is a brief outline of the tasks this thesis work focuses on to address the challenges and solve the thesis problem outlined above :

1. *Model and learn context motifs.* Build a framework to model and identify context motifs from the dataset. Consider both univariate and multivariate conditioning factors for specifying the context motif. Use corrected multiple hypothesis testing for determining statistical significance of context motifs in order to minimize false positives. Explore integration of different types of

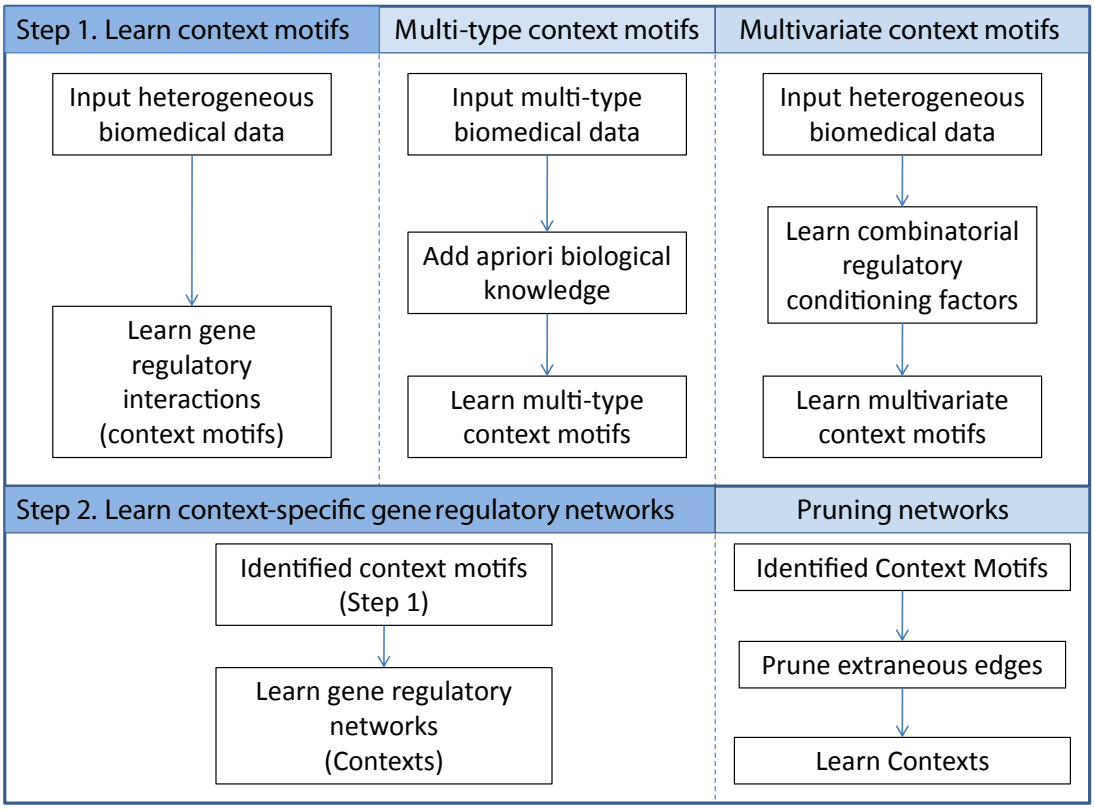


Figure 1.2: Tasks outlined for learning context-specific GRNs from heterogeneous biological data. Learn context-specific interactions and integrate them to form the context-specific GRN.

data to increase confidence in results. This task is presented as Step 1 in Figure 1.2.

2. *Learn contexts.* Use graph based approach to combine context motifs and learn regulatory interaction subnetworks associated with different set of common conditions. Develop context specific network pruning methods to extract a reduced set of characteristic interactions from large scale contextual gene regulatory networks.This task is presented as Step 2 in Figure 1.2.

3. *Validate.* Generate artificial context-specific gene regulatory networks (aGRN), and produce gene expression data. This will allow proper validation and comparison of results of inference methods and study the system characteristics. Apply the framework to extract context-specific GRNs from simulated data

and real world (cancer) data. Work with biologists for biological validation and interpretation of results.

1.4 Outline of Dissertation

The dissertation is divided into different chapters. Chapter 2 provides a brief overview of available biomedical data, GRN modeling formalisms, GRN learning algorithms. Chapter 3 introduces the mathematical model on contextual genomic regulation and presents the context motif inference algorithm developed for the framework. Chapter 4 introduces a method to learn contexts from context motifs and applies it to a refractory cancer dataset. Chapter 5 presents different context-specific GRN pruning strategies to remove extraneous edges. Chapters 4 and 5 also compares the performance of the framework with popular GRN reverse engineering algorithms on artificial contextual networks. Chapter 6 presents an innovative method for multi data type integration to identify multi-type context motifs. Chapter 7 presents a method to identify multivariate drivers of context motifs. Chapter 8 summarizes the key contributions of the dissertation work and future directions.

Chapter 2

BACKGROUND

Biological systems are inherently complex, the study of which entails observation, modeling, simulation and learning. The observation of biological systems could be quantitative measurements such as concentrations, or qualitative such as presence or absence of a gene product/protein. The diversity of available biomedical data thus provides different perspectives of the biological system under study (according to the source and type of data). This chapter provides an outline of some of the biomedical data technologies being employed in the study of GRNs. Next, we discuss current frameworks and approaches to modeling and simulation of gene interactions. Finally, we describe some of the methods in use to learn gene interactions from biomedical data.

2.1 Biomedical Data

Briefly, in eukaryotic organisms, the genetic information is encoded as DNA (Deoxyribonucleic acid) residing in the nucleus. DNA gets transcribed to mRNA (messenger Ribonucleic acid) which goes into the cytoplasm from the nucleus. The mature mRNA finds its way to a ribosome where it gets translated to a protein. The subsequence of the DNA that encodes for any protein is referred to as the gene.

For cellular contexts, the genetic interactions are not restricted between the encoded proteins or between the encoded proteins and the genes. As cellular contexts represent deterministic transcriptional activity, interactions are considered between any regulatory element and elements regulated by it. We extend the domain of conditioning factors (regulating elements) from only genes, to elements which influence, regulate or act specific to the existing cellular state. Any such factor would also be bound by the constraints in place due to cellular contextual state. With the advent of new technologies there are many types of biomedical data that are

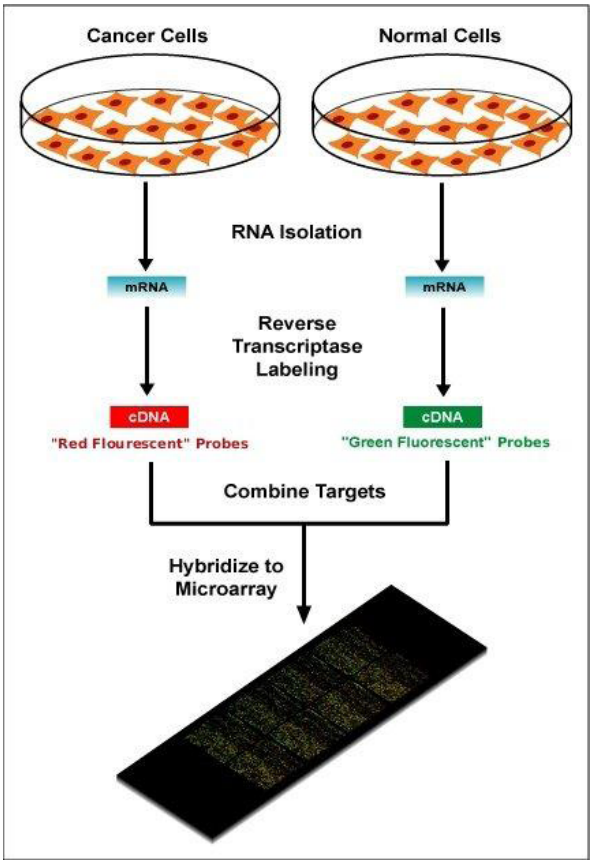


Figure 2.1: cDNA microarrays can measure life-stage and tissue specific patterns of gene expression. Reference Wikipedia.[2]

available. In the following sections we describe some popular current biomedical technologies and data on which we apply our framework. In chapter 6 we exploit the directionality of regulatory influence between different data types to integrate these data types in our framework (Figure 6.2). This approach can be easily extended to incorporate other types of biomedical data too, for instance, microRNA [3] data and clinical data.

Gene Expression Data

Gene expression microarrays are used to study the relative mRNA concentrations (in lieu of relative protein concentrations). As mRNAs are unstable, cDNAs (complementary DNA) are generated and used for the microarray experiment. The microar-

[2]http://en.wikipedia.org/wiki/File:Microarray-schema.jpg

ray chip is spotted with probes - single stranded DNA which on hybridization with its complementary cDNA strand would give off florescence. Relative gene expression is measured as the ratio of the two fluorescences: “up-regulation” of the experimental transcriptome relative to the control as red pseudo-color, “down-regulation” as green, and constitutive expression (1:1 versus control) as neutral black. The intensity of color is proportional to the expression differential. Gene expression microarrays display results in a matrix of gene activity in each experimental sample. However, due to the disproportionate number of probes (genes) and samples, deciphering the interactions between them becomes an ill-posed problem. Gene expression data is extensively used by GRN inference methods [2, 4, 5, 6, 7, 8, 9] to obtain genetic interactions.

Array-Based Comparative Genomic Hybridization

Array-based comparative genomic hybridization (aCGH) is a technique to detect genomic copy number variations (CNVs) in DNA. DNA from a test sample and normal reference sample are labelled using different fluorophores, and hybridized to several thousand probes printed on a glass slide. The probes are derived from most of the known genes and non-coding regions of the genome. Measure of the CNVs for a particular location in the genome is calculated as the ratio of the fluorescence intensity of the test to the reference DNA.

Like other types of genetic variation, some CNVs have been associated with susceptibility or resistance to disease. For instance, the epidermal growth factor receptor (EGFR) copy number has been found to be higher than normal in non-small cell lung cancer [10]. Thus, CNVs can be considered as conditioning factors, possibly influencing gene activity and resultant genetic interactions in the cellular context. As a constituent of a cellular context interactions it would provide insights into disease specific CNVs for disease associated cellular contexts. Figure 2.2 shows the outcome of an array CGH experiment with the intensity of different spots

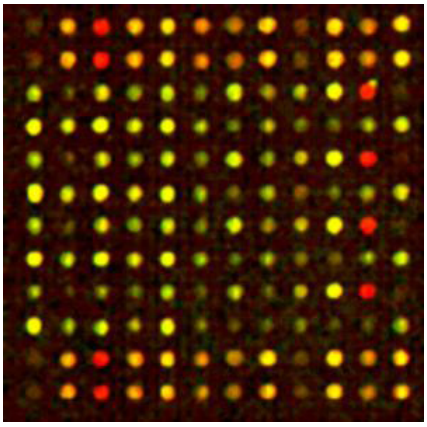


Figure 2.2: Array comparative genomic hybridization (CGH) experiment outcome.

denoting amplification or deletion of corresponding regions of the genome.

Drug Activity Data

Drugs are defined as chemical substances used in the treatment, cure, prevention, or diagnosis of disease or used to otherwise enhance physical or mental well-being. However, in most cases, the mechanism of action of drugs at cellular level is not known. Pharmaceutical companies are investing in expensive clinical trials to study the effects of different drugs on different subtype of diseases for treatment. Understanding the mechanism of action become imperative for prescribing personalized medicine. The experiments measure drug activity – a measure of the physiological response a drug produces. One of the measurements of drug activity is GI50, the concentration needed to reduce the growth of treated cells to half that of untreated (i.e., control) cells. For example, Scherf *et al.* [11] studied the gene expression data across National Cancer Institute cell lines (NCI60) and the drug activity of 1400 drugs on those cell lines.

Prevalent cellular context conditions would dictate or impact molecular level drug interactions. Therefore, targeted genes and pathways would be better realized when considering drug activity within a cellular context. In Chapter 3 we demonstrate the use of drug activity as possible conditioning entities in interactions within

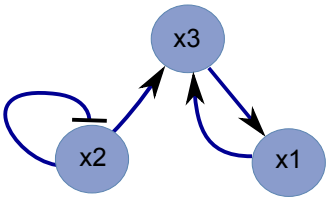
cellular contexts.

2.2 Modeling Gene Regulatory Networks

Existing formalisms to model GRNs include Boolean Networks, Probabilistic Boolean Networks (PBNs), Bayesian Networks and Ordinary Differential Equations (ODEs). Among these, the ODE formalism involves the construction of a set of differential equations to relate the rate of change of active gene concentrations. Different values of production and degradation constants in the equation are set based on biological observations. ODE formalism provides a finer granularity (as compared to Boolean and Bayesian Network formalisms) to model and analyze GRNs. In the next sections, we briefly describe some popular formalisms.

Boolean Networks

In the Boolean network formalism, the genes are allowed binary expression levels, with Boolean functions deterministically describing the relationships between the genes. Although simplistic in concept and approach, random Boolean networks were found to display properties similar to the yeast transcriptional network [12]. A Boolean network of n genes, $B = (V, F)$, is defined by the set of nodes $V = \{x_1, x_2, ..., x_n\}$ and their corresponding set of Boolean functions $F = (f_1, f_2, ..., f_n)$. The functions $f_i : \{0, 1\}^n \rightarrow \{0, 1\}, i = 1, ..., n$ are the *predictor functions* for gene i [13]. The value of x_i represents the state/expression of the gene i, where 0 means gene i is OFF and 1 means gene i is ON. At each time step the states of all genes are updated synchronously according to their predictor function. The gene activity profile (GAP) is the state of the network at that instant, given by $x(t) = (x_1(t), x_2(t), ..., x_n(t))$. A small example of Boolean network is shown in Figure 2.3.



$$x_1(t+1) = x_3(t),$$
$$x_2(t+1) = 0,$$
$$x_3(t+1) = \bar{x}_1(t)x_2(t) + x_1(t)\bar{x}_2(t)$$

x(*t*)	000	001	010	011	100	101	110	111
x(*t*+1)	000	100	001	101	001	101	000	100

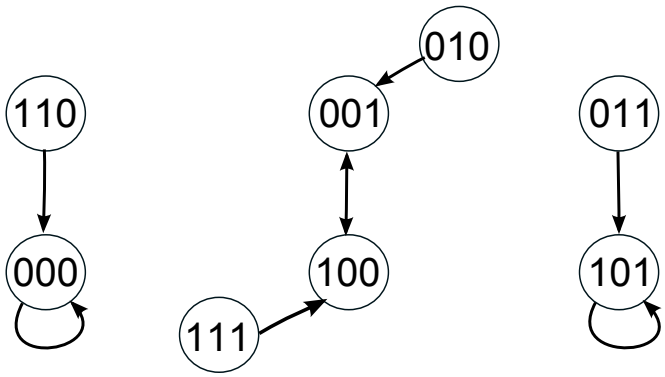


Figure 2.3: Boolean network as a graph with corresponding Boolean functions and state transition diagram.

Probabilistic Boolean Networks

The PBN formalism [14] extends the Boolean Network by introducing stochasticity in choosing the functions describing the gene relationships. A PBN consists of a set of nodes $V = \{x_1, x_2, ..., x_n\}$ and their corresponding set of vector-valued network functions $F = (\mathbf{f_1}, \mathbf{f_2}, ..., \mathbf{f_n})$ governing the state transitions of the genes. PBN associate multiple predictor functions and corresponding selection probabilities with each gene. Thus, at any instance, the realization of the PBN is determined by the selection of predictor functions for each gene, resulting in a probabilistically determined Boolean network. Figure 2.4 shows a basic building block of a PBN. In the figure, a number of predictors share common inputs while their outputs are synthesized, in this case by random selection, into a single output. The wiring

diagram for the entire PBN would consist of n such building blocks. Although the 'wiring' of the inputs to each function is shown to be quite general, in practice, each function (predictor) has only a few input variables.

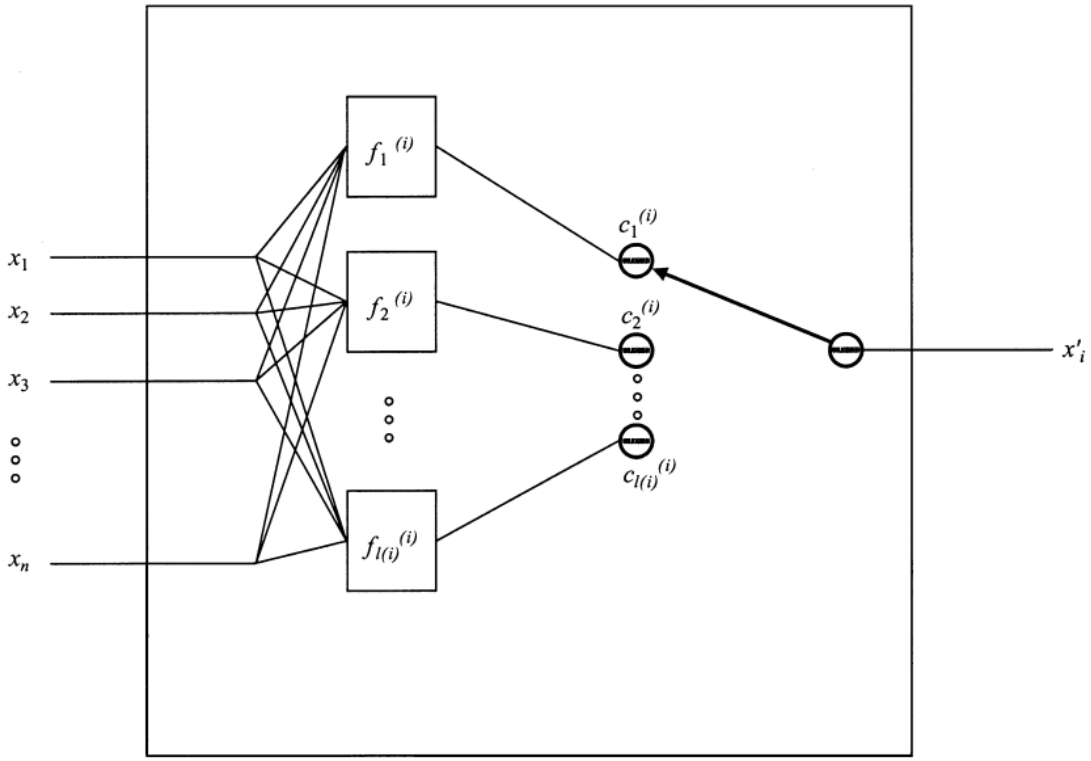


Figure 2.4: A basic building block of a Probabilistic Boolean Network.

Probabilistic Causal Models

In order to explain Probabilistic Causal Models (PCMs) we first need to explain causal models. A *causal model* [18] is defined as a triple $M = \langle U, V, F \rangle$ where:

1. U is a set of *background* variables, (also called exogenous), that are determined by factors outside the model.

2. V is a set $\{V_1, V_2, \ldots, V_n\}$ of variables, called endogenous, that are determined by variables in the model - that is, variables in $U \cup V$; and

3. F is a set of functions $\{f_1, f_2, \ldots, f_n\}$ such that each f_i is a mapping from (the respective domains of) $U \cup (V \setminus V_i)$ to V_i and such that the entire set F forms a mapping from U to V . In other words, each f_i tells us the value of V_i given the values of all other variables in $U \cup V$, and the entire set F has a unique solution $V(u)$. Symbolically, the set of equations F can be represented by writing

$$v_i = f_i(pa_i, u_i),\ i = 1, \ldots, n$$

where pa_i is any realization of the unique minimal set of variables PA_i in $V \setminus V_i$ (connoting *parents*) sufficient for representing f_i. Likewise, $U_i \subseteq U$ stands for the unique minimal set of variables in U sufficient for representing f_i.

Next, to define a submodel we assume M is a causal model, X is a set of variables in V, and x is a particular realization of X. A submodel M_x of M is the causal model $M_x = \langle U, V, F_x \rangle$ where $F_x = \{f_i : V_i \notin X\} \cup \{X = x\}$. Finally, a probabilistic causal model (PCM) [18] is defined as a pair $\langle M, P(u) \rangle$ where M is a causal model and $P(u)$ is a probability function defined over the domain of U.

Although PCM and PBNs are both stochastic in nature, there are vital differences. PCM do not have the restriction of being Boolean. Also, PCMs are constructed to allow interventions in the form of $do(x)$. By using the intervention $do(x)$ in PCM one can easily predict the effects on other variables in the model through observation of the corresponding submodels M_x. In PBNs the focus is more on finding the steady state distribution and identify how to reach a desired state as early as possible [14]. PCMs provide a more intuitive method of causal representation, not possible with PBNs. Thus in PCMs embedded causal information can be exploited in case of counterfactual reasoning where the model itself might need to be mutated to realize the scenario. PBNs would not be an adequate model for

counterfactual reasoning. However, inference of probabilistic causal models from high throughput real world data, such as steady state gene expression data is still a difficult problem. Identification of distinct context-specific gene regulatory interactions and networks adds another layer of complexity to the learning that is yet to be explored.

Bayesian Network

Bayesian network formalism[15, 16, 8] models the GRNs as directed acyclic graphs, where, a gene is assumed to be conditionally independent of all other genes given its parents. Bayesian network learning builds the joint probability distribution using the conditional dependencies in the network. Bayesian networks can be associational or causal. Formally, a Bayesian network for a set of random variables X is a pair $B = (G, Q)$. The first component, G, is a directed acyclic graph (DAG) whose vertices correspond to the random variables $x_1, \ldots, x_n$, and whose edges represent direct dependencies between the variables. The graph G encodes the following set of independence statements: each variable x_i is independent of its nondescendants given its parents in G. The second component of the pair, namely Q, represents the set of parameters that quantifies the network and describes a conditional distribution for each variable, given its parents in G. Together, these two components specify a unique distribution on $x_1, \ldots, x_n$. Bayesian networks not necessarily causal, a directed edge from u to v does not require that X_v is causally dependent on X_u. For instance, Bayesian networks on the graphs:

$$a \longrightarrow b \longrightarrow c \qquad \text{and} \qquad a \longleftarrow b \longleftarrow c$$

are equivalent as they impose exactly the same conditional independence requirements. In case of genetic regulatory networks, it is difficult to determine directions of causality and thus association based Bayesian networks are more

prevalent. Recently, dynamic Bayesian networks [17] have been used to study time series data and unravel the cyclic dependencies across time points.

Causal Bayesian Network

A *causal Bayesian network* [18] is a Bayesian network with an added requirement that parents of each node are its direct causes. The additional semantics of the causal Bayesian networks specify that if a node X is actively caused to be in a given state x (an action written as $do(X = x)$), then the probability density function changes to the one of the network obtained by cutting the links from X's parents to X, and setting X to the caused value x [18]. Using these semantics, one can predict the impact of external interventions from data obtained prior to intervention.

A *causal structure* D of a set of variables V is a DAG in which each node corresponds to a distinct element of V, and each link represents direct functional relationship among the corresponding variables. A *causal model* is a pair $M = \langle D, \Theta_D \rangle$ consisting of a causal structure D and a set of parameters Θ_D compatible with D. The parameters Θ_D assign a function $x_i = f_i(pa_i, u_i)$ to each $X_i \in V$ and a probability measure $\Pr(u_i)$ to each u_i, where PA_i are the parents of X_i in D and where each U_i is a random disturbance distributed according to $\Pr(u_i)$, independently of all other u. Causal Bayesian Networks [18] are formally defined as:

Let $P(v)$ be a probability distribution on a set V of variables, and let $P_x(v)$ denote the distribution resulting from the intervention $do(X = x)$ that sets a subset X of variables to constants x. Denote by $\mathbf{P}_*$ the set of all interventional distributions $\mathbf{P}_x(v)$, $X \subset V$, including $P(v)$, which represents no intervention (i.e., $X = \phi$). A DAG G is said to be a *causal Bayesian network* compatible with $\mathbf{P}_*$ if and only if the following three conditions hold for every $P_x \in \mathbf{P}_*$:

1. $P_x(v)$ is Markov relative to G;

2. $P_x(v_i) = 1$ for all $V_i \in X$ whenever v_i is consistent with $X = x$;

3. $P_x(v_i|pa_i) = P(v_i|pa_i)$ for all $V_i \notin X$ whenever pa_i is consistent with $X = x$.

The graph G represents conditional independence assumptions that allow the joint distribution to be decomposed, economizing on the number of parameters. The graph G encodes the Markov Assumption: (Each variable X_i is independent of its nondescendants, given its parents in G).

In causal Bayesian networks, the dependencies between nodes are represented by the directed edges between nodes and the local conditional probability distribution of those nodes. The causal interactions are assumed invariant with respect to time. The dependencies in causal Bayesian networks need to be invariant with time, so that the interventions of type $do(X = x)$ can be resolved without considering any extra parameter such as time. If the dependencies in the causal Bayesian networks were temporal then the local conditional probability tables would have to account for the time component too or the model would have to incorporate sets of conditional probability tables for different instances of time. Thus causal Bayesian network assumes a single underlying causal network that can explain the obtained data. Also, they cannot be used to model cyclic dependencies or model dynamic or temporal processes.

Artificial GRNs

Modeling of GRNs is complemented by simulation of artificial GRNs (aGRNs). Methods to simulate synthetic data (e.g., A-BIOCHEM [19], GRENDEL [20] and SynTReN [21]) focus mainly on the topological aspects and mRNA concentration dependancies. The selected topologies in the aGRNs define network characteristics such as average clustering coefficient, average path length and marginal degree distributions [21]. Some well known topologies used for generating GRNs are Erdös-Rényi random networks [22], Kauffman networks (restricted by the num-

ber of connections per gene) [12], Watts-Strogatz small-world networks and Albert-Barabási scale-free networks (gene connectivity follows the power law). Briefly, A-BIOCHEM [19] uses coupled differential equations to represent mRNA concentrations and allows different topologies. GRENDEL [20] decorrelates the activities of mRNAs, proteins and environmental stimuli and uses topology generation and kinetic parameterization to initialize the continuous time dynamical systems. SynTReN [21] samples subgraphs of known transcriptional networks to generate realistic biological topologies and uses Michaelis-Menten and Hill kinetic equations to model the interaction kinetics. However, most aGRN simulators assume a single global regulatory network accounting for the genetic activity profile of the cellular system. Thus the corresponding generated dataset is representative of a single network. This is a simplistic reduction of real world datasets as biological data represent highly heterogeneous set of biological samples.

2.3 Learning Gene Regulatory Networks

Computational approaches such as clustering, classification and feature selection yield interesting preliminary results in identifying co-expressed and possibly co-regulated genes. Clustering manifests similarity between cancer cell lines that were generated from the same tissue [11], implying successful capture of phenotypic similarity. Classification finds subtypes of tumors such as in case of breast cancer data predicting better survival [23]. Feature selection as a dimensional reduction approach effectively reduces the number of genes considered as relevant. Saeys *et al.* provide a review of different feature techniques applied in bioinformatics [24].

Initial work to infer interactions between genes were based on linear and nonlinear correlation measures. For instance, using Pearson's correlation of gene activity profile to determine if the genes are possibly co-expressed or co-regulated. REVerse Engineering ALgorithm (REVEAL) [25] uses mutual information (MI) [26], a nonlinear correlation, to analyze the data generated from Boolean models of ge-

netic networks to identify sets of input genes controlling each gene in the network. While many association-based and classification-based approaches have proven useful, one must look among all of the associated genes and attempt to group them on the basis of prior knowledge about the activities of the individual genes to identify particular processes. As the method tries to look for more specific relationships among genes, it can find smaller groups of interacting genes, defined by the kinds of behaviors that arise from the way in which transcriptional regulation operates, improving the likelihood that such sets do represent interpretable hypothesis.

This brings to attention methods to find modules, possibly functional, of the regulation network. Popular methods include biclustering [7], the Signature Algorithm [6, 5], and Segal's module map method [27]. Biclustering [7] identifies gene-sample grouping that have similar expression patterns. Biclustering considers coherent gene-sample patterns but struggles with evaluating the separation between the identified biclusters, making its output not as easily interpretable. Tanay et al. [28] provide a comprehensive review of biclustering methods applied to gene expression data. Shi et al. [29] extended the biclustering approach to generate super biclusters by combining biclusters, allows for more comprehensive results. Bhattarcharya et al. [30] developed bi-correlation clustering algorithm (BCCA) to yield a diverse set of biclusters of co-regulated genes over a subset of samples where all the genes in a bicluster have a similar change of expression pattern over the subset of samples. Unfortunately, these methods are time consuming (BCCA's time complexity is $O(n^5)$), and yield a very high number of biclusters, making further study or interpretation difficult.

The Signature Algorithm [6, 5] uses an initial gene set to identify and score conditions where the genes are differentially active and iteratively choose genes to maximize the score. The necessity for initial gene list limits the exploratory power of this algorithm. As the algorithm proceeds, dependent upon the genes/conditions

included in progressive iterations, it may allow convergence to a separate module altogether, thereby losing the signal present in the initial list.

Another popular method, module map method by Segal et al. [27], utilizes predefined gene sets with prominent expression signature within different arrays (compiled set of biological labels) to identify core set of genes showing similar signature across arrays, i.e., functional modules participating in a common biological process. Segal et al. applied their method to Saccharomyces cerevisiae expression data set to identify regulatory modules and their condition-specific regulators from gene expression data [31]. They also applied the method to perform an integrated analysis of 1,975 published microarrays spanning 22 tumor types to develop cancer module maps [27]. This method starts from initial partitions generated from clustering and utilizes prior biological knowledge such as Gene Ontology [32], KEGG (Kyoto Encyclopedia of Genes and Genomes) [33] and Gene MicroArray Pathway Profiler [34], if available, in that study. Details of other comparable module methods – LeMoNe [35], CONEXIC [36] and COALESCE [37] are provided in related works of chapter 4.

There are also methods which learn regulatory networks directly instead of functional modules. Relevance Networks [38] uses a threshold MI and only gene-gene associations at or above the threshold are used to construct the interaction networks. Extensions of Relevance Networks approach led to Context likelihood of relatedness (CLR) [9] and ARACNE [39] algorithms. CLR applies an adaptive background correction step to eliminate false correlations and indirect influences. ARACNE, instead, uses the Data Processing Inequality (DPI) to remove indirect interactions , i.e., interaction with lowest MI in any triplet of fully connected gene interactions. As discussed earlier, Bayesian networks have also been applied to learn gene regulatory networks from gene expression data. Among them, BANJO [8, 40] learns a dynamic Bayesian network and assigns positive or negative signs to the

directed interactions to denote promotory or inhibitory effects between genes. Also, network learning using network clustering method [41] uses ℓ_1 penalized network clustering assuming Sparse Gaussian Markov Random Fields. This uses undirected graphs and assumes the data is jointly Gaussian. Such assumptions are not always valid for biological datasets and thus may provide erroneous results.

2.4 Summary

This chapter provides the introduction and background to different biomedical technologies and data used in the framework. It also briefly describes comparable formalisms and methods to model and reverse engineer gene regulatory networks. The next chapter establishes the framework by describing the model, formal definitions and methods developed in this dissertation.

Chapter 3

IDENTIFYING CONTEXT MOTIFS

The correct identification of biological patterns or interactions from the data is the first step in learning GRNs. However, most methods that try to identify biological patterns assume all samples within the data represent the same set of interactions. This assumed homogeneity of datasets may cause these methods to miss interactions specific to different contexts. Here we discuss the motivation behind identifying context specific biological patterns (, i.e., context motifs) and formally describe the model and methodology developed in this dissertation work.

3.1 Motivation

As discussed in chapter 2, there are many algorithms that try to identify gene co-expression as putative co-regulation or biological interaction. In all such cases, the assumption is that there is one underlying gene regulatory network consisting of a set of interactions that needs to be identified. However, as outlined in Section 1.1, cellular systems can be modeled in terms of contexts. Any comparison of healthy versus diseased cell can now be comprehended as a shift in the underlying mechanism regulating the activities of the cell, i.e., a contextual shift. Thus different cellular contexts (for example subtypes of diseases) would constitute of different (possibly overlapping) sets of interactions. Our aim thus becomes to identify these sets of interactions (context motifs) first and then learn the underlying GRN specific to that context.

3.2 Methodology

Contextual Genomic Regulation Modeling

It is important to select a mathematical model of a cell's regulatory activity that accounts for regulation which very actively adjusts to differing internal and external environmental factors. Rather than models which infer connections between single genes, or between genes and phenotypes, we need to select a model which can

find subsets of samples where it is possible to attribute the states of all the members of a set of controlled genes to a single gene, or to a small set of regulatory genes which have expression properties that could be the source of control.

The framework needs to identify context specific relationships from the system realizations, by defining relationships as a functional mapping of sets of regulatory (driver) elements to the activities of the regulated (driven) elements which constitute the system. For the cellular domain, this has been captured by the mathematical model for contextual genomic regulation [42]. The context specific GRN learning algorithm in Expattern is based on this model, to identify novel cellular contexts, sets of genes whose expression pattern is significantly consistent within a specific biological context.

In our framework [2], we introduce a mathematical model to approximate contextual genomic regulation. Formally, the model assumes there are m sets, $G_1, G_2, ..., G_m$, of driver genes and m corresponding sets, $S_1, S_2, ..., S_m$, of driven genes. For each set of driven genes S_j, there is a corresponding set G_j of driver genes regulating their behavior. $G_1, G_2, ..., G_m$ are not necessarily disjoint, neither are $S_1, S_2, ..., S_m$ necessarily disjoint; thus some driver gene may regulate more than one driven set, and some driven gene may be regulated by more than one driver gene set.

Two parameters are essential to the definition of the contextual genomic regulation model. To define these parameters, consider a single set of driver genes G_i and its driven set of regulated genes S_i. For the set of drivers, still assuming a binary model (without loss of generality), there exists a state vector $\mathbf{Y_i} = (Y_{i_1}, Y_{i_2}, ..., Y_{i_q})$ where $Y_{i_k} (1 \leq k \leq q)$ gives the value of $g_{i_k} \in G_i$. Let regulation by the driver genes be such that for a state $\mathbf{y}$ of the driver gene state vector $\mathbf{Y_i}$ (for G_i), when $\mathbf{Y_i} = \mathbf{y}$, all genes in S_i are switched ON, that is , without loss of generality genes in S_i take on the value 1 with high probability.

Similarly, without loss of generality, let $\mathbf{y}$ be the state in which all members of $\mathbf{Y_i}$ have the value 1, denoted by $\mathbf{1}$; we will consider two situations for G_i, namely the situation where $\mathbf{Y_i} = \mathbf{1}$ and the situation where $\mathbf{Y_i} \neq \mathbf{1}$. Similarly to $\mathbf{Y_i}$ for G_i, let $X_i = (X_{i_1}, X_{i_2}, ..., X_{i_r})$ be the state vector for S_i where $X_{i_k}(1 \leq k \leq r)$ gives the value of $s_{i_k} \in S_i$. In the first case, where $\mathbf{Y_i} = \mathbf{1}$, although the driver is ON, there may be other regulatory activities within the context affecting the driven genes. This would be captured by a parameter that measures for any driven gene $s_{i_j} \in S_i$, the conditional probability of s_{i_j} being ON.

Definition 1 Conditioning parameter (δ_{ij}) *depends on the extent that contextual effects diminish the influence of the driver G_i on the driven gene s_{i_j}.*

$$P(X_{i_j} = 1 | \mathbf{Y_i} = \mathbf{1}) = 1 - \delta_{ij} \tag{3.1}$$

If $\mathbf{Y_i} \neq \mathbf{1}$, then the probability that some driven state $X_{i_j} = 1$ depends on contextual effects alone and not the effects of drivers is captured by the second parameter.

Definition 2 Crosstalk parameter (η_{ij}) *depends on the extent that contextual effects outside of the drivers activate the driven genes.*

$$P(X_{i_j} = 1 | \mathbf{Y_i} \neq \mathbf{1}) = \eta_{ij} \tag{3.2}$$

The paper by Dougherty et al. [42] discusses further considerations of the model, including prediction accuracy and error representation. Figure 3.1 depicts the expression patterns corresponding to different combinations of high and low values of crosstalk and conditioning parameters.

Here, we define some other terms that would be used throughout this dissertation thesis.

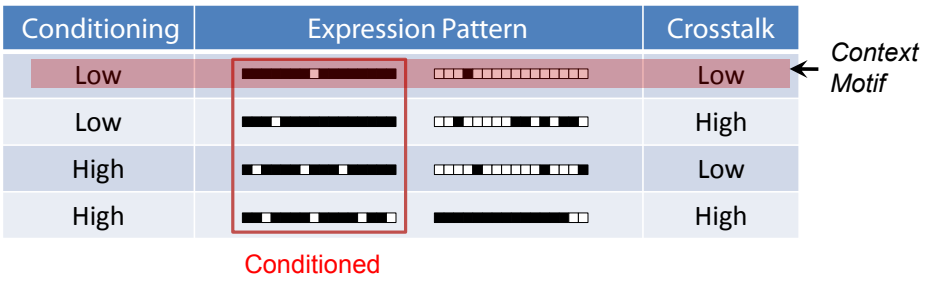


Figure 3.1: Schematic diagram of expression patterns outside of and within cellular context motifs corresponding to different levels of conditioning and crosstalk.

Definition 3 Biological entity *in cellular contexts denotes any cellular constituent, product or external experimental factors (such as environment or drug treatment) that influence, regulate or act specific to the existing cellular state.*

Definition 4 Interactions *are direct or indirect regulatory influences or activity involving two or more biological entities. Biological entity* A *at state(activity) level* Y *is said to have a regulatory influence on another biological entity* B *when the conditioning* δ_{AB} *and crosstalk* η_{AB} *values are lesser than user-specified threshold* δ_θ *and* η_θ*.*

Definition 5 *A* context motif C_i *is a set of interactions between biological entity set* G_i *at state given by vector* Y_i*, and biological entity set* S_i *under the set of common conditions (samples)* T_i*. Thus,* $C_i = \{G_i, Y_i, S_i, T_i\}$ *where* T_i *is defined as a list of conditions where the entity set* G_i *is in state* Y_i*.*

Definition 6 *A* context H *is a network of context motifs corresponding to a set of common conditions* W*,* $W \subset T$*.* $H = (V, E, W)$ *where* V *is the set of vertices which represent interacting biological entities and* E *is the set of edges which represent the interactions pertinent to set of conditions* W*.*

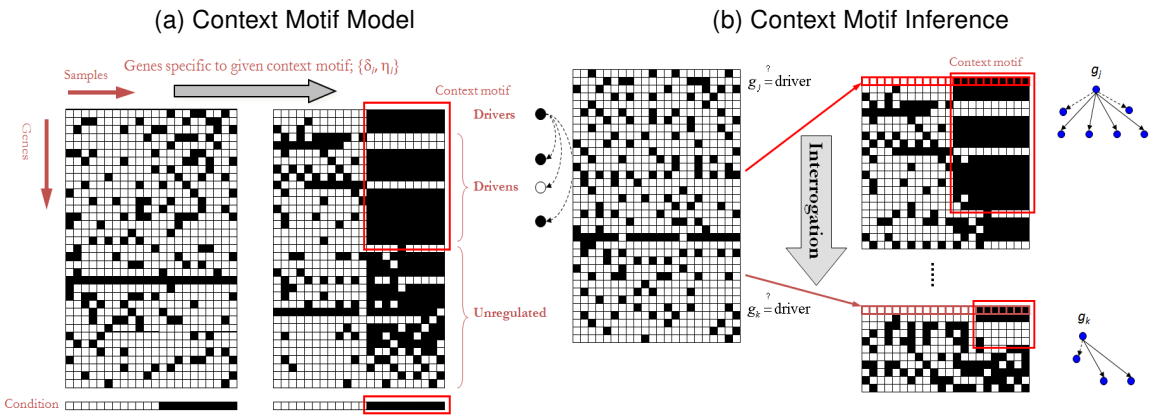


Figure 3.2: Modeling and inference of context motifs. a) Depicts the context motif model with crosstalk and conditioning parameters. b) Provides a schematic for context motif inference algorithm (outlined in Figure 1).

Identification of Context Motifs

A context motif is a set of genes, one or more of which function as drivers and the others as driven genes, which exhibit consistent transcriptional behavior across a subset of samples. We use two statistics — conditioning and crosstalk (defined in Section 3.2) to identify context motifs from gene expression data. The context motifs are used as building blocks for the contexts representing the gene regulatory network.

We apply *in-silico* conditioning [2], a method designed to be similar to a biologist manipulating the status of a gene or conditioning the cells in an experiment with techniques including ectopic expression or gene silencing. With *in-silico* conditioning, the conditionings are not performed manually as the data is collected, but rather computationally on the observations after the data has been collected, hence the name. In this chapter, we demonstrate the method by only considering a single gene driver (conditioner/regulator) at a time for conditioning, although the model allows for more. Figure 3.2a displays context motif modeling and Figure 3.2b inference of context motif by interrogating each gene as a possible driver.

The advantage of the context motif mining method is that it is built upon a biologically-inspired mathematical model, which gives strong meaning to the direction of the edges, i.e., driver (gene) regulating driven (gene). Also, context motif mining identifies each context motif with a corresponding driver gene and a set of samples, thereby ensuring the identification of unique cellular context motifs. Algorithm 1 outlines the algorithm to identify the context motifs.

Input: Gene Set G, Sample Set T, Quantized Dataset $D = G \times T$, Conditioning Threshold δ_θ, Crosstalk Threshold η_θ, Resampling Iterations K
Output: List of Context Motifs $ContextMotifList$

$ContextMotifList \leftarrow$ null;
for *Gene or clinical parameter* g_i *in state* y_i **do**
  $T_i^{y_i} \leftarrow$ Samples where gene g_i is in state y_i;
  /* Identify $Driven_i^{y_i}$ - Genes regulated by g_i in state y_i */
  $Driven_i^{y_i} \leftarrow$ null;
  forall the *Genes or clinical parameter* $g_j, g_j \neq g_i$ **do**
    $\eta_{ij} \leftarrow$ Crosstalk of g_j regulated by g_i in $T_i^{y_i}$;
    $\delta_{ij} \leftarrow$ Conditioning of g_j regulated by g_i in $T_i^{y_i}$;
    if *(*$(\eta_{ij} < \eta_\theta$ *in* $T_i^{y_i}$*) AND (*$\delta_{ij} < \delta_\theta$ *in* $T_i^{y_i}$*))* **then**
      Add g_j to $Driven_i^{y_i}$;
    end
  end
  if $Context\_Motif = \{g_i, y_i, Driven_i^{y_i}, T_i^{y_i}\}$ *is statistically significant* **then**
    Add $Context\_Motif$ to $ContextMotifList$;
  end
end

Algorithm 1: Context motif identification algorithm

If m and n denote the total number of genes and samples in data set, and k denotes the user specified number of iterations for bootstrap sampling to calculate statistical significance (outlined in next section), then the complexity to identify a single context motif is $O(n^3m)$. The complexity to identify context motifs for all possible driver genes is $O(n^3m^2)$. Bootstrap resampling to calculate the statistical significance is $O(n^4mk)$. Thus the complexity of Algorithm 1 is $O(n^4mk + n^3m^2)$. If $nk > m$ then the complexity becomes $O(n^4mk)$ else $O(n^3m^2)$.

Statistical Significance of Context Motifs

For each pairing of driver and driven gene within a context motif, both the conditioning and crosstalk parameters are estimated from the observations. Thus, we consider the statistical significance of the context motifs in order to avoid highly possible false discoveries. We assess the probability of finding a context motif with the same or more number of genes tightly regulated across the same number of samples by chance. We use re-sampling based approach to calculate this hypergeometric probability. If this probability is very low, such as less than 0.05, it is rare to find those estimated values by chance, i.e., it is significantly different from what can be found by chance.

Let (M, N) denote data size where M is the total number genes and N is the number of samples in data set. We also let m and n denote the number of co-regulated genes and the number of observations in an identified context motif, respectively. We estimate $\Pr(m' \geq m | n' = n)$, the probability that a context motif contains larger or equal number of genes than m, given the sub-sample size n. This probability is estimated via re-sampling method. More specifically, we randomly split given data set into two groups of which the one is of sample size n (context motif candidate) and the other of $N - n$. We then apply the same set of statistics (Eqs. 3.1, 3.2) to identify the number of genes filtered by the same thresholds for conditioning (δ_θ) and crosstalk (η_θ). By repeating this procedure many times, we estimate $\Pr(m' \geq m | n' = n)$. The accuracy of the estimation is based on the number of repetitions. In typical setting, 1,000 repetitions are required to provide distribution with enough statistical power. Using this re-sampling-based approach, we assess the statistical significance of identified context motifs. We then apply Benjamini and Hochberg multiple testing correction [43] to the statistical significance values and consider only the filtered context motifs for further analysis.

3.3 Summary

In this chapter, we presented the model and method we developed in this dissertation work to identify putative cellular context motifs via in-silico conditioning. These, if applied to a study of cancer, could lead to the discovery of subtypes of the disease not obvious at the histological level but possibly explained at molecular levels and carry prognostic relevance. We present in later chapters the application of the developed method to the experimental data with disparate data sources to improve understanding of the multilayer interactivity of biological components and help direct further studies. In the next chapter we shift our focus from individual interactions (context motifs) to networks (contexts) and establish a systematic way of agglomerating the context motifs into contexts. We then apply the framework to biological datasets and present corresponding results.

Chapter 4

LEARNING CONTEXT-SPECIFIC GENE REGULATORY NETWORKS

Here we present a method to learn context specific GRNs and assert that the GRNs produced by using context motif mining results exhibit biological advantages absent in related techniques. We use the notations defined in chapter 3, Section 3.2, and develop the network from the context motifs identified by the cellular context mining method outlined in Section 3.2.

4.1 Motivation

Comparable methods such as biclustering [7] or functional module identification [27, 6, 5] algorithm stop once the modules or biclusters have been identified. However, generation of a thousand or more modules identified mainly by numeric identity tags do not easily add to the understanding of how regulatory networks function. The identified bi-clusters or modules may have a lot of overlap in terms of genes, samples or both. Thus to summarize and classify the identified context motifs, we use the structural definition of contexts. The notion of contexts captures the underlying common conditioning in a set of samples and its associated biological entity (e.g., gene-gene) interactions.

Correct representation of context specific GRN requires the description of the cellular context. As we understand, the set of interactions or underlying mechanisms at the cellular level is closely associated with the cellular context. The context motifs capture one step regulation between the driver entity and the driven entity. However, in order to study the cascading controls we need to learn the context itself, i.e., in terms of context motif clusters. All cells belonging to the same cellular context would have a common characteristic set of conditions and biological entity patterns that would distinguish it from other cellular contexts. We want to identify this characteristic set of common conditions and interactions that would define the

cellular context. This would help in the classification or subtyping of different diseases such as cancer and provide insights into their treatment. Here we present the methodology to use context motifs and learn the contexts, assign enrichment values to the context and then apply this to a heterogeneous refractory cancer dataset.

4.2 Related Work

Many studies have found Bayesian networks to be good models for gene regulatory networks, and their popularity has grown in recent years. This makes Bayesian networks a good candidate for comparison with newer network inference algorithms. BANJO (Bayesian Network Inference with Java Objects)[3] is one of the popular Bayesian based GRN reverse engineering method. It searches for the graphical structure in the space of acyclic networks satisfying the conditional dependencies observed in the data. Although Bayesian networks cannot represent cycles, they have been used to define dynamic Bayesian networks which uses a pair of Bayesian networks and transition tables to represent cycles. However, as the structure and number of parameters to be learned increase, the task of learning Bayesian networks and dynamic Bayesian networks from data becomes more difficult.

Other methods to reverse engineer GRNs sometimes employ concepts similar to contexts (as defined in this thesis). For instance, network clustering method [41] uses the definition of subtypes differing in terms of network phenotype. CONEXIC (Copy number and expression in cancer) [36] isolates genes that influence cellular phenotype via changes in driver's expression. LeMoNe (Learning Module Networks) [35] identifies regulatory modules with condition-dependent coherent activity. There are subtle differences in the definition of the problem and the methodologies are structured accordingly.

LeMoNe [35], a module-based algorithm, uses probabilistic ensemble based optimization techniques to infer high quality module networks, where the genes are

[3] Available at http://www.cs.duke.edu/ amink/software/banjo/

first partitioned into coexpression modules and regulators are assigned to modules based on how well they explain the condition-dependent expression behavior of the module. Relevance network based learning CLR(Context Likelihood of relatedness) [9] considers all possible pairwise regulator-target interactions and scores these interactions based on the mutual information of their expression profiles as compared to an interaction specific background distribution. A paper by Michoel et al. [44] compares LeMoNe, module-based method, and CLR, direct pairwise interaction method, and shows that global comparison of results using recall versus precision curves hides the topologically distinct nature of the inferred networks. It distinguished the specific subtasks for which each method is most suited, CLR being 'regulator-centric' (true predictions for higher number of regulators) and LeMoNe being 'target-centric' (higher number of known targets for fewer regulators).

Combinatorial Algorithm for Expression and Sequence-based Cluster Extraction (COALESCE) [37] allows discovery of regulatory motifs and modules from large collections of genomic data such as gene expression and DNA Sequence Data. CONEXIC [36] influence cellular phenotype via changes in driver's expression. It integrates matched copy number changes (amplifications and deletions) and corresponding gene expression data from tumor samples to identify driver mutations and the processes that they influence. A score-guided search identifies the combination of modulators that best explains the behavior of a gene expression module across tumor samples and searches for those with the highest score within the amplified or deleted regions. The method has been developed to reduce the selection of modulators that are not drivers. To gain this specificity, they do not detect all genes and pathways that drive tumors. CONEXIC only identifies candidate drivers that are encoded in amplified or deleted regions. The main limitation of such methods is the way that the amplified or deleted regions are getting associated with the gene. Recent work by Li et al. [45] show that the RNA sequences do not cor-

respond exactly to the DNA sequences. The identified differences were shown to be nonrandom, as many mismatched exonic sites were found in multiple individuals and in different cell types, including primary skin cells and brain tissues. Therefore, unless we have a way of accurately determining the correspondence between the DNA and RNA sequences, methods such as COALESCE and CONEXIC may miss important cellular phenotypes or create false positives in the module.

4.3 Methodology

The normalized, quantized dataset is input to the cellular context motif mining algorithm. Each conditioning of a gene set G_i (driver) on a vector expression value $Y_i = y_i$ yields a subset of samples T_i within which a set of genes $S_i = \{s_{i_1}, ..., s_{i_k}\}$ appears to be tightly regulated, so a cellular context motif is defined as $C_i = \{G_i, Y_i, S_i, T_i\}$. A re-sampling approach is used to determine the most statistically significant context motifs represented in the data. A user-set threshold is used to filter out the statistically significant context motifs.

Contexts from Context Motifs

Note that each context motif defines regulatory relationships $g_i \rightarrow g \in S_i$, specific to T_i with G_i (driver) conditioned on a value $Y_i = y_i$. Thus, gene g_i at state y_i uniquely defines a set of samples and is included as a driven in the context motif C_x only if there is a high overlap (low value of conditioning) with the samples of C_x. This ensures that the gene g_i is a part of context motifs only when it has a significant overlap with the sample set of the context motif driver. If we group all the context motifs which share g_i at state y_i, we naturally obtain a high number of common samples between these context motifs. As contexts capture common interactions in specific sets of samples, these implicit relationships leads to the construction of context specific regulatory networks.

As shown in Figure 4.1, the driver g_j of the context motif C_j might be driven

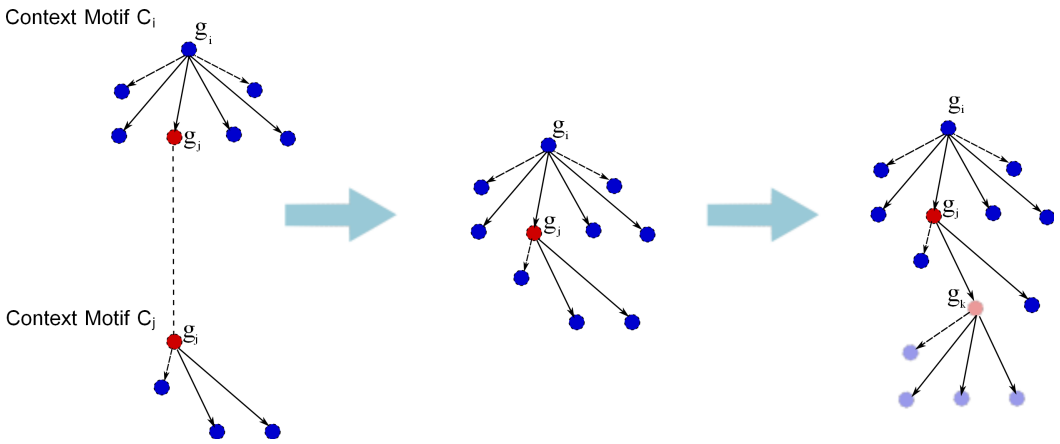


Figure 4.1: Chaining of context motifs to form the context motif network. Clustering of the context motif network yields the contexts

by another driver g_i of context motif C_i. The chaining of such regulatory relationships $g_i \rightarrow g_j$, in addition to implicit driver-driven relationships $g_i \rightarrow g \in S_i$, results in an interesting graphical structure, representing relationships between context motifs. We call this a context-specific gene regulatory network (GRN) as each regulatory relationship $g_i \rightarrow g \in S_i$ is specific to corresponding subset of samples, T_i.

A context-specific GRN differs from other representations not in its graphical structure, but by the fact that context motifs connected to one another in a network differ in their sample composition. Formally, a context-specific GRN H is represented as $H = (V, E, W)$, where V is a set of genes (biological entities) representing vertices, E is a set of edges oriented from genes (biological entities) designated as drivers to genes (biological entities) designated as driven and W is the set of associated conditions; thus H is a directed graph structure, though not necessarily acyclic, since a driven gene in one context motif may be a driver in another.

Again, note that each edge e_{i*} is specific to only its corresponding subset of samples, T_i, where e_{i*} refers to $g_i \rightarrow g \in S_i$. We observe that not only do context-

specific GRNs report verifiable (and possibly novel) relationships between genes, but moreover the overall network structure groups itself into biologically meaningful and readily annotated context motif clusters, i.e., contexts. For proof of concept, we present the results of application of this technique to the refractory cancer Target Now (TN) data set, which includes gene expression profiles of 146 patients with refractory cancer (section 4.4).

Markov Clustering

Any clustering method can be applied to identify the contexts from the context motif network. In some cases, the contexts are easily visually separable. However, to cater for cases with very dense and large context motif networks, there is a need to use clustering algorithms that would automate the identification of contexts from the context motif graphs. An important constraint is we cannot predefine the number of clusters. Thus, a viable option is using Markov clustering (MCL). The MCL algorithm [46] 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 MCL process causes flow to spread out within natural clusters and disappear in between different clusters.

Enrichment Analysis

We use biological enrichment analysis to validate the contexts obtained from context motifs to determine if we get biologically meaningful results using the method. For instance, one approach would be to use Gene Ontology [32] term analysis of genes belonging to the same context for testing functional enrichment. Gene

Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases. However, this approach does not take into account any of the conditions (in terms of associated samples) for analysis. We would need to determine if obtained contexts indeed display specificity in terms of samples or tissues. Here, we develop association scores of contexts [47] based on the specificity of the samples associated with context motifs constituting the context.

Sample Association to Context

A context is a set of context motifs connected to one another through intra-context regulation. It would be informative to associate a set of samples to each context based on its strength of association with the member context motifs. For example, in a heterogeneous cancer dataset, the sample association to contexts would allow annotation of the context motif cluster as a partial representative of cancer type. Since one context is comprised of potentially many context motifs, each representing a particular subset of biological samples, it is of interest which of those samples appears in more than one context motif in the context. Samples are scored on the basis of occurrence within the context motif, over all the context motifs found in the context. We developed a sample association score [47], associating a sample s, with a context C consisting of m context motifs $\{C_1, C_2, \ldots, C_m\}$, with the scoring:

$$SAS(s, C) = \sqrt[m]{\prod_{i=1}^{m} f_i(s)}, \text{ where } f_i(s) = \begin{cases} k_i/N & s \in C_i \\ 1 & \text{otherwise} \end{cases} \tag{4.1}$$

where k_i is number of samples within context motif C_i and N is the total number of samples in the gene expression data. The sample which occurs in all context motifs of the context cluster would then have the least score, and the sample which is not present in any of the context motifs will have a score of 1. The samples with score less than 0.5 would be associated with the corresponding context cluster. Continuing the heterogeneous cancer dataset example, the selected samples can

be used for the calculation of the distributions and tumor types across all context clusters.

Probability of sample s associated with context C, can be estimated using the following formula:

$$\Pr(s \in C) = 1 - \prod_{C_i \in C} p_i^{I(s)}, \text{ where } p_i = k_i/N \text{ and } I(s) = \begin{cases} 1 & s \in C_i \\ 0 & \text{otherwise} \end{cases} \quad (4.2)$$

where k_i is number of samples within context motif C_i and N is the total number of samples in the gene expression data.

Tumor Type Enrichment

After sample association to specific context cluster, each cluster can be subjected to a statistical test for enrichment of specific types of tumors. The Yates corrected chi square test for significance would be applied (if numbers are less than 5) to each tumor type-context cluster pair. This will allow annotation of significantly enriched tumor type association with each context.

Comparison with other methods

Not all methods for reverse engineering GRNs are directly comparable to our method. The first reason is context specificity, accounting for associated subsets of samples instead of assuming that relations between genes extend across all samples. This makes comparison with methods like Bayesian networks and mutual information networks difficult. The second is, accounting for the state of the gene - over expressed, under expressed or no-change, associated to distinct context motifs and thereby contexts. In case of biclustering it is implicitly captured by the biclusters but not so in the case of mutual information based relevance networks or Signature Algorithms [6, 5]. Thus for comparison we would have to choose a dataset which will allow us to test contextual network learning by different methods. Biological real world data have embedded context specific information, but as the true under-

lying networks are unknown, the data cannot be used to directly verify the results obtained from our framework. In the next section, we engineer synthetic data sets which mimic the heterogeneous nature of the true biological system. Such data sets will have context-specific regulation pre-determined and embedded so that the identified interactions can be quantifiably validated. In order to use these synthetic data sets to validate the context-specific GRNs produced through the cellular context mining technique, we must avoid bias by generating the networks by a method other than that which we want to validate.

Artificial Contextual Networks and Data

Artificial networks and corresponding datasets generated by biochemical simulator A-biochem [19] were used as individual contextual networks for comparison. These datasets have been used as benchmark datasets and cited by different groups to verify reverse engineering GRNs (Dialogue for Reverse Engineering Assessments and Methods (DREAM)[4], ARACNE [4]). The datasets have a strong mathematical model used to create the relationships between genes, TFs, mRNA etc. within the network. They use rate laws of transcription and mRNA degradation to determine the dynamics of gene networks. The rate laws are mathematical expressions, which relate the rate of reaction (transcription, etc.) to the concentration of several substances (effectors). The rate of transcription responds to the concentrations of nucleotides, RNA polymerase, and transcription factors. The models ignore the effect of the nucleotides and polymerase but use other effects from other gene products that could be positive (activation) or negative (inhibition). Induction and repression over time steps ensure change in the networks. These account for the combinatorial effect of activation and inhibition influences to direct network data generation.

[4] http://wiki.c2b2.columbia.edu/dream/index.php/The_DREAM_Project

We used the Century data series as generated through A-biochem[5]. Each Century network consists of 100 genes with a total of 200 gene interactions (on average each gene has 2 modulators). All networks are composed of genes with similar kinetics, the only difference between networks is how the gene interactions are organized (i.e. which genes induce and repress which other genes). The networks belong to three major groups according to their topologies. Each set of fifty networks was based on different topologies – scale-free network topology, random network topology and small world topologies. For our application we used the first fifty networks which were based on scale-free network topology. To pre-process the dataset we converted dataset to ratios by dividing each null mutant gene expression with the corresponding wild type gene expression. Next, we used a quantization threshold of 2 standard deviations from mean to identify differentially expressed genes and discretize the expression values. Finally, considering each artificial network as a separate context, we randomly combined discretized datasets and using the corresponding original artificial graph information we created the composite of the edges represented in the combined dataset.

4.4 Results

Application to Artificial Contextual Network Data

In order to compare the performance of Expattern with other comparable methods, we chose popular methods ARACNE [39] (undirected graph, based on MI) and BANJO [8, 40] (directed graph, based on Bayesian networks).

Undirected Edges Comparison

We applied ARACNE [39] to the artificial contextual network and found the best performance in terms of precision and f-measure at Mutual Information (MI) threshold and DPI settings at 0.15 and 0.01 respectively. We used these settings of ARACNE throughout all runs. Expattern used the discretized version of the dataset

[5]Data available at“http://www.comp-sys-bio.org/AGN/data.html”.

at 2 standard deviations for quantization. Figure 4.4 shows Expattern performs better than ARACNE in f-measure values in case of multiple contexts. Noticeably, the f-measure and precision values for the single context ARACNE runs are better than the multi-context runs of ARACNE. This depicts how methods such as ARACNE are better suited to find interactions in case of homogeneous datasets (single contexts) but the performance deteriorates if we introduce heterogeneity (multiple contexts). This is understandable as ARACNE looks for a single underlying network, whereas Expattern separates out the different contextual networks. We also calculated paired t-test p-values to compare the performance of Expattern with ARACNE. We found very low p-values for precision ($1.20e^{-46}$) and recall ($2.45e^{-07}$), denoting significant differences in results obtained by Expattern and ARACNE for multiple context cases.

Directed Edges Comparison

BANJO (Bayesian Network Inference with Java Objects) is one of the popular Bayesian based GRN reverse engineering method. It searches for the graphical structure in the space of acyclic networks satisfying the conditional dependencies observed in the data. We used BANJO to compare directed edges found by Bayesian inference methods with the directed edges found by Expattern. We applied BANJO and Expattern to the same discretized dataset. BANJO used simulated annealing search with 50000 restarts and 1000 minimum networks to search for Bayesian network structures before checking for consensus. The ten highest scoring networks were retained and combined into a consensus network (represented as BanC in figures). While this consensus network may lose the Bayesian structure properties (namely acyclicity), it gives a generalized summary of the putative interactions over the highest scoring network alone.

Figure 4.4 presents performance comparisons, as applied to artificial contextual networks, between highest scoring network found by BANJO, BANJO Con-

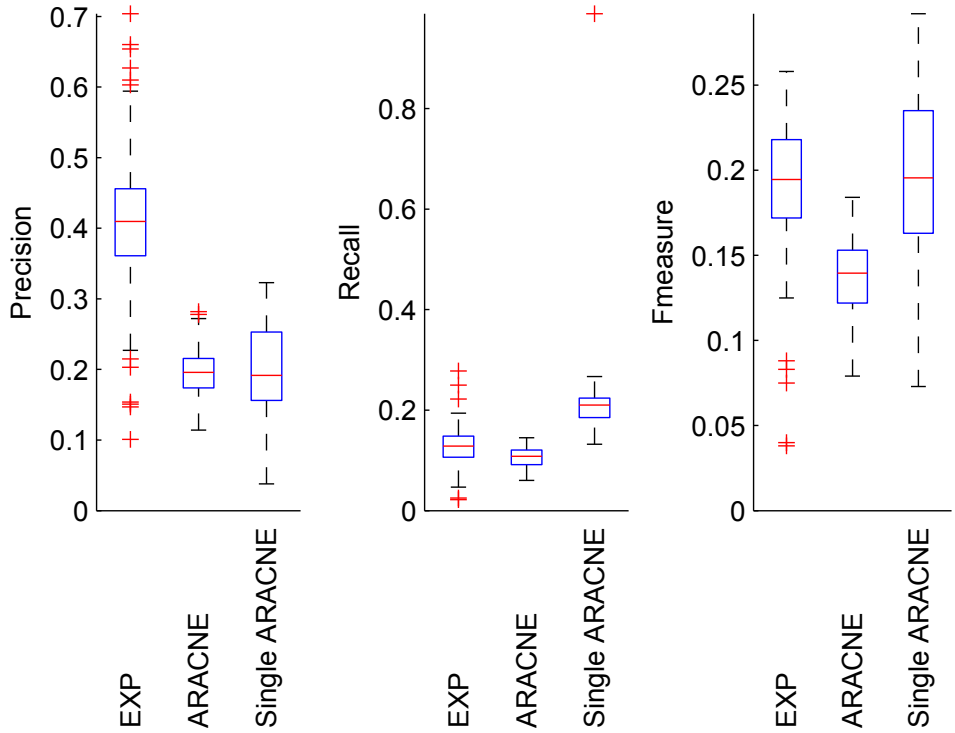


Figure 4.2: Precision, recall and f-measure comparison for undirected edges of two context networks. The results depicted are averaged over 100 pairs of randomly combined context networks. Single ARACNE depicts the average result for 50 single context network runs.

Table 4.1: Paired t-test p-values: Precision, recall and f-measure p-value comparison of BANJO, BANJO consensus with Expattern results.

Method vs Expattern	Precision p-value	Recall p-value	F-measure p-value
BANJO	0.0020	1.92E-60	7.53E-49
BANJO Consensus	0.0825	4.49E-61	5.99E-49

sensus network and Expattern context motif network. Expattern performs well in terms of recall and f-measure when compared to both BANJO and BANJO consensus network. However, the precision values of Expattern were only marginally better than others. Table 4.1 presents the paired t-test p-values of performance metrics (precision, recall and f-measure) of BANJO and BANJO Consensus network compared with Expattern. The differences in f-measures and recall between the methods were more significant than the precision differences. Figure 4.4 provides a

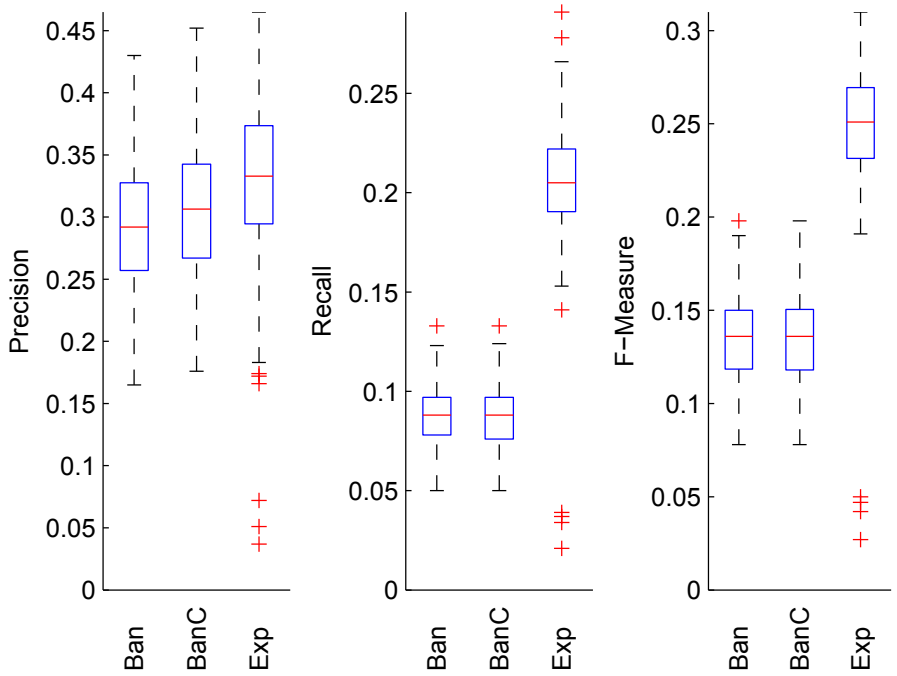


Figure 4.3: Precision, recall and f-measure comparison for directed edges of two context networks (BANJO, BANJO Consensus and Expattern). The results depicted are averaged over 100 randomly combined pairs of artificial context networks.

closer examination of the distribution of true positive and false positive edges identified by each of the methods. We observed that high recall of Expattern ensures not only a higher number of identified true positive edges, but also, a very high number of false positive edges. This raises the question whether we can develop any strategies to reduce the number of false positives and improve precision of Expattern output. The next chapter explores different strategies to reduce false positive edges in Expattern context motif network.

Application to Refractory Cancer Data

We applied our framework to the refractory cancer Target Now (TN) data set, which includes gene expression profiles of 146 patients with refractory cancer. The motivation of the Target Now (TN) study (http://www.targetnow.com) is to determine whether patients with refractory cancer, who had not received a benefit from the

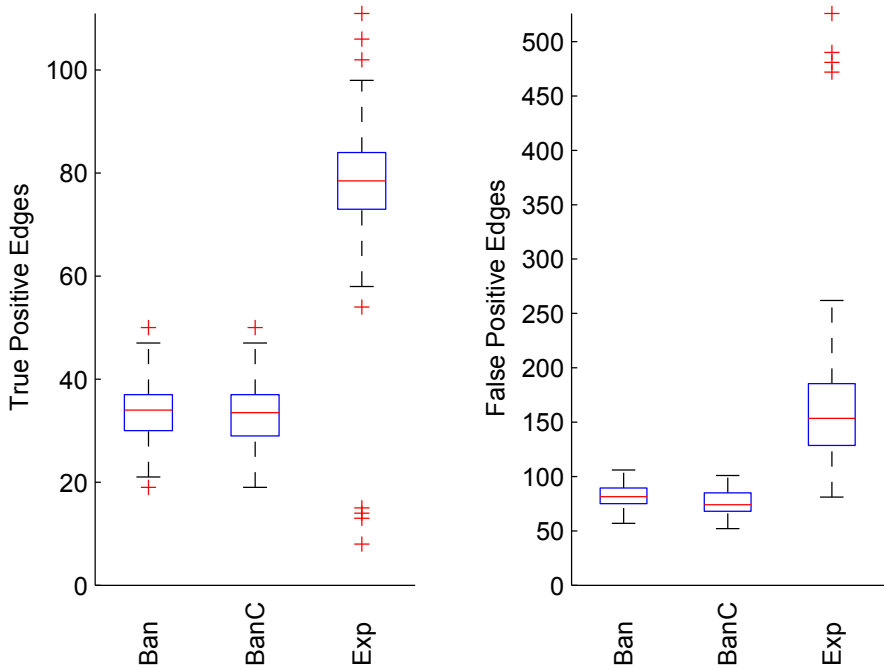


Figure 4.4: True Positive and False Positive edges averaged over 100 randomly combined pairs of artificial context networks.

standard types of treatment, could derive benefit from therapy with a drug not normally used for their particular form of cancer. The therapeutic to apply is one that has activity against a gene target that is found to be altered in that patient's cancer. The cancer patients contributing to the TN study all have late stages cancer. Late stage cancer is very frequently very de-differentiated, having lost a great deal of the specialized functions present in the tissue from which it arose. Due to this biological simplification of the system, those genes whose abundance is found to be altered from the normal tissue of origin and whose change of abundance is found in other refractory cancers of the same type or of different types may be representatives of changes that are necessary to support a particular molecular subtype of cancer.

The TN dataset, which consists of 17,085 unique probes from 146 patients with different types of refractory cancer, was used to learn context-specific GRNs.

Table 4.2: Target Now Dataset Sample Distribution with the number of samples associated with different cancer tumor types.

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 dataset was pre-filtered based on transcription activity of each gene across the samples to be reduced to only 4,000 probes. The distribution of the 146 samples between different cancer tumor types is listed in Table 4.2.

Running the context mining algorithm with a strict statistical significance threshold resulted in 205 context motifs (p-value < 0.0005). Using these context motifs, the method described to create context specific GRNs yielded a directed graph with 1,790 vertices (genes) and 9,566 edges (regulatory relationships), as shown in Figure 4.5. This graph had an interesting property of being systematically fragmented into four separate contexts, which were identified by locating the weakly connected components in the graph. These contexts provide a useful approach to interpreting the context motifs found by the context motif mining algorithm. The contexts, typically displayed significant overlaps among their subsets of samples. This is due to complex inter-connections among drivers that result from common cellular processes being shared among them.

When investigating the four disjoint contexts, we noticed the two largest context clusters consisted of densely connected parts loosely bound to one another. Seeking to further characterize the data on the basis of very dense connectivity, we investigated the connections within the two largest contexts. On the basis of density of connection and directionality of control, we resolved the four original contexts into seven biologically separable ones. In Figure 4.5, bottom right, we segregated

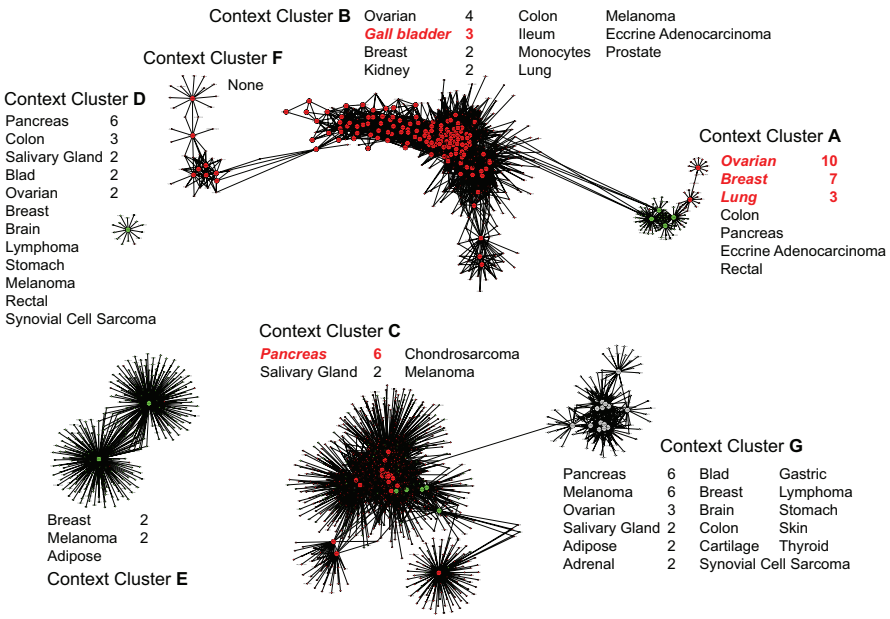


Figure 4.5: Context-specific GRNs - each context motif cluster is annotated with the corresponding set of samples and highlights significantly enriched tumor types in red. See Table 4.2 for cancer tumor sample distribution in the dataset. In the graph, red vertices represent over-expressed genes, green under-expressed, and grey neither under- nor over-expressed. Edge orientation is driver genes (large vertices) to driven genes (small vertices).

the first large cluster into contexts C and G. Context G is easily separable as all its genes are neither under- nor over-expressed (unlike C), and only one edge exists between the contexts C and G (C drives a gene also driven by G). These characteristics convinced us that C and G should be analyzed as separate contexts. The weak connection may have been rooted in tissue of origin similarity, as between them they account for two-thirds of the pancreatic samples in the data with six members in each.

Next, we segregated the top large cluster in Figure 4.5 into context clusters A, B and F. All driver-to-driver edges between A and B are oriented from A to B, implying a hierarchical regulatory relationship from A to B. Also, like C and G, their connection in the graph is explained by the fact that both A and B represent significant numbers of both breast and ovarian tumor types. Contexts B and F share

Table 4.3: Chi-square enrichment test p-values of tumor types in different context clusters

Context Cluster A		Context Cluster B		Context Cluster C	
Tumor Type	p-value	Tumor Type	p-value	Tumor Type	p-value
Ovarian	2.3E-05	Gallbladder	1.6E-04	Pancreas	8.2E-05
Breast	0.0057				
Lung	0.012				

four edges, two involve genes driven by drivers in both B and F. The two remaining edges are both directed from F to drivers in B, indicating a possible hierarchical regulatory relationship between them.

Each of the seven contexts are visible in Figure 4.5 and have the associated tumor types next to them. Enriched tumor types are highlighted in red and the numbers next to all tumor types correspond to the number of samples distinguished as significant by the scoring function (Equation 4.1, discussed in next section). Table 4.2 contains the TN dataset sample composition. Sample enrichment is depicted in Figure 4.5 displaying the tumor types having nonzero sample counts corresponding to each context cluster. Significant Tumor Type enrichment results are summarized in Table 4.3. Figure 4.5 highlights (in red) the tumor type considered enriched within the corresponding context cluster.

Intriguingly, context A showed significant tumor enrichment of ovarian cancer, breast cancer and lung cancer. A literature survey shows breast cancer drugs are being used in the treatment of lung cancer [48], because of vital role of estrogen in lung development and subsequently cancer pathway. Literature survey verified some known gene interactions and relationships to diseases within context clusters. Context A involved breast cancer, ovarian cancer and lung cancer, and included genes such as TNFRSF1A, known to promote breast cancer[48]; CD74, usually expressed in ovarian and lung cancers, considered as a target for Multiple Myeloma treatment therapy [49]; HLA-DM, its expression when combined with

that of HLA-DR, is considered to influence breast tumor progression and patient outcome [50]. Context C, related to pancreatic cancer, contained GP2, a protein specifically expressed in pancreatic acinar cells and considered as a diagnostic marker in animals [51].

Conventional approaches such as clustering and Bayesian Network learning provide some ability to observe sample enrichment, but they do not exploit the association of particular expression behaviors in subsets of the samples to the fullest extent. Since clustering and Bayesian Network learning implicitly assume that the observed data is from a single distribution, their results are always diluted approximations relative to results that assume the observed data to have come from various different distributions and evaluate them in appropriate isolation.

We compared our method to hierarchical clustering and k-means clustering using similarity metrics of correlation and Euclidean distance, in Cluster version 3.0 [52, 53], to group samples with similar gene expression profiles together. We verified that in cases where a similar number of clusters (six or seven) were identified by Cluster 3.0, the conventional clusters display significant overlap (ranging from 40% to 90% overlap) with context clusters in terms of samples (and thus tumor type enrichment). Conventional clustering algorithms do not however provide a quantitative evaluation with which to isolate vital gene markers or describe the genes' activity for the subtype of disease described by the sample subset. The context motif cluster approach has a distinct advantage of extracting relevant genes pertaining to the particular disease type.

4.5 Summary

In this chapter, we presented a method to consolidate similar context motifs and build the network of context motifs. We also introduced an innovative score – Sample Association Score and evaluated (in TN dataset case) the Tissue Enrichment Score. The sample association score characterizes the context motif networks with

enriched samples, an effective method for classification in the unsupervised learning of subtypes of cancer. We can use any clustering method to cluster the context motif network to obtain the contexts. We applied the framework to both artificial datasets and refractory cancer data. We presented a unique method to create artificial contextual gene expression data and network using Century series data (from A-biochem), a widely used repository of artificial datasets. We validated our framework results on the artificial contextual networks and compared the performance of this framework with two very popular methods - ARACNE (undirected edges) and BANJO (directed edges). We observed that, although Expattern shows a higher number of True Positives than BANJO, its overall precision and f-measure are low because of a greater percentage of False Positives. Thus to improve the performance of Expattern, we need to develop strategies to reduce the number of false positives in the contextual network. The next chapter explores and develops different approaches to do the same.

Chapter 5

PRUNING CONTEXT-SPECIFIC GENE REGULATORY NETWORKS

Context-specific gene regulatory networks use probabilistic measures of consistency to infer gene regulatory relationships. Consequently, the algorithm captures indirect influences between genes, resulting in several false positives, as well as redundant edges. While statistical p-value thresholds reduce the network to the most significant regulatory relationships, overlap between contexts still leads to a large number of redundant edges within the network.

In this chapter, we present a method for pruning context-specific GRNs, derived from the relationship between the consistency metrics used to learn the regulatory interactions. Apart from a theoretical proof of concept, we assess the performance of our methods based on the sizes of the reduced network as well as its ability to capture biologically relevant regions of the network. We apply our methods to a cancer dataset and show how the reduction strategy is able to remove redundant edges while preserving the functional enrichment of the network. Further, we compare the performance of several variants of the pruning strategy with the transitive reduction method and show how our method is superior in terms of both performance as well as biologically significant clusters.

5.1 Motivation

In last few chapters, we presented our framework to learn context-specific GRNs from gene expression data. In context-specific regulatory networks, an edge between two nodes corresponds to the two nodes being consistently regulated; consistency being defined based on probabilistic measures of similarity in expression values. Unlike conventional GRNs, 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 method discretizes

the data at a predefined quantization level and then learns the network from the discretized data. However, the problem of selecting the quantization level is not a trivial one and greatly affects the final network constituents and structure. Too stringent quantization levels removes the associations between genes and tolerant or lax quantization levels introduces false associations between genes in the discretized data.

Also, GRNs learned by the framework (method outlined in Chapter 4) are often made of a few thousand nodes (genes) and tens of thousands of interactions rendering interpretation of the network almost impossible. Overlapping contexts adds to the difficulty in interpretation. In order to compensate for quantization effects and statistical analysis artifacts on increased false positive edges found through our method, we propose context-specific GRN pruning methods. These remove extraneous edges, exploiting relationships between the consistency metrics - crosstalk and conditioning (section 3.2). We propose scale-free topology based pruning, random network topology based pruning and several variants of network pruning for the removal of reverse edges, sibling edges and transitive edges. The different strategies are described in the methodology Section 5.3.

5.2 Related Work

Graph pruning methods have been prevalent for the last few decades and can be roughly categorized into generic and domain dependent pruning strategies based on the pruning objective. The goal of generic pruning strategies is to reduce the graph size while maintaining certain properties of the graph. Methods include graph spanners, where the complex graph is replaced with a sparser one, while preserving all or most pairwise distances [54],[55] and tree approximation algorithms, where the graph is replaced by a tree or a set of trees while preserving pairwise distances. Another method – transitive reduction, removes edges while ensuring that connectivity between nodes is maintained. Transitive reduction has

been applied to both social networks [56], gene regulatory networks [57] and signal transduction networks [58]. Graph pruning is also considered as a variation of the feedback node set problem, where the goal is to remove vertices and edges of the graph with the aim of breaking existing cycles [59].

Interestingly, there are also several domain-dependent methods for pruning. Graph pruning is often applied in speech recognition, in order to prune word graphs. Such methods include forward-backward pruning strategies by [60] as well dynamic programming methods by [61]. In the world of gene regulatory network learning, pruning strategies are less prevalent. Transitive reduction strategy has been applied to gene regulatory networks [57] and signal transduction networks [58]. ARACNe [39], based on Mutual Information, uses the Data Processing Inequality (DPI) criteria to prune interactions corresponding to conditionally independent genes. Others, such as, network inference algorithms allow for controlling network sizes through various ways. Boolean network learning methods allow control of the size of the network by considering all Boolean functions of no more than k variables [62]. Similarly, in Bayesian network learning, network sizes are usually controlled by constraining the search space using either graph properties [63] and/or prior biological knowledge [64]. Boolean networks, Bayesian networks and context-specific gene regulatory networks [2] also use functional annotations to reduce the learned gene regulatory networks to a handful of interesting pathways. Depending upon the nature in which the search space is constrained, such methods could lead to errors of omission.

5.3 Methodology

We propose an approach to reduce the redundancy in context-specific GRNs by removing extraneous edges, exploiting relationships between the consistency metrics — crosstalk and conditioning [2]. We propose several variants of network pruning for the removal of reverse edges, sibling edges, transitive edges focusing on two

different topology based pruning strategies — Random Topology Based Pruning and Scale-free Topology Based Pruning.

Given a context-specific GRN H, our goal is to remove edges in the network that may be artifacts of indirect dependencies in the evidence data. For example, Figure 5.1a depicts a case where gene A regulates gene B, which in turn regulates gene C, and an edge is incorrectly inferred between genes A and C, in other words a redundant transitive edge. Figure 5.1b depicts the case where gene A regulates both genes B and C, and an edge is incorrectly inferred between genes B and C, that is, a redundant sibling edge. Finally, the case shown in Figure 5.1c is where gene A regulates B with very low level of latent intervention, this can lead to low conditioning and crosstalk from B to A, thus, resulting in an incorrect edge from B to A, that is, a redundant reverse edge. We derive the expected crosstalk and conditioning values using the crosstalk and conditioning values of all other edges in the triad (triplet of connected genes such as A, B, C in Figure 5.1). The expected values are then used to construct the conditions for edge pruning. It is important to understand that determination of the correct directionality of the edge, from the data itself, is not always possible. Thus, we assume for each directed edge in a triad of genes such as in Figure 5.1, if the crosstalk and conditioning values are lesser than expected values, then they are considered to be true edges. Otherwise, they are treated as redundant/extraneous edges and need to be removed. The algorithms we present here prune extraneous edges from the graph. The derivations for all

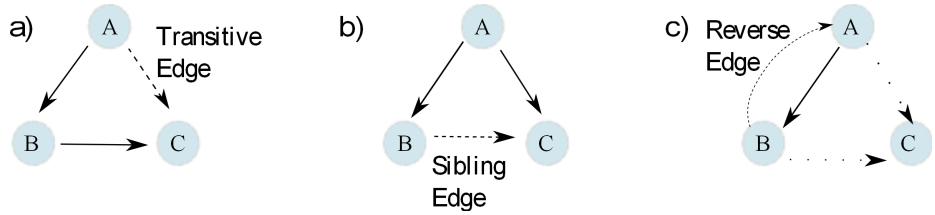


Figure 5.1: Edge Types — Transitive Edge, Sibling Edge and Reverse Edge in a basic interaction triad of genes A, B, C.

the cases can be found in Appendix A, we present only the final results in the corresponding sections.

Once we obtain the context-specific GRN, we use combinations of two methods, edge removal and pruning order determination to remove extraneous edges. We apply edge removal methods in the order determined by the chosen pruning order strategy. Next, we apply Markov Clustering (MCL) [46, 65] to the reduced context motifs and obtain the reduced contexts. Finally, we run enrichment analysis and compare the reductions across different strategies.

Edge Removal

Transitive Edge Removal

Transitive edge would be deemed false positive if there is strong evidence of indirect conditioning between the driver and any of its drivens' driven. For example as in Figure 5.1a, let us assume A to be the true driver gene and B to be the true driven gene in one context motif and B to be the true driver gene and C to be the true driven gene in another context motif. Then both edges AB, BC are present in GRN H, and the expected values of crosstalk and conditioning of edge AC can be calculated. The following theorem presents the result.

Theorem 5.3.1 (*Transitive Conditioning*) *Given the values of conditioning and crosstalk of edges AB, BC as below,*

$$\delta_{ab} = 1 - \Pr(y_b = 1 | y_a = 1) \equiv 1 - \Pr(B|A)$$

$$\eta_{ab} = \Pr(y_b = 1 | y_a \neq 1) \equiv \Pr(B|A')$$

$$\delta_{bc} = 1 - \Pr(y_c = 1 | y_b = 1) \equiv 1 - \Pr(C|B)$$

$$\eta_{bc} = \Pr(y_c = 1 | y_b \neq 1) \equiv \Pr(C|B')$$

where, y_a denotes the state vector of gene A and $\Pr(A)$ denotes the probability of gene A to be in state $y_a = 1$ (ON or active). Also $\Pr(A, B)$ is used to represent the

probability of gene A to be in state $y_a = 1$ and gene B in state $y_b = 1$. Then, the expected conditioning on edge AC is

$$\delta_{ac} \geq [\delta_{ab}(1-\eta_{bc}) + \delta_{bc}(1-\delta_{ab})] + [\eta_{bc}(1-\eta_{ab}) - \eta_{ab}\delta_{bc}]\,\gamma_A \tag{5.1}$$

where

$$\gamma_A = \frac{\Pr(A')}{\Pr(A)} = \frac{1-\Pr(A)}{\Pr(A)}.$$

Proof. Provided in Appendix A. ■

Now, given conditioning and crosstalk values as above, we can derive the expected value of transitive crosstalk as presented in the following theorem.

Theorem 5.3.2 (Transitive Crosstalk) *Assume if $\delta_{ac} > \delta_{bc}$,i.e., $\Pr(C|A) < \Pr(C|B)$, then the expected crosstalk on edge AC is*

$$\eta_{ac} > \frac{\eta_{bc}\cdot\gamma_B + \alpha_{bc}\cdot\{\alpha_{ab}-1\}}{\gamma_A} \tag{5.2}$$

where

$$\alpha_{ab} = \frac{\Pr(B)}{\Pr(A)},$$

and γ_A and γ_B given as above.

Proof. Provided in Appendix A. ■

Algorithm 2 with complexity of $O(n)$ tests and removes redundant transitive edges.

Sibling Edge Removal

Sibling edges might be wrongly inferred if one of the target becomes a driver and closely mimics the activity of any of its siblings. As in Figure 5.1b, if we assume A

Input: Ordered Vertices A, B, C, Edges AB, BC, AC
Output: List of Retained Edges

```
removeEdgeAC ← true ;
if { (η_ac ≥ η_ab) & Equation 5.1 not satisfied} then
    if { (δ_ac > δ_bc) & Equation 5.2 not satisfied} then
        removeEdgeAC ← false ;
    end
end
if (removeEdgeAC) then
    Remove Edge AC ;
end
```

Algorithm 2: Transitive edge removal algorithm

to be the true driver gene and B, C to be the true driven genes, with edges AB, AC present in GRN H, the expected values of crosstalk and conditioning of edge BC can be calculated. Results are presented in the following theorem.

Theorem 5.3.3 (Sibling Conditioning) *Given the values of conditioning and crosstalk of edges AB and AC, if we assume $\eta_{bc} \geq \eta_{ac}$, i.e., $\Pr(C|B') \geq \Pr(C|A')$, then the expected conditioning on edge BC is*

$$\delta_{bc} \geq 1 - \{(1 - \delta_{ac}) \cdot \alpha_{ba} + \eta_{ac} \cdot (1 - \alpha_{ba})\} \tag{5.3}$$

where α_{ba} is defined similarly as above.

Proof. Provided in Appendix A. ■

Theorem 5.3.4 (Sibling Crosstalk) *Assume if $\delta_{bc} \geq \delta_{ac}$, i.e., $\Pr(C|B) \leq \Pr(C|A)$, then the expected crosstalk on edge BC is*

$$\eta_{bc} \geq \frac{\eta_{ac} \cdot \gamma_A - \alpha_{ac}\,(1 - \alpha_{ba})}{\gamma_B} \tag{5.4}$$

where $\alpha_{ac}, \alpha_{ba}, \gamma_A$ and γ_B are defined similarly as above.

Proof. Provided in Appendix A. ■

Algorithm 3 tests and removes redundant sibling edges with complexity of $O(n)$.

Input: Ordered Vertices A, B, C, Edges AB, BC, AC
Output: List of Retained Edges

```
removeEdgeBC ← true ;
if { (η_bc > η_ac) & Equation 5.3 not satisfied} then
    if { (δ_bc ≥ δ_ac) & Equation 5.4 not satisfied } then
        removeEdgeBC ← false ;
    end
end
if (removeEdgeBC) then
    Remove Edge BC;
end
```

Algorithm 3: Sibling edge removal algorithm

Reverse Edge Removal

Reverse Edges would be wrongly inferred if the crosstalk and conditioning values between the driver and the driven are very low. In such cases it is hard to determine which is the true driver. As in Figure 5.1c, if we assume A to be the true driver gene and B to be the true driven gene, with edge AB present in GRN H, the induced values of conditioning and crosstalk are provided as below.

Theorem 5.3.5 (Reverse Conditioning) *Given conditioning value δ_{ab} of edge AB, the expected conditioning on edge BA is*

$$\delta_{ba} = 1 - (1 - \delta_{ab}) \cdot \frac{\Pr(A)}{\Pr(B)} \tag{5.5}$$

(5.6)

Proof. Provided in Appendix A. ■

Theorem 5.3.6 (Reverse Crosstalk) *Given crosstalk value η_{ab} of edge AB, the expected crosstalk on edge BA is*

$$\eta_{ba} = 1 - (1 - \eta_{ab}) \cdot \frac{\Pr(A')}{\Pr(B')} \tag{5.7}$$

Proof. Provided in Appendix A. ■

An edge BA would be included in the context graph H, given that edge AB is already present in H iff

$$\delta_{ba} < \delta_\theta \text{ AND } \eta_{ba} < \eta_\theta$$

For the conditioning of BA to be less than the threshold,

$$\begin{aligned} &\delta_{ba} < \delta_\theta \\ &\Rightarrow \frac{1-\delta_\theta}{1-\delta_{ab}} < \frac{\Pr(A)}{\Pr(B)} \end{aligned}$$

For the crosstalk of BA to be less than the threshold,

$$\begin{aligned} &\eta_{ba} < \eta_\theta \\ &\Rightarrow \frac{1-\eta_\theta}{1-\eta_{ab}} < \frac{\Pr(A')}{\Pr(B')} \end{aligned}$$

When both the above conditions are true, reverse edge BA, i.e., gene B regulating gene A is thought to be a possible regulation and included in the graph. However we find that $\delta_{ab} < \delta_{ba} \Rightarrow \eta_{ba} < \eta_{ab}$ and $\eta_{ab} < \eta_{ba} \Rightarrow \delta_{ba} < \delta_{ab}$. Thus we use a third gene C seemingly regulated by both A and B to determine precedence of the drivers in algorithm as outlined in Algorithm 4, and confirm if the reverse edge needs to be pruned or not, with complexity of $O(n)$.

Input: Ordered Vertices A, B, C, Edges AB, BA, BC, AC
Output: List of Retained Edges
removeEdgeBA $\leftarrow$ true;
if *{* $(\eta_{ba} < \eta_{ab})$ *&* $(\delta_{bc} \geq \delta_{ac})$*}* **then**
  removeEdgeBA $\leftarrow$ false;
end
if *(removeEdgeBA)* **then**
  Remove edge BA;
end

Algorithm 4: Reverse edge removal algorithm

Pruning Order

Given the context-specific GRN H, the edge removal algorithm would first select the vertices present as an ordered triad in the graph as depicted in Figure 5.1. Next, it would check the edges in turn for the satisfiability of the edge retention criterion. The edge retention criteria query relevant edges as reverse edge, sibling edge or transitive edge. However, the order of the checks influence the final structure of the graph. We propose a topology based edge pruning to determine the order of edge checks. It has been shown in different studies that biological networks such as metabolic networks display scale-free topology and small-world topology than random topology [66], [67]. In order to conserve the characteristic topology of the GRN, we would use corresponding topology based edge pruning. Here we present the random network and scale-free network based edge pruning. In Section 5.3 we compare these to pruning orders of sibling edge checks followed by transitive edge checks (Sibling-Transitive) and transitive edge checks followed by sibling edge checks (Transitive-Sibling), completed by reverse edge checks.

Random Topology Based Pruning

For the random topology based edge pruning, after the selection of each triad such as shown in Figure 5.1, we let the order of edges and edge checks be random. The algorithm is outlined in Algorithm 5. If we denote number of genes and number of samples as m and n respectively and e as number of edges in the context motif graph then the complexity of Algorithm 5 is $O(mne)$.

Scale-free Topology Based Pruning

For the scale-free topology based edge pruning, the aim is the conserve the hubs in the networks. Thus the node with higher outgoing edge would be taken as the driver vertex v_1. In this method, the first check removes extraneous edges of hubs, i.e., transitive edge checking and removal reduces the outdegree of vertex v_1. These

Input: Context motif graph $H = (V, E)$
Output: Pruned context motif graph $H' = (V', E')$

forall the $v_1 \in V$ **do**
  $driven(v_1) \leftarrow$ all genes driven by v_1 ;
  forall the *(*$(v_2 \in driven(v_1))$ *&* $(v_3 \in driven(v_1) \cap driven(v_2))$ *&* $(v_3 \neq v_1, v_2)$*)* **do**
    Randomly select ordering of checks below;
    if *checkTransitiveEdgeRemoval(*v_1, v_2, v_3*);*
    then
      continue;
    end
    if *checkSiblingEdgeRemoval(*v_1, v_2, v_3*);*
    then
      continue;
    end
    if *checkReverseEdgeRemoval(*v_1, v_2, v_3*);*
    then
      continue;
    end
  end
end
$H' = (V' \leftarrow$ All retained vertices, $E' \leftarrow$ All retained edges);

Algorithm 5: Random topology based pruning algorithm

checks are followed by sibling and reverse edge checks. The algorithm is outlined in Algorithm 6. If we denote number of genes and number of samples as m and n respectively and e as number of edges in the context motif graph then the complexity of Algorithm 6 is $O(mne)$.

Comparison with Other Graph Pruning Methods

Traditional methods such as transitive reduction and Data Processing Inequality (DPI) criteria cannot be directly applied to the context-specific GRN. For example, transitive reduction algorithm application requires the input graph to be a directed acyclic graph. However, the context-specific GRN has no such restriction on its structure, thus possibly contains cycles. For the sake of comparison a special case of the transitive reduction has been applied. The algorithm checks end points of each edge, to determine if there is any two-step path between the end points. If

Input: Context motif graph $H = (V, E), V$ ordered by descending outdegree
Output: Pruned context motif graph $H' = (V', E')$

```
1  forall the v_1 ∈ V do
2      driven(v_1) ← all genes driven by v_1;
3      forall the ((v_2 ∈ driven(v_1)) & (v_3 ∈ driven(v_1) ∩ driven(v_2)) &
       (v_3 ≠ v_1, v_2)) do
4          if checkTransitiveEdgeRemoval(v_1, v_2, v_3);
5          then
6              continue;
7          end
8          if checkSiblingEdgeRemoval(v_1, v_2, v_3);
9          then
10             continue;
11         end
12         if checkReverseEdgeRemoval(v_1, v_2, v_3);
13         then
14             continue;
15         end
16     end
17 end
18 H' = (V' ← All retained vertices, E' ← All retained edges);
```

Algorithm 6: Scale-free topology based pruning algorithm

such a path exists, the edge is removed, else the edge is retained in the graph. The idea is that the transitive closure of original GRN and of this reduced graph would be the same.

ARACNE [39] uses DPI, that is, the algorithm chooses connected triplets of genes and removes the edge with minimum value of Mutual Information [26]. In the context-specific GRN, as the nodes represent genes in particular gene expression level, and not just the genes themselves, each edge corresponds to a different set of samples. In order to calculate the Mutual Information, the values of joint probabilities and marginal probabilities would be calculated on a subset of samples and not all the samples. Therefore it becomes a challenging problem to determine the correct ordering of Mutual Information values calculated across different sets of samples. Thus this approach was not used for the comparison here. Note, as the triad of vertices considered in Expattern share a common set of samples (by

virtue of chaining of context motifs for graph formation), we are able to use the consistency metrics conditioning and crosstalk defined on that set of samples for our estimations of expected values of the metrics.

Inter and Intra Context Edges

We apply the MCL algorithm [46, 65] to the reduced context motif networks in order to study modular structure within the networks. Once we identify the contexts, we can observe the outcome of the different pruning methods on the graphs. We want to confirm the hypothesis that there is a higher chance of observing extraneous edges as statistical artifacts in highly connected portions of the graph. So removal of the extraneous edges would not only make the graph less dense, but it might allow the emergence of smaller cliques within the larger contexts indistinguishable earlier. By tagging the edges as inter context or intra context, we evaluate the number and type of edges retained after the application of different pruning methods.

5.4 Results

Application to Artificial Contextual Network Datasets

We observed in the last chapter, Section 4.4, that ARACNE does poorly in case of multiple context data. Thus we did not consider the undirected edges comparison here but focused only on comparison of Expattern with (directed) Bayesian network learning method BANJO. BANJO reports the consensus network and the highest scoring network at the end of each run. As before, to maintain consistency, we applied BANJO and Expattern to the same discretized datasets. BANJO was specified to run using simulated annealing with 50000 restarts and 1000 minimum networks before checking. We present both the highest scoring network found by BANJO and the BANJO consensus network. As earlier, we noticed that even though the number of true positive edges found in (unpruned GRN) Expattern was higher than found by BANJO, Expattern also identified a very high number of false positive edges. Here, we present the results of applying the different pruning methods to the Expattern

GRN.

Figure 5.4 compares the performance of BANJO, BANJO Consensus, Expattern without pruning, Expattern with pruning – sibling transitive (ST), random topology (RN), scale free (SF), transitive sibling (TS) and transitive reduction(TR) as described in the methods section. Figure 5.4 displays the improvement in precision on applying the different pruning methods. There is some decrease in recall values but as shown in Table 5.1, the change in f-measure is not significant.

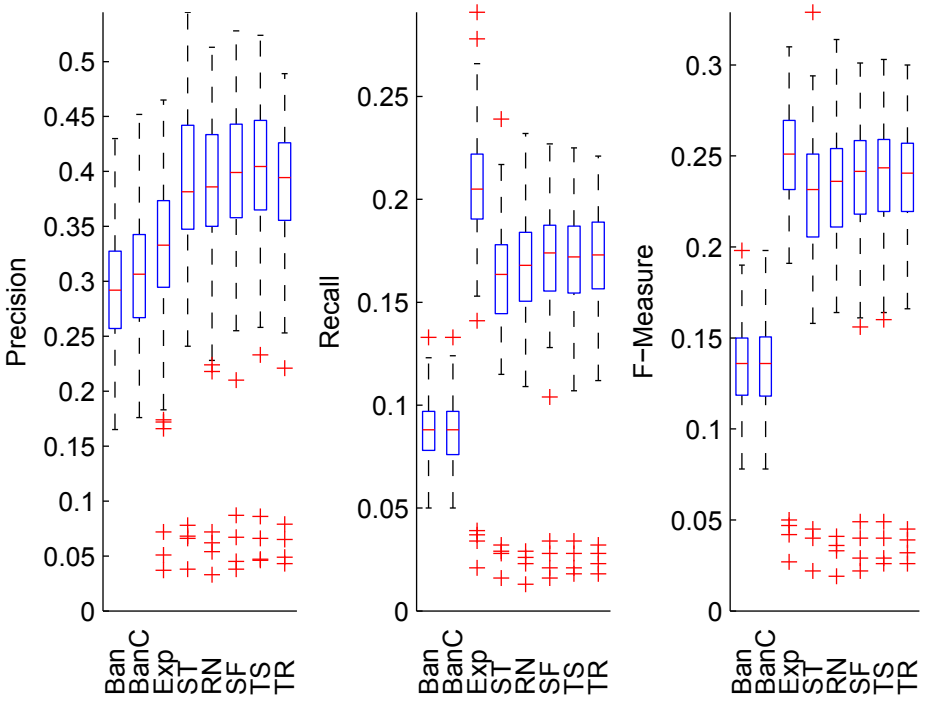


Figure 5.2: Precision, recall and f-measure comparison for directed edges of two context artificial networks. The results depicted are averaged over 100 pairs of randomly combined artificial context networks. The different methods presented here – Ban (BANJO), BanC (BANJO Consensus), Exp(unpruned), ST (Sibling-Transitive), RN (Random Topology), SF (Scale Free), TS (Transitive-Sibling) and TR (Transitive Reduction).

The comparison between scale-free pruning versus sibling-transitive or transitive-sibling shows that ordering of the vertices in each considered triad of genes gives

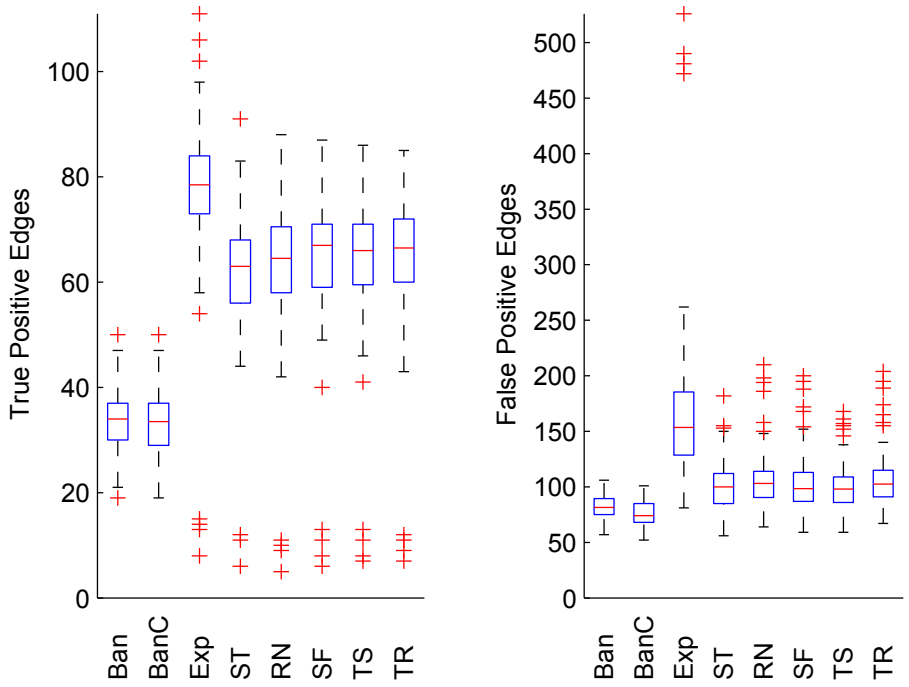


Figure 5.3: True positive and false positive edge comparison for directed edges of two context networks. The results depicted are averaged over 100 pairs of randomly combined context networks. The different methods presented here – Ban (BANJO), BanC (BANJO Consensus), Exp(unpruned), ST (Sibling-Transitive), RN (Random Topology), SF (Scale Free), TS (Transitive-Sibling) and TR (Transitive Reduction).

rise to differences in pruning results. In scale free approach, we identify hub genes, and verify that the out-degree of the hub vertex is not due to extraneous edges and finally prune extra edges in the neighborhood of the hub gene. In the random topology network approach, we randomly order removal of edges. Scale-free pruning aims to retain the connectivity of hub genes before checking for true sibling or transitive edges (Figure 5.1).

Application to Refractory Cancer Dataset

Here, we present the findings on applying the algorithms to a refractory cancer data based on Target Now study. The Target Now gene expression profiling experiments were conducted using the Agilent 011521 Human 1A Microarray G4110A platform

Table 5.1: Paired t-test p-values: Comparison of methods with Expattern. Precision, recall and f-measure p-value comparison of BANJO, BANJO consensus and different pruning methods with Expattern results.

Method vs Expattern	Precision p-value	Recall p-value	F-measure p-value
BANJO	0.0020	1.92E-60	7.53E-49
BANJO Consensus	0.0825	4.49E-61	5.99E-49
ST	1.56E-05	1.50E-11	0.0039
RN	3.03E-05	1.45E-09	0.0183
SF	3.89E-07	9.79E-08	0.1529
TS	3.05E-08	2.75E-08	0.1496
TR	3.67E-06	1.35E-07	0.1152

and consists of 146 patients, spanning 35 different types of tumor. The dataset was filtered based on the transcription activity of each gene across samples, and reduced to 3,851 genes by eliminating genes with a low variance across samples. The context motif mining algorithm was applied to the reduced dataset, to extract context motifs with 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 of 0.01 were considered. Pruning was applied to the dataset and Table 5.2 shows the resulting number of contexts obtained in each case. Of these, we eliminated contexts with fewer than 10 genes and fewer than 10 samples and studied the biological enrichment of these contexts. Single sample tumors were omitted from the phenotypic enrichment analyses and a statistical significance threshold of 0.05 was used for all functional enrichment.

For TN Dataset, Table 5.2 compares the functional enrichment of contexts after MCL clustering of reduced context-specific TN dataset graphs. We observe in Table 5.2 that different methods of pruning let different enriched contexts emerge from the unpruned network. Interestingly, scale-free, transitive reduction and transitive-sibling methods find enriched breast cancer contexts whereas sibling-transitive and random-topology methods retained the enrichments found by the unpruned net-

Table 5.2: Number of contexts found by Markov Clustering at inflation of 1.4 on the unpruned and pruned networks with enriched tissue types. Exp (Unpruned network), SF (Scale Free), TS (Transitive-Sibling), ST (Sibling-Transitive), TR (Transitive Reduction) and RN (Random Topology).

Pruning Strategy	Number of Contexts	Enriched Tissue
Exp	59	RCT,OV
ST	41	RCT,OV
RN	42	RCT,OV
SF	46	BR,OV
TR	72	BR,RCT,OV,PANC
TS	46	BR,OV

work. We observed the pancreatic cancer enrichment found by transitive reduction was associated with one of the smaller contexts with low number of genes and associated samples. We believe that the very high number of contexts found by transitive reduction causes more fragmentation of the contexts and allow us a view of a finer granularity of contexts. Whereas methods like scale free and sibling transitive retain the overall structured of the unpruned network and allow a global overview. If the question posed is to identify as many as possible subtypes of diseases a finer granularity may be desired but if the question aims to broadly classify say patients or disease type we can use a coarser granularity method such as scale-free.

Table 5.3: Inter and intra context edges as retained (removed) by different methods. Exp (Unpruned network), SF (Scale Free), TS (Transitive-Sibling), ST (Sibling-Transitive), TR (Transitive Reduction) and RN (Random Topology).

Method	Total edges	Inter edges	Intra edges	Total Removed	% Removed Inter edges	% Removed Intra edges
Exp	19948	321	19627	-	-	-
ST	9033	222	8811	10915	30.84	55.10
RN	8808	216	8592	11140	32.71	56.22
SF	8493	224	8269	11455	30.21	57.87
TR	8767	236	8531	11181	26.47	56.53
TS	8393	221	8172	11555	31.15	58.36

Table 5.3 displays the inter and intra context edges with respect to contexts found by the unpruned network. It show the retained and removed inter context and intra context edges as identified through Markov clustering of the unpruned

network. Interestingly, although the number of removed edges are comparable, the emergent graph structures are different (as presented in Table 5.2).

Application to Glioma Cancer Dataset

Here, we present the findings on applying the algorithms to a Glioma cancer data based on The Cancer Genome Atlas (TCGA)[6] study. TCGA is a joint effort of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), two of the 27 Institutes and Centers of the National Institutes of Health, U.S. Department of Health and Human Services. In this study, we focus on Glioblastoma multiforme (GBM) which is the most common and most aggressive malignant primary brain tumor in humans.

Total 301 samples from GBM gene expression data (from TCGA) were used after screening out samples from cell lines and replicates. 10 normal samples were used for 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. All z-score values in GBM samples were quantized to one of three discrete values - '1' for over-expression, '0' for no-change and '-1' for under-expression compared to the normal case. Genes with low variance were excluded from the analysis and 13,822 genes were analyzed in this work.

The context motif mining algorithm was applied to this dataset, to extract context motifs with 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 of 0.05 were considered. Pruning was applied to the dataset and Table 5.2 shows the resulting number of contexts obtained in each case. Of these, we eliminated contexts with fewer than 10 genes and fewer than 15 samples and studied the biological enrichment of these contexts. Single sam-

[6]http://cancergenome.nih.gov/abouttcga

ple tumors were omitted from the phenotypic enrichment analyses and a statistical significance threshold of 0.05 was used for all functional enrichment.

Table 5.4: Inter and intra context edges as retained (removed) by different methods. Exp (Unpruned network), SF (Scale Free), TS (Transitive-Sibling), ST (Sibling-Transitive), TR (Transitive Reduction) and RN (Random Topology).

Method	Total edges	Inter edges	Intra edges	Total Removed	% Removed Inter edges	% Removed Intra edges
Exp	227576	8963	218613			
TR	15605	215	15390	211971	97.60	92.96
TS	11743	144	11599	215833	98.39	94.69
ST	11403	142	11261	216173	98.41	94.85
SF	11882	144	11738	215694	98.39	94.63
RN	10814	141	10673	216762	98.43	95.12

Table 5.4 displays the inter and intra context edges with respect to contexts found by the unpruned network. It show the retained and removed inter context and intra context edges as identified through Markov clustering of the unpruned network. Interestingly, although the number of removed edges are comparable, the emergent graph structures are different (as presented in Table 5.5).

Table 5.5: Number of contexts found by Markov Clustering at inflation of 1.4 on the unpruned and pruned networks with enriched tissue types. Exp (Unpruned network), SF (Scale Free), TS (Transitive-Sibling), ST (Sibling-Transitive), TR (Transitive Reduction) and RN (Random Topology). Contexts were filtered at p-value 0.05 with number of genes > 10 and number of samples > 15.

Method	Filtered Contexts	Enriched Contexts	Proneural	Neural	Classical	Mesenchymal
Exp	97	7	0	4	0	3
TR	55	16	7	2	2	5
TS	22	4	0	2	0	2
ST	21	3	0	1	0	2
SF	22	4	0	2	0	2
RN	21	3	0	1	0	2

Table 5.5 shows the enrichment of identified contexts with subtypes of GBM. Transitive Reduction method creates more fragmented contexts than the other methods which output comparable number of contexts. We show some enriched path-

ways in contexts found by Scale Free pruning in Table 5.6. We observe that Mesenchymal subtype enriched context are associated with mesoderm development, signaling pathways and leukocyte regulation. Also the Neural subtype enriched contexts were associated to pathways related to neuroblastoma.

Table 5.6: Enriched pathways identified in Scale Free Contexts. The enriched subtypes are displayed within parentheses.

SF Contexts	Pathway Enrichment
Context 4,7 (Mesenchymal)	Mesoderm_development, Sig_PIP3_signaling_in_B_lymphocytes, Sig_BCR_signaling_pathway, Positive_regulation_of_secretion, Pattern_recognition_receptor_activity, Leukocyte_chemotaxis, GA12_pathway, B_cell_antigen_receptor,Leukocyte_migration, Wong_endometrial_cancer_late, Defense_response_to_bacterium
Context 6,18 (Neural)	White_neuroblastoma_with_1p36.3_deletion, Chr19q12

5.5 Summary

Here we present a unique approach to graph pruning of context-specific gene regulatory networks based on consistency metrics used in the learning of the networks. We also implemented a variant of the popularly used graph reduction method – transitive reduction pruning, adapted for use in context-specific GRNs instead of only on directed acyclic graphs. We show the proposed context-specific graph pruning strategies reduce the number of extraneous edges and allow emergence of the functional enrichment of the context specific GRNs. We introduced different strategies to determine the order of edge pruning. We applied these methods to TN and GBM cancer datasets and ran obtained reduced networks through Markov clustering. Subsequently, we calculated enrichment within the set of clusters and interpreted the obtained results. The simple variant of transitive reduction that was applied removed maximum edges and produced maximum number of clusters. We observe that the best pruning strategy to be employed depends on the dataset and the question posed, e.g., the level of desired granularity and functional enrichment.

Chapter 6

INTEGRATING MULTI SOURCE DATA

We extend the concept of finding conditioning factors (regulating elements) from only genes, to elements which influence, regulate or act specific to the existing cellular state. Any such factor would also be bound by the constraints in place due to cellular contextual state. Applying our method to such disparate datasets such as aCGH, gene expression and/or drug activity data, would allow us to witness the possible underlying patterns of the inter- and intra-relationships between the different datatypes.

6.1 Motivation

Gene expression profiling experiments are a popular research and screening tool for differentially expressed genes. The experiments are designed to simultaneously measure the expression of thousands of genes. This high throughput biomedical data is stored and readily available in public repositories. Thus, initially we applied our framework to gene expression data. However, it is known that gene expression data suffers from issues of reproducibility and reliability [68], and small sample size [69]. Some of the issues are partially mitigated by careful design of experiment [68] and application of appropriate statistical methods [69]. However, in order to increase the confidence in obtained results from our framework, we propose the use of multiple types of biomedical data to yield multi-type interactions, for instance, gene-drug, gene-phenotype and gene-environment interactions.

The multiple types of biomedical data originate from different platforms which collect data at different stratas of abstraction according to specific purposes, e.g., aCGH at genome sequence level for DNA copy number changes; and microarray data at the mRNA level for gene expression values. We use the framework to integrate data extracted from different types of regulatory factors, originating from

different experimental measurements, for example, CGH (copy number changes), microRNA (miRNA) expression, and clinical data such as survival or metabolite levels.

6.2 Related Work

Comparable GRN reverse engineering approaches which integrate multiple sources of data, such as by using Bayesian network [70] and [71] treat all data types equally in the mathematical framework to generate corresponding interactions. Methods such as COALESCE [37] and CONEXIC [36] distinguish between gene expression and DNA Sequence Data but restrict the use of amplified or deleted matched DNA sequence to gene probes for finding regulatory interactions (section 4.2).

Intuitively, different data types (corresponding to the disparate sources of data) measure different components of the biological system, providing unique perspectives to the cellular system. Understandably, each pair of data type would interact uniquely. Thus, considering all data types with equal weights or inference parameter thresholds cannot be biologically meaningful. In our framework, we allow users to include apriori knowledge as meta-rules to guide the search of context motifs and thereby contexts. For example, we observe that interactions between entities of the same data types has a stronger correlation (low crosstalk and conditioning) than interactions between entities of different types. Thus any method to integrate data from different biological domains for learning, needs to taking into account this difference. Also, recent work by Li et al. [45] found the RNA sequences do not correspond exactly to the DNA sequences. These nonrandom differences were found in multiple individuals and in different cell types. Thus methods such as COALESCE and CONEXIC which assume a one to one DNA to RNA sequence correspondence may miss important indirect DNA to RNA influences under different cellular phenotypes. Our algorithm, Expattern does not assume pre-existing correspondence and thus does not restrict the relationships between DNA sequences

and RNA sequences. This allows us to capture indirect relationships not identified through CONEXIC and COALESCE.

6.3 Methodology

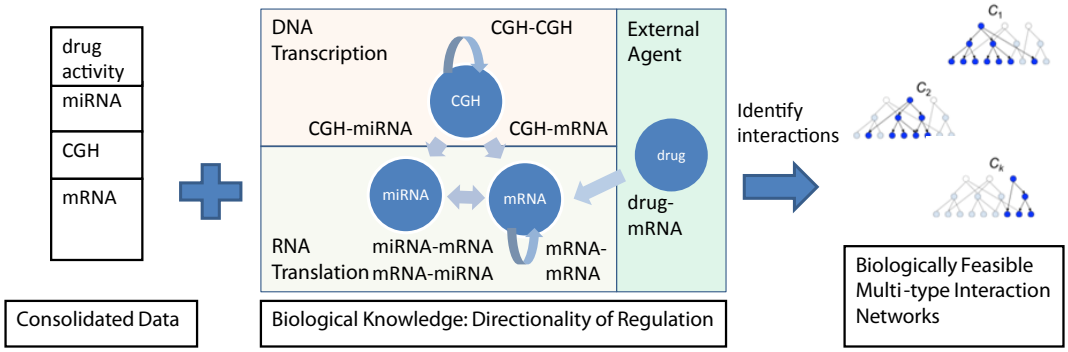


Figure 6.1: Directionality of edges in multi-domain conditioning.

Directionality of Regulatory Interactions

The central dogma in biology highlights the direction of information flow from DNA to RNA to protein. Understandably, the DNA sequence affects its transcription to RNA, linking copy number changes found in aCGH data to the amount of transcribed RNA observed in gene expression data. However, this influence direction is unidirectional, i.e., the amount of RNA cannot dictate or modify the structure of the DNA sequence or the copy number changes. One can also understand the mRNA-mRNA interaction where the transcribed mRNA of say transcription factors triggers the production of a target mRNA to yield a final protein product. Similarly, we can extrapolate the aCGH-mRNA, aCGH-miRNA, miRNA-mRNA and mRNA-miRNA interactions. Exploiting the directionality of regulatory influences across different data source stratas such as DNA to RNA to protein, we can identify pertinent interactions that are biologically meaningful instead of relying solely on statistical significance. Figure 6.1 depicts a directionality of regulatory influences between aCGH, gene expression and drug activity data. The shaded rectangles in the figure represent the different data domains or strata with respect to the cellular system -

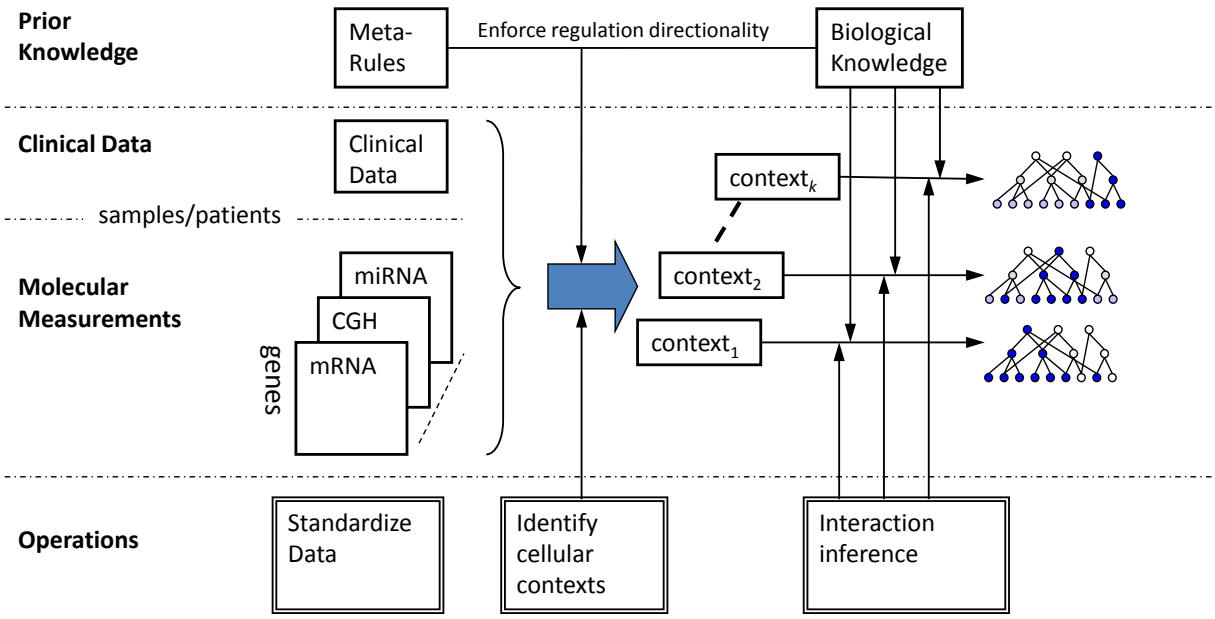


Figure 6.2: Integrated approach to learn context-specific interaction networks via multi-domain conditioning.

DNA transcription, RNA Translation and External Agent/Influence. We incorporate this directionality as apriori knowledge in the algorithm to guide the identification of context motifs satisfying the biological constraints.

Incorporating apriori Knowledge in Expattern

Figure 6.2 provides an outline of the integrated approach to learning context-specific interaction networks via multi-domain conditioning. We preprocess and quantize each dataset separately. Each input file corresponds to a different data type, and is read in separately, associating the data type (for example: mRNA, aCGH) with each dataset. In order to integrate the datasets, the current version of the algorithm assumes all datasets correspond to the same sample set. To guide the multi-type context-motif identification meta-rules are incorporated in the algorithm. The meta-rules specify the combinations of data type interactions to be identified and optionally the required crosstalk, conditioning and statistical strength thresholds for the interactions, both within and across different data types.

Meta-rules capture user intuitions about possible regulatory signal/influence flow between domains. For example, we use a meta-rule to identify context motifs which have aCGH driver and mRNA passengers. aCGH data measures copy number changes of the DNA sequence and pinpoints aberrations (amplification/deletions); mRNA data measures the gene expression levels (overexpression/underexpression). Through the central dogma in biology, we know the flow of information happens from DNA to RNA. It follows that amplified regions of the DNA containing a gene could cause an increase its corresponding gene expression making it over expressed. To test this we now have the option in the framework to use the meta-rule where aCGH drives mRNA expression. Similarly, we can test whether miRNA presence could point to underexpression of mRNA by silencing the corresponding region containing the gene. In summary, the meta rule contains information as displayed below:-

$$MetaRule^{(i)} = \langle Driver_{DT}^{(i)} \rangle \langle Driven_{DT}^{(i)} \rangle \langle \delta_{Th}^{(i)} \rangle \langle \eta_{Th}^{(i)} \rangle \langle \rho_{Th}^{(i)} \rangle.$$

where, $\langle Driver_{DT}^{(i)} \rangle$ is the driver entity's data type, $\langle Driven_{DT}^{(i)} \rangle$ is the Driven entity's domain type, and $\langle \delta_{Th}^{(i)} \rangle$, $\langle \eta_{Th}^{(i)} \rangle$, $\langle \rho_{Th}^{(i)} \rangle$ are the optional conditioning threshold, crosstalk threshold, and statistical significance threshold of meta-rule i respectively.

The algorithm for identification of multi-type interactions is outlined in Algorithm 7. Each interaction is tested based on the user specified meta-rules. If the multi-type interaction satisfies any of the meta-rules, it is saved, else discarded. Using the saved interactions we build the context motifs and consequently obtain the contexts. Figure 6.2 depicts the process flow for integrating multi data type and learning context-specific interaction networks using the meta-rules.

We use m and n to denote the total number of genes and samples in data

Input: Genes G, Samples T, Dataset $D = G \times T$, Meta-Rules Array $mArr$
Output: List of Multi-Type Context Motifs $MTCMList$

$MTCMList \leftarrow$ null;
for *Gene or clinical parameter* g_i *in state* y_i **do**
  $allRules \leftarrow$ mArr.getRulesWithDriverType(g_i.getType());
  if *($allRules$==null)* **then**
    break;
  end
  $T_i^{y_i} \leftarrow$ Samples where gene g_i is in state y_i;
  /* $Driven_i^{y_i}$ = Genes regulated by g_i in state y_i */
  $Driven_i^{y_i} \leftarrow$ null;
  forall the *Genes or clinical parameter* $g_j, g_j \neq g_i$ **do**
    $mRule \leftarrow$ allRules.getRuleWithPassengerType(g_j.getType());
    if *($mRule$==null)* **then**
      break;
    end
    $\eta_{ij} \leftarrow$ Crosstalk of g_j regulated by g_i in $T_i^{y_i}$;
    $\delta_{ij} \leftarrow$ Conditioning of g_j regulated by g_i in $T_i^{y_i}$;
    if *($(\eta_{ij} < mRule.\eta_\theta$ in $T_i^{y_i}$) AND ($\delta_{ij} < mRule.\delta_\theta$ in $T_i^{y_i}$))* **then**
      Add g_j to $Driven_i^{y_i}$;
    end
  end
  if $Context\_Motif = \{g_i, y_i, Driven_i^{y_i}, T_i^{y_i}\}$ *is statistically significant* **then**
    Add $Context\_Motif$ to $MTCMList$;
  end
end

Algorithm 7: Identification of multi-type context motifs algorithm

set, r as the number of meta-rules, and k as the user specified number of iterations for bootstrap sampling to calculate statistical significance. Then the complexity to identify a single multi-type context motif is $O(n^3mr)$. The complexity to identify multi-type context motifs for all driver genes is $O(n^3m^2r)$. Bootstrap resampling to calculate the statistical significance is $O(n^4mk)$. Thus the complexity of Algorithm 7 is $O(n^4mk + n^3m^2r)$. The requirement of apriori knowledge to set the parameters may be perceived as a limitation, but this option in the framework provides an exploratory tool for hypotheses formulation about regulations between different (inter) or same (intra) data types. Once the user sets a criterion for measuring the context quality, the next step would have the software identify optimal multitype contexts.

For example, using a criteria such as a biological score or a cluster validity measure, the framework execution can be automated to find inter-datatype parameter thresholds.

6.4 Results

Cancer Cell Line Data

We apply the method for context motif identification to NCI60 gene expression-drug activity dataset [11]. We show that the identified context motifs can further guide studies of drug effectiveness and mechanism of action. Here we illustrate how multiple types of data, for instance, gene expression and drug activity data, can be combined to identify interesting patterns of interactions not only among genes but, for instance, between genes and drugs. To provide an example of exploratory functionality possible by context motif mining method of Expattern we applied it to the NCI60 drug data. The NCI60 is a set of human cancer cell lines derived from diverse tissues; brain, blood and bone marrow, breast, colon, kidney, lung, ovary, prostate and skin. The dataset consisted of the drug activity data of 118 drugs and the gene expression data of 1375 genes across the NCI60 cell lines [11]. Drug activity is represented in a matrix with $\log$ GI50 values, where GI50 is an indicator of the growth inhibition of the compound on the cell line. The original paper [11] related this data to sensitivity to therapy rather than to molecular consequences of the therapy, as the gene expression patterns were determined in untreated cells.

We scaled the different datasets to comparable form (normalization), combined these forms, and applied the method to obtain context motifs corresponding to the different conditioning factors. The data matrices were normalized by subtracting their row-wise mean and dividing by their row-wise standard deviation. Next, matrix entries were quantized on the basis of two-fold changes, for statistical significance. Then all quantized matrices were used as the input data for the context analysis.

The context motif analysis on the NCI60 drug activity and gene expression

data resulted in 4153 context motifs. Among those, we focused on the context motifs where at least one drug and one gene were included, which resulted in 243 context motifs. On filtering, only 27 context motifs were found to be statistically significant with p-value less than 0.01.

We observed that the majority of the context motifs reflected patterns found in the original paper [11]. For example, the two breast cancer cell lines positive for oestrogen receptor, T-47D and MCF7, clustered together in the original paper, were also found to be grouped together in our analysis. The context motif identified showed higher activity of drug 11-formyl Camptothecin (RS) than its counterpart Camptothecin, 11-HOMe(RS).

For the two cell lines (MDA-MB-435 and MDA-N), there were two filtered context motifs of interest. In the first context motif with only these two cell lines grouped, drug 7-Epi-10-deacetylbaccatin III (Taxol Analog NSC No. 656178), Paclitaxel and other Taxol analog drugs with the mechanism of action as Tubulin-active antimitotic agents (TU) displayed highly active status. In the second context motif, conditioned by gene RAB7, these two cell lines were grouped together with Melanoma cell lines (MALME-3M, SK-MEL-5 and UACC-62).

Interestingly the drugs identified in this context motif as being consistent were Cyclocytidine and Cyctarabine(araC), belonging to DNA synthesis inhibitor mechanism (Ds). However, they did not display high activity across all these samples, while Taxol analog drugs were highly active in these two breast cancer cell lines. In the original paper, MDA-MB-435 and MDA-N cell lines clustered closely with Melanoma cell lines [11]. The authors discussed that the MDA-MB-435 and its Erb/B2 transfectant MDA-N expressed large number of genes characteristic of melanoma, and the recent findings now group these two as a subtype of Melanoma itself [72, 73, 74]. However, the finding in our study may indicate they still do not use the same mechanisms in drug responses.

Table 6.1: Top 27 context motifs identified from combined drug data and gene expression data, with statistical significance 0.01. The first column represents the conditioning factors (gene/drug/disease) of the context motif. Genes are represented by gene symbols. In case of drugs, the drug name is shown with the mechanism of action e.g. [TU]. The second, third and fourth columns lists the number of conditioning factors, number of cell lines and total number of drug/gene elements respectively identified for the context motif. The fifth column reports the p-value of finding such a context motif. The final column reports the number of drugs that were found to be highly active in that context motif.

Type	SW	S	G	$Pr(G+\|S)$	Drug
Gene					
PTK2	2	4	184	0.00163	33
RAB7	1	5	132	0.00163	0
GJA4	2	3	241	0.00169	39
HEXB	1	2	200	0.00172	37
MMP14	1	3	173	0.00253	5
TWF1	1	5	102	0.00327	3
CORO1A	1	5	102	0.00327	5
TDG	1	3	164	0.00337	16
GLUL	1	3	145	0.00422	3
ISGF3G	4	4	159	0.00489	3
KCNQ4	2	5	93	0.00490	4
-	1	3	139	0.00506	9
-	1	2	174	0.00517	44
MYL3	2	4	145	0.00653	9
RP6-213H19.1	1	4	120	0.00734	3
MAPRE2	2	3	118	0.00759	3
KLF6	1	3	118	0.00759	12
-	1	4	117	0.00816	5
-	1	3	114	0.00843	5
IRX3	1	4	107	0.00897	2
REEP5	3	2	144	0.00948	6
-	1	4	100	0.00979	0
Drugs					
7-Epi-10-deacetylbaccatin III [TU]	1	2	190	0.00259	7
Camptothecin,20-ester (S) [T1]	1	42	7	0.00382	0
Camptothecin,11-HOMe (RS) [T1]	1	2	160	0.00603	42
Taxol analog [TU]	1	54	4	0.00771	0
Disease					
Leukemia	2	6	78	0.00485	2

Assignment of Drug Mechanism of Action: Many of the context motifs include drugs that have different mechanism of action. Every context motif depicts the common transcriptional activities of given cell lines, for example, subtypes of cancers with shared transcriptional behavior. It is possible that in order to stop proliferation of the cell, different points of the regulatory mechanisms present in cancer cells are targeted. Thus depending upon drug target point, varying degree of potency of drug would be established, effective in arresting the cancer development. Our initial purpose of being able to attribute the drug to a particular mechanism seemed thwarted by the inclusion of drug in multiple context motifs, showing more than one type of mechanisms active in each context. Considering the previous argument, we improved the prediction of mechanism of action of drug by finding maximum overlap between biological processes (Gene Ontology [32] terms) of the genes targeted by drug with unknown action and those of drugs with known action. Gene Ontology (GO) organizes genes into hierarchical categories based on biological process, molecular function and subcellular localization. Greater overlap between GO terms would imply similarity in mechanism of action.

We used this approach to assign the mechanism of action of drug Inosine-glycodialdehyde (Inox) by studying other drugs in all context motifs which included Inox. In the context motif conditioned by IRX3, Inox showed similar activity to 11-Formyl-20(RS)-Camptothecin, of mechanism T1, topoisomerase 1 inhibitor. In the context motif conditioned by gene TWF1, it showed high activity along with drugs Dichloroallyl-lawsone and Pyrazofurin of mechanism Rs, RNA synthesis inhibitor. This context motif consisted of Leukemia cell lines CCRF-CEM, K-562, MOLT-4, HL-60 and RPMI-8226.

We extracted for each drug the corresponding target genes from PubGene [75], and ran the obtained lists through GoMiner [76]. GoMiner is a program package that organizes lists of 'interesting' genes (for example, under- and overex-

pressed genes from a microarray experiment) for biological interpretation as GO terms. Instead of analyzing microarray results with a gene-by-gene approach, GoMiner classifies the genes into biologically coherent categories and assesses these categories.

On matching the significant GO terms (with p-value$<$0.05) we found that although there are less than 10 exact matches but the terms display coherency in terms of function to Rs mechanism derived GO terms. For Inosine-glycodialdehyde, we found GO terms which relate to negative regulation of transcription (from RNA polymerase II promoter and DNA- dependant, namely, GO:0000122, GO:0045892). GO terms matching those from Pyrazofurin and dichloroallyl-lawsone (Rs mechanism) related to nucleotide metabolism and biosynthesis (terms include GO:0006220, GO:0009058, GO:0009165 and GO:00-44249). There was no significant GO term match between those derived from Inox and those from Camptothecin.

Some context motifs group different cell lines possibly implying an underlying similarity in the regulatory mechanism in place, irrespective of the tissue of origin. This allows identification of drugs which could be potent in these particular cancer subtypes, allowing us to span and target a greater range of cancer types using the same drug. By finding targeted mechanisms by concentrating on annotations such as GO terms would allow greater power in ability to prescribe potent drugs. Here, we showed a need for context motif identification and possible application of this method to decipher the mechanism of action for drugs or propose alternative drugs for treatment according to cancer type and/or patient profile.

Application to Glioma Cancer Dataset

The Cancer Genome Atlas (TCGA)is a comprehensive and coordinated effort to accelerate our understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing.

TCGA is a joint effort of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), two of the 27 Institutes and Centers of the National Institutes of Health, U.S. Department of Health and Human Services. In this study, we focus on Glioblastoma multiforme (GBM) which is the most common and most aggressive malignant primary brain tumor in humans.

Total 301 samples from GBM gene expression data (from TCGA) were used after screening out samples from cell lines and replicates. 10 normal samples were used for 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. All z-score values in GBM samples were quantized to one of three discrete values - '1' for over-expression, '0' for no-change and '-1' for under-expression compared to the normal case. Genes with low variance were excluded from the analysis and 13,822 genes were analyzed in this work.

Next, matching aCGH samples (Agilent 244K) were obtained from TCGA portal after filtering to remove duplicate vials and whole genome amplified samples resulting in a total of 265 samples. Adjacent probes were collapsed into segments using a circular segmentation algorithm DNACopy[7]. The smoothed 244K probes were further discretized to ternary (-1=values less than -2, 1=values greater than 1, 0=all intermediary values) and compressed to 362 probes using CGHAnalysis[8].

We were interested in studying three types of interactions for the TCGA data (mRNA - mRNA, co-aberrations and DNA copy number changes-mRNA). We preprocessed and quantized each dataset separately as described above. The two datasets were read in separately and the domain types were associated with each dataset: in our case mRNA and aCGH. To guide the multi-type context-motif identification meta-rules were incorporated in the algorithm as explained in Section 6.3.

[7] www.bioconductor.org/packages/2.3/bioc/html/DNAcopy.html
[8] http://public.tgen.org/ckingsle/

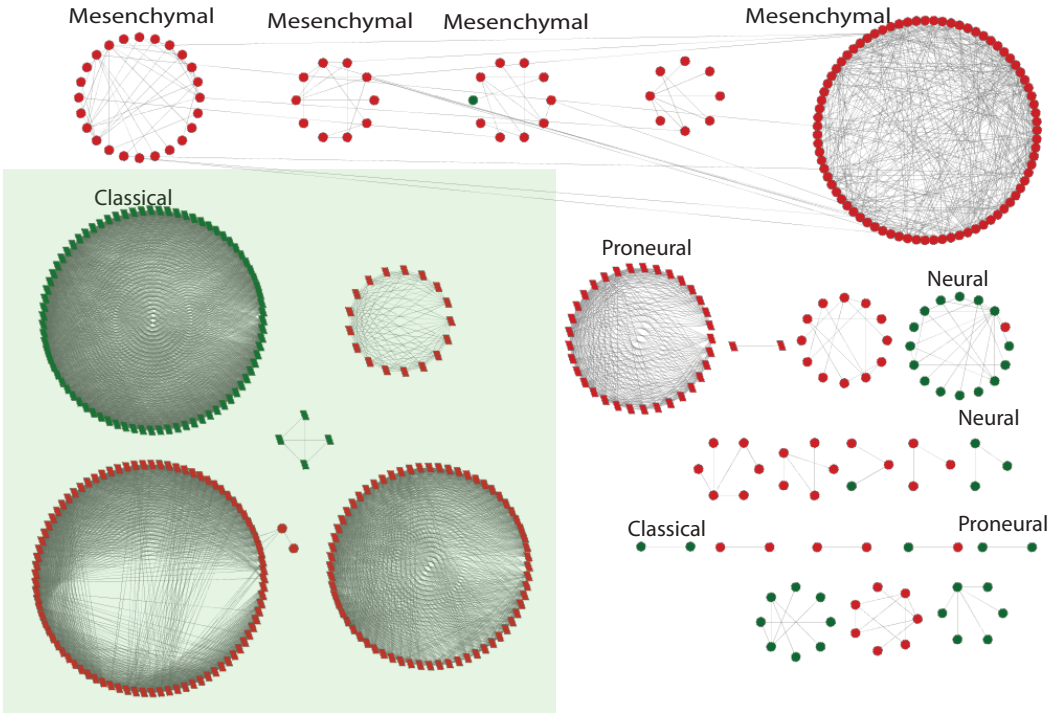


Figure 6.3: Contexts with enriched subtypes (on top) and mutations (on bottom). Contexts that have mRNA as a driver are shown as circles and contexts driven by aCGH are shown as parallelograms. Red contexts are up-regulated and green are down-regulated with respect to non-tumor samples. The contexts in shaded area are contexts with aCGH drivers and mRNA drivens.

The saved interactions were identified as corresponding multi-type context motifs and we obtained the multi-type contexts consequently.

For enrichment analysis we used the software Enrichment - Annotation Pipeline for Cellular Contexts (EPICC) to read context motifs from the result of context mining, which is done by a separate cluster computer program for its scalability, and applies several steps to identify and annotate contexts. This software can be easily extended to use results from multi-type context-mining.

Biological Interpretation

Quantized copy number data and expression data for 265 GBM samples was input to a parallel implementation of context-mining algorithm, Expattern, for multi-source

data. Rules were specified to find mRNA-mRNA, aCGH-mRNA and aCGH-aCGH interactions. A total of 39K context-motifs were obtained at a corrected p-value of 0.05 for genes or regions belonging to a context-motif. After filtering out context-motifs with p-value <0.05 (corrected), 3,567 context motifs were obtained out of which 3,529 were mRNA-mRNA, 167 aCGH-aCGH and 140 were aCGH-mRNA. Context-motif results were further analyzed using EPICC to form 84 contexts via graph clustering and annotated with subtype and mutation information. Figure 6.3 illustrates 26 selected contexts that have samples sizes between 5% and 75% of the total 265 samples. The figure also illustrates subtypes associated with these contexts.

A total of 11 contexts were found driven by aCGH from which 5 contexts contained mixed regulation of aCGH and mRNA data types. Such results would not have been easily identified by methods like CONEXIC and COALESCE when focusing only on drivers with associated amplified or deleted regions. Multi-type contexts are listed in Table 6.2, these contexts were regulated by aCGH and have aCGH /mRNA as drivens. The DNA to expression regulation of genes MTAP in Context 6, TSFM in Context 7, GSTT1 in Context 35 and SLC35E3 in Context 15 confirm that amplifications /deletions of these regions cause the overexpression/underexpression of these genes. These results also display other key players in the contexts which influence or are influenced by the downstream effects of the aCGH drivers.

Table 6.3 shows the comparison of context subtype enrichment when applied to single data type (gene expression data, described in last chapter) than when applied to multi-type data (gene expression and aCGH). Contexts were filtered at p-value 0.05 with number of genes > 10 and number of samples > 15. We observe a higher number of enriched contexts identified on using multi-type data. This is because context motifs identified through multi-type integration of data (here,

Table 6.2: Contexts in which aCGH regions regulate aCGH and mRNA. The entries in parenthesis (s;e) denote (number of samples; number of entities) identified in that context. Driver state is represented as D (Deleted) or A (Amplified). Subtype associated with the context is denoted by italicized text.

Context ID (s;e)	Drivers	Drivens	Driver State	Mutations
Context 6 (18;85) *Classical*	KIAA1797,IFNB1,IFNW1, IFNE1,DKFZp781D1719 IFNA8,FLJ42400,IFNA14, MTAP,CDKN2A,DMRTA1, FLJ35282,KLHL9,IFNA2	MTAP	D	CENTG1, DST,TNK2,
Context 7 (12;77)	KIAA1002,INHBE,DDIT3, ARHGAP9, MARS, GLI1, MBD6,DCTN2,TSPAN31, GEFT,FLJ39081,INHBC, CENTG1, CDK4, KIF5A, METTL1,FAM119B,OS9, PIP5K2C, DTX3, GEFT, FAM119B,XRCC6BP1, SLC26A10, B4GALNT1, TSFM, AVIL, CTDSP2, LOC283387,CYP27B1	39149, TSFM	A	DDIT3,FGFR1, MAG,OR5P2, PDGFRA
Context 4 (11;88)	LANCL2,ECOP,PSPH, ZNF713, FLJ44060	CHCHD2, LANCL2	A	
Context 15 (15;18)	NUP107,MDM2, SLC35E3	SLC35E3	A	
Context 35 (10;6)	FKSG58, RABIF, KLHL12, ADIPOR1, CYB5R1, TMEM183A, GSTT1	GSTT1	D	COL11A1,MAG, PDGFRA,RYR3, MYO3A,PLAG1

Table 6.3: Subtype enriched contexts identified using single type data (SType) versus multi-type (MType) GBM data.

Method	Filtered Contexts	Enriched Contexts	Proneural	Neural	Classical	Mesenchymal
SType Exp	97	7	0	4	0	3
MType Exp	94	20	8	3	3	6

using meta-rules) would be more specific to capture interactivity between different entity types, e.g., DNA copy number change influencing gene expression within specific subtype of GBM.

6.5 Summary

This chapter focuses on integrating multi-source data to learn context specific GRNs. Most approaches to integrate different types of data usually ignore apriori knowledge of the user and instead treat all data types equally, irrespective of the directionality of influence between the different sources of data. We provide an innovative way to incorporate information about different data domains as meta rules in the framework. These guide the learning of context motifs and thereby contexts. The flow of information is different between different data perspectives. We successfully applied the multi-source implementation to aCGH-mRNA data of TCGA for GBM and found interesting contexts with aCGH drivers. These might have been lost or not readily identified within the numerous purely aCGH or purely mRNA interactions generated by other comparable methods.

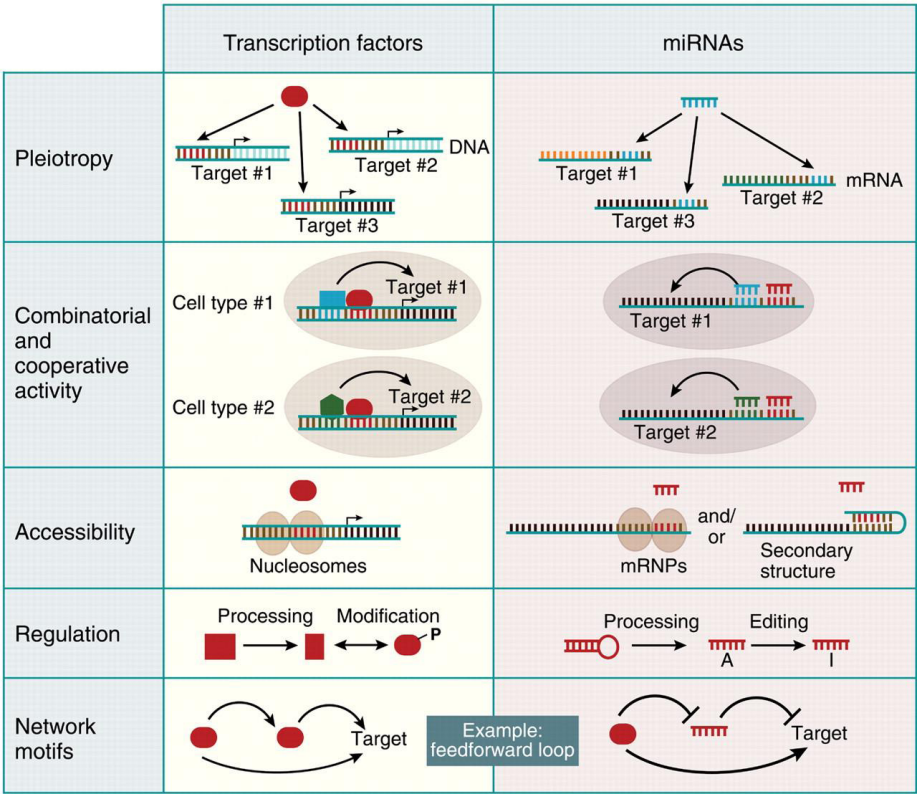
Chapter 7

IDENTIFYING MULTIVARIATE DRIVER CONTEXT MOTIFS

The framework, in previous chapters, used only one (gene) entity as the driver, i.e., applied univariate in-silico conditioning for identifying context motifs from the data. However, most processes in biology are known to be triggered by the concordant activity of multiple entities. For example, transcription factors need co-transcription factors to bind to the promoter region of the gene. To capture such concordant and combinatorial activity, there is a need to extend the univariate conditioning to study the combined conditioning effects of multiple entities, i.e., multivariate in-silico conditioning. The following sections describe some related work, our method of identifying multivariate drivers and application to Drosophila Melanogaster datset results.

7.1 Motivation

In eukaryotes, combinatorial and cooperative activity of regulatory factors such as transcription factors (TFs) and microRNAs (miRNA) regulate the gene expression [77]. For example, in humans, the regulation of more than 25000 genes is carried out by less than 2000 TFs. Eukaryotic TFs, individually considered, display only a modest degree of specificity and affinity in their interactions with ligands [78]. The specificity of transcriptional interactions are instead enforced by transciption complexes, created through cooperative multiple interactions. It is observed that transcription complexes have high specificity, even if the constituent interactions are of low specificity [78]. Similarly, cooperative action of miRNAs has also been defined through reporter gene assays [79]. Cooperativity therefore provides the mechanistic basis for reading out combinatorial expression patterns of both TFs and miRNAs as shown in Figure 7.1 [77]. In this chapter we focus on the combinatorial conditioning factors, specifically - Transcription Factors(TF).

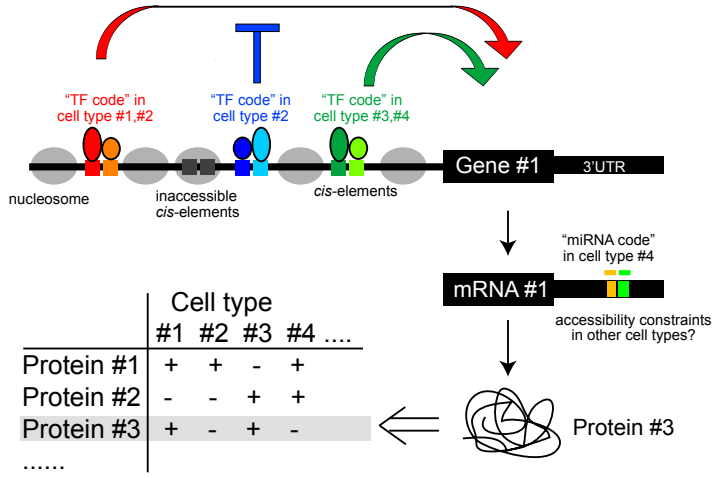


From O Hobert Science 2008;319:1785-1786. Reprinted with permission from AAAS.

Figure 7.1: A schematic visualization of some shared principles of TF and miRNA action [77].

Hobert et al. [77] discuss and compare gene regulation mechanisms by TFs and miRNAs (Figure 7.1). For this study, we are interested in combinatorial and cooperative activity of TF regulation (second row in Figure 7.1). Figure 7.2 (from the same paper [77]) depicts a cooperative scenario of joint regulation by TF and miRNA. Such a joint regulatory state would shape the expression profiles of individual cell types. Individual cell types and/or cellular states are defined by the expression of gene batteries and their encoded protein products. The specific composition of individual gene batteries would be controlled by the regulatory milieu of a cell, composed of specific combinations of TFs and miRNAs. In the example shown in Figure 7.2 [77], a combination of TFs can activate expression of a gene, Gene #1, in two cell types, #1 and #2, but a combinatorial code of repressors pre-

vents activation in cell type #2. Similarly, another transcription factor code activates transcription in cell type #3 and #4, but the miRNA code specific to cell type #4, restricts the product of gene #1 to be only expressed in #3. Understandably, through combinatorial action, a relatively limited set of trans-acting factors is able to define a vast number of distinct regulatory states [77].



From O Hobert Science 2008;319:1785-1786. Reprinted with permission from AAAS.

Figure 7.2: TF and miRNA codes to shape expression profile [77].

7.2 Related Work

Experimental studies are expensive and time consuming to understand the intricate and complex control of genes. Especially in case of the combinatorial analysis of regulatory factors, computational analysis becomes a very attractive alternative to identify regulatory factors acting in concert. Computational approach by Yu et al.[80] demonstrate that tissue-specific gene expression is generally regulated by more than a single transcription factor; where non-tissue specific TFs play a large role in regulation of tissue-specific genes. Furthermore, they show that individual TFs can contribute to tissue specificity in different tissues by interacting with distinct TF partners.

Growing interest in combinatorial activity of TFs is also evidenced by the

popularity of databases such as TRANSCompel® [81]. TRANSCompel® [81] focuses on composite elements (CE), the smallest units of combinatorial transcriptional regulation, characterizing the synergistic or antagonistic effects between the two transcription factors binding to two or more neighboring binding sites.

Over the years, different computational approaches have been proposed to identify combinatorial regulations. Among them, Sach Mukherjee et al. [82] used a statistical model for sparse, noisy Boolean functions and methods to identify combinatorial regulation. This was applied to a study of signaling proteins in cancer biology. More recent work has been done by Park et al. [83] on extracting combinatorial Boolean rules of synergistic gene sets from cancer microarray datasets. However, computational methods to reverse engineer gene regulatory networks commonly use a simplistic view of single regulatory factor and its target, even though (as outlined in Section 7.1) biology has many examples of several variables jointly influencing an output or response. In this chapter, we extend the framework to model and capture such consorted regulations which cannot be identified through univariate in-silico analysis.

7.3 Methodology

In the framework, each gene is first tested as a univariate conditioning factor. If true, the framework obtains the corresponding context motif. If not, i.e., the gene does not seem to be driving any other genes on its own, it is added it to the list of candidate multivariate drivers. As regulatory factors are known to have low specificity individually [78], thus coherent activity by corresponding drivens may not be displayed in data when evaluated via consistency metrics only (possibly not statistically significant). However, if we can combine the regulatory elements that constitute the regulatory complex, the combined effect would display the consistent regulation in its corresponding drivers.

Once we have the list of candidate multivariate drivers, we can select and

combine them, iteratively searching for i-conditioner sets of combinations among i different genes ($i = 2, \ldots, k$). We use different Boolean operators to obtain the i-conditioners as explained in the next section. As before, the algorithm will filter out combinations predicted to yield non-statistically significant conditioned sample sets before proceeding. This is necessary to ensure the algorithm restricts its output to significant results only, which will be subsequently considered for biological interpretation. Using the filtered list of i-conditioners, the algorithm will identify corresponding context motifs from the data. These steps will be repeated until a stopping criterion such as the number of i-conditioners or the number of identified context motifs is satisfied. Exhaustive search can be considered with the use of the parallel processing version of the implementation if needed.

Input: Genes G, Samples T, Dataset $D = G \times T$, Number of drivers combined combinatorially c
Output: List of Multivariate Context Motifs $MvarCMList$

```
1  i ← 1;
2  while {i ≤ c} do
3      if i == 1 then
4          Obtain contexts from univariate in-silico conditioning;
5      else
           /* Create i-conditioner set */
6          Generate new i-conditioners using Boolean operators on i-1
           conditioners ;
7          if new i-conditioner drives a context motif then
8              Check significance of derived sample set based on new
               i-conditioner;
9              if significant then
10                 add to filtered i-conditioner list;
11             end
12         end
13         Use filtered i-conditioner list to identify multivariate context motifs;
14         if i == c then
15             MvarCMList ← i-conditioner multivariate context motifs;
16         end
17     end
18     i ← i + 1;
19 end
```

Algorithm 8: Algorithm for the identification of multivariate context motifs

We use m and n to denote the total number of genes and samples in data set, c is the number of conditioners to be combined, l is the list size of candidates for multivariate i-conditioner set, and k as the user specified number of iterations for bootstrap sampling to calculate statistical significance. Then the complexity to identify a single multivariate context motif is $O(n^3m)$. The complexity to identify multivariate context motif for every set of candidate multivariate conditioners is $O(n^3ml^c)$. Bootstrap resampling to calculate the statistical significance is $O(n^4mk)$. Thus the complexity of Algorithm 8 is $O(n^4mk + n^3ml^c)$.

Combining Conditioners Using Boolean Operators

We refer to the samples T_i, where the conditioning factor G_i is observed as active, as the conditioned sample set or context sample set. We apply Boolean operations such as AND, OR in order to combine gene states and obtain combined conditioners. The AND operation will be a context motif sample set constraining operation and OR operation would be a context motif sample set expanding operation. For instance, if we have two context motifs $C_1 = \{G_1, Y_1, S_1, T_1\}$ and $C_2 = \{G_2, Y_2, S_2, T_2\}$, where $T_1 = \{t_1, t_2, t_3\}$ and $T_2 = \{t_2, t_3, t_4\}$. Then combined context motif $C_{1.2} = \text{AND}(C_1, C_2)$ would be the effect of G_1 AND G_2 on sample set of combined context motif $T_{1.2} = T_1 \cap T_2 = \{t_2, t_3\}$. As $T_{1.2} \subset T_1$ and $T_{1.2} \subset T_2$, the AND operation effectively reduces the context motif sample set. Similarly, we can show that the OR operation effectively increases the context motif sample set.

Transcription Factor Enrichment Ratio

In order to evaluate the number of regulatory factors found by the multivariate analysis, we investigate if the conditioning factors in the multivariate context motifs are enriched with regulatory elements such as transcription factors (TFs). Only when the TFs are flagged as conditioning factors in the context would it count as evidence to assert an enrichment of regulatory elements in the conditioning factor of the context.

7.4 Results

Combining conditioners using Boolean operator OR provides alternate drivers or conditioning pathways that yield the same output expressed by a common set of drivens. Combining conditioners using Boolean operator AND, however, focuses on the combinatorial and concerted activity of the conditioning factors. In this section, for applying the framework to biological data, we focused on AND operator to find new contexts that cannot be identified by univariate analysis alone. Also, we restricted our conditioning set size to i = 2 ,i.e., to combinations of two drivers but the model is extensible to accommodate any number of driver combinations.

We applied the method to a dataset of *D. Melanogaster* gene-expression profiles from 88 experimental conditions hybridized to a total of 267 GeneChip Drosophila Genome Arrays (Affymetrix, Santa Clara, CA) [84]. It consisted of six independent investigations studying five different experimental questions namely aging, immune response, DNA-damage response, resistance to DDT, and embryonic development. The dataset had 13,165 genes with RNA samples from both embryos and adults. The downloaded data was in the form of mean-normalized log ratios (see [84] for more detail). The data was discretized based-on fold-change against mean of each gene. In other words, for each gene, if the log ratio differs from the mean by more than 1.5 folds, it was assigned +1 or -1 depending its direction of changes. Otherwise, it was assigned 0.

Combinatorial Drivers

We used settings of crosstalk 0.2 and conditioning 0.1 to identify multivariate context motifs (Mvar CMs). We filtered the context motifs at different significance levels and obtained the results presented in Table 7.1. We noticed that we get a very high number of multivariate context motifs with same p-values. Thus we could not rank the multivariate context motifs by p-values. In order to examine the characteristics

of the identified context motifs, we used an approximation of the number of genes times the number of samples to rank the multivariate context motifs. Table 7.1 lists the number of identified multivariate context motifs at different p-value thresholds. We also list the range of number of drivers found in each case.

Table 7.1: Drosophila Melanogaster data multivariate context motifs identified at crosstalk of 0.2, conditioning of 0.1 and statistical significance of 0.05.

P(G+|S)	Identified Mvar CMs	Number of driven genes
10^{-4}	79	[2492, 3728]
10^{-3}	101,494	[142, 3768]
10^{-2}	1,592,953	[5, 3768]
0.05	4,038,021	[4, 3768]

Transcription Factor Enrichment in Conditioning Factors

In order to evaluate the number of regulatory factors found by the multivariate analysis, we investigated if the conditioning factors (drivers) in the multivariate context motifs C_i are enriched with transcription factors (TFs). The TF list for *Drosophila Melanogaster* was taken from the FlyTF database [85]. It initially consisted of 753 site-specific TFs but was shortened to 658 according to total gene matches found across the gene expression data set.

The lists were used to check the conditioning factors of the context motifs for at least one occurrence of TF. The computed enrichment ratio was the success rate of context motifs that contain at least one TF in their conditioning factors. TF regulation is considered when flagged as conditioning factors of the cellular context. The same property ceases to hold when the gene is considered to be under the influence of other cellular factors.

Simulation

In order to compare if TF presence is higher in the context motifs extracted by the algorithm than that by pure chance, re-sampling based simulations were carried out as follows. We first estimated empirical distribution of the number of conditioning

factors in each context motif from the identified multivariate context motifs for each set of p-values $(10^{-2}, 10^{-3}, 10^{-4})$ using the drivers of top k = 100, 200, 300 and 1000 multivariate context motifs. For each set we found out the list of unique drivers in that set. Once identified, we mapped them to find out the percentage of TF presence. Then, N_r (=10,000) random sets of genes of unique driver set sizes were sampled, using non-repeating Bootstrap sampling. Following the empirical distribution, the enrichment percentage of TFs was calculated. This was used to find the p-value of observing as many or more TFs in the driver set as found in the real case.

Table 7.2: Drosophila Melanogaster data multivariate context motifs. Probability of obtaining that many or more TF matches among uniquely chosen drivers via non-repeating Bootstrap resampling.

P-value	Top k Mvar CMs	TF driven Mvar CMs	Number of unique drivers	TF match unique drivers	P(TFMatch+)
10^{-4}	79	28	56	7	0.0103
10^{-3}	100	34	48	5	0.0538
	200	69	85	7	0.0404
	500	150	189	16	0.0117
	1000	273	274	22	0.0106
10^{-2}	100	18	44	5	0.0256
	200	33	66	6	0.0578
	500	111	124	9	0.0504
	1000	200	187	13	0.0808

As observed in Table 7.2, the probability of observing as many or higher number of TFs by chance via random sampling is very low. This shows that these statistically significant multivariate context motifs identified more regulatory factors (TFs) than possible by random chance alone. To provide a comparison with the univariate case, Expattern only returned 37 univariate context motifs at p-value of 10^{-4}, of which only three context motifs had a TF as a driver. Thus univariate context motifs cannot effectively capture combinatorial regulatory interactions. Table 7.2 shows Expattern can be applied to situations where the combinatorial effect of different regulatory factors are at play. The contexts identified through this

method would give us a better in-silico insight into possible regulatory pathways and combinations hitherto unknown.

7.5 Summary

In this chapter we identified multivariate conditioning factors belonging to statistically significant context motifs. We used the gene expression data of Drosophila Melanogaster, and mapped the identified multivariate conditioning factors to transcription factors, a class of known regulatory elements. Comparison results show that there is enrichment of TFs in the actual contexts sorted by p-values than random contexts, generated by re-sampling based simulation. This approach can be applied to the study of any organism or complex disease where the interplay of myriad factors propels the system to different phenotypic outcomes. For example, the study of human cancer data to identify conditioning factors required to function in concert for cancer development. The multivariate contextual outcome of the framework can be employed to advance hypotheses about hitherto unknown biological entities (e.g.: miRNA, kinases, gene expression) required to be working in concert to yield particular cellular pathologies.

Chapter 8

CONCLUSION

Inference of context specific gene regulatory networks from available biological data is a highly challenging problem of computational systems biology. Briefly, the challenges are four fold – identifying an appropriate mathematical model, learning individual interactions, building a network from identified interactions and developing a method for validation. In this dissertation we have developed and presented a computational framework that meets these challenges to model and learn context specific GRNs from multi-source data. In this chapter, we summarize the key contributions of this dissertation and discuss directions for future research.

8.1 Key Contributions

Developed Framework To Learn csGRNs

One of the main contributions of this dissertation is the development of a systematic computational framework, ExPattern, to learn context specific GRNs. This entailed many critical tasks, the first of which was the formulation of the biological problem into an equivalent computational problem (chapter 1). We used contextual genomic regulation mathematical model, an appropriate mathematical model for capturing the notions of contextual consistency. We formally defined a set of constructs in the framework – measures of consistency (conditioning and crosstalk), cellular context motifs (interactions) and cellular contexts (networks)(section 3.2). This model differs from other gene set dependency models in that it learns about the gene modules and their dependency structure simultaneously from data rather than from predefined gene sets.

Comparable module identification or biclustering methods just output a collection of thousands of modules or biclusters without integrating the results to provide a comprehensible set of network or graph solution. We developed the in-silico conditioning method to identify cellular context motifs and agglomerated these into

sample annotated comprehensible set of context specific networks. We also defined a unique way of scoring the contexts and identify functionally or biologically significant contexts. The score, Sample Association Score, focuses on the enriched presence of tissue type from samples (in chapter 4, Section 4.3). Other GRN reverse engineering methods do not take into account the sample composition of the modules or rather assume the interactions are valid across the entire sample set. In our case however, each context represents a different configuration of samples and it is interesting to observe enriched sample or tissue type capturing the phenotypic basis of the context specific GRN.

Created Artificial Contextual Networks for Framework Validation

In this thesis we present an innovative method to create artificial contextual GRNs and its associated data to validate the results of different reverse engineering algorithms (Chapters 4,5). In order to use these synthetic data sets to validate the context-specific GRNs produced through the cellular context mining technique, we avoid bias by generating the networks by a method other than that which we want to validate. We used the benchmark artificial network generator A-biochem to create contextual networks. A-biochem has been used extensively for testing the performance of reverse engineering algorithms in Dialogue for Reverse Engineering Assessments and Methods (DREAM). However, popular artificial network generators assume a single underlying network and do not have contextual information embedded in the network or the data. To our knowledge, no one else has designed or created artificial contextual networks. We randomly combined individual artificial networks generated by A-biochem, treating each as a unique context, and combined the corresponding data to generate contextual gene expression data. We compared the performance of our method with other popular GRN reverse engineering methods (Chapter 4).

Developed Innovative Strategies To Prune Extraneous Edges

We introduced unique ways of pruning extraneous edges to reduce the false positives from the context specific GRNs (Chapter 5). GRNs learned by the framework (method outlined in Chapter 4) are often made of a few thousand nodes (genes) and tens of thousands of interactions rendering interpretation of the network almost impossible. Large amount of redundancy in the network, especially with overlapping contexts adds to the difficulty in interpretation. We observed on comparison of the performance of our framework with other methods that even though Expattern identified a higher percentage of true edges the overall scores of precision and f-measure were low because of high number of false positive edges. In order to compensate for quantization effects and network redundancy on increased false positive edges found through our method, we developed context-specific GRN pruning methods. In our experiments with artificial contextual networks, the strategies successfully removed extraneous edges, reducing false positives (by 50% - 70%) without losing more than 10% of true positives by exploiting relationships between the consistency metrics - crosstalk and conditioning (Chapter 5).

Integrated Multiple Sources of Data

Popular approaches to integrate data from multiple sources do not distinguish between data types, treating all equally in the mathematical framework to generate corresponding interactions. For obtaining biologically meaningful interactions we allow the users to include apriori knowledge as meta-rules (Chapter 6) in Expattern to guide the search of context motifs and thereby contexts. For example, we observe that interactions between entities of the same data types has a stronger correlation (low crosstalk and conditioning) than interactions between entities of different types. Thus any method to integrate data from different biological domains for learning, needs to taking into account this difference. We applied our method

to NCI60 drug activity and gene expression data and inferred mechanism of action for a drug with unknown mechnism of action. We also applied our method to TCGA data and obtained interesting contexts which combined aCGH and gene expression data (Chapter 6).

Identified Combinatorial Conditioning Factors

We extended the framework to identify complex conditioning factors, using Boolean operations to identify combinatorial conditioning factors and restricting the search space using the sizes of samples to be considered. Focusing on transcription factors we found that univariate contexts do not show much enrichment of TFs. This is understandable as research shows that TFs are not specific, it is only in combination with its co-factors that the TFs show deterministic transcriptional activity with high specificity. In order to identify such combinatorial regulatory factors, we applied the multivariate contextual analysis of Expattern to a dataset of Drosophila M. We used Bootstrap resampling method to estimate the probability of finding as many or more transcriptional factors as found by Expattern. Our results show very low p-values, making the multivariate contextual networks identified by Expattern as statistically significant with TF enriched conditioning factors (Chapter 7).

8.2 Future Directions

We discuss several interesting directions for future work here.

1. Currently our framework quantizes the data as a preprocessing step before calculating the consistency metrics. This step enforces a dependency on the quantization and normalization method used. Using the continuous values instead of discrete values of data in the framework might restrict the loss of information by quantization step. A necessary prerequisite would be the assumption of an underlying probability distribution of the data and mathematical redefinitions of consistency metrics conditioning, crosstalk and statistical

significance of context motifs for continuous values.

2. We applied the framework to finding combinatorial conditioners and integrating multi-source data separately. An interesting next step would be to combine the methods of both multivariate and multi-type runs to guide the identification of combinatorial multi type conditioning factors. It would require an assessment of the kind of apriori knowledge to be included for identifying context motifs. For example, what kind of multivariate combinations would be allowed given the types of data sources and strengths of interactions. For such a scenario, the next challenge would be the validation of the obtained results. Also, if we want to reduce extraneous edges in this network, some modifications to pruning strategies would have to be made. For example, whether the data type composition of extraneous edges would influence the edges to be removed? Finally, the interpretation of results would pose a challenge as – what would the edges and contexts mean when observing combinatorial drivers of different types of data ?

3. Currently we validated our results using different artificial datasets. Understandably, different network characteristics would influence the obtained structure of graphs when we apply different pruning strategies. A more rigorous study of different artificial networks and run outcomes would provide us a better picture of which types of networks would be more suitable for the different pruning strategies. For example, scale-free pruning uses the assumption of the original graph having scale free characteristics. How poorly would the method perform on networks that are not scale free?

4. Expattern executions for multi-source data used apriori knowledge of biologists to assign strengths and relationships between different data types. We can also run Expattern in an exploratory mode, by automating runs for differ-

ent settings and evaluating contexts based on different criteria. This can help us determine optimal settings for multi-type interactions. It is important to realize that the context evaluation is very highly dependent on the question the biologist is interested in and the outcomes (optimal settings) would change depending on the evaluation criteria.

Current computational approaches employed for bioinformatics research are geared towards unraveling underlying regulatory mechanisms from high throughput data. Purely logical or statistical approaches employed for hypothesis generation needs to be corroborated with biological evidence to confirm validity of assumed models. This thesis provides an approach to bridge statistical output with biological interpretation. To summarize, in this dissertation work we successfully modeled and learned context specific Gene Regulatory Networks from multi-source data. We applied it to both artificial datasets as well as real world biological data, focusing mainly on heterogeneous disease such as cancer. Functional enrichment and annotation helped validate the results obtained. We developed a unique method using apriori biological knowledge to integrate different data sources and identify contextual networks. We introduced an innovative way to reduce extraneous edges from the context specific GRNs and extended the framework to identify combinatorial conditioning factors to build contexts. We believe the computational framework Expattern is a unique and powerful exploratory tool for computational biologists, easily extensible to integrate multiple sources of data and identify combinatorial conditioning factors at play. We believe that this tool will definitely aid in understanding the underlying systems biology mechanisms of heterogeneous diseases such as cancers.

REFERENCES

[1] H. Kitano, "Systems biology: a brief overview," *Science*, vol. 295, no. 5560, p. 1662, 2002.

[2] S. Kim, I. Sen, and M. Bittner, "Mining molecular contexts of cancer via in-silico conditioning." in *Computational Systems Bioinformatics: Proceedings of the CSB 2007 Conference*, 2007, pp. 169–79.

[3] D. Bartel, "MicroRNAs genomics, biogenesis, mechanism, and function," *Cell*, vol. 116, no. 2, pp. 281–297, 2004.

[4] K. Basso, A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano, "Reverse engineering of regulatory networks in human B cells," *Nature genetics*, vol. 37, no. 4, pp. 382–390, 2005.

[5] 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. 370–378, 2002.

[6] J. Ihmels, S. Bergmann, and N. Barkai, "Defining transcription modules using large-scale gene expression data," pp. 1993–2003, 2004.

[7] Y. Cheng and G. Church, "Biclustering of Expression Data," in *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology table of contents*. AAAI Press, 2000, pp. 93–103.

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

[9] 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.

[10] F. Cappuzzo, F. Hirsch, E. Rossi, S. Bartolini, G. Ceresoli, L. Bemis, J. Haney, S. Witta, K. Danenberg, I. Domenichini *et al.*, "Epidermal growth factor receptor gene and protein and gefitinib sensitivity in non-small-cell lung cancer," *JNCI Journal of the National Cancer Institute*, vol. 97, no. 9, p. 643, 2005.

[11] U. Scherf, D. Ross, M. Waltham, L. Smith, J. Lee, L. Tanabe, K. Kohn, W. Reinhold, T. Myers, D. Andrews *et al.*, "A gene expression database for the molecular pharmacology of cancer," *Nature Genetics*, vol. 24, pp. 236–244, 2000.

[12] 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*, vol. 100, no. 25, pp. 14 796–14 799, 2003.

[13] S. Kauffman, "The Origins of Order: Self-organization and Selection in Evolution (1993)." New York: Oxford University Press, 1993.

[14] 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.

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

[16] 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.

[17] M. Zou and S. Conzen, "A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data," *Bioinformatics*, vol. 21, no. 1, pp. 71–79, 2005.

[18] J. Pearl, *Causality: Models, reasoning, and inference*. Cambridge Univ Pr, 2000.

[19] P. Mendes, W. Sha, and K. Ye, "Artificial gene networks for objective comparison of analysis algorithms," *Bioinformatics*, vol. 19, no. 90002, pp. 122–129, 2003.

[20] B. Haynes and M. Brent, "Benchmarking regulatory network reconstruction with GRENDEL," *Bioinformatics*, vol. 25, no. 6, p. 801, 2009.

[21] T. Van den Bulcke, K. Van Leemput, B. Naudts, P. van Remortel, H. Ma, A. Verschoren, B. De Moor, and K. Marchal, "SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms," *BMC bioinformatics*, vol. 7, no. 1, p. 43, 2006.

[22] P. Erdos and A. Renyi, "On random graphs," *Publ. Math. Debrecen*, vol. 6, no. 290-297, p. 156, 1959.

[23] A. Bild, G. Yao, J. Chang, Q. Wang, A. Potti, D. Chasse, M. Joshi, D. Harpole, J. Lancaster, A. Berchuck *et al.*, "Oncogenic pathway signatures in human cancers as a guide to targeted therapies," *Nature*, vol. 439, no. 7074, p. 353, 2006.

[24] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, p. 2507, 2007.

[25] 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, no. 18-29. Citeseer, 1998, p. 22.

[26] T. Cover and J. Thomas, *Elements of information theory*. Wiley, 2006.

[27] E. Segal, N. Friedman, D. Koller, and A. Regev, "A module map showing conditional activity of expression modules in cancer," *Gene expression*, vol. 1, p. 45, 2004.

[28] A. Tanay, R. Sharan, and R. Shamir, "Discovering statistically significant biclusters in gene expression data," *Bioinformatics*, vol. 18, no. suppl 1, p. S136, 2002.

[29] F. Shi, C. Leckie, G. MacIntyre, I. Haviv, A. Boussioutas, and A. Kowalczyk, "A bi-ordering approach to linking gene expression with clinical annotations in gastric cancer," *BMC bioinformatics*, vol. 11, no. 1, p. 477, 2010.

[30] A. Bhattacharya and R. De, "Bi-correlation clustering algorithm for determining a set of co-regulated genes," *Bioinformatics*, vol. 25, no. 21, p. 2795, 2009.

[31] E. Segal, M. Shapira, A. Regev, D. PeŠer, D. Botstein, D. Koller, and N. Friedman, "Module networks: Discovering regulatory modules and their condition specific regulators from gene expression data," *Nature genetics*, vol. 34, no. 2, pp. 166–176, 2003.

[32] 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.

[33] M. Kanehisa and S. Goto, "Kegg: Kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, p. 27, 2000.

[34] 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.

[35] 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.

[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] 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.

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

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

[40] 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.

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

[42] E. Dougherty, M. Brun, J. Trent, and M. Bittner, "Conditioning-Based Modeling of Contextual Genomic Regulation," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 29, 2007.

[43] 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.

[44] T. Michoel, R. De Smet, A. Joshi, Y. Van de Peer, and K. Marchal, "Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks," *BMC systems biology*, vol. 3, no. 1, p. 49, 2009.

[45] M. Li, I. Wang, Y. Li, A. Bruzel, A. Richards, J. Toung, and V. Cheung, "Widespread rna and dna sequence differences in the human transcriptome," *Science*, 2011.

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

[47] 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 Conference*, 2009.

[48] M. Rivas, R. Carnevale, C. Proietti, C. Rosemblit, W. Beguelin, M. Salatino, E. Charreau, I. Frahm, S. Sapia, P. Brouckaert *et al.*, "TNFα acting on TNFR1 promotes breast cancer growth via p42/P44 MAPK, JNK, Akt and NF-κB-dependent pathways," *Experimental Cell Research*, vol. 314, no. 3, pp. 509–529, 2008.

[49] J. Burton, S. Ely, P. Reddy, R. Stein, D. Gold, T. Cardillo, and D. Goldenberg, "CD74 is expressed by multiple myeloma and is a promising target for therapy," pp. 6606–6611, 2004.

[50] S. Oldford, J. Robb, D. Codner, V. Gadag, P. Watson, and S. Drover, "Tumor cell expression of HLA-DM associates with a Th1 profile and predicts improved survival in breast carcinoma patients," *International immunology*, vol. 18, no. 11, p. 1591, 2006.

[51] Y. Hao, J. Wang, N. Feng, and A. Lowe, "Determination of plasma glycoprotein 2 levels in patients with pancreatic disease," *Archives of Pathology and Laboratory Medicine*, vol. 128, pp. 668–674, 2004.

[52] M. Eisen, P. Spellman, P. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," pp. 14 863–14 868, 1998.

[53] M. de Hoon, S. Imoto, J. Nolan, and S. Miyano, "Open source clustering software," *Bioinformatics*, vol. 20, no. 9, p. 1453, 2004.

[54] S. Baswana, “Dynamic algorithms for graph spanners,” *Lecture Notes in Computer Science*, vol. 4168, p. 76, 2006.

[55] D. Peleg and A. Schäffer, “Graph spanners,” *Journal of graph theory*, vol. 13, no. 1, pp. 99–116, 1989.

[56] V. Dubois, C. Bothorel, R. France Telecom, and F. Lannion, “Transitive reduction for social network analysis and visualization,” in *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, 2005, pp. 128–131.

[57] K. Kyoda, M. Morohashi, S. Onami, and H. Kitano, “A gene network inference method from continuous-value gene expression data of wild-type and mutants,” *Genome Informatics Series*, pp. 196–204, 2000.

[58] R. Albert, B. DasGupta, R. Dondi, S. Kachalo, E. Sontag, A. Zelikovsky, and K. Westbrooks, “A novel method for signal transduction network inference from indirect experimental evidence,” *Journal of Computational Biology*, vol. 14, no. 7, pp. 927–949, 2007.

[59] Y. Saab, “A fast and effective algorithm for the feedback arc set problem,” *Journal of Heuristics*, vol. 7, no. 3, pp. 235–250, 2001.

[60] A. Sixtus and S. Ortmanns, “High quality word graphs using forward-backward pruning,” *vectors*, vol. 10, p. 1, 1999.

[61] T. Kuhn, P. Fetter, A. Kaltenmeier, and P. Regel-Brietzmann, “DP-based word-graph pruning,” in *IEEE International Conference On Acoustics Speech And Signal Processing*, vol. 2, 1996.

[62] H. Lähdesmäki, 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.

[63] B. Perrin, L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, and F. d’Alche Buc, “Gene networks inference using dynamic Bayesian networks,” *Bioinformatics-Oxford*, vol. 19, no. 2, pp. 138–148, 2003.

[64] S. Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, and S. Miyano, “Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks,” *Journal of Bioinformatics and Computational Biology.*, vol. 2, no. 1, pp. 77–98, 2004.

[65] A. Ramesh, R. Trevino, D. Von Hoff, and S. Kim, "Clustering Context-Specific Gene Regulatory Networks." in *Proceedings of the Pacific Symposium on Biocomputing*, 2010, pp. 444–455.

[66] R. Albert, "Scale-free networks in cell biology," *Journal of cell science*, vol. 118, no. 21, p. 4947, 2005.

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

[68] S. Draghici, P. Khatri, A. Eklund, and Z. Szallasi, "Reliability and reproducibility issues in DNA microarray measurements," *TRENDS in Genetics*, vol. 22, no. 2, pp. 101–109, 2006.

[69] E. Dougherty, "Small sample issues for microarray-based classification," *Comparative and Functional Genomics*, vol. 2, no. 1, pp. 28–34, 2001.

[70] O. Troyanskaya, K. Dolinski, A. Owen, R. Altman, and D. Botstein, "A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 14, p. 8348, 2003.

[71] C. Myers and O. Troyanskaya, "Context-sensitive data integration and prediction of biological networks," *Bioinformatics*, vol. 23, no. 17, p. 2322, 2007.

[72] D. Ross, U. Scherf, M. Eisen, C. Perou, C. Rees, P. Spellman, V. Iyer, S. Jeffrey, M. Van de Rijn, M. Waltham *et al.*, "Systematic variation in gene expression patterns in human cancer cell lines," *nature genetics*, vol. 24, no. 3, pp. 227–235, 2000.

[73] J. Rae, S. Ramus, M. Waltham, J. Armes, I. Campbell, R. Clarke, R. Barndt, M. Johnson, and E. Thompson, "Common origins of MDA-MB-435 cells from various sources with those shown to have melonoma properties," *Clinical and Experimental Metastasis*, vol. 21, no. 6, pp. 543–552, 2004.

[74] J. Rae, C. Creighton, J. Meck, B. Haddad, and M. Johnson, "MDA-MB-435 cells are derived from M14 Melanoma cells—-a loss for breast cancer, but a boon for melanoma research," *Breast cancer research and treatment*, vol. 104, no. 1, pp. 13–19, 2007.

[75] T. Jenssen, A. Lægreid, J. Komorowski, and E. Hovig, "A literature network of human genes for high-throughput analysis of gene expression," *Nature Genetics*, vol. 28, no. 1, pp. 21–28, 2001.

[76] 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.

[77] O. Hobert, "Gene regulation by transcription factors and microrna," *Science*, vol. 319, no. 5871, p. 1785, 2008.

[78] A. Frankel and P. Kim, "Modular structure of transcription factors: implications for gene regulation," *Cell*, vol. 65, no. 5, pp. 717–719, 1991.

[79] P. Sætrom, B. Heale, O. Snøve, L. Aagaard, J. Alluin, and J. Rossi, "Distance constraints between microrna target sites dictate efficacy and cooperativity," *Nucleic Acids Research*, vol. 35, no. 7, p. 2333, 2007.

[80] X. Yu, J. Lin, D. Zack, and J. Qian, "Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues," *Nucleic acids research*, vol. 34, no. 17, p. 4925, 2006.

[81] O. Kel-Margoulis, A. Kel, I. Reuter, I. Deineko, and E. Wingender, "TransCompel®: a database on composite regulatory elements in eukaryotic genes," *Nucleic acids research*, vol. 30, no. 1, p. 332, 2002.

[82] S. Mukherjee, S. Pelech, R. Neve, W. Kuo, S. Ziyad, P. Spellman, J. Gray, and T. Speed, "Sparse combinatorial inference with an application in cancer biology," *Bioinformatics*, vol. 25, no. 2, p. 265, 2009.

[83] I. Park, K. Lee, and D. Lee, "Inference of combinatorial boolean rules of synergistic gene sets from cancer microarray datasets," *Bioinformatics*, vol. 26, no. 12, p. 1506, 2010.

[84] P. Spellman and G. Rubin, "Evidence for large domains of similarly expressed genes in the drosophila genome," *Journal of Biology*, vol. 1, no. 1, p. 5, 2002.

[85] B. Adryan and S. Teichmann, "Flytf: a systematic review of site-specific transcription factors in the fruit fly drosophila melanogaster," *Bioinformatics*, vol. 22, no. 12, p. 1532, 2006.

Appendix A

MATHEMATICAL PROOFS

A.1 Mathematical definitions

Here, we use the shorthand of gene name (say X) to denote gene X is in active state, i.e. $y_x = 1$, where y_x is the activity state of the gene X and 1 represents the activity level. Similarly, X' denotes gene X is not in active state, i.e. $y_x \neq 1$. If we assume A to be the true driver gene and B, C to be driven genes in GRN H, the values of crosstalk and conditioning of induced edges could be calculated on the basis of the following formulas and definitions. Let δ_θ = threshold for δ (conditioning) and η_θ = threshold for η (crosstalk). Crosstalk and Conditioning formulas:

$$\delta_{ab} = 1 - \Pr(y_b = 1 | y_a = 1) \equiv 1 - \Pr(B|A)$$

$$\eta_{ab} = \Pr(y_b = 1 | y_a \neq 1) \equiv \Pr(B|A')$$

$$\delta_{bc} = 1 - \Pr(y_c = 1 | y_b = 1) \equiv 1 - \Pr(C|B)$$

$$\eta_{bc} = \Pr(y_c = 1 | y_b \neq 1) \equiv \Pr(C|B')$$

$$\delta_{ac} = 1 - \Pr(y_c = 1 | y_a = 1) \equiv 1 - \Pr(C|A)$$

$$\eta_{ac} = \Pr(y_c = 1 | y_a \neq 1) \equiv \Pr(C|A')$$

A.2 Transitive Edges

We assume A to be the true driver gene and B to be the true driven gene in one context motif and B to be the true driver gene and C to be the true driven gene in another context motif. Then both edges AB, BC present in GRN G, the induced values of crosstalk and conditioning of edge AC can be calculated as follows. To estimate:

$$\delta_{ac} = 1 - \Pr(y_c = 1 | y_a = 1) \equiv 1 - \Pr(C|A)$$

$$\eta_{ac} = \Pr(y_c = 1 | y_a \neq 1) \equiv \Pr(C|A')$$

Theorem A.2.1 (*Transitive Edge Conditioning*) *Given the values of conditioning and crosstalk of edges AB, BC as above,*

$$\delta_{ac} \geq [\delta_{ab}(1-\eta_{bc}) + \delta_{bc}(1-\delta_{ab})] + [\eta_{bc}(1-\eta_{ab}) - \eta_{ab}\delta_{bc}]\,\gamma_A \tag{A.1}$$

where

$$\gamma_A = \frac{\Pr(A')}{\Pr(A)} = \frac{1-\Pr(A)}{\Pr(A)}.$$

Proof. In order to find δ_{ac}, we first compute the expected value of $\Pr(C|A)$.

$$\Pr(C|A) = \frac{\Pr(A,C)}{\Pr(A)} = \frac{\Pr(A,B,C)}{\Pr(A)} + \frac{\Pr(A,B',C)}{\Pr(A)}$$

Expanding,

$$\begin{aligned}\Pr(A,B,C) &= \Pr(C,B) + \Pr(A) - \Pr\left(A \cup (B,C)\right)\\ &= \Pr(C|B)\cdot\Pr(B) + \Pr(A) - \Pr\left(A \cup (B,C)\right)\end{aligned}$$

Similarly,

$$\Pr(A,B',C) = \Pr(C|B')\cdot\Pr(B') + \Pr(A) - \Pr\left(A \cup (B',C)\right) \tag{A.2}$$

Substituting,

$$\begin{aligned}\Pr(C|A) &= \Pr(C|B)\cdot\frac{\Pr(B)}{\Pr(A)} + 1 - \frac{\Pr(A\cup(B,C))}{\Pr(A)} + \Pr(C|B')\cdot\frac{\Pr(B')}{\Pr(A)} + 1 - \frac{\Pr(A\cup(B',C))}{\Pr(A)}\\ &= \Pr(C|B)\cdot\frac{\Pr(B)}{\Pr(A)} + \Pr(C|B')\cdot\frac{\Pr(B')}{\Pr(A)} + 2 - \frac{\Pr\left(A\cup(B,C)\right) + \Pr\left(A\cup(B',C)\right)}{\Pr(A)}.\end{aligned}$$

As

$$\Pr(A\cup(B\cap C)) + \Pr(A\cup(B'\cap C)) = \Pr(A) + \Pr(A\cup C), \tag{A.3}$$

$$\begin{aligned}\Pr(C|A) &= \Pr(C|B)\cdot\frac{\Pr(B)}{\Pr(A)} + \Pr(C|B')\cdot\frac{\Pr(B')}{\Pr(A)} + 2 - \frac{\Pr(A) + \Pr(A\cup C)}{\Pr(A)}\\ &= (1-\delta_{bc})\cdot\frac{\Pr(B)}{\Pr(A)} + \eta_{bc}\cdot\frac{\Pr(B')}{\Pr(A)} + \left(1 - \frac{\Pr(A\cup C)}{\Pr(A)}\right)\end{aligned}$$

If we assume $\eta_{ac} \geq \eta_{ab}$ then:

$$
\begin{aligned}
&\eta_{ac} \geq \eta_{ab} \\
&\Leftrightarrow \Pr(C|A') \geq \Pr(B|A') \\
&\Leftrightarrow \Pr(C|A') \cdot \Pr(A') \geq \Pr(B|A') \cdot \Pr(A') \\
&\Leftrightarrow \Pr(A', C) \geq \Pr(A', B) \\
&\Leftrightarrow \Pr(C) - \Pr(A, C) \geq \Pr(B) - \Pr(A, B) \\
&\Leftrightarrow \Pr(A) + \Pr(C) - \Pr(A, C) \geq \Pr(A) + \Pr(B) - \Pr(A, B) \\
&\Leftrightarrow \Pr(A \cup C) \geq \Pr(A \cup B) \\
&\Leftrightarrow \frac{\Pr(A \cup C)}{\Pr(A)} \geq \frac{\Pr(A \cup B)}{\Pr(A)} \\
&\Leftrightarrow 1 - \frac{\Pr(A \cup C)}{\Pr(A)} \leq 1 - \frac{\Pr(A \cup B)}{\Pr(A)}.
\end{aligned}
$$

Using $1 - \frac{\Pr(A \cup C)}{\Pr(A)} \leq 1 - \frac{\Pr(A \cup B)}{\Pr(A)}$,

$$
\begin{aligned}
\Pr(C|A) &\leq (1 - \delta_{bc}) \cdot \frac{\Pr(B)}{\Pr(A)} + \eta_{bc} \cdot \frac{\Pr(B')}{\Pr(A)} + \left(1 - \frac{\Pr(A \cup B)}{\Pr(A)}\right) \\
1 - \delta_{ac} &\leq (1 - \delta_{bc}) \cdot \frac{\Pr(B)}{\Pr(A)} + \eta_{bc} \cdot \frac{\Pr(B')}{\Pr(A)} + \left(1 - \frac{\Pr(A) + \Pr(B) - \Pr(A, B)}{\Pr(A)}\right) \\
1 - \delta_{ac} &\leq (1 - \delta_{bc}) \cdot \frac{\Pr(B)}{\Pr(A)} + \eta_{bc} \cdot \frac{\Pr(B')}{\Pr(A)} + \left(\frac{-\Pr(B) + \Pr(A, B)}{\Pr(A)}\right) \\
1 - \delta_{ac} &\leq (1 - \delta_{bc} - 1) \cdot \frac{\Pr(B)}{\Pr(A)} + \eta_{bc} \cdot \frac{\Pr(B')}{\Pr(A)} + \Pr(B|A) \\
1 - \delta_{ac} &\leq -\delta_{bc} \cdot \frac{\Pr(B)}{\Pr(A)} + \eta_{bc} \cdot \frac{\Pr(B')}{\Pr(A)} + (1 - \delta_{ab}) \\
\therefore \delta_{ac} &\geq \delta_{ab} + \delta_{bc} \cdot \frac{\Pr(B)}{\Pr(A)} - \eta_{bc} \left(\frac{1 - \Pr(B)}{\Pr(A)}\right).
\end{aligned}
$$

To simplify it further,

$$
\begin{aligned}
\eta_{ab} = \Pr(B|A') &= \frac{\Pr(B, A')}{\Pr(A')} = \frac{\Pr(B) - \Pr(A, B)}{\Pr(A')} \\
&= \frac{\Pr(B) - (1 - \delta_{ab}) \Pr(A)}{\Pr(A')} \\
\therefore \Pr(B) &= \eta_{ab} \Pr(A') + (1 - \delta_{ab}) \Pr(A).
\end{aligned}
$$

Substituting,

$$\begin{aligned}
\therefore \delta_{ac} &\geq \delta_{ab} + \delta_{bc} \cdot \frac{\eta_{ab}\Pr(A') + (1-\delta_{ab})\Pr(A)}{\Pr(A)} - \eta_{bc}\left(\frac{1-\eta_{ab}\Pr(A') - (1-\delta_{ab})\Pr(A)}{\Pr(A)}\right) \\
&= \delta_{ab} + \delta_{bc} \cdot \left\{\eta_{ab}\frac{\Pr(A')}{\Pr(A)} + (1-\delta_{bc})\right\} - \frac{\eta_{bc}}{\Pr(A)} + \eta_{ab}\eta_{bc}\frac{\Pr(A')}{\Pr(A)} + (1-\delta_{ab})\,\eta_{bc} \\
&= \delta_{ab} + (\delta_{bc} + \eta_{bc})\left(\eta_{ab}\frac{\Pr(A')}{\Pr(A)} + 1 - \delta_{ab}\right) - \frac{\eta_{bc}}{\Pr(A)}.
\end{aligned}$$

Let

$$\gamma_A = \frac{\Pr(A')}{\Pr(A)} = \frac{1-\Pr(A)}{\Pr(A)},$$

then

$$\Pr(A') = \frac{\gamma_A}{1+\gamma_A}, \text{ and } \Pr(A) = \frac{1}{1+\gamma_A}.$$

Again, substituting, and rearranging terms

$$\begin{aligned}
\therefore \delta_{ac} &\geq \delta_{ab} + (\delta_{bc} + \eta_{bc})\,(\eta_{ab}\gamma_A + 1 - \delta_{ab}) - \eta_{bc}\,(1+\gamma_A) \\
&= \delta_{ab} + (\delta_{bc} + \eta_{bc})\,\eta_{ab}\gamma_A + (\delta_{bc} + \eta_{bc})\,(1-\delta_{ab}) - \eta_{bc} - \eta_{bc}\gamma_A \\
&= [\delta_{ab}\,(1-\eta_{bc}) + \delta_{bc}\,(1-\delta_{ab})] - [\eta_{ab}\,(\delta_{bc} + \eta_{bc}) - \eta_{bc}]\,\gamma_A \\
&= [\delta_{ab}\,(1-\eta_{bc}) + \delta_{bc}\,(1-\delta_{ab})] + [\eta_{bc}\,(1-\eta_{ab}) - \eta_{ab}\delta_{bc}]\,\gamma_A.
\end{aligned}$$

■

Now, given conditioning and crosstalk values as above, we can derive the expected value of transitive crosstalk as follows.

Theorem A.2.2 (Transitive Edge Crosstalk) *Assume if $\delta_{ac} > \delta_{bc}$, i.e., $\Pr(C|A) < \Pr(C|B)$, then*

$$\eta_{ac} > \frac{\eta_{bc}\cdot\gamma_B + \alpha_{bc}\cdot\{\alpha_{ab} - 1\}}{\gamma_A} \tag{A.4}$$

where

$$\alpha_{ab} = \frac{\Pr(B)}{\Pr(A)},$$

and γ_A and γ_B given as above.

Proof.

$$
\begin{aligned}
&\delta_{ac} > \delta_{bc} \\
&\Rightarrow 1 - \Pr(C|A) > 1 - \Pr(C|B) \\
&\Rightarrow \Pr(C|B) > \Pr(C|A) \\
&\Rightarrow \frac{\Pr(C,B)}{\Pr(B)} > \frac{\Pr(C,A)}{\Pr(A)} \\
&\Rightarrow \Pr(C,B) > \Pr(C,A) \cdot \frac{\Pr(B)}{\Pr(A)} \\
&\Rightarrow \Pr(C) - \Pr(C|B') \cdot \Pr(B') > \{\Pr(C) - \Pr(C|A')\Pr(A')\} \cdot \frac{\Pr(B)}{\Pr(A)} \\
&\Rightarrow \Pr(C) \cdot \left\{1 - \frac{\Pr(B)}{\Pr(A)}\right\} + \Pr(C|A') \cdot \Pr(A') \cdot \frac{\Pr(B)}{\Pr(A)} > \Pr(C|B') \cdot \Pr(B') \\
&\Rightarrow \Pr(C) \cdot \left\{1 - \frac{\Pr(B)}{\Pr(A)}\right\} + \eta_{ac} \cdot \frac{\Pr(A')}{\Pr(A)} \cdot \Pr(B) > \eta_{bc} \cdot \Pr(B') \\
&\Rightarrow \eta_{ac} > \frac{\eta_{bc} \cdot \Pr(B') - \Pr(C) \cdot \left\{1 - \frac{\Pr(B)}{\Pr(A)}\right\}}{\Pr(B) \cdot \frac{\Pr(A')}{\Pr(A)}} \\
&\Rightarrow \eta_{ac} > \frac{\eta_{bc} \cdot \frac{\Pr(B')}{\Pr(B)} + \frac{\Pr(C)}{\Pr(B)} \cdot \left\{\frac{\Pr(B)}{\Pr(A)} - 1\right\}}{\frac{\Pr(A')}{\Pr(A)}}
\end{aligned}
$$

Let

$$\alpha_{ab} = \frac{\Pr(B)}{\Pr(A)} \tag{A.5}$$

If we are given $\delta_{ab} < \delta_\theta$ where δ_θ is the conditioning threshold value, then $\alpha_{ab} > 1 - \delta_\theta$ as

$$
\begin{aligned}
&\delta_{ab} < \delta_\theta \\
&1 - \Pr(B|A) < \delta_\theta \\
&\Pr(B|A) > 1 - \delta_\theta \\
&\Pr(A,B) > \{1 - \delta_\theta\} \cdot \Pr(A) \\
&\Pr(B) > \Pr(A,B) > \{1 - \delta_\theta\} \cdot \Pr(A) \\
&\frac{\Pr(B)}{\Pr(A)} > 1 - \delta_\theta \\
&\alpha_{ab} > 1 - \delta_\theta
\end{aligned}
$$

Substituting back we can write,

$$\eta_{ac} > \frac{\eta_{bc} \cdot \gamma_B + \alpha_{bc} \cdot \{\alpha_{ab} - 1\}}{\gamma_A}$$

■

A.3 Sibling Edges

We assume A to be the true driver gene and B, C to be the true driven genes, with edges AB, AC present in GRN H, the induced values of crosstalk and conditioning of edge BC can be calculated as follows. To estimate:

$$\delta_{bc} = 1 - \Pr(y_c = 1 | y_b = 1) \equiv 1 - \Pr(C|B)$$
$$\eta_{bc} = \Pr(y_c = 1 | y_b \neq 1) \equiv \Pr(C|B')$$

Theorem A.3.1 (*Sibling Edge Conditioning*) *Given the values of conditioning and crosstalk of edges AB and AC, if we assume $\eta_{bc} \geq \eta_{ac}$, i.e. $\Pr(C|B') \geq \Pr(C|A')$, then*

$$\delta_{bc} \geq 1 - \{(1 - \delta_{ac}) \cdot \alpha_{ba} + \eta_{ac} \cdot (1 - \alpha_{ba})\} \tag{A.6}$$

where α_{ba} is defined similarly as above.

Proof. In order to find δ_{bc}, we first compute the expected value of $\Pr(C|B)$.

$$\Pr(C|B) = \frac{\Pr(B, C)}{\Pr(B)} = \frac{\Pr(A, B, C)}{\Pr(B)} + \frac{\Pr(A', B, C)}{\Pr(B)}$$

Expanding,

$$\begin{aligned}\Pr(A, B, C) &= \Pr(A, C) + \Pr(B) - \Pr\left(B \cup (A, C)\right) \\ &= \Pr(C|A) \cdot \Pr(A) + \Pr(B) - \Pr\left(B \cup (A, C)\right)\end{aligned}$$

Similarly,

$$\Pr(A', B, C) = \Pr(C|A') \cdot \Pr(A') + \Pr(B) - \Pr\left(B \cup (A', C)\right)$$

Substituting,

$$\begin{aligned}\Pr(C|B) &= \Pr(C|A) \cdot \frac{\Pr(A)}{\Pr(B)} + 1 - \frac{\Pr(B \cup (A, C))}{\Pr(B)} + \Pr(C|A') \cdot \frac{\Pr(A')}{\Pr(B)} + 1 - \frac{\Pr(B \cup (A', C))}{\Pr(B)} \\ &= \Pr(C|A) \cdot \frac{\Pr(A)}{\Pr(B)} + \Pr(C|A') \cdot \frac{\Pr(A')}{\Pr(B)} + 2 - \frac{\Pr\left(B \cup (A, C)\right) + \Pr\left(B \cup (A', C)\right)}{\Pr(B)}.\end{aligned}$$

As

$$\Pr(B \cup (A \cap C)) + \Pr(B \cup (A' \cap C)) = \Pr(B) + \Pr(B \cup C), \tag{A.7}$$

$$\begin{aligned}\Pr(C|B) &= \Pr(C|A) \cdot \frac{\Pr(A)}{\Pr(B)} + \Pr(C|A') \cdot \frac{\Pr(A')}{\Pr(B)} + 2 - \frac{\Pr(B) + \Pr(B \cup C)}{\Pr(B)} \\ &= (1 - \delta_{ac}) \cdot \frac{\Pr(A)}{\Pr(B)} + \eta_{ac} \cdot \frac{\Pr(A')}{\Pr(B)} + \left(1 - \frac{\Pr(B \cup C)}{\Pr(B)}\right)\end{aligned}$$

Now,

$$1 - \frac{\Pr(B \cup C)}{\Pr(B)} = 1 - \frac{\Pr(B) + \Pr(B', C)}{\Pr(B)} = -\frac{\Pr(B', C)}{\Pr(B)} \tag{A.8}$$

If we assume $\eta_{bc} \geq \eta_{ac}$ then:

$$\begin{aligned}&\eta_{bc} \geq \eta_{ac} \\ &\Leftrightarrow \Pr(C|B') \geq \Pr(C|A') \\ &\Leftrightarrow \frac{\Pr(B', C)}{\Pr(B')} \geq \Pr(C|A') \\ &\Leftrightarrow -\frac{\Pr(B', C)}{\Pr(B)} \cdot \frac{\Pr(B)}{\Pr(B')} \leq -\Pr(C|A') \\ &\Leftrightarrow -\frac{\Pr(B', C)}{\Pr(B)} \leq -\Pr(C|A') \cdot \frac{\Pr(B')}{\Pr(B)}\end{aligned}$$

Using the above results and substituting back,

$$\begin{aligned}\Pr(C|B) &\leq (1 - \delta_{ac}) \cdot \frac{\Pr(A)}{\Pr(B)} + \eta_{ac} \cdot \frac{\Pr(A')}{\Pr(B)} - \Pr(C|A') \cdot \frac{\Pr(B')}{\Pr(B)} \\ &= (1 - \delta_{ac}) \cdot \frac{\Pr(A)}{\Pr(B)} + \eta_{ac} \cdot \frac{\Pr(A') - \Pr(B')}{\Pr(B)}\end{aligned}$$

Let

$$\alpha_{ba} = \frac{\Pr(A)}{\Pr(B)} \tag{A.9}$$

Then,

$$\begin{aligned}
&\Pr(C|B) \leq (1-\delta_{ac}) \cdot \alpha_{ba} + \eta_{ac} \cdot (1-\alpha_{ba}) \\
&\therefore 1-\delta_{bc} \leq (1-\delta_{ac}) \cdot \alpha_{ba} + \eta_{ac} \cdot (1-\alpha_{ba}) \\
&\quad \Leftrightarrow \delta_{bc} \geq 1 - \{(1-\delta_{ac}) \cdot \alpha_{ba} + \eta_{ac} \cdot (1-\alpha_{ba})\}
\end{aligned}$$

■

Theorem A.3.2 (Sibling Edge Crosstalk) *Assume if* $\delta_{bc} \geq \delta_{ac}$*, i.e.,* $\Pr(C|B) \leq \Pr(C|A)$*, then*

$$\eta_{bc} \geq \frac{\eta_{ac} \cdot \gamma_A - \alpha_{ac}\,(1-\alpha_{ba})}{\gamma_B} \tag{A.10}$$

where $\alpha_{ac}, \alpha_{ba}, \gamma_A$ *and* γ_B *are defined similarly as above.*

Proof.

$$\begin{aligned}
&\delta_{bc} \geq \delta_{ac} \\
&\Rightarrow 1 - \Pr(C|B) \geq 1 - \Pr(C|A) \\
&\Rightarrow \Pr(C|A) \geq \Pr(C|B) \\
&\Rightarrow \frac{\Pr(C,A)}{\Pr(A)} \geq \frac{\Pr(C,B)}{\Pr(B)} \\
&\Rightarrow \Pr(C,A) \geq \Pr(C,B) \cdot \frac{\Pr(A)}{\Pr(B)} \\
&\Rightarrow \Pr(C) - \Pr(C,A') \geq (\Pr(C) - \Pr(C,B')) \cdot \frac{\Pr(A)}{\Pr(B)} \\
&\Rightarrow \Pr(C) - \Pr(C|A') \cdot \Pr(A') \geq (\Pr(C) - \Pr(C|B') \cdot \Pr(B')) \cdot \frac{\Pr(A)}{\Pr(B)} \\
&\Rightarrow (1-\alpha_{ba}) \Pr(C) + \eta_{bc} \cdot \Pr(A) \cdot \frac{\Pr(B')}{\Pr(B)} \geq \eta_{ac} \cdot \Pr(A') \\
&\Rightarrow \eta_{bc} \cdot \Pr(A) \cdot \frac{\Pr(B')}{\Pr(B)} \geq \eta_{ac} \cdot \Pr(A') - (1-\alpha_{ba}) \Pr(C)
\end{aligned}$$

$$\Rightarrow \eta_{bc} \geq \frac{\eta_{ac} \cdot \Pr(A') - (1 - \alpha_{ba}) \Pr(C)}{\Pr(A) \cdot \frac{\Pr(B')}{\Pr(B)}}$$

$$\Rightarrow \eta_{bc} \geq \frac{\eta_{ac} \cdot \frac{\Pr(A')}{\Pr(A)} - (1 - \alpha_{ba}) \cdot \frac{\Pr(C)}{\Pr(A)}}{\frac{\Pr(B')}{\Pr(B)}}$$

Let

$$\gamma_A = \frac{\Pr(A')}{\Pr(A)} = \frac{1 - \Pr(A)}{\Pr(A)},$$

Substituting,

$$\eta_{bc} \geq \frac{\eta_{ac} \cdot \gamma_A - \alpha_{ac} (1 - \alpha_{ba})}{\gamma_B} \tag{A.11}$$

■

A.4 Reverse Edges

We assume A to be the true driver gene and B to be the true driven gene, with edge AB present in GRN H, the induced values of crosstalk and conditioning of edge BA can be calculated as follows. Let δ_θ = threshold for δ (conditioning) and η_θ = threshold for η (crosstalk).

Theorem A.4.1 (Reverse Edge Conditioning) *Given conditioning value δ_{ab} of edge AB, we have*

$$\delta_{ba} = 1 - (1 - \delta_{ab}) \cdot \frac{\Pr(A)}{\Pr(B)} \tag{A.12}$$

Proof.

$$\begin{aligned} \delta_{ba} &= 1 - \Pr(A|B) \\ &= 1 - \Pr(B|A) \cdot \frac{\Pr(A)}{\Pr(B)} \\ &= 1 - (1 - \delta_{ab}) \cdot \frac{\Pr(A)}{\Pr(B)} \end{aligned}$$

■

Theorem A.4.2 (Reverse Crosstalk) *Given crosstalk value η_{ab} of edge AB, we have*

$$\eta_{ba} = 1 - (1 - \eta_{ab}) \cdot \frac{\Pr(A')}{\Pr(B')} \tag{A.13}$$

Proof.

$$\begin{aligned}
\eta_{ba} &= \Pr(A|B') \\
&= \frac{\Pr(A \cap B')}{\Pr(B')} \\
&= \frac{\Pr(A) - \Pr(A \cap B)}{\Pr(B')}
\end{aligned}$$

Now,

$$\eta_{ab} = \Pr(B|A') = \frac{\Pr(A' \cap B)}{\Pr(A')}$$

$$\eta_{ab} = \frac{\Pr(B) - \Pr(A \cap B)}{\Pr(A')}$$

$$\Rightarrow \Pr(A \cap B) = \Pr(B) - \Pr(A') \cdot \eta_{ab}$$

Substituting,

$$\begin{aligned}
\eta_{ba} &= \frac{\Pr(A) - \Pr(A \cap B)}{\Pr(B')} \\
&= \frac{\Pr(A) - \Pr(B) + \Pr(A') \cdot \eta_{ab}}{\Pr(B')} \\
&= \frac{\Pr(A) - 1 + 1 - \Pr(B) + \Pr(A') \cdot \eta_{ab}}{\Pr(B')} \\
&= \frac{\Pr(B') - (1 - \eta_{ab}) \cdot \Pr(A')}{\Pr(B')} \\
&= 1 - (1 - \eta_{ab}) \cdot \frac{\Pr(A')}{\Pr(B')}
\end{aligned}$$

■

Reverse Edge in Context Graph

An edge BA would be included in the context graph H, given that edge AB is already present in H iff

$$\delta_{ba} < \delta_\theta \text{ AND } \eta_{ba} < \eta_\theta \tag{A.14}$$

For the conditioning of BA to be less than the threshold,

$$\begin{aligned}
&\delta_{ba} < \delta_\theta \\
&\Rightarrow 1 - (1 - \delta_{ab}) \cdot \frac{\Pr(A)}{\Pr(B)} < \delta_\theta \\
&\Rightarrow \frac{1 - \delta_\theta}{1 - \delta_{ab}} < \frac{\Pr(A)}{\Pr(B)}
\end{aligned}$$

For the crosstalk of BA to be less than the threshold,

$$\begin{aligned}
&\eta_{ba} < \eta_\theta \\
&\Rightarrow 1 - (1 - \eta_{ab}) \cdot \frac{\Pr(A')}{\Pr(B')} < \eta_\theta \\
&\Rightarrow \frac{1 - \eta_\theta}{1 - \eta_{ab}} < \frac{\Pr(A')}{\Pr(B')}
\end{aligned}$$

When both the above conditions are true, reverse edge BA, i.e., gene B regulating gene A is thought to be a possible regulation and included in the graph.

Parameter Relationship between Forward and Reverse Edges

If we assume $\delta_{ab} < \delta_{ba}$, we find

$$\begin{aligned}
\delta_{ab} < \delta_{ba} \quad &\Rightarrow \quad \delta_{ab} < 1 - (1 - \delta_{ab}) \cdot \frac{\Pr(A)}{\Pr(B)} \\
&\Rightarrow \quad (1 - \delta_{ab}) \cdot \frac{\Pr(A)}{\Pr(B)} < (1 - \delta_{ab}) \\
&\Rightarrow \quad (1 - \delta_{ab}) \cdot \left(1 - \frac{\Pr(A)}{\Pr(B)}\right) > 0 \\
&\Rightarrow \quad \Pr(A) < \Pr(B) \\
&\Rightarrow \quad 1 - \Pr(A) > 1 - \Pr(B) \\
&\Rightarrow \quad \frac{1 - \Pr(A)}{1 - \Pr(B)} > 1 \\
&\Rightarrow \quad (1 - \eta_{ab}) \cdot \left(\frac{1 - \Pr(A)}{1 - \Pr(B)}\right) > (1 - \eta_{ab}) \\
&\Rightarrow \quad 1 - (1 - \eta_{ab}) \cdot \left(\frac{1 - \Pr(A)}{1 - \Pr(B)}\right) < 1 - (1 - \eta_{ab}) \\
&\Rightarrow \quad \eta_{ba} < \eta_{ab}
\end{aligned}$$

If we assume $\eta_{ab} < \eta_{ba}$ then,

$$
\begin{aligned}
\eta_{ab} < \eta_{ba} \quad &\Rightarrow \quad \eta_{ab} < 1 - (1 - \eta_{ab}) \frac{\Pr(A')}{\Pr(B')} \\
&\Rightarrow \quad (1 - \eta_{ab}) \cdot \frac{\Pr(A')}{\Pr(B')} < (1 - \eta_{ab}) \\
&\Rightarrow \quad (1 - \eta_{ab}) \cdot \left(1 - \frac{\Pr(A')}{\Pr(B')}\right) > 0 \\
&\Rightarrow \quad \Pr(A') < \Pr(B') \\
&\Rightarrow \quad 1 - \Pr(A') > 1 - \Pr(B') \\
&\Rightarrow \quad \Pr(A) > \Pr(B) \\
&\Rightarrow \quad (1 - \delta_{ab}) \cdot \frac{\Pr(A)}{\Pr(B)} > (1 - \delta_{ab}) \\
&\Rightarrow \quad 1 - (1 - \delta_{ab}) \cdot \frac{\Pr(A)}{\Pr(B)} < 1 - (1 - \delta_{ab}) \\
&\Rightarrow \quad \delta_{ba} < \delta_{ab}
\end{aligned}
$$

Thus $\delta_{ab} < \delta_{ba} => \eta_{ba} < \eta_{ab}$ and $\eta_{ab} < \eta_{ba} => \delta_{ba} < \delta_{ab}$. By only comparing the values of $\delta_{ab}, \delta_{ba}, \eta_{ab}, \eta_{ba}$ the true directionality of the edges cannot be determined. In such a case, we use a third gene C seemingly regulated by both A and B to determine precedence of the drivers in algorithm as outlined in Algorithm 4, and confirm if the reverse edge needs to be pruned or not.