Gene Regulatory Networks:

Modeling, Intervention and Context

by

Michael Verdicchio

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

Approved May 2013 by the
Graduate Supervisory Committee:

Seungchan Kim, Chair
James Collofello
Gustavo Stolovitzky
Chitta Baral

ARIZONA STATE UNIVERSITY

August 2013

ABSTRACT

Biological systems are complex in many dimensions as endless transportation and communication networks all function simultaneously. Our ability to intervene within both healthy and diseased systems is tied directly to our ability to understand and model core functionality. The progress in increasingly accurate and thorough high-throughput measurement technologies has provided a deluge of data from which we may attempt to infer a representation of the true genetic regulatory system.

A gene regulatory network model, if accurate enough, may allow us to perform hypothesis testing in the form of computational experiments. Of great importance to modeling accuracy is the acknowledgment of biological contexts within the models – i.e. recognizing the heterogeneous nature of the true biological system and the data it generates. This marriage of engineering, mathematics and computer science with systems biology creates a cycle of progress between computer simulation and lab experimentation, rapidly translating interventions and treatments for patients from the bench to the bedside.

This dissertation will first discuss the landscape for modeling the biological system, explore the identification of targets for intervention in Boolean network models of biological interactions, and explore context specificity both in new graphical depictions of models embodying context-specific genomic regulation and in novel analysis approaches designed to reveal embedded contextual information. Overall, the dissertation will explore a spectrum of biological modeling with a goal towards therapeutic intervention, with both formal and informal notions of biological context, in such a way that will enable future work to have an even greater impact in terms of direct patient benefit on an individualized level.

To Jacey

You believed in me enough to marry me while I was a graduate student.

To My Family

None of this stuff matters if you don’t have your family.

ACKNOWLEDGEMENTS

I must first acknowledge God. Through this process He has helped me become so much more, both inside and outside of this work.

I met Dr. Seungchan Kim in my third year of college, where he encouraged me to begin an undergraduate research program under his mentorship. Now, nearly ten years later, we've finished the work. For his care, his candor, and for pushing me to become more, I am deeply grateful. Dr. Chitta Baral, Dr. Jim Collofello, Dr. Gustavo Stolovitzky, and Dr. Pat Langley have all provided valuable guidance and insight during the dissertation process. For their time, energy and expertise, I am also deeply grateful.

My wonderful wife Jacey, until now, has only known me as a graduate student. Yet she not only married me nearly five years ago, but she has been my best friend and greatest encourager. Finishing a Ph.D. requires a sort of herculean effort; Jacey has been on the front lines of that for years now and, without her love and care, I would never have finished.

My father Mike and my mother Kathy, throughout their 42 years of marriage, have created such a loving and supportive environment to grow up in. They taught me how to love people. They taught me not just how to work hard, but how to take joy in diligent effort. They supported me while I lived under their roof, and continue to be a great source of encouragement and support today.

My father-in-law Jay and mother-in-law Mary Lynne are not normal in the world of in-laws. After trusting me with their eldest daughter just two years into graduate school, they have continued to support, encourage and love me during an extended process that could have created mutual discomfort for those with lesser amounts of grace.

My sister Paula, my brother-in-law Tyson, and my sister-in-law Haley have been the most wonderful siblings during this process. They've all believed in me and encouraged me, which meant more than they ever knew.

My lab siblings are people with whom I locked arms and marched along this journey. Dr. Ina Sen, Dr. Archana Ramesh, Dr. Sungwon Jung, Dr. Sara Nasser, and Rob Trevino: thank you for being in the trenches, for working alongside me, and for being tremendous friends and individuals.

Among others, I also owe a debt of gratitude to Faye Navabi, Martha Vander Berg, Dr. Andy Wuensche, Dr. Dan Bryce, Dr. David Pheanis, Dr. John Moore, Dr. Lok Lew Yan Voon, Dr. George Rudolph, Dr. Shankar Banik and Dr. Deepti Joshi. To Andrew, MaryGene, Dane, and Meghan, thank you for your love, prayers and encouragement.

TABLE OF CONTENTS

Page

LIST OF TABLES xii

LIST OF FIGURES xiv

CHAPTER

1 INTRODUCTION 1

1.1 Gene Regulatory Networks 1

Modeling 1

Intervention 2

Context 3

1.2 Problem Definition 4

Challenges 4

1.3 Overview of Work 5

1.4 Outline of Dissertation 6

1.5 Individual Contributions of Dissertation 7

2 IDENTIFYING TARGETS FOR INTERVENTION BY ANALYZING BASINS OF ATTRACTION 9

2.1 Motivation 10

2.2 Related Work 11

2.3 Background 13

2.4 Methodology 14

Logic Minimization 14

Single Variable Interventions for Small Networks 17

Popularity, Term Power and Variable Power 17

Related Measures 18

Example Application 20

CHAPTER Page

2.5 Results . . . 22

Application to Systems Biology of Human Aging . . . 22

Network Construction . . . 24

Network and State Space Properties . . . 26

Top Targets for Human Aging . . . 27

Summary . . . 28

Application to the Study of the Yeast Cell Cycle . . . 29

Network and State Space Properties . . . 29

Top Targets for Yeast Cell Cycle . . . 30

Summary . . . 31

Application to WNT5A Network for Melanoma . . . 31

Network and State Space . . . 32

Top Targets for WNT5A Network . . . 33

Summary . . . 34

Related Study: Artificial Intelligence (AI) Planning . . . 34

AI Planning in the Aging Network . . . 34

AI Planning in the Yeast Network . . . 35

2.6 Chapter Summary . . . 36

3 TEMPLATE-BASED INTERVENTION IN BOOLEAN NETWORK MODELS OF BIOLOGICAL SYSTEMS . . . 39

3.1 Motivation . . . 40

3.2 Related Work . . . 41

3.3 Background . . . 41

3.4 Methodology . . . 43

Template Based Intervention Targets . . . 43

Template-Based Scores for Intervention Target Selection . . . 45

CHAPTER Page

Intervention Targets . . . 46
Illustration . . . 47
Sampling Large State Spaces . . . 48
Evaluating Intervention Success . . . 49
Simulation Study . . . 50
In-silico Network Models . . . 50
Performing Interventions . . . 51
3.5 Results . . . 51
Simulation Study for Template Methods . . . 51
The Effect of Sampling on Template Interventions . . . 52
The Effect of Increasing Network Size on Template Interventions . 55
Summary of Simulation Study . . . 56
Application to T-LGL Leukemia Network . . . 58
Network Construction . . . 58
Network and State Space Properties . . . 59
Templates Analysis . . . 61
Top Templates for Healthy and T-LGL Attractors . . . 62
3.6 Chapter Summary . . . 66
4 CONTEXT SPECIFICITY IN MODELING BIOLOGICAL INTERACTIONS . 69
4.1 Motivation . . . 69
4.2 Related Work . . . 70
4.3 Background . . . 71
Mathematical Model for Contextual Genomic Regulation . . . 71
Identifying Context Motifs . . . 72
4.4 Methodology . . . 73
Context-Specific Gene Regulatory Networks . . . 74

CHAPTER Page

Identifying Context Motif Clusters 74

Sample Association to Context Motif Clusters 76

Bayesian Network Analysis for Comparison 77

4.5 Results 77

Target Now Data Set 78

Context Motif Clusters in Target Now Data 79

Tumor Type Enrichment 81

Context Motif Clusters Identify Disease-Pertinent Genes 81

Comparison to Bayesian Network Analysis 83

4.6 Chapter Summary 84

5 CONTEXTUAL GENE SET INTERACTION NETWORKS 87

5.1 Motivation 87

5.2 Related Work 88

5.3 Background 91

5.4 Methodology 92

STEP I: Identifying Contextual Gene Sets 95

STEP II: Summarizing Contextual Gene Sets 95

STEP III: Learning Contextual Gene Set Interaction Networks 96

STEP IV: Identifying Context-Specific Network Components 96

Annotation of Gene Sets 97

5.5 Results 98

Example of Identifying Condition Specificity 98

Advantage of Contextual Gene Sets over Biclusters 99

Contextual Gene Set Interaction Network from the Refractory Cancer Gene Expression Data 100

CHAPTER Page

Contextual Gene Set Interaction Network of Refractory Cancer with Tissue Type Specificity . . . 101

Discrepancy in Biological Functions Between Cancer-Generic and Tissue-Centric Contextual Gene Sets . . . 102

Cancer-Generic Network Region . . . 104

Tissue-Specific Regions . . . 105

Contextual Gene Set Interaction Network from the GBM Data of TCGA . . 107

Contextual Gene Set Interaction Network with Phenotype/Genotype Specificity . . . 108

Functional Difference Between GBM-Generic and Subtype-Centric Contextual Gene Sets . . . 108

Comparison of Contextual Gene Sets with Other Gene Signatures of GBM Subtypes . . . 111

GBM-Generic Network Region . . . 112

GBM Subtype-Specific Regions . . . 114

Other Genotype/Phenotype-Specific Regions . . . 117

5.6 Chapter Summary . . . 117

6 CONCLUSION . . . 119

6.1 Key Contributions . . . 120

Contribution 1 . . . 120

Contribution 2 . . . 121

Contribution 3 . . . 122

Contribution 4 . . . 122

Contribution 5 . . . 124

Contribution 6 . . . 126

6.2 Future Directions . . . 126

CHAPTER Page

6.3 List of Publications 127

REFERENCES 131

APPENDIX

A BOOLEAN NETWORKS 143

A.1 Formal Description 144

State Transition Diagram 144

Attractors and Basins 145

A.2 Tools and Algorithms for Modeling and Simulating 147

Exhaustive Enumeration of Attractors and Basins 147

Sampling of Attractors and Basins 148

Template Counting Algorithm 150

B ROBUSTNESS OF SINGLE VARIABLE MEASURES 152

B.1 Problem Statement and Hypothesis 153

B.2 Experiment 153

Restrictions and Omitted Interventions 153

Types of Interventions Performed 154

B.3 Methodology of the Experiment 155

B.4 Results of Experiment 158

B.5 Conclusions 159

C OTHER NETWORK CENTRALITY AND TOPOLOGICAL MEASURES . . . 162

C.1 Influence and Sensitivity 163

C.2 Network Topology and Centrality Measures 164

Degree 164

Betweenness 165

Centroid 165

Eccentricity 166

CHAPTER Page

C.3 Topology and Centrality Measures for Four Networks 166

D T-LGL LEUKEMIA NETWORK 169

D.1 Network Simplification 170

D.2 Boolean Rules 172

D.3 Attractors 172

D.4 Top Templates Results 172

E CONTEXTUAL GENE SET INTERACTION SUPPLEMENTS 188

E.1 Proof of Theorem 189

E.2 Cholesterol Figure 190

E.3 Refractory Cancer Figures 190

E.4 GBM Figures 190

F BAYESIAN NETWORKS 196

F.1 Introduction 197

F.2 Bayesian Structure Learning 198

Scoring Functions 199

Search-and-Score 200

Constraint-Based 201

Hybrid Algorithms 201

Application to Contextual Gene Set Interaction Networks 202

LIST OF TABLES

Table Page

2.1 8-Variable Example Network Metrics and Minimized Terms 21

2.2 Abbreviations and Functions for Aging Network 25

2.3 Three Attractors with Small Network Measures for Human Aging 27

3.1 Connectivity Distribution for Random Network 50

3.2 P-Values for Two-Proportion Z-Tests 53

3.3 Summary of Four Attractor Basins for the 43-Variable Leukemia Network . . . 59

3.4 Notable Leukemia Network Templates 62

3.5 Frequent Templates are More Successful than Others 63

4.1 Target Now Sample Distribution 79

4.2 Context Motif Cluster Tumor Enrichment 82

5.1 Refractory Cancer Specific Tissue Type Interactions 102

5.2 Significant GO Terms for Refractory Cancer Data Set 105

5.3 Significant Annotations for Contextual Gene Sets 106

5.4 Specific GBM Subtype and Sample Condition Interactions 110

5.5 GBM Subtype-Centric Contextual Gene Sets 110

5.6 Comparison of GBM Subtype-Centric Contextual Gene Sets and GBM Subtype Signature Genes 113

5.7 Comparison of Contextual Gene Sets and MSigDB Gene Sets Identified with GSEA 113

B.1 Single Variable Network Measures Compared 160

C.1 Centrality Measures for the 8-Variable Random Network 166

C.2 Centrality Measures for the Human Aging Network 167

C.3 Centrality Measures for the Yeast Cell Cycle Network 167

C.4 Centrality Measures for the Wnt5a Melanoma Network 168

D.1 Boolean Rules for T-LGL Leukemia Network 174

Table Page

D.2 Attractor States for T-LGL Leukemia Network 175
D.3 Top 30 1-Templates for Leukemia Network Basin 1 176
D.4 Top 30 1-Templates for Leukemia Network Basin 2 177
D.5 Top 30 1-Templates for Leukemia Network Basin 3 178
D.6 Top 30 1-Templates for Leukemia Network Basin 4 179
D.7 Top 30 2-Templates for Leukemia Network Basin 1 180
D.8 Top 30 2-Templates for Leukemia Network Basin 2 181
D.9 Top 30 2-Templates for Leukemia Network Basin 3 182
D.10 Top 30 2-Templates for Leukemia Network Basin 4 183
D.11 Top 30 3-Templates for Leukemia Network Basin 1 184
D.12 Top 30 3-Templates for Leukemia Network Basin 2 185
D.13 Top 30 3-Templates for Leukemia Network Basin 3 186
D.14 Top 30 3-Templates for Leukemia Network Basin 4 187

LIST OF FIGURES

Figure Page

2.1 Five-Variable Example Boolean Network and State Transitions 15

2.2 Example Boolean Network Basin State Minimizations 16

2.3 8-Variable Random Boolean Network 19

2.4 Chart of Human Aging 23

2.5 Boolean Network for Aging 24

2.6 Yeast Network and Key Basin of Attraction 29

2.7 Yeast Network Terms and Small Network Measures 31

2.8 Melanoma Network 32

2.9 Melanoma Basin of Attraction Field 33

3.1 Success Rates Between Exhaustive and Sampled State Spaces 52

3.2 Success Rates for D and F Templates 57

4.1 Modeling and Inference of Context Motifs 73

4.2 Chaining Context Motifs to Form a Context-specific gene regulatory network . 75

4.3 Annotated Context Motif Clusters in a Context-specific gene regulatory network 80

4.4 Bayesian Network Analysis of TN Data 85

5.1 Schematic Overview of Learning Contextual Gene Set Interaction Networks and Identifying Condition Specificity 94

5.2 Jaccard Similarity Heat Maps of 339 Gene Sets 101

5.3 Annotated Contextual Refractory Cancer Network 103

5.4 View of *RNF11* and *RNF139* Expressions Across Refractory Cancer Samples . 107

5.5 Annotated Contextual GBM Network 109

5.6 Selected Context-Specific Regions from GBM Network 115

B.1 Small Network Robustness Experiment Results Plot 161

D.1 43-Variable T-LGL Leukemia Network 171

E.1 Boolean Network Model for the Cholesterol Regulatory Pathway 191

Figure Page

E.2 Summarized Gene Set Expression Data for the Refractory Cancer Data Set . . . 192

E.3 Regional Heat Maps from the Refractory Cancer Contextual Gene Set Interaction Network 193

E.4 The Summarized Gene Set Expression Data of GBM Samples from TCGA . . . 194

E.5 Heat Maps from the GBM Contextual Gene Set Interaction Network 195

CHAPTER 1

INTRODUCTION

Biological systems are complex in many dimensions as endless transportation and communication networks all function simultaneously. Our ability to intervene within both healthy and diseased systems is tied directly to our ability to understand and model core functionalities. The progress in increasingly accurate and thorough high-throughput measurement technologies (e.g. DNA microarray) has provided a deluge of data from which we may attempt to infer a representation of the true underlying system. Such a representation (or model), if accurate enough, may allow us to perform hypothesis testing in the form of computational experiments, including intervention and control studies.

1.1 Gene Regulatory Networks

Gene regulatory network models of biological systems are graphical structures describing the interactions of genes and/or gene products governing the transcription of messenger RNA. Many different models of gene regulation have been studied, with ultra-specific ordinary differential equation models on one end of the spectrum to the coarse Boolean model on the other end. With increasing interest in other biological entities, including proteins, cellular components, biological processes, and even drugs, these same models have also been applied to more general biological interaction networks. In our work we first spend time with a classic model for biological systems, namely the Boolean network. The work then evolves towards newer models for gene regulation and biological interaction – models which specifically incorporate the idea of biological context.

Modeling

There is no shortage of models to choose from for approximating the behavior of biological systems. Depending on whether one wishes to simply simulate behavior, make predictions and/or provide a framework for analysis, an appropriate model must be used. When model-

ing biological systems one is faced with a choice as to how to represent the steady behavior inherent in a system. Whether that steady behavior takes the form of a phenotype, a disease state, a subtype of disease, or a regulatory process robust to perturbation, we must have a way to represent these biological contexts accurately if we are to learn from it and/or use it to make novel predictions or inferences. We explore these choices with the end goal of being able to design interventions to shift us to desirable steady behavior states. This journey begins within the Boolean network model, whose steady state attractors have been show to adequately represent phenotypes of disease. The analysis in this framework, while elucidating, reveals a modeling decision: do we take an existing modeling framework, such as the Boolean formalism or perhaps Bayesian networks, model some biological system and then attempt to extrapolate contextual understanding? Or do we instead work within a modeling framework explicitly designed to represent biological contexts? The answer, of course, is "it depends", as we find tremendous insight from both sides in this dissertation.

Intervention

Before venturing into new context-based models with the end goal of designing interventions, we first explore the classic Boolean network model. The foundation for our work on intervention analyses has been laid gradually over the last few decades. Stuart Kauffman was the first to apply the Boolean Network model to biological systems, as he showed that despite the inherent simplicity, the framework was sufficient to model many complex dynamics [1]. Later, work by Andy Wuensche made special exploration of the steady states of Boolean network dynamics, known as attractors, and was able to correlate them with biological phenotype [2]. More recently the stochastic extension to Boolean networks, namely Probabilistic Boolean Networks (PBNs), was proposed by Shmulevich *et al.*; the addition of uncertainty and switching behavior to the Boolean model made this framework even more suitable for biological system modeling [3]. The PBN framework was utilized for studies on external system control [4, 5], as well as studies on intervention to avoid

undesirable states[6, 7]. With this motivation we contribute approaches for identifying key variables likely to function well as intervention targets first as single variables, second as multiple variable cohorts.

Context

Reserving formal discussion for later, the idea of a biological context is the idea that the behavior of a particular biological system alters under the effect of stimuli–perhaps from a desirable steady state of health to an undesirable steady state of disease. Interest in modeling this behavior has been steadily increasing with our understanding of the biological system and with the advances in high-throughput measurement technology. Initial attempts at modeling this behavior naturally started with existing models; the switch-like behavior of the PBNs mentioned above as well as the basins of attraction in Boolean networks can be interpreted with contextual implications. However the desire to make contextual interpretations based on biological modeling is not limited to discrete dynamics. Work by Segal [8] used existing knowledge sources as well as clustering to find sets of genes that are either functionally co-related or expressed, which can lead to the idea of a context. Using another model, Bayesian networks[9], Grzegorczyk *et al.*proposed a way to learn non-homogeneous Bayesian networks that can represent multiple conditions[10], which fits with the idea of context as well. Other recent works show interest in identifying patterns specific to certain biological contexts [11, 12], in relationships between sets of genes and biological/clinical characterization of data samples [13, 14], and in identifying context-specific activity of known biological networks [15, 16, 17, 18]. As these context-specific models stabilize in the coming years, we will be able to learn from our previous path of intervention studies on classical models and design new intervention strategies to take advantage of this new perspective of context specificity. With this motivation we contribute both graphical depictions of a context-specific mathematical model for genomic regulation as well as analysis approaches detailing how to identify significant behavior in the system.

1.2 Problem Definition

This section formulates the problems of modeling, intervention and context in gene regulatory networks. We attempt to answer the following questions:

- "Can we identify targets for intervention to change or preserve biological steady states (i.e. contexts) within the classic framework of Boolean networks?"
- "Can we extend the idea of a gene regulatory network to explicitly model context specificity?"
- "Can we provide an analysis approach for context-modeling networks such that we can identify biologically meaningful content?"

Challenges

To answer the above questions we must overcome the following challenges:

1. Navigate the exponential state space of the Boolean network model
    a) We must design measures and analysis methodologies which reveal the key variables in a Boolean network through analysis of the basins of attraction. Oft neglected in favor of topological or attractor-only analysis, the potentially immense basins of attraction can reveal additional insight.
    b) Because the size of the basin of attraction field quickly becomes too large to process in full, we must design further measures and analysis methodologies designed to work despite partial information concerning the basins of attraction.
2. Create graphical depictions of models for contextual genomic regulation
    a) Significant work has been put into a powerful mathematical model for contextual genomic regulation, however the results are not in network form and thus

cannot benefit from network-based approaches. We must first find a way to represent the idea of a biological context graphically.

b) Once a graphical depiction of a contextual model is established, we must ensure the analysis capability of the model and refine or extend the representation as necessary.

3. Design methods for extracting biologically meaningful contextual information

a) With a graphical depiction of a model for contextual genomic regulation, we must define a methodology to stratify the important information contained in the representation.

b) Due to the widespread availability of curated biological knowledge, we must define a methodology to incorporate this plethora of additional insight into our models to further elucidate contextual information.

1.3 Overview of Work

Presented here is a brief outline of the tasks focused on within this dissertation to address the challenges and solve the problem questions.

1. *Identify targets for intervention in Boolean networks*: Exploit the logical nature of the Boolean network states to reduce complexity and form minimized representations of attractor basins. Define measures to identify key variables from minimized representations. Demonstrate how key variables have significant network influence and may function well as targets for intervention. Extend methodology to larger networks for which complete knowledge of the basins of attraction is not possible. Demonstrate methodology on synthetic networks. Apply methodology to a well-studied biological network.

2. *Create graphical depictions of models for contextual genomic regulation*: Extend existing mathematical model for contextual genomic regulation into a graphical representation. Apply the model to real networks based on high-throughput biological data. Check capabilities against related approaches. Refine and extend the model as needed to ensure feasibility of analysis.

3. *Provide an analysis approach for context-modeling networks to identify biologically meaningful content*: Design an analysis approach for new graphical depictions of models of contextual genomic regulation. Identify biologically meaningful contextual information. Validate findings applied to real networks based on high-throughput biological data. Compare findings to related approaches.

1.4 Outline of Dissertation

This dissertation is divided into four main content chapters, as well as several supplementing appendices. The first question of identifying targets for intervention in Boolean networks is contained within Chapter 2 and Chapter 3, as well as their referenced appendices. Within Chapter 2, the methodology section refers to the minimization of our representations of basins of attraction and describes new single variable measures for small networks. The results section details application of the methodology to various biological systems, including human aging, the yeast cell cycle, melanoma, leukemia, and related work in artificial intelligence planning. Within Chapter 3, the methodology section details the extension of the ideas in the previous chapter to a multivalued intervention target called a template. It goes on to define intervention templates formally and give an algorithm for their discovery, measures to quantify the top performing templates, and the setup for a simulation experiment to validate their robustness. The results section details the results and implications of the template simulation experiment and demonstrates application of the contribution on a large T-LGL Leukemia network.

The second and third questions are each addressed within two chapters. Chapter 4 describes our initial work extending the mathematical model of contextual genomic regulation into a graphical depiction. It first covers the requisite background information regarding the model, followed by a detailed description of the methodology for creating the network. In the results section we apply our work to a biological data set for cancer and compare our approach and findings with related techniques.

Chapter 5 again answers the second and third questions, but with the next generation of our work in Chapter 4. While the first graphical depiction of the contextual model for genomic regulation resulted in significant findings, the analysis approach was complex. Building upon the work established in Chapter 4, Chapter 5 extends to a new graph-based model designed to model contextual behavior, specifically subtypes of disease. The methodology section details the technical approach while the results section demonstrates our application to two cancer networks.

Chapter 6 summarizes the results and concludes the dissertation, however supplementary information referenced in the chapters is included within the appendices.

1.5 Individual Contributions of Dissertation

As with all research, the work in this dissertation was collaborative. To clarify and delineate the individual contributions of the author, the following list of individual contributions is given for convenience. The works are described as a whole to facilitate understanding, however this reference identifies exactly which aspects of the works presented are the contributions of the author.

- New method for identifying targets for intervention by analyzing Boolean network basins of attraction (Chapter 2)

- New method for identifying multi-variable targets for intervention in large Boolean networks requiring a sampled view of intractable state spaces (Chapter 3)
- New graphical depiction of a mathematical model for contextual genomic regulation (Chapter 4)
- New *ad hoc* analysis approach for the graphical depiction of contextual genomic regulation to identify biologically meaningful context clusters (Chapter 4)
- New graphical depiction of a context-based model: a contextual gene set interaction network, specifically STEP III and STEP IV of the approach concerning Bayesian network learning and identifying condition specificity (Chapter 5)
- New knowledge-based annotation and validation approach for contextual gene set interaction networks to reveal contextual information embedded within (Chapter 5)

CHAPTER 2

IDENTIFYING TARGETS FOR INTERVENTION BY ANALYZING BASINS OF ATTRACTION

A grand challenge in the modeling of biological systems is the identification of key variables which can act as targets for intervention. Boolean networks are among the simplest of models, yet they have been shown to adequately model many of the complex dynamics of biological systems.

In this chapter we explore a novel analysis approach based on the Boolean network model for biological systems (see Appendix A for a formal description of Boolean networks). With their discretization of variable behavior to ON/OFF or 1/0 values, Boolean networks are a coarse model, yet their usefulness as models for various systems, especially biological systems, has been well established.

Our work shows that logic minimization (i.e. classical circuit design) of the collections of states in Boolean network basins of attraction reveals key players in the network. Furthermore, we demonstrate that the key players identified by this method are often excellent targets for intervention given a network modeling a biological system. More importantly, we show that the key players identified are not apparent from the attractor states alone, from existing Boolean network measures, or from other common network measurements. We demonstrate these claims by applying our method to a well-studied yeast cell cycle network, to a WNT5A network for melanoma computationally predicted from gene expression data, and to a model for human aging.

With our modeling groundwork laid at the Boolean end of the spectrum, we will have a useful motivation and reference point for the more complex techniques and models described in later chapters.

2.1 Motivation

The very nature of medicine is to know when and how to intervene in order to shift the steady behavior of a system to a more desirable state[19]. Ideally such interventions would be as minimally damaging as possible; however we know that especially with diseases such as cancer, interventions like chemotherapy are anything but minimal. In the path towards personalized medicine and individualized treatments with minimal collateral damage, designing and studying interventions that take advantage of our systems-level understanding and available data is and will remain of paramount importance, as working with computational models allows us to perform tests, execute simulations, and make predictions in inexpensive ways that require no human subjects[20].

Biological systems are complex in many dimensions as endless transportation and communication networks all function simultaneously[2]. Despite its simplicity, the Boolean network model has proven to be quite viable at approximating certain aspects of biological processes[19]. For example, it has been used to simulate the yeast cell cycle [21], which we looked at closely in our work [22]. It has also been used to simulate the expression pattern of segment polarity genes in *Drosophila melanogaster* [23], as well as the vocal communication system of the songbird brain [24, 25]. Since Kauffman's seminal work[1] there have been countless variations and extensions of the use of Boolean networks for modeling biological systems, and various inference procedures have been proposed for them [26, 27, 28].

Yet despite the great progress that has been made in Boolean network modeling of biological systems, much of the analysis has been focused on network topology, or on the attractor states themselves. Often neglected, however, are the exponentially sized basins of attraction. We wish to determine what can be elucidated from these parts of the state space in terms of identifying targets for intervention.

2.2 Related Work

Within the world of *in silico* modeling and intervention studies, significant groundwork has been laid. Boolean networks allow modeling at the most simplified extreme of the spectrum due to their coarse discretization of values to 0 and 1 and their simplified, rule-based update mechanism, yet have still been shown to adequately model complex behaviors seen in the biological system, as previously mentioned. In Appendix A we give formal descriptions of Boolean networks, the basin of attraction field they generate, as well as pertinent algorithms. Over thirty years after Kauffman's seminal work [1], Shmulevich *et al.*[29] pioneered work on a stochastic extension to the model called probabilistic Boolean networks (PBNs), which share the rule-based nature of Boolean networks but also handle uncertainty well. Within this extended framework of PBNs, studies were performed by Datta *et al.*[4, 5], which focused on external system control; studies by Pal *et al.*[6] and Choudhary *et al.*[7] explored intervention in PBNs to avoid undesirable states.

One major challenge in using Boolean networks is the exponential growth of the basin of attraction field with the linear growth of the number of variables, prompting others to work in the Boolean framework itself to achieve some kind of improvement. The reader is referred to Appendix A for formal descriptions of basins of attraction, also called the state space or state transition diagram. The approach of Richardson [30] attempted to shrink the size of the state space through the careful removal of "frozen nodes" and network leaf nodes. The smaller state space then lent itself more readily to the discovery of attractors and basins by sampling methods. Dubrova *et al.*[31] explored properties of random Boolean networks, particularly their robustness in the face of topological changes and the removal of "redundant vertices", thus shrinking the state space. While effective in shrinking the state space and removing extraneous nodes, neither of these methods is looking for key

players in a system or possible intervention targets; in fact both methods have the chance of eliminating such variables.

Willadsen and Wiles [32] reduce Boolean network state space by creating what they call *schemas*. Using a ternary representation with ones, zeros and wildcards similar to the don't-cares of logic minimization, they are able to create an abstract representation of Boolean network basins of attraction, which they use to quantify dynamics and robustness. These schemas provide the authors with a convenient way of representing groups of related, neighboring states as they compute a state space robustness metric called *structural coherency*. While powerful in exploring relationships between state space structure and robustness in random Boolean networks of up to a couple dozen variables, the approach is not intended to identify standout variables that can function as intervention targets.

In an attempt to achieve certain analysis goals, Saez *et al.*[33] as well as Schlatter *et al.*[34] convert their Boolean models of biological systems into hypergraphs, which generalize graphs with edges connecting sets of vertices instead of just pairs or singletons, thus lending themselves to representing Boolean functions. Both papers use analysis techniques to identify important pathways, network motifs and feedback loops. The work of Schlatter *et al.*also mentions the discovery of relevant hubs in the network. Steggles *et al.*[35] convert to a different graphical structure, namely Petri nets. In making this conversion, they use the logic minimization technique we employ, albeit in a different way. While each of these methods achieves some kind of improvement on what we can see with a classical Boolean network, like PBNs, each is forced to add to the Boolean formalism and translate to another framework where analysis can be performed.

Maji [36] does not use a Boolean network but nonetheless works with the notion of state transition using a related discrete model: fuzzy cellular automata. His work uses multi-valued logic and presents a new way of identifying attractor basins; however it does

not focus on the identification of intervention targets in the system. Mar *et al.*[37] also employ the notion of a state transition space without the direct use of a Boolean network. Using their regression model they strive to classify core variables (genes in their case) as they decompose state space trajectories. Their method, however, is dependent on time-course data, and furthermore its primary focus is at the pathway level and not the variable level.

Wuensche [2] and others also have studied the basins of attraction in traditional Boolean network models of genomic regulation, specifically the relationship of their structures to the stability of attractors (cell types) in the face of perturbations. Despite the usefulness of basins of attraction in biological modeling, they are often neglected in favor of the attractor states when it comes to analysis due to their transient nature and size complexity. Despite the neglect concerning the basins of attraction themselves, as they are collections of states leading into corresponding attractors, i.e. phenotype, careful analysis of these basins could reveal interesting biological characteristics that determine cell fate. Furthermore, such analysis would not require the translation of the Boolean network into any other type of model; In this section we describe a new analysis approach to using the Boolean network model in identifying biological targets for intervention through the basins of attraction.

2.3 Background

A Boolean network $\mathbf{B}(V,\mathbf{f})$ is made of a set of binary nodes V, and a set of functions $\mathbf{f}$, where the binary value of each variable at time $(t+1)$ is determined by the status of other variables at time (t) by means of its personal Boolean update function. The wiring of function input variables to output variables creates a "wiring diagram", or network topology, and the progression of each variable's value as Boolean update functions are applied creates the state transition dynamics of the network. These elements together allow Boolean networks to be used as discrete dynamical models of the biological system.

Formal definitions and formulas governing the Boolean network formalism, as well as the algorithm for enumerating its state transition dynamics, are given in full in Appendix A.

2.4 Methodology

Our approach employs a logic minimization algorithm to reduce the Boolean states comprising Boolean network basins of attraction to minimal representations, and it is from these minimizations that we identify single-variable intervention targets. We first describe logic minimization and how it relates to the Boolean network formalism. Next we introduce measures for small networks to identify single variable intervention targets from the results of the logic minimization; we demonstrate the usefulness of these measures on 5-variable and 8-variable random networks. Once our measures and analysis approach are established, in the Results section we make application to several studies, including human aging, the yeast cell cycle, a melanoma cancer study, a related work involving artificial intelligence planning, and on a large leukemia network.

Logic Minimization

Logic minimization (or reduction) is a classic problem from digital circuit design employed to reduce the number of actual logic gates needed to implement a given function [38]. With careful logic minimization one can reduce the number of gates required and thus include more functionality on a single chip. Minimization identifies variables which have no influence on the outcome of a function and marks them appropriately as a *don't-care*. As a simple example, we take the Boolean function: $(A \wedge B) \vee (\neg A \wedge B)$ (2 signals, 4 gates). Since the role of A changes while B remains *ON* with the same output, it is clear to see that the only influencing variable is B, which can be given with just that signal itself (0 gates).

In this study, we employ the Espresso tool [39], which is a heuristic logic minimizer designed to efficiently reduce logic complexity even for large problems. We supply as

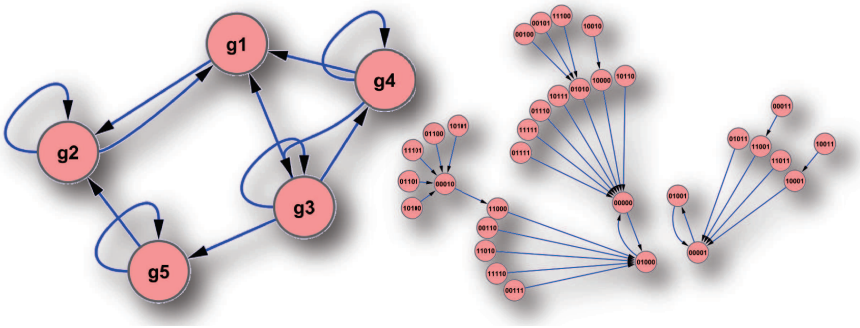


Figure 2.1: Five-Variable Example Boolean Network and State Transitions: On the left is the network topology and Boolean function inputs for each variable. On the right are the 32 states comprising the state transition diagram partitioned into two attractor basins.

input the set of states in a particular basin of attraction B_i (the complete state space is comprised entirely of all basins (i.e. $\bigcup_i B_i = S$)); this input comprises the *ON-cover* (or truth table) in disjunctive normal form (DNF) for a Boolean function whose output is *ON* for the states of B_i ($\{\mathbf{x}_{i_1} \vee \mathbf{x}_{i_2} \vee \cdots \vee \mathbf{x}_{i_M}\} \mapsto ON$) and whose output is *OFF* for the states of $S \setminus B_i$. Espresso analyzes this cover and returns a minimal (though not necessarily unique) DNF set comprised of one or more terms, denoted $T_i = \{\mathbf{t}_{i_1}, \mathbf{t}_{i_2}, \cdots, \mathbf{t}_{i_N}\}$, where $N \leq M$. These $\mathbf{t}_i$ have some variables set to *ON* (denoted 1), some set to *OFF* (denoted 0), and some set as *don't-care* (denoted "-"). The presence of these don't-care variables in some terms is what allows the reduction.

To demonstrate logic reduction on an example related to our methodology, consider a 5-variable Boolean network (Figure 2.1, left). We exhaustively enumerate its state space following Algorithm A.1, which partitions the state space into two basins of attraction, each leading to cyclic attractors of length 2 (Figure 2.1, right). The larger basin (basin 1) is made up of 24 of the 32 states and the smaller basin (basin 2) contains the remaining 8 states. Logic minimization reduces the 24- and 8-state basins to only 2 and 1 term(s), respectively (Figure 2.2). In Figure 2.2, the dashes in the reduced terms represent don't-cares, and the bits in the bit strings represent the states of the five variables in the order g_1, g_2, g_3, g_4, g_5.

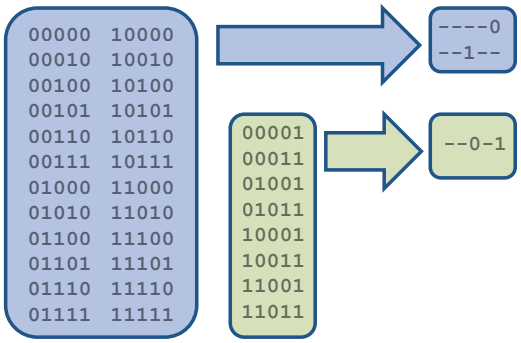


Figure 2.2: Example Boolean Network Basin State Minimizations: On the left are two boxes containing each state in the two basins shown on the right of Fig. 2.1. On the right are the terms produced from applying logic minimization to each basin. The larger basin is reduced to two terms with one value each (OR relationship between the variables) and the smaller basin is reduced to one term with two variables set (AND relationship between the variables).

Inspection of reduced terms immediately identifies the importance of g_3 and g_5. To demonstrate their utility as targets for intervention, we assume that one of our two attractors is some kind of undesirable state (e.g. the presence of tumors) and that the other attractor is some kind of desirable state (e.g. the absence of tumors). Thus, if we are in an undesirable basin of attraction we would be interested in the minimal intervention needed to "jump" to the desirable one. The variables identified by our method give us precisely those targets.

To see this, let us fit our previous hypothetical with g_3 as a tumor suppressor gene and g_5 as an oncogene, and let basin 1 be the desirable, healthy steady state and basin 2 be undesirable. Take the first of the two terms in basin 1: `[----0]`. What this tells us is that if we take any state in either basin and negate g_5 (i.e. suppress the oncogene), we will end up in a state somewhere in basin 1, which will then lead us to the corresponding desirable steady state attractor. A similar trigger is found with the second term: `[--1--]` (i.e. activating the tumor suppressor gene). The term `[--0-1]` tells us that to jump from anywhere to the undesirable basin 2 we need to negate g_3 and activate g_5 simultaneously (i.e. inhibit the tumor suppressor gene and activate the oncogene). While intuitive, this be-

havior could not be found by simple topological inspection, or by studying either attractors themselves. For example, ignoring self-regulation, the two variables with the highest number of incident edges in the original Boolean network are g_1 and g_3. We have reported on g_3's importance in the network through our method, and this seems to be corroborated by its large degree. However our method has revealed the relative unimportance of g_1, which is given the same ranking as g_3 by topological considerations. So while topological inspection may lead is in the right direction, it may also lead us astray. Regarding the attractor states themselves, we first cannot know in advance what the attractors will look like. If an attractor analysis depended on point attractors, or long-cycle attractors with many states, it could never count on seeing those types of attractors every time. In the case of our toy example, the two attractors are given as $A_1 = (\texttt{[00000]}, \texttt{[01000]})$ and $A_2 = (\texttt{[00001]}, \texttt{[01001]})$. We may be able to notice the abnormal behavior of g_2, though it is actually not significant, and we may be able to recognize g_5 as a differentiator, though it would be difficult to draw such a conclusion based on attractor states alone. Like topological inspection, attractor state analysis may reveal some items of interest from time to time, but it may also be misleading.

Single Variable Interventions for Small Networks

After applying logic minimization to a set of Boolean functions one is left with a minimal DNF representation comprised of a set of terms containing ones, zeros, and don't-cares. We have shown how to identify important variables in a very small example, but a more formalized method is needed to identify key variables and possible targets for intervention from the minimized terms in larger problems.

Popularity, Term Power and Variable Power

To this end we introduce three simple measures. The first is to measure how frequently a variable *v* is required to be *ON* or *OFF* across different terms, called *Popularity* (*p*), and is

defined below, where x represents the number of times v is set in T_i and y represents the total number of terms in T_i. Next, we define terms which are powerful due to the combinatorial effect of their few set variables (e.g. if a five-variable term has one variable set and four listed as don't-cares, that one set variable controls 16 (2^4) configurations covered by the don't-care variables (half of the state space)). This term would be more powerful than a term with two variables set and three don't-cares, which covers only 8 configurations (one-fourth of the state space). *Term Power* (P_T) is defined below, where a is the number of variables set in a term ($\mathbf{t}$) and b is the number of variables in ($\mathbf{t}$). Given the notion of term power, one can also consider variables which preside over powerful terms to be potentially important and powerful intervention targets. *Variable Power* (P_V) is defined below as the average term power over the terms in which v is explicitly configured, where n is the number of terms where v is set, y is the total number of terms in T_i, and v is the variable in question.

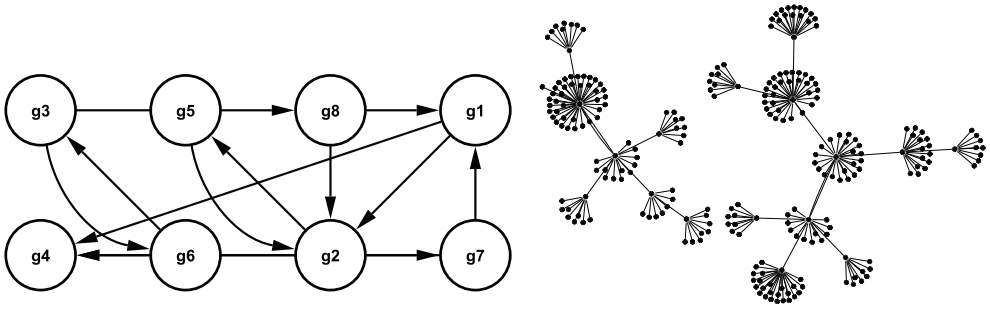
$$p(v) = x/y, \tag{2.1}$$

$$P_T(\mathbf{t}) = 1 - a/b, \tag{2.2}$$

$$P_V(v) = \sum_{i=1}^{y} P_T(\mathbf{t_i}) | (v \textit{ is set in } \mathbf{t_i}), n \geq 1, (\textit{otherwise } 0), \tag{2.3}$$

Related Measures

There are various network centrality measures often used in network studies, particularly concerning biological networks, to identify important variables. We have already touched on the degree of a node, but we also consider, for comparison, the network centrality measures of *betweenness*, *centroid value*, and *eccentricity*. High betweenness indicates that a variable is crucial in maintaining connections between other variables. The centroid value for a variable provides a weighted centrality index. A high eccentricity measure indicates that all other nodes are in proximity. Full definitions as well as biological explanations



(a) Network Topology (b) Basin of Attraction Field

Figure 2.3: 8-Variable Random Boolean Network: On the left we see the network topology and Boolean function inputs; on the right we see the state transition diagram partitioned into two attractor basins of sizes 160 and 96, respectively.

can be found in Appendix C, as well as in the supplementary information of Scardoni *et al.*[40], but in short, network nodes with the most extreme values for these measures can be correlated with biologically significant nodes, possibly even intervention targets.

For the Boolean network formalism, there are also variable-specific measures known as *Influence* and *Sensitivity* for a variable x_i, denoted $r(x_i)$ and $s(x_i)$, respectively. The reader is referred to Shmulevich *et al.*[29] for formal definitions and Appendix C for further discussion. In short, in biological Boolean networks, variables with high influence have the potential to regulate the dynamics of the network, and so they are of interest to this study. Sensitivity represents the degree to which a variable is affected by other variables, and so of the most interest are variables with the highest influence and the lowest sensitivity. Since our measures p and P_V are specific to each basin, this presents an unfair advantage over the network-generality of $r(x_i)$ and $s(x_i)$. Thus, we extend the measures shown in Shmulevich *et al.* [29] to be specific to a particular basin of attraction by manipulating the joint probability distribution of the state space; while Shmulevich *et al.*assign a uniform probability to all states, we simply assign a zero probability to any state not in the basin and assign a uniform probability to states within.

Example Application

In order to demonstrate the use of our single variable measures and also to reinforce our claim regarding the utility of our methodology over other measures, we offer larger example random network. We create the 8-variable network shown in Fig. 2.3(a) in which we assign one, two or three random inputs to random Boolean functions for a balance of simplicity and dynamic behavior. This network was simulated using Algorithm A.1 and resulted in two basins of attraction, shown in Fig. 2.3(b). Basin 1 included 160 states converging on a cyclic attractor of length two (`[01011101]` and `[11011100]`), and Basin 2 included the remaining 96 states, which converged on another cyclic attractor of length two (`[00011100]` and `[11011101]`). After applying logic reduction, the 160 states in Basin 1 were reduced to a set of three terms, and the 96 states of Basin 2 were reduced to a set of four terms: $T_1 = \{$`[0-----00]` $\vee$ `[1-----10]` $\vee$ `[1-----01]`$\}$, $T_2 = \{$`[1-----00]` $\vee$ `[0-----1-]` $\vee$ `[0------1]` $\vee$ `[------11]`$\}$.

After analysis with the measures defined in the previous section, we find, based on high p and P_V, g_1, g_7 and g_8 to be of interest. Because each of g_1, g_7 and g_8 are explicitly configured in each of three terms for the larger basin and in 3 out of 4 terms in the smaller basin, their scores for p and P_V are each identical and overshadow the remaining variables. These results can be seen in Table 2.1. In this example, we again observe that simply identifying vertices in the graph with high degree does not necessarily reveal important variables. With self-loops removed to prevent inflation of degree counts, the variables with the highest degree are g_2 with six incident edges and g_1 with four. From our analysis, g_1 is one of the most important variables. However the variable with the highest degree, g_2, has been shown to have no influence at all in our analysis. When the network centrality measures of betweenness, centroid value, eccentricity and node degree are calculated for this toy network (Appendix C), we find that g_8 is frequently reported with high scores, just

Table 2.1: 8-Variable Example Network Metrics and Minimized Terms: In this table we observe several measures across the eight variables. The narrow colored boxes represent the minimized terms from which the first measures are calculated. In those green indicates 1, red indicates 0, and yellow indicates a don't-care value. The first measures shown are popularity and variable power, and the later measures are influence and sensitivity. While influence and sensitivity reported g_1, they sometimes missed or incorrectly reported others, such as g_4, g_7 and g_8.

	Basin 1 (Larger)			Basin 2 (Smaller)				
Variables		$p(v)$	$P_V(v)$		$p(v)$	$P_V(v)$	$r(x_i)$	$s(x_i)$
g_1		1.00	0.63		0.75	0.71	1.75	1.75
g_2		0.00	0.00		0.00	0.00	1	1
g_3		0.00	0.00		0.00	0.00	0	0
g_4		0.00	0.00		0.00	0.00	1.5	0
g_5		0.00	0.00		0.00	0.00	1	0.5
g_6		0.00	0.00		0.00	0.00	0	0
g_7		1.00	0.63		0.75	0.71	0	1.25
g_8		1.00	0.63		0.75	0.71	1	1.75

like our approach. $r(x_i)$, in fact, identifies g_1, g_7 and g_8 as important, which match our three best. However, several of the measures, including $s(x_i)$, incorrectly dismiss g_7, and many measures also elevate g_2, which is shown mathematically to have no real intervention capabilities. Other network centrality and topology measures and results are also discussed for this network in Appendix C.

Thus for networks of a size manageable enough to exhaustively enumerate the states space with Algorithm A.1, popularity and variable power can be used to identify key variables which make excellent candidate intervention targets. In Appendix B we perform a simulation based experiment to identify the best single variable measures between popularity, power, and those referenced in Appendix C. We found that popularity, power, and their combination in the form of an harmonic mean, are consistently the top performing single variable measures. As we will see in the coming section, larger networks present some problems for these measures and require a different yet related approach.

The contributed analysis approach for identifying targets for intervention in Boolean network models has many opportunities for insight and exploits the combinatoric nature of the state space of Boolean networks while adapting to their size. First, logic minimization is used to eliminate unimportant variables and form minimal representations including don't-care values, which are promptly exploited by the contributed measures. We identify variables overlooked or incorrectly reported by topological measures, centrality measures, or attractor state analysis. Additionally we are able to associate this standout behavior with specific values and specific attractor basins, telling us not just what to intervene upon, but precisely how and precisely where.

2.5 Results

In this section we summarize the results of applying our analysis approach on a variety of networks, from a hand-made model for the systems biology of aging, to a hand-made gene regulatory network for the yeast cell cycle, to a computationally-predicted gene regulatory network for a WNT5A melanoma network. We also summarize the validation and usefulness of our findings in our related study which uses AI search techniques to design intervention plans.

Application to Systems Biology of Human Aging

The systems biology of human aging is a complex, quantitative process. Many theories regarding senescence involve the roles of cellular components, such as mitochondria and lysosomes, as well the transportation and accumulation of various entities. In recent years, work by Furber[41] has amalgamated the research of many prominent biogerontologists into a large chart illustrating many of the leading theories on human aging. The chart is organized into cellular components and describes many intricate processes qualitatively, along with their input and output entities. The chart is shown in Figure 2.4. Using this network we construct and simulate a Boolean network and apply our small networks analysis approach to the basins of attraction.

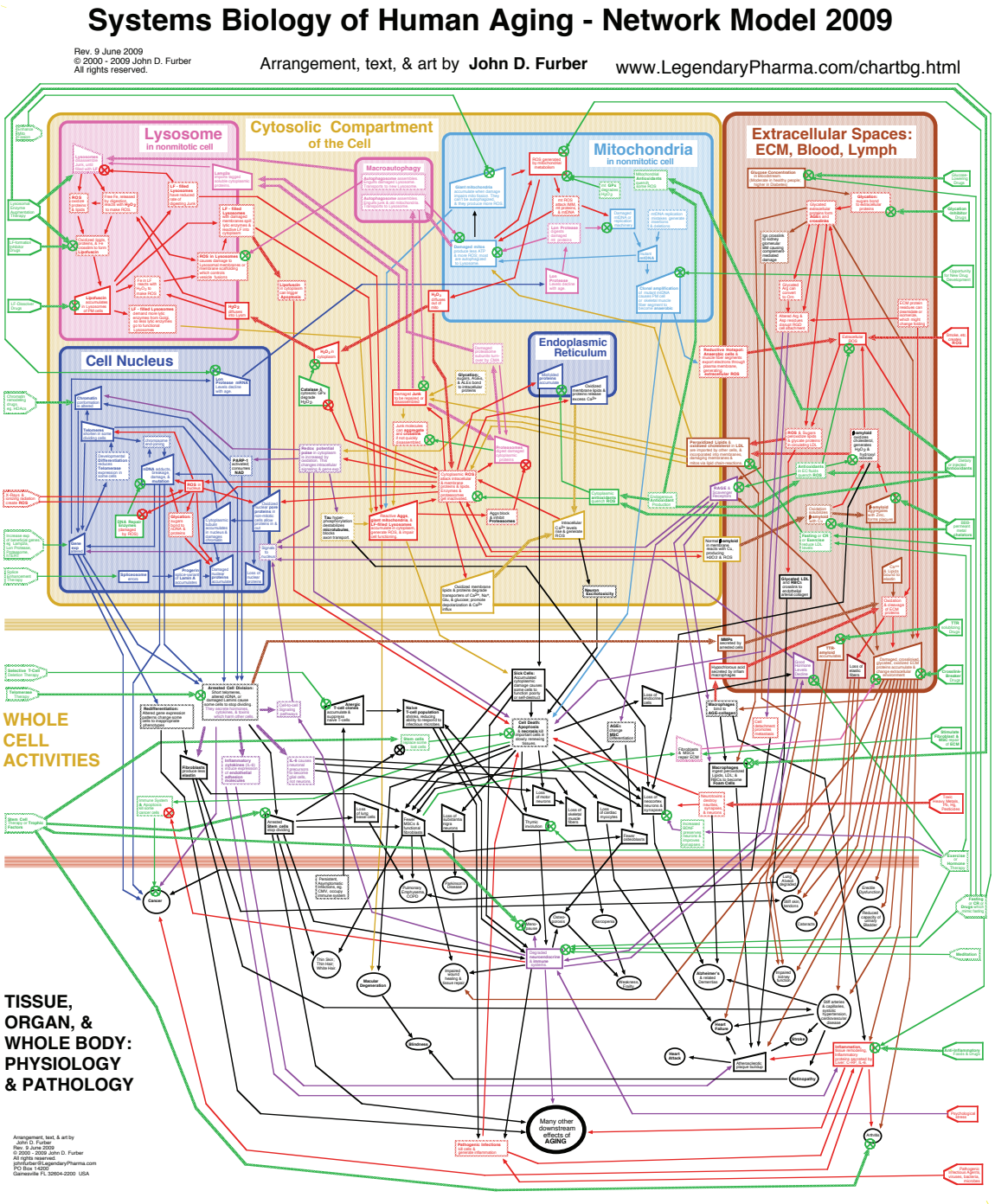


Figure 2.4: Chart of Human Aging: Original Image Copyright ©2010 John D. Furber. All rights reserved. Used by permission.

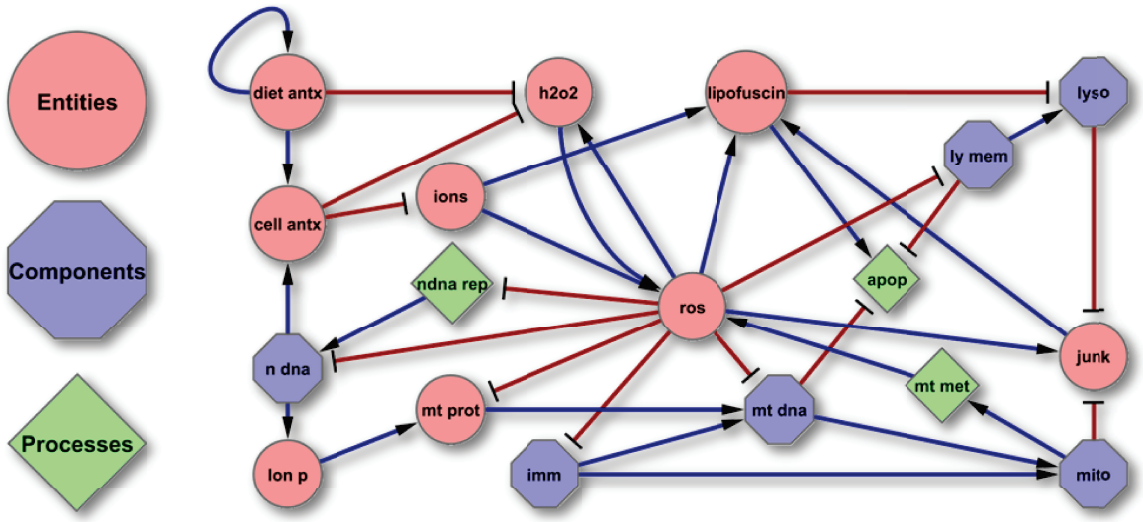


Figure 2.5: Boolean Network for Aging: On the left we have a legend for the shapes of the nodes in the connected network, with circles for biological entities, hexagons for cellular components, and diamonds for biological processes. Boolean function input values are represented by the edge endpoint shapes, with arrowheads representing logical 1 and inhibition marks representing logical 0. The actual functions are shown in Table 2.2.

Network Construction

For this study we chose to mainly model the mitochondria and lysosome cellular components along with a few involved entities in the cytoplasm and nucleus. First the aforementioned sections of the chart were examined for essential entities and their involved processes, each of which constituted a variable in the Boolean network representation. For our study we ended up with 18 variables. The 18-variable Boolean network can be seen in Fig. 2.5.

Once variables were formed, Boolean functions had to be determined from information encoded in the chart. A Boolean conversion of the chart representation, which describes events at several different levels of granularity at a qualitative level, is difficult since entities and processes must be distilled down to very simplified interpretations. To simplify interpretation, a Boolean state of 1, or *ON*, was designated as “present”, “healthy”,

Table 2.2: Abbreviations and Functions for Aging Network. Note that $\wedge$, $\vee$ and $\neg$ represent logical AND, OR and NOT, respectively.

Full Name	Short	"Good"	Function
apoptosis	apop	0	$\neg$mtdna $\vee$ ($\neg$lymem $\wedge$ lipo)
lysosome membrane	ly mem	1	$\neg$ros
lon protease	lon p	1	ndna
mito. proteins	mt prot	1	lp $\vee$ $\neg$ros
mito. dna	mt dna	1	(mt prot $\wedge$ imm) $\vee$ $\neg$ros
inner mito. membrane	imm	1	$\neg$ros
mitochondria	mito	1	mt dna $\vee$ imm
lipofuscin	lipofuscin	0	ros $\wedge$ (junk $\vee$ ions)
metal ions	ions	0	$\neg$cell antiox
reactive oxygen species	ros	0	mt met $\vee$ (h2o2 $\wedge$ ions)
mito. metabolism	mt met	1	mito
nucleus dna	ndna	1	ndna rep $\vee$ $\neg$ros
cellular antioxidants	cell antx	1	ndna $\vee$ diet antx
hydrogen peroxide	h2o2	0	(ros $\wedge$ (cell antx $\vee$ diet antx))
dietary antioxidants	diet antx	1	self regulated
junk	junk	0	ros $\wedge$ ($\neg$lyso $\vee$ $\neg$mito)
nucleus dna repair	ndna rep	1	$\neg$ros
lysosome	lyso	1	lymem $\vee$ $\neg$lipofuscin

"functioning", "intact", etc., while a Boolean state of 0, or *OFF*, was designated as "absent", "damaged", etc. It should be noted that *ON* does not always correspond to "good" and *OFF* to "bad"; there are many undesirable entities and processes which are "good" when they are *OFF* and vice versa. Whether an entity, process or component is desirable or not depends on the context of the study. For example, apoptosis is a necessary process to maintain healthy cells. However, in this particular context apoptosis is taken to mean the premature death of a damaged cell unable to repair itself, and so it is considered undesirable in this model's context. Based on these designations, Boolean functions were determined for the state of each variable, with some assumptions made for the effects of combinatorial inputs (in this study up to three inputs per variable) when the chart's descriptions did not give enough direct information. The functions assigned are shown in Table 2.2, along with which state of the variable is desirable.

Network and State Space Properties

The encoding of the Boolean network was then processed in MATLAB [42], producing the state transition matrix for all 262,144 states, leading to three attractors–one point attractor (fed by a basin with 131,072 states), a cycle of 9 states (basin size 114,176), and a cycle of 3 states (basin size 16,896). Each attractor was scored based on how many of its variables were found to be in a desirable state. In this way, a "degree of health" was assigned to each attractor.

The first and largest attractor had the singleton (point attractor) with a basin size of 131,072 states. The attractor was interesting in that the state of every variable was undesirable; thus, this attractor could represent the worst case of aging, namely death. Unfortunately it is less interesting to study such an attractor as there is not much intervention to do at that point. The third attractor was a cycle of three states, fed by the smallest basin of size 16,896, which represented a stable cycle of somewhat hindered health. The second attractor was a cycle of nine states, fed by a basin of size 114,176 which, like the third attractor, represented a stable cycle of relative health. Unlike the third attractor, the "good" states in this one were better and the "bad" states were not as bad. Thus we took this attractor to represent a cycle of relative health and one the most desirable to be in. We were able to minimize the 131,072 states of the first basin to just one single term. We minimized the 16,896 states of the third basin to just 12 terms, and we minimized the 114,176 states of the first basin to just 12 terms also. For the minimized basins of attraction we calculate our small network measures of variable popularity and power for each variable. These values and the attractor state classifications can be seen in Table 2.3. Other network centrality and topology measures and results are discussed for this network in Appendix C.

Table 2.3: Three Attractors with Small Network Measures for Human Aging: Good and Bad columns are given for visual reference to the attractor states (since 0/1 values differ).

			Singleton			Cycle (3)			Cycle (9)	
Variables	Good	Bad	$p(v)$	$P_V(v)$		$p(v)$	$P_V(v)$		$p(v)$	$P_V(v)$
apop	0	1	0.00	0.00		0.00	0.00		0.00	0.00
ly mem	1	0	0.00	0.00		0.00	0.00		0.00	0.00
lon p	1	0	0.00	0.00		0.00	0.00		0.00	0.00
mt prot	1	0	0.00	0.00		0.50	0.60		0.33	0.76
mt dna	1	0	0.00	0.00		0.50	0.62		0.25	0.76
imm	1	0	0.00	0.00		1.00	0.61		0.58	0.76
mito	1	0	0.00	0.00		0.67	0.61		0.33	0.76
lipofuscin	0	1	0.00	0.00		0.00	0.00		0.00	0.00
ions	0	1	0.00	0.00		0.50	0.60		0.33	0.74
ros	0	1	0.00	0.00		1.00	0.61		0.33	0.79
mt met	1	0	0.00	0.00		0.83	0.62		0.50	0.76
n dna	1	0	0.00	0.00		0.00	0.00		0.00	0.00
cell antx	1	0	0.00	0.00		0.50	0.60		0.25	0.74
h2o2	0	1	0.00	0.00		0.50	0.60		0.33	0.74
diet antx	1	0	1.00	0.94		1.00	0.61		1.00	0.76
junk	0	1	0.00	0.00		0.00	0.00		0.00	0.00
ndna rep	1	0	0.00	0.00		0.00	0.00		0.00	0.00
lyso	1	0	0.00	0.00		0.00	0.00		0.00	0.00

Top Targets for Human Aging

For the singleton attractor basin the measures are trivial since there is only one term in the reduction. What we find, however, is that dietary antioxidants control the fate of this basin. If they are in an undesirable state (i.e. absent), then no matter the influence of the other variables, we will end up in the "bad" attractor. This behavior is clearly extreme. While critical to any person, not taking one's dietary antioxidants certainly has a more subtle response in real life. This behavior can likely be attributed to the fact that dietary antioxidants are represented by the only root node (i.e. no input function) in the Boolean network, which gives it much control and heavy influence over the other variables. We note one must always consider the coarseness of the Boolean model, and in this case, the multiple levels of abstraction inherent to a general model of the perpetual event of aging.

Moving on to the 3-cycle attractor, we find more stable results. In addition to the dietary antioxidants we find two other popular variables, namely the inner mitochondrial membrane (IMM) and the damaging reactive oxygen species (ROS). In this case it is the popularity which reveals our standouts as all variables with any popularity have roughly the same power. Biologically, the identification of these two variables is perfectly reasonable, especially that of ROS. ROS clearly plays a central role in the topology, but the importance of the health of the inner mitochondrial membrane is revealed by our method; its importance is not so evident given the attractor states and the topology.

The 9-cycle attractor seems to be quite similar to the 3-cycle, except more robust due to the size of its basin. The main difference is the increased importance of the mitochondrial metabolism, whose popularity is elevated above ROS. Not to be forgotten, however, ROS sees a slightly elevated power metric, which is enough to make it stand out from the others and be included among the most important. The inclusion of the mitochondrial metabolism into this upper echelon of variables is biologically appropriate given its role in the production of ROS.

Summary

We have demonstrated the utility of our analysis approach on a network size approaching the limits of exhaustively enumerable state spaces and found key variables with relevant biological significance. Demonstrated also is the dominant behavior of control nodes, which can sometimes mask the behavior of lesser variables we may be interested in. This lesson led to one of the motivations to remove the control nodes of the T-LGL Leukemia network discussed later in §3.5. We have extended this work by applying the results of this analysis to a related study on Artificial Intelligence Planning, detailed later in this results section.

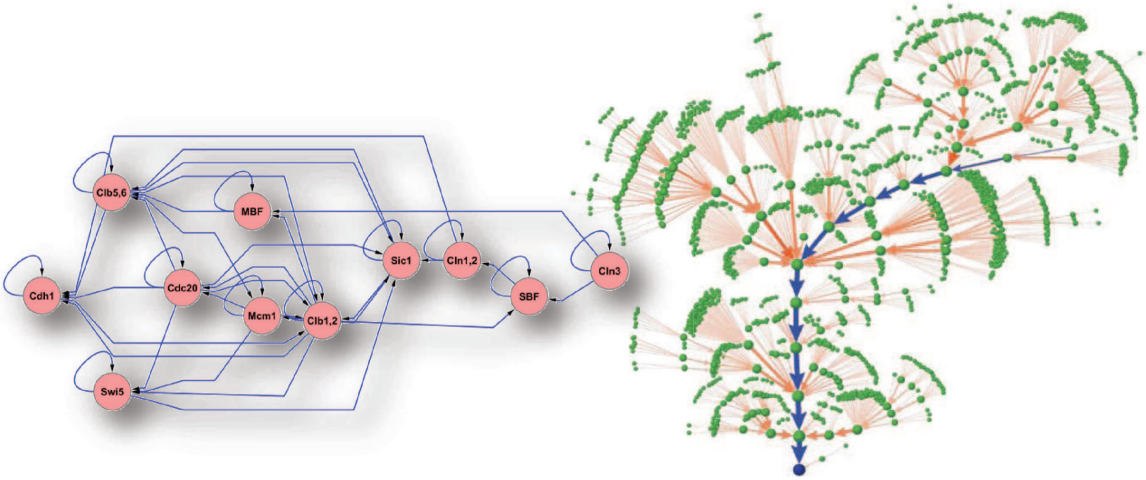


Figure 2.6: Yeast Network and Key Basin of Attraction

Application to the Study of the Yeast Cell Cycle

As another proof of concept on a gene regulatory network we applied our methodology to a well-studied Boolean network model of the yeast cell cycle [21] and show that key variables described in the manuscript are identified by our approach.

Network and State Space Properties

In their paper, Li *et al.*manually construct a Boolean network modeling the yeast cell cycle using 11 of the most important genes out of the approximately 800 known to play a role in the process. This network is simulated and results in seven basins of attraction, one of which is by far the largest and was studied exclusively in the paper. The network and this basin are shown in Fig. 2.6. In the large basin of attraction, which included 1,764 states and covers 86.1% of the state space, Li *et al.*were able to trace the trajectory of the yeast cell cycle from one of the fringe, or "Garden of Eden" [2], states down to the eventual point attractor state. The Boolean network adapted from Li *et al.*along with the state transition diagram it yields is shown in Fig. 2.6.

Top Targets for Yeast Cell Cycle

After applying logic minimization to the 1,764 states of the large basin we are left with a sum of 39 product terms. An abstraction of these terms can be seen in Fig. 2.7. The p and P_V measures were calculated for each of the eleven genes in the network, with our measures revealing the popular variables to include Clb5,6, Clb1,2, and the most powerful variable to be Cln3. Since the Clb5 gene (part of the Clb5,6 variable) is described as being responsible for driving the cell into the S phase (in which the DNA is synthesized and chromosomes are replicated), it seems reasonable to find it strongly represented in the minimized basin. If the role of Clb5 were not known beforehand, analysis of the basin in the manner described could identify it as important (and in the *ON* state) even though it is *OFF* in the eventual attractor state. The Clb2 gene (part of the Clb1,2 variable) is stated as being responsible for the transition in and out of the M phase (in which chromosomes are separated and the cell is divided into two). Thus, like Clb5,6, it is not surprising to find it among the most frequently specified variables in the basin representing the cell cycle. The most powerful variable, cyclin Cln3, was described in the paper as the trigger committing the cell to the division process. Despite its importance, we find it only explicitly configured in 2 of the 39 terms in the reduced basin (once for *OFF* and once for *ON*), which ranks it lowest in the p measure. However, because these two terms are the most powerful, Cln3's P_V score is quickly elevated.

When traditional network centrality measures of betweenness, centroid value, eccentricity and node degree are calculated for this yeast network (see Appendix C), we find that Clb1,2 and Clb5,6 are frequently reported with high scores, just like we find using our approach. This is also the case when influence[29] is calculated based on the Boolean network properties underlying the topology. However, the centrality measures also report variables such as Clb1,2, SBF and MBF, which are shown mathematically by our method to

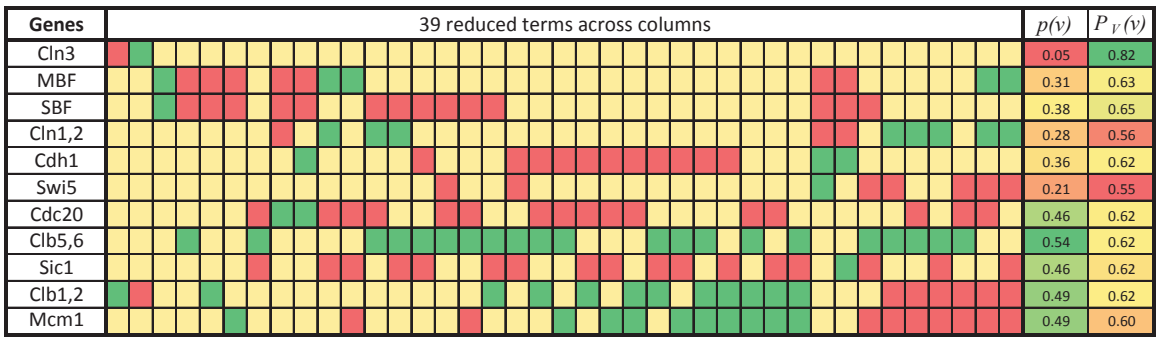


Figure 2.7: Yeast Network Terms and Small Network Measures: Across the central columns we see the 39 minimized terms vertically for the 11 variables, with yellow meaning a don't-care value.

have little intervention power whatsoever. Furthermore, these measures give little consideration to other key variables, including Cln3 and Mcm1, which our approach mathematically shows to have some intervention capabilities.

Summary

In summary, our single variable approaches report the key variables described by Li *et al.* and missed by traditional measures, and avoids reporting mathematically weak variables reported strongly by traditional measures.

Application to WNT5A Network for Melanoma

After applying our approach to hand-made networks, we apply our methodology to a well-studied WNT5A network computationally predicted from a melanoma data set [43, 44, 45]. In our previous work [45], the original data set was narrowed down to the ten most critical variables; these were selected out of 587 total on the basis of their strong interactive connectivity and either their known or likely roles in WNT5A driven induction of an invasive phenotype in melanoma cells, or their close predictive relationship with these genes. For each of the ten variables, we were able to identify the three most ideal predictors out of the remaining nine. Using this connectivity and a binary quantization of the original data set, the best binary logic functions were inferred for each target minimizing the Bayes

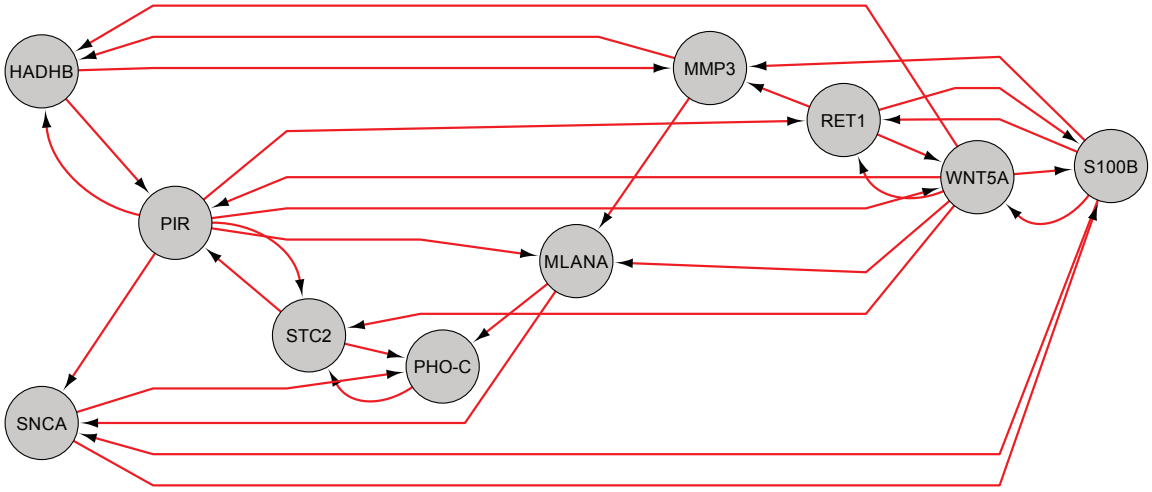


Figure 2.8: Melanoma Network

error[46, 47]. From these functions, the Boolean network and state space was enumerated, and attractor states identified (See Fig. 2.8 and Fig. 2.9). The reader is referred to the cited publications for detailed information on the data and connectivity.

Network and State Space

This network is unique among all studied in this chapter in that it contains only one basin of attraction leading to an attractor cycle of four states. This presents a rare and interesting situation, namely, that our analysis approaches no longer apply. Because logic minimization of the complete state space will minimize into a trivial tautology (all variables are don't-cares), our popularity measures cannot be calculated. The reason for this situation is likely the small size and generalization of the WNT5A system provided by the network (and mentioned in our previous work[45]). However there is still some insight to be gathered by our approaches.

Leading into the four state cyclic attractor are four basin subtrees, comprised of 714, 278, 28 and 4 states, respectively and shown in Fig. 2.9. If we consider these basin subtrees as though they were individual and separate basins, we can see if any key variables govern access to the attractor in one subtree over the others. This concept naturally extends

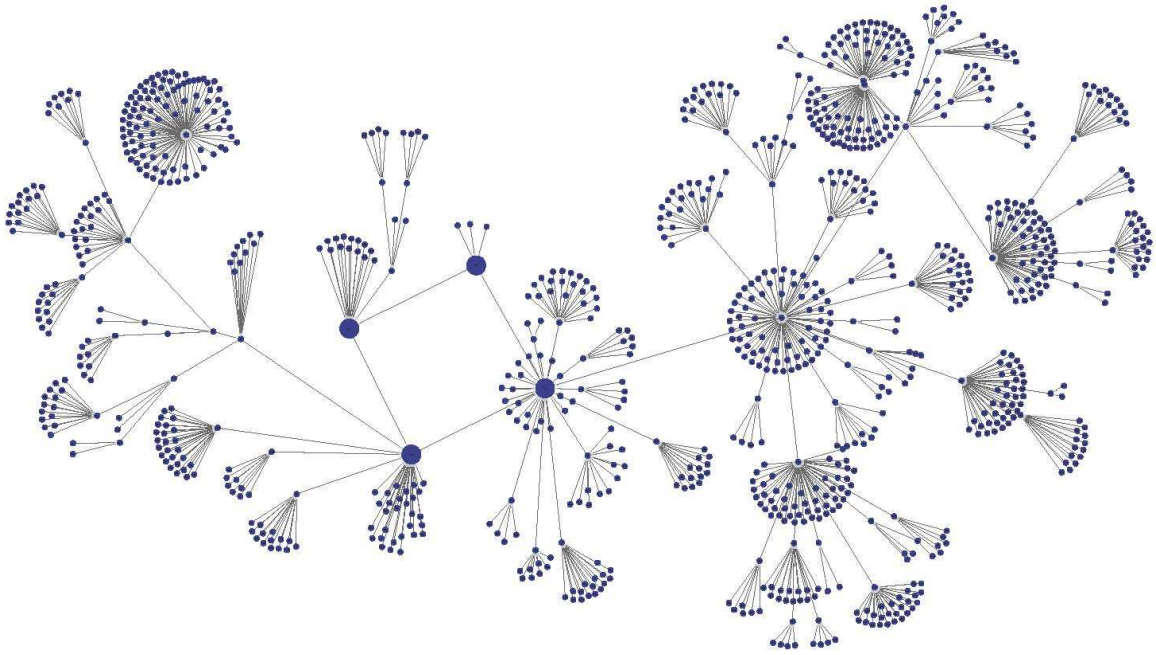


Figure 2.9: Melanoma Basin of Attraction Field: Just left of center we see the diamond-shaped four-state cyclic attractor. The top node in this cycle has the smallest basin subtree of just four states (including the attractor state). Clockwise to the next attractor state is the largest subtree with 714 states occupying the entire right side of the figure. Next is the 278-state subtree along the bottom and left of the figure. Finally the final attractor state roots the 28-state basin subtree extending upward.

to actually separate basins, and in that way this investigation will lend further insight into established results.

Top Targets for WNT5A Network

For the largest basin subtree, the popularity measure identified PIR as the top scorer, and since it is known that “cytoplasmic localization of PIR may represent a characteristic of WNT5A network for melanoma progression”[48], we are not surprised to see PIR atop the largest basin subtree. For the second largest basin subtree, our popularity measure reports PIR and also RET-1 as the top scoring variables, but not with as significant stand-out behavior as the first basin. The third and fourth basins have only 28 and 4 states, respectively, and thus their results do not reveal any clear stand-out variables.

When compared to the network centrality measures, as well as influence and sensitivity [29], PIR and WNT5A were the top ranked standouts. However, some measures dismissed RET-1 by including it among the mid to lower scoring variables, despite influence potential discovered by the popularity measure. While all variables in such a small, carefully selected set will bear some significance, our approach simply reveals those with true intervention potential given the network dynamics.

Summary

In a rare situation where a small network results in a single basin of attraction, we have observed a couple interesting trends despite the inherent inability to apply our full methodology. By isolating basin subtrees we are able to see that when there are enough terms to work with, the popularity measure successfully identifies influential variables, some of which are looked over by centrality measures.

Related Study: Artificial Intelligence (AI) Planning

In our recent work [49] using the previously discussed aging and yeast networks, we formulated intervention plans designed to trigger moves between the steady states of our systems. The work was an extension of an earlier study [50], and because of our unique approach we were able to identify basin-specific intervention strategies with the AO* search algorithm. These intervention strategies were found to improve upon prior AI planning and search techniques based on dynamic programming, produce biologically meaningful results, and furthermore these results aligned with the identified intervention targets from the small network approach of the previous sections.

AI Planning in the Aging Network

Regarding the aging network, in investigating intervention plans designed to move between all pairings of basins, we are able to create intervention plans which simulate treatments as well as natural aging and deterioration. In our previous study [51], one characteristic of the

"unhealthy" basin was that every one of the states had the variable for dietary antioxidants in the zero state. This fact was capitalized upon in all related intervention plans, which quickly complemented dietary antioxidants to get in/out of the basin. Our previous study [51] showed that the importance and role of dietary antioxidants was elevated by certain network topology characteristics, and its importance was confirmed and utilized in the formulation of intervention plans as well. To get to and from the "average health" basin, which had the smallest share of the complete state space, additional interventions needed to be devised. After dietary antioxidants, reactive oxygen species (ROS) was the variable next recognized as a critical intervention target as it was used from the unhealthy to the average basin and from the average basin to the healthy basin. ROS, which had maximal variable popularity scores and high power scores, was used in standard and complemented forms to make those specific adjustments needed to get to and from the average health basin. The selection of these two key variables is quite relevant biologically, as they are found to be at cross purposes within the cell. As ROS are damaging cellular entities which accumulate with age and deterioration, dietary antioxidants work to neutralize their effects. Thus, it is quite biologically relevant to see these two variables used in various combinations to intervene both as treatment and as aging effects.

AI Planning in the Yeast Network

The same approaches were applied to the yeast network to design interventions to get to the principle basin containing the entire cell cycle. Like in the aging network, our high-scoring variables were featured prominently in the plans, especially with those involving the largest basins, but we also saw the criticality of less powerful variables in navigating interventions from the smallest (rarest) basins. Thus, the intervention targets identified by our logic reduction methods were validated by our AI planning approaches; furthermore the roles of intervention targets with lower values of popularity and power were revealed

by the planning results to be the key players in navigating to smaller, more precise basins of attraction.

Regarding the yeast network, with over 86 percent of the yeast cell cycle network state space covered by Basin 1, planning for intervention in this network was less a question of healthy vs. unhealthy, as it was with the aging network, but more a question of normal v. abnormal. As [21] show that the entire cell cycle can be found within Basin 1, we aim to find out how to get to and from Basin 1 in our intervention plans. The variables selected include SBF, MBF, MCM1, Cln1,2, and Swi5. In the previous study [51], SBF and MBF were found to be variables with high power in Basin 1. As such their selection is quite warranted. Furthermore, the selection of MCM1 makes great sense as MCM1 has relatively high popularity scores in Basin 1. However, the selection of Cln1,2 and Swi5 are slightly more puzzling as they had low scores reported in the same study. We do see a trend in the selection of low-scoring variables for intervention plans involving rare basins (i.e. Basins 4-7, which combined total just over one percent of the total state space). This is because these low-scoring variables are what end up separating Basins of such small sizes; they are the difference makers. In one case, namely intervening on Basin 1 to get to Basin 7, no solution was able to be found, even with the use of low-scoring variables. For the converse case, from Basin 7 to Basin 1, we see the use of the low-scoring Swi5 variable, as Basin 7 is only one single state, to make the difference to Basin 1. Finally, we see the widespread use of the high-scoring variables for intervention plans to and from the larger basins as those variables define the biggest differences in state space trends between Basins.

2.6 Chapter Summary

Our work thus far has clearly established a usefulness in analyzing basins of attraction in identifying intervention targets. Our use of logic minimization significantly reduces

the representation of basins of attraction, and the proposed measures stratify the terms, revealing the key players in the system and how to manipulate them.

Our smaller networks, both hand-made and inferred, have provided clearly identified intervention targets that are not always evident in attractor states themselves, in the network topology, or even from various existing measures, both graph-theoretic and Boolean-network-specific. Perhaps the most important aspect of our revealed intervention targets is the fact that they are both basin- and value-specific; in other words, we providing not just a target, but how exactly to intervene (value) and also a context in which the intervention is appropriate (basin).

In our toy networks with 5 and 8 variables, we demonstrated the ideas behind the measures, exactly what information they can extract, and some clear distinctions from related measures. In making application to the handmade network modeling some of the systems biology behind human aging, we were able to identify biologically meaningful intervention targets as well as determine for which desired system steady states (i.e. disease, health, etc.) they are most effective. Our work on the systems biology of human aging also showed the applicability of this approach to models of scales approaching the limits of exhaustive state space enumeration. The analysis of the yeast cell cycle network also demonstrated that our methodology can identify key variables in the system. We were able to systematically identify three important variables described specifically by the original study and propose others for further study. Our application to the WNT5A network for melanoma demonstrated the applicability of our approach beyond hand-created networks to networks inferred from biological data; furthermore our targets identified for intervention had been previously validated by laboratory studies. Finally, the application of our small network approaches to AI planning showed how this approach can assist existing intervention planning strategies.

With small network sizes, however, it is likely that many variables will either be important in some way (known beforehand) or may even represent an amalgamation of multiple entities. Thus, intervention targets revealed may be true, but they may also be obvious depending on the study. This, along with the fact that the single variable measures begin to fail in larger networks requiring a sampled state space motivates us to expand our approach towards larger networks with dozens of variables and more, allowing us to include variables which are less well known and that may not be obvious intervention targets. In Chapter 3 we introduce a multiple variable intervention target approach to supersede the small network popularity and power measures, thus making possible the identification of powerful intervention targets despite sampled state spaces.

CHAPTER 3

TEMPLATE-BASED INTERVENTION IN BOOLEAN NETWORK MODELS OF BIOLOGICAL SYSTEMS

The identification of key variables which can act as targets for intervention remains a grand challenge in the modeling of biological systems. In the previous chapter we utilized a logic minimization approach to identify quality single variable targets for intervention from the state space of a Boolean network. However, as the number of variables in a network increases, the more likely it is that a successful intervention strategy will require multiple variables. Thus, for larger networks, such an approach is required in order to identify more complex intervention strategies while working within the limited view of the network's state space. Specifically we address three primary challenges for the large network arena: the first challenge is how to consider many subsets of variables, the second is to design clear methods and measures to identify the best targets for intervention in a systematic way, and the third is to work with an intractable state space through sampling.

We introduce a multiple variable intervention target called a template and show through simulation studies of random networks that these templates are able to identify top intervention targets in increasingly large Boolean networks. We first show that, when other methods show drastic loss in performance, template methods show no significant performance loss between fully explored and partially sampled Boolean state spaces. We also show that, when other methods show a complete inability to produce viable intervention targets in sampled Boolean state spaces, template methods maintain significantly consistent success rates even as state space sizes increase exponentially with larger networks. Finally we show the utility of the template approach on a real world Boolean network modeling T-LGL leukemia. Overall, these results demonstrate how template based approaches now

effectively take over for our previous single variable approaches and produce quality intervention targets in larger networks requiring sampled state spaces.

3.1 Motivation

The very nature of medicine is to know when and how to intervene in order to shift the steady behavior of a system to a more desirable state[19]. Ideally such interventions would be as minimally damaging as possible; however we know that especially with diseases such as cancer, interventions like chemotherapy are anything but minimal. In the path towards personalized medicine and individualized treatments with minimal collateral damage, designing and studying interventions that take advantage of our systems-level understanding and available data is and will remain of paramount importance, as working with computational models allows us to perform tests, execute simulations, and make predictions in inexpensive ways that require no human subjects[20].

In our recent work [22], detailed in the previous chapter, we employed a logic reduction algorithm to reduce the Boolean states comprising the basins of attraction to minimal representations, and from those minimizations we identified high quality intervention targets comprised of single variables. These approaches remain incomplete on their own for one key reason. With small network sizes (less than 20 nodes), it is likely that many variables will either be important in some way (known beforehand) or may even represent an amalgamation of multiple entities. Thus, intervention targets revealed may be true, but they may also be obvious depending on the study. This motivates our work to expand our approach to larger networks with dozens of variables and more, allowing us to include variables which are less well known and that may not be obvious intervention targets. Thus, for larger networks, a new approach is required beyond our previous work in order to identify coherent, multivalued intervention targets while working in with the limited view of the network's state space. In this chapter we introduce the template based approach to supersede

the small network popularity and power measures, thus making possible the identification of key players from sampled state spaces.

3.2 Related Work

In §2.2, we cover some key contributions to the application of Boolean networks to modeling biological systems. These summarized works also relate to our work in this chapter, as we seek to extend our previous approach to larger networks and multiple variable intervention targets.

Over thirty years after Kauffman's seminal work [1], Shmulevich *et al.*[29] pioneered work on a stochastic extension to the model called probabilistic Boolean networks (PBNs), which share the rule-based nature of Boolean networks but also handle uncertainty well. Within this extended framework of PBNs, studies were performed by Datta *et al.*[4, 5], which focused on external system control; studies by Pal *et al.*[6] and Choudhary *et al.*[7] explored intervention in PBNs to avoid undesirable states.

Work by Bryce *et al.*[50] progressed along the idea of connecting biological intervention design to AI planning problems, expanding on work from Datta *et al.*[4, 5]. These intervention plans can be perceived as multiple variables functioning as an intervention strategy, which is a parallel goal with our work in this chapter. Our previous work [49] formally mapped the biological intervention planning problem to a finite horizon partially observable Markov decision process (POMDP), using results from Chapter 2 as inputs for developing intervention plans. While this formulation generated high quality sequentially administered intervention plans, it takes as input a set of variables upon which to intervene, and is not designed to elucidate the intervention targets themselves.

3.3 Background

This chapter is an extension of the methods and results presented in Chapter 2, especially regarding logic minimization and the single variable intervention target measures of pop-

ularity and power. It is also based upon the Boolean network framework and associated algorithms presented in Appendix A.

For a reasonable number of variables, enumerating all 2^n states in the state transition diagram is not an issue. By starting at each state and evolving the network forward, each attractor and its basin can be enumerated. Exhaustive enumeration is the best possible situation for logic minimization because with more states, more common values can be identified and summarized in the reduction. In contrast, when the number of variables grows large and requires a partial enumeration obtained by a sampling approach, this greatly hinders the reduction step and results in many remaining terms with fewer don't-cares. Enumerating the full state transition diagram runs in time exponential in n, specifically $O(2^n)$, due to computing the next state for each of the 2^n states.

We found in the previous chapter that, for networks of a size manageable enough to exhaustively enumerate the states space, popularity and variable power can be used to identify key variables which make excellent candidate intervention targets. As described in Appendix B, we have performed a simulation experiment to identify the best single variable measures between popularity, power, and related measures described in our previous work[22]. We found in a simulation study over thousands of random networks with between 7 and 16 variables, that popularity, power, and their combination in the form of an harmonic mean showed the most statistically significant differences in intervention success rates of all 14 methods compared. Included in the comparison were Boolean network measures as well as graph theoretic measures.

Unfortunately, however, larger networks present some problems for these measures and require a different approach. The problems in larger networks manifest because popularity and power depend heavily upon the reducibility of the basin of attraction field. In larger networks, where we are forced to explore the basin of attraction field by sampling,

the reducibility of the state space is greatly hindered and our single variable measures are rendered unusable.

In this chapter we contribute a multiple variable intervention target called a template and show how, even in large networks with sampled state spaces, that they are still able to identify powerful targets for intervention. Thus we will see the template approach effectively taking over for the former single variable measures, especially in larger networks. Finally we contribute an example templates application to a T-LGL leukemia network and analyze the implications of our approach on this real world scenario.

3.4 Methodology

As the number of variables in a biological interaction network increases, the more likely it is that a successful intervention will require multiple variables. In fact, in our work in AI planning[49], we found that a planned sequence of interventions was an effective way to transition to a desired steady state. Our previous measures of popularity and power are capable of identifying multiple high value intervention targets separately in smaller networks. In this section, we will introduce intervention templates to take into account the multivariate effects of gene regulation and propose an approach to address larger networks. We will be faced with several challenges such as: 1) how to consider many subsets of variables in each basin of attraction, 2) to design clear methods and measures to identify the best template based intervention targets in a systematic way, and 3) to work with an intractable state space through sampling to cope with larger networks. We end by outlining a robust simulation study designed to illustrate achievement in these three areas.

Template Based Intervention Targets

Let the term *template* indicate a subset of variables (ordering not important) in a specific 0/1 (OFF/ON) configuration. Let the term *k-template* refer to a template with k variables; call the maximum value of k being considered K. Thus, for n variables there exist $\sum_{k=1}^{K} 2^k \binom{n}{k}$

templates, where $k = 2$ implies doubles, $k = 3$ implies triples, etc. This follows from $\binom{n}{k}$ ways of selecting k unique sets of variables from n total, 2^k binary value combinations for each of those sets, and K values of k.

Since a template is a subset of n network variables assigned to a specific Boolean configuration, each template with k variables covers 2^{n-k} other states. The smallest extreme is a template with $k = 1$, or a 1-template, which is a single variable assigned to *ON* or OFF. The largest extreme would be a template with $k = n$ — i.e. a single state in the state space. Such a template would cover no additional states, would not provide any further insights, and would be trivial to count. In practice, k is typically small, in the range 1-5 depending on n. Because there are $(2^k)\binom{n}{k}$ templates for every k, counting (and studying) quickly becomes intractable. However this is not typically an issue when seeking to identify intervention targets in biological networks since the difficulty of intervening increases with the number of variables required in the actual intervention.

Our combinational analysis involves counting the occurrence of each template remaining in the minimized DNF terms (T_i) of the original basins of attraction, (B_i), and is described in Algorithm 3.1. The algorithm executes with a runtime exponential on n and the size of the templates.

```
1 foreach Basin B_i : B do
2     for k = 1 : K do
3         foreach Template template_{j,i} for j = 1 :2^k (n choose k) do
4             foreach Term t_l : T_i do
5                 if Template template_{j,i} is found in t_l then
6                     increment count(template_{j,i}) for k
```

Algorithm 3.1: Template Analysis in Attractor Basins

With the vast number of templates we require ways to identify ones most likely to make the best intervention targets. After counting the occurrence of all templates in all basins, we begin by analyzing the most frequently counted templates as potential top intervention targets. Since logic minimization can greatly reduce the representation of the attractor basins (and thus the overall template counts), we provide a second measure to identify top templates with frequencies diminished by the logic minimization step.

Template Frequency (F): The most frequently counted templates in a particular basin are the first place to look for templates likely to make top intervention targets. By examining the set of terms T_i (reduced states with some don't-care variables) of a basin B_i, we can, for each of the $j = 1 : (2^k)\binom{n}{k}$ templates, count how many times that template appears in the terms of T_i. Formally,

$F_{j,i}$ is the final count of occurrences of $template_{j,i}$ in the set of terms T_i (corresponding to basin B_i). The maximum value of $F_{j,i}$ is 2^{n-k}, where k is the number of variables in the template, though in practice the value of $F_{j,i}$ is much lower, especially when state spaces are exhaustively enumerated and then greatly reduced by logic minimization. We can rank templates by $F_{j,i}$ to find top template candidates.

Template Basin Distribution Percentage (D): Just because a k-template is the most frequently counted template in a particular basin does not necessarily mean that it is the most significant. Sometimes, especially when logic minimization is able to significantly reduce the number of states in a basin to a much smaller set of terms, the frequencies of important templates will be diminished due to the introduction of many don't-care values. Thus we need a way to identify these high-value templates despite their lower frequencies of occurrence. To do this we consider the distribution of a particular template across all basins, or in other words, the affinity of a template to a particular basin. Since templates

can and often do appear in terms of multiple basins, we will calculate the ratio of occurrence in each basin and of the total number of occurrences. Formally,

$$D_{j,i} = \frac{F_{j,i}}{\sum_{m=1}^{L} F_{j,m}}, \tag{3.1}$$

where L is the total number of basins, j is the template number, and i is the basin number. Like frequency, for each basin, we can rank the templates by this ratio to find top template candidates.

These template measures provide two ways to identify a subset of templates warranting further investigation. By analyzing templates among the highest values of $F_{j,i}$ we will identify variables occurring together in a particular basin most often. By analyzing templates among the highest values of $D_{j,i}$ we will identify variables that may not be the most frequent, but retain the most affiliation with a particular basin even after logic minimization.

Intervention Targets

For each basin of attraction, full or partial, we can compute the best intervention targets using the following eleven methods. The first three methods comprise the best of the previous small network measures, namely, popularity, power, and their harmonic mean (abbreviated POP, POW, and HPP). The next four are the top templates of sizes 1, 2, 3 and 4 computed according to D (abbreviated K1TBDP, K2TBDP, K3TBDP, K4TBDP). The final four are the top templates of sizes 1, 2, 3 and 4 computed according to F (abbreviated K1FREQ, K2FREQ, K3FREQ, K4FREQ). For the simulation study described later in this section, we compare the single best template identified by each of these eleven methods. For our application to T-LGL leukemia, we examine the sets of top templates identified by K1FREQ, K2FREQ and K3FREQ.

Illustration

To illustrate the concept simply, let us consider for now the *unreduced* state space of the 5-variable example network shown in Fig. 2.1. The states from the diagram are collected and listed in Fig. 2.2 (left). We will count a few specific templates visually from the complete set (on the left), but later on in practice we will count templates from the set reduced by logic minimization (on the right). We can quickly see from the dramatic reduction (Fig. 2.2 (right)) that g_3 and g_5 are the key players in the network and that they display contrasting behavior between the two basins. As such let us examine only 1- and 2-templates involving g_3 and g_5 and observe their frequency and template basin distribution patterns. We begin with $k = 1$, and follow its discussion with $k = 2$.

We will first consider the four 1-templates involving g_3 and g_5 and look for any disproportionate patterns among the two basins. By inspection of the states of basin 1 we can quickly see that each of the 16 occurrences of $g_3 = 1$ and of $g_5 = 0$ are counted there, and that these are the only 1-templates with a maximal count of 16 in either basin. These counts are contrasted in basin 2 where we find the 1-templates for $g_3 = 0$ and $g_5 = 1$ (opposite configuration) counted the maximum of 8 times (since there are only 8 states in the basin). With no logic reduction in this example, these F values produce proportional D values that elevate $g_3 = 1$ and $g_5 = 0$ for basin 1 and $g_3 = 0$ and $g_5 = 1$ for basin 2.

Next we consider the 2-templates involving g_3 and g_5. Basin 2 counts just one 2-template the maximum of 8 times – precisely the template remaining after logic reduction: $g_3 = 0$ and $g_5 = 1$. Basin 1 counts 15 of the 40 total 2-templates 8 times, 12 of them 6 times, 12 of them 4 times, and 1 of them zero times, but counts none exclusively. Of the 15 templates counted 8 times in basin 1, the 3 templates for g_3 and g_5 dissimilar to basin 2 are found among them.

With the increased value of k we observe more dramatic template basin distributions and now reveal a template with complete affinity for Basin 2, which we did not see for $k = 1$. We now find all eight 2-template occurrences of $g_3 = 1$ and $g_5 = 0$ simultaneously exclusively in Basin 2 (100%), and this is corroborated by the logic minimization of Basin 2 which left us one single term with the same variables remaining in the same values we now find. Thus we observe that the templates with the highest F and/or D values correspond to the variables shown to exhibit the most network influence by logic minimization. Since single variable measures based on the results of logic minimization produced good targets in our previous work, we are motivated to further investigate the intervention viability of templates given their correlation with logic minimization results.

Sampling Large State Spaces

As the size of the network grows, i.e. $n > 20$, exhaustive enumeration of the state space (size 2^n) quickly becomes intractable and a sampling approach is required. In this approach we randomly and uniformly sample a number of initial starting states from the state space range $[1 : 2^n]$. From each starting state the network is then run forward to the corresponding attractor, collecting any states visited along the *transient* path. Attractors are noted and all corresponding states are collected into a partial basin, and it is from these partial basins that we identify our intervention targets. This approach will sometimes miss attractors with very small basins leading to them, but it certainly finds the largest ones, and for a large number of samples, gives us a significant set of member states to analyze. We can also approximate the percentage of the total state space occupied by each basin based on the percentage of total samples associated with it. It should be carefully noted that when sampling, the identified partial basins are themselves proper subsets of the complete basins. In other words, the sampling approach creates no incorrect assignments of basin states to attractors. With exhaustive enumeration, we complete the state space exploration and acquire all states

in each basin. Thus both sampling and exhaustive enumeration provide correct basin states, just with sampling being an incomplete picture and exhaustive being a complete picture.

Evaluating Intervention Success

Abstractly, an intervention should shift the steady behavior of a system to a different (usually more desirable) state. In the context of the Boolean network formalism, this is represented by shifting the steady behavior of the system, represented by an attractor state or cycle, into a different basin of attraction. A successful intervention needs only to shift the state from a starting attractor state into any state in the basin of the goal attractor, as the network dynamics will naturally bring the state to the attractor itself. With the top intervention targets determined for each basin, we attempt interventions to each basin as a goal destination, starting from attractor states from each outside basin, recording each attempt as a success or failure.

Across many Boolean networks we will find a range of the number of attractor basins, from one or two to dozens, and within those basins, attractor cycles of various periods, from 1 to 5 to 100 or more. In estimating intervention success we should not let abnormally large attractor cycles bias the results by providing too many of the intervention starting states. For example, if we have a network with a point attractor in one basin and an attractor cycle of 50 states in the second basin, our intervention success estimates would be based on an unfair distribution of starting states. In practice, if an attractor cycle is longer than ten states, we randomly sample ten attractor states for that basin upon which to apply candidate interventions. If it is less than ten, we use them all.

In our simulation study (described next), we must compare the performance of top intervention methods across various network sizes and between the methods themselves. As we will be comparing pairs of success proportions for independent interventions, which

qualify as Bernoulli trials under the binomial model, we will use a two-proportion Z-test with the null hypothesis of having equal success proportions[52].

Simulation Study

Our two main challenges with the template approach involve performance as we transition from fully enumerated state spaces to sampled state spaces, and also performance as we increase the size of the networks. To address these concerns we design a simulation study over hundreds of randomly generated Boolean networks within which we compare performance of the former single variable measures and the new template based approach.

In-silico Network Models

In order to test interventions over an adequate range of network sizes, we create multiple random networks with 10, 12, 18, 20, 25 and 40 variables each, for a total of 200 networks. In each network we randomly generate Boolean update rules for each variable, thus also creating random network connectivity. In order to create biologically inspired networks, we adopt the per-variable connectivity distribution from Albert *et al.*[23] used originally for a *Drosophila melanogaster* network, shown in Table 3.1.

Once network connectivity and rules are determined, the basin of attraction fields must be generated in full or in part. For the network of size 10, 12 and 18, we use half the networks for exhaustive enumeration of the basins of attraction and for the other half of networks we enumerated partial basins through sampling. For networks of size 20,25 and 40, only sampled state spaces were used.

Table 3.1: Connectivity Distribution for Random Network: In the table we see the probability of being assigned various numbers of inputs to the Boolean functions.

Number of Inputs	1	2	3	4	5	6	7
Probability	.101	.233	.267	.183	.083	.083	.050

Performing Interventions

An intervention is defined as a modification to an attractor state of a source basin attempting to reach any state in a goal basin. For the simulation study we do not attempt interventions where the goal and source basins are the same, since these are more likely to succeed and would inflate our results. Likewise we do not attempt interventions to goal basins estimated to occupy less than 15% of the total state space, since reaching these very rare basins is the most difficult and has little biological relevance. Our 200 original random networks, through their various numbers of attractor basins and attractor cycle lengths, produced 4,223 individual intervention attempts, each applied separately with all eleven methods.

3.5 Results

Next we present the results of the simulation study described in the methods section, which reveals the ability of template-based interventions to maintain performance between exhaustive and sampled state spaces and also in increasingly large random networks with sampled state spaces. We then provide a demonstration of the approach on a real world network modeling T-LGL leukemia, originally hand-created by domain experts.

Simulation Study for Template Methods

In order to demonstrate the robustness of template-based interventions we present the results of the simulation study described in the Methods section. The study addresses two main questions: 1) What effect does the change to sampled Boolean network state spaces have on the performance of template-based interventions, and 2) What effect does increasing network size have on the performance of template-based interventions? To address these questions we analyze hundreds of randomly generated Boolean networks for which we compare performance of single variable measures and the variations of the new template based approach.

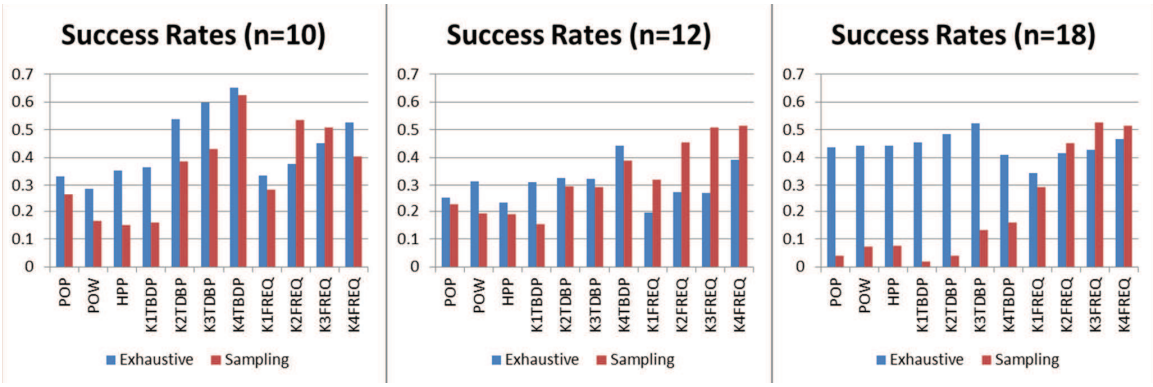


Figure 3.1: Success Rates Between Exhaustive and Sampled State Spaces: For three network sizes we see proportions of successful interventions across single variable and template based approaches. The first three sets of bars in each sub-figure are the single variable measures of popularity, power, and their harmonic mean. The latter eight are the two template approaches of D and F across four template sizes. P-values reflecting proportion differences can be found in Table 3.2.

The Effect of Sampling on Template Interventions

To address the challenge of whether or not template based approaches remain robust as we transition from a fully enumerated state space to a sampled one, we compare the performance of each measure between full and sampled versions of the 10, 12 and 18 variable networks. Since these networks are small enough to exhaustively enumerate the full state space, the comparison will provide a full assessment of any performance degradation due to sampling. For each network size we compare the proportions of successful interventions of each type between exhaustive and sampled networks. Interventions that show a significant change in proportion between exhaustive and sampled state spaces will be noted. Success rates can be seen in Fig. 3.1, (with significance values in Table 3.2) for which we note the following observations:

In the networks with 10 variables, nine of eleven methods showed a decrease in performance in from exhaustive to sampled, with six of those being statistically significant changes. Interestingly, two of the eleven methods showed an increase in performance,

Table 3.2: P-Values for Two-Proportion Z-Tests: In the table we see the p-values for two-proportion Z-tests across eleven intervention targets between exhaustively enumerated (E) and sampled (S) networks with 10, 12 and 18 variables. For p-values below the $\alpha = 0.05$ significance level we reject the null hypothesis and conclude that there exists a statistically significant difference in success proportions between the exhaustive and the sampled state space cases. These differences can be inspected visually in Fig. 3.1

	10E/10S	12E/12S	18E/18S
POP	0.0920	0.4619	0.0000
POW	0.0013	0.0003	0.0000
HPP	0.0000	0.2030	0.0000
K1TBDP	0.0000	0.0000	0.0000
K2TDBP	0.0007	0.4421	0.0000
K3TDBP	0.0002	0.4391	0.0000
K4TBDP	0.5424	0.1385	0.0000
K1FREQ	0.2197	0.0002	0.2038
K2FREQ	0.0004	0.0000	0.3964
K3FREQ	0.2015	0.0000	0.0192
K4FREQ	0.0062	0.0014	0.2310

with one of those being statistically significant at the 0.05 alpha level; both of those cases involved frequency based templates. While showing decreases in performance between exhaustive and sampling with each template size, template basin distribution percentage templates registered among the highest success proportions overall, indicating that they might be a suitable replacement for single variable methods in exhaustive cases if the computational cost of templates can be afforded.

In the networks with 12 variables, all four template sizes for frequency based templates showed an increase in performance in the sampled cases, with all four differences in proportions being statistically significant. All seven other methods showed visible decreases in performance, with two of them being considered statistically significant. Again, template basin distribution percentage templates had among the best performances in the exhaustive cases.

In the networks with 18 variables we see the clear resistance of frequency based template approaches to the side effects of sampled state spaces. With one statistically significant increase in performance and no significant decreases, frequency based templates overwhelmingly dominate performance in sampled cases while all other methods drastically decrease.

It is very interesting to observe any increase in performance between a fully enumerated space and a sampled space. In our case the cause of this behavior is because, in the sampled condition, the intervention targets are computed from a larger amount of terms. This is due to the fact that the logic minimization ability is highest when every state is known beforehand and it is hindered greatly by not knowing all basin states ahead of time. Only the templates (especially the larger ones) benefit from this situation – single variable measures like popularity and power suffer when there are *fewer* don't-cares in the terms, and some template measures suffer when there are *too many* don't-cares (i.e. exhaustive case). Thus, while drastic logic minimization on exhaustive state spaces allows popularity and power to quickly reveal decent targets, a less effective logic minimization leaving many more terms behind after reduction benefits the template approach by providing more information from which to identify the best templates while still eliminating the least important variables. In some cases, this benefit outweighs the benefit of a full logic minimization.

While we expected to avoid recording, for template approaches, significant success proportion decreases for sampled network state spaces, we not only failed to detect that trend altogether in frequency based templates, but in many cases detected significant increases. We also began to see the template basin distribution percentage templates as most ideally suited to smaller networks in exhaustive cases, perhaps as a more thorough alternative to the former single variable measures. From these data we conclude that frequency based template methods are much more robust in sampled state spaces than their single

variable counterparts. The next step is to observe if this trend continues with increasing network size.

The Effect of Increasing Network Size on Template Interventions

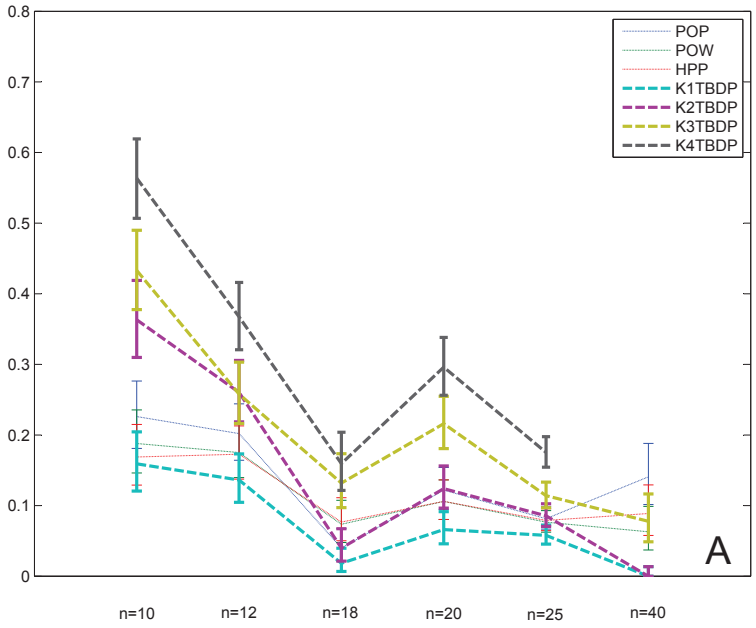
Satisfied that template measures remain robust in networks with sampled state spaces, we now investigate larger random networks. It is expected that any measure or technique will decrease in performance as the size and/or complexity of the network increases. However, with the knowledge that the single variable measures fail completely with sampled state spaces even for small networks, we need to be assured that template performance remains robust. We generate further random networks with 20, 25 and 40 variables, which produce state space sizes of 2^{20}, 2^{25}, and 2^{40}. We sample these enormous state spaces with about 1% or less coverage in initial states and estimate the intervention success of the top scoring intervention targets from our eleven methods (3 classic single variable, 4 sizes of D templates and 4 sizes of F templates). For the 40-variable network we do not include size-4 templates for computational considerations of the simulation. In Fig. 3.2 (a) we show the performance of template basin distribution (D) based templates at template sizes of 1-4, and the single variable measures of variable popularity (POP), variable power (POW), and the harmonic mean of the two measures (HPP). In Fig. 3.2 (b) we show the performance of frequency (F) based templates of sizes 1-4 against the same single variable methods. Both sub-figures show performance over networks of size 10, 12, 18, 20, 25 and 40, all with sampled basins of attraction. Error bars shown reflect the 95% binomial confidence intervals.

In Fig. 3.2 (a) we observe that template basin distribution percentage templates do not consistently show significant differences in success proportions with POP, POW or HPP. We do note that K3TDBP and K4TBDP do show significant performance over POP, POW and/or HPP in all networks up to 25 variables, but they are significantly outperformed

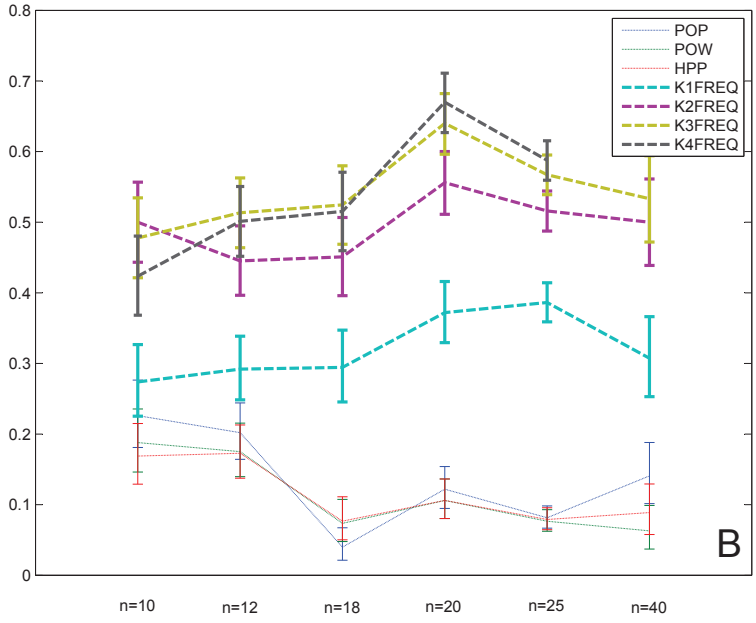
by their frequency based counterparts in Fig. 3.2 (b). In Fig. 3.2 (b) we observe that the only time a frequency based template does not show a significant difference in proportion is in the 10 variable network for the template with only one variable (the most extreme case), though in practice we would not apply template analysis on such small networks. A complete separation of 95% confidence intervals surely indicates a significant separation in success proportions[53], but to further reinforce these observations and to reveal any significant differences in proportions not obvious from confidence intervals, we computed two proportion Z-tests for independence for each pair of methods for each networks size. These pairwise matrices of p-values revealed even stronger conclusions than the graphs in Fig. 3.2, further confirming the statistically significant differences beyond what is obvious by visual inspection of the error bars. As they do not reveal any critical trends not visible in Fig. 3.2, we omit these lengthy p-value matrices from this report.

Summary of Simulation Study

Over all network sizes, it is clear that frequency based templates not only maintain performance between exhaustive and sampled networks, but also provide consistent success rates with increasing network sizes despite the exponential explosion of state space sizes. We were also interested to observe that in some cases the inhibited reducibility of sampled state spaces actually contributed additional information to the computation of the larger template targets–in some cases actually improving their performance in sampled networks over their performance in maximally reduced state spaces. Template basin distribution percentage based templates are sometimes useful in smaller networks, and are the most effective in smaller, exhaustive networks as a more thorough alternative to the simpler, single variable measures of POP, POW and HPP, albeit at an increased computational cost.



(a) Template Basin Distribution (D) Rates



(b) Frequency (F) Rates

Figure 3.2: Success Rates for D and F Templates: In (a) we see the proportions of successful intervention attempts for D-based templates of four sizes across six different network sizes. In (b) we see the same for F-based templates. Both figures include the performance of the single variable approaches for comparison. While D-based templates perform reasonably well in the smallest network sizes, it is clear that F-based templates show consistent and significant performance across all network sizes. Note that 4-variable templates are not included on the 40 variable network for computational reasons.

Application to T-LGL Leukemia Network

In our previous work [22] we identified useful intervention targets using the single variable measures in real world networks for melanoma, the yeast cell cycle, and for human aging. Because we saw in our simulation study that no significant new information is revealed by template approaches in exhaustive state spaces, we do not apply our template approaches to those previously explored networks here. Instead, because of the robustness of the templates approach for large networks with sampled state spaces, we make application to a 43-variable network for large granular lymphocytic (T-LGL) leukemia where the single variable measures have no usefulness.

Zhang *et al.*[54] have methodically constructed a model of the blood cancer T cell large granular lymphocyte (T-LGL) leukemia from hundreds of literature sources. The original study, as well as others based on variations of this large network[55], have searched for therapeutic targets and have even validated some experimentally. However these predictions required expert-level topological reduction and simplification of the network in various ways. But because of the validated findings, this network makes an ideal situation in which to apply our approach, which is purely computational and requires no expert-level knowledge of the disease system. If our results on the larger, less simplified network are reasonable, our approach will be shown useful and applicable on large networks for which we may not have expert-level knowledge and/or the ability to systematically simplify.

Network Construction

The original network [54] created from the literature contained 128 nodes and 287 edges, but was simplified by the authors through software and manual adjustments to 60 nodes and 142 regulatory edges. After collaborating with a principle author from [54], we performed further reductions on the network according to techniques described in related work involving this same network [56, 55]. The goal of further reduction was to remove nodes which

Table 3.3: Summary of Four Attractor Basins for the 43-Variable Leukemia Network: We observe the attractor length, numbers of states identified, and the estimated state space coverage for each basin. Attractor type classification criteria are described in the text.

	Basin Number	1	2	3	4
50,000 Sampled +	Attractor Period (p)	4	1	1	4
93,618 Transient =	Number of States	58713	57359	17099	10441
143,618 Total States	Estimated % Coverage	40.88	39.94	11.91	7.27
	Attractor Type	Healthy	T-LGL	T-LGL	Healthy

mask the dynamic behavior of the network variables (i.e. the overarching influence of the apoptosis node as well as control nodes); since steady-state analysis will be performed over many randomly generated states, control variables are not necessary since variables they control will be forced to take on different values through random starting state assignment. After the simplifications described in Appendix D.1, we obtain the 43-variable version shown in Fig. D.1. A list of the Boolean functions is also given in Appendix D.

Network and State Space Properties

With 50,000 randomly and uniformly sampled initial states, the partial basin of attraction field was enumerated, resulting in five basins of attraction, four of which comprise greater than 99% of the state space (estimated), and will thus be the focus of the analysis. These four basins are summarized in Table 3.3. In other sampling runs up to seven attractors were identified, but these additional two, when discovered, were estimated to occupy thousandths, if not tens of thousandths of one percent of the state space, and would be discarded for analysis along with the fifth basin. Due to the massive size of the full state space (2^{43}, over 8 trillion states), 50,000 initial sampled states was chosen first due to it being large enough to proportionally reveal all major basins and, second, because choosing more samples, such as 100,000 or 200,000 would only marginally increase the coverage of the full state space.

We have categorized the four basins to either Healthy (i.e. normal apoptosis function) or T-LGL (i.e. cancer state) based on the values of certain key variables in the steady attractor states/cycles. In the original 60-variable network, the presence of control nodes, including one for apoptosis, simplified classification of attractor states. Since the apoptosis node was wired to nearly every other node for purposes of the original study[54], its behavior dominated the dynamics of the entire network. Since we have stripped the network of the apoptosis node, as well as other control nodes, we must interpret the attractor states based on other criteria. These criteria involve precisely the series of input regulatory nodes controlling the former apoptosis node. Nodes found in identical steady states across all four attractors were not considered, but several variables, namely FasT, Fas, Ceramide, and FLIP, were used in their boolean functions to effectively determine whether or not an apoptosis node would have been active or inactive in the attractor. Full attractor states are given in Appendix D materials. In [55], extreme simplifications (down to just six nodes) to the same original network resulted in two very simple and easily classifiable basins of attraction. Here we have four attractor basins, with two classified as healthy and two as T-LGL. Among the two in each category one exhibited stronger, more consistent behavior. Specifically, Basin 1, as the largest basin in the space, was classified as healthy but had oscillatory values for DISC and FLIP, direct influences on apoptosis. Basin 4, on the other hand, had consistent behavior for both DISC and FLIP despite the cyclic nature of the attractor. Thus, due to the small size of Basin 4 and its pure behavior even with a cyclic attractor, it is considered the healthiest attractor. For complimentary reasons, Basin 2 is the larger and thus more general T-LGL attractor, and B3 is the smaller, more precise T-LGL attractor. A total of eleven attractor states were identified: a cycle of 4 states in Basins 1 and 4, and singleton attractor states in Basins 2, 3 and 5, the latter of which was not considered as a goal state.

Templates Analysis

Next we identify key variables from k-templates for $k = 1, 2$ and 3. The 43 network variables give 86 1-templates, 3,612 2-templates, and 98,728 3-templates to count across the minimized terms in our four basins. We stop with $k = 3$ since $k = 4$ provides nearly two million 4-templates to count, but only across less than 120,000 terms. Future work intends to parallelize and split up the computational burden of the sequential counting algorithm and offer conditions under which we can test the efficacy of counting orders of magnitude more templates across relatively few terms.

After counting k-templates for $k = 1, 2$ and 3 we estimated the intervention success rate for the 50 most frequent templates (highest values for $F_{j,i}$) in each size and for each basin. Full listings of these templates can be found in Appendix D, but a listing of interesting templates is found here in Table 3.4. To estimate a success rate, we apply each top template intervention to all eleven attractor states across all basins in the T-LGL network and compute the distribution of basins reached. If a template causes a significant number of these attractor states to jump to (or remain in) a desired basin, such a template is of great interest. In general we expect an attractor state to remain robust to perturbation but expect the larger templates to have the best chance at changing the steady state of the system. Since Basins 1 and 4 each contain a 4-cycle of attractor states, we expect at least 4 of the 11 destination states (36%) to remain in those basins (due to the expectation of robustness mentioned). Likewise the remaining basins each have a single attractor state each and we expect 1 of the 11 destination states (9%) to remain in these basins. Of the top 50 templates for each k and each basin, we indeed saw these expected distributions of goal states very frequently. Numbers above and beyond these expectations warrant closer inspection, which we provide for the most interesting templates listed in Table 3.4.

Table 3.4: Notable Leukemia Network Templates: Listed below are a selection of interesting templates from the analysis of the Leukemia network. Complete lists of the top frequency templates for each basin and each template size can be found in Appendix D.

Basin	RANK	Variables	Values	$F_{j,i}$	SUCCESS
1	18	Ceramide PDGFR S1P	0 1 0	22936	0.73*
2	27	PDGFR S1P TBET	0 0 1	28057	1.00
3	1	Ceramide PDGFR S1P	0 1 1	14404	1.00
4	6	TBET	0	6404	0.73

*Remaining 27% of interventions lead Basin 4 (also healthy).

While the simulation study was clear that intervening with a template of high frequency was sufficient to best comparable methods, before continuing we provide a comparison demonstrating that more successful templates and fewer powerless templates are found among the most frequently counted. To do this we compute success rates for 3-templates in three categories: the 1000 most frequent, 1000 random, and the 1000 least frequent templates. These rates are shown in Table 3.5 and, as expected, we see with statistical confidence that the top 1000 most frequent templates have higher proportions of success than the other groups.

Top Templates for Healthy and T-LGL Attractors

We discuss two templates with high success rates leading to healthy attractor states followed by two templates with high rates of leading to T-LGL states. Lastly we observe the single template which best differentiates the healthy and T-LGL states.

The first healthy intervention template of interest is for Basin 1, where we find a 3-template (Ceramide, PDGFR and S1P set $0/1/0$, respectively) with a high frequency and

Table 3.5: Frequent Templates are More Successful than Others: Listed below are intervention success counts, Z-scores and p-values for two-proportion z-tests between the top 1000 most frequent 3-templates and 1000 random templates or the 1000 least frequent templates. With eleven intervention starting states over 1000 templates, each number of successes is out of 11000 intervention attempts. In each case we find, with statistical confidence, that the top frequency template group provides a higher proportion of successful interventions than the other groups.

	Top 1000	Random 1000	Z-Score	p-value
Basin 1	4493	4293	2.753	0.00298
Basin 2	3838	1783	31.767	0.00000
Basin 3	2534	1073	26.605	0.00000
Basin 4	4520	3378	16.050	0.00000
	Top 1000	Bottom 1000	Z-Score	p-value
Basin 1	4493	1979	37.196	0.00000
Basin 2	3838	8	67.986	0.00000
Basin 3	2534	1000	28.166	0.00000
Basin 4	4520	2315	32.124	0.00000

the ability to guide 73% of intervention attempts into Basin 1. In fact, the 27% remaining attempts not leading to Basin 1 lead to Basin 4, which is the other attractor classified as "healthy". Thus, this template completely avoids the T-LGL basins. S1P is known to be 1 (ON) in a T-LGL state[57], and so it is biologically consistent to find it set to 0 (OFF) in the template.

The second healthy intervention template of interest is found in Basin 4–the smallest classified basin considered in this study. In it we find a powerful 1-template, TBET 0, which transitions 73% of intervention attempts to this very small basin covering only about 7% of the total state space. TBET is known to be 1 in T-LGL[54], so as a healthy basin this setting is biologically consistent. Due to the small size of the basin, none of the top 2- or 3-templates were able to improve on this rate, and any that matched it included TBET 0 as part of the template. Any time a single variable can exert such a high degree of influence on a network it is noteworthy.

The first T-LGL "intervention" template is found in Basin 2. While is may seem contradictory to describe an "intervention" which leads to a disease state, we nonetheless consider the power in the variable combination, perhaps as a trigger to avoid. In Basin 2 we quickly find the standout behavior of the 3-template PDGFR/S1P/TBET, which is set $0/0/1$, respectively. This template had a 100% intervention success rate, guiding network dynamics to Basin 2 upon every application. As in the previous template, we again see TBET, however this time it is set 1, which is biologically consistent with the T-LGL state.

For Basin 3, which is also classified as a T-LGL attractor, we find another very powerful template, namely Ceramide, PDGFR and S1P set $0/1/1$, respectively. Not only does this template have a 100% intervention success rate, but each of its variable settings is known to be biologically consistent[57, 58, 59]. As the attractor state for Basin 3 is classified to be the stronger of the two T-LGL basins, and as Basin 3 is estimated to occupy only about 12% of the total state space, such a powerful template with perfect biological consistency is significant indeed.

Perhaps most interesting is the observation that this 3-template in Basin 3 and the 3-template for Basin 1 share the same three variables, with one going to 100% healthy attractors and the other going 100% to a T-LGL attractor. The templates have the same settings for Ceramide and PDGFR but differ in the setting for S1P, which is the biologically consistent setting across both basins. This reveals that, while Ceramide and PDFGR do not have biologically meaningful settings in Basin 1, for Basins 1 and 3 they still open the path, in terms of network dynamics, for the biologically consistent behavior of S1P to accurately and powerfully shift the network between healthy and T-LGL attractors. Thus we conclude that, for this network, S1P is the key differentiator between healthy and T-LGL steady states, assisted by the combinatorial power of Ceramide and PDGFR.

Our work with the Leukemia network has produced some notable findings. First, we note that while all single variable measures were unable to produce helpful intervention targets because of the 43 network variables, the template based approaches did produce single and multi-variable intervention targets with observable separation in per-basin frequencies and in intervention success rates. Second, we observed that identifying the best template can benefit greatly from expert assignment of the basins of attraction to biological contexts (e.g. health v. disease).

Finally, we saw that the biological significance of the results depends quite heavily on network rules. Basin 2 produced some biologically unexpected advice (i.e. S1P set to 0) within the templates, while the large, healthy Basin 1 and also the smaller T-LGL Basin 3 contained templates with immense power and great biological significance. This discrepancy can be attributed to many causes, including a network more focused on modeling the disease state (relatively rare) v. the healthy state, which may be acceptable depending on the application. In our case, we saw two variables fixed between healthy and T-LGL interventions while the biologically consistent setting of the third variable, enabled by the combinatorial power of the first two, was able to dictate network fate.

Thus, while templates are capable of revealing novel biological insights, they may also reveal or confirm sensitivities in the network rule system that may or may not be desirable for a particular biological model. In the end, template based analysis reveals the most powerful triggers for altering network dynamics into desired attractor basins strictly based upon the given Boolean rules. In our look into the T-LGL leukemia network, our templates were realized, and in most cases biologically reinforced, on a network with over 8 trillion states in the basin of attraction field based sampling only 50,000 initial states.

3.6 Chapter Summary

Our work in Chapter 2 and in this chapter has clearly established a usefulness in analyzing basins of attraction in identifying intervention targets. Our use of logic minimization reduces the representation of basins of attraction, and the template measures stratify the terms, revealing not only the key players in the system but also how to manipulate them. Perhaps the most important aspect of our revealed intervention targets is the fact that they are both basin- and value-specific; in other words, we providing not just targets, but how exactly to intervene (value) and also a context in which the intervention is appropriate (basin).

With small network sizes (less than 20), it is likely that many variables will either be important in some way (known beforehand) or may even represent an amalgamation of multiple entities. Thus, intervention targets revealed may be true, but they may also be obvious depending on the study. This, along with the fact that the single variable measures fail in larger networks requiring a sampled state space, motivated our work to expand our approach for larger networks with dozens of variables and more, allowing us to include variables which are less well known and that may not be obvious intervention targets. By introducing the template counting approach to supersede the small network popularity and power measures, we have made possible the identification of powerful intervention targets despite sampled state spaces.

We first demonstrated the maintained success proportions of frequency based template interventions between exhaustively enumerated and sampled state spaces. Convinced that key information was preserved by the measures despite sampling, we next showed the consistent success proportions across networks of increasing sizes as other methods fell away in performance. These investigations into robustness convinced us that the template approach was sure to provide the critical information needed regarding intervention targets.

We have also demonstrated the efficacy of the approach on a larger T-LGL Leukemia network crafted by domain experts. We note that when all single variable measures were unable to produce helpful intervention targets, the template based approaches did produce single and multi-variable intervention targets with high intervention success rates. In the end, the template based analysis revealed the most powerful triggers for altering network dynamics into desired attractor basins, and these results were realized, and in many cases biologically corroborated, on a network with over 8 trillion states in the basin of attraction field based sampling only 50,000 initial states.

Despite progress in sampling large state spaces, we will always be limited by the exponential growth of the state space with the number of variables. Fortunately as network sizes race into intractability, so too does the reliability of such networks, which is a direct influence on the quality of our results. In the end, our measures will always reveal the true triggers of network dynamics based on the given rules of the system. Thus, while they are capable of revealing novel biological insights, they may also reveal or confirm sensitivities in the network rule system that may or may not be desirable for a particular biological model. Since there are quality handmade networks with sizes into the dozens of variables, such as our T-LGL Leukemia network, a Drosophila network from Albert *et al.*[23], and others, our leap to the 40-50 variable size level is significant. With improvements to algorithm implementation and with the incorporation of parallelization, we plan to improve the large networks approach in terms of speed and network size capability, ideally towards the 75-100 variable mark.

Frequently in Chapter 2 and in this chapter we classify Boolean network attractors as phenotypes or contexts and the basins of attraction as the gateways to those phenotypes. This classic perspective on Boolean networks lends itself well towards intervention design, but it is undoubtedly restricted in what it can model. While our basins of attraction can be taken to represent some idea of a biological context in certain situations, perhaps in

addition to the contribution of our new methodology, this work also reveals some key limitations concerning these classic models. Specifically, it is interesting to make context-based interpretations using the Boolean model (or other models), but if our goal is to really model systems specific to some biological context, it would be better to work within a framework designed to explicitly model context-specificity, rather than attempt interpretations on more general models. To this end we strive to use our progress here in the Boolean framework as a springboard towards a new context-based perspective of biological systems modeling.

CHAPTER 4

CONTEXT SPECIFICITY IN MODELING BIOLOGICAL INTERACTIONS

Motivated by the progress in identifying targets for intervention through novel analysis techniques for the classic Boolean framework, but at the same time recognizing some inherent limitations, we strive to create a new framework from the ground-up – one designed to more accurately model the contextual behavior of the biological system. Based upon such a model, future methods for identifying intervention targets will be more effective and helpful.

4.1 Motivation

Of great importance to modeling accuracy is the model's acknowledgment of biological contexts within the data – i.e. recognizing the heterogeneous nature of the true biological system and the data it generates. One may consider the context of a biological system to be any steady state we find ourselves in (e.g. a phenotype); a coarse example of this may be a diseased state versus a healthy state. Any samples taken from the context of health will reflect the results of healthy interactions and homeostasis, while the samples taken from the diseased context will reflect the proliferation of undesirable interactions. The problem surfaces when we realize that many of our models, for instance Bayesian networks, assume that the data we observe has one underlying distribution, be it Gaussian or otherwise. This biases any network inference we perform to disregard the natural separations of biological context and incorrectly treat the data as homogeneous. Many other computational and mathematical techniques have been developed to infer molecular patterns of biological and translational interactions from gene expression data. As most of such methods are dependent on simple correlation of changes in the expression, their sensitivity and specificity are limited by the highly heterogeneous nature of tumor expression patterns.

Current efforts are asking different questions than asked in the previous decade. Rather than taking a high-level approach to investigate systematic differences in healthy versus diseased samples, researchers are now asking questions regarding subtypes of disease. Resources such as the International Genomics Consortium's exp*O* biospecimin repository (http://www.intgen.org/expo/), which as of this writing contains over 16,000 cancer biosamples, allow the kind of data researchers need to investigate the similarities and differences across different cancer subtypes. The availability of exp*O* and resources like it demands the creation of mathematical models which take into account the context-specific behavior of the true biological system.

Motivated by these demands and the limitations of other approaches, we first detail background work on a mathematical model for contextual genomic regulation and a method to identify cellular context motifs from biological data. This will provide the framework for us to describe our specific contributions: a new graphical representation of the contextual genomic regulation model called a *context-specific gene regulatory network*, and a new analysis methodology based upon this graphical representation. We show the results of our work on the Target Now data set and conclude the section with a discussion of related work and future directions.

4.2 Related Work

For the reasons discussed, identifying patterns specific to certain biological contexts is gaining more interest [11, 12]. One of the early approaches to identify context-specific patterns is searching for the co-regulated sets of genes and showing the relationships between the gene sets and the biological/clinical characterization of samples [13, 14]. However, these approaches still require the knowledge of the context to identify patterns when such knowledge is often incomplete. In terms of the data sets themselves, we can think of *contexts* as intelligent groupings of subsets of variables with associated subsets of samples. When considering a grouping of a subset of data samples with a subset of the variables, bicluster-

ing [60] is an approach that immediately comes to mind. While biclustering is effective at identifying significant groupings of samples and variables, most methods lack the ability to group similar biclusters and give overlapping results. Great progress was made with Segal *et al.*'s module networks [8], which used existing knowledge sources as well as clustering to find gene sets that are either functionally correlated or expressed in a similar way in each set, and then identified their specificity to the given conditions of samples. However, these approaches do not find gene sets that show co-regulation under certain conditions, neither do they provide information on the relevance between gene sets. Grzegorczyk *et al.*proposed a method to learn a non-homogeneous Bayesian network[10] that can represent multiple different conditions, but the complexity of the model limited the scalability of its application to eleven biological entities (e.g., genes) at the most. In our work we strive to overcome some of these limitations.

4.3 Background

In this section we first describe the mathematical model for contextual genomic regulation originated by Dougherty *et al.*[61] from which we base our work. Following this overview we briefly describe related algorithmic work and terminology from Kim and Sen [62, 63, 64] relevant to the contributions outlined in this dissertation.

Mathematical Model for Contextual Genomic Regulation

The framework used here defines context-specific relationships as a functional mapping of sets of regulatory (driver) elements to the activities of the regulated (driven) elements constituting the system. It is important to select a mathematical model for a cell's regulatory activity that actively adjusts to differing internal and external environmental factors. Rather than models which infer connections between single genes, or between genes and phenotypes, we wish 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. For the cellular domain, this has been pioneered by the mathematical model for contextual genomic regulation [61, 62].

Formally, this 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. Under a set of common conditions T_i (samples), the driver genes, driven genes and their samples form a unit known as a *context motif*.

Two parameters are essential to the definition of the contextual genomic regulation model, namely *conditioning* and *crosstalk*. These parameters are defined formally in related papers [61, 62, 63, 64], but in general, conditioning depends on the extent that contextual effects diminish the influence of the driver on the driven gene, and crosstalk depends on the extent that contextual effects outside of the drivers activate the driven genes. The paper by Dougherty *et al.*[61] discusses further considerations of the model, including prediction accuracy and error representation.

Identifying Context Motifs

As described, 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 the two statistics – conditioning and crosstalk – to identify context motifs in gene expression data.

Kim and Sen describe an algorithm, *Cellular Context Mining* to identify context motifs, where context motifs have significantly low conditioning and crosstalk values across the associated samples[62, 64]. One major step in cellular context mining is *in-silico* conditioning, a step designed to be similar to a biologist manipulating the status of a gene or

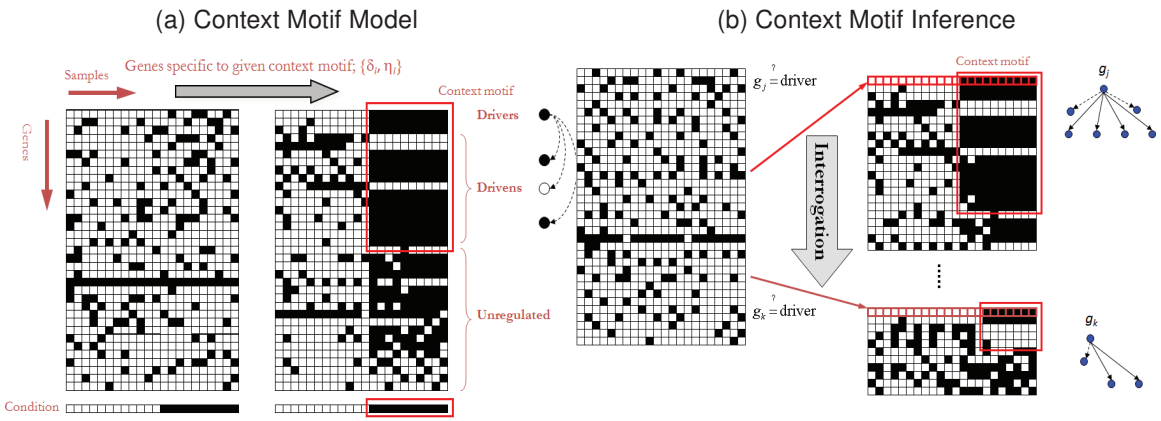


Figure 4.1: Modeling and Inference of Context Motifs: a) Depicts the context motif model with crosstalk and conditioning parameters. b) Provides a schematic for the cellular context mining algorithm. Figure ©Ina Sen[64]. Used with permission.

conditioning cells in an experiment with techniques including ectopic expression or gene silencing. With *in-silico* conditioning, the conditioning and the observations are not performed manually as the data is collected, but rather computationally after the data has been collected, hence the name. In the work described in this chapter we only consider a single gene driver at a time for conditioning, although the model allows for more. Figure 4.1 displays context motif modeling and inference of context motifs by interrogating each gene as a possible driver.

4.4 Methodology

While cellular context mining is indeed successful at producing coherent context motifs, it was lacking in interpretability due to the reporting of identified context motifs in either a list, tree or tabular format. In a recent work, Kim *et al.*[65] present a graphical depiction of up to two context motifs for a small cancer network. To expand, systematize, and to address shortcomings in visualization and interpretation, we developed the formalized graphical representation of the results of cellular context mining called context-specific gene regulatory networks, as well as an intuitive analysis approach based on the new graphical depiction.

Context-Specific Gene Regulatory Networks

Our recent study [63] asserts that the context-specific gene regulatory networks produced by cellular context mining exhibit biological advantages absent in related techniques. We first note that driver g_j in the context motif C_j might be driven by a g_i in another context motif C_i. The chaining of such regulatory relationships $g_i \rightarrow g_j$, in addition to implicit driver-driven relationships $g_i \rightarrow g \in S_i$, produces an interesting graphical structure as shown in Fig. 4.2, representing relationships between context motifs. We call this a context-specific gene regulatory network because each regulatory relationship $g_i \rightarrow g \in S_i$ is specific to corresponding subset of samples, T_i.

Originally described[1] by Kim *et al.*[65], a context-specific gene regulatory network 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 gene regulatory network H is a pair $H = (V,E)$, where V is a set of gene-representing vertices and E is a set of edges oriented from genes designated as drivers to genes designated as driven; 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 $g_i \rightarrow g \in S_i$.

Identifying Context Motif Clusters

While replete with contextual information, the context-specific gene regulatory network alone is not readily interpretable. The large number of context motifs will likely have some topological areas of interest, such as dense subgraphs or components. In order to make any meaningful discoveries, however, the context-specific gene regulatory network must

[1] Kim *et al.*[65] describe the general idea of chaining context motifs and show preliminary graphs of one and two context motifs including only one expression level per gene per graph. Our work formalizes and extends this concept and also contributes the analysis approach.

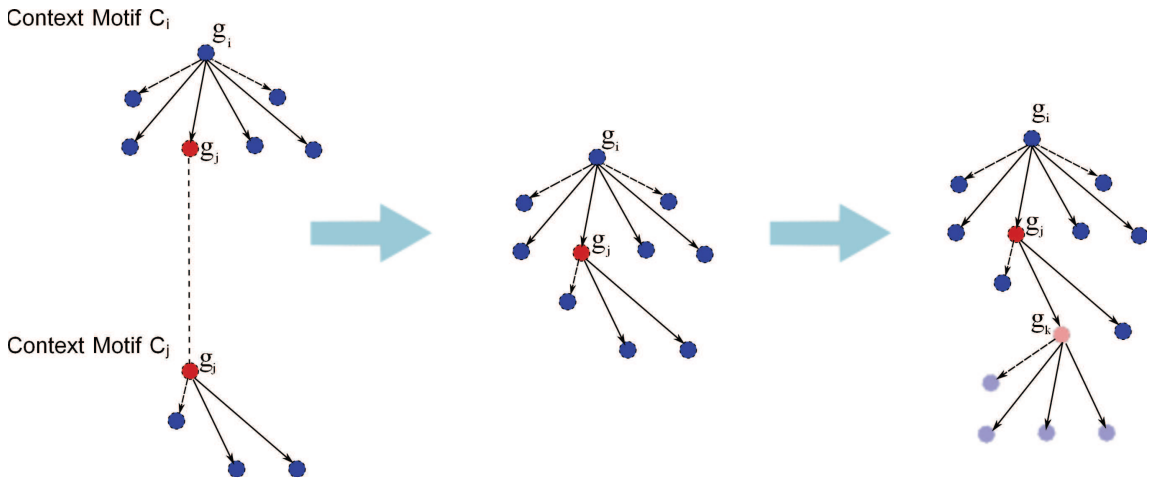


Figure 4.2: Chaining Context Motifs to Form a Context-specific gene regulatory network: The chaining together of context motifs forms the context-specific gene regulatory network. Analysis of the context-specific gene regulatory networkwill reveal context motif clusters. Figure ©Ina Sen[64]. Used with permission.

be segregated into not just topologically interesting components, but rather *biologically meaningful* components. We call these components or subgraphs *context motif clusters*, and denote a context motif cluster C to be a set of m context motifs $C_1, C_2, \cdots, C_m$. To obtain the best context motif clusters, we contribute the following *ad hoc* approaches: In general one first considers any topological clues: an analysis of the weakly connected or strongly connected (if present) components in the network topology is a natural place to begin segregating into subnetworks. Within these components, the regulatory control inherent in the driver-driven relationships is useful in identifying separation points between portions of weak or strong components. Additionally, expression values lend insight; because each variable in a context-specific gene regulatory network built from a quantized data set is present in each level of quantization (e.g. up to three nodes for a ternary gene variable found to be up- or down-regulated or no change), one can quickly identify portions of the network exhibiting similar expression behavior. Our example following these approaches is described in our preliminary results section for the Target Now data set. In general, context motif clusteridentification steps include the following separation criteria:

- Weakly Connected Components (topological)
- Strongly Connected Components (topological)
- Large Subgraphs Separated by Few Edges (topological)
- Subgraphs Differing in Expression Value Consensus (biological)
- Subgraphs Differing in Disease Subtype Enrichment (biological)

Sample Association to Context Motif Clusters

Depending on the data set under study it is possible to compute sample association for context motif clusters. Using the Sample Association Score (Eq. 4.1) one can identify the subset of samples featured most prominently among the included context motifs in a particular context motif cluster. Additionally, if further sample category information is available, further enrichment scores of particular categories is also possible through the Yates corrected Chi square test for significance[63]. As the Target Now data set (see §4.5) has readily annotated samples each associated with a particular tissue type, both of these calculations are reported in the results of our analysis of the Target Now data.

Formally, the sample association score was developed [63] to associate a sample s with a context motif cluster C consisting of m context motifs $C_1, C_2, \cdots, C_m$, with the score:

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

where k_i is the number of samples within context motif C_i and N is the total number of samples in the gene expression data. Thus, the sample which occurs in all context motifs of the context motif cluster would have the least score and the sample not present in any would have the perfect score of 1. Samples with scores less than 0.5 are associated to the corresponding context motif cluster and are used for computing the tumor type distributions across all context motif clusters.

Bayesian Network Analysis for Comparison

The Bayesian network model is a popular tool for modeling gene regulatory networks and thus serves as a good point of comparison for new graph-based methods such as ours. Here, we used a hybrid algorithm of hierarchical clustering and order restriction (H-CORE)[66] and sparse candidate (SC)[67] to learn genome-wide Bayesian network structures from a dataset. To compensate for the inadequate amount of observed data, we applied a k-fold bootstrapping in learning gene regulatory networks with Bayesian network learning.

We built K subsets $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_K$ by randomly selecting $(k-1)/k$ proportion (for example, 90% if $k = 10$) of samples with replacement from $\mathbf{D}$. We applied the hybrid algorithm of H-CORE and SC to learn a Bayesian network structure H_i from each $\mathbf{D}_i$, and built K Bayesian network structures. The likelihood of an edge $g_i \rightarrow g_j$ to be in the final network H_{BN} was evaluated as given in Eq. 4.2. The final H_{BN} only included the edges having likelihood L larger than 0.5.

$$L(g_i \rightarrow g_j) = \frac{\sum_{\forall H_l \text{ with } g_i \rightarrow g_j} Pr(H_l \mid \mathbf{D}_l)}{\sum_{k=1}^{K} Pr(H_k \mid \mathbf{D}_k)} \tag{4.2}$$

4.5 Results

We applied our methodology to the Target Now data set, which includes gene expression profiles of 146 patients with refractory cancer, and inferred a context-specific gene regulatory network using the Cellular Context Mining algorithm. Amalgamation of identified cellular context motifs yielded a context-based network structure. Biological annotation according to sample composition of context motif clusters as well as literature verification of gene-disease relevance was performed. Our study shows that not only do context-specific gene regulatory networks report verifiable (and possibly novel) relationships between genes, but moreover, the overall network structure can be formed into bi-

ologically meaningful and readily annotated context motif clusters.This is unique among network learning techniques, which we illustrate with a comparison to Bayesian network approaches.

Target Now Data Set

In applying our method we used the gene expression profile of the Target Now[2] (TN) study. TN study aims to determine if patients with refractory cancer, who did not benefit from the 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 stage cancer. Late stage cancer is very frequently 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 (Agilent-011521 Human 1A Microarray G4110A) from 146 patients with different types of refractory cancer, was used to learn context-specific gene regulatory networks. For each tumor type, its (normal) tissue of origin was used as a baseline and the ratio of the tumor to its tissue of origin was computed and the ratio value was ternarized using a statistical model [68]. The dataset was then pre-filtered based on transcription activity of each gene across the samples to be reduced to only 4,000 probes. The Target Now dataset sample distribution of the 146 samples between different cancer tumor types is listed in Table 4.1.

[2]http://www.caristargetnow.com/

Table 4.1: Target Now Sample Distribution: Numbers of samples associated with different cancer tumor types are shown in the table. *Single Sample Types*: Appendix, Cartilage, Chondrosarcoma, Eccrine Adenocarcinoma, Glioma, Gastric, Ileum, Lymphoma, Momocytes, Prostate, Uterus, Rhabdomyosarcoma, Synovial Cell Sarcoma, Skeletal Muscle, Testicular

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

Context Motif Clusters in Target Now Data

Running the context mining algorithm with a strict statistical significance threshold resulted in 205 context motifs with p-value < 0.0005. Using these context motifs, the method described to create context-specific gene regulatory networks yielded a directed graph with 1,790 vertices (genes) and 9,566 edges (regulatory relationships), as shown in Fig. 4.3. This graph had an interesting property of being systematically fragmented into four separate context motif clusters, which were identified by locating the weakly connected components in the graph. These context motif clusters provide a useful approach to interpreting the context motifs found by context mining. The context motif clusters typically display significant overlaps among their subsets of samples. This is due to complex inter-connections among drivers that result from particular commonly shared cellular processes.

When investigating the four separate context motif clusters, we noticed the two largest context motif 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 context motif clusters. In Fig. 4.3, bottom right, we segregated the first large cluster into context motif clusters C and G. Context motif cluster G is easily separable as all its genes are neither under- nor over-expressed (unlike C), and only one edge exists between the context motif clusters C and G

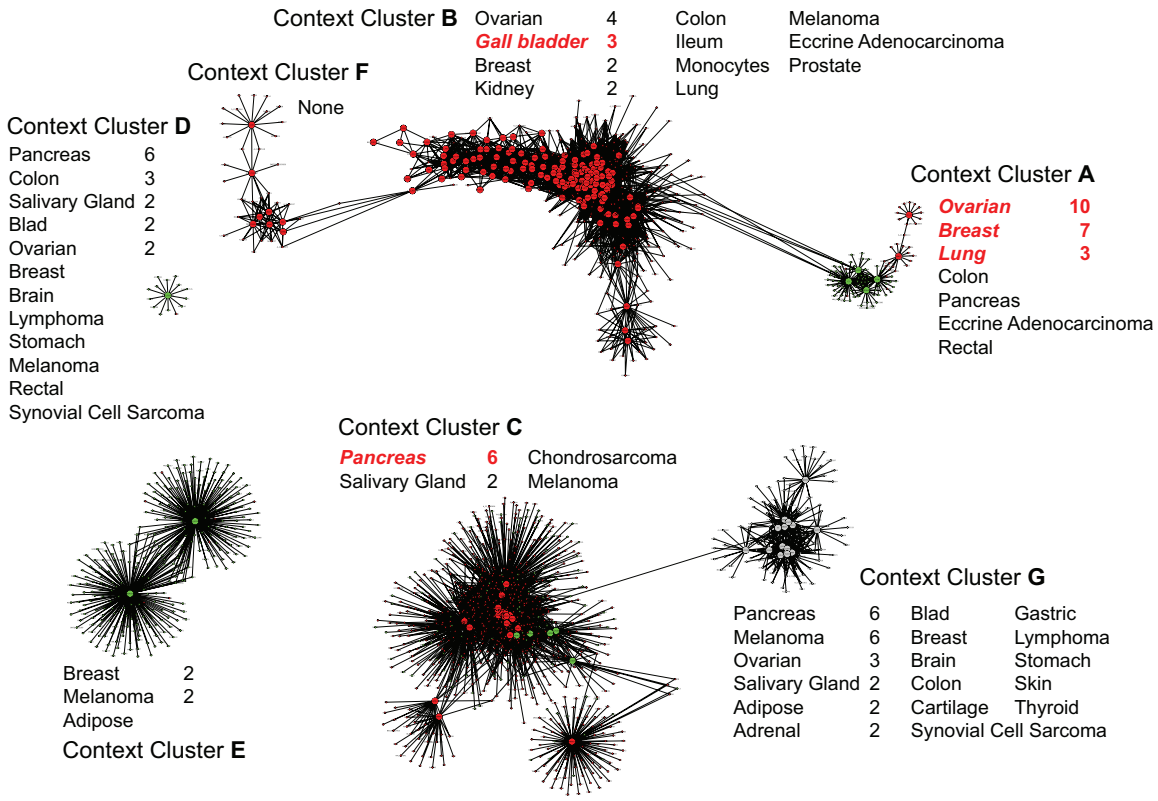


Figure 4.3: Annotated Context Motif Clusters in a Context-specific gene regulatory network: Each context motif cluster (abbreviated "Context Cluster") is annotated with the corresponding set of samples and highlights significantly enriched tumor types in red. See Table 4.1 for cancer tumor sample distribution in the data set. In the graph itself, red vertices represent over-expressed genes, green under-expressed, and gray neither under- nor over-expressed. Edges are oriented from driver genes (large vertices) to driven genes (small vertices).

(C drives a gene also driven by G). These characteristics convinced us that C and G should be analyzed as separate context motif clusters. 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 Fig. 4.3 into context motif 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. Context motif

clusters B and F share 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 again a possible hierarchical regulatory relationship between the two.

On the basis of density of connection and directionality of control, we resolved the four original context motif clusters into seven biologically separable ones. Each of these seven context motif clusters are visible in Fig. 4.3 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 (Eq. 4.1). See Table 4.1 for details of the data set sample composition.

Tumor Type Enrichment

Recall that for the Sample Association Score of Eq. 4.1, a sample which occurs in all context motifs of the context motif cluster would have the smallest score, and a sample which is not present in any of the context motifs will have a maximal score of 1. TN samples with scores less than 0.5 were thus associated to their corresponding context motif clusters. Using only the selected samples, sample distributions and tumor type enrichments were calculated across all context motif clusters as described next. Fig. 4.3 depicts the tumor types having non-zero sample counts corresponding to each context motif cluster.

After sample association to specific context motif clusters, each cluster was subjected to a statistical test for enrichment of specific types of tumors. The Yates corrected chi square test for significance was applied (since some numbers were less than 5) to each tumor type-context motif cluster pair. Some of the significant results are summarized in Table 4.2.

Context Motif Clusters Identify Disease-Pertinent Genes

Fig. 4.3 highlights (in red) the tumor types considered to be enriched within the corresponding context motif cluster. Intriguingly, context motif cluster A showed significant

Table 4.2: Context Motif Cluster Tumor Enrichment: Chi-square enrichment test p-values of tumor types in different context motif clusters are shown in the table.

Context Motif Cluster A		Context Motif Cluster B		Context Motif Cluster C	
Tumor Type	p-value	Tumor Type	p-value	Tumor Type	p-value
Ovarian	2.30E-05	Gallbladder	1.60E-04	Pancreas	8.20E-05
Breast	0.0057				
Lung	0.0120				

tumor enrichment of ovarian cancer, breast cancer and lung cancer. A literature survey shows that breast cancer drugs are being used in the treatment of lung cancer [69], because of the vital role of estrogen in lung development, and subsequently the cancer pathway.

Conventional approaches such as clustering and Bayesian network learning provide some ability to observe sample enrichment, but they do so in ways that 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 some conventional clustering algorithms, such as hierarchical clustering and k-means clustering, using the similarity metrics of correlation and Euclidean distance, using Cluster version 3.0 [70], to group samples with similar gene expression profiles. We were able to verify that in cases where a similar number of context motif clusters (six or seven) were identified by conventional methods, the conventional clusters display significant overlap (ranging from 40% to 90% overlap) with context motif 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 gene activity for the subtypes 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.

Some examples of known gene interactions and relationships to diseases within context motif clusters were verified through a literature survey. Context motif cluster A involved breast cancer, ovarian cancer and lung cancer, and included genes such as TNFRSF1A, which is known to promote breast cancer [71]; CD74, usually expressed in ovarian and lung cancers, is being considered as a target for Multiple Myeloma treatment therapy [72]; HLA-DM, when its expression is combined with that of HLA-DR, is considered to influence breast tumor progression and patient outcome [73]. Context motif cluster C, related to pancreatic cancer, contained GP2, a protein specifically expressed in pancreatic acinar cells and considered as a diagnostic marker in animals [74].

Comparison to Bayesian Network Analysis

Context motif clusters (A $\sim$ G) were compared to H_{BN}, which is the result from the Bayesian network analysis. H_{BN} was composed of many subgraphs ($H_{CC}s$), which were connected components, and the 32 largest $H_{CC}s$ were chosen for comparison. The degree of overlap was evaluated for each pair (context motif cluster, H_{CC}) using the geometric mean of common gene ratios for the context motif cluster and the H_{CC}. After considering the empirical distribution of the degree of overlap, a pair with a degree of overlap larger than 0.162 was determined to share a significant amount of genes. To figure out the difference between two results, enrichment analyses for Gene Ontology (GO)[75] terms were conducted for every context motif cluster and every H_{CC} using GoMiner [76].

Fig. 4.4 compares context-specific gene regulatory network and Bayesian network analysis. The comparison revealed 10 $H_{CC}s$ with a significant number of shared genes with at least one context motif cluster (Shared = $H_{CC,1}$, $H_{CC,2}$, $H_{CC,3}$, $H_{CC,7}$, $H_{CC,16}$, $H_{CC,17}$, $H_{CC,19}$, $H_{CC,26}$, $H_{CC,30}$, $H_{CC,32}$). When $H_{CC}s$ in Shared were subject to GO term enrichment

analysis, most of such $H_{CC}s$ did not have any enriched GO term (8 of 10). On the contrary, other $H_{CC}s$ with no significantly shared genes (NonShared) were often enriched with GO terms (19 of 22). The reason can be as follows: Bayesian network learning assumes that the observed data is from a single distribution, and attempts to capture information consistent across all samples in the observed data. On the contrary, analysis of context-specific gene regulatory networks captures information consistent in subsets of samples. If an H_{CC} shares a significant amount of genes with a context motif cluster , it means that a significant portion of information in the H_{CC} is consistent only in some subset of samples. This is against the viewpoint of Bayesian network learning, because the H_{CC} was built with a bias toward the process of capturing information consistent across all samples. For this reason, the shared genes in an H_{CC} may have inconsistent information with those unshared genes in the H_{CC}, eventually making it hard for the H_{CC} to have enriched GO terms. The conventional Bayesian network learning is therefore not an optimum choice for identifying context-specific information from some subset of samples.

This comparison and the tumor enrichment studies of context motif clusters in previous sections shows that the context-specific gene regulatory network has a novel ability to represent context-specific information from subset of samples while conventional approaches assuming the entire sample set to be from a single distribution have a harder time recognizing this kind of behavior.

4.6 Chapter Summary

Our work presents a formal approach to generate context motif clusters, i.e. disease-pertinent (cancer tumor type in case of TN dataset) gene regulatory networks, using our novel analysis methodology applied to the context-specific gene regulatory networks. Our study asserts that these gene regulatory networks exhibit biological advantages absent in related techniques. The mapping of inter- and intra-context edges results in a graph of context motif clusters which is found to have interesting properties, such as specific can-

	HCC1	HCC2	HCC3	HCC4	HCC5	HCC6	HCC7	HCC8	HCC9	HCC10	HCC11
A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.02
B	0.32	0.00	0.03	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00
C	0.02	0.19	0.01	0.00	0.11	0.00	0.00	0.00	0.00	0.02	0.05
D	0.42	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E	0.00	0.00	0.00	0.14	0.00	0.01	0.00	0.00	0.00	0.14	0.00
F	0.00	0.00	0.53	0.00	0.00	0.00	0.51	0.00	0.00	0.00	0.00
G	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

	HCC12	HCC13	HCC14	HCC15	HCC16	HCC17	HCC18	HCC19	HCC20	HCC21	HCC22
A	0.05	0.00	0.00	0.00	0.00	0.25	0.05	0.00	0.00	0.00	0.03
B	0.00	0.00	0.00	0.00	0.42	0.00	0.00	0.22	0.00	0.00	0.00
C	0.00	0.00	0.01	0.08	0.05	0.07	0.11	0.11	0.03	0.01	0.00
D	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08
F	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
G	0.00	0.00	0.00	0.00	0.39	0.00	0.00	0.04	0.00	0.00	0.00

	HCC23	HCC24	HCC25	HCC26	HCC27	HCC28	HCC29	HCC30	HCC31	HCC32
A	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.03	0.00
B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.00	0.17
C	0.01	0.00	0.00	0.01	0.00	0.08	0.07	0.00	0.00	0.00
D	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.02	0.00
F	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
G	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.31	0.00	0.21

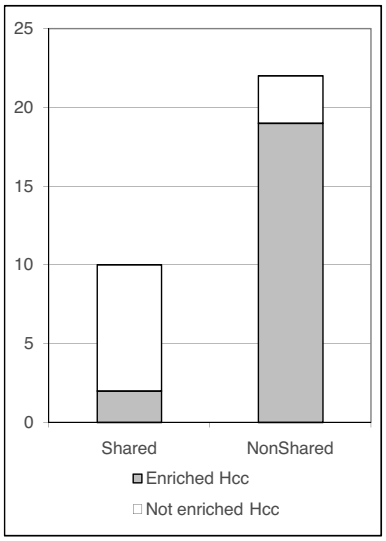


Figure 4.4: Bayesian Network Analysis of TN Data: The left table shows the degree of overlap for each pair of context motif clusters A $\sim$ G and a subgraph ($H_{CC,1} \sim H_{CC,32}$) from Bayesian network analysis. A shaded cell in the table represents the pair with a significant level of shared genes. A shaded context motif cluster or H_{CC} has at least one enriched GO term while non-shaded ones have no enriched GO terms. The right graph shows the portion of H_{CC}s with enriched GO terms for H_{CC}s with a significant number of shared genes (**Shared**) and H_{CC}s with no significant number of shared genes (**NonShared**).

cer tumor type enrichment (used for context motif cluster annotation) and occurrence of genes relevant to the annotated diseases. Comparison of this approach with conventional clustering algorithms demonstrated its advantage for relevant gene subset identification. When compared with Bayesian network analysis, we noted that context motif clusters can capture regulatory relationships specific to subsets of samples while conventional Bayesian network learning rarely captures meaningful context-specific information.

While our work on this topic has concluded, the results have already laid the ground work for related research. Based on our preliminary work the terms *context motif* and *context motif cluster* have been formalized. The *context motif clusters* from our manuscript[63] have now evolved into the general term *context* thanks to work by Ramesh *et al.*[77] and by Sen[64]; Work by Ramesh *et al.* replaced the original *ad hoc* analysis approach with systematic graph clustering, comparing the results of Markov clustering with two variants of

spectral clustering. Work by Sen *et al.* has simplified context-specific gene regulatory networks with novel pruning techniques, and has advanced them to incorporate multi-source data as well as multivariate drivers[64].

Alongside these works, our work on developing contextual gene set interaction networks (detailed in the coming chapter) also uses the work of this chapter as a branching off point towards improved interpretation capabilities and thus patient benefit. Based on clustered context-specific gene regulatory networks and addressing remaining interpretability limitations, we reduce the quantity of reported interactions with annotation-based summarization of gene sets and reveal which network components are general and which are context-specific.

CHAPTER 5

CONTEXTUAL GENE SET INTERACTION NETWORKS

As mentioned, researchers are now asking specific questions regarding subtypes of disease. Resources such as the International Genomics Consortium's exp*O* biospecimin repository (http://www.intgen.org/expo/), which as of this writing contains over 16,000 cancer biosamples, allow the kind of data researchers need to investigate the similarities and differences across different cancer subtypes. The availability of exp*O* and resources like it demands the representation of the context-specific behavior of the true biological system.

We have developed methods to learn and annotate context-specific gene regulatory networks based on a contextual genomic regulation model [62, 63, 77]. While the model[63] defined in the previous chapter provides an interesting and useful context-specific gene regulatory network and concordant context motif clusters, its weakness is in interpretability. Because of its inherent usefulness, however, we are motivated to utilize it as the basis for a more readily interpretable network model. As such we describe our contribution of learning contextual gene set interaction networks and identifying context specificity.

5.1 Motivation

Identifying similarities and differences in the molecular constitutions of various types of cancer is one of the key challenges in cancer research. The appearances of a cancer depend on complex molecular interactions, including gene regulatory networks and gene-environment interactions. This complexity makes it challenging to decipher the molecular origin of the cancer. In recent years, many studies reported methods to uncover heterogeneous depictions of complex cancers, which are often categorized into different subtypes. The challenge is to identify diverse molecular contexts within a cancer and to relate them to different subtypes so that we can recommend context-specific treatment to patients.

In this study, we describe a method to discern molecular interactions specific to certain molecular contexts. The method first identifies sets of genes of which each gene set shares coherent transcriptional patterns across a subset of samples, termed *contextual gene set*. Unlike conventional approaches to building modular genetic networks, our focus is to identify cancer-generic and subtype-specific interactions between contextual gene sets by quantifying the effect of the samples from each subtype on the calculated strength of interactions observed. Two cancer data sets were analyzed to support the validity of identifying condition-specificity. The results revealed that network components specific to different types of cancer are related to different biological functions from cancer-generic network components. We also found results that are consistent with previous studies, as well as new hypotheses on the biological mechanisms specific to certain cancer types.

The analysis on the contextual gene sets and characterization of networks of interaction composed of these sets discovered distinct functional differences underlying various types of cancer. The results show that our method successfully reveals many subtype-specific regions in the identified maps of biological contexts, which well represent biological functions that can be connected to specific subtypes.

5.2 Related Work

Many computational and mathematical techniques have been developed to infer molecular patterns of biological and translational interest from gene expression data profiled from human tumors. As most of these methodologies are highly dependent on simple correlation of changes in mRNA abundance as the primary measure of relatedness, they are intrinsically limited in their sensitivity and specificity by the highly heterogeneous, idiosyncratic nature of tumor gene expression patterns. Early expectations were that the molecular pathology of tumors arising from a particular tissue of origin would show striking similarities due to very common sets of oncogenic molecular processes accounting for each such tumor type's initiation and progression. The finding that samples of tumors taken at different points in

the course of an individual's disease were more similar to each other than to any other tumor in the study was therefore quite surprising and stands as a noteworthy finding from early expression profiling studies [78]. One tumor type, chronic myelogenous leukemia (CML), has been found to have a large number of genes behaving in a very homogeneous fashion [79], however this kind of behavior has been the exception rather than the rule. The relative homogeneity of CML is probably in line with the expectation that a cancer type would exhibit high homogeneity if there was a high similarity in the process of oncogenesis. In the case of CML, it is true that the means of transformation is simple and constant. As the biochemical mechanisms of tumor growth and survival have been subjected to ever more detailed analysis, it has now become clear that for most tumor types there is substantial variation in how tumors use available normal and altered cellular functions to achieve relentless growth and disproportionate survival.

In recent studies, the identification of genomic patterns that are specific to certain biological contexts is gaining more interest as the heterogeneity in biological data becomes better embraced. *Biological contexts* of interest can derive from subtypes of diseases or different clinical outcomes within the same subtype, such as responses to therapy. One of the early approaches to identify context-specific patterns involved searching for the co-regulated sets of genes and depicting the relationships between the gene sets and the biological or clinical characterization of samples. Gasch and Eisen [13] used a modified fuzzy *k*-means clustering method to find gene sets and showed correlation between those gene sets and the experimental conditions that determined how yeast cells respond to environmental changes. Segal et al. [14] used existing knowledge sources as well as clustering techniques to find gene sets that are either functionally co-related or coherently expressed in each set, then determined their specificity to particular types of tumors.

More recently, new strategies to identify context specificity of biological interactions are being proposed, as it is being widely accepted that biological interactions are co-

ordinated in systematic ways but distinctively so, depending on biological contexts. These approaches are often based on network models representing biological interactions. Identifying context specificity in biological interactions can reveal environmental conditions under which the activities of components of biological networks vary, and this can make significant contribution to reinforcing confidence in the network as well as to improving the understanding of more exact mechanisms of transcriptional or translational regulations. Several studies have shown that it is possible to identify context-specific activity in known biological networks [15, 16, 17, 18]. In these studies, biological networks were built from existing prior knowledge, and context-specific gene expression or protein expression data was used to annotate biological interactions' context specificities. While this approach is useful for understanding the context specificity in already characterized networks, its utility is limited by the scope of the network. Because context-specificity is identified on top of networks from prior knowledge, it cannot make findings related to novel interactions. Because of such limitation, a more desirable approach is to simultaneously identify biological interaction networks and their context specificity from high-throughput data. Grzegorczyk et al. [80] proposed a method to learn a non-homogeneous Bayesian network that can represent multiple different conditions, by using a mixture model that can unify networks of different conditions. This approach of using mixture models can be ideal in situations where enough amount of samples and computing resources are available. However, the scalability of its application is significantly limited in practice due to the complexity of the model. Indeed, their study only show the results from networks of up to 11 genes. Another possible approach is applying conventional network learning methods to subtype-specific samples and comparing the results. However, this approach requires many samples for each condition to achieve reliable results, and comparing inferred models from different conditions can be arguable especially when different amount of samples were available across conditions.

5.3 Background

In this study, we use a method to learn contextual gene set interaction networks that can represent maps of functional modules in target biological systems, and identify condition/subtype-specificity on the inferred networks. Our method takes two novel approaches; the first is using context-specific gene sets as the nodes of networks, and the second is measuring condition-specificity with a formulation based on the Bayesian network model. Using gene sets instead of individual genes as nodes in networks can significantly increase the scalability of the application. However, using conventional gene clusters that show similar expression across all samples will make it difficult to identify context specificity of interactions in networks, as such gene clusters will have mostly the same value across the samples. In our approach, we use contextual gene sets that show similar expression under certain subset of samples, where such subsets of samples are identified as potential biological contexts by using the method for context-specific genomic regulations [81, 82, 77]. Contextual gene sets have specific expression values only for a subset of samples where they have statistically significant coherency, which helps to identify networks and relevant condition-specificity. Another key aspect of our approach is using the conventional homogeneous Bayesian network model to learn networks and to measure condition-specificity. Using homogeneous network models requires significantly less computational cost than using non-homogeneous mixture models. However, the homogeneous model does not represent any condition specificity by itself. To overcome this limitation, we designed a formulation to quantify the effect of the samples from different conditions/subtypes on the formation of networks to measure the degree and the statistical significance of condition specificity. (The brief results of a simulation study will also be given to show the validity of identifying condition-specificity.)

5.4 Methodology

Learning contextual gene set interaction networks and identifying condition specificity involves several steps of data transformation, as illustrated in Fig. 5.1. It consists of four major steps[1]: (STEP I) identifying contextual gene sets as basic functional modules, (STEP II) summarizing contextual gene sets to transform the data of genes to the data of contextual gene sets, (STEP III) learning contextual gene set interaction networks where each interaction represents dependency in expression (specific expression status of a gene set depends on the expression status of the other gene set) between two contextual gene sets, and (STEP IV) identifying condition specificity of each interaction. For the first step of identifying contextual gene sets, we define a contextual gene set as a set of genes that show consistent expression pattern under a biological context, i.e. a subset of samples. This is based on the assumption that once a biological context achieves a steady state, genes involved in the process show consistent transcriptional patterns under the biological context. Identifying contextual gene sets first requires the identification of samples where biological contexts are involved, and we use the *context mining* algorithm[81, 77] to find such contextual conditions, where a contextual condition is a subset of samples where groups of closely related coherent expression patterns are found. Under each contextual condition, sets of genes with similar over-expression or under-expression are identified as contextual gene sets (STEP I in Fig. 5.1).

To infer networks of contextual gene sets, each contextual gene set is represented as a single variable/node. This requires the original gene expression matrix to be transformed to a gene set expression matrix, where the value of a contextual gene set for a sample is a representative value of all genes in the contextual gene set. Expression values of genes in a contextual gene set for a sample are summarized to either UP or DOWN if

[1]While the entire approach was a collaborative process, the specific personal contributions of the dissertation author were to STEP III, STEP IV and to the validation and annotation work thereafter.

the majority of the genes are over-expressed or under-expressed, and NOCHANGE value is given otherwise (STEP II). We are going to focus on the cases of statistically significant up-regulation or down-regulation, and most results from this study are from the cases of up or down-regulations.

A contextual gene set interaction network is learned from the summarized contextual gene set expression data, by evaluating the likelihood of dependency between each pair of contextual gene sets given all samples and building a connection if the dependency likelihood is larger than a given threshold (STEP III). Inference of interaction networks from the summarized data has a few advantages over traditional approach where all genes are used. Since the number of variables (nodes) is significantly smaller in this approach as all the genes in contextual gene set are aggregated to a single variable, the method suffers less in both computational complexity and statistical estimation accuracy.

A resultant interaction between two contextual gene sets represents that there is a probabilistic dependency in their summarized expressions. Gene sets with dependency are expressed in coordinated manners, where the expression status of a gene set depends on the expression status of the other gene set. However, the effect of the dependency from the samples can be different for diverse conditions, as they can imply different activities of biological functions. Based on this idea, we identify condition-specific regions in the built network by measuring the effect from the samples of each condition on the likelihood of dependency. To measure the effect of a condition on a dependency, we evaluated the likelihood of the dependency without the samples of the condition and computed its difference with the original likelihood obtained using all available samples (STEP IV). If the original likelihood is significantly higher than the likelihood without the samples from the condition, it means that the samples under the condition have made significant contribution to the dependency. This implies that the dependency exists mainly due to the samples from the condition, thus it is declared as a condition-specific dependency.

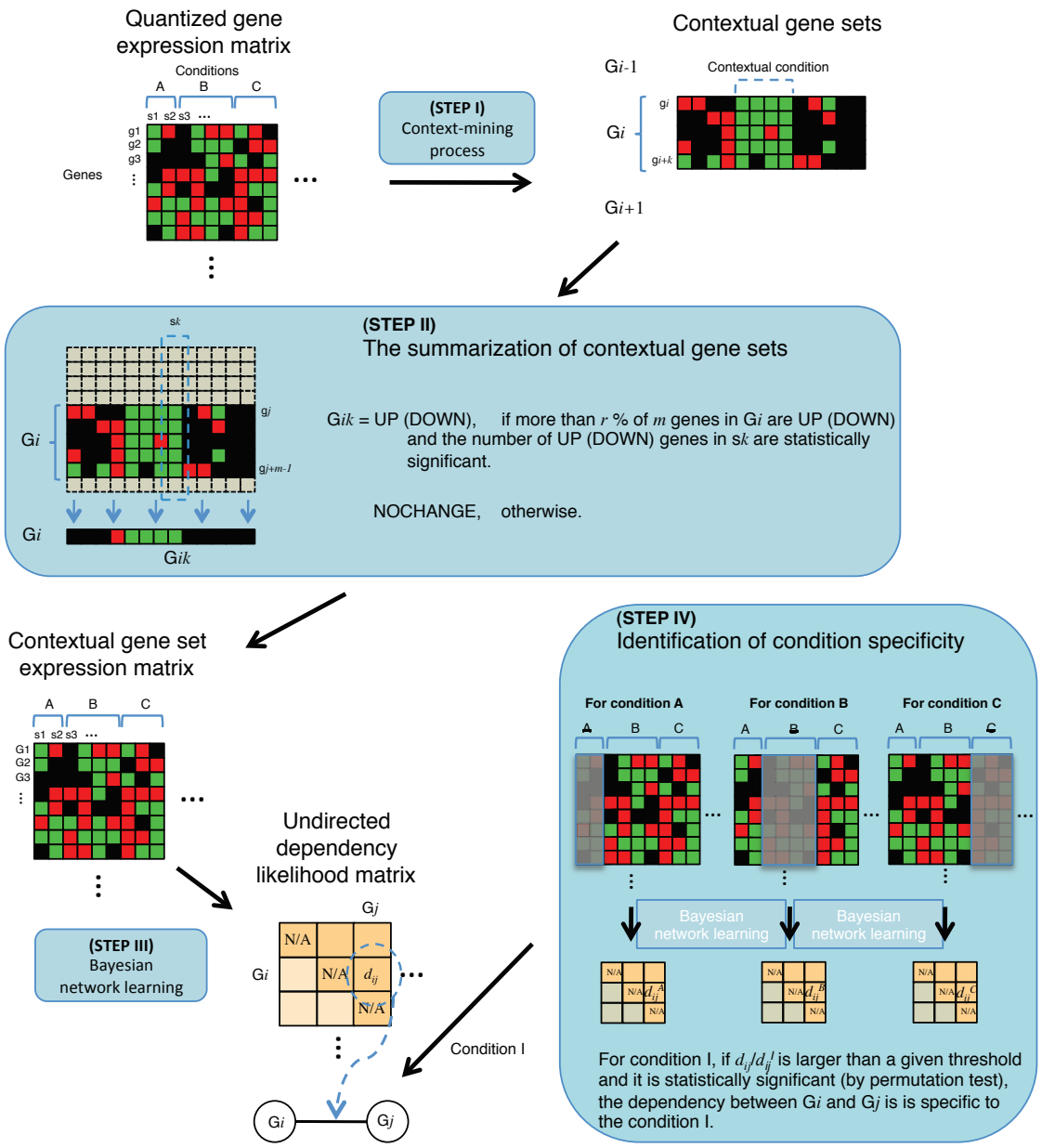


Figure 5.1: Schematic Overview of Learning Contextual Gene Set Interaction Networks and Identifying Condition Specificity: From the gene expression matrix, contextual gene sets are identified through the *context mining* process. The expression values of genes in each contextual gene set for each sample are summarized into one major representative value, and a contextual gene set expression matrix is built as a result. Multiple Bayesian networks are learned from this matrix and their consensus network (undirected dependency likelihood matrix) is built while ignoring the direction of connections. For each condition, a subset of data is built by discarding the samples of the condition from the original data and a new dependency likelihood matrix is built from it. If the dependency likelihood of the interaction between $\mathbf{G}_i$ and $\mathbf{G}_j$ from all samples is significantly larger than the dependency likelihood from a data without a condition I, the interaction is specific to the condition I.

STEP I: Identifying Contextual Gene Sets

We define a contextual gene set as a set of genes that show consistent expression pattern under a biological context. This is based on the assumption that once a biological context reaches a steady state, genes involved in the process show consistent patterns under the biological context. It requires the identification of subsets of samples, which are representations of biological contexts. We use the *context mining* algorithm [81, 77] to find contextual conditions based on these samples, where a contextual condition is a subset of samples that have groups of closely related coherent expression patterns (such patterns are called *context motifs* in the algorithm). In the process of context-mining, two consistency statistics *conditioning* (δ) and *crosstalk* (η) are used to find context motifs, and a permutation test is applied to check their statistical significance. In our study, $\delta = 0.1$, $\eta = 0.3$ and the significance $P < 0.05$ (Benjamini and Hochberg corrected) were used. With the graphical representation of context-motifs, the Markov cluster (MCL) algorithm [83] is used to cluster closely related context motifs into biological contexts. The $inflation$ parameter was set to 2 in our study, as suggested by the developer of MCL. A representative set of samples from each identified context is determined as a contextual condition using the sample association score (SAS, Eq. 4.1), and we used SAS < 0.5 in our study. For each contextual condition, consistency statistics δ and η were used again to find contextual gene sets that include genes with consistent over-expression or under-expression. The same $\delta = 0.1$, $\eta = 0.3$ and $P < 0.05$ were used in this step.

STEP II: Summarizing Contextual Gene Sets

To infer networks of contextual gene sets, each contextual gene set needs to be represented as a single variable as most network models assume each node as a single random variable. For each sample s_k, the expression values of m genes in a contextual gene set $\mathbf{G}_i$ is summarized to a single representative value G_{ik}, where G_{ik} is UP if more than r % of the

genes in $\mathbf{G}_i$ are over-expressed with statistical significance of a hypergeometric P lower than a given threshold (vice versa for the case of DOWN). Otherwise, G_{ik} is given a value of NOCHANGE. In this study, $r = 50\%$ and a P threshold 0.05 were used.

STEP III: Learning Contextual Gene Set Interaction Networks

A contextual gene set interaction network is built from $\mathbf{S}_U$, which is the set of all samples, after the expression summarization of the original gene expression matrix $\mathbf{D}$ by computing the likelihood of dependency $d_{ij} = \Pr(\mathbf{G}_i \leftrightarrow \mathbf{G}_j | \mathbf{S}_U)$ $(= d_{ji})$ between each pair of contextual gene sets $\mathbf{G}_i$ and $\mathbf{G}_j$. $\mathbf{G}_i \leftrightarrow \mathbf{G}_j$ is a connection between two contextual gene sets $\mathbf{G}_i$ and $\mathbf{G}_j$ in any direction. We used the Bayesian network model to identify the dependency between contextual gene sets, and the BANJO software[2] was used to learn Bayesian networks. From R independent runs of BANJO with $\mathbf{S}_U$, d_{ij} was computed as a frequency of the undirected connection $\mathbf{G}_i \leftrightarrow \mathbf{G}_j$ as follows:

$$d_{ij} = \frac{\sum_{k=1}^{R} F(BN_k, \mathbf{G}_i \leftrightarrow \mathbf{G}_j)}{R}, \tag{5.1}$$

where BN_k is a Bayesian network structure from kth run of BANJO, and F is a function that returns 1 if BN_k has $\mathbf{G}_i \leftrightarrow \mathbf{G}_j$, or 0 otherwise. The direction of connections in the Bayesian networks was ignored as we consider either direction of a connection to represent the same existence of dependency between two contextual gene sets. If d_{ij} is larger than a given threshold d_θ, we declared that a dependency exists between $\mathbf{G}_i$ and $\mathbf{G}_j$. In our study, we used $R = 1,024$ and $d_\theta = 0.5$.

STEP IV: Identifying Context-Specific Network Components

Our approach to identify the specificity of a dependency relationship to a condition is measuring the effect by the samples of a condition on the likelihood of the dependency. When $\mathbf{S}_{T_k}$ is a set of samples of a condition T_k, the amount of effect γ by $\mathbf{S}_{T_k}$ on a dependency

[2] http://www.cs.duke.edu/ amink/software/banjo/

relationship $\mathbf{G}_i \leftrightarrow \mathbf{G}_j$ is defined by the relative ratio of the dependency likelihood with $\mathbf{S}_U$ to the likelihood with $\mathbf{S}_U - \mathbf{S}_{T_k}$. More specifically, γ can be defined as follows:

$$\gamma(\mathbf{G}_i \leftrightarrow \mathbf{G}_j; \mathbf{S}_U, \mathbf{S}_{T_k}) = \frac{d_{ij}}{d_{ij}^{T_k}}, \tag{5.2}$$

where $d_{ij}^{T_k} = \Pr(\mathbf{G}_i \leftrightarrow \mathbf{G}_j | \mathbf{S}_U - \mathbf{S}_{T_k})$. One characteristic of γ is that $\gamma(\mathbf{G}_i \leftrightarrow \mathbf{G}_j; \mathbf{S}_U, \mathbf{S}_{T_k})$ is proportional to $\Pr(\mathbf{G}_i \leftrightarrow \mathbf{G}_j | \mathbf{S}_{T_k})$ (Proof given in Appendix E), which is a conventional measure for the relevance of $\mathbf{G}_i \leftrightarrow \mathbf{G}_j$ to the type T_k.

The benefit of this characteristic is that γ can be used instead of the conventional measure $\Pr(\mathbf{G}_i \leftrightarrow \mathbf{G}_j | \mathbf{S}_{T_k})$, especially when the direct measurement of $\Pr(\mathbf{G}_i \leftrightarrow \mathbf{G}_j | \mathbf{S}_{T_k})$ can be unreliable due to the limited number of samples for each condition, which is the case of many biological applications.

To measure the statistical significance of $\gamma(\mathbf{G}_i \leftrightarrow \mathbf{G}_j; \mathbf{S}_U, \mathbf{S}_{T_k})$, a permutation test is done by using $\mathbf{S}_{T_k}^r$ instead of $\mathbf{S}_{T_k}$, which is built by randomly selecting $|\mathbf{S}_{T_k}|$ samples from $\mathbf{S}_U$. If H out of M permutations gave γ greater than or equal to $\gamma(\mathbf{G}_i \leftrightarrow \mathbf{G}_j; \mathbf{S}_U, \mathbf{S}_{T_k})$, the statistical significance P of $\gamma(\mathbf{G}_i \leftrightarrow \mathbf{G}_j; \mathbf{S}_U, \mathbf{S}_{T_k})$ is H/M. When $\gamma(\mathbf{G}_i \leftrightarrow \mathbf{G}_j; \mathbf{S}_U, \mathbf{S}_{T_k})$ is larger than a given threshold γ_θ and its P is lower than a threshold, $\mathbf{G}_i \leftrightarrow \mathbf{G}_j$ is declared to be specific to the condition T_k. In our study, we limited M to 100 for each significance test of γ due to the computational cost of repeating Bayesian network learning many times for each permutation. $\gamma_\theta = 2$, which indicates two-fold or higher increase of dependency relationship by adding the sample of the condition T_k, and $P = 0.05$ were used as threshold values.

Annotation of Gene Sets

GATHER [84] was used to identify GO terms that are associated to each contextual gene set with statistical significance, where the terms from molecular function and biological

process categories were considered. $P = 0.01$ was used as a significance threshold. For annotating specific regions in a contextual gene set interaction network, we computed the overlap of genes (all, over-expressed, or under-expressed) in a region with the pathway gene sets in MSigDB [85], and evaluated its statistical significance with a hypergeometric P. After the false discovery rate (FDR) correction of P values using Benjamini and Hochberg's method, FDR-corrected $P = 0.01$ was used as a significance threshold.

5.5 Results

We first provide an example of identifying condition specificity of interactions in biological networks by way of application to a Boolean model of a cholesterol regulatory pathway. Next we describe the advantages of contextual gene sets over the comparable biclustering approach. Following, two cancer data sets were used as applications to show the benefit of identifying condition/subtype-specificity, 1) a refractory cancer gene expression data with 113 cancer patient samples of 32 different tissue types [77] and 2) The Cancer Genome Atlas (TCGA) glioblastoma multiforme (GBM) gene expression data. Each resultant contextual gene set interaction network shows both cancer-generic and subtype-specific interactions. We found that the identified cancer-generic and subtype-specific sub-networks have different functional roles. Besides the comparison of functional annotations, we also related the identified subtype-specific interactions to supporting evidence from other knowledge sources. These results show that our approach to identify condition specificity in learned networks can provide novel information about biological functions specific to the given conditions.

Example of Identifying Condition Specificity

One key component of our approach is identifying the condition specificity of interactions in biological networks. To show the applicability of our method of identifying condition specificity, we conducted a simulation experiment as an example. We used a Boolean network model of the cholesterol regulatory pathway [86] to generate synthetic data sets of

two different conditions. The cholesterol regulatory pathway describes the synthesis of cholesterol from acetyl CoA. This process can be prohibited by drugs such as statins, and the Boolean network model also includes the statins and its regulatory path. From the Boolean network model of the cholesterol regulatory pathway, two synthetic data sets were generated with/without statins perturbation, where statins-perturbed data sets were generated by setting the state of statins to the value 1, and the statins-free data were generated by setting the state of statins to the value 0. For each case of with/without statins perturbation, 100 samples were generated. The specificity to the statins perturbation and its statistical significance were evaluated for each regulatory relationship from the cholesterol pathway model, by using the method for identifying condition-specific interactions described in the Methodology section. The same parameter values were used as described in the Methodology section except for the number of permutations, as 1,000 permutations were used in this simulation example. Fig. E.1 in Appendix E illustrates a network diagram of the cholesterol regulatory pathway, and its statins perturbation-specific regulations were highlighted with red color. From the figure, the inhibitive regulation from statins to HMG-CoA reductase and its downstream regulations from HMG-CoA reductase to Mevalonic acid were identified as statins-perturbation-specific regulations with statistical significance $P < 0.05$. All other regulatory relationships did not pass the significance threshold $P = 0.05$ (regulations represented with black connections), thus were not declared to be statins perturbation-specific. This simulation example shows that the proposed method successfully identifies the regulatory path from statins to its downstream elements, which is affected by the condition of statins existence.

Advantage of Contextual Gene Sets over Biclusters

We showed that the identification of condition-specificity in biological networks can be possible by the proposed method. However, applying that method to gene-level networks is computationally challenging due to their sizes, thus considering a set of genes as a func-

tional module in the target network is a viable approach. We used the context mining method to find contextual gene sets as such functional modules. Compared to the approach of context mining, conventional methods to cluster genes cannot identify gene sets that show coherent expression across a subset of samples, and thus final networks of gene sets will have only generic interactions across all samples, without any condition/subtype-specific interaction. Biclustering can be an alternative approach to find gene sets that show coherent expression under certain subset of samples, because it searches combinations of a subset of genes and a subset of samples, where the genes show similar patterns in the corresponding subset of samples. However, most biclustering methods lack the ability to group similar biclusters and often give significantly overlapping results. We compared our context mining based method and the Iterative Signature Biclustering Algorithm (ISA) [87] to see the overlap between their identified gene sets. From the refractory cancer data detailed in the coming section, 339 contextual gene sets were identified by our context mining based method, and the same number of biclusters were identified using ISA. The context mining based method gave significantly lower overlap between gene sets than ISA (Fig. 5.2), thus using gene sets from the context mining based method will have less chance of inferring undesirable generic interactions between similar gene sets. This was confirmed by learning a gene set interaction network and identifying tissue-type specificity from the refractory cancer data, with gene sets from both methods. From the gene set interaction network with the ISA gene sets, no interaction was determined to be tissue-specific (result is not shown), while 87 tissue-specific interactions were identified with the contextual gene sets, which will be described in the later sections.

Contextual Gene Set Interaction Network from the Refractory Cancer Gene Expression Data

For the case of analyzing the Target Now refractory cancer data [77], gene expression data of 21,073 probes (from Agilent-011521 Human 1A Microarray G4110A) and 113

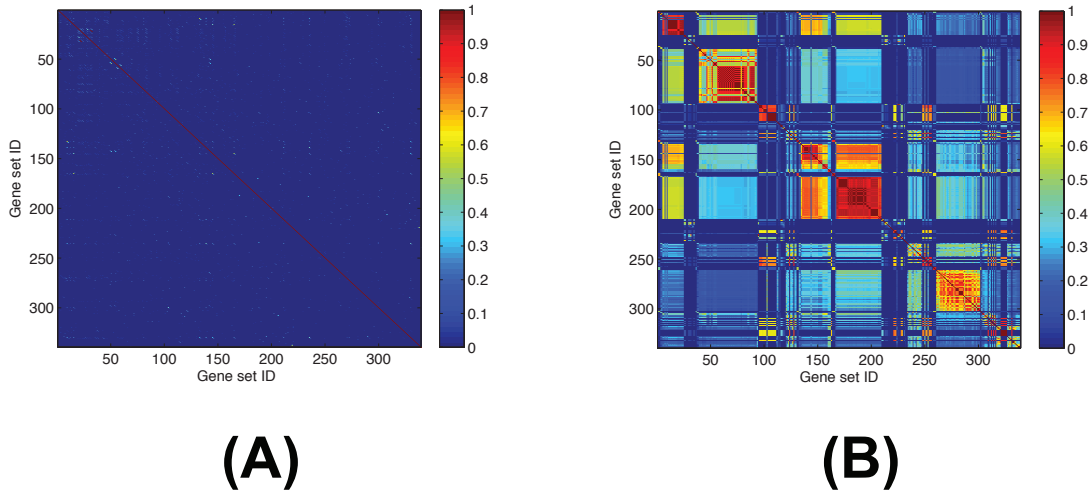


Figure 5.2: Jaccard similarity heat maps of 339 gene sets from two different methods: (A) Context Mining Based Method, (B) ISA.

patient samples (32 different types of refractory cancer) were used in this study. For each type, its normal tissue of origin was used as a baseline and the ratio of the cancer to its tissue of origin was computed. The ratio value was then ternarized using the 2-fold change threshold, to three discrete values of UP, DOWN and NOCHANGE. The distribution of the 113 samples among different cancer tumor types is listed in Table 4.1.

Contextual Gene Set Interaction Network of Refractory Cancer with Tissue Type Specificity

We identified contextual gene sets and a network of the gene sets from the gene expression data of refractory cancer. From the gene expression data, 339 contextual gene sets were identified as functional modules. The gene expression data was summarized to an assortment of 339 contextual gene sets as shown in Fig. E.2 in Appendix E. By learning Bayesian networks from the summarized expression data, a dependency likelihood d_{ij} $(= d_{ji})$ was evaluated for each pair of contextual gene sets $\mathbf{G}_i$ and $\mathbf{G}_j$, where $0 \leq d_{ij} \leq 1$. A contextual gene set interaction network of 285 interactions and 278 contextual gene sets was constructed by connecting gene sets with dependency likelihood larger than 0.5, which

Table 5.1: Refractory Cancer Specific Tissue Type Interactions: The number of specific interactions for each tissue type from the refractory cancer data. Tissue types with no specific interaction were not shown.

Tissue type	Specific interactions
Stomach	11
Pancreas	10
Melanoma, Adrenal	9
Ovary	8
Gall Bladder	7
Kidney	6
Breast, Brain, Testicular	4
Adipose tissue	3
Esophagus, Salivary Gland, Skin, Chondrosarcoma, Smooth muscle from uterus	2
Colon, T cell lymphoma, Glioma	1

implies that they are more "likely" to exist. For each of 285 interactions in the contextual gene set interaction network, its specificity to each of 32 tissue types was evaluated and 88 interactions (31%) were identified to be tissue specific. The number of specific interactions for each tissue type is summarized in Table 5.1. Fig. 5.3 shows the contextual gene set interaction network and tissue type specificity, with contextual gene sets as nodes and interactions as edges. Edges have different styles and colors based on their tissue type specificity. In this network, tissue-specific interactions are associated with 19 tissue types and the other 13 tissue types did not have significant effect on the interactions. We also highlighted sub-networks enriched with interactions of certain tissue types by dotted ellipses. This result represents that the samples of different tissue types have different effect on the dependency between contextual gene sets.

Discrepancy in Biological Functions Between Cancer-Generic and Tissue-Centric Contextual Gene Sets

Each contextual gene set was annotated with Gene Ontology (GO) [88] terms of biological functions and pathways to further elucidate the meaning of the network. GATHER [84] was

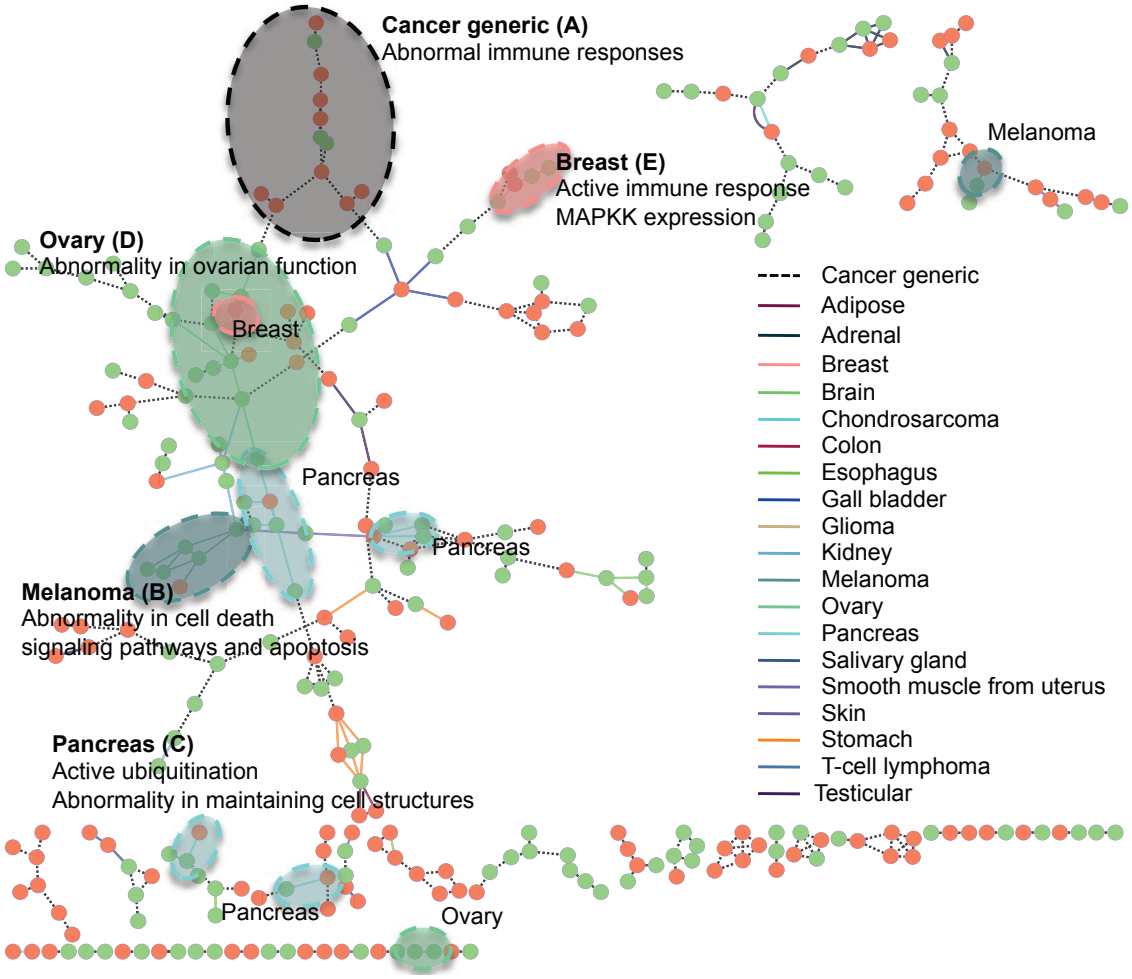


Figure 5.3: Annotated Contextual Refractory Cancer Network: The refractory cancer contextual gene set interaction network annotated with the identified tissue type specificity is shown. Only 278 contextual gene sets with at least one interaction are shown. The contextual gene sets under-expressed across their corresponding contextual conditions are colored with green and over-expressed contextual gene sets are colored with red.

used to find associated annotations with statistical significance, with $P = 0.01$ as a threshold of significance. To validate the discrepancy in functional meaning between cancer-generic interactions and tissue-specific interactions, the associated GO terms were compared between contextual gene sets with different types of interactions. From the contextual gene set interaction network in Fig. 5.3, contextual gene sets with only interactions specific to a tissue type T_k were declared as T_k-centric contextual gene sets. Similarly, contextual gene sets with only cancer-generic interactions were declared as cancer-generic contextual gene sets. Representing the set of GO terms associated with T_k-centric contextual gene sets as $\mathbf{Annot}(T_k)$, we compared the union of $\mathbf{Annot}(T_k)$ from all tissue types to the GO terms associated with cancer-generic contextual gene sets, $\mathbf{Annot}(\text{Cancer})$. The number of all tissue-centric contextual gene sets were 51 with $|\cup_{T_k} \mathbf{Annot}(T_k)| = 169$, while 175 contextual gene sets were cancer generic with $|\mathbf{Annot}(\text{Cancer})| = 413$ (the rest 52 gene sets with mixed tissue type interactions were not considered in this process). $\cup_{T_k}\mathbf{Annot}(T_k)$ and $\mathbf{Annot}(\text{Cancer})$ have 93 common GO terms. Table 5.2 lists the most frequent GO terms associated with either only cancer-generic contextual gene sets or only tissue-centric contextual gene sets. Many GO terms associated with cancer-generic contextual gene sets are basic biological mechanisms and well-known biological functions related to general cancer, such as signal transduction, cell cycle and RNA processing. For tissue-specific functions, a few tissue-specific network regions will be discussed in detail in the later subsection with their functional annotations.

Cancer-Generic Network Region

The region (A) in Fig. 5.3 is one example of a cancer-generic region. As this region includes many interactions that are not specific to certain tissue types, the expressions of the contextual gene sets within this region shows correlation across all tissue type samples. This region is mainly related to immune systems, as the 10 most statistically significant an-

Table 5.2: Significant GO Terms for Refractory Cancer Data Set: Significantly associated GO terms to only cancer-generic or tissue-centric contextual gene sets. The frequency indicates the number of contextual gene sets associated with the corresponding GO term. Only 20 most frequent terms are shown for each case.

GO terms relevant only to cancer-generic contextual gene sets		GO terms relevant only to tissue-centric contextual gene sets	
GO term	Frequency	GO term	Frequency
GO:0007166: cell surface receptor signal transduction	11	GO:0006163: purine nucleotide metabolism	2
GO:0007049: cell cycle	9	GO:0006164: purine nucleotide biosynthesis	2
GO:0006396: RNA processing	7	GO:0006629: lipid metabolism	2
GO:0008283: cell proliferation	7	GO:0009150: purine ribonucleotide metabolism	2
GO:0016070: RNA metabolism	7	GO:0009152: purine ribonucleotide biosynthesis	2
GO:0043207: response to external biotic stimulus	7	GO:0009259: ribonucleotide metabolism	2
GO:0000375: RNA splicing, via transesterification	6	GO:0009260: ribonucleotide biosynthesis	2
GO:0000377: RNA splicing, via transesterification	6	GO:0044255: cellular lipid metabolism	2
GO:0000398: nuclear mRNA splicing, via spliceosome	6	GO:0046148: pigment biosynthesis	2
GO:0006397: mRNA processing	6	GO:0000904: cellular morphogenesis	1
GO:0006959: humoral immune response	6	GO:0006099: tricarboxylic acid cycle	1
GO:0008380: RNA splicing	6	GO:0006119: oxidative phosphorylation	1
GO:0009613: response to pest, pathogen or parasite	6	GO:0006144: purine base metabolism	1
GO:0016071: mRNA metabolism	6	GO:0006188 : IMP biosynthesis	1
GO:0030333: antigen processing	6	GO:0006189: 'de novo' IMP biosynthesis	1
GO:0000067: DNA replication and chromosome cycle	5	GO:0006510: ATP-dependent proteolysis	1
GO:0000075: cell cycle checkpoint	5	GO:0006554: lysine catabolism	1
GO:0006950: response to stress	5	GO:0006570: tyrosine metabolism	1
GO:0016064: humoral defense mechanism	5	GO:0006582: melanin metabolism	1
GO:0000279: M phase	4	GO:0006583: melanin biosynthesis from tyrosine	1

notations are listed in Table 5.3. This can imply abnormal activity in immune mechanisms for the patients corresponding to the contextual conditions.

Tissue-Specific Regions

The regions (B) – (E) in Fig. 5.3 are the examples of network regions where many tissue-specific interactions exist. For each region, the expression patterns of contextual gene sets show strong correlation for the corresponding tissue type samples (see Supplementary Fig. E.3 (B-E)).

The melanoma-specific region (B) in Fig. 5.3 showed association with apoptosis. The five under-expressed contextual gene sets were related to abnormality in pigmentation, cell death signaling pathways and apoptosis. Individual contextual gene sets show further details in tissue-specific functional abnormalities. For example, $\mathbf{G}_{122}$ and $\mathbf{G}_{223}$ in the region

Table 5.3: Significant Annotations for Contextual Gene Sets: Top 10 most significant annotations for 12 contextual gene sets of region (A), Fig. 5.3.

MSigDB annotation (Source)	*P*
Graft versus host disease (KEGG)	5.39E-08
Type I diabetes mellitus (KEGG)	9.56E-08
Natural killer cell mediated cytotoxicity (KEGG)	1.08E-07
Generation of second messenger molecules (REACTOME)	1.24E-07
Allograft rejection (KEGG)	1.69E-07
Viral myocarditis (KEGG)	9.33E-07
Translocation of ZAP70 to immunological synapse (REACTOME)	2.58E-06
Signaling in immune system (REACTOME)	3.97E-06
Leishmania infection (KEGG)	4.64E-06
Autoimmune thyroid disease (KEGG)	6.20E-06

(B) are related to the metabolism of nicotinamide and nicotinamide adenine dinucleotide (NAD^+) metabolism. Through these metabolisms, the coenzyme compound NAD^+ accepts or donates electrons in redox reactions [89] that play significant roles in releasing energy from nutrients by generating ATPs. This possible abnormal activity of energy generation can be related to the fact that melanoma (and also other cancers) is intensively positive in positron emission tomography (PET) scans due to their intense demand for energy, where tumors have up-regulated receptors that take in glucose and subsequently have high levels of glycolysis. One over-expressed contextual gene set in this region was related to GTP binding ($P = 1.71E-5$) and guanyl nucleotide binding ($P = 1.37E-5$), and this is supported by a report of the expression of small GTP-binding protein genes of the RAS family in melanoma [90].

For the pancreas-specific region (C) in Fig. 5.3, the over-expressed $\mathbf{G}_{198}$ was associated with post-translational modification, such as ubiquitination. Ring finger proteins *RNF11* (7 out of 16 pancreas samples, $P = 0.001$) and *RNF139* (9 out of 16 pancreas samples, $P = 0.0085$) are over-expressed for several pancreas samples in $\mathbf{G}_{198}$ (Fig. 5.4), and there is a report that the ubiquitin-editing enzyme *A20* downregulates *NF-κB* signal-

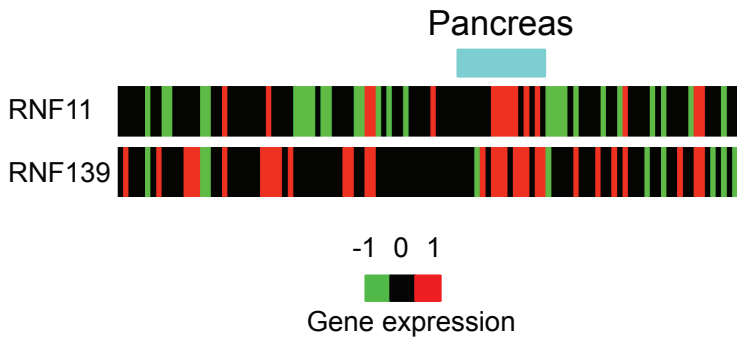


Figure 5.4: View of *RNF11* and *RNF139* Expressions Across Refractory Cancer Samples: Gene expression data were transformed to $\log_2$ ratios compared to expressions from normal tissue samples.

ing in the presence of *RNF11* [91]. While the over-expressed $\mathbf{G}_{198}$ was related to post-translational modification with ubiquitination, the genes in the two under-expressed contextual gene sets ($\mathbf{G}_{199}$ and $\mathbf{G}_{210}$) were related to maintaining cell structures, which can facilitate cell motility and invasion.

The ovary-specific region (D) in Fig. 5.3 includes six contextual gene sets under-expressed in most of the ovarian samples, where they were associated with ovary-specific functional annotations such as reproduction and pregnancy. This can be related to the loss of normal ovarian function from the ovarian cancer patients. Besides the ovary-specific annotations, they are also related to the β-Arrestin pathway and the caspase mediated cleavage of cytoskeletal proteins. Arrestins can block G protein-mediated signaling, and redirect signaling to alternative G protein-independent pathways. Regarding this annotation, there is a report that caspase mediated cleavage of cytoskeletal actin plays a positive role in the morphological changes of apoptosis [92].

Contextual Gene Set Interaction Network from the GBM Data of TCGA

From the TCGA-GBM gene expression data, 202 samples with four known subtypes (54 Classical, 58 Mesenchymal, 33 Neural and 57 Proneural [93]) were used in this study.

The expression of 17,814 genes in the GBM samples were converted to z-scores using 10 normal samples as a reference. The standardized expression values were quantized to three levels of UP, DOWN and NOCHANGE by using one standard deviation as a threshold.

Contextual Gene Set Interaction Network with Phenotype/Genotype Specificity

We also identified a contextual gene set interaction network from the gene expression data of TCGA GBM. From 202 GBM patient samples with four subtypes (Classical, Mesenchymal, Neural and Proneural) reported by Verhaak et al. [93], 316 contextual gene sets were identified. The gene expression data was summarized for the 316 contextual gene sets as shown in Fig. E.4 in Appendix E. Based on this summarized contextual gene set expression data, a contextual gene set interaction network (with 296 interactions and 247 contextual gene sets with at least one interaction) was built with the same methods and parameters applied to the case of the refractory cancer data analysis. For each interaction, its specificity to each of four subtypes was evaluated and 77 interactions (26%) were declared to be subtype specific. In addition to the four subtypes of GBM, specificities to the mutations of selected genes (*EGFR*, *NF1*, *PDGFRA*, *PIK3CA*, *PIK3R1* and *TP53*), the methylation of *MGMT*, and age$<$ 40 were also evaluated. The number of specific interactions for each subtype and condition is given in Table 5.4. Fig. 5.4 shows the contextual gene set interaction network with identified phenotype/genotype specificity. Regions enriched with interactions of certain subtypes were highlighted with dotted ellipses. Among the tested genetic mutations, only *EGFR* mutation showed associated interactions.

Functional Difference Between GBM-Generic and Subtype-Centric Contextual Gene Sets

Each contextual gene set was annotated with GO terms of biological functions and pathways, and significantly associated GO terms were compared for the contextual gene sets that have different types of interactions. From the GBM contextual gene set interaction network in Fig. 5.4, the annotations of subtype T_k-centric contextual gene sets were compared

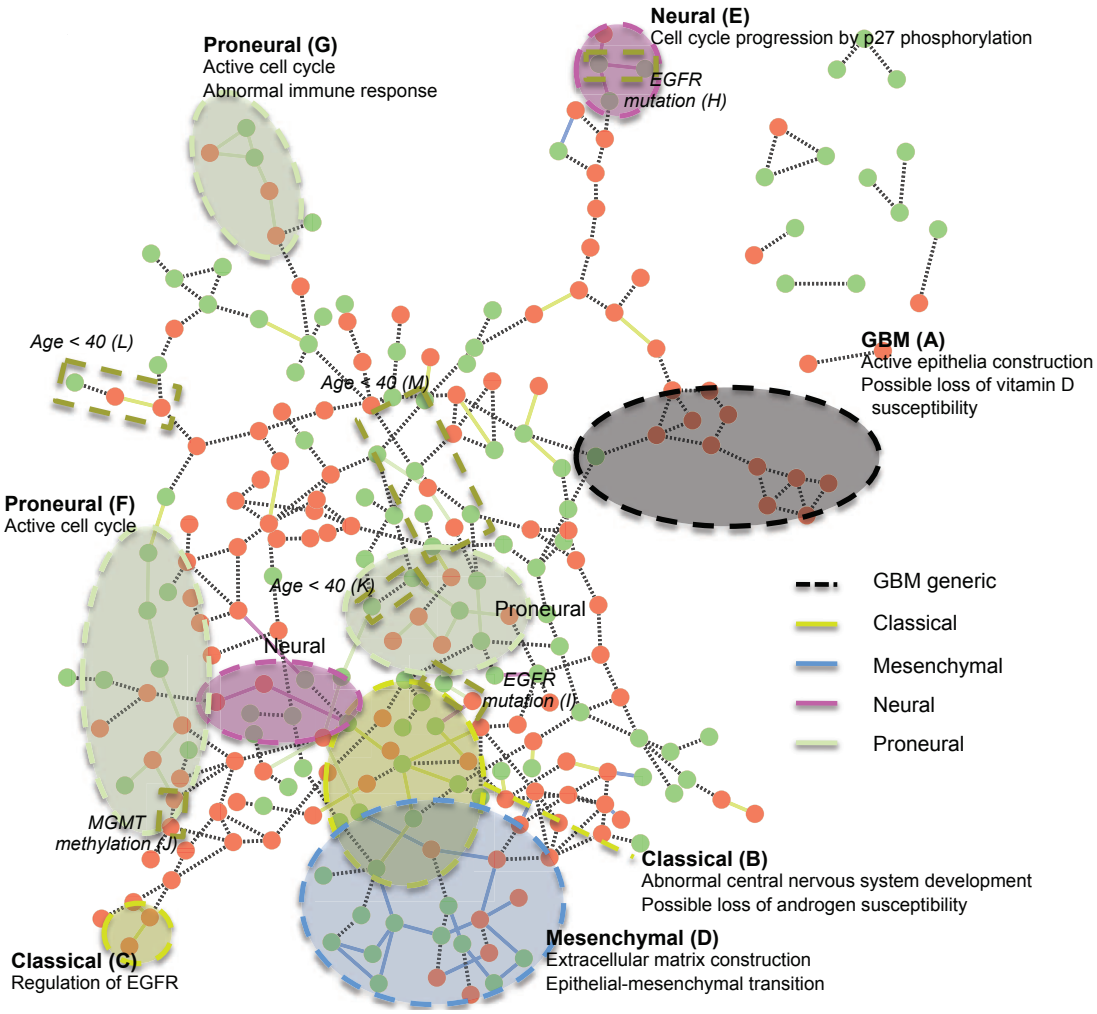


Figure 5.5: Annotated Contextual GBM Network: The GBM contextual gene set interaction network is annotated with the identified phenotype/genotype specificity. Only 247 contextual gene sets with at least one interaction are shown. The contextual gene sets under-expressed across their corresponding contextual conditions are colored with green and over-expressed contextual gene sets are colored with red.

Table 5.4: Specific GBM Subtype and Sample Condition Interactions: The number of specific interactions for each GBM subtype and sample condition. Among the investigated conditions, conditions with no specific interaction are not shown.

Subtype/Condition	Number of specific interactions
Classical	24
Mesenchymal	20
Neural	8
Proneural	24
EGFR mutation	2
MGMT methylation	1
Age < 40	7

Table 5.5: GBM Subtype-Centric Contextual Gene Sets: The number of subtype-centric contextual gene sets and associated annotation terms are given.

T_k	T_k-centric gene sets	$\lvert\mathbf{Annot}(T_k)\rvert$	$\lvert\mathbf{Annot}(T_k) \cap \mathbf{Annot}(GBM)\rvert$ (Overlap %)
Classical	6	4	0 (0 %)
Mesenchymal	8	34	6 (17.6 %)
Neural	4	4	4 (100 %)
Proneural	14	35	31 (88.6 %)

to the annotations of GBM-generic contextual gene sets. Table 5.5 lists the number of each subtype-centric contextual gene sets with their associated annotation terms. From Table 5.5, all four (100 %) GO terms associated with Neural-centric contextual gene sets and 31 out of 35 (88.6 %) GO terms associated with Proneural-centric contextual gene sets were also associated with GBM-generic contextual gene sets. This can imply the closeness of the abnormalities in Neural and Proneural subtypes to GBM-generic abnormalities. Compared to the cases of Neural and Proneural subtypes, none or few of the GO terms from Classical (none out of four, 0 %) and Mesenchymal (6 out of 34, 17.6 %) subtypes were overlapping with the GO terms associated with GBM-generic contextual gene sets, which can imply these two subtypes are more differentiated forms of GBM than Neural and Proneural subtypes. These findings of Classical and Mesenchymal subtypes being more differentiated GBM are consistent with the results of a previous GBM study [93].

Comparison of Contextual Gene Sets with Other Gene Signatures of GBM Subtypes

We compared the genes in subtype-centric contextual gene sets with the GBM subtype signature genes reported by Verhaak et al. [93]. Table 5.6 lists the number of over/under-expressed genes for each subtype, from two methods. Subtype-centric contextual gene sets have completely different list of genes compared to the gene signatures reported by Verhaak et al., as there was no overlap across all subtypes. This significant difference is due to the different approaches of identifying genes from two methods. The GBM subtype signature genes by Verhaak et al. were identified in two steps – they first used a consensus clustering to group the patient samples into four stable clusters, and then, differentially over-expressed genes in each subtype were identified using the combination of significance analysis of microarrays (SAM), receiver operating characteristic (ROC) methods and ClaNC, a nearest centroid-based classifier [93]. In contrast, contextual gene sets were identified with more focus on biological processes rather than differentially expressed signatures of subtypes. Genes in each contextual gene set are grouped as they show statistical interactions amongst them based on their expression levels, and each subtype-centric contextual gene set has its interactions strongly associated with a specific subtype. Thus, subtype-specific contextual gene sets represent genes whose interactions are strongly associated with biological processes underlying certain specific subtypes, rather than gene signatures for certain subtypes. In addition, as some genes may function in multiple different biological processes, a gene may belong to multiple contextual gene sets, and this is another discrepancy between contextual gene sets and subtype gene signatures.

The identified contextual gene sets were also compared to the result of Gene Set Enrichment Analysis (GSEA) on MSigDB [85], which is a method to identify differentially expressed gene sets for each condition, based on prior knowledge. For GSEA, the standardized gene expression data was used without quantization. GSEA analysis was done

by comparing "samples of one subtype versus the rest of the samples" for each subtype, to identify subtype-specific gene sets. Out of 2,101 gene sets of canonical pathway gene sets and GO gene sets of biological process and molecular function from MSigDB, 2,067 gene sets (98.4 %) with up to 500 genes were tested using GSEA. In running GSEA for each gene set, 1,000 permutations were applied. From the result, P values were FDR-corrected using Benjamini and Hochberg's method, and FDR-corrected $P = 0.05$ was used for statistical significance. Table 5.7 lists the number of identified gene sets by using two methods, where GBM-generic or each subtype-centric gene sets are shown for contextual gene sets, and subtype-specific gene sets are shown for GSEA. Our method could identify GBM-generic contextual gene sets as well as multi-type contextual gene sets, which show interactions with other contextual gene sets across multiple subtypes of GBM. Such types of gene sets cannot be identified using conventional GSEA, which works in a supervised manner based on the given subtype information. Another finding is that GSEA returns very biased number of gene sets between subtypes, where the Mesenchymal subtype dominates the number of findings. A possible hypothesis of GSEA returning biased results for Mesenchymal is that Mesenchymal is a most differentiated form of GBM (physiologically or genotypically) [93] and many genes are differentially expressed in Mesenchymal compared to other subtypes. Compared to GSEA, our method gives less biased results to a specific subtype, and this is because subtype-centric contextual gene sets are identified based on subtype-specific interactions. As subtype-specific interactions are focused on consistent interactions instead of consistent expression levels, they are relatively free from the bias of consistent differential expression in each subtype.

GBM-Generic Network Region

The network region (A) in Fig. 5.4 is one example of a GBM-generic region with 12 contextual gene sets in it. The heat map (Fig. E.5 (A) in Appendix E) of the contextual gene sets show that their expressions are closely correlated across all patient samples, making

Table 5.6: Comparison of GBM Subtype-Centric Contextual Gene Sets and GBM Subtype Signature Genes: The table lists the number of over-expressed genes (UP) and under-expressed genes (DOWN)

Gene set	Classical		Mesenchymal		Neural		Proneural	
	UP	DOWN	UP	DOWN	UP	DOWN	UP	DOWN
Subtype-centric contextual gene set	309	8	27	537	362	215	612	796
Verhaak et al.	162	0	216	0	129	0	178	0
Overlap	0		0		0		0	

Table 5.7: Comparison of Contextual Gene Sets and MSigDB Gene Sets Identified with GSEA: The number of gene sets are listed. For the contextual gene sets, GBM-generic or each subtype-centric contextual gene sets are shown. *Multi-type* indicates gene sets with various subtype-specific interactions.

	GBM-generic	Classical	Mesenchymal	Neural	Proneural	Multi-type
Contextual gene set	144	6	8	4	14	71
GSEA	N/A	1	245	6	3	N/A

the interactions among these contextual gene sets to be GBM-generic. From the associated annotations of the over-expressed contextual gene sets in their corresponding contextual conditions, this network region was mainly related to the tight junction and the intercellular adhesion, which occur in epithelia and brain endothelia. Also by considering other annotations such as epithelial cell differentiation and morphogenesis of an epithelium, this network component can represent active epithelia construction in GBM, which can imply active blood vessel construction. For the MSigDB annotations of the under-expressed contextual gene set, the presence of a transcriptional start site motif was statistically significant, which matches annotations for vitamin D receptor *VDR*. Considering the function of vitamin D killing GBM cells [94], a possible hypothesis is that the loss of vitamin D susceptibility in GBM patients can be related to the low activities of genes targeted by *VDR*, while the main cause of such low activities remains for further studies.

GBM Subtype-Specific Regions

The regions (B) – (G) in Fig. 5.4 show the examples of subtype-specific regions for four subtypes of GBM. In each example, the expression patterns of contextual gene sets show strong correlation for the corresponding subtype samples (see Fig. 5.6 (A) and (B), and Fig. E.5).

The Classial-specific region (B) in Fig. 5.4 includes six contextual gene sets, where one is over-expressed and five of them are under-expressed across Classical samples. From the MSigDB analysis, one of the significant annotations related to the over-expressed contextual gene set is a transcriptional start site motif that matches annotation for a member of *ETS* oncogene family, *ELK1*, which implies genes targeted by this oncogene are over-expressed and the oncogene *ELK1* is enabling many of its downstream genes. The genes in the under-expressed contextual gene sets were related to central nervous system development ($P = 2.58E-4$). Another transcriptional start site motif was also related, which matches annotation for the androgen receptor *AR*. These can imply the abnormalities in neural system development and in the susceptibility of GBM to androgen. The Classical-specific region (C) reveals the presence of *N-Myc* over-expression, which directly regulates a number of genes associated with the classical phenotype gene signature including *EGFR*.

For the Mesenchymal-specific region (D) in Fig. 5.4, five over-expressed contextual gene sets for Mesenchymal samples are related to cell surface interactions (integrin cell surface interactions, $P = 7.3E-5$) and several signaling pathways related to cancer. We could see up-regulation of Collagen I, IV ECM components (*COL1A2* up-regulation in 33 out of 58 Mesenchymal samples, $P = 1.22E-15$; *COL4A5* up-regulation in 10 Mesenchymal samples, $P = 0.1069$) that signify increased ECM production. TGF beta receptor II is also up-regulated (46 out of 58 Mesenchymal samples, $P < 8.12E-21$), which is associated with epithelial-mesenchymal transition (EMT). Jak-STAT signaling

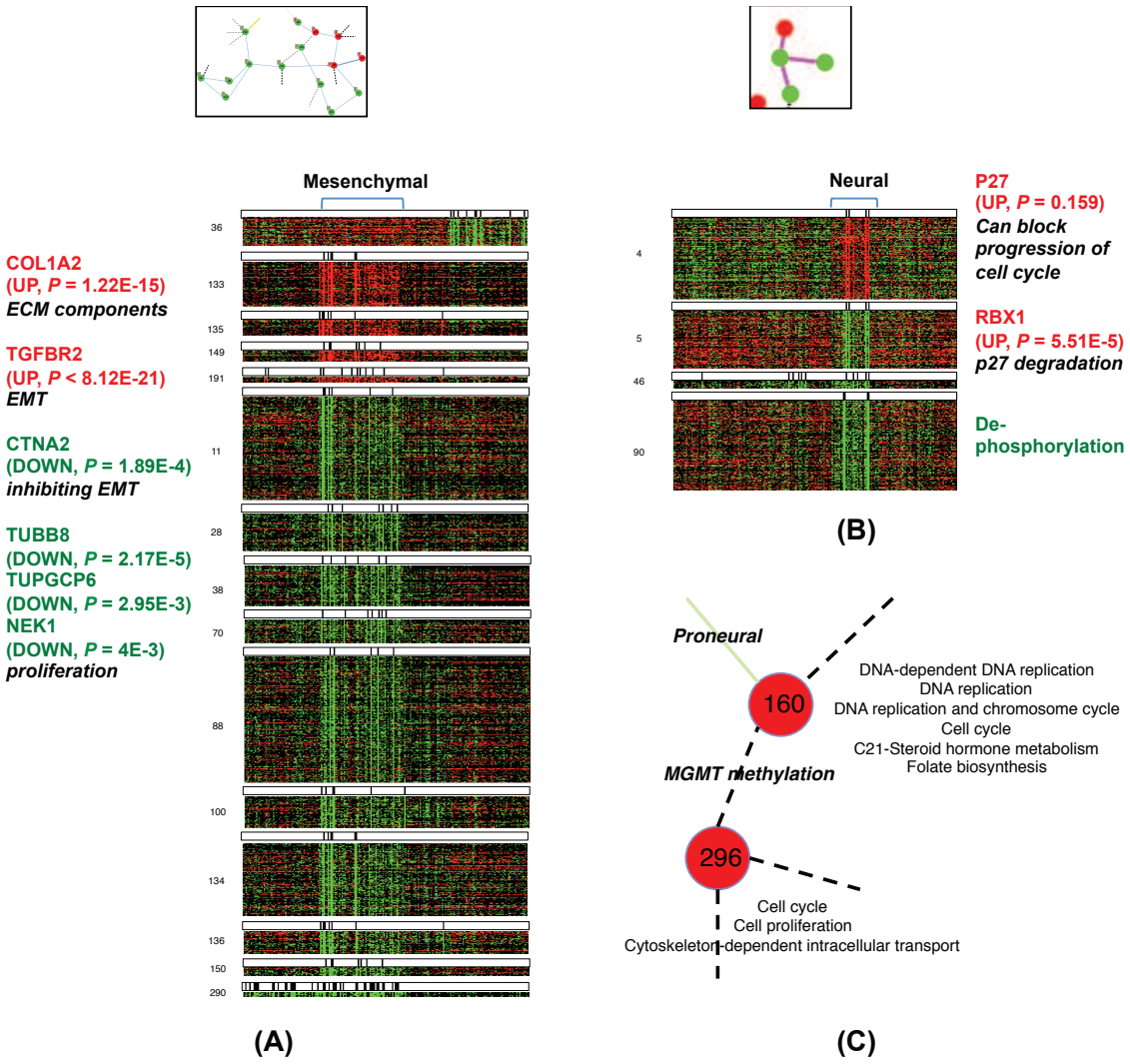


Figure 5.6: Selected Context-Specific Regions from GBM Network: A closer look on selected condition-specific regions from the GBM contextual gene set interaction network is shown here. (A) Heat maps of the Mesenchymal-specific region (D), with a brief summary of its relatedness to Mesenchymal features. (B) Heat maps of the Neural-specific region (E), where this region can represent a cell cycle progression by p27 phosphorylation. (C) *MGMT* methylation-specific region (J), where *MGMT* methylation is related to the repair of damaged DNA in the process of cell cycle.

pathway was related, too ($P = 1.84E-4$), with *PIK3CD* and *JAK2* up-regulation. Genes involved in integrin cell surface interactions (*COL1A2*, *ITGA11*, *RAP1B*, *COL4A5* and *APBB1IP*) were also included and these can lead to *MAPK* signaling. The other 10 contextual gene sets under-expressed in many Mesenchymal samples include down-regulated Catenins (*CTNNA2* down-regulation in 18 of 58 Mesenchymal samples, $P = 1.89E-4$), where Catenins are often associated with inhibiting EMT. These gene sets are also related to microtubule cytoskeleton organization and biogenesis ($P = 1E-5$). Besides, we could observe the down-regulation of tubulins (*TUBB8* down-regulation in 30 of 58 Mesenchymal samples, $P = 2.17E-5$; *TUBGCP6* down-regulation in 23 of 58 Mesenchymal samples, $P = 2.95E-3$) and *NEK1* (25 of 58 Mesenchymal samples, $P = 4E-3$) and *NEK11* (16 of 58 Mesenchymal samples, $P = 0.263$), which are components of proliferation, and it can imply that the proliferation activity is low in Mesenchymal cells since their high migratory behavior. A brief summary and a heat map of gene expressions of this region is shown in Fig. 5.6 (A), and the findings mentioned above can imply active EMT, abnormalities in maintaining cell structures, and low proliferative activity, which well fit the characteristics of the Mesenchymal subtype.

The Neural-specific region (E) in Fig. 5.4 can represent the cell cycle progression by p27 phosphorylation. The over-expressed $\mathbf{G}_4$ includes p27 (12 of 33 Neural samples, $P = 0.159$) and *RBX1* (16 of 33 Neural samples, $P = 5.51E-5$), where p27 can block progression of cell cycle. However, F box protein binds with phosphorylated p27 with the involvement of *RBX1*, causes p27 degradation and cell cycle progression. Accordingly, the three under-expressed contextual gene sets were related to dephosphorylation ($P = 2.45E-3$). A summary and a heat map of this region is also shown in Fig. 5.6 (B).

The Proneural-specific region (F) in Fig. 5.4 includes over-expressed contextual gene sets that are related to cell cycle, and under-expressed contextual gene sets related

to homeostatic processes. The Proneural-specific region (G) also shows active cell cycle processes, and degraded immune responses.

Other Genotype/Phenotype-Specific Regions

In addition to the four subtypes of GBM, other condition-specific interactions were also identified. Among genetic mutations, only *EGFR* mutation has associated interactions in regions (H) and (I). The condition of ages < 40 is associated with regions (K), (L) and (M). An interesting result is the region (J) associated with the methylation of *MGMT* (Fig. 5.6 (C)). In the region (J), the two over-expressed contextual gene sets, which are connected with a *MGMT* methylation-specific interaction, were related to cell cycle, especially DNA replication and checkpoint (DNA dependent DNA replication with $P = 7.84E-10$, mitotic cell cycle with $P = 3.47E-9$ and mitotic M-M/G1 phases with $P = 1.59E-7$). This specificity of *MGMT* methylation to the DNA replication and checkpoint in cell cycle is evident, by considering that *MGMT* is involved in the process of repairing damaged DNA during the replication process. The expression of *MGMT* could have been disturbed by methylation and eventually lost its function in the cell cycle checkpoint. The review by Casorelli et al. [95] also covers the role of the *MGMT* repair protein in cancer.

5.6 Chapter Summary

High heterogeneity in cancer has been evident since early studies, and approaches to reveal such heterogeneity embedded in genomic profiling data are showing promising results. In this work, we used a method to measure the effect from expression samples of certain conditions on the components in contextual gene set interaction networks for identifying condition-specificity. In addition to the simulation experiment, two cancer data sets were analyzed to support the validity of identifying condition-specificity, which are the refractory cancer data with 32 tissue types and TCGA GBM data with four subtypes. Contextual gene set interaction networks were built with tissue/phenotype/genotype-specificities. The resultant contextual gene set interaction networks with specificities showed different inter-

action patterns across conditions, and they provided new hypotheses as well as consistency with previous studies. Bayesian network learning was used in this work to more correctly estimate the likelihood of dependency, but simpler measures such as correlation or mutual information can be also used for the same formulation. Thorough analysis will follow this study for condition-specific interactions and related contextual gene sets to further analyze biological mechanisms of condition specificity in cancer.

Thus we have extended the context-specific gene regulatory network idea with a new graph-based representation of contextual genomic information and also contribute a novel, knowledge-based analysis approach to extract that contextual information.

CHAPTER 6

CONCLUSION

In this work we have visited a spectrum of biological modeling with a goal towards therapeutic intervention, with both formal and informal notions of biological context. In doing so we have extended the utility of the Boolean network model, formalized a graphical depiction of a model for contextual genomic regulation, and provided an analysis framework for said graphical depiction. In addition to providing a branching off point for various extensions, we have built upon the context-specific gene regulatory network towards the elucidating contextual gene set interaction network, along with a knowledge-based annotation and validation approach to reveal embedded contextual information.

In accomplishing these tasks we have answered our three problem statement questions.

Question 1 Can we identify targets for intervention to change or preserve biological steady states (i.e. contexts) within the classic framework of Boolean networks? We have demonstrated a novel approach to analyzing the Boolean network basins of attraction to discover targets for intervention, as well as a sampling-based extension to larger networks with inexhaustible state spaces.

Question 2 Can we extend the idea of a gene regulatory network to explicitly model context specificity? We have formalized the graphical depiction of a model for contextual genomic regulation, called a context-specific gene regulatory network, and extended this idea into the more powerful contextual gene set interaction networks. Both of these explicitly model the context-specific information embedded within biological data.

Question 3 Can we provide an analysis approach for context-modeling networks such that we can identify biologically meaningful content? For both context-specific gene

regulatory networks and contextual gene set interaction networks we have provided in-depth analysis approaches to reveal biologically meaningful content embedded within the contextual data.

In accomplishing these tasks we have also overcome the three stated challenges: 1) navigate the exponential state space of the Boolean network model, 2) create graphical depictions of models for contextual genomic regulation, and 3) design methods for extracting biologically meaningful contextual information.

6.1 Key Contributions

Here the key individual contributions of the dissertation are summarized.

Contribution 1

New method for identifying targets for intervention by analyzing Boolean network basins of attraction: Our work clearly established a usefulness in analyzing basins of attraction in identifying targets for intervention, and our use of logic minimization significantly reduced the representation of basins of attraction, and the proposed measures stratified the terms, revealing both the key players in the system and how to manipulate them.

Our smaller networks, both hand-made and inferred, have provided clearly identified intervention targets that are not always evident in attractor states themselves, in the network topology, or even from various existing measures, both graph-theoretic and Boolean-network-specific. In making application to the handmade network modeling some of the systems biology behind human aging, we were able to identify biologically meaningful intervention targets as well as determine for which desired system steady states (i.e. disease, health, etc.) they are most effective. The analysis of the yeast cell cycle network also demonstrated that our methodology can identify key variables in the system. We were able to systematically identify three important variables described specifically by the original

study and propose others for further study. Our application to the WNT5A network for melanoma demonstrated the applicability of our approach beyond hand-created networks to networks inferred from biological data; furthermore our targets identified for intervention had been previously validated by laboratory studies. Finally, the application of our small network approaches to AI planning showed how this approach can assist existing intervention planning strategies. Perhaps the most important aspect of our revealed intervention targets is the fact that they are both basin- and value-specific; in other words, we providing not just a target, but how exactly to intervene (value) and also a context in which the intervention is appropriate (basin).

Contribution 2

New method for identifying multi-variable targets for intervention in large Boolean networks requiring a sampled view of intractable state spaces: By introducing the template counting approach to supersede the small network popularity and power measures, we have made possible the identification of key players from the sampled state spaces of larger networks through our template analysis. We demonstrated the robustness of templates as intervention targets through a simulation study comprising thousands of trial interventions over various random networks. In these simulations we found that templates maintain performance in sampled state spaces and in increasingly large networks. We also demonstrated the application of the template-based approach on a much larger T-LGL Leukemia network crafted by domain experts. We note that when all single variable measures were unable to produce helpful intervention targets, the template based approaches did produce single and multi-variable intervention targets with high estimated success rates. We also observed that identifying the best template can benefit greatly from expert classification of the basins of attraction since the best template may not always be the one with the highest score. In the end, the template based analysis revealed the most powerful triggers for altering network dynamics into desired attractor basins, and these results were realized, and in certain cases

biologically reinforced, on a network with over 8 trillion states in the basin of attraction field based sampling only 50,000 initial states.

Contribution 3

New graphical depiction of a mathematical model for contextual genomic regulation: While cellular context mining is indeed successful at producing coherent context motifs, it was lacking in interpretability due to the reporting of identified context motifs in either a list, tree or tabular format. To expand, systematize, and to address shortcomings in visualization and interpretation, we developed the formalized graphical representation of the results of cellular context mining called context-specific gene regulatory networks, building off a mathematical framework which recognizes the idea of biological contexts, where we chain together of context motifs to form a graphical depiction of the mathematical model. This graphical depiction models not only genes in varied expression states, but also per-context-motif sample information, and thus provides fertile ground for context-based analysis. Formally, a context-specific gene regulatory network H is a pair $H = (V, E)$, where V is a set of gene-representing vertices and E is a set of edges oriented from genes designated as drivers to genes designated as driven; 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 in our representation is specific to only its corresponding subset of samples.

Contribution 4

New *ad hoc* analysis approach for the graphical depiction of contextual genomic regulation to identify biologically meaningful context clusters: Our work presents a step-by-step approach to generate context motif clusters, i.e. disease-pertinent gene regulatory networks, using our novel analysis methodology applied to the context-specific gene regulatory networks. To obtain the best context motif clusters, we contribute the following

ad hoc approaches: In general one first considers any topological clues: an analysis of the weakly connected or strongly connected (if present) components in the network topology is a natural place to begin segregating into subnetworks. Within these components, the regulatory control inherent in the driver-driven relationships is useful in identifying separation points between portions of weak or strong components. Additionally, expression values lend insight; because each variable in a context-specific gene regulatory network built from a quantized data set is present in each level of quantization (e.g. up to three nodes for a ternary gene variable found to be up- or down-regulated or no change), one can quickly identify portions of the network exhibiting similar expression behavior. In general, context motif cluster identification steps include the following separation criteria:

- Weakly Connected Components (topological)
- Strongly Connected Components (topological)
- Large Subgraphs Separated by Few Edges (topological)
- Subgraphs Differing in Expression Value Consensus (biological)
- Subgraphs Differing in Disease Subtype Enrichment (biological)

Our study shows how these gene regulatory networks exhibit biological advantages absent in related techniques. We demonstrate the utility of the approach on a refractory cancer data set, the mapping of inter- and intra-context edges for which results in a graph of context motif clusters which is found to have interesting properties, such as specific cancer tumor type enrichment (used for context motif cluster annotation) and occurrence of genes relevant to the annotated diseases. Comparison of this approach with conventional clustering algorithms demonstrated its advantage for relevant gene subset identification. When compared with Bayesian network analysis, we noted that context motif clusters can

capture regulatory relationships specific to subsets of samples while conventional Bayesian network learning rarely captures meaningful context-specific information.

In the end this work provided a branching-off point not only for our extensions in Contributions 5 and 6, but for other work by Sen[64] and Ramesh[77].

Contribution 5

New graphical depiction of a context-based model: a contextual gene set interaction network, specifically STEP III and STEP IV of the approach concerning Bayesian network learning and identifying condition specificity: The usefulness of the context-specific gene regulatory network despite its interpretation difficulties motivated out next step in context-specific approaches, in which we detailed the formulation and preliminary results of contextual gene set interaction networks. The gene set interaction networks reduce the dimensionality of our results by creating summarizing gene sets and providing a simpler network structure whose edges are identified with the context responsible for their presence. Like context-specific gene regulatory networks, the contextual gene set interaction networks are replete with sample information, but now have the newfound simplicity provided by gene sets. Combined, these two features enable the incorporation of existing biological knowledge to provide rich, contextual annotations.

Learning contextual gene set interaction networks and identifying condition specificity involves several steps of data transformation, as illustrated in Fig. 5.1. It consists of four major steps: (STEP I) identifying contextual gene sets as basic functional modules, (STEP II) summarizing contextual gene sets to transform the data of genes to the data of contextual gene sets, (STEP III) learning contextual gene set interaction networks where each interaction represents dependency in expression (specific expression status of a gene set depends on the expression status of the other gene set) between two contextual gene sets, and (STEP IV) identifying condition specificity of each interaction. For the first step

of identifying contextual gene sets, we define a contextual gene set as a set of genes that show consistent expression pattern under a biological context, i.e. a subset of samples. While the entire approach was a collaborative process, the specific personal contributions of the dissertation author were to STEP III, STEP IV and to the validation and annotation work thereafter.

(Step III) A contextual gene set interaction network is learned from the summarized contextual gene set expression data, by evaluating the likelihood of dependency between each pair of contextual gene sets given all samples and building a connection if the dependency likelihood is larger than a given threshold. Inference of interaction networks from the summarized data has a few advantages over traditional approach where all genes are used. Since the number of variables (nodes) is significantly smaller in this approach as all the genes in contextual gene set are aggregated to a single variable, the method suffers less in both computational complexity and statistical estimation accuracy.

(Step IV) A resultant interaction between two contextual gene sets represents that there is a probabilistic dependency in their summarized expressions. Gene sets with dependency are expressed in coordinated manners, where the expression status of a gene set depends on the expression status of the other gene set. However, the effect of the dependency from the samples can be different for diverse conditions, as they can imply different activities of biological functions. Based on this idea, we identify condition-specific regions in the built network by measuring the effect from the samples of each condition on the likelihood of dependency. To measure the effect of a condition on a dependency, we evaluated the likelihood of the dependency without the samples of the condition and computed its difference with the original likelihood obtained using all available samples. If the original likelihood is significantly higher than the likelihood without the samples from the condition, it means that the samples under the condition have made significant contribution to

the dependency. This implies that the dependency exists mainly due to the samples from the condition, thus it is declared as a condition-specific dependency.

Contribution 6

New knowledge-based annotation and validation approach for contextual gene set interaction networks to reveal contextual information embedded within: We also contributed a knowledge-based analysis approach on these new contextual gene set interaction networks incorporating Gene Ontology as well as pathway information. Using this approach we have achieved valuable analysis results on both the Target Now data set and with a glioblastoma data set from TCGA, and we have demonstrated advantages over the related biclustering method.

GATHER [84] was used to identify GO terms that are associated to each contextual gene set with statistical significance, where the terms from molecular function and biological process categories were considered. $P = 0.01$ was used as a significance threshold. For annotating specific regions in a contextual gene set interaction network, we computed the overlap of genes (all, over-expressed, or under-expressed) in a region with the pathway gene sets in MSigDB [85], and its statistical significance was evaluated with a hypergeometric *P*. After the false discovery rate (FDR) correction of *P* values using Benjamini and Hochberg's method, FDR-corrected $P = 0.01$ was used as a significance threshold.

6.2 Future Directions

Future work based off of this dissertation will proceed in two main directions.

First, the Boolean network progress of Chapter 2 must be extended for even larger networks and must be adapted for parallel computing solutions for efficiency. Despite progress in sampling large state spaces, we will always be limited by the exponential growth of the state space with the number of variables. Fortunately as network sizes race into intractability, so too does the reliability of such networks, which is a direct influence on the

quality of our results. Since there are quality handmade networks with sizes into the dozens of variables, such as our T-LGL Leukemia network, a Drosophila network from Albert *et al.*[23], and others, our leap to this size level is significant. With improvements to algorithm implementation and with the incorporation of parallelization, we plan to improve the large networks approach in terms of speed and network size capability, ideally towards the 100 variable mark. Additionally, tools for biologists must be created to make our advancements accessible to a wide audience.

Second, regarding the contextual gene set interaction networks, thorough analysis will follow our study for condition-specific interactions and for related contextual gene sets to further analyze biological mechanisms of condition specificity in cancer. As biologists' analysis needs are further understood in correlation with contextual information available for use, systematic analysis software will be designed to enable the use of our approaches by other biological researchers themselves.

6.3 List of Publications

The following are the publications, presentations and work relevant to this dissertation.

Publications[1]

1. Verdicchio, M. P. and Kim, S. (2013) "Template-Based Intervention in Boolean Network Models of Biological Systems." *BMC Bioinformatics, Submitted*. (Chapter 2, Chapter 3)

2. Jung S., Verdicchio M. P., Kiefer, J., Von Hoff, D., Berens, M., Bittner, M., and Kim, S. (2013) "Learning Contextual Gene Set Interaction Networks of Cancer with Condition Specificity." *BMC Genomics* 14(1), 110+, 2013. (Chapter 5)

[1] All journal and conference papers were peer reviewed. The TIST paper was invited and peer reviewed.

3. Jung S., Verdicchio M. P., Kiefer, J., Berens, M., and Kim, S. (2011) "Learning Contextual Gene Set Interaction Networks of Glioblastoma and Identifying Subtype Specificity." *In Eighth International Workshop on Computational Systems Biology (WCSB 2011)*, June 6-8, 2011, Zurich, Switzerland. (Chapter 5)

4. Verdicchio, M. P. and Kim, S. (2011) "Identifying Targets for Intervention by Analyzing Basins of Attraction." *Pacific Symposium on Biocomputing*, 16, Jan 3-7 2011, Hawaii. (Chapter 2, Chapter 3)

5. Bryce, D., Verdicchio M.P., Kim, S. (2010). "Planning Interventions in Biological Networks." *ACM Transactions on Intelligent Systems and Technology: Special Issue on Applications of Automated Planning,* Volume 1, Issue 2, Paper 11, pages 1-26, 2010 (Chapter 2, Chapter 3)

6. Verdicchio, M. P. and Kim, S. (2010). "Reduction of Boolean Network Basins of Attraction Reveals Intervention Targets. Technical report, Arizona State University, Tempe, Arizona. (Chapter 2, Chapter 3)

7. Sen, I., Verdicchio M. P., Jung, S., Trevino, R., Bittner, M., Kim, S. (2009). "Context-Specific Gene Regulations in Cancer Gene Expression Data" [2]." *Pacific Symposium on Biocomputing*, 14:75-86, Jan 5-9 2009, Hawaii. (Chapter 4)

Oral and Poster Presentations

1. Jung S., Verdicchio M. P., Kiefer, J., Berens, M., and Kim, S. (2011) "Learning Contextual Gene Set Interaction Networks of Glioblastoma and Identifying Subtype Specificity." *Eighth International Workshop on Computational Systems Biology (WCSB 2011)*, June 6-8, 2011, Zurich, Switzerland. Invited oral and poster presentation of accepted paper (S. Jung presented). (Chapter 5)

[2]Ina Sen and Michael P. Verdicchio contributed equally to this work as first authors.

2. Verdicchio, M. P., Kim, S. (2011). "Identifying Targets for Intervention by Analyzing Basins of Attraction." *Pacific Symposium on Biocomputing*, 16, Jan 7 2011, Hawaii. Invited oral and poster presentation of accepted paper.

3. Alterovitz, G., Verdicchio, M.P., Cavalcanti, S., Wang, M., Ramoni, M.[3] (2011). "Reverse Engineering and Synthesis of Biomolecular Systems." *Pacific Symposium on Biocomputing*, 16, Jan 3 2011, Hawaii. Tutorial presentation with G. Alterovitz.

4. Verdicchio, M., Kim, S. (2009). "Boolean Network Models of Human Aging." Invited oral and poster presentations at the *Rocky '09 Bioinformatics Conference*, December 10-12, Aspen, Colorado.

5. Bidaye, D., Dzifcak, J., Stracuzzi, D., Chimera, R., Verdicchio, M. P., Kim, S., Langley, P. "An Interactive Environment for Visualizing, Interpreting, and Revising Biological Process Models." Poster presentation at *RECOMB 2009 Conference*, May 17-21, Tucson, Arizona.

6. Verdicchio, M., Zhang, X., Baral, C., Kim, S. (2008). "Learning Causal Relationships Between Genes from Steady State Data: Algorithms, Simulation and Application." Invited oral and poster presentations at the *Rocky '08 Bioinformatics Conference*, December 4-7, Aspen, Colorado.

7. Zhang, X., Verdicchio, M., Baral, C., Kim, S. (2008). "Learning Causal Relationships Between Genes from Steady State Data: Algorithms, Simulation and Application." Poster presentation at the *2008 Translational Genomics Research Institute (TGen) Scientific Retreat*, Phoenix, Arizona.

8. Verdicchio, M., Kim, S. (2006). "Graphical Representation of Gene Regulatory Networks." Poster presentation at the *Fulton Undergraduate Research Initiative Un-*

[3]Deceased

dergraduate Research Symposium, Tempe, Arizona. (Earlier versions presented at similar symposia, November, April, 2005)

Works in Progress

- Verdicchio, M. P. “Parallel Processing Tools for Analyzing Basins of Attraction with Application to Large Networks.” (Expected Submission Fall 2013).
- Jung, S., Verdicchio, M. P., Bittner, M., Kim, S. “Review Paper: Context-Specific Gene Regulatory Networks.” (Expected Submission Fall 2013).

REFERENCES

[1] Kauffman S: **Metabolic stability and epigenesis in randomly constructed genetic nets**. *Journal of Theoretical Biology* 1969, **22**(3):437–467, [[http://dx.doi.org/10.1016/0022-5193(69)90015-0]].

[2] Wuensche A: **Genomic regulation modeled as a network with basins of attraction.** In *Pacific Symposium on Biocomputing* 1998:89–102, [[http://view.ncbi.nlm.nih.gov/pubmed/9697174]].

[3] Shmulevich I, Dougherty ER, Zhang W: **From Boolean to probabilistic Boolean networks as models of genetic regulatory networks**. *Proceedings of the IEEE* 2002, **90**(11):1778–1792, [[http://dx.doi.org/10.1109/JPROC.2002.804686]].

[4] Datta A, Choudhary A, Bittner ML, Dougherty ER: **External Control in Markovian Genetic Regulatory Networks**. *Machine Learning* 2003, **52**:169–191, [[http://dx.doi.org/10.1023/A:1023909812213]].

[5] Datta A, Choudhary A, Bittner ML, Dougherty ER: **External control in Markovian genetic regulatory networks: the imperfect information case**. *Bioinformatics* 2004, **20**(6):924–930, [[http://dx.doi.org/10.1093/bioinformatics/bth008]].

[6] Pal R, Datta A, Bittner ML, Dougherty ER: **Intervention in context-sensitive probabilistic Boolean networks**. *Bioinformatics* 2005, **21**(7):1211–1218, [[http://dx.doi.org/10.1093/bioinformatics/bti131]].

[7] Choudhary A, Datta A, Bittner ML, Dougherty ER: **Intervention in a family of Boolean networks**. *Bioinformatics* 2006, **22**(2):226–232, [[http://dx.doi.org/10.1093/bioinformatics/bti765]].

[8] Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data**. *Nat Genet* 2003, **34**(2):166–176, [[http://dx.doi.org/10.1038/ng1165]].

[9] Neapolitan RE: *Learning Bayesian Networks*. Prentice Hall, illustrated edition edition 2003, [[http://www.worldcat.org/isbn/0130125342]].

[10] Grzegorczyk M, Husmeier D, Edwards KD, Ghazal P, Millar AJ: **Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian**

network and the allocation sampler. *Bioinformatics* 2008, **24**(18):2071–2078, [[http://dx.doi.org/10.1093/bioinformatics/btn367]].

[11] Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: **Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes**. *Molecular Biology of the Cell* 2000, **11**(12):4241–4257, [[http://www.molbiolcell.org/cgi/content/abstract/11/12/4241]].

[12] Goh K, Cusick M, Valle D, Childs B, Vidal M, Barabási A: **The human disease network**. *Proc Natl Acad Sci USA* 2007, **104**(21):8685–8690.

[13] Gasch A, Eisen M: **Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering**. *Genome Biology* 2002, **3**(11):Research0059.1–0059.22.

[14] Segal E, Friedman N, Koller D, Regev A: **A module map showing conditional activity of expression modules in cancer.** *Nat Genet* 2004, **36**(10):1090–1098.

[15] Luscombe N, Madan Babu M, Yu H, Snyder M, Teichmann S, Gerstein M: **Genomic analysis of regulatory network dynamics reveals large topological changes**. *Nature* 2004, **431**(7006):308–312, [[http://dx.doi.org/10.1038/nature02782]].

[16] Shlomi T, Cabili M, Herrgard M, Palsson B, Ruppin E: **Network-based prediction of human tissue-specific metabolism**. *Nat Biotechnol* 2008, **26**(9):1003–1010, [[http://dx.doi.org/10.1038/nbt.1487]].

[17] Bossi A, Lehner B: **Tissue specificity and the human protein interaction network**. *Mol Syst Biol* 2009, **5**, [[http://dx.doi.org/10.1038/msb.2009.17]].

[18] Shiraishi T, Matsuyama S, Kitano H: **Large-Scale Analysis of Network Bistability for Human Cancers**. *PLoS Comput Biol* 2010, **6**(7):e1000851–, [[http://dx.doi.org/10.1371%2Fjournal.pcbi.1000851]].

[19] Kauffman SA: *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, first edition 1993.

[20] Ekins S, Wrighton SA: **Application of in silico approaches to predicting drug-drug interactions**. *Journal of Pharmacological and Toxicological Methods* 2001, **45**:65–69, [[http://dx.doi.org/10.1016/s1056-8719(01)00119-8]].

[21] Li F, Long T, Lu Y, Ouyang Q, Tang C: **The yeast cell-cycle network is robustly designed**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(14):4781–4786, [[http://dx.doi.org/10.1073/pnas.0305937101]].

[22] Verdicchio MP, Kim S: **Identifying targets for intervention by analyzing basins of attraction**. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 2011, :350–361, [[http://view.ncbi.nlm.nih.gov/pubmed/21121062]].

[23] Albert R, Othmer HG: **The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in Drosophila melanogaster**. *Journal of Theoretical Biology* 2003, **223**:1–18, [[http://dx.doi.org/10.1016/S0022-5193(03)00035-3]].

[24] Yu J, Smith VA, Wang PP, Hartemink AJ, Jarvis ED: **Advances to Bayesian network inference for generating causal networks from observational biological data**. *Bioinformatics* 2004, **20**(18):bth448–3603, [[http://dx.doi.org/10.1093/bioinformatics/bth448]].

[25] Smith VA, Jarvis ED, Hartemink AJ: **Evaluating functional network inference using simulations of complex biological systems**. *Bioinformatics* 2002, **18**(suppl_1):S216–224, [[http://dx.doi.org/10.1093/bioinformatics/18.suppl_1.S216]].

[26] Akutsu T, Miyano S, Kuhara S: **Identification of genetic networks from a small number of gene expression patterns under the Boolean network model.** *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 1999, :17–28, [[http://view.ncbi.nlm.nih.gov/pubmed/10380182]].

[27] Shmulevich I, Saarinen A, Yli-Harja O, Astola J: **Inference of Genetic Regulatory Networks Via Best-Fit Extensions**. In *Computational and Statistical Approaches to Genomics*. Edited by Zhang W, Shmulevich I, Boston: Kluwer Academic Publishers 2003:197–210, [[http://dx.doi.org/10.1007/0-306-47825-0_11]].

[28] Lähdesmäki H, Shmulevich I, Yli-Harja O: **On Learning Gene Regulatory Networks Under the Boolean Network Model**. *Machine Learning* 2003, **52**:147–167, [[http://dx.doi.org/10.1023/A:1023905711304]].

[29] Shmulevich I, Dougherty ER, Kim S, Zhang W: **Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks**. *Bioinformatics* 2002, **18**(2):261–274, [[http://dx.doi.org/10.1093/bioinformatics/18.2.261]].

[30] Richardson KA: **Simplifying Boolean Networks**. *Advances in Complex Systems* 2005, **8**(4):365–382.

[31] Dubrova E, Teslenko M, Tenhunen H: **A Computational Scheme Based on Random Boolean Networks**. In *Transactions on Computational Systems Biology X, Volume 5410 of* Lecture Notes in Computer Science. Edited by Priami C, Dressler F, Akan O, Ngom A, Springer Berlin / Heidelberg 2008:41–58, [[http://dx.doi.org/10.1007/978-3-540-92273-5_3]]. [10.1007/978-3-540-92273-5_3].

[32] Willadsen K, Wiles J: **Robustness and state-space structure of Boolean gene regulatory models**. *Journal of Theoretical Biology* 2007, **249**(4):749–765, [[http://dx.doi.org/10.1016/j.jtbi.2007.09.004]].

[33] Saez-Rodriguez J, Alexopoulos LG, Epperlein J, Samaga R, Lauffenburger DA, Klamt S, Sorger PK: **Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction**. *Molecular Systems Biology* 2009, **5**, [[http://dx.doi.org/10.1038/msb.2009.87]].

[34] Schlatter R, Schmich K, Avalos Vizcarra I, Scheurich P, Sauter T, Borner C, Ederer M, Merfort I, Sawodny O: **ON/OFF and Beyond - A Boolean Model of Apoptosis**. *PLoS Comput Biol* 2009, **5**(12):e1000595+, [[http://dx.doi.org/10.1371/journal.pcbi.1000595]].

[35] Steggles LJ, Banks R, Shaw O, Wipat A: **Qualitatively modelling and analysing genetic regulatory networks: a Petri net approach**. *Bioinformatics* 2007, **23**(3):336–343, [[http://dx.doi.org/10.1093/bioinformatics/btl596]].

[36] Maji P: **On Characterization of Attractor Basins of Fuzzy Multiple Attractor Cellular Automata**. *Fundam. Inf.* 2008, **86**(1,2):143–168, [[http://portal.acm.org/citation.cfm?id=1466468]].

[37] Mar JC, Quackenbush J: **Decomposition of Gene Expression State Space Trajectories**. *PLoS Comput Biol* 2009, **5**(12):e1000626+, [[http://dx.doi.org/10.1371/journal.pcbi.1000626]].

[38] Marcovitz A: *Introduction to Logic Design*. McGraw-Hill Science/Engineering/Math, first edition 2002, [[http://www.worldcat.org/isbn/0072951761]].

[39] Rudell RL, Sangiovanni-Vincentelli AL: **ESPRESSO-MV: Algorithms for Multiple Valued Logic Minimization**. In *Proc. of the IEEE Custom Integrated Circuits Conference* 1985.

[40] Scardoni G, Petterlini M, Laudanna C: **Analyzing biological network parameters with CentiScaPe**. *Bioinformatics* 2009, **25**(21):2857–2859, [[http://dx.doi.org/10.1093/bioinformatics/btp517]].

[41] Furber JD: **Systems biology of human aging: Network model of biochemical and physiological interactions in human senescence** 2010, [[http://www.legendarypharma.com/chartbg.html]].

[42] Shmulevich I, Dougherty ER, Zhang W: **Gene perturbation and intervention in probabilistic Boolean networks**. *Bioinformatics* 2002, **18**(10):1319–1331, [[http://dx.doi.org/10.1093/bioinformatics/18.10.1319]].

[43] Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Sampas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V: **Molecular classification of cutaneous malignant melanoma by gene expression profiling.** *Nature* 2000, **406**(6795):536–540, [[http://dx.doi.org/10.1038/35020115]].

[44] Weeraratna AT, Jiang Y, Hostetter G, Rosenblatt K, Duray P, Bittner M, Trent JM: **Wnt5a signaling directly affects cell motility and invasion of metastatic melanoma**. *Cancer Cell* 2002, **1**(3):279–288, [[http://dx.doi.org/10.1016/S1535-6108(02)00045-4]].

[45] Kim S, Li H, Dougherty ER, Cao N, Chen Y, Bittner M, Suh EB: **Can Markov Chain Models Mimic Biological Regulation?** *Journal of Biological Systems* 2002, **10**(4):337–357, [[http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.2.1381]].

[46] Dougherty ER, Kim S, Chen Y: **Coefficient of determination in nonlinear signal processing**. *Signal Processing* 2000, **80**:2219–2235, [[http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.2.4244]].

[47] Kim S, Dougherty ER, Bittner ML, Chen Y, Sivakumar K, Meltzer P, Trent JM: **General nonlinear framework for the analysis of gene interaction via multivariate expression arrays.** *Journal of biomedical optics* 2000, **5**(4):411–424, [[http://view.ncbi.nlm.nih.gov/pubmed/11092429]].

[48] Licciulli S, Luise C, Zanardi A, Giorgetti L, Viale G, Lanfrancone L, Carbone R, Alcalay M: **Pirin delocalization in melanoma progression identified by high content immuno-detection based approaches.** *BMC cell biology* 2010, **11**:5+, [[http://dx.doi.org/10.1186/1471-2121-11-5]].

[49] Bryce D, Verdicchio MP, Kim S: **Planning Interventions in Biological Networks**. *ACM Transactions on Intelligent Systems and Technology* To Appear.

[50] Bryce D, Kim S: **Planning for gene regulatory network intervention**. In *International Joint Conference on Artificial Intelligence* 2007:1834–1839.

[51] Verdicchio MP, Kim S: **Reduction of Boolean Network Basins of Attraction Reveals Intervention Targets**. Tech. rep., Arizona State University, Tempe, AZ 2010.

[52] D'agostino RB, Chase W, Belanger A: **The Appropriateness of Some Common Procedures for Testing the Equality of Two Independent Binomial Populations**. *The American Statistician* 1988, **42**(3):198–202, [[http://dx.doi.org/10.1080/00031305.1988.10475563]].

[53] Knezevic A: **Overlapping Confidence Intervals and Statistical Significance**. Cornell Statistical Consulting Unit 2008.

[54] Zhang R, Shah MV, Yang J, Nyland SB, Liu X, Yun JK, Albert R, Loughran TP: **Network model of survival signaling in large granular lymphocyte leukemia**. *Proceedings of the National Academy of Sciences* 2008, **105**(42):16308–16313, [[http://dx.doi.org/10.1073/pnas.0806447105]].

[55] Saadatpour A, Wang RS, Liao A, Liu X, Loughran TP, Albert I, Albert R: **Dynamical and Structural Analysis of a T Cell Survival Network Identifies Novel Candidate Therapeutic Targets for Large Granular Lymphocyte Leukemia**. *PLoS Comput Biol* 2011, **7**(11):e1002267+, [[http://dx.doi.org/10.1371/journal.pcbi.1002267]].

[56] Saadatpour A, Albert I, Albert R: **Attractor analysis of asynchronous Boolean models of signal transduction networks.** *Journal of theoretical biology* 2010, **266**(4):641–656, [[http://dx.doi.org/10.1016/j.jtbi.2010.07.022]].

[57] Weinberg RA: *The Biology of Cancer HB*. Garland Science, 1 edition 2006, [[http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0815340788]].

[58] Yang J, Liu X, Nyland SB, Zhang R, Ryland LK, Broeg K, Baab KTT, Jarbadan NRR, Irby R, Loughran TP: **Platelet-derived growth factor mediates survival of leukemic large granular lymphocytes via an autocrine regulatory pathway.** *Blood* 2010, **115**:51–60, [[http://dx.doi.org/10.1182/blood-2009-06-223719]].

[59] Lamy T, Liu JH, Landowski TH, Dalton WS, Loughran TP: **Dysregulation of CD95/CD95 ligand-apoptotic pathway in CD3(+) large granular lymphocyte leukemia.** *Blood* 1998, **92**(12):4771–4777, [[http://view.ncbi.nlm.nih.gov/pubmed/9845544]].

[60] Tanay A, Sharan R, Shamir R: **Biclustering Algorithms: A Survey**. In *In Handbook of Computational Molecular Biology Edited by: Aluru S. Chapman and Hall/CRC Computer and Information Science Series* 2005.

[61] Dougherty E, Brun M, Trent J, Bittner M: **Conditioning-Based Modeling of Contextual Genomic Regulation**. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on : Accepted for future publication* 2008.

[62] Kim S, Sen I: **Mining Molecular Contexts of Cancer via in-silico Conditioning**. In *Proceedings of Sixth International Conference on Comput Syst Bioinf*, San Diego, CA 2007:169–179.

[63] Sen I, Verdicchio M, Jung S, Trevino R, Bittner M, Kim S: **Context-Specific Gene Regulations in Cancer Gene Expression Data**. In *Proceedings of Pac Symp Biocomput* 2009.

[64] Sen I: **A Computational Framework to Model and Learn Context-Specific Gene Regulatory Networks from Multi-Source Data**. *PhD thesis*, Arizona State University 2011.

[65] Kim S, Roy I, Raghavan S, Dougherty ER, Bittner M: **Learning Context-Specific Gene Regulatory Networks via In-Silico Conditioning**. In *2007 IEEE International Workshop on Genomic Signal Processing and Statistics*, IEEE 2007:1–4, [[http://dx.doi.org/10.1109/GENSIPS.2007.4365831]].

[66] Jung S, Lee KH, Lee D: **H-CORE: Enabling genome-scale Bayesian analysis of biological systems without prior knowledge.** *Biosystems* 2007, **90**:197–210, [[http://dx.doi.org/10.1016/j.biosystems.2006.08.004]].

[67] Friedman N, Nachman I, Pe'er D: **Learning Bayesian Network Structure from Massive Datasets: The "Sparse Candidate" Algorithm**. In *Proc. Fif-*

teenth Conf. on Uncertainty in Artificial Intelligence (UAI) 1999:206–215, [[http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.37.5605]].

[68] Chen Y, Kamat V, Dougherty ER, Bittner ML, Meltzer PS, Trent JM: **Ratio statistics of gene expression levels and applications to microarray data analysis.** *Bioinformatics (Oxford, England)* 2002, **18**(9):1207–1215, [[http://view.ncbi.nlm.nih.gov/pubmed/12217912]].

[69] Weinberg OK, Marquez-Garban DC, Fishbein MC, Goodglick L, Garban HJ, Dubinett SM, Pietras RJ: **Aromatase Inhibitors in Human Lung Cancer Therapy**. *Cancer Research* 2005, **65**(24):11287–11291, [[http://dx.doi.org/10.1158/0008-5472.CAN-05-2737]].

[70] Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proceedings of the National Academy of Sciences* 1998, **95**(25):14863–14868, [[http://dx.doi.org/10.1073/pnas.95.25.14863]].

[71] Rivas MA, Carnevale RP, Proietti CJ, Rosemblit C, Beguelin W, Salatino M, Charreau EH, Frahm I, Sapia S, Brouckaert P, Elizalde PV, Schillaci R: **TNF alpha acting on TNFR1 promotes breast cancer growth via p42/P44 MAPK, JNK, Akt and NF-kappa B-dependent pathways.** *Experimental cell research* 2008, **314**(3):509–529, [[http://dx.doi.org/10.1016/j.yexcr.2007.10.005]].

[72] Burton JD, Goldenberg DM, Blumenthal RD: **Potential of peroxisome proliferator-activated receptor gamma antagonist compounds as therapeutic agents for a wide range of cancer types.** *PPAR research* 2008, **2008**, [[http://dx.doi.org/10.1155/2008/494161]].

[73] Oldford SA, Robb JD, Codner D, Gadag V, Watson PH, Drover S: **Tumor cell expression of HLA-DM associates with a Th1 profile and predicts improved survival in breast carcinoma patients**. *International Immunology* 2006, **18**(11):1591–1602, [[http://dx.doi.org/10.1093/intimm/dxl092]].

[74] Hao Y, Wang J, Feng N, Lowe AW: **Determination of plasma glycoprotein 2 levels in patients with pancreatic disease.** *Arch Pathol Lab Med* 2004, **128**(6):668–74+, [[http://ukpmc.ac.uk/abstract/MED/15163232]].

[75] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontol-**

ogy: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 2000, **25**:25–29, [[http://dx.doi.org/10.1038/75556]].

[76] Zeeberg B, Qin H, Narasimhan S, Sunshine M, Cao H, Kane D, Reimers M, Stephens R, Bryant D, Burt S, Elnekave E, Hari D, Wynn T, Rundles CC, Stewart D, Nelson D, Weinstein J: **High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID)**. *BMC Bioinformatics* 2005, **6**:168+, [[http://dx.doi.org/10.1186/1471-2105-6-168]].

[77] Ramesh A, Trevino R, Von Hoff D, Kim S: **Clustering Context-Specific Gene Regulatory Networks**. In *Proceedings of Pac Symp Biocomput, Volume 15*, Hawaii 2010:444–455.

[78] Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D: **Molecular portraits of human breast tumours**. *Nature* 2000, **406**(6797):747–752.

[79] Nowicki MO, Pawlowski P, Fischer T, Hess G, Pawlowski T, Skorski T: **Chronic myelogenous leukemia molecular signature**. *Oncogene* 2003, **22**(25):3952–3963.

[80] Grzegorczyk M, Husmeier D, Edwards K, Ghazal P, Millar A: **Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian network and the allocation sampler**. *Bioinformatics* 2008, **24**(18):2071–2078.

[81] Kim S, Sen I: **Mining Molecular Contexts of Cancer via in-silico Conditioning**. In *Proceedings of Sixth International Conference on Comput Syst Bioinf*, San Diego, CA 2007:169–179.

[82] Dougherty E, Brun M, Trent J, Bittner M: **Conditioning-Based Modeling of Contextual Genomic Regulation**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2009, **6**(2):310–320.

[83] van Dongen S: **Graph Clustering by Flow Simulation**. *PhD thesis*, University of Utrecht 2000.

[84] Chang J, Nevins J: **GATHER: a systems approach to interpreting genomic signatures**. *Bioinformatics* 2006, **22**(23):2926–2933.

[85] Subramanian A, Tamayo P, Mootha V, Mukherjee S, Ebert B, Gillette M, Paulovich A, Pomeroy S, Golub T, Lander E, Mesirov J: **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci USA* 2005, **102**(43):15545–15550.

[86] Kervizic G, Corcos L: **Dynamical modeling of the cholesterol regulatory pathway with Boolean networks**. *BMC Systems Biology* 2008, **2**(99).

[87] Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N: **Revealing modular organization in the yeast transcriptional network**. *Nature Genetics* 2002, **31**(4):370–377.

[88] Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, Harris M, Hill D, Issel-Tarver L, Kasarskis A, Lewis S, Matese J, Richardson J, Ringwald M, Rubin G, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25–29.

[89] Belenky P, Bogan K, Brenner C: **NAD+ metabolism in health and disease**. *Trends Biochem Sci* 2007, **32**:12–19.

[90] Chen D, Guo J, Gahl W: **RAB GTPases expressed in human melanoma cells**. *Biochimica et Biophysica Acta (BBA)/Molecular Cell Research* 1997, **1355**:1–6.

[91] Shembade N, Parvatiyar K, Harhaj NS, Harhaj EW: **The ubiquitin-editing enzyme A20 requires RNF11 to downregulate NF-[kappa]B signalling**. *EMBO J* 2009, **28**(5):513–522.

[92] Mashima T, Naito M, Tsuruo T: **Caspase-mediated cleavage of cytoskeletal actin plays a positive role in the process of morphological apoptosis**. *Oncogene* 1999, **18**:2423–2430.

[93] Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, Alexe G, Lawrence M, O'Kelly M, Tamayo P, Weir BA, Gabriel S, Winckler W, Gupta S, Jakkula L, Feiler HS, Hodgson JG, James CD, Sarkaria JN, Brennan C, Kahn A, Spellman PT, Wilson RK, Speed TP, Gray JW, Meyerson M, Getz G, Perou CM, Hayes DN: **Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1**. *Cancer cell* 2010, **17**:98–110, [[http://linkinghub.elsevier.com/retrieve/pii/S1535610809004322]].

[94] Magrassi L, Adomi L, Montorfano G, Rapelli S, Butti G, Berra B, Milanesi G: **Vitamin D metabolites activate the sphingomyelin pathway and induce death of glioblastoma cells**. *Acta Neurochir* 1998, **140**(7):707–713.

[95] Casorelli I, Russo TM, Bignami M: **Role of Mismatch Repair and MGMT in Response to Anticancer Therapies**. *Anti-Cancer Agents in Medicinal Chemistry* 2008, **8**:368–380.

[96] Harvey I, Bossomaier T: **Time Out of Joint: Attractors in Asynchronous Random Boolean Networks**. In *Proceedings of the Fourth European Conference on Artificial Life (ECAL97* 1997:67–75, [[http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.48.6693]].

[97] Gershenson C: **Updating Schemes in Random Boolean Networks: Do They Really Matter?** In *Artificial Life IX, Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems*. Edited by Pollack J, Bedau M, Husbands P, Ikegami T, Watson RA, MIT Press 2004:238–243, [[http://arxiv.org/abs/nlin/0402006]].

[98] Wuensche A: **Discrete Dynamics Lab**. In *Artificial Life Models in Software*. Edited by Komosinski M, Adamatzky A, London: Springer London 2009:215–258, [[http://dx.doi.org/10.1007/978-1-84882-285-6_8]].

[99] Cormen TH, Leiserson CE, Rivest RL, Stein C: *Introduction to Algorithms, Third Edition*. The MIT Press, 3rd edition 2009, [[http://portal.acm.org/citation.cfm?id=1614191]].

[100] Pearl J: *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Santa Mateo, CA, USA: Morgan Kaufmann 1997, [[http://www.worldcat.org/isbn/1558604790]].

[101] Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian networks to analyze expression data.** *Journal of Computational Biology* 2000, **7**(3-4):601–620, [[http://dx.doi.org/10.1089/106652700750050961]].

[102] Simons KT, Kooperberg C, Huang E, Baker D: **Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions.** *Journal of molecular biology* 1997, **268**:209–225, [[http://dx.doi.org/10.1006/jmbi.1997.0959]].

[103] Carvalho AM: **Scoring functions for learning Bayesian networks**. Tech. rep., Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento 2009, [[http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.157.3919]].

[104] Tsamardinos I, Brown L, Aliferis C: **The max-min hill-climbing Bayesian network structure learning algorithm**. *Machine Learning* 2006, **65**:31–78, [[http://dx.doi.org/10.1007/s10994-006-6889-7]].

APPENDIX A

BOOLEAN NETWORKS

Despite its simplicity, the Boolean network model has proven to be quite viable at approximating certain aspects of biological processes. For example, it has been used to simulate the yeast cell cycle [21], which we looked at closely in our work [22]. It has also been used to simulate the expression pattern of segment polarity genes in *Drosophila melanogaster* [23], as well as the vocal communication system of the songbird brain [24, 25]. Since Kauffman's seminal work[1] there have been countless variations and extensions of the use of Boolean networks for modeling biological systems, and various inference procedures have been proposed for them [26, 27, 28].

A.1 Formal Description

A Boolean network $\mathbf{B}(V,\mathbf{f})$ is made of a set of binary nodes $V = \{x_1, x_2, \cdots, x_n\}$, and a set of functions $\mathbf{f} = \{f_1, f_2, \cdots, f_n\}$. The binary value of variable $x_i \in V$ at time $(t+1)$ is determined by other variables $x_{j_1(i)}, x_{j_2(i)}, ..., x_{j_{k_i}(i)}$ at time (t) by means of a Boolean function $f_i \in \mathbf{f}$. That is, there are k_i variables assigned to x_i, and the mapping $j_k : \{1, \cdots, n\} \rightarrow \{1, \cdots, n\}, k = 1, \cdots, k_i$ determines the "wiring" of variable x_i. Thus k_i is called the *connectivity* of x_i, which is to say the number of inputs to its particular Boolean function. The values of the variables at time $(t+1)$ are always a function of the values of the k_i respective input variables at time t. Formally,

$$x_i(t+1) = f_i(x_{j_1(i)}(t), x_{j_2(i)}(t), ..., x_{j_{k_i}(i)}(t)) \tag{A.1}$$

State Transition Diagram

The state of a Boolean network at time t refers to the state vector for all variables, $\mathbf{x}(t) = (x_1(t), x_2(t), ..., x_n(t))$, where a specific state can be expressed as an n-dimensional binary vector $(z_1, z_2, ..., z_n)$, where $z_i \in \{0,1\}, 1 \leq i \leq n$, or in a more compact form as $z_1 z_2 \cdots z_n$. The state space of the network is $\{0,1\}^n = \{00\cdots0, 00\cdots1, ..., 11\cdots1\}$, whose size is 2^n. Letting $\mathbf{x}(t)$ take on the value of each of the possible 2^n states and obtaining the next states

$\mathbf{x}(t+1)$ gives a set of 2^n one-step transitions that completely characterize the dynamics of the system. Let this set of all states be called S, such that $S = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{2^n}\}$, and let the set of all transitions between the states of S be called E. The state transition diagram $\mathbf{G}(S,E)$ for a Boolean network $\mathbf{B}(V,\mathbf{f})$ with n nodes is a directed graph where $|\mathrm{S}| = |\mathrm{E}| = 2^n$. Each of the vertices $\mathbf{x} \in S$ represents one possible configuration of $x_1, x_2, ..., x_n$ and each of the directed edges represents one of the one-step transitions between two states as we synchronously apply Boolean functions to all variables.

We adhere to the traditional, synchronous update scheme due to its origins in relating to biological cell types[1], and because its determinism is exploited by our analysis approach. Some correctly claim that real biological systems do not "march in step" and that asynchronous update mechanisms are more appropriate [96]. However, synchronous networks are able to compute and to reduce complexity better than asynchronous non-deterministic ones and that synchronicity is still related to living systems[97].

Attractors and Basins

Wuensche [2] and others have studied the basins of attraction in Boolean network models of genomic regulation, specifically the relationship of their structures to the stability of attractors (cell types) in the face of perturbations. However, because of the transient nature and size complexity of basins of attraction, they are often neglected in analysis in favor of the attractor states. As a basin of attraction is a collection of states leading into a corresponding attractor, i.e. phenotype, careful analysis of these basins could reveal interesting biological characteristics that determine cell fate, which our recent study addressed[22].

In the absence of interventions or perturbations, beginning in any initial state, repeated application of transition functions will bring the network to a finite set of states and cycle among them forever in fixed sequence. This set of states is known as an *attractor*, denoted A_i. An attractor with just one state is called a singleton attractor and an attractor

with more than one state is called a cyclic attractor. Boolean networks may have anywhere from one cyclic attractor comprised of 2^n states to 2^n point attractors, although most commonly a network will have just a handful of singleton or short-cycle attractors. The complete set of states from which a network will eventually reach A_i is known as the *basin of attraction* for A_i, denoted B_i. Formally, the states of basin B_i are precisely those that, given $w \leq 2^n$ applications of Boolean functions to an evolving state, end up in attractor A_i: $B_i = \{\mathbf{x} \mid f^{(w)}(\mathbf{x}) \in A_i\}, i, w \leq 2^n$. The basins of attraction correspond precisely to the *weakly connected components* of the state transition diagram (i.e. a subgraph such that every pair of vertices *u* and *v* is connected by a directed path from *u* to *v* and an undirected path from *v* to *u*), and the attractors correspond precisely to the *strongly connected components* of the state transition diagram (i.e. a subgraph such that every pair of vertices *u* and *v* is connected by a directed path from *u* to *v* and also from *v* to *u*).

An individual basin B_i and its attractor A_i can be described in terms of the collection of states comprising them, $A_i \subseteq B_i \subseteq S = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{2^n}\}$. Let the size of each attractor $\|A_i\| = p_i$, where p_i is the period, or length of the attractor cycle. An attractor with $p = 1$ is called a point, or singleton attractor, and an attractor with $p > 1$ is called a cyclic attractor (with cycle length equal to p). If $\mathbf{x}_i$ is a state in A_i, we can describe the next state of a point attractor as $\mathbf{x}_i(t+1) = \mathbf{x}_i(t)$, and the behavior of a cyclic attractor as $\mathbf{x}_i(t+p) = \mathbf{x}_i(t)$.

All attractors are subsets of their basins (i.e. $A_i \subseteq B_i, \forall_i$), all basins (and concordantly all attractors also) are mutually exclusive (i.e. $B_i \bigcap B_j = \emptyset, \forall_{i \neq j}$), and the complete state space is comprised entirely of all basins (i.e. $\bigcup_i B_i = S$). For referencing specific basins and attractors, the set of all basins is denoted $\mathbf{B} = \{B_1, B_2, \cdots, B_L\} = S$, and the set of all corresponding attractors is denoted $\mathbf{A} = \{A_1, A_2, \cdots, A_L\} \subseteq S, 1 \leq L \leq 2^n$.

A.2 Tools and Algorithms for Modeling and Simulating

In our studies we use our own software in conjunction with the BN/PBN Toolbox [42], as well as DDLab[98], for Boolean network simulation and processing. While implementation details differ, the underlying fundamentals are the same between tools. One must store the topology of the Boolean network, which is contained within the functions $\mathbf{f}$). For each V, one must also maintain the truth table of values for each combination of its input variables. The elementary operation in simulation studies is *NextState*, which updates all variables based on the current values in the network.

The operation of the algorithms described was initially given by Kauffman [1] over forty years ago and is formalized here with a consistent notation. The correctness of the algorithms is evident by the simplicity of the *NextState* operation and by the definitions of attractors and attractor basins given in the previous section. Trivial proofs by contradiction would assert that any incorrectly assigned basin or attractor states along a transient path from any state towards an attractor could not be produced by the *NextState* operation, which synchronously updates all variables based on their current values.

Exhaustive Enumeration of Attractors and Basins

For a reasonable number of variables, enumerating all 2^n states in the state transition diagram is not an issue. By starting at each state and evolving the network forward, each attractor and its basin can be enumerated, as shown in Algorithm A.1. Exhaustive enumeration is the best possible situation for logic minimization. This is because with more states, more common values can be identified and summarized in the reduction. This algorithm's run time is exponential in n, and specifically is $O(2^n)$. This is due to the outer loop running once for each of the 2^n states, while the inner loop executes at most 2^n times across all instances of the outer loop.

Input: $\mathbf{B}(V,\mathbf{f})$

Output: $G(S,E)$

1 $S = [1:2^n]$, $k = 1$;

2 **foreach** *State s : S* **do**

3 **while** *s is previously unseen* **do**

4 $s.attractorID \leftarrow k$;

5 $E.addEdge(s, NextState(s))$;

6 $s \leftarrow NextState(s)$;

7 **end**

8 **switch** $s.attractorID$ **do**

9 **case** $\neq k$*:* old attractor found, update path, break;

10 ;

11 **case** $= k$*:* new attractor found, $k = k+1$, break;

12 ;

13 **endsw**

14 **end**

Algorithm A.1: Exhaustive Enumeration of Attractor Basins

Sampling of Attractors and Basins

Unfortunately, as values of n pass 20+ variables, exhaustive enumeration of the state space (size 2^n) quickly becomes intractable and a sampling approach is required. In this approach we simply replace line 1 of Algorithm A.1 with a vector of states uniformly sampled from the space $[1:2^n]$. For the sake of reference this adapted algorithm is shown in in Algorithm A.2. This approach will sometimes miss attractors with very small basins leading to them, but it certainly finds the largest ones, and for a large number of samples, gives us a significant set of member states to analyze. Additionally we can also approximate the percentage of the total state space occupied by each basin based on the percentage of total samples

associated with it. Since the worst case of this algorithm is having exactly 2^n samples, it is in fact a special case of Algorithm A.1 and thus the run time complexity is the same, namely, $O(2^n)$.

Input: $\mathbf{B}(V,\mathbf{f})$, *numSamples*

Output: $G(S,E)$

$S = \emptyset$, $k = 1$;
for *i=1* **to** *numSamples* **do**
    Let x be a discrete uniform random variable with value from the range $[1, 2^n]$;
    $S = S \cup \{x\}$;
end
foreach *State s : S* **do**
    while *s is previously unseen* **do**
        $s.attractorID \leftarrow k$;
        $E.addEdge(s, NextState(s))$;
        $s \leftarrow NextState(s)$;
    end
    switch $s.attractorID$ **do**
        case $\neq k$*:* old attractor found, update path, break;
        ;
        case $= k$*:* new attractor found, $k = k + 1$, break;
        ;
    endsw
end

Algorithm A.2: Exhaustive Enumeration of Attractor Basins

Template Counting Algorithm

The template counting algorithm from Ch. 3 is reproduced here for a discussion of run time complexity.

```
1 foreach Basin B_i : B do
2     for k = 1 : K do
3         foreach Template template_{j,i} for j = 1 :2^k (n choose k) do
4             foreach Term t_l : T_i do
5                 if Template template_{j,i} is found in t_l then
6                     increment count(template_{j,i}) for k
```

Algorithm A.3: Template Analysis in Attractor Basins, reproduced from Algorithm 3.1

Step 1 of this algorithm is a simple partitioning of the total iterations, and is thus given a constant overhead c_1. Step 2 of this algorithm is governed by the value K. Step 3 executes once for each unique template, namely, $2^K \binom{n}{K}$ times. Step 4 executes once for each term, where the number of terms is at most 2^n (i.e. no logic minimization at all), but according to Table 3.3 (and observation in the simulation study in Chapter 3) is typically a fraction of 2^n proportional to the sampling coverage, we will assign it $2^n/c_2$. Steps 5 and 6 are a constant time operation, and so we assign a cost c_3. Because all steps are nested, our run time is a product:

$$c_1 * K * 2^K * \binom{n}{K} * \left(\frac{2^n}{c_2}\right) * c_3$$

It is known that the binomial coefficient $\binom{n}{k}$ is bounded above by $(n*e/k)^k$[99]. Thus,

$$c_1 * K * 2^K * \left(\frac{n * e}{K}\right)^K * \left(\frac{2^n}{c_2}\right) * c_3$$

Because we have a product, we can remove all constants, namely c_1, c_2, c_3, and e. Thus, with O notation,

$$O(K * 2^K * \left(n/K\right)^K * 2^n)$$

In practice, K will be a small number, and certainly much less than n. For the sake of formality, let K be bounded by a constant, c, and not dependent upon n. We can then simplify:

$$O(K * 2^K * \left(n/K\right)^K * 2^n) \leq$$

$$O(c * 2^c * (n/c)^c * 2^n) \leq$$

$$O(2^c * 2^n).$$

APPENDIX B

ROBUSTNESS OF SINGLE VARIABLE MEASURES

We claim that our small network measures of popularity and power identify targets for intervention in biological interaction networks modeled in the Boolean network framework. Furthermore we claim that the targets identified by our measures are better than those that can be identified by related approaches for identifying key network nodes, including two measures specific to Boolean networks. To justify this claim we present the following simulation experiment.

B.1 Problem Statement and Hypothesis

Are the popularity and power measures (Eqs. 2.1, 2.3) superior to related network measures for finding targets for intervention? We hypothesize that the popularity and power measures will lead to desired intervention outcomes with a greater rate of success than other methods and design the following experiment to test this prediction.

B.2 Experiment

Abstractly, an intervention should shift the steady behavior of a system to a different (usually more desirable) state. In the context of the Boolean network formalism, this is represented by shifting the steady behavior of the system, represented by an attractor state or cycle, into a different basin of attraction. A successful intervention needs only to shift the state from the starting attractor state into any state in the basin of the goal attractor, as the network dynamics will naturally bring the state to the attractor itself. Thus, we define a successful intervention as a single-variable modification to an attractor state which generates a state in the goal attractor basin.

Restrictions and Omitted Interventions

We make a few restrictions on our experiment to ensure a fair outlook on the results. First, attractor states are extremely stable and difficult to break out of with only a single variable perturbation. Nonetheless we restrict interventions to one single variable to induce a challenge and to simplify the experiment. Second, a “self-intervention” where the source and

goal basins of attraction are the same could be considered a “maintenance” intervention if the basin is a desirable one; however this is a special case at best, and the robustness of attractor states will generate higher intervention success rates, inflating the overall performance. Thus we do not include interventions were the source and goal basins are the same, which also means that we discard networks with only one attractor basin. Third, our artificial intelligence planning work[49] showed that reaching the smallest, rarest basins often requires a series of interventions. Attempting these interventions with a single variable perturbation would unfairly bring scores down, so we omit interventions with a goal basin occupying less than 15% of the total state space.

Types of Interventions Performed

To create a simple comparison across a variety of approaches, we perform a single-variable perturbation upon the top variable candidate from each measure. If more than one variable shares the top rank for a particular measure, we choose between them randomly. If the measure includes a way to provide a value for the perturbation, we use the given value, otherwise we choose a value at random to be shared across any measure requiring one; in this way if more than one measure wishes to perturb the same variable but does not specify a value, they will all perturb the variable with the same random value so that no one method is advantaged within a particular instance.

We offer 14 methods for selecting an intervention target for comparison, with some of those being combinations of related measures. These methods are listed in Table B.1 and are described here. First we offer our popularity and power measures (Eqs. 2.1, 2.3). Empirical observations have shown us that sometimes the best variable can be identified by a collaboration between related measures. For example, let variable x have the highest popularity score (but a low power score), and let variable y have a high power score (but a low popularity score). Another variable z may not score higher in popularity or power

than *x* or *y*, but it may have somewhat high values for both. In this case the harmonic mean of popularity and power may reveal *z* to be the top variable over *x* or *y*; thus we include the harmonic mean of popularity and power as the third way to identify an intervention target. These three measures are unique in that they have a specific value for each variable to be used in the intervention. This value used is based on the frequency of *ON* values in the reduced sum-of-product terms for the particular basin of attraction; if the frequency is greater than or equal to 0.5, a value of 1 is used, else, a value of 0 is used.

The next six methods for identifying an intervention target are variations of the influence and sensitivity measures (Described in §2.4 and in Appendix C), which are designed specifically for Boolean networks[3]. We use influence, sensitivity, and their harmonic mean to identify the best target for intervention. Following, we use our basin-specific versions of influence and sensitivity (and their harmonic mean) to offer a basin-specific advantage akin to our popularity and power measures.

The remaining methods are related to the original network topology and are described in detail in Appendix C. First, we use the node with the highest degree, including both inbound and outbound edges, but not double-counting for self-regulation. Next, we use the nodes with the top centroid value, eccentricity value and betweenness centrality value. We also include an intervention using a randomly chosen target for comparison purposes.

B.3 Methodology of the Experiment

We use Algorithm B.1 to automate our experiments, repeating it for three different small network sizes. Each network is randomly generated using the biologically inspired connectivity probability distribution given in Table 3.1. Rather than intervening upon randomly selected states, we intervene only upon attractor states. This is because we expect to find the system in a steady state, so it is the logical subset of states upon which to intervene.

While some basins have a singleton attractor state, many basins have cyclic attractors with many states. To maximize the efficiency of the simulation, we will attempt an intervention from every attractor state in every basin in the state space to every other basin of significant size. This will allow us to attempt many interventions for each random network generated.

Input: Number of Variables, *n*; Number of Networks, *nets*

for $i = 1 : nets$ **do**
    Generate Random Boolean Network $\mathbf{BN}(V, \mathbf{f})$ with *n* variables;
    Enumerate Basin of Attraction Field $\mathbf{G}(S, E)$;
    if *Less than Two Attractor Basins* **then** Discard Network, Continue;
    ;
    Compute all measures to identify intervention targets;
    foreach *Basin* $B_i \in \mathbf{B}$ *as the SOURCE* **do**
        foreach *Attractor State* $\mathbf{a_i} \in A_i \in B_i$ **do**
            foreach *Basin* $B_j \neq B_i$ *as the GOAL* **do**
                if B_j *occupies* $<$ *15% of* **B then** Discard Intervention, Continue;
                ;
                foreach *Intervention Target Method* **do**
                    Select one top intervention target;
                    Select or generate value for target;
                    Apply intervention to $\mathbf{a_i}$ to produce new state $\mathbf{x_k}$;
                    Identify *DESTINATION* basin B_k such that $\mathbf{x_k} \in A_k \in B_k$;
                    if $\mathbf{B_j} == \mathbf{B_k}$ **then**
                        Record Success;
                    else
                        Record Failure;

Algorithm B.1: Algorithm for Small Network Measures Comparison: Notation from the algorithm is defined in Appendix A. The *Continue* statement indicates skipping the remainder of the specified loop's current iteration and continuing the same loop at the next iteration.

We apply Algorithm B.1 with the following input parameters: $n = \{7, 8, 9, 10, 11, 12, 13, 14, 15, 16\}$, and $nets = 400$ for each network size. Due to the varying numbers of attractor basins as well as the varying lengths of attractor cycles, these input combinations resulted in 95,870 attempted interventions.

B.4 Results of Experiment

Overall we observed the dominance of the popularity and power measures (along with their harmonic mean) in identifying targets for intervention over other strategies, which is shown in Table B.1 and in Fig. B.1. The requested interventions were increasingly difficult with network size due to their restriction to a single variable. The benefits of counting easy "self-interventions" (i.e. staying in the same basin) were not included, and the detriments of attempting interventions to very rare basins were also omitted to get a fair idea of performance. The influence measure showed the best behavior after the popularity and power measures, and the topological measures performed poorly.

On the smaller networks, our measures dominate with a great deal of separation in performance. As the network size increases, however, the performance of the various methods begins to converge. This is the primary motivation for our introduction of template based approaches. We note that besides our methods, the Boolean-network-specific measure of influence (and it's basin-specific counterpart) perform the best on average of the related measures.

Sensitivity is a measure such that variables with low sensitivity can maintain their states better than others with higher sensitivity[29]. However this attribute does not imply the potential for regulatory control like the influence measure does. Thus it is not surprising to see influence dominate the sensitivity measure. It has been said that a variable with a low sensitivity but a high influence would make the best intervention targets[29]; however while our results certainly shown an improvement over sensitivity alone by the harmonic mean

of influence and (low) sensitivity, the combination of the two was not able to outperform influence alone in either the whole-network or basin-specific cases. Of the topological measures, only node degree and betweenness passed, on average, the performance of a random intervention target. Centroid and eccentricity failed completely, oftentimes being outperformed by a random intervention.

Intervention success counts and numbers of trials were used to calculate confidence intervals according to the Binomial test, since our interventions qualified as Bernoulli trials. While we can visually see separation in the figure and elevated average and individual performances, the confidence intervals of our small network measures did indeed overlap with some of the best scoring related measures. But, while topological measures can certainly lend insight at the top level of networks, the underlying complexities of the Boolean model are clearly better elucidated by measures designed specifically for the Boolean model. Of those, the robustness of the popularity and power measures is clearly demonstrated, despite their inability to statistically separate in the given trial. We hypothesize that if the interventions were to begin from random states and not just from attractor states that we would gain the separation needed for a statistical significance; however our motivation to move to larger networks and introduce template based approaches renders this point moot.

B.5 Conclusions

We believe the dominating success of popularity and power measures is due to the inherent inclusion of a specific value to use in the intervention, and that both target variables and their values are specific to the goal basin. In each intervention instance, both the target and the value were chosen from the popularity and power measures specific to the goal basin. These advantages absent in other methods clearly resulted in the higher performance. It is also not surprising to see the influence measures appearing next in the performance after our measures for two reasons. First, like our measures, it is a measure designed in and thus taking advantage of the Boolean network model. Second, the basin-specific influence we

Table B.1: Single Variable Network Measures Compared: We have shown the ability of the single variable measures of popularity and power to identify intervention targets missed by other related measures. In this table we report average success estimates for intervening with a single variable chosen by a variety of methods and their combinations. Random networks with between 7 and 16 variables were generated and those networks with between 2 and 7 basins of attraction were intervened upon. Overall, we observed popularity, power, and their harmonic mean as supplying the most effective intervention targets. After these, the Boolean-network-specific measures of influence and sensitivity performed best, especially their basin-specific versions invented for this study. Thus we conclude that popularity, power and the harmonic mean of popularity and power produce the best single variable intervention targets.

Variables:	7	8	9	10	11	12	13	14	15	16	AVG
Popularity	0.297	0.325	0.296	0.302	0.292	0.292	0.289	0.318	0.312	0.314	0.304
Power	0.315	0.311	0.289	0.281	0.286	0.294	0.277	0.314	0.320	0.314	0.300
$H(p(v), P_V(v))$	0.295	0.317	0.288	0.299	0.291	0.293	0.281	0.314	0.314	0.314	0.301
Influence	0.266	0.281	0.278	0.285	0.280	0.276	0.292	0.299	0.307	0.314	0.288
Sensitivity	0.250	0.265	0.265	0.250	0.245	0.275	0.249	0.268	0.275	0.276	0.262
$H(Infl., Sens.)$	0.272	0.276	0.273	0.287	0.272	0.277	0.284	0.274	0.304	0.310	0.283
Basin Influence	0.280	0.273	0.276	0.283	0.278	0.277	0.291	0.298	0.307	0.314	0.288
Basin Sensitivity	0.251	0.272	0.257	0.249	0.241	0.272	0.246	0.271	0.279	0.272	0.261
$H(B. Infl., B. Sens.)$	0.260	0.278	0.273	0.285	0.269	0.279	0.284	0.273	0.299	0.303	0.280
Degree	0.262	0.267	0.270	0.250	0.269	0.283	0.263	0.285	0.297	0.306	0.275
Centroid	0.234	0.244	0.262	0.230	0.233	0.262	0.242	0.252	0.253	0.298	0.251
Eccentricity	0.235	0.246	0.255	0.221	0.239	0.273	0.237	0.256	0.256	0.293	0.251
Betweenness	0.258	0.274	0.269	0.253	0.276	0.283	0.273	0.285	0.289	0.309	0.277
Random	0.254	0.264	0.255	0.237	0.247	0.280	0.252	0.254	0.277	0.304	0.262

designed to add a degree of fairness demonstrated a benefit from that modification. The topological measures without either basin or value specificity naturally performed the most poorly.

Despite a lack of complete separation of confidence intervals, but based on the consistent top performance of the popularity and power measures across various network sizes and over thousands of interventions, we fail to reject our hypothesis and conclude that the popularity and power measures lead to desired intervention outcomes with a greater rate of success than other topological, centrality-based, and even Boolean network specific methods.

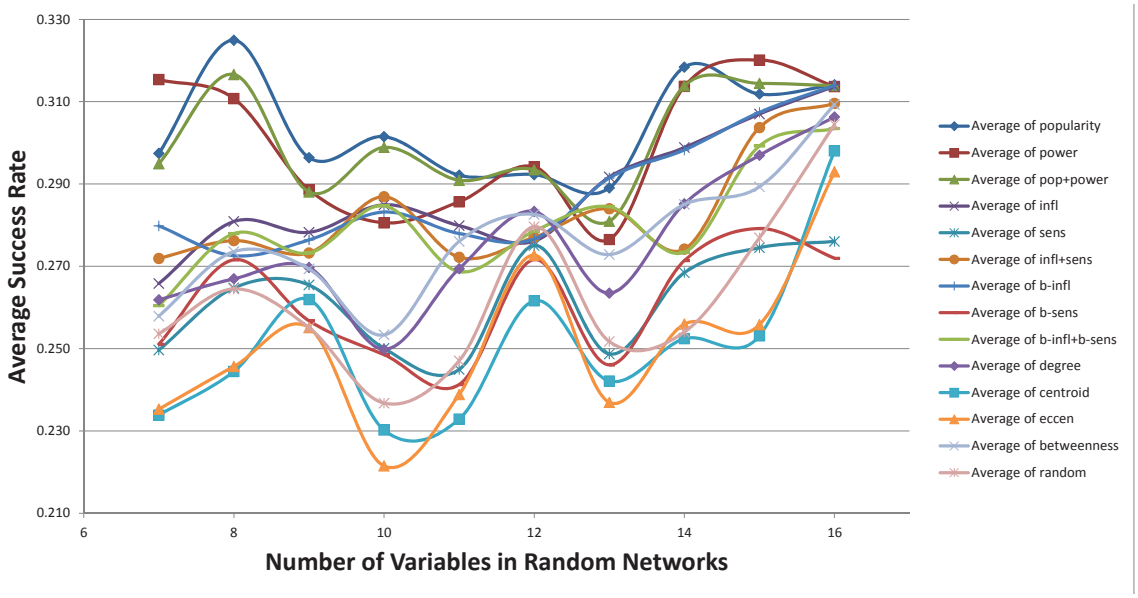


Figure B.1: Small Network Robustness Experiment Results: Here we see a graphical representation of Table B.1. We observe the dominance of popularity, power, and their harmonic mean over other related measures..

APPENDIX C

OTHER NETWORK CENTRALITY AND TOPOLOGICAL MEASURES

Other network measures, including topological ones as well as formalized centrality concepts, can certainly help identify key players in a network. We found, however, for each network we looked at, that these measures added no new insight into the reporting of key players and potential intervention targets. There were many instances of overlaps in reporting, however many times variables with no mathematical control in the network were elevated by these measures, and many times influential variables were left out by these measures. Thus we report findings for some of the networks in this appendix, rather than Ch. 2, to document and discuss these values, their usefulness and their limitations.

C.1 Influence and Sensitivity

We have seen the ability of popularity (p) and variable power (P_V) (Eq. 2.1, Eq. 2.3) to identify variables with great combinatorial control over the state space of a Boolean network. We have further demonstrated how those variables identified are often known to be suitable targets for intervention. In demonstrating this we have compared p and P_V to influence and sensitivity ($r(x_i)$ and $s(x_i)$), as well as network centrality measures, and here we discuss some differences in these measures.

While $r(x_i)$ and $s(x_i)$ are based on Boolean functions, p and P_V are based on Boolean states. Influence[29] is computed by variable pairs in a matrix and summed by rows and columns to get $r(x_i)$ and $s(x_i)$, where p and P_V are independent measurements on variables and do not depend on pairs. $r(x_i)$ and $s(x_i)$ are general measures, where p and P_V are specific to each basin of attraction. To level the field of comparison, we created a basin-specific version of $r(x_i)$ and $s(x_i)$ ($r_k(x_i)$ and $s_k(x_i)$ for basin k), but they were not able to offer any new insight that $r(x_i)$ and $s(x_i)$ were not already able to.

There are additional advantages over $r(x_i)$ and $s(x_i)$. p and P_V are not only basin-specific, but they are also value-specific. While we can adapt an influence matrix to be basin-specific, it still cannot be made value-specific. Thus, with p and P_V, because of the

minimized terms, we not only know where to intervene, but precisely how to do so. These values, or how we should intervene, can be and often are different than the values in the attractor state (if we're lucky enough to not have a cyclic attractor where values toggle), and furthermore the same target may be viable for more than one basin, but with different values. This kind of information is not available with an influence matrix or the derived measures $r(x_i)$ and $s(x_i)$.

Furthermore, p and P_V allow us to find the minimal effective intervention. Any computational aid to intervention studies will always be human-reviewed in the end, so it need not give one definitive answer. We can say with mathematical certainty that setting certain variables together will force a basin (and thus attractor) to be selected. With a set of minimized terms we can find the smallest interventions (highest P_T) using the most effective targets (high p and/or P_V) which are suitable for intervention with current medical abilities (human evaluation of mathematical possibilities).

C.2 Network Topology and Centrality Measures

Several common network topology measures exist and can reveal key players in networks. While they report some of our key players, they miss certain critical ones and also report others shown (mathematically) to have no intervention control capabilities. While useful in identifying important variables topologically speaking, they may be misleading when searching for suitable intervention targets. For the biological context, Scardoni *et al.*[40] detail these measures in the supplementary information of their manuscript, and we reproduce and summarize parts of it here.

Degree

The degree allows an immediate evaluation of the regulatory relevance of the node. For instance, in signaling networks, proteins with very high degree are interacting with several other signaling proteins, thus suggesting a central regulatory role, that is they are likely

to be regulatory ”hubs”. For instance, signaling proteins encoded by oncogenes, such as HRAS, SRC or TP53, are hubs. Depending on the nature of the protein, the degree could indicate a central role in amplification (kinases), diversification and turnover (small GTPases), signaling module assembly (docking proteins), gene expression (transcription factors), etc. Signaling networks have typically a scale-free architecture.

Betweenness

The S.-P. Betweenness of a node in a biological network, for instance a protein-signaling network, can indicate the relevance of a protein as functionally capable of holding together communicating proteins. The higher the value the higher the relevance of the protein as an organizing regulatory molecule. The S.-P. Betweenness of a protein effectively indicates the capability of a protein to bring in communication distant proteins. In signaling modules, proteins with high S.-P. Betweenness are likely crucial to maintain functionally and coherence of signaling mechanisms.

Centroid

The centrality value of a node in a biological network, for instance a protein-signaling network, can be interpreted as the probability of a protein to be functionally capable of organizing discrete protein clusters or modules. Thus, a protein with high centroid value, compared to the average centroid value of the network, will be possibly involved in coordinating the activity of other highly connected proteins, altogether devoted to the regulation of a specific cell activity (for instance, cell adhesion, gene expression, proliferation etc.). Accordingly, a signaling network with a very high average centroid value is more likely organizing functional units or modules, whereas a signaling network with very low average centroid value will behave more likely as an open cluster of proteins connecting different regulatory modules. It can be useful to compare the centroid value to other algorithms detecting dense regions in a graph, indicating protein clusters, such as, for instance, MCODE.

Eccentricity

The eccentricity of a node in a biological network, for instance a protein-signaling network, can be interpreted as the easiness of a protein to be functionally reached by all other proteins in the network. Thus, a protein with high eccentricity, compared to the average eccentricity of the network, will be more easily influenced by the activity of other proteins (the protein is subject to a more stringent or complex regulation) or, conversely could easily influence several other proteins. In contrast, a low eccentricity, compared to the average eccentricity of the network, could indicate a marginal functional role (although this should be also evaluated with other parameters and contextualized to the network annotations).

C.3 Topology and Centrality Measures for Four Networks

Here we present measures for the 8-variable random network, along with the networks for Aging, Yeast and Melanoma from Chapter 2. In the values, as discussed in Chapter 2, we observe some overlap in the reporting of our key variables. However we also notice three disadvantages: first, some irrelevant variables are elevated; second, some important variables are not elevated; third, for the network sizes reported it is difficult to establish a cutoff point as to which variables display standout behavior versus those which do not.

Table C.1: Centrality Measures for the 8-Variable Random Network discussed in §2.4.

Variable	Betweenness	Centroid	Eccentricity	Degree
g8	20	-1	0.50	4
g2	18	1	0.33	6
g3	12	-4	0.33	3
g1	6	-1	0.33	5
g4	0	-4	0.25	3
g7	0	-4	0.25	3
g5	0	-6	0.25	3
g6	0	-6	0.25	2

Table C.2: Centrality Measures for the Human Aging Network discussed in §2.5.

Variable	Betweenness	Centroid	Eccentricity	Degree
ros	158.95	8	0.50	12
h2o2	25.95	-12	0.33	4
lipofuscin	23.89	-8	0.33	5
n dna	23.78	-10	0.33	4
mt dna	21.86	-9	0.33	5
junk	15.75	-10	0.33	4
mt prot	14.67	-11	0.33	3
ions	13.60	-10	0.33	3
cell antx	13.33	-10	0.25	4
ly mem	10.83	-12	0.33	3
mito	6.83	-10	0.25	4
imm	3.69	-12	0.33	3
mt met	3.69	-14	0.33	2
apop	3.67	-11	0.25	3
lyso	2.00	-12	0.25	3
lon p	1.50	-13	0.25	2
ndna rep	0.00	-13	0.33	2
diet antx	0.00	-13	0.25	3

Table C.3: Centrality Measures for the Yeast Cell Cycle Network discussed in §2.5.

Variable	Betweenness	Centroid	Eccentricity	Degree
Clb1,2	28.58	2	0.50	13
Clb5,6	9.17	-2	0.50	8
SBF	8.92	-5	0.50	4
MBF	8.67	-6	0.33	4
Sic1	6.83	-4	0.33	8
Cdh1	4.08	-5	0.33	6
Cdc20	3.67	-3	0.33	8
Cln1,2	3.50	-5	0.33	4
Cln3	1.33	-6	0.33	3
Swi5	0.75	-5	0.33	5
Mcm1	0.50	-6	0.33	6

Table C.4: Centrality Measures for the Wnt5a Melanoma Network discussed in §2.5.

Variable	Betweenness	Centroid	Eccentricity	Degree
PIR	11.94	0	0.50	9
WNT5A	11.24	0	0.50	9
MLANA	9.11	-1	0.50	5
SNCA	4.90	-2	0.50	5
MMP3	4.29	-2	0.33	5
S100B	3.70	-2	0.50	7
STC2	2.90	-4	0.33	5
RET1	2.33	-3	0.33	6
PHO-C	1.92	-4	0.33	4
HADHB	1.67	-4	0.33	5

APPENDIX D

T-LGL LEUKEMIA NETWORK

This appendix provides additional information for the T-LGL Leukemia network described in Chapter 3, §3.5.

D.1 Network Simplification

The original network [54] created from the literature contained 128 nodes and 287 edges, but was simplified by software and manual adjustments to 60 nodes and 142 regulatory edges. In order to comply with current algorithmic implementations we were required to further reduce this network to have less than 53 variables; this was due to a dependency on storing information in a double-precision 64-bit variable only capable of storing unsigned integers up to $2^{53} - 1$. Our future work includes updating the software around this limitation, which is entirely possible with the use of variable precision integers.

After collaborating with Dr. Réka Albert, a principle author from [54], we performed the following reductions on the network according to techniques described in related work involving this network [56, 55]. The goal of the reduction is twofold. First, we wish to remove nodes which mask the dynamic behavior of the network variables (i.e. the overarching influence of Apoptosis as well as root nodes); since steady-state analysis will be performed over many randomly generated states, control variables are not necessary since variables they control will be forced to take on different values through random starting state assignment. Second we wish to reduce the size of the network to less than 53 nodes in order to work with the current algorithm implementations. After the simplifications described below we obtain the 43-variable version shown in Fig. D.1.

1. Delete all references to Apoptosis. Since it can turn off every cell it dominates the Boolean logic and focuses the study on the stationary states of the living cell.

2. Delete all root nodes: Stimuli, IL15, PDGF, CD45, TAX, Stimuli2 and P2. P2 becomes root after deleting Stimuli2.; These control nodes emerge as the dominant

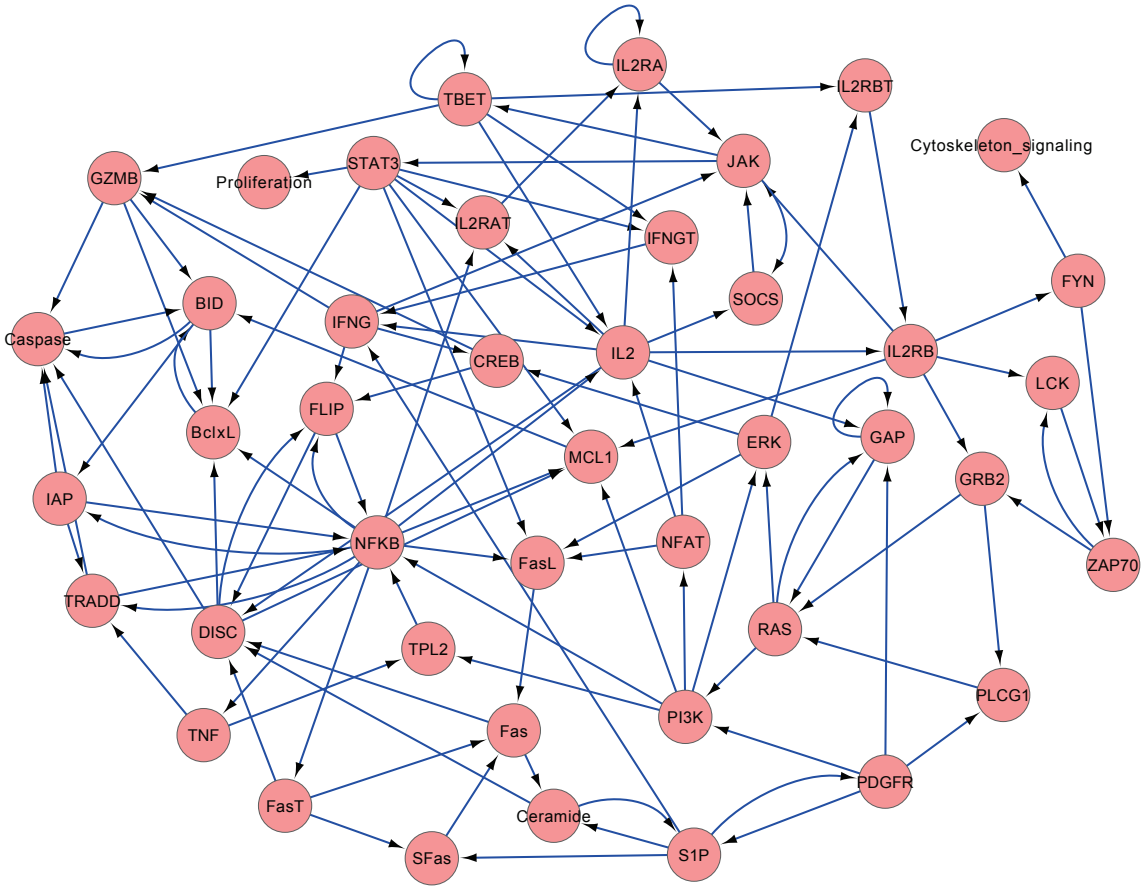


Figure D.1: 43-Variable T-LGL Leukemia Network after simplifications.

variables upon network analysis. Removing them reveals which OTHER variables are powerful and may be targets for intervention.

3. Delete node A20. A20 has only one input edge and one output edge forming a closed loop from NFKB to A20 to TRADD and back to NFKB. Replace A20 in the rule for TRADD with the same value of NFKB because the rule for A20 is A20 = NFKB.

4. Delete node CTLA4. CTLA4 has only one input and one output edge and forms a simple loop with TCR.

5. Delete TCR. After deleting CTLA4, and Stimuli, TCR has no inputs.

6. Delete GPCR and SMAD. The chain goes from S1P, to GPCR, to SMAD, to IFNG. GPCR and SMAD have one input and one output, so connect S1P to IFNG, replacing SMAD in the rule.

7. Delete MEK, which has only 1 input and 1 output. Connect RAS directly to ERK instead of through MEK.

8. Delete P27. Delete it from the rule in Proliferation as well. Proliferation already has input from STAT3, which is P27's parent.

9. Delete RANTES. Delete it from the rule in JAK as well. JAK already has input from NFKB, which is RANTES's parent.

10. Delete SPHK1. Connect PDGFR directly to S1P1 creating a direct loop instead of one interrupted by SPHK1.

D.2 Boolean Rules

Boolean rules after simplification are shown in Table D.4.

D.3 Attractors

Identified attractors are shown in Table D.4.

D.4 Top Templates Results

With reference to §3 and §3.5, here we summarize the top thirty most frequent intervention templates from each template size from each of the four main basins in the network. To estimate a success rate, we apply each top template intervention to all eleven attractor states across all basins in the T-LGL network and compute the distribution of basins reached. If a template causes a significant number of these attractor states to jump to (or remain in) a desired basin, such a template is of great interest. In general we expect an attractor state to remain robust to perturbation but expect the larger templates to have the best chance at changing the steady state of the system. Since Basins 1 and 4 each contain a 4-cycle

of attractor states, we expect at least 4 of the 11 destination states (36%) to remain in those basins (due to the expectation of robustness mentioned). Likewise the remaining basins each have a single attractor state each and we expect 1 of the 11 destination states (9%) to remain in these basins. Of the top 50 templates for each k and each basin, we indeed saw these expected distributions of goal states very frequently. Numbers above and beyond these expectations warrant closer inspection, which we provide for the most interesting templates listed in Table 3.4. Templates listed in Table 3.4 and discussed in §3.5 are highlighted with horizontal rule lines:

- Basin 1, 3-Templates, Rank 18, Table D.4
- Basin 2, 3-Templates, Rank 27, Table D.4
- Basin 3, 3-Templates, Rank 1, Table D.4
- Basin 4, 1-Templates, Rank 6, Table D.4

Table D.1: Boolean Rules for T-LGL Leukemia Network: After network simplification from Zhang *et al.*[54] based on techniques from Saadatpour *et al.*[56, 55], the following update rules were formed. The first column represents the next state of the listed variables and the second column refers to the current state of the variable(s). $\wedge$, $\vee$ and $\neg$ represent logical AND, OR and NOT, respectively.

Variable	Boolean Function
BID	(Caspase $\vee$ GZMB) $\wedge \neg$ (BclxL $\vee$ MCL1)
BclxL	(NFKB $\vee$ STAT3) $\wedge \neg$ (BID $\vee$ GZMB $\vee$ DISC)
CREB	(ERK $\wedge$ IFNG)
Caspase	((((TRADD $\vee$ GZMB) $\wedge$ BID) $\wedge \neg$ IAP) $\vee$ DISC)
Ceramide	Fas $\wedge \neg$ S1P
Cytoskeleton_signaling	FYN
DISC	(FasT $\wedge$ ((Fas $\wedge$ IL2) $\vee$ Ceramide $\vee$ (Fas $\wedge \neg$ FLIP)))
ERK	(RAS $\wedge$ PI3K)
FLIP	(NFKB $\vee$ (CREB $\wedge$ IFNG)) $\wedge \neg$ DISC
FYN	IL2RB
Fas	(FasT $\wedge$ FasL) $\wedge \neg$ SFas
FasL	(STAT3 $\vee$ NFKB $\vee$ NFAT $\vee$ ERK)
FasT	NFKB
GAP	(RAS $\vee$ (PDGFR $\wedge$ GAP)) $\wedge \neg$ IL2
GRB2	(IL2RB $\vee$ ZAP70)
GZMB	((CREB $\wedge$ IFNG) $\vee$ TBET)
IAP	NFKB $\wedge \neg$ BID
IFNG	(IL2 $\wedge$ IFNGT) $\wedge \neg$ S1P
IFNGT	(TBET $\vee$ STAT3 $\vee$ NFAT)
IL2	(NFKB $\vee$ STAT3 $\vee$ NFAT) $\wedge \neg$ TBET
IL2RA	(IL2 $\wedge$ IL2RAT) $\wedge \neg$ IL2RA
IL2RAT	(IL2 $\wedge$ (STAT3 $\vee$ NFKB))
IL2RB	IL2RBT $\wedge$ IL2
IL2RBT	(ERK $\wedge$ TBET)
JAK	(IL2RA $\vee$ IL2RB $\vee$ IFNG) $\wedge \neg$ SOCS
LCK	IL2RB $\wedge \neg$ ZAP70
MCL1	MCL1 = (IL2RB $\wedge$ STAT3 $\wedge$ NFKB $\wedge$ PI3K) $\wedge \neg$ DISC
NFAT	PI3K
NFKB	((TPL2 $\vee$ PI3K) $\vee$ (FLIP $\wedge$ TRADD $\wedge$ IAP))
PDGFR	S1P
PI3K	(PDGFR $\vee$ RAS)
PLCG1	(GRB2 $\vee$ PDGFR)
Proliferation	STAT3
RAS	(GRB2 $\vee$ PLCG1) $\wedge \neg$ GAP
S1P	PDGFR $\wedge \neg$ Ceramide
SFas	FasT $\wedge$ S1P
SOCS	JAK $\wedge \neg$ IL2
STAT3	JAK
TBET	(JAK $\vee$ TBET)
TNF	NFKB
TPL2	PI3K $\wedge$ TNF
TRADD	TNF $\wedge \neg$ (IAP $\vee$ NFKB)
ZAP70	LCK $\wedge \neg$ FYN

Table D.2: Attractor States for T-LGL Leukemia Network: The four basins are classified as either Normal (i.e. healthy, normal apoptosis function) or T-LGL (i.e. cancer state) based on the values of certain key variables in the steady attractor states/cycles, as described in §3.5.

Variable	Basin 1				Basin 4				Basin 2	Basin 3
BID	1	1	1	1	1	1	1	1	1	1
BclxL	0	0	0	0	0	0	0	0	0	0
CREB	0	0	0	0	0	0	0	0	0	0
Caspase	1	1	1	1	1	1	1	1	1	1
Ceramide	0	1	0	1	1	0	1	0	0	0
Cytoskeleton_signaling	0	0	0	0	0	0	0	0	0	0
DISC	1	0	1	0	1	1	1	1	0	0
ERK	0	0	0	0	0	0	0	0	0	0
FLIP	1	0	1	0	0	0	0	0	0	1
FYN	0	0	0	0	0	0	0	0	0	0
Fas	1	0	1	0	0	1	0	1	0	0
FasL	1	1	1	1	1	1	1	1	0	1
FasT	1	1	1	1	1	1	1	1	0	1
GAP	0	1	0	0	0	0	1	0	0	1
GRB2	0	0	0	0	0	0	0	0	0	0
GZMB	1	1	1	1	1	1	1	1	1	1
IAP	0	0	0	0	0	0	0	0	0	0
IFNG	0	0	0	0	0	0	0	0	0	0
IFNGT	1	1	1	1	1	1	1	1	1	1
IL2	0	0	0	0	0	0	0	0	0	0
IL2RA	0	0	0	0	0	0	0	0	0	0
IL2RAT	0	0	0	0	0	0	0	0	0	0
IL2RB	0	0	0	0	0	0	0	0	0	0
IL2RBT	0	0	0	0	0	0	0	0	0	0
JAK	0	0	0	0	0	0	0	0	0	0
LCK	0	0	0	0	0	0	0	0	0	0
MCL1	0	0	0	0	0	0	0	0	0	0
NFAT	1	0	1	0	0	1	0	1	0	1
NFKB	1	1	1	1	1	1	1	1	0	1
PDGFR	1	0	1	0	0	1	0	1	0	1
PI3K	0	1	0	1	1	0	1	0	0	1
PLCG1	0	1	0	1	1	0	1	0	0	1
Proliferation	0	0	0	0	0	0	0	0	0	0
RAS	1	0	0	0	0	1	0	0	0	0
S1P	0	1	0	1	1	0	1	0	0	1
SFas	1	0	1	0	0	1	0	1	0	1
SOCS	0	0	0	0	0	0	0	0	0	0
STAT3	0	0	0	0	0	0	0	0	0	0
TBET	1	1	1	1	1	1	1	1	1	1
TNF	1	1	1	1	1	1	1	1	0	1
TPL2	1	0	1	0	0	1	0	1	0	1
TRADD	0	0	0	0	0	0	0	0	0	0
ZAP70	0	0	0	0	0	0	0	0	0	0

Table D.3: Top 30 1-Templates for Leukemia Network Basin 1

RANK	Variable	Value	FREQ	SUCCESS	B1	B2	B3	B4	B5
1	MCL1	0	37204	0.36	0.36	0.09	0.09	0.36	0.09
2	FasL	1	35721	0.36	0.36	0.09	0.09	0.36	0.09
3	TBET	1	35405	0.36	0.36	0.18	0.09	0.36	0.00
4	IFNGT	1	34977	0.36	0.36	0.09	0.09	0.36	0.09
5	BclxL	0	34861	0.36	0.36	0.09	0.09	0.36	0.09
6	Ceramide	0	34861	0.36	0.36	0.09	0.09	0.36	0.09
7	TRADD	0	34840	0.36	0.36	0.09	0.09	0.36	0.09
8	IFNG	0	34792	0.36	0.36	0.09	0.09	0.36	0.09
9	IL2RA	0	33928	0.36	0.36	0.09	0.09	0.36	0.09
10	Fas	0	33697	0.55	0.55	0.09	0.09	0.18	0.09
11	ZAP70	0	33070	0.36	0.36	0.09	0.09	0.36	0.09
12	IL2RB	0	32936	0.36	0.36	0.09	0.09	0.36	0.09
13	CREB	0	32750	0.36	0.36	0.09	0.09	0.36	0.09
14	PLCG1	1	31966	0.36	0.36	0.18	0.09	0.36	0.00
15	LCK	0	31537	0.36	0.36	0.09	0.09	0.36	0.09
16	SOCS	0	31438	0.36	0.36	0.09	0.09	0.36	0.09
17	NFKB	1	31197	0.36	0.36	0.09	0.09	0.36	0.09
18	ERK	0	30968	0.36	0.36	0.09	0.09	0.36	0.09
19	TPL2	0	30842	0.55	0.55	0.09	0.09	0.18	0.09
20	DISC	0	30814	0.55	0.55	0.09	0.09	0.18	0.09
21	PI3K	1	30798	0.36	0.36	0.18	0.09	0.36	0.00
22	IL2RBT	0	30794	0.36	0.36	0.09	0.09	0.36	0.09
23	GAP	0	30120	0.36	0.36	0.09	0.09	0.36	0.09
24	GZMB	1	29936	0.36	0.36	0.09	0.09	0.36	0.09
25	IAP	0	29855	0.36	0.36	0.09	0.09	0.36	0.09
26	IL2	0	29322	0.36	0.36	0.09	0.09	0.36	0.09
27	GRB2	1	29180	0.36	0.36	0.18	0.09	0.36	0.00
28	SFas	0	29023	0.36	0.36	0.09	0.09	0.36	0.09
29	BID	0	28779	0.36	0.36	0.09	0.09	0.36	0.09
30	PDGFR	1	28268	0.45	0.45	0.00	0.09	0.45	0.00

Table D.4: Top 30 1-Templates for Leukemia Network Basin 2

RANK	Variable	Value	FREQ	SUCCESS	B1	B2	B3	B4	B5
1	S1P	0	46989	0.45	0.27	0.45	0.00	0.18	0.09
2	PDGFR	0	40833	0.45	0.27	0.45	0.00	0.18	0.09
3	SFas	0	37512	0.09	0.36	0.09	0.09	0.36	0.09
4	MCL1	0	36675	0.09	0.36	0.09	0.09	0.36	0.09
5	FasL	1	35071	0.09	0.36	0.09	0.09	0.36	0.09
6	TRADD	0	34577	0.09	0.36	0.09	0.09	0.36	0.09
7	BclxL	0	33922	0.09	0.36	0.09	0.09	0.36	0.09
8	IFNGT	1	33405	0.09	0.36	0.09	0.09	0.36	0.09
9	ZAP70	0	33168	0.09	0.36	0.09	0.09	0.36	0.09
10	IL2RB	0	32959	0.09	0.36	0.09	0.09	0.36	0.09
11	IL2RA	0	32580	0.09	0.36	0.09	0.09	0.36	0.09
12	CREB	0	31818	0.09	0.36	0.09	0.09	0.36	0.09
13	IL2RBT	0	31236	0.09	0.36	0.09	0.09	0.36	0.09
14	SOCS	0	31202	0.09	0.36	0.09	0.09	0.36	0.09
15	TBET	1	31043	0.18	0.36	0.18	0.09	0.36	0.00
16	LCK	0	30996	0.09	0.36	0.09	0.09	0.36	0.09
17	NFKB	1	30689	0.09	0.36	0.09	0.09	0.36	0.09
18	GAP	0	30321	0.09	0.36	0.09	0.09	0.36	0.09
19	ERK	0	30080	0.09	0.36	0.09	0.09	0.36	0.09
20	TPL2	0	29021	0.09	0.55	0.09	0.09	0.18	0.09
21	Fas	0	28701	0.09	0.55	0.09	0.09	0.18	0.09
22	IFNG	0	28414	0.09	0.36	0.09	0.09	0.36	0.09
23	IAP	0	27628	0.09	0.36	0.09	0.09	0.36	0.09
24	GZMB	1	27597	0.09	0.36	0.09	0.09	0.36	0.09
25	PLCG1	1	27306	0.18	0.36	0.18	0.09	0.36	0.00
26	FLIP	0	27280	0.09	0.18	0.09	0.09	0.55	0.09
27	IL2	0	27011	0.09	0.36	0.09	0.09	0.36	0.09
28	GRB2	1	26652	0.18	0.36	0.18	0.09	0.36	0.00
29	BID	0	26610	0.09	0.36	0.09	0.09	0.36	0.09
30	DISC	0	26371	0.09	0.55	0.09	0.09	0.18	0.09

Table D.5: Top 30 1-Templates for Leukemia Network Basin 3

RANK	Variable	Value	FREQ	SUCCESS	B1	B2	B3	B4	B5
1	Ceramide	0	14404	0.09	0.36	0.09	0.09	0.36	0.09
2	PDGFR	1	14404	0.09	0.45	0.00	0.09	0.45	0.00
3	S1P	1	14404	0.45	0.27	0.00	0.45	0.27	0.00
4	IFNG	0	11323	0.09	0.36	0.09	0.09	0.36	0.09
5	PLCG1	1	11278	0.09	0.36	0.18	0.09	0.36	0.00
6	PI3K	1	11256	0.09	0.36	0.18	0.09	0.36	0.00
7	MCL1	0	11018	0.09	0.36	0.09	0.09	0.36	0.09
8	FasL	1	10664	0.09	0.36	0.09	0.09	0.36	0.09
9	TRADD	0	10376	0.09	0.36	0.09	0.09	0.36	0.09
10	BclxL	0	10289	0.09	0.36	0.09	0.09	0.36	0.09
11	IFNGT	1	10043	0.09	0.36	0.09	0.09	0.36	0.09
12	IL2RA	0	9994	0.09	0.36	0.09	0.09	0.36	0.09
13	DISC	0	9927	0.09	0.55	0.09	0.09	0.18	0.09
14	NFKB	1	9911	0.09	0.36	0.09	0.09	0.36	0.09
15	Fas	0	9867	0.09	0.55	0.09	0.09	0.18	0.09
16	ZAP70	0	9785	0.09	0.36	0.09	0.09	0.36	0.09
17	IL2RB	0	9748	0.09	0.36	0.09	0.09	0.36	0.09
18	CREB	0	9733	0.09	0.36	0.09	0.09	0.36	0.09
19	TBET	1	9389	0.09	0.36	0.18	0.09	0.36	0.00
20	SOCS	0	9286	0.09	0.36	0.09	0.09	0.36	0.09
21	LCK	0	9270	0.09	0.36	0.09	0.09	0.36	0.09
22	IL2RBT	0	9187	0.09	0.36	0.09	0.09	0.36	0.09
23	ERK	0	8662	0.09	0.36	0.09	0.09	0.36	0.09
24	IAP	0	8640	0.09	0.36	0.09	0.09	0.36	0.09
25	GRB2	1	8483	0.09	0.36	0.18	0.09	0.36	0.00
26	GAP	0	8439	0.09	0.36	0.09	0.09	0.36	0.09
27	TPL2	0	8436	0.09	0.55	0.09	0.09	0.18	0.09
28	GZMB	1	8355	0.09	0.36	0.09	0.09	0.36	0.09
29	BID	0	8341	0.09	0.36	0.09	0.09	0.36	0.09
30	NFAT	1	8244	0.09	0.36	0.18	0.09	0.36	0.00

Table D.6: Top 30 1-Templates for Leukemia Network Basin 4

RANK	Variable	Value	FREQ	SUCCESS	B1	B2	B3	B4	B5
1	NFKB	1	7260	0.36	0.36	0.09	0.09	0.36	0.09
2	FasL	1	6735	0.36	0.36	0.09	0.09	0.36	0.09
3	MCL1	0	6594	0.36	0.36	0.09	0.09	0.36	0.09
4	PI3K	1	6428	0.36	0.36	0.18	0.09	0.36	0.00
5	TRADD	0	6417	0.36	0.36	0.09	0.09	0.36	0.09
6	TBET	0	6404	0.73	0.00	0.00	0.09	0.73	0.18
7	IL2RBT	0	6342	0.36	0.36	0.09	0.09	0.36	0.09
8	JAK	0	6318	0.36	0.36	0.09	0.09	0.36	0.09
9	FasT	1	6212	0.36	0.36	0.09	0.09	0.36	0.09
10	IL2RB	0	6091	0.36	0.36	0.09	0.09	0.36	0.09
11	BclxL	0	5955	0.36	0.36	0.09	0.09	0.36	0.09
12	IFNG	0	5941	0.36	0.36	0.09	0.09	0.36	0.09
13	ZAP70	0	5928	0.36	0.36	0.09	0.09	0.36	0.09
14	TNF	1	5925	0.36	0.36	0.09	0.09	0.36	0.09
15	CREB	0	5888	0.36	0.36	0.09	0.09	0.36	0.09
16	PLCG1	1	5845	0.36	0.36	0.18	0.09	0.36	0.00
17	SOCS	0	5758	0.36	0.36	0.09	0.09	0.36	0.09
18	STAT3	0	5756	0.36	0.36	0.09	0.09	0.36	0.09
19	IL2RA	0	5737	0.36	0.36	0.09	0.09	0.36	0.09
20	LCK	0	5720	0.36	0.36	0.09	0.09	0.36	0.09
21	Ceramide	0	5622	0.36	0.36	0.09	0.09	0.36	0.09
22	GAP	0	5474	0.36	0.36	0.09	0.09	0.36	0.09
23	IFNGT	1	5397	0.36	0.36	0.09	0.09	0.36	0.09
24	IL2	1	5356	0.55	0.18	0.09	0.09	0.55	0.09
25	NFAT	1	5253	0.36	0.36	0.18	0.09	0.36	0.00
26	GZMB	0	5101	0.36	0.36	0.09	0.09	0.36	0.09
27	SFas	0	4965	0.36	0.36	0.09	0.09	0.36	0.09
28	PDGFR	1	4898	0.45	0.45	0.00	0.09	0.45	0.00
29	Proliferation	0	4849	0.36	0.36	0.09	0.09	0.36	0.09
30	BID	0	4840	0.36	0.36	0.09	0.09	0.36	0.09

Table D.7: Top 30 2-Templates for Leukemia Network Basin 1

RANK	Variables		Values		FREQ	SUCCESS	B1	B2	B3	B4	B5
1	FasL	MCL1	1	0	29352	0.36	0.36	0.09	0.09	0.36	0.09
2	BclxL	MCL1	0	0	28755	0.36	0.36	0.09	0.09	0.36	0.09
3	IFNG	MCL1	0	0	28511	0.36	0.36	0.09	0.09	0.36	0.09
4	Ceramide	MCL1	0	0	28485	0.36	0.36	0.09	0.09	0.36	0.09
5	IFNGT	MCL1	1	0	28412	0.36	0.36	0.09	0.09	0.36	0.09
6	MCL1	TBET	0	1	28410	0.36	0.36	0.18	0.09	0.36	0.00
7	MCL1	TRADD	0	0	28348	0.36	0.36	0.09	0.09	0.36	0.09
8	FasL	IFNGT	1	1	28311	0.36	0.36	0.09	0.09	0.36	0.09
9	IL2RA	MCL1	0	0	27759	0.36	0.36	0.09	0.09	0.36	0.09
10	FasL	TRADD	1	0	27617	0.36	0.36	0.09	0.09	0.36	0.09
11	FasL	TBET	1	1	27444	0.36	0.36	0.18	0.09	0.36	0.00
12	Ceramide	FasL	0	1	27420	0.36	0.36	0.09	0.09	0.36	0.09
13	Fas	MCL1	0	0	27402	0.55	0.55	0.09	0.09	0.18	0.09
14	FasL	IFNG	1	0	27354	0.36	0.36	0.09	0.09	0.36	0.09
15	IFNGT	TBET	1	1	27268	0.36	0.36	0.18	0.09	0.36	0.00
16	Ceramide	IFNG	0	0	27028	0.36	0.36	0.09	0.09	0.36	0.09
17	BclxL	FasL	0	1	26988	0.36	0.36	0.09	0.09	0.36	0.09
18	IFNG	IFNGT	0	1	26865	0.36	0.36	0.09	0.09	0.36	0.09
19	MCL1	ZAP70	0	0	26861	0.36	0.36	0.09	0.09	0.36	0.09
20	BclxL	IFNG	0	0	26845	0.36	0.36	0.09	0.09	0.36	0.09
21	IL2RB	MCL1	0	0	26696	0.36	0.36	0.09	0.09	0.36	0.09
22	IFNGT	TRADD	1	0	26678	0.36	0.36	0.09	0.09	0.36	0.09
23	TBET	TRADD	1	0	26619	0.36	0.36	0.18	0.09	0.36	0.00
24	Ceramide	TRADD	0	0	26591	0.36	0.36	0.09	0.09	0.36	0.09
25	Ceramide	TBET	0	1	26568	0.36	0.36	0.18	0.09	0.36	0.00
26	FasL	IL2RA	1	0	26532	0.36	0.36	0.09	0.09	0.36	0.09
27	IFNG	TBET	0	1	26530	0.36	0.36	0.18	0.09	0.36	0.00
28	Ceramide	IFNGT	0	1	26512	0.36	0.36	0.09	0.09	0.36	0.09
29	CREB	MCL1	0	0	26439	0.36	0.36	0.09	0.09	0.36	0.09
30	IFNG	TRADD	0	0	26428	0.36	0.36	0.09	0.09	0.36	0.09

Table D.8: Top 30 2-Templates for Leukemia Network Basin 2

RANK	Variables		Values		FREQ	SUCCESS	B1	B2	B3	B4	B5
1	PDGFR	S1P	0	0	40833	0.91	0.00	0.91	0.00	0.00	0.09
2	S1P	SFas	0	0	37512	0.45	0.27	0.45	0.00	0.18	0.09
3	MCL1	S1P	0	0	36675	0.45	0.27	0.45	0.00	0.18	0.09
4	FasL	S1P	1	0	35071	0.45	0.27	0.45	0.00	0.18	0.09
5	S1P	TRADD	0	0	34577	0.45	0.27	0.45	0.00	0.18	0.09
6	PDGFR	SFas	0	0	34475	0.45	0.27	0.45	0.00	0.18	0.09
7	BclxL	S1P	0	0	33922	0.45	0.27	0.45	0.00	0.18	0.09
8	MCL1	PDGFR	0	0	33606	0.45	0.27	0.45	0.00	0.18	0.09
9	IFNGT	S1P	1	0	33405	0.45	0.27	0.45	0.00	0.18	0.09
10	S1P	ZAP70	0	0	33168	0.45	0.27	0.45	0.00	0.18	0.09
11	IL2RB	S1P	0	0	32959	0.45	0.27	0.45	0.00	0.18	0.09
12	IL2RA	S1P	0	0	32580	0.45	0.27	0.45	0.00	0.18	0.09
13	FasL	PDGFR	1	0	31990	0.45	0.27	0.45	0.00	0.18	0.09
14	MCL1	SFas	0	0	31931	0.09	0.36	0.09	0.09	0.36	0.09
15	CREB	S1P	0	0	31818	0.45	0.27	0.45	0.00	0.18	0.09
16	PDGFR	TRADD	0	0	31488	0.45	0.27	0.45	0.00	0.18	0.09
17	IL2RBT	S1P	0	0	31236	0.45	0.27	0.45	0.00	0.18	0.09
18	S1P	SOCS	0	0	31202	0.45	0.27	0.45	0.00	0.18	0.09
19	S1P	TBET	0	1	31043	0.55	0.27	0.55	0.00	0.18	0.00
20	LCK	S1P	0	0	30996	0.45	0.27	0.45	0.00	0.18	0.09
21	BclxL	PDGFR	0	0	30821	0.45	0.27	0.45	0.00	0.18	0.09
22	NFKB	S1P	1	0	30689	0.45	0.27	0.45	0.00	0.18	0.09
23	FasL	SFas	1	0	30382	0.09	0.36	0.09	0.09	0.36	0.09
24	IFNGT	PDGFR	1	0	30344	0.45	0.27	0.45	0.00	0.18	0.09
25	GAP	S1P	0	0	30321	0.45	0.27	0.45	0.00	0.18	0.09
26	ERK	S1P	0	0	30080	0.45	0.27	0.45	0.00	0.18	0.09
27	PDGFR	ZAP70	0	0	30062	0.45	0.27	0.45	0.00	0.18	0.09
28	IL2RB	PDGFR	0	0	29864	0.45	0.27	0.45	0.00	0.18	0.09
29	SFas	TRADD	0	0	29762	0.09	0.36	0.09	0.09	0.36	0.09
30	FasL	MCL1	1	0	29544	0.09	0.36	0.09	0.09	0.36	0.09

Table D.9: Top 30 2-Templates for Leukemia Network Basin 3

RANK	Variables		Values		FREQ	SUCCESS	B1	B2	B3	B4	B5
1	Ceramide	PDGFR	0	1	14404	0.45	0.27	0.00	0.45	0.27	0.00
2	Ceramide	S1P	0	1	14404	0.45	0.27	0.00	0.45	0.27	0.00
3	PDGFR	S1P	1	1	14404	0.64	0.18	0.00	0.64	0.18	0.00
4	Ceramide	IFNG	0	0	11323	0.09	0.36	0.09	0.09	0.36	0.09
5	IFNG	PDGFR	0	1	11323	0.09	0.45	0.00	0.09	0.45	0.00
6	IFNG	S1P	0	1	11323	0.45	0.27	0.00	0.45	0.27	0.00
7	Ceramide	PLCG1	0	1	11278	0.09	0.36	0.18	0.09	0.36	0.00
8	PDGFR	PLCG1	1	1	11278	0.09	0.45	0.00	0.09	0.45	0.00
9	PLCG1	S1P	1	1	11278	0.45	0.27	0.00	0.45	0.27	0.00
10	Ceramide	PI3K	0	1	11256	0.09	0.36	0.18	0.09	0.36	0.00
11	PDGFR	PI3K	1	1	11256	0.09	0.55	0.00	0.09	0.36	0.00
12	PI3K	S1P	1	1	11256	0.45	0.27	0.00	0.45	0.27	0.00
13	Ceramide	MCL1	0	0	11018	0.09	0.36	0.09	0.09	0.36	0.09
14	MCL1	PDGFR	0	1	11018	0.09	0.45	0.00	0.09	0.45	0.00
15	MCL1	S1P	0	1	11018	0.45	0.27	0.00	0.45	0.27	0.00
16	Ceramide	FasL	0	1	10664	0.09	0.36	0.09	0.09	0.36	0.09
17	FasL	PDGFR	1	1	10664	0.09	0.45	0.00	0.09	0.45	0.00
18	FasL	S1P	1	1	10664	0.45	0.27	0.00	0.45	0.27	0.00
19	Ceramide	TRADD	0	0	10376	0.09	0.36	0.09	0.09	0.36	0.09
20	PDGFR	TRADD	1	0	10376	0.09	0.45	0.00	0.09	0.45	0.00
21	S1P	TRADD	1	0	10376	0.45	0.27	0.00	0.45	0.27	0.00
22	BclxL	Ceramide	0	0	10289	0.09	0.36	0.09	0.09	0.36	0.09
23	BclxL	PDGFR	0	1	10289	0.09	0.45	0.00	0.09	0.45	0.00
24	BclxL	S1P	0	1	10289	0.45	0.27	0.00	0.45	0.27	0.00
25	Ceramide	IFNGT	0	1	10043	0.09	0.36	0.09	0.09	0.36	0.09
26	IFNGT	PDGFR	1	1	10043	0.09	0.45	0.00	0.09	0.45	0.00
27	IFNGT	S1P	1	1	10043	0.45	0.27	0.00	0.45	0.27	0.00
28	Ceramide	IL2RA	0	0	9994	0.09	0.36	0.09	0.09	0.36	0.09
29	IL2RA	PDGFR	0	1	9994	0.09	0.45	0.00	0.09	0.45	0.00
30	IL2RA	S1P	0	1	9994	0.45	0.27	0.00	0.45	0.27	0.00

Table D.10: Top 30 2-Templates for Leukemia Network Basin 4

RANK	Variables		Values		FREQ	SUCCESS	B1	B2	B3	B4	B5
1	FasL	NFKB	1	1	6105	0.36	0.36	0.09	0.09	0.36	0.09
2	MCL1	NFKB	0	1	5885	0.36	0.36	0.09	0.09	0.36	0.09
3	NFKB	TRADD	1	0	5766	0.36	0.36	0.09	0.09	0.36	0.09
4	NFKB	PI3K	1	1	5653	0.36	0.36	0.18	0.09	0.36	0.00
5	IL2RBT	NFKB	0	1	5625	0.36	0.36	0.09	0.09	0.36	0.09
6	FasT	NFKB	1	1	5618	0.36	0.36	0.09	0.09	0.36	0.09
7	FasL	MCL1	1	0	5617	0.36	0.36	0.09	0.09	0.36	0.09
8	FasL	TRADD	1	0	5509	0.36	0.36	0.09	0.09	0.36	0.09
9	MCL1	TRADD	0	0	5463	0.36	0.36	0.09	0.09	0.36	0.09
10	NFKB	SOCS	1	0	5454	0.36	0.36	0.09	0.09	0.36	0.09
11	FasL	FasT	1	1	5391	0.36	0.36	0.09	0.09	0.36	0.09
12	FasL	IL2RBT	1	0	5367	0.36	0.36	0.09	0.09	0.36	0.09
13	BclxL	NFKB	0	1	5336	0.36	0.36	0.09	0.09	0.36	0.09
14	IL2RB	NFKB	0	1	5333	0.36	0.36	0.09	0.09	0.36	0.09
15	FasL	PI3K	1	1	5326	0.36	0.36	0.18	0.09	0.36	0.00
16	IL2RBT	MCL1	0	0	5307	0.36	0.36	0.09	0.09	0.36	0.09
17	JAK	TBET	0	0	5306	0.73	0.00	0.00	0.09	0.73	0.18
18	NFKB	ZAP70	1	0	5282	0.36	0.36	0.09	0.09	0.36	0.09
19	NFKB	TNF	1	1	5258	0.36	0.36	0.09	0.09	0.36	0.09
20	FasT	TNF	1	1	5247	0.36	0.36	0.09	0.09	0.36	0.09
21	NFKB	TBET	1	0	5230	0.73	0.00	0.00	0.09	0.73	0.18
22	CREB	NFKB	0	1	5221	0.36	0.36	0.09	0.09	0.36	0.09
23	MCL1	PI3K	0	1	5201	0.36	0.36	0.18	0.09	0.36	0.00
24	IFNG	NFKB	0	1	5179	0.36	0.36	0.09	0.09	0.36	0.09
25	NFKB	PLCG1	1	1	5168	0.36	0.36	0.18	0.09	0.36	0.00
26	FasL	SOCS	1	0	5155	0.36	0.36	0.09	0.09	0.36	0.09
27	FasL	TNF	1	1	5149	0.36	0.36	0.09	0.09	0.36	0.09
28	JAK	NFKB	0	1	5144	0.36	0.36	0.09	0.09	0.36	0.09
29	IL2RBT	TRADD	0	0	5143	0.36	0.36	0.09	0.09	0.36	0.09
30	FasT	TRADD	1	0	5139	0.36	0.36	0.09	0.09	0.36	0.09

Table D.11: Top 30 3-Templates for Leukemia Network Basin 1

RANK	Variables			Values			FREQ	SUCCESS	B1	B2	B3	B4	B5
1	FasL	IFNGT	MCL1	1	1	0	24553	0.36	0.36	0.09	0.09	0.36	0.09
2	FasL	MCL1	TRADD	1	0	0	23890	0.36	0.36	0.09	0.09	0.36	0.09
3	FasL	IFNG	MCL1	1	0	0	23812	0.36	0.36	0.09	0.09	0.36	0.09
4	Ceramide	FasL	MCL1	0	1	0	23772	0.36	0.36	0.09	0.09	0.36	0.09
5	BclxL	FasL	MCL1	0	1	0	23607	0.36	0.36	0.09	0.09	0.36	0.09
6	Ceramide	IFNG	MCL1	0	0	0	23494	0.36	0.36	0.09	0.09	0.36	0.09
7	FasL	MCL1	TBET	1	0	1	23484	0.36	0.36	0.18	0.09	0.36	0.00
8	BclxL	IFNG	MCL1	0	0	0	23482	0.36	0.36	0.09	0.09	0.36	0.09
9	IFNGT	MCL1	TBET	1	0	1	23257	0.36	0.36	0.18	0.09	0.36	0.00
10	IFNG	IFNGT	MCL1	0	1	0	23234	0.36	0.36	0.09	0.09	0.36	0.09
11	FasL	IL2RA	MCL1	1	0	0	23113	0.36	0.36	0.09	0.09	0.36	0.09
12	FasL	IFNGT	TBET	1	1	1	23044	0.36	0.36	0.18	0.09	0.36	0.00
13	FasL	IFNGT	TRADD	1	1	0	23018	0.36	0.36	0.09	0.09	0.36	0.09
14	BclxL	Ceramide	MCL1	0	0	0	22998	0.36	0.36	0.09	0.09	0.36	0.09
15	FasL	IFNG	IFNGT	1	0	1	22970	0.36	0.36	0.09	0.09	0.36	0.09
16	IFNG	IL2RA	MCL1	0	0	0	22950	0.36	0.36	0.09	0.09	0.36	0.09
17	BclxL	IFNGT	MCL1	0	1	0	22937	0.36	0.36	0.09	0.09	0.36	0.09
18	Ceramide	PDGFR	S1P	0	1	0	22936	0.73	0.73	0.00	0.00	0.27	0.00
19	Fas	FasL	MCL1	0	1	0	22860	0.55	0.55	0.09	0.09	0.18	0.09
20	Ceramide	FasL	IFNGT	0	1	1	22851	0.36	0.36	0.09	0.09	0.36	0.09
21	Ceramide	MCL1	TRADD	0	0	0	22838	0.36	0.36	0.09	0.09	0.36	0.09
22	IFNG	MCL1	TRADD	0	0	0	22825	0.36	0.36	0.09	0.09	0.36	0.09
23	Ceramide	IFNGT	MCL1	0	1	0	22820	0.36	0.36	0.09	0.09	0.36	0.09
24	BclxL	MCL1	TRADD	0	0	0	22809	0.36	0.36	0.09	0.09	0.36	0.09
25	IFNGT	MCL1	TRADD	1	0	0	22784	0.36	0.36	0.09	0.09	0.36	0.09
26	BclxL	MCL1	TBET	0	0	1	22735	0.36	0.36	0.18	0.09	0.36	0.00
27	IFNG	MCL1	TBET	0	0	1	22684	0.36	0.36	0.18	0.09	0.36	0.00
28	BclxL	IL2RA	MCL1	0	0	0	22651	0.36	0.36	0.09	0.09	0.36	0.09
29	Ceramide	MCL1	TBET	0	0	1	22631	0.36	0.36	0.18	0.09	0.36	0.00
30	Ceramide	FasL	IFNG	0	1	0	22599	0.36	0.36	0.09	0.09	0.36	0.09

Table D.12: Top 30 3-Templates for Leukemia Network Basin 2

RANK	Variables			Values			FREQ	SUCCESS	B1	B2	B3	B4	B5
1	PDGFR	S1P	SFas	0	0	0	34475	0.91	0.00	0.91	0.00	0.00	0.09
2	MCL1	PDGFR	S1P	0	0	0	33606	0.91	0.00	0.91	0.00	0.00	0.09
3	FasL	PDGFR	S1P	1	0	0	31990	0.91	0.00	0.91	0.00	0.00	0.09
4	MCL1	S1P	SFas	0	0	0	31931	0.45	0.27	0.45	0.00	0.18	0.09
5	PDGFR	S1P	TRADD	0	0	0	31488	0.91	0.00	0.91	0.00	0.00	0.09
6	BclxL	PDGFR	S1P	0	0	0	30821	0.91	0.00	0.91	0.00	0.00	0.09
7	MCL1	PDGFR	SFas	0	0	0	30409	0.45	0.27	0.45	0.00	0.18	0.09
8	FasL	S1P	SFas	1	0	0	30382	0.45	0.27	0.45	0.00	0.18	0.09
9	IFNGT	PDGFR	S1P	1	0	0	30344	0.91	0.00	0.91	0.00	0.00	0.09
10	PDGFR	S1P	ZAP70	0	0	0	30062	0.91	0.00	0.91	0.00	0.00	0.09
11	IL2RB	PDGFR	S1P	0	0	0	29864	0.91	0.00	0.91	0.00	0.00	0.09
12	S1P	SFas	TRADD	0	0	0	29762	0.45	0.27	0.45	0.00	0.18	0.09
13	FasL	MCL1	S1P	1	0	0	29544	0.45	0.27	0.45	0.00	0.18	0.09
14	IL2RA	PDGFR	S1P	0	0	0	29501	0.91	0.00	0.91	0.00	0.00	0.09
15	BclxL	S1P	SFas	0	0	0	29136	0.45	0.27	0.45	0.00	0.18	0.09
16	MCL1	S1P	TRADD	0	0	0	28953	0.45	0.27	0.45	0.00	0.18	0.09
17	FasL	PDGFR	SFas	1	0	0	28859	0.45	0.27	0.45	0.00	0.18	0.09
18	CREB	PDGFR	S1P	0	0	0	28723	0.91	0.00	0.91	0.00	0.00	0.09
19	IFNGT	S1P	SFas	1	0	0	28720	0.45	0.27	0.45	0.00	0.18	0.09
20	BclxL	MCL1	S1P	0	0	0	28564	0.45	0.27	0.45	0.00	0.18	0.09
21	S1P	SFas	ZAP70	0	0	0	28382	0.45	0.27	0.45	0.00	0.18	0.09
22	PDGFR	SFas	TRADD	0	0	0	28250	0.45	0.27	0.45	0.00	0.18	0.09
23	IL2RB	S1P	SFas	0	0	0	28227	0.45	0.27	0.45	0.00	0.18	0.09
24	IL2RBT	PDGFR	S1P	0	0	0	28104	0.91	0.00	0.91	0.00	0.00	0.09
25	FasL	S1P	TRADD	1	0	0	28090	0.45	0.27	0.45	0.00	0.18	0.09
26	PDGFR	S1P	SOCS	0	0	0	28072	0.91	0.00	0.91	0.00	0.00	0.09
27	PDGFR	S1P	TBET	0	0	1	28057	1.00	0.00	1.00	0.00	0.00	0.00
28	FasL	MCL1	PDGFR	1	0	0	28007	0.45	0.27	0.45	0.00	0.18	0.09
29	LCK	PDGFR	S1P	0	0	0	27944	0.91	0.00	0.91	0.00	0.00	0.09
30	IL2RA	S1P	SFas	0	0	0	27889	0.45	0.27	0.45	0.00	0.18	0.09

Table D.13: Top 30 3-Templates for Leukemia Network Basin 3

RANK	Variables			Values			FREQ	SUCCESS	B1	B2	B3	B4	B5
1	Ceramide	PDGFR	S1P	0	1	1	14404	1.00	0.00	0.00	1.00	0.00	0.00
2	Ceramide	IFNG	PDGFR	0	0	1	11323	0.45	0.27	0.00	0.45	0.27	0.00
3	Ceramide	IFNG	S1P	0	0	1	11323	0.45	0.27	0.00	0.45	0.27	0.00
4	IFNG	PDGFR	S1P	0	1	1	11323	0.64	0.18	0.00	0.64	0.18	0.00
5	Ceramide	PDGFR	PLCG1	0	1	1	11278	0.45	0.27	0.00	0.45	0.27	0.00
6	Ceramide	PLCG1	S1P	0	1	1	11278	0.45	0.27	0.00	0.45	0.27	0.00
7	PDGFR	PLCG1	S1P	1	1	1	11278	0.64	0.18	0.00	0.64	0.18	0.00
8	Ceramide	PDGFR	PI3K	0	1	1	11256	0.45	0.36	0.00	0.45	0.18	0.00
9	Ceramide	PI3K	S1P	0	1	1	11256	0.45	0.27	0.00	0.45	0.27	0.00
10	PDGFR	PI3K	S1P	1	1	1	11256	0.64	0.18	0.00	0.64	0.18	0.00
11	Ceramide	MCL1	PDGFR	0	0	1	11018	0.45	0.27	0.00	0.45	0.27	0.00
12	Ceramide	MCL1	S1P	0	0	1	11018	0.45	0.27	0.00	0.45	0.27	0.00
13	MCL1	PDGFR	S1P	0	1	1	11018	0.64	0.18	0.00	0.64	0.18	0.00
14	Ceramide	FasL	PDGFR	0	1	1	10664	0.45	0.27	0.00	0.45	0.27	0.00
15	Ceramide	FasL	S1P	0	1	1	10664	0.45	0.27	0.00	0.45	0.27	0.00
16	FasL	PDGFR	S1P	1	1	1	10664	0.64	0.18	0.00	0.64	0.18	0.00
17	Ceramide	PDGFR	TRADD	0	1	0	10376	0.45	0.27	0.00	0.45	0.27	0.00
18	Ceramide	S1P	TRADD	0	1	0	10376	0.45	0.27	0.00	0.45	0.27	0.00
19	PDGFR	S1P	TRADD	1	1	0	10376	0.64	0.18	0.00	0.64	0.18	0.00
20	BclxL	Ceramide	PDGFR	0	0	1	10289	0.45	0.27	0.00	0.45	0.27	0.00
21	BclxL	Ceramide	S1P	0	0	1	10289	0.45	0.27	0.00	0.45	0.27	0.00
22	BclxL	PDGFR	S1P	0	1	1	10289	0.64	0.18	0.00	0.64	0.18	0.00
23	Ceramide	IFNGT	PDGFR	0	1	1	10043	0.45	0.27	0.00	0.45	0.27	0.00
24	Ceramide	IFNGT	S1P	0	1	1	10043	0.45	0.27	0.00	0.45	0.27	0.00
25	IFNGT	PDGFR	S1P	1	1	1	10043	0.64	0.18	0.00	0.64	0.18	0.00
26	Ceramide	IL2RA	PDGFR	0	0	1	9994	0.45	0.27	0.00	0.45	0.27	0.00
27	Ceramide	IL2RA	S1P	0	0	1	9994	0.45	0.27	0.00	0.45	0.27	0.00
28	IL2RA	PDGFR	S1P	0	1	1	9994	0.64	0.18	0.00	0.64	0.18	0.00
29	Ceramide	DISC	PDGFR	0	0	1	9927	0.45	0.27	0.00	0.45	0.27	0.00
30	Ceramide	DISC	S1P	0	0	1	9927	0.45	0.45	0.00	0.45	0.09	0.00

Table D.14: Top 30 3-Templates for Leukemia Network Basin 4

RANK	Variables			Values			FREQ	SUCCESS	B1	B2	B3	B4	B5
1	FasL	MCL1	NFKB	1	0	1	5203	0.36	0.36	0.09	0.09	0.36	0.09
2	FasL	NFKB	TRADD	1	1	0	5147	0.36	0.36	0.09	0.09	0.36	0.09
3	MCL1	NFKB	TRADD	0	1	0	5051	0.36	0.36	0.09	0.09	0.36	0.09
4	FasL	FasT	NFKB	1	1	1	5031	0.36	0.36	0.09	0.09	0.36	0.09
5	FasL	IL2RBT	NFKB	1	0	1	4954	0.36	0.36	0.09	0.09	0.36	0.09
6	FasL	NFKB	SOCS	1	1	0	4938	0.36	0.36	0.09	0.09	0.36	0.09
7	FasL	MCL1	TRADD	1	0	0	4920	0.36	0.36	0.09	0.09	0.36	0.09
8	FasL	NFKB	PI3K	1	1	1	4904	0.36	0.36	0.18	0.09	0.36	0.00
9	FasT	NFKB	TNF	1	1	1	4869	0.36	0.36	0.09	0.09	0.36	0.09
10	MCL1	NFKB	SOCS	0	1	0	4852	0.36	0.36	0.09	0.09	0.36	0.09
11	IL2RBT	MCL1	NFKB	0	0	1	4848	0.36	0.36	0.09	0.09	0.36	0.09
12	FasL	FasT	TNF	1	1	1	4823	0.36	0.36	0.09	0.09	0.36	0.09
13	FasT	NFKB	TRADD	1	1	0	4779	0.36	0.36	0.09	0.09	0.36	0.09
14	FasL	NFKB	TNF	1	1	1	4764	0.36	0.36	0.09	0.09	0.36	0.09
15	MCL1	NFKB	PI3K	0	1	1	4725	0.36	0.36	0.18	0.09	0.36	0.00
16	NFKB	SOCS	TRADD	1	0	0	4720	0.36	0.36	0.09	0.09	0.36	0.09
17	FasL	FasT	TRADD	1	1	0	4718	0.36	0.36	0.09	0.09	0.36	0.09
18	IL2RBT	NFKB	TRADD	0	1	0	4712	0.36	0.36	0.09	0.09	0.36	0.09
19	FasL	IL2RBT	MCL1	1	0	0	4691	0.36	0.36	0.09	0.09	0.36	0.09
20	NFKB	PI3K	TRADD	1	1	0	4683	0.36	0.36	0.18	0.09	0.36	0.00
21	FasL	MCL1	SOCS	1	0	0	4682	0.36	0.36	0.09	0.09	0.36	0.09
22	FasT	TNF	TRADD	1	1	0	4677	0.36	0.36	0.09	0.09	0.36	0.09
23	FasL	IL2RB	NFKB	1	0	1	4658	0.36	0.36	0.09	0.09	0.36	0.09
24	BclxL	FasL	NFKB	0	1	1	4638	0.36	0.36	0.09	0.09	0.36	0.09
25	IL2RBT	NFKB	SOCS	0	1	0	4632	0.36	0.36	0.09	0.09	0.36	0.09
26	FasL	NFKB	ZAP70	1	1	0	4621	0.36	0.36	0.09	0.09	0.36	0.09
27	FasT	MCL1	NFKB	1	0	1	4619	0.36	0.36	0.09	0.09	0.36	0.09
28	IL2RB	MCL1	NFKB	0	0	1	4612	0.36	0.36	0.09	0.09	0.36	0.09
29	NFKB	TNF	TRADD	1	1	0	4612	0.36	0.36	0.09	0.09	0.36	0.09
30	FasL	SOCS	TRADD	1	0	0	4600	0.36	0.36	0.09	0.09	0.36	0.09

APPENDIX E

CONTEXTUAL GENE SET INTERACTION SUPPLEMENTS

This appendix provides additional information for the Contextual Gene Set Interaction Network Chapter (Ch. 5).

E.1 Proof of Theorem

Below is the proof referenced in §5.4.

Theorem 1. *Once* $\mathbf{S}_{T_k}$ *is given,* $\gamma(\mathbf{G}_i \leftrightarrow \mathbf{G}_j; \mathbf{S}_U, \mathbf{S}_{T_k})$ *is proportional to* $\Pr(\mathbf{G}_i \leftrightarrow \mathbf{G}_j | \mathbf{S}_{T_k})$.

Proof. With the Bayes' theorem,

$$\gamma(\mathbf{G}_i \leftrightarrow \mathbf{G}_j; \mathbf{S}_U, \mathbf{S}_{T_k}) = \frac{\Pr(\mathbf{G}_i \leftrightarrow \mathbf{G}_j | \mathbf{S}_U)}{\Pr(\mathbf{G}_i \leftrightarrow \mathbf{G}_j | \mathbf{S}_U - \mathbf{S}_{T_k})} \tag{E.1}$$

$$= \frac{\Pr(\mathbf{S}_U | \mathbf{G}_i \leftrightarrow \mathbf{G}_j) \Pr(\mathbf{G}_i \leftrightarrow \mathbf{G}_j)}{\Pr(\mathbf{S}_U)} \times \frac{\Pr(\mathbf{S}_U - \mathbf{S}_{T_k})}{\Pr(\mathbf{S}_U - \mathbf{S}_{T_k} | \mathbf{G}_i \leftrightarrow \mathbf{G}_j) \Pr(\mathbf{G}_i \leftrightarrow \mathbf{G}_j)} \tag{E.2}$$

$$= \frac{\Pr(\mathbf{S}_U | \mathbf{G}_i \leftrightarrow \mathbf{G}_j)}{\Pr(\mathbf{S}_U - \mathbf{S}_{T_k} | \mathbf{G}_i \leftrightarrow \mathbf{G}_j)} \times \frac{\Pr(\mathbf{S}_U - \mathbf{S}_{T_k})}{\Pr(\mathbf{S}_U)} \tag{E.3}$$

By assuming independence in observing each sample given $\mathbf{G}_i \leftrightarrow \mathbf{G}_j$,

$$\gamma(\mathbf{G}_i \leftrightarrow \mathbf{G}_j; \mathbf{S}_U, \mathbf{S}_{T_k}) = \Pr(\mathbf{S}_{T_k} | \mathbf{G}_i \leftrightarrow \mathbf{G}_j) \times \frac{\Pr(\mathbf{S}_U - \mathbf{S}_{T_k})}{\Pr(\mathbf{S}_U)} \tag{E.4}$$

With the Bayes' theorem,

$$\Pr(\mathbf{G}_i \leftrightarrow \mathbf{G}_j | \mathbf{S}_{T_k}) = \frac{\Pr(\mathbf{S}_{T_k} | \mathbf{G}_i \leftrightarrow \mathbf{G}_j) \Pr(\mathbf{G}_i \leftrightarrow \mathbf{G}_j)}{\Pr(\mathbf{S}_{T_k})} \tag{E.5}$$

Thus,

$$\Pr(\mathbf{S}_{T_k} | \mathbf{G}_i \leftrightarrow \mathbf{G}_j) = \frac{\Pr(\mathbf{S}_{T_k})}{\Pr(\mathbf{G}_i \leftrightarrow \mathbf{G}_j)} \times \Pr(\mathbf{G}_i \leftrightarrow \mathbf{G}_j | \mathbf{S}_{T_k}) \tag{E.6}$$

From Equations (E.4) and (E.6),

$$\gamma(\mathbf{G}_i \leftrightarrow \mathbf{G}_j; \mathbf{S}_U, \mathbf{S}_{T_k}) = \frac{\Pr(\mathbf{S}_{T_k})}{\Pr(\mathbf{G}_i \leftrightarrow \mathbf{G}_j)} \times \Pr(\mathbf{G}_i \leftrightarrow \mathbf{G}_j | \mathbf{S}_{T_k}) \times \frac{\Pr(\mathbf{S}_U - \mathbf{S}_{T_k})}{\Pr(\mathbf{S}_U)} \quad \text{(E.7)}$$

$$= \frac{1}{\Pr(\mathbf{G}_i \leftrightarrow \mathbf{G}_j)} \times \frac{\Pr(\mathbf{S}_U - \mathbf{S}_{T_k}) \Pr(\mathbf{S}_{T_k})}{\Pr(\mathbf{S}_U)} \times \Pr(\mathbf{G}_i \leftrightarrow \mathbf{G}_j | \mathbf{S}_{T_k}) \quad \text{(E.8)}$$

By assuming a uniform prior for $\Pr(\mathbf{G}_i \leftrightarrow \mathbf{G}_j)$, $\frac{1}{\Pr(\mathbf{G}_i \leftrightarrow \mathbf{G}_j)}$ is a constant. And once $\mathbf{S}_{T_k}$ is given, $\frac{\Pr(\mathbf{S}_U - \mathbf{S}_{T_k}) \Pr(\mathbf{S}_{T_k})}{\Pr(\mathbf{S}_U)}$ is also a constant. Thus, the equation (E.8) can be written as follows:

$$\gamma(\mathbf{G}_i \leftrightarrow \mathbf{G}; \mathbf{S}_U, \mathbf{S}_{T_k}) = C \times \Pr(\mathbf{G}_i \leftrightarrow \mathbf{G}_j | \mathbf{S}_{T_k}), \quad \text{(E.9)}$$

where C is a constant. Therefore,

$$\gamma(\mathbf{G}_i \leftrightarrow \mathbf{G}; \mathbf{S}_U, \mathbf{S}_{T_k}) \propto \Pr(\mathbf{G}_i \leftrightarrow \mathbf{G}_j | \mathbf{S}_{T_k}) \quad \text{(E.10)}$$

□

E.2 Cholesterol Figure

Here is the figure related to the example of identifying condition specificity on the cholesterol network.

E.3 Refractory Cancer Figures

Here are figures relating to the results for the Refractory Cancer data.

E.4 GBM Figures

Here are figures relating to the results for the GBM TCGA data.

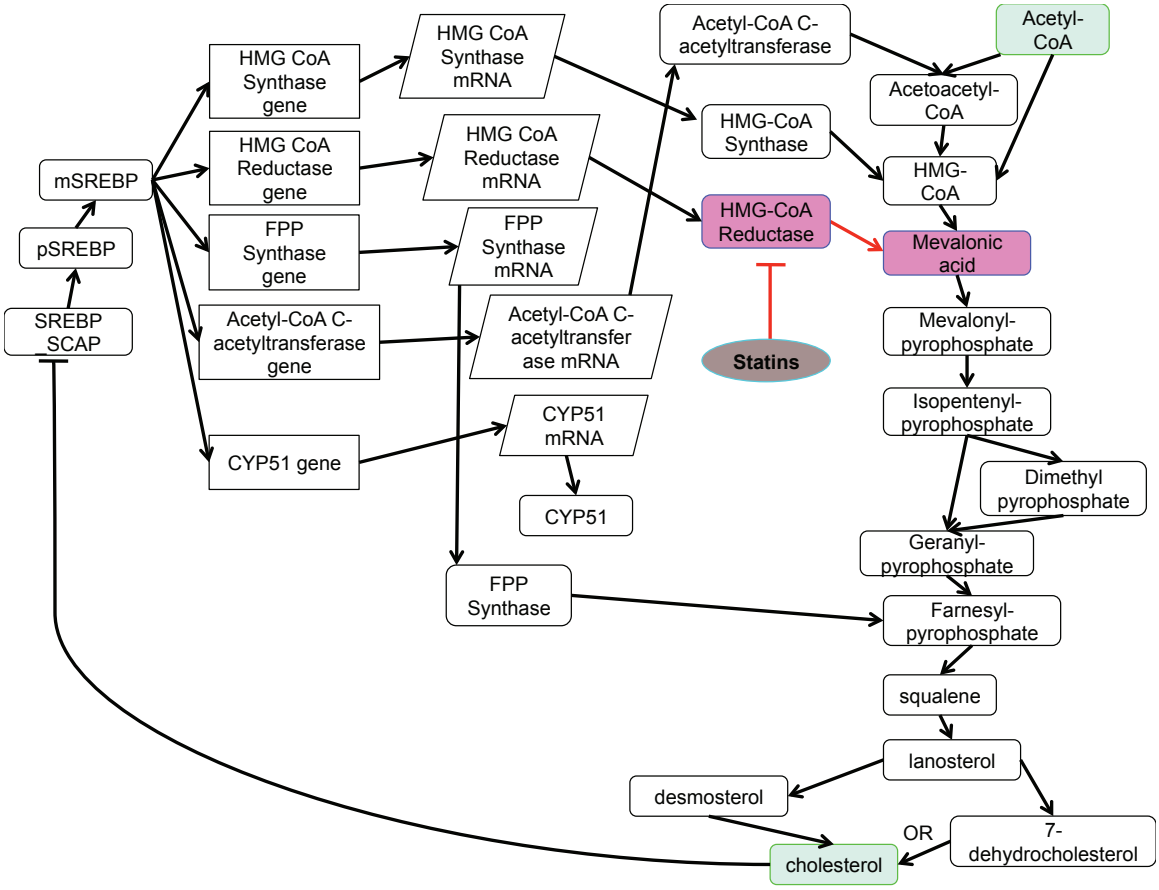


Figure E.1: Boolean Network Model for the Cholesterol Regulatory Pathway: A network diagram of the Boolean network model for the cholesterol regulatory pathway is given. All incoming connections into a node constitute AND logic except cholesterol that has OR logic. A connection ending with a bar indicates NOT logic. The cholesterol synthesis pathway is shown from the precursor Acetyl-CoA to the final product cholesterol. There is a feedback from cholesterol to *SREBP-SCAP*. Statins inhibits HMG-CoA reductase, and regulates the synthesis of cholesterol. After evaluating the specificity of each regulation to the statins perturbation, statins perturbation-specific regulations were colored with red.

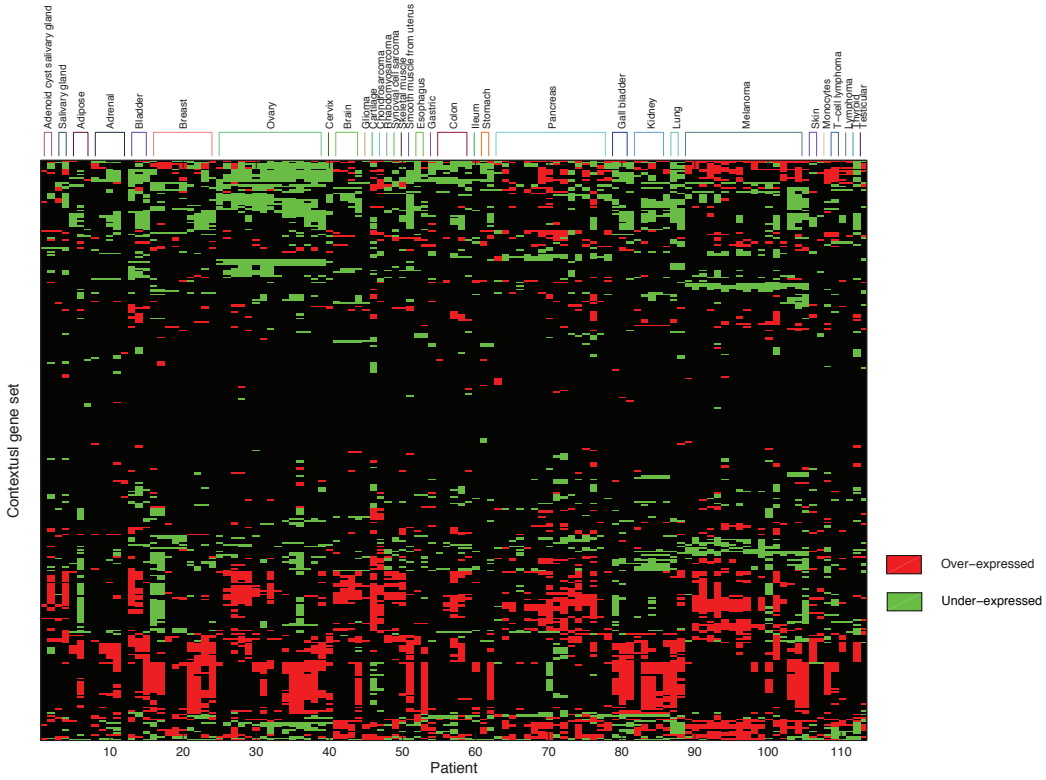


Figure E.2: The summarized gene set expression data of refractory cancer patients from §5.5. There are 339 contextual gene sets for 113 cancer patient samples of 32 tissue types. The expressions of genes in each contextual gene set were summarized to one of three discrete values UP (red), DOWN (green), and NOCHANGE (black).

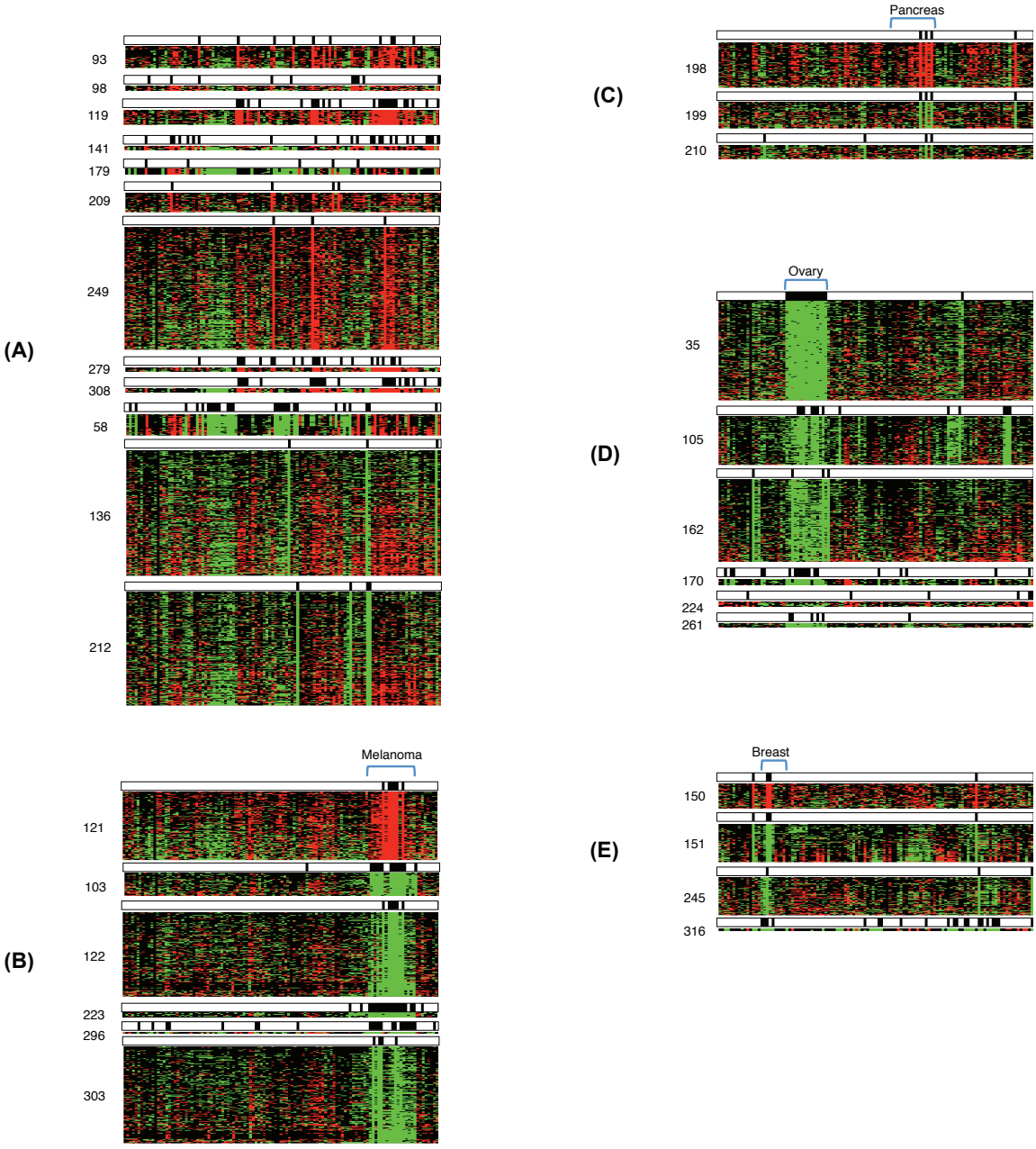


Figure E.3: Regional Heat Maps from the Refractory Cancer Contextual Gene Set Interaction Network: From §5.5, the heat maps of cancer-generic region (A) and tissue specific regions (B) - (E) from the refractory cancer contextual gene set interaction network. (A) Cancer-generic. (B) Melanoma. (C) Pancreas. (D) Ovary. (E) Breast. The numbers are IDs of contextual gene sets. The corresponding contextual condition of each contextual gene set is marked with black color above the heat map. Over-expression is colored with red and under-expression is colored with green.

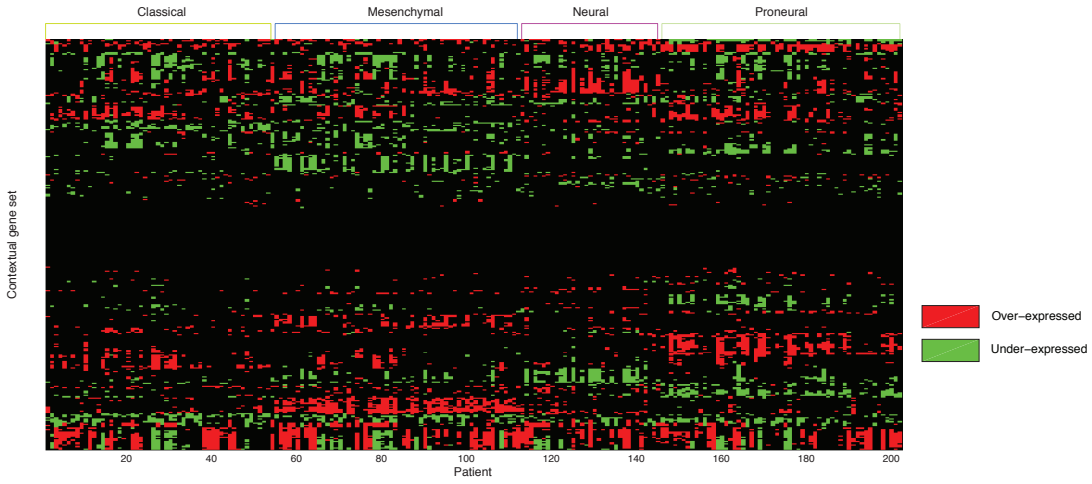


Figure E.4: The Summarized Gene Set Expression Data of GBM Samples from TCGA: From §5.5, we show the summarized gene set expression data of GBM samples from TCGA. There are 316 contextual gene sets for 202 GBM patient samples of four subtypes. The expressions of genes in each contextual gene set were summarized to one of three discrete values UP (red), DOWN (green), and NOCHANGE (black).

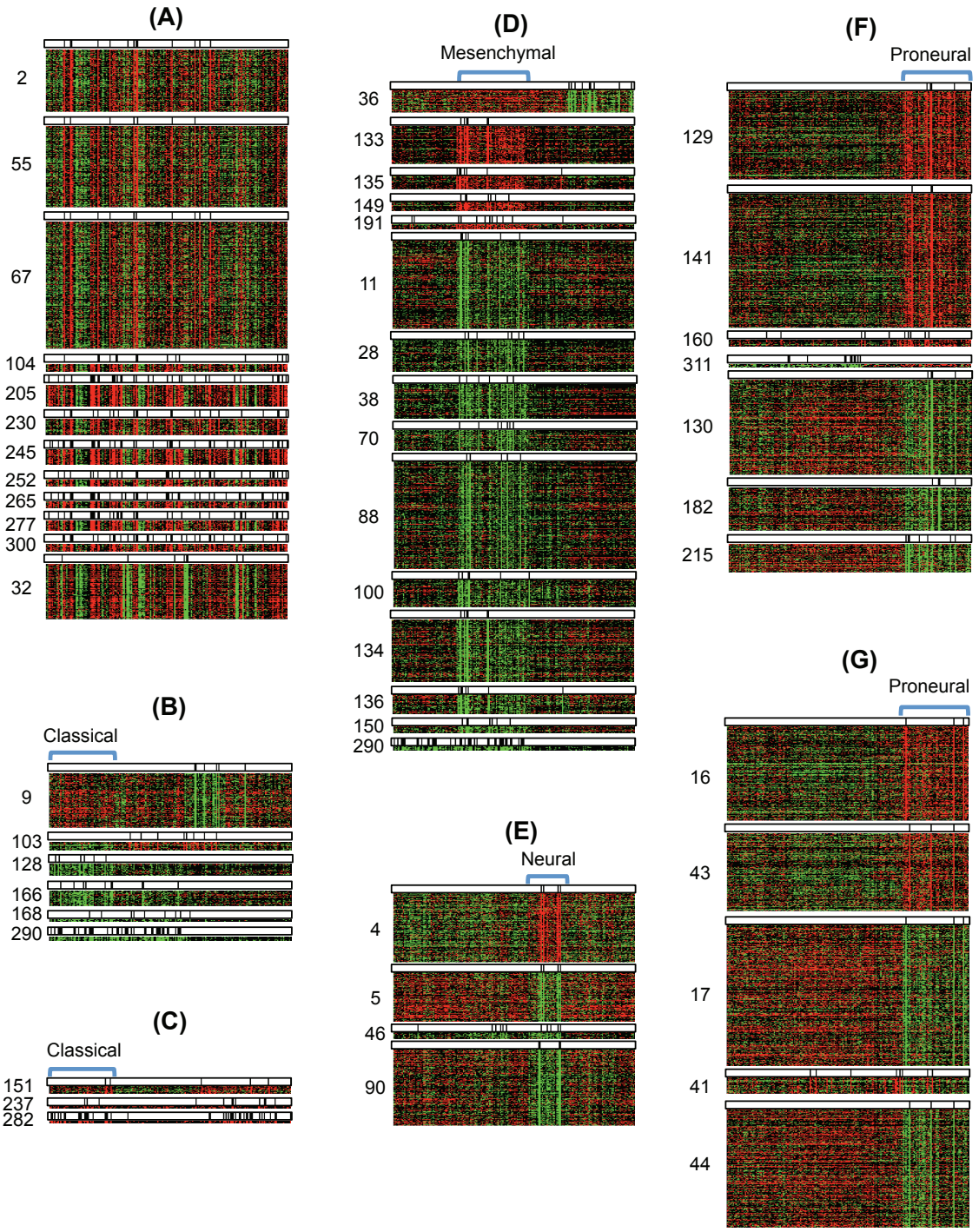


Figure E.5: Heat Maps from the GBM Contextual Gene Set Interaction Network: The heat maps of GBM-generic region (A) and GBM subtype-specific regions (B) (G) from the TCGA-GBM contextual gene set interaction network. (A) GBM-generic. (B) and (C) Classical. (D) Mesenchymal. (E) Neural. (F) and (G) Proneural. The corresponding contextual condition of each contextual gene set is marked with black color above the heat map. Over-expression is colored with red and under-expression is colored with green.

APPENDIX F

BAYESIAN NETWORKS

F.1 Introduction

Bayesian networks are a popular graphical model used in the biological domain due to their ability to encode uncertainty through probabilities and because the conditional independence relationships encoded in the topologies are often useful in modeling biological interactions. Generally they are a probabilistic graphical model comprised of a graph with a node (vertex) for each variable, directed edges (creating no directed cycles) encoding dependence relationships between the nodes, and, for each node, conditional probability tables. These conditional probability tables give the values of the nodes conditioned on the values of their respective parents (input nodes). Formally, a set of random variables with respect to a directed, acyclic graph (DAG) is called a Bayesian network if and only if it satisfies the Markov Property, which states that each variable is conditionally independent of its non-descendent variables given the values of its parent variables.

The recent popularity of Bayesian networks can be attributed to Judea Pearl's seminal text [100], and Neapolitan's text [9] provides a complete overview of learning Bayesian networks from data. Nir Friedman *et al.*introduced the use of Bayesian networks to the biomedical domain [101] and since then they have been rapidly adopted in various application approaches, for example as a comparison to probabilistic Boolean networks as models for genomic regulation [29]. Bayesian networks have are attractive to use as models for biological interactions due to their handling of uncertainty and the ability to infer the network structures from given data. Structure inference alone is a challenging problem as the number of possible structures in a Bayesian network is super-exponential based on the number of variables. However heuristic search approaches such as simulated annealing[102] allow searching the space of structures to learn a suitable structure. While this is attractive, the nature of the biological data itself presents several challenges.

One of the major limitations is the small-sample nature of most biological data sets. Since oftentimes the samples in our biological data sets represent actual patients suffering from a particular disease, there is an obvious hindrance in increasing their numbers. Nonetheless, structure inference for Bayesian networks suffers because conditional independence relationships must be inferred for sometimes thousands of variables across sometimes as few as one or two dozen samples when the ideal ratio of variables and samples is inverted. This issue is addressed many times by reducing the number of variables included in the interaction network through various pre-processing methodologies, for example excluding variables lacking statistically significant fold changes across available samples. Another limitation, and perhaps the one with further-reaching consequences is the main assumption of Bayesian networks, namely that all data is sampled from the same underlying distribution. This is simply not the case of the inherently heterogeneous biological system. Over the past decade we have indeed seen significant results in keeping the assumption of homogeneous data and learning Bayesian networks – we identify more obvious, global and consistent interactions. But with the increasing interest in comparing not healthy versus cancerous, for example, but rather comparing across various subtypes of cancer, we need a model that acknowledges the true homogeneity in the data in order to identify the distinguishing interactions.

F.2 Bayesian Structure Learning

Learning a Bayesian network involves inferring both the structural topology and the parameters of each conditional probability distribution. Depending on the size and the intended usage of the learned network, one may be interested in learning just the structure, which is common in bioinformatics research.

Methods to infer a structure or topology from the data can be broadly grouped into three categories: 1) methods which search the space of possible models using some kind of scoring function (search-and-score), 2) methods which estimate whether or not

conditional independence relationships between variables hold (constraint-based), and 3) methods which are a mixture of the first two (hybrid).

Scoring Functions

Search-and-score approaches use a variety of scores to judge the degree to which the proposed topology fits the data. The scoring functions fall into two categories: Bayesian and information-theoretic. The organization in this section is based on a recent technical report by Carvalho[103].

Under the Bayesian category there are several functions which evaluate the posterior probability of the structure given the data based on some prior probability. In the search, the best network is the one that maximizes this posterior probability. The first is the *Bayesian Dirichlet* (BD), from Heckerman *et al.*in 1995. It assumes multinomial samples, a Dirichlet distribution for parameters, independent parameters and parameter modularity. While the BD score is not so useful in practice, a few special cases of it are. The K2 scoring function, proposed before BD by Cooper and Herskovitz in 1992 is actually a particular case of BD, simplifying the pseudo-count hyperparameter. Heckerman *et al.*also in 1995 address the problem of hyperparameter specification by adding assumptions regarding likelihood equivalence and structure possibility, leading to the *likelihood equivalence Bayesian Dirichlet* (BDe) scoring function, however like BD it has practical difficulties. However, a special case of BDe (actually proposed by Buntine in 1991) called *likelihood equivalence Bayesian Dirichlet with uniform joint distribution* (BDeu), which depends on only one parameter and is much more useful in practice, as it gives the same score to equivalent structures.

Under the information theory category, scoring functions are based on the compression that can be achieved over the data with an optimal code induced by the Bayesian network. These codes are based on Shannon entropy and the Shannon source coding the-

orem which states that as the number of independent and identically distributed variables tends to infinity, no data compression is possible beyond the total Shannon entropy without loss of information. The first score in this category is achieved by taking the logarithm of the likelihood of the data given the network, appropriately called the *log-likelihood* (LL). However this score tends to over fit. The *minimum description length* (MDL), from Lam and Bacchus in 1994 and Suzuki in 1993, is based on LL and prefers simpler solutions (i.e. Occam's Razor). The *Akaike Information Criterion* (AIC) (Akaike 1974) and *Bayesian Information Criterion* (BIC) (based on Schwarz Information Criterion, Schwarz 1978), generalize the MDL scoring function. And recently, a scoring function based on mutual information, called *mutual information tests* (MIT) was proposed by deCampos in 2006.

Search-and-Score

The following overview is organized according to Tsamardinos *et al.*[104]. Using one of the scoring functions above, a search-and-score method search over the space of possible structures, which is super exponential in the number of variables. A basic search method is *greedy hill climbing*, which traverses the search space by a sequence of edge removals, additions, and reversals, but can easily become trapped in a local maximum. To alleviate this symptom of greedy search, the search method can be augmented with random restarts, simulated annealing, or TABU lists. To address the complexity of the search space, *Greedy Bayesian Pattern Search* (GBPS, Acid and de Campos 2003) transforms the traditional search space to the equivalence class of partially directed acyclic graphs (PDAGs), which summarize statistically indistinguishable structures. *Greedy equivalent search* (GES, Meek 1997 and Chickering 2002) is quite similar to GBPS, but with the property of being guaranteed to identify in the sample limit the most probable a posteriori Bayesian network. Besides reducing the state space using PDAGs, another way to address the complexity is the *K2 algorithm* (Cooper and Herskovitz 1992), which combines greedy hill climbing with the *K2* metric and requires total variable ordering. The *Sparse Candidate Algorithm*

from Friedman *et al.*in 1999 estimates a maximum number of candidate parent nodes before beginning the greedy search. The *Optimal Reinsertion* (OR) algorithm takes a unique approach, starting from an initial structure and severing/re-adding edges subject to constraints. There are other algorithms for sure but this list summarizes many of the main ones.

Constraint-Based

The following overview is also organized according to Tsamardinos *et al.*[104]. Generally speaking constraint-based approaches use tests for conditional independence to infer the final graph. The best example of this category is arguably the *PC Algorithm* (Named for Peter Spirtes and Clark Glymour, from their 2000 text with Scheines), which itself is an evolution from the original SGS algorithm from 1990. PC starts with a complete, undirected graph and uses conditional independence tests to iteratively remove edges, then it orients those edges. *Three Phase Dependency Analysis* (TPDA) from Cheng *et al.*in 2002 introduces an information theoretic metric in the conditional independence tests.

Hybrid Algorithms

The final category contains algorithms that combine approaches from both search-and-score and constraint-based. The *CB Algorithm* from Singh *et al.*in 1993 uses SGS to get a variable ordering then applies K2 to give the edges direction. The PC+GBPS (from Spirtes and Meek, 1995) uses PC to create a pattern used to seed the GBPS algorithm. The *Essential Graph Search* (EGS) from Dash and Druzdzel, 1999) repeatedly calls PC with random orders for nodes and various thresholds, and the *Belief Networks Discovery using Cut-set Techniques* (BENEDICT, Acid and de Campos, 2001) uses d-separation and cut-sets to define a scoring metric. Lastly, Tsamardinos *et al.*introduce *Max-Min Hill Climbing* (MMHC)[104], which first reconstructs a skeleton and then performs a greedy hill climbing search to orient the edges.

Other algorithms not in these categories exist, including exact algorithms, but none of them are able to scale to realistic numbers of parents or variables–usually at most one or two dozen variables.

Application to Contextual Gene Set Interaction Networks

Bayesian network learning plays a key role in our contextual gene set interaction networks in Chapter 4 of the dissertation. We first apply Bayesian network learning 1000+ times on our gene set data across all samples, seeding the search space so as to return as many diverse, high-scoring network structures as possible. The edges of these networks are counted, rendering a frequency and concordant likelihood of dependence between gene sets. Following the procedure outlined in the chapter, we systematically excise subsets of samples from the gene set data and run 1000+ instances of Bayesian network learning again on the data without the specific samples of interest (usually corresponding to some disease subtype), and again calculate the likelihood of dependence between gene sets by means of the edge frequencies. The ratio between the likelihood of the edge in the full data and the data without the sample group in question represents the gain of interaction likelihood by including those samples (i.e. a measure contribution by that sample group). In order to evaluate the significance of the edge frequency found by removing subgroups of samples we also follow a permutation approach, removing random sample groups of the same size and performing Bayesian network learning on these permutated data sets.

Parallel Bayesian network learning with a greedy hill climbing/simulated annealing approach allows us to identify many high-scoring networks around the super exponential-sized space of possible structures for our edge frequency counts while avoiding local maxima with simulated annealing and random restarts. The Bayesian framework was chosen for identifying frequency counts because the next step (sample effects on the edges of a contextual gene set interaction network) was also done based on a probabilistic graphi-

cal model, and so the two tied together well. However, since the directionality of edges is not used in our calculations, other network learning approaches could be substituted in for the Bayesian learning, such as mutual-information-based approaches or other correlation/association approaches.