PREDICTIVE NETWORK MODELING AND EXPERIMENTATION IN COMPLEX BIOLOGICAL SYSTEMS: APPLICATIONS TO CANCER AND INFECTIOUS DISEASE

A Dissertation in
Molecular Medicine
by
Steven Nathaniel Steinway

© 2015 Steven Nathaniel Steinway

Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

August 2015
The dissertation of Steven Nathaniel Steinway was reviewed and approved* by the following:

Thomas P. Loughran, Jr.
Adjunct Professor of Microbiology & Immunology
F. Palmer Weber-Smithfield Foods Professor of Oncology Research and Professor of Medicine, University of Virginia
Co-Chair of Committee
Dissertation Co-Advisor

Reka Albert
Distinguished Professor of Physics & Biology
Co-Chair of Committee
Dissertation Co-Advisor

David J. Feith
Adjunct Assistant Professor of Cellular and Molecular Physiology
Associate Professor of Medicine, University of Virginia

James R. Broach
Distinguished Professor of Biochemistry & Molecular Biology
Chair of Biochemistry & Molecular Biology

Diane Thiboutot
Professor of Dermatology
Vice-Chair for Research for Dermatology

Charles H. Lang
Professor of Cellular and Molecular Physiology
Co-Chair of Molecular Medicine Graduate Program

*Signatures are on file in the Graduate School
ABSTRACT

Biology is incredibly complex – at the molecular, cellular, tissue, and population level, there exists a tremendous number of discrete interacting components tightly regulating the processes that sustain life. Biological systems have traditionally been viewed in a reductionist manner often literally (and metaphorically) through a magnifying glass, leading to insight into how the individual parts work. Network theory, on the other hand, can be used to put the pieces together, to understand how complex and emergent behaviors arise from the totality of interactions in complex systems, such as those seen in biology. Network theory is the study of systems of discrete interacting components and provides a framework for understanding complex systems. A network-focused investigation of a complex biological system allows for the understanding of the system’s emergent properties, for example its function and dynamics. Network dynamics are of particular interest biologically because biological systems are not static but are constantly changing in response to perturbations and environmental stimuli in space and time.

Systems level biological analysis has been aided by the recent explosion of high throughput data. This has led to an abundance of quantitative and qualitative information related to the activation of biological systems, but frequently there is still a paucity of kinetic and temporal information. Discrete dynamic modeling provides a means to create predictive models of biological systems by integrating fragmentary and qualitative interaction information. Using discrete dynamic modeling, a structural (static) network of biological regulatory relationships can be translated into a mathematical model without the use of kinetic parameters. This model can describe the dynamics of a biological system (i.e. how it changes over time), both in normal and in perturbation (e.g. disease) scenarios. In this dissertation we present the application of network theory and discrete dynamic modeling integrated with experimental laboratory analysis to understand biological diseases in three contexts.

The first is the construction of a network model of epidermal derived growth factor receptor (EGFR) signaling in cancer. We translate this model into two types of discrete models: a Boolean model and a three-state model. We show how the effects of an EGFR inhibitor (such as the drug gefitinib) can suppress tumor growth, and we model how genomic variants can augment the effect of EGFR inhibition in tumor growth. Importantly, we compare discrete modeling outcomes to an alternative modeling framework, which relies on detailed
kinetic information, called ordinary differential equation (ODE) modeling and show that both models achieve similar findings. Our results demonstrate that discrete dynamic model can accurately model biomedical systems and make important predictions about the effect a drug will have on a disease (e.g. tumor growth) in the context of various perturbations. Importantly, discrete dynamic models can be employed in the absence of kinetic parameters, making this modeling approach suitable for the many biological systems in which detailed kinetic information is not available.

Second, we construct a network model of epithelial-to-mesenchymal transition (EMT), a developmental process hijacked by cancer cells to leave the primary tumor site, invade surrounding tissue, and establish distant metastases. We demonstrate that the EMT network model recapitulates known dysregulations during the induction of EMT and predicts the activation of the Wnt and Sonic hedgehog (SHH) signaling pathways during this process. We confirm the cross-talk between TGFβ, Wnt and SHH signaling in vitro in multiple human liver cancer cell lines and tumor samples. Next, we use the EMT network model to systematically explore perturbations that suppress EMT, with the ultimate goal of identifying therapeutic interventions that suppress tumor invasion. We computationally explore close to half a million individual and combination perturbations to the EMT network and identify that only a dozen suppress EMT. We test these interventions experimentally and our findings suggest that many predicted interventions suppress the EMT process.

Lastly, we construct a model of the enormous ecological community of bacteria that live in our intestines, collectively called the gut microbiome. This model is used to understand the effect of antibiotic treatment and opportunistic *C. difficile* infection (a devastating and highly prevalent disease entity) on the native microbiome and predict therapeutic probiotic interventions to suppress *C. difficile* infection. We integrate this modeling with another type of modeling, genome scale metabolic network reconstructions, to understand metabolic differences between community members and to identify the role of metabolism in the observed microbial interactions. In vitro experimental data validate a key result of my computational model, that *Barnesiella intestinohominis* can in fact suppress *C. difficile* growth. This novel result suggests that *Barnesiella* could potentially be used as a probiotic to suppress *C. difficile* growth.
Taken together, the studies presented in this thesis demonstrate the tremendous capacity of network modeling to elucidate biomedical systems. We build networks, construct mathematical models, study network dynamics, and use network-directed insight to guide experiments in critical biomedical areas. The ultimate goal of this work has been to translate network-directed insight into actionable biomedical findings that lead to improved understanding of human disease, enhanced patient care, and a betterment of the human condition.
TABLE OF CONTENTS

List of Tables ........................................................................................................... x
List of Figures ........................................................................................................... xi
List of Abbreviations .................................................................................................. xiii
Acknowledgments ...................................................................................................... xvi

CHAPTER 1. Introduction to Discrete Dynamic Modeling: A Network Approach
for Systems Pharmacology ................................................................. 1
1.1 Abstract ............................................................................................................. 1
1.2 Introduction ...................................................................................................... 1
1.3 Constructing the Network ............................................................................. 3
1.4 Determining the Boolean Functions ............................................................. 6
1.5 Selecting Time Implementation for State Transitions .................................. 7
1.6 Evaluating the Dynamics and Steady States of the System ......................... 9
1.7 Network Reduction Techniques .................................................................. 13
1.8 Testing the Validity of the Dynamic Model .................................................. 14
1.9 Introducing Model Perturbations .................................................................. 14
1.10 Non-Boolean Discrete Dynamic Modeling (Multi-State Discrete Models) .... 15
1.11 Conclusions .................................................................................................. 16
1.12 Acknowledgments ......................................................................................... 16
1.13 References .................................................................................................... 16

CHAPTER 2. Discrete Dynamic Modeling of EGFR Signaling in Tumor Growth:
A Comparison to Ordinary Differential Equation Modeling ....................... 20
2.1 Introduction .................................................................................................... 20
2.2 A Boolean Model of EGFR Inhibition in Tumor Growth .............................. 20
2.3 A Multi-State Discrete Model of EGFR Inhibition in Tumor Growth .......... 24
2.4 Conclusions .................................................................................................. 25
2.5 Acknowledgments ......................................................................................... 26
2.6 References .................................................................................................... 26

CHAPTER 3. Network Modeling of TGFβ Signaling in Hepatocellular Carcinoma
Epithelial-to-Mesenchymal Transition Reveals Joint Sonic Hedgehog and Wnt Pathway Activation ......................................................... 27
3.1 Abstract ......................................................................................................... 27
3.2 Quick Guide to Equations and Assumptions .............................................. 27
  3.2.1 Major Assumptions of the Model ......................................................... 27
  3.2.2 Key Equations ..................................................................................... 28
3.3 Introduction .................................................................................................. 30
3.4 Materials & Methods ................................................................................... 35
  3.4.1 Cell Culture ......................................................................................... 35
  3.4.2 Cell Line Authentication ...................................................................... 35
  3.4.3 Quantitative real-time polymerase chain reaction ................................ 35
  3.4.4 Western blot analysis ......................................................................... 35
  3.4.5 Microarray gene set enrichment analysis (GSEA) .............................. 36
  3.4.6 Patient Samples .................................................................................. 36
CHAPTER 4. Combinatorial Interventions Lend Insight into the Inhibition of TGFβ Driven Epithelial-to-Mesenchymal Transition and the Existence of Hybrid Epithelial-Mesenchymal Phenotypes

4.1 Abstract ........................................................................................................... 73
4.2 Introduction ...................................................................................................... 74
4.3 Materials & Methods ..................................................................................... 76
  4.3.1 Cell Culture ............................................................................................... 76
  4.3.2 siRNA transfections ................................................................................ 76
  4.3.3 Quantitative real-time polymerase chain reaction ..................................... 77

  4.3.4 Immunoblot analysis ................................................................................ 77
  4.3.5 Detection of cell migration using automated cell imaging acquisition and analysis .......................................................... 78
  4.3.6 Quantification of cell migration with ImageJ ............................................. 78
  4.3.7 Network modeling framework .................................................................. 79
  4.3.8 Dynamic Analysis ...................................................................................... 79
  4.3.9 Perturbation Analysis ............................................................................... 80

  4.3.10 Network Control Methodology ................................................................. 80
4.4 Results ............................................................................................................... 80
  4.4.1 Systematic exploration of perturbations in the EMT network reveal combinatorial inhibition of specific nodes can suppress the TGFβ-driven EMT .................................................. 80
4.4.2 Node combinations identified by systematic perturbation analysis are consistent with a network motif based control methodology applied to the dynamic network model ................................................................. 84
4.4.3 Mesenchymal phenotype cells are enriched for nodes predicted by the EMT network model to inhibit TGFβ-driven EMT ........................................ 86
4.4.4 A multi-faceted siRNA screen to test predicted node knockout combinations in vitro ........................................................................................................ 86
4.4.5 Analysis of motifs that arise after SMAD inhibition in the EMT network reveal an attractor landscape that is distinct from the SMAD unperturbed network ....... 92
4.4.6 Quantitative analysis of steady states in the EMT network reveals a putative spectrum of EMT phenotypes ................................................................. 92
4.5 Discussion ........................................................................................................ 93
4.6 Acknowledgments ............................................................................................... 97
4.7 References ........................................................................................................... 98

CHAPTER 5. Inference of Network Dynamics and Metabolic Interactions in the Gut Microbiome ................................................................. 101
5.1 Abstract .............................................................................................................. 101
5.2 Introduction ...................................................................................................... 101
5.3 Methods .......................................................................................................... 103
  5.3.1 Data Sources .................................................................................................. 103
  5.3.2 Interpolation of Missing Genus Abundance information ................................. 103
  5.3.3 Network Modeling Framework ..................................................................... 104
  5.3.4 Binarization ................................................................................................... 105
  5.3.5 Inference of Boolean Rules from Time Series Genus Abundance Information 106
  5.3.6 Dynamic Analysis ........................................................................................ 107
  5.3.7 Perturbation Analysis ..................................................................................... 108
  5.3.8 Generating Genus-Level Genome-Scale Metabolic Reconstructions ............. 108
  5.3.9 Subsystem Enrichment Analysis ................................................................... 108
  5.3.10 Identifying Seed Sets and Defining Metabolic Competition Score ............. 109
  5.3.11 Co-culture and Spent Media Experiments ................................................ 110
5.4 Results ................................................................................................................ 111
  5.4.1 Construction of a Microbial Genus Abundance Dataset for Network Inference111
  5.4.2 Binarization of Gut Microbiome Genus Abundance Information for Construction of a Dynamic Network Model ...................................................... 114
  5.4.3 Construction of a Dynamic Network Model from Binarized Time Series Microbial Genus Abundance Information ...................................................... 115
  5.4.4 Perturbation Analysis ..................................................................................... 118
  5.4.5 Generating Genus-Level Metabolic Reconstructions .................................... 118
  5.4.6 Subsystem Enrichment Analysis ................................................................... 119
  5.4.7 Generating Metabolic Competition Scores .................................................. 125
5.5 Discussion .......................................................................................................... 127
  5.5.1 Network Structure ........................................................................................... 128
  5.5.2 Experimental Validation of *Barnesiella* Inhibition of *C. difficile* ................. 129
  5.5.3 Network Dynamics and Perturbation Analysis ............................................. 130
5.5.4 Metabolic Competition Scores Point Towards a Non-Metabolic Interaction Mechanism ................................................................. 130
5.6 Conclusions .................................................................................. 133
5.7 Acknowledgments ......................................................................... 133
5.8 References ..................................................................................... 134

CHAPTER 6. Conclusions & Outlook....................................................... 138
LIST OF TABLES

Table 2-1. Boolean Rules for the Reduced (7 node) EGFR Network 22
Table 2-2. Genes and Perturbations Modeled in the EGFR ePD Model ....................... 23
Table 2-3. Comparison of the Outcomes of the ePD Model and of the Boolean Model for the Effect of an EGFR Inhibitor on Tumors with Specific Genome Profiles. ...... 24
Table 3-1. Evidence for interactions among nodes in the EMT network......................... 36
Table 3-2. The full names of components in the EMT network corresponding to the abbreviated node labels used in Figure 3-7 and Figure 3-9A......................... 51
Table 3-3. Key experimental outcomes reproduced by the EMT network model............... 55
Table 3-4. Boolean update rules and initial state of the 68 node EMT network ............... 56
Table 3-5. Boolean rules and initial state for the 19 node reduced EMT network ............ 61
Table 4-1. Results of the EMT network model perturbation screen............................ 84
Table 4-2. Components of the epithelial control set............................................. 85
Table 5-1. Boolean update rules for the gut microbiome network............................... 116
Table 5-2. Basin size as % of total state space (unique basin size) for experimentally realized network attractors ................................................................. 117
Table 5-3. Genus level genome scale metabolic network reconstructions ..................... 119
Table 5-4. Unique reactions within genera. Each row has n reactions that column does not have........................................................................................................... 120
Table 5-5. Reaction overlap between genera.................................................................. 120
LIST OF FIGURES

Figure 1-1. A simple signaling network and its Boolean representations ............................................. 5
Figure 1-2. Fraction of simulations where O=ON as a function of time........................................ 10
Figure 1-3. The state transition network of the four node network given in Figure 1-1
obtained using the synchronous update method............................................................. 12
Figure 1-4. The state transition network of the four node network given in Figure 1-1
obtained using a random order asynchronous updating scheme ........................................ 13
Figure 2-1. The EGFR signaling network used to model EGFR inhibitor therapeutics ........ 21
Figure 2-2. Truth tables for the multi-state discrete model of EGFR inhibition on tumor
growth .................................................................................................................................. 25
Figure 3-1. The ratio between the timescale of the slow and fast processes does not matter as
long as it is significantly larger than one ........................................................................ 29
Figure 3-2. Exploration of alternative rules to describe nuclear localization of β-catenin .... 32
Figure 3-3. Flowchart of computational algorithms ..................................................................... 33
Figure 3-4. All stable motifs (feedback loops) stabilizing the TGFβ-driven EMT
phenotype ............................................................................................................................ 43
Figure 3-5. Active TGFβ signaling correlates with mesenchymal phenotype in murine and
human models of HCC ....................................................................................................... 48
Figure 3-6. TGFβR inhibition suppresses mesenchymal features in human HCC cell lines .. 49
Figure 3-7. The 70 node, 135 edge network model representing the epithelial-
to-mesenchymal transition (EMT) signaling network in HCC ......................................... 50
Figure 3-8. Network modeling reproduces established signaling abnormalities seen in TGFβ-
driven epithelial-to-mesenchymal transition and predicts novel cross-talk
mechanisms .......................................................................................................................... 59
Figure 3-9. Network reduction and state space analysis of the EMT network...................... 60
Figure 3-10. AXIN2 and GLI are induced by Wnt and SHH signals in the epithelial-to-
mesenchymal transition network model ............................................................................ 62
Figure 3-11. Network model-directed testing validates activation of Sonic hedgehog and Wnt
signaling pathways by TGFβ ............................................................................................... 64
Figure 3-12. Network motifs that stabilize in the EMT=ON steady state ....................... 66
Figure 4-1. An *in silico* combinatorial knockout screen in the EMT network reveals specific node combinations can suppress the TGFβ-driven EMT ........................................... 82

Figure 4-2. Mesenchymal phenotype cells are enriched for nodes predicted by the EMT network model to inhibit TGFβ-driven EMT ................................................................. 88

Figure 4-3. Node knockdown efficiency for siRNAs used ........................................ 89

Figure 4-4. Western blots used to quantitate the effect of node knockdown combinations on E-cadherin expression ............................................................... 89

Figure 4-5. A multi-faceted siRNA screen to test predicted node knockdown combinations *in vitro* ........................................................................................................... 90

Figure 4-6. Analysis of motifs that arise after SMAD inhibition in the EMT network reveal an attractor landscape that is distinct from the SMAD unperturbed network ....... 93

Figure 4-7. Principal components analysis of single perturbations in the EMT network model supports the existence of an EMT spectrum ......................................................... 94

Figure 5-1. Dynamic analysis workflow .................................................................... 103

Figure 5-2. Bacterial genera abundances over time in response to clindamycin treatment and /or *C. difficile* inoculation .................................................................................. 104

Figure 5-3. Averaged binarized genera abundances using iterative k-means binarization ... 111

Figure 5-4. Averaged binarized genera abundances using iterative k-means binarization were rounded to the most probable binarized state ....................................................... 112

Figure 5-5. Construction of a network model of the gut microbiome from time course metagenomic genus abundance information .............................................................................. 115

Figure 5-6. All possible steady states of the Boolean model of the gut microbiome ....... 121

Figure 5-7. Steady states and node perturbations in the gut microbiome model ............. 122

Figure 5-8. Subsystem enrichment analysis highlights metabolic differences between taxa 123

Figure 5-9. Metabolic competition scores and *in vitro* data indicate a non-metabolic interaction mechanism ........................................................................................................ 124

Figure 5-10. Competition and mutualism scores by edge and path type in Boolean network .......................................................................................................................... 125

Figure 5-11. Computational models can bring us closer to true interaction networks ..... 132
### LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AKT</td>
<td>v-akt murine thymoma viral oncogene</td>
</tr>
<tr>
<td>AUC</td>
<td>area under the growth curve</td>
</tr>
<tr>
<td>β-catenin_memb</td>
<td>membrane-bound β-catenin</td>
</tr>
<tr>
<td>β-catenin_nuc</td>
<td>nuclear β-catenin</td>
</tr>
<tr>
<td>βTrCP</td>
<td>beta-transducin repeat containing E3 ubiquitin protein ligase</td>
</tr>
<tr>
<td>CD44</td>
<td>CD44 molecule (Indian blood group)</td>
</tr>
<tr>
<td>CDC42</td>
<td>cell division cycle 42</td>
</tr>
<tr>
<td>CDH1</td>
<td>E-cadherin gene</td>
</tr>
<tr>
<td>c-fos</td>
<td>FBJ murine osteosarcoma viral oncogene homolog</td>
</tr>
<tr>
<td>CHD1L</td>
<td>1-like</td>
</tr>
<tr>
<td>CMB</td>
<td>chopped meat medium</td>
</tr>
<tr>
<td>CML</td>
<td>met proto-oncogene (hepatocyte growth factor receptor)</td>
</tr>
<tr>
<td>CML</td>
<td>recombination signal binding protein for</td>
</tr>
<tr>
<td>CSL</td>
<td>immunoglobulin kappa J region</td>
</tr>
<tr>
<td>CSN</td>
<td>COP9 signalosome</td>
</tr>
<tr>
<td>CTNNB1</td>
<td>catenin (cadherin-associated protein), beta 1; β-catenin</td>
</tr>
<tr>
<td>DELTA</td>
<td>DELTA-Like ligand</td>
</tr>
<tr>
<td>Dest_compl</td>
<td>destruction complex (adenomatous polyposis coli; axin 2; glycogen synthase kinase 3 beta)</td>
</tr>
<tr>
<td>DSH</td>
<td>dishevelled, dsh homolog 1</td>
</tr>
<tr>
<td>E-cadherin</td>
<td>cadherin 1, type 1, E-cadherin</td>
</tr>
<tr>
<td>EGF</td>
<td>epidermal growth factor</td>
</tr>
<tr>
<td>EGFR</td>
<td>epidermal derived growth factor receptor</td>
</tr>
<tr>
<td>EGR1</td>
<td>early growth response 1</td>
</tr>
<tr>
<td>EMT</td>
<td>epithelial-to-mesenchymal transition</td>
</tr>
<tr>
<td>ERK</td>
<td>mitogen-activated protein kinase</td>
</tr>
<tr>
<td>FGF</td>
<td>fibroblast growth factor 2</td>
</tr>
<tr>
<td>FGFR</td>
<td>fibroblast growth factor receptor</td>
</tr>
<tr>
<td>FOXC2</td>
<td>forkhead box C2 (MFH-1, mesenchyme forkhead 1)</td>
</tr>
<tr>
<td>Frizzled</td>
<td>frizzled family receptor</td>
</tr>
<tr>
<td>FUS</td>
<td>serine/threonine kinase 36</td>
</tr>
<tr>
<td>GLI</td>
<td>GLI family zinc finger</td>
</tr>
<tr>
<td>Goosecoid</td>
<td>goosecoid homeobox</td>
</tr>
<tr>
<td>GSEA</td>
<td>gene set enrichment analysis</td>
</tr>
<tr>
<td>GSK3β</td>
<td>glycogen synthase kinase 3 beta</td>
</tr>
</tbody>
</table>
HCC  hepatocellular carcinoma
    hairy/enhancer-of-split related with YRPW
HEY1  motif 1
HGF  hepatocyte growth factor
    hypoxia inducible factor 1, alpha subunit (basic
HIF1α  helix-loop-helix transcription factor
IGF1  Insulin-like growth factor 1
IGF1R  insulin-like growth factor 1 receptor
IKKα  conserved helix-loop-helix ubiquitous kinase
IKM  iterative k-means binarization
ILK  integrin-linked kinase
Jagged  jagged 1
KEGG  Kyoto Encyclopedia of Genes and Genomes
KM3  iterative k-means binarization at a depth of 3
    Solute Carrier Family 39 (Zinc Transporter),
LIV1  Member 6
LOXL23  lysyl oxidase-like
LPS  lipopolysaccharide
MAPK  Mitogen-activated protein kinase
MEK  mitogen-activated protein kinase
miR200  microRNA 200b
mir-221  microRNA 221
NFκB  nuclear factor kappa beta
NOTCH  NOTCH (Drosophila) Homolog 1
NOTCH_ic  intracellular NOTCH
ODE  ordinary differential equation
OUT  operational taxonomic unit
PAK1  p21 protein (CDC42/Rac)-activated kinase 1
Patched  PTCH Homolog 1 (Drosophila)
PDF  probability distribution function
PDGF  platelet-derived growth factor beta polypeptide
PDGFR  platelet-derived growth factor receptor
PEBP1  Phosphatidylethanolamine-binding protein 1
PI3K  phosphoinositide-3-kinase
PVDF  polyvinylidene difluoride
RAF  raf oncogene
RAS  ras oncogene
RASAL1  RAS protein activator like 1
Rasgap  RAS p21 protein activator 1
RAST  Rapid Annotations based on Subsystem
Technology
Rkip  Raf kinase inhibitor protein
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHH</td>
<td>Sonic hedgehog</td>
</tr>
<tr>
<td>SMAD</td>
<td>SMAD family member complex 2,3,4</td>
</tr>
<tr>
<td>SMO</td>
<td>smoothened, frizzled family receptor</td>
</tr>
<tr>
<td>SNAI1</td>
<td>snail homolog 1 (Drosophila)</td>
</tr>
<tr>
<td>SNAI2</td>
<td>snail homolog 2 (Drosophila)</td>
</tr>
<tr>
<td>SOS/GRB2</td>
<td>son of sevenless homolog 1 (Drosophila) &amp; growth factor receptor-bound protein 2</td>
</tr>
<tr>
<td>SRC</td>
<td>v-src sarcoma (Schmidt-Ruppin A-2) viral oncogene homolog (avian)</td>
</tr>
<tr>
<td>STAT</td>
<td>signal transducer and activator of transcription</td>
</tr>
<tr>
<td>SUFU</td>
<td>suppressor of fused homolog</td>
</tr>
<tr>
<td>TCF/LEF</td>
<td>transcription factor 4, a basic helix-loop-helix transcription factor; lymphoid enhancer-binding factor 1</td>
</tr>
<tr>
<td>TGFβ</td>
<td>transforming growth factor beta</td>
</tr>
<tr>
<td>TGFβR</td>
<td>transforming growth factor beta receptor</td>
</tr>
<tr>
<td>TLR</td>
<td>toll-like receptor</td>
</tr>
<tr>
<td>TWIST1</td>
<td>twist basic helix-loop-helix transcription factor 1</td>
</tr>
<tr>
<td>VRE</td>
<td>vancomycin-resistant Enterococci</td>
</tr>
<tr>
<td>Wnt</td>
<td>wingless-type MMTV integration site family, member 1</td>
</tr>
<tr>
<td>ZEB1</td>
<td>zinc finger E-box binding homeobox 1</td>
</tr>
<tr>
<td>ZEB2</td>
<td>zinc finger E-box binding homeobox 2</td>
</tr>
<tr>
<td>β-catenin_memb</td>
<td>membrane-bound β-catenin</td>
</tr>
<tr>
<td>β-catenin_nuc</td>
<td>nuclear β-catenin</td>
</tr>
<tr>
<td>βTrCP</td>
<td>beta-transducin repeat containing E3 ubiquitin protein ligase</td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENTS

The work I present here was far from an individual effort and would not have been possible without the help and encouragement of so many people.

I am forever grateful to my parents, whose guidance, support, and encouragement has shaped who I am today, to my brother and sister for biggest supporters, and to the rest of my friends and family for being so supportive throughout my training.

I would next like to thank my mentors Dr. Reka Albert, Dr. Thomas Loughran, and Dr. David Feith who have all been exemplary role models to me. To Reka, I am forever grateful for your unwavering support through both chaotic and exciting moments. You always seemed to have the answers when I did not, and I have learned and grown tremendously through your guidance. To Dr. Loughran, I am truly grateful for your mentorship and support, allowing me the freedom to explore my creative endeavors in your lab. I could not have asked for a better “physician scientist” mentor. To Dave, I am grateful for your keen questions, attention to detail, and help through all the day-to-day difficulties of being a graduate student. I am also extremely grateful for my committee for their valuable guidance throughout my training.

I have had the honor of working closely with two brilliant PhD candidates during my training. I collaborated with Jorge Gomez Tejeda Zañudo (in Reka Albert’s group) on the dynamic network modeling of liver cancer invasion and metastasis covered in chapters 3 and 4 of my dissertation. My work in chapter 5, describing modeling the gut microbiome was a collaborative effort with Matt Biggs (in Jason Papin’s Lab at University of Virginia). I am additionally grateful to all the members (past and present) of the Albert group and Loughran lab for their support and assistance throughout my training.
CHAPTER 1.
Introduction to Discrete Dynamic Modeling: A Network Approach for Systems Pharmacology

1.1 Abstract

Systems pharmacology is an interdisciplinary field that aims to apply the theoretical and experimental tools of systems biology to drug development. The goal is to go beyond the interaction between a drug and the target to which it binds to explore drug effects on the cellular networks affected by disease. Over the years, vast amounts of information about the regulatory relationships among genes, proteins, and small molecules have been acquired. Similarly, there is much known about the deregulation of these systems during disease. However, many knowledge gaps still exist. There is an abundance of qualitative or relative information related to the activation of signaling pathways, but a paucity of kinetic and temporal information. Discrete dynamic modeling provides a means to create predictive models of signal transduction pathways by integrating fragmentary and qualitative interaction information. Using discrete dynamic modeling, a structural (static) network of biological regulatory relationships can be translated into a mathematical model without the use of kinetic parameters. This model can describe the dynamics of a biological system over time, both in normal and in perturbation scenarios. In this chapter, we discuss the fundamentals of discrete dynamic modeling as it pertains to systems pharmacology. As an example, we apply this methodology to a previously constructed pharmacodynamic model of epidermal derived growth factor receptor (EGFR) signaling. We (1) translate this model into two types of discrete models, a Boolean model and a three-state model, (2) show how the effects of an EGFR inhibitor (such as gefitinib) can suppress tumor growth, and (3) model how genomic variants can augment the effect of EGFR inhibition in tumor growth. We argue that discrete dynamic models can be used to facilitate many of the goals of systems pharmacology. These include understanding how individual differences contribute to variability in drug response and determining which drugs would be best depending on individual genetic differences.

1.2 Introduction
Systems biology involves the reconstruction of a system from individual biological interactions. Small scale and high throughput experiments over a long period amass to systems-level information. Collecting this information into holistic models is a fundamental goal of systems biology, because it allows the study of biological processes at a systems level, realizing emergent properties that are critical to the biological process and that may not be recognized from traditional reductionist views. Systems biology relies on a combination of experimental and computational techniques. Experimental techniques include the acquisition of “omics” level analysis of DNA, mRNA, proteins, and metabolites related to specific biological processes and disease states. Computational techniques are required to analyze and interpret these massive amounts of data.

Systems pharmacology is an emerging interdisciplinary field that aims to use the tools of systems biology to improve the development of drugs and to understand drug effects on the body. Traditional drug discovery has largely focused on the drug-target interaction. It is apparent now that this is insufficient, as it is recognized that diseases and drug targets are connected to networks of proteins that regulate drug response. As a result, it has become quite difficult for traditional pharmacological approaches to yield promising drug candidates (1). Understanding how drug effects propagate from the site of action through the signaling network it regulates is a critical aspect of pharmacological development.

One aspect of systems biology is network modeling, which is a formalism that can be used to represent signaling networks. In a network, nodes represent the components of the network (e.g. proteins, genes, and small molecules) and the edges are the interactions between nodes. Edges can be directed (indicating the direction of information flow, for example from a regulator to a target) and can be positive (activating) or negative (inhibitory). Construction of a network allows for topological analysis, based on measures related to the structure of the network. Many measures have been developed to describe the features of networks based on their structure. For example, centrality measures have been used to identify the relative importance of a node within a network (2).

An additional layer of modeling is the dynamic model, which characterizes each network node with a state variable and describes how these states change in time due to the interactions among nodes. Both quantitative and qualitative models exist to describe the dynamics of systems. Quantitative dynamic models generally use systems of differential equations to describe
the continuous change of the system over time. These models can be highly accurate, but the limited kinetic and temporal information about individual nodes in the network limits their feasibility and widespread use, especially for large-scale systems (3).

Discrete dynamic modeling represents a class of qualitative dynamic models used to study signal transduction processes because of its computational feasibility and capacity to be constructed with qualitative biological data (4). The simplest kind of discrete models are Boolean models. In these models, nodes can have two qualitative states, ON (above an activity threshold) or OFF (below an activity threshold). The biological relationships defined by the structure of the network can be translated into mathematical equations using Boolean operators (5). Network-based discrete dynamic models can be used to generate testable hypotheses and are particularly useful in poorly characterized biological systems (6). Boolean network models have led to new insight into signal transduction and gene regulatory networks in numerous organisms (7-10).

In this chapter, we introduce discrete dynamic modeling in the context of systems pharmacology. We first describe how to reconstruct a network from available experimental information, which is often disparate and incomplete. We then describe how we can translate this network framework into a dynamic model that can predict the behavior of a biological system in response to some kind of signal (e.g. a drug) and in the presence of model perturbations (e.g. genomic or epigenomic alterations). We explain in detail how to construct Boolean dynamic models, including synchronous and asynchronous updating schemes and model reduction techniques. We further discuss how to analyze the dynamics of a system, including attractors, initial conditions, basins of attraction and the network’s state space. Based on these models, predictive and testable hypotheses can be obtained. As an example, we have constructed a Boolean and a multi-state discrete model of EGFR signaling and of the effect of an EGFR inhibitor on tumor growth in the context of different tumor mutations. We demonstrate that this model is consistent with existing knowledge and that it can be used to predict tumor response to drug treatment in the context of different mutations.

1.3 Constructing the Network

In order to develop a dynamic Boolean model, a network must be first constructed, which represents the players to be modeled and their interactions or regulatory relationships. Importantly, networks provide a unifying representation of heterogeneous biological
information. This allows diverse regulatory relationships from protein-protein interactions, post-translational modifications, to mRNA transcriptional changes and/or sub-cellular localization to be modeled in a unified language.

Networks are commonly constructed from pre-existing literature. Primary literature and review papers provide a valuable resource for identifying 1) the components that regulate a cellular process and 2) the qualitative regulatory relationships among these components. For example, a putative component might be added to a network if a known node in the network of interest alters the activity or expression of the putative network component. This information may be gleaned experimentally by studying the effects of activating mutations or over-expression of a network component. Specifically, if over-expressing the known network component leads to up-regulation or increased activity of the putative network component, it suggests that an activating relationship exists between the two nodes. However, if over-expressing the known node produces down-regulation or decreased activity of the component, then this suggests that an inhibitory relationship exists between the two nodes.

Determining the response of a gene or molecular entity after mutating or over-expressing a regulator of that gene provides genetic evidence for the involvement of the regulator in a signal transduction event. Over-expression vectors, the use of dominant negatives, RNA interference, and genome editing are common genetic manipulations to produce this kind of evidence. Chemical inhibitors may be used to provide pharmacological evidence about the relationship between two nodes. It is important to note that most chemical inhibitors block a specific interaction, whereas many genetic techniques (e.g. over-expression or knockdown) are producing node effects. It should also be noted that chemical inhibitors frequently have off-target effects, possibly affecting other network components, so these inhibitors should be used cautiously when trying to understand their effect on a network. Other evidence that may be used when constructing networks includes enzymatic activity, protein-protein interactions, post-translational modifications (11), transcription factor binding (12), and mRNA transcriptional changes (13). In addition to primary literature, curated databases of signaling pathways may also be of use in building networks (14). Software packages exist to aid in the rapid acquisition of model-producing information from databases like Pubmed (15, 16). One such software package, Chilibot, mines Pubmed for causal and parallel relationships between two or more gene or protein names (17, 18). Furthermore, methods have been developed to reverse engineer networks...
from high throughput expression-type datasets (18, 19).

<table>
<thead>
<tr>
<th>Nodes</th>
<th>Causal Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>I activates A</td>
</tr>
<tr>
<td></td>
<td>I activates B</td>
</tr>
<tr>
<td>A</td>
<td>A inhibits B</td>
</tr>
<tr>
<td></td>
<td>A activates O</td>
</tr>
<tr>
<td>B</td>
<td>B activates A</td>
</tr>
<tr>
<td></td>
<td>B inhibits O</td>
</tr>
<tr>
<td>O</td>
<td></td>
</tr>
</tbody>
</table>

Genetic, pharmacological, and biochemical evidence can be represented as component-to-component relationships such as “I activates A” (denoted as I → A) or “B inhibits O” (denoted by B ──| O), which correspond to directed edges from an upstream regulator to a downstream target in a graph representing a signaling network (Figure 1-1A and Figure 1-1B). In some situations, experimental evidence leads to double causal inferences such as “X activates the process through which Y activates Z”. In some cases these inferences can be logically broken down to two separate component-to-component relationships. Some experimental evidence, however, represents indirect causal relationships that are not easy to interpret. A method has been developed that uses techniques from combinatorial optimization to find the sparsest network consistent with all experimental observations (20). This method was implemented in the software NET-SYNTHESIS (http://www.cs.uic.edu/~dasgupta/network-synthesis). The input to NET-SYNTHESIS is a text file of positive or negative relationships among nodes in the biological network. The software produces a simplified network representation and a text file with the edges of the inferred signaling network (21).
1.4 Determining the Boolean Functions

The network representation of a signal transduction process is static, while biological processes happen over time. In order to understand the dynamic behavior of a system, each node needs to be characterized by a state variable that can change in time. Specifically, each node’s state variable is determined by the state variable of the nodes that regulate it. This dependency is expressed through a node-specific Boolean function. This function aggregates the effect of upstream nodes (activation or inhibition) on a regulated node as well as the relationships among the regulators for a given node. A popular way of representing Boolean functions is through the Boolean operators NOT, AND, OR. The Boolean operator NOT is used to specify inhibitory effects; AND and OR are used to define the relationships among regulators. If node Y has a single activator X, represented by $X \rightarrow Y$ and rooted in experimental evidence that X leads to activation or up-regulation of Y after a period of time, the state of node Y can be represented by the Boolean function $Y^* = X$. In this representation, the state of the nodes is represented by the node labels and $Y^*$ denotes the state of node Y at a future time point. The rule indicates that the future state of node Y will be equal to the current state of node X. Biologically, this means that a high level or activity of X will lead to high level or activity of Y. Nodes can be inhibited by upstream nodes (e.g. $X \rightarrow \neg Y$), indicating that the activation of the target node Y requires a low level or activity of the inhibitory node X. The effect of a single negative regulation is thus represented by the Boolean rule $Y^* = \neg X$. Here the NOT represents logical negation, where $\neg \text{ON} = \text{OFF}$ and $\neg \text{OFF} = \text{ON}$.

Cases of combinatorial regulation can be explored using the hypothetical Boolean relationships in Figure 1-1C. The Boolean function for the state of node A:

$$A^* = I \text{ OR } B,$$

uses the OR operator to represent that node A can be independently activated by node I or by node B. In general, an OR relationship represents the converge of two independent and individually effective pathways on a target node. The AND operator is used to represent situations where the synergistic activity of multiple regulators is required to activate the target node. For example, in Figure 1-1:

$$B^* = I \text{ AND } \neg A$$
indicates that node B will be active (ON) if there is a high level (ON state) of I and low level (OFF state) of A in the cell. Similarly to B, the output O of the hypothetical network is regulated by an activating node A and an inhibitory node B:

\[ O^* = A \text{ AND NOT } B \]

O is activated under the condition that A is ON and B is OFF. Under circumstances where more than two regulators exist for a node, the Boolean function can be comprised of a combination of AND, OR, and NOT operators, depending on the biological regulation involved.

Each Boolean function can also be represented by a truth table, which lists all possible future states of a node resulting from all possible states of its regulators. Each row in the truth table lists a combination of values of the Boolean variables of the node’s regulators and the associated output value (i.e. future state of the regulated node). “1” represents a node being ON and “0” represents a node being OFF. The truth table of a Boolean function with x regulators has \(2^x\) rows and \(x + 1\) columns. Figure 1-1D shows the truth tables corresponding to the Boolean functions of the four nodes in the network.

Boolean functions are usually determined based on known relationships in the published literature. If a node’s Boolean function cannot be unequivocally determined from the literature, then multiple Boolean functions can be created and tested against known biological outcomes. The equation that best recapitulates known network outcomes can then be used. Boolean rule construction could thus yield information about regulatory relationships that were not previously known. For example, if an AND rule between two regulators best recapitulates the known responses for a network, this suggests a putative biological synergy between the two regulators. Lastly, approaches have been developed to produce Boolean networks and functions from high throughput expression data (22-24).

1.5 Selecting Time Implementation for State Transitions

Boolean models and discrete dynamic models in general focus on state transitions instead of following the system in continuous time. Thus, time is an implicit variable in these models. After constructing the static signal transduction network and translating it into a system of Boolean functions, the next step is to choose an algorithm for the time implementation. Time implementation refers to the timing of transitions from one network state to the next.
Synchronous models are the simplest update method: all nodes are updated at multiples of a common time step based on the previous state of the system:

\[ X_i(t + 1) = F_i(X_1(t), X_2(t), \ldots, X_N(t)) \]

Here \( X_i(t) \) represents the state of node \( i \) at timestep \( t \) of a network with \( N \) nodes, given by \( i = 1, 2, \ldots, N \). The state of node \( i \) at a future timestep, \( X_i(t+1) \), is a function of the previous states of the nodes that regulate it. The synchronous model is deterministic in that the sequence of state transitions is definite for identical initial conditions of a model (25). This update method is valid if the time scales of all signal transduction processes represented as edges of the model are quite similar.

The kinetics of biological reactions and processes has been shown to vary substantially. For example, post-translational modifications like phosphorylation events can occur in thousandths of a second, whereas transcriptional events may take hundreds of seconds (26). Thus updating algorithms that account for differences in process timing must be used. In asynchronous models, the nodes are updated individually, depending on the timing information, or lack thereof, of individual biological events. In this section we discuss three asynchronous updating models: the random order asynchronous, general asynchronous, and deterministic asynchronous models.

In a random order asynchronous model, at each time step every node is updated in a random order chosen from all possible node orders with equal probability (27). In this model, the state of node \( i \) at the next time step, \( t+1 \), is obtained from the most recently updated states of its input nodes:

\[ X_i(t + 1) = F_i(X_1(\tau_{i1}), \ldots, X_N(\tau_{iN})) \]

where \( \tau_{ij} \in \{t,t+1\} \), for any node \( i \) and \( j \), where \( j = 1,2,\ldots, N \). If node \( j \) is before node \( i \) in the randomly chosen update order, then \( \tau_{ij} = t+1 \), otherwise, \( \tau_{ij} = t \). With this update method each node is updated once in each round of updating.

In the general asynchronous model (27), a single node is randomly updated at each time step. Under this approach, it is possible that a node chosen in the current time step will be chosen in a future time step. The unit of time in the general asynchronous model is \((1/N)\) of the time unit of the random order asynchronous model.

Both random order and general asynchronous models are stochastic, reflecting the variability in the timing of signal transduction events at the population level. If there is a priori knowledge
about the relative timescales over which biological processes in the signaling network of interest occur, it can be incorporated as a constraint of the updating scheme. This may also be accomplished through a deterministic asynchronous model. In this model each node $i$ is associated with an intrinsic time unit $\gamma_i$ and is updated at multiples of that unit, $t_i=k\gamma_i$ (28). At a time $t+1$, the node $i$ whose $t_i=t+1$ is updated and all other nodes remain in the same state:

$$X_i(t+1) = \begin{cases} F_i(X_1(t),\ldots,X_N(t)) & \text{if } t+1 = k\gamma_i \\ X_i(t) & \text{otherwise} \end{cases}$$

In the deterministic asynchronous model, nodes with longer time units will have less updates than nodes with shorter time units. This update mode is best suited for cases when the relative frequency or rate of biological events is known or can be estimated from biological knowledge. As this is a deterministic model, there is a guaranteed trajectory for state transitions.

### 1.6 Evaluating the Dynamics and Steady States of the System

Choosing a time implementation for a Boolean model allows one to explore the dynamics of the system; that is, how it changes over time. Initial conditions can be set to represent the initial biological phenotype of the system and any signals or environmental conditions that may cause the system to change over time. By varying the initial conditions, the effect of different initial conditions on an output can be modeled.

Initial conditions for a model can be set based on a priori biological knowledge. For example, in the hypothetical signal transduction network of Figure 1-1, if we are curious about the effect of the input (signal) $I$ on the output, $O$, and we know that prior to receiving the signal node $A$ is ON and node $B$ is OFF, we can use the initial conditions $I=ON$, $A=ON$, $B=OFF$, $O=OFF$, to study the predicted effects by the model. It is also good to compare this system to the case where the signal, $I=OFF$, to determine the state of $O$ when no signal is present. If the information is not sufficiently specific, multiple initial conditions can be used as starting points of replicate simulations. The percentage of replicate simulations where a node (e.g. the network’s output node) is ON can be calculated for each time step, representing the probability that a node will be activated or an outcome will occur given a stimulus. Additionally, patterns for initial conditions that lead to $O=ON$ versus $O=OFF$ can be explored, to determine the conditions that allow the signal to induce the output.
Starting from a plausible initial condition and updating the nodes’ states according to the Boolean rules in the model, the system’s state will change over time until it reaches a stable outcome called an attractor. An attractor can be a fixed point (steady state) or a set of states that repeat indefinitely (a complex attractor). The basin of attraction refers to the initial conditions that lead the system to a specific attractor.

There are several software applications to simulate the trajectory of the system from the initial condition to the attractor(s). BooleanNet (http://code.google.com/p/booleannet/) is a Python library that facilitates Boolean simulations of biological regulatory networks. This software requires a text file containing Boolean rules as input, and users can choose from various updating schemes for their models (29). BoolNet, SimBoolnet, and ChemChains are other software tools for Boolean modeling (30-32).

Under the hypothetical circumstance that nothing is known about the initial conditions of the signal transduction network in Figure 1-1, and that we are curious about the effect an input I on the output node, O, we can set I to ON, randomly set nodes A, and B, and have node O OFF in the initial conditions. Using BooleanNet, we can simulate this using a random order asynchronous model (29) and determine the fraction of O= ON in 100 replicate simulations as shown in Figure 1-2.

![Figure 1-2](image)

**Figure 1-2.** Fraction of simulations where O=ON as a function of time, given I=ON and nodes A and B are randomly set to ON or OFF in the initial condition for the Boolean model given in Figure 1-1 (100 replicates; random order asynchronous update).

Analysis of attractors is important in signaling networks because they represent the long term effect of a signal on an output. The attractors of gene regulatory networks that do not have
external signals represent different cell fates. We can analytically find all possible fixed points of a network by recalling that in the fixed point the future state of any node equals its current state. Thus the Boolean update functions become equations. The solutions of these equations are the fixed points of the system. If the system of equations does not have a solution, the model does not have fixed points, just one or several complex attractors. For the four node signaling network illustrated in Figure 1-1, if we assume I=0, then the Boolean equations simplify to A=0 OR B=B, B=0 AND NOT A=0, O =A AND NOT B. Substituting B=0 in the first and third equation we find the solution I,A,B,O = 0000. If we substitute this state back into the set of equations, we see the same result; thus 0000 is a fixed point of the system. If we assume I=1, then A=1 OR B = 1, B=1 AND NOT A=NOT A = NOT 1=0, and O=1 AND NOT 0 =1, thus we get the result 1101. Thus, there are two fixed points, one for each value of the signal. The state of the output node O equals the state of the input node I in each fixed point, which is representative of a response to a signal.

For a Boolean network with N nodes, there are $2^N$ possible states. In the absence of specific information, each of these states can be considered as an initial state, and the trajectory of the system, starting from this state and ending in one of the attractors, can be determined. A compact representation of all possible trajectories is given by the state transition network. The nodes of this network are the $2^N$ states of the system, and a directed edge is drawn from state $S_1$ to state $S_2$ if applying the Boolean rules to state $S_1$ transforms it into state $S_2$ (i.e. there is a state transition from $S_1$ to $S_2$). Figure 1-3 illustrates the state transition network for the Boolean network of Figure 1-1 using synchronous update. The state transition network indicates the two previously found steady states as nodes that do not have edges that point toward other nodes. It also indicates that all states in which the signal is OFF converge into the 0000 state and all states in which the signal is ON converge into the 1101 state.
Synchronous and asynchronous Boolean models have the same fixed points, because fixed points are independent of the implementation of time. However, the basins of attraction that lead to the fixed points may differ between synchronous and asynchronous models. Synchronous and deterministic asynchronous models have a definite trajectory between states, whereas stochastic asynchronous models may have multiple trajectories depending on the specific update sequence chosen. Specifically, each state can have up to $N$ successors when following a general asynchronous update, as any of the nodes of the system could be updated. In random order update, each state can have up to $N!$ successors. Thus, initial conditions that lead to a single attractor in deterministic models may enter multiple attractors in asynchronous models. Lastly, certain complex attractors may disappear when switching from synchronous to stochastic asynchronous models. This is because those attractors (called limit cycles) depend on multiple nodes changing states at the same time (33).

As an illustration, Figure 1-4 indicates the random order asynchronous state transition network of the system in Figure 1-1. As expected, the fixed points of the system are the same as in Figure 1-3. In this case the basins of attraction of the steady states remain the same as well, because they are determined by the state of the input node. Multiple trajectories from a single initial condition may exist in Figure 1-4, as this updating scheme is non-deterministic. State transition networks for both updating schemes are implemented in BooleanNet (29).
In large Boolean networks, it is computationally unfeasible to map the network’s state space (34). Methods have been developed to reduce the size of networks to make them easier to handle algorithmically, while maintaining the attractor repertoire of the larger network. Two methods have been developed that search for “frozen nodes,” which are nodes that stay constant regardless of the initial condition. Because these nodes do not change, they are irrelevant for differentiating attractors and thus can be removed from attractor and state space analyses (35, 36). Two other computational methods iteratively remove non-auto-regulatory components (i.e. nodes without a self-loop). These methods have been proven to conserve fixed points but may produce complex attractors that do not exist in the larger network (37, 38). Another reduction method involves a two-step process, which identifies nodes which stabilize due to the presence of a sustained signal, followed by removing simple mediators (nodes with a single incoming and outgoing edge) (39). Lastly, a network reduction approach has been developed that identifies
(possibly nested) feedback loops that have a defined steady state associated with them, then uses this steady state to simplify the network as in Saadatpour et al. (40).

1.8 Testing the Validity of the Dynamic Model

Discrete dynamic models are constructed to understand the dynamics of biological systems, or how a biological system might change or respond to a stimulus or signal. In order to explore unknown features of biological systems, it is first important to assess whether the model reproduces known features of the system being studied. In order to do this, one needs to study the long-term behaviors (e.g. steady states) of the system, as well as the intermediate states leading to them. If there is a baseline, unstimulated, or signal-free state in the system, it is important to determine whether that state exists in the dynamic model as a steady state. If the baseline condition does not exist, or is transient, the cause is probably an error in the Boolean logic (for example the use of OR instead of AND) or the incompleteness of the model.

Once a baseline condition is reproduced, it is important to check whether the model recapitulates a known relative order of events after a stimulus. If it does not, the erroneous state changes can be traced back and errors in the Boolean logic or incompleteness of the model may be discovered. It is also important to determine whether a biologically realistic steady state exists. If in addition to a biologically realistic steady state the model also indicates another steady state that has never been realized biologically before, the new steady state should not necessarily be discounted. It is possible that this steady state could represent a previously undiscovered biological outcome for the system. In these circumstances, the initial conditions and intermediate states that lead to this novel steady state can be explored and their biological feasibility can be determined.

1.9 Introducing Model Perturbations

Discrete dynamic models allow for the exploration of the effect of network perturbations on the dynamics of the biological system. Perturbations are the network equivalent of a deviation from biological homeostasis. Essentially all diseases are a deviation from a homeostatic or “normal” state. These perturbations may be the direct result of single or multiple mutations (e.g. cancer or genetic diseases). Other causes of biological perturbations include environmental influences like pharmacological use and abuse or traumatic injury. Certain environmental cues
may cause genetic aberrations so these modes of biological perturbations are not mutually exclusive. Discrete dynamic models can also be used to model the effect of pharmacological interventions, in the presence of no, single or multiple genomic abnormalities, and how these effects percolate through a network to produce a biological phenotype.

There are several ways to study the effect of perturbations on biological networks. Biological over-expression or over-activity of a node can be modeled by keeping its state as ON irrespective of its regulators. Biological knockouts or down-regulation of a node can be modeled by keeping the state of a node as OFF, irrespective of the state of its regulators. The effect of these perturbations can be determined by performing simulations or state space analysis in their presence, and comparing perturbed outcomes to their normal states. Drug or growth factor treatments can be simulated by inhibiting or activating certain nodes transiently (setting the node state and allowing it to update based on its biological regulation) or permanently (over-riding the existing rules). BooleanNet has built-in functions to facilitate network perturbations in Boolean models (29).

Perturbations to discrete dynamic models can facilitate the efforts of systems pharmacology by facilitating the development of personalized networks representing individuals in disease. We can model the effect of drugs in the presence of individual network differences (i.e. individual genetic differences) and can predict the effect of pharmacological intervention in specific individuals. Discrete dynamic models can be used in these ways to choose pharmacological interventions that will work for specific patients and to avoid treatments that would have no benefit.

1.10 Non-Boolean Discrete Dynamic Modeling (Multi-State Discrete Models)

Until this point we have discussed the simplest kind of discrete dynamic model, the Boolean model. Multi-state discrete dynamic models can also be constructed to analyze biological systems and in certain cases may be more appropriate than Boolean models. As with Boolean models, truth tables can be constructed to represent the regulatory relationships among nodes. For example, in a three-state model, nodes can be assigned three states (e.g. -1,0,1 or 0,1,2) to represent under-activity (down-regulation), normal activity, and over-activity (up-regulation). The value of a regulated node will depend on the logical constraints designated by the modeler in the truth tables.
There are alternative mathematical formalisms for a more compact representation of multi-state discrete models: logical models (41), implemented in the software GINSim (42) and polynomial dynamical systems (43) implemented in the software ADAM (44). The logical functions used in the logical model framework specify the conditions for which the regulated node’s state is different from the baseline (41). Polynomial dynamical systems represent each truth table by a polynomial function. Polynomial algebra can then be used to identify the steady states of the model. One disadvantage of this method is that the polynomial representation (even for Boolean models) is less intuitive than other forms, e.g. logical models (43).

1.11 Conclusions

Very frequently the kinetic details of biological processes are not known because they are very difficult to measure experimentally. Discrete dynamic modeling is formidable in the absence of such details. Here we constructed a discrete dynamic model of EGFR signaling and compared the results to an established ODE model of EGFR signaling, which was constructed using experimentally derived parameters. Our results demonstrate that discrete dynamic models are highly consistent with ODE models without the requirement of difficult to obtain and often unavailable experimental parameters. Our results support that discrete dynamic modeling is a formidable approach in systems where experimental parameters are unavailable, which is very common in biological systems.

1.12 Acknowledgments

This work was partially supported by grants NIH Ruth L. Kirschstein National Research Service Award (F30 DK093234) and NSF PHY-1205840. This work was adapted from a book chapter: “Steinway SN, Wang RS, Albert R. Discrete dynamic modeling: A network approach for systems pharmacology. Systems Pharmacology and Pharmacodynamics. In Press.”

1.13 References

CHAPTER 2.
Discrete Dynamic Modeling of EGFR Signaling in Tumor Growth: A Comparison to Ordinary Differential Equation Modeling

2.1 Introduction

Previously an enhanced pharmacodynamic (ePD) model was constructed by Iyengar et al. to model the effects of an EGFR inhibitor (e.g. gefitinib) on \( EGFR^+ \) tumor growth (1). This model was novel because it merged traditional PK/PD models, which focus on the site of drug action (EGFR protein) with ordinary differential equation models of signaling events downstream. This is important, essentially encapsulating the goals of systems pharmacology, because many factors beyond the direct effect of the drug on the receptor contribute to the effect of a drug in treating disease. As with other differential equation based models, the EGFR model used a large number of parameters, whose value was estimated based on experiments reported in the literature. The model demonstrated that three genes (\( RASAL1 \), \( PEBP1 \), and \( miR221 \)), all of which are downstream of the site of drug action (EGFR), affect whether a tumor responds to gefitinib treatment. The model also quantified how sensitive a tumor is to gefitinib; whether the tumor was resistant, went into a partial remission, or complete remission (1). As an illustration, we translate this ePD model of EGFR inhibition into two types of discrete dynamic models: 1) a Boolean model and 2) a three-state discrete model. Furthermore, we demonstrate the efficacy of discrete models at reproducing the results obtained from the ePD model.

2.2 A Boolean Model of EGFR Inhibition in Tumor Growth

The ePD model represents the EGFR signaling pathway as a linear pathway augmented with a feed-forward motif from EGFR to Raf1 (Figure 2-1A). Active EGFR signaling leads to the induction of proliferation (output node). Tumor cell proliferation regulates tumor growth and thus tumor size. Since we have the network structure, we can move on to creating Boolean rules for each
Most rules are simple to determine: nodes with a single input (regulator) can be represented as Node*=Regulator. For example, from Figure 2-1A, the rule for Pkc is the following: Pkc*=Ply. There are three nodes with multiple regulators, Kras, Raf1, and Cdk4/6. As discussed in Section 5.2.2, most biological regulators act independently of one another and can be combined by an OR rule; however, it is important to understand the biology behind the network and to determine if synergistic relationships exist between regulators of the same node (AND rule). In the case of the well-studied EGFR signaling network, the only evidence of biological synergy occurs for the Cdk4/6 node, indicating that p27kip binds to CyclinD and prevents its interaction with Cdk4, inhibiting cell proliferation (2). Thus, Cdk4/6*=CyclinD AND NOT P27kip (Figure 2-1A).
Using the reduction of simple mediator nodes as described in Section 5.2.5 and previously (3), we reduced this network to an equivalent 7 node and 7 edge network by collapsing nodes with an in-degree and out-degree of 1 (Figure 2-1B). For example, EGFR → Pley → Pkc –|Rkip was collapsed to EGFR –|Rkip. We subsequently substituted upstream regulators into rules for downstream nodes. For example, Rkip* = NOT EGFR would be produced from collapsing the previously described rules. Using these methods, we constructed a set of Boolean equations for the reduced EGFR network (Table 2.1).

In the ePD model of an EGFR inhibitor, perturbations in three different gene products were the focus. RasGAP and miR221 were assumed to have three states: over-active, normal, and under-active. Rkip was assumed to have two states: normal and over-active (see Table 2.2). In the ePD model, combinations of these mutations were integrated with the effect of an EGFR inhibitor to explore individual patients’ response to treatment. As examples, R-,K0,M+ patients were predicted to have EGFR inhibitor resistant tumors, R0,K+,M- led to tumors that are EGFR inhibitor sensitive (full remission), and R0,K0,M0 tumors were predicted to lead to a partial remission phenotype.

In order to test the fidelity of our Boolean model at reproducing these results we have to make some assumptions and simplifications. The ePD model had three possible states for Rasgap and miR221 (e.g. R-,R0, and R+ for Rasgap) and also three possible model outputs: resistant, partial remission, and full remission; however, Boolean models can only handle two states. Let us first make the assumption that there are two model outputs: proliferation and no proliferation, which correspond to resistance to EGFR inhibitor treatment and remission (response to EGFR inhibitor treatment), respectively. Now we have to determine how to handle the Rkip and Rasgap states. In order for the model to reproduce tumor cell proliferation when EGFR signaling is active (EGFR=ON) in the absence of EGFR inhibitor and no mutations, normal miR221 needs to correspond to the ON state, and thus decreased miR221 is the OFF state. The increased miR221

<table>
<thead>
<tr>
<th>Table 2-1. Boolean Rules for the Reduced (7 node) EGFR Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rkip* = NOT EGFR</td>
</tr>
<tr>
<td>Kras* = EGFR OR NOT Rasgap</td>
</tr>
<tr>
<td>Raf1* = Kras OR NOT Rkip</td>
</tr>
<tr>
<td>Proliferation* = Raf1 AND miR221</td>
</tr>
</tbody>
</table>
level cannot be distinguished from the normal level in a Boolean model so \( M_0 = M^+ = \text{ON} \). If we assume that the normal Rasgap state (\( R_0 \)) is closer to the \( R^+ \) state; that is, that under normal circumstances, Rasgap suppresses EGFR signaling and tumor cell proliferation, then we can say \( R_0 = R^+ = \text{ON} \).

<table>
<thead>
<tr>
<th>Gene</th>
<th>Symbol: Perturbation</th>
<th>Normal function</th>
<th>Effect on tumor growth</th>
</tr>
</thead>
</table>
| RASAL1 (Rasgap) | R\(-\): hypermethylation  
R0: normal  
R+: hypomethylation | Suppresses Ras signaling (catalyzes Ras-GTP to Ras-GDP) | Suppresses tumor growth by inhibiting MAPK signaling (Ras) |
| PEBP1 (Rkip) | K0: Responsive to PKC  
K+: Non-responsive to PKC | Rkip Inhibits Raf (MAPK signaling).  
Pkc inhibits Rkip. | Suppresses tumor growth by inhibiting MAPK signaling (Raf) |
| miR221 | M\(-\): decreased miR221  
M0: normal miR221  
M+: increased miR221 | miR221 suppresses p27. p27 suppresses CDK4. | \( \text{miR221} \) induces tumor growth |

Using the above assumptions for converting node perturbations for Rasgap (\( R_0 = R^+ = \text{ON} \)) and miR221 (\( M_0 = M^+ = \text{ON} \)) we can determine the steady state of the Proliferation node for a Boolean model for each tumor genotype by substituting these values into the Boolean equations or by using BooleanNet. We can then compare with the results of the ePD model by making a correspondence between proliferation and a resistant tumor, and between the absence of proliferation and full remission (Table 2.3).

The agreement between the ePD and Boolean model of an EGFR inhibitor is quite strong. We see that 100% (4 out of 4) of the tumors that are resistant in the ePD model undergo proliferation in the Boolean model and that 67% (4 out of 6) of tumors that achieve full remission in the ePD model undergo no proliferation in the Boolean model. Interestingly, even though there is no partial remission category in the Boolean model, it seems that the Boolean model stratifies into proliferation versus no proliferation groups right in the middle of the partial remission category. Assuming that the Boolean model is correct if it attains the proliferation outcome to the left of this mark and the no proliferation outcome to the right of this mark, then the model correctly predicts the ePD outcomes in 83% of the cases.
2.3 A Multi-State Discrete Model of EGFR Inhibition in Tumor Growth

The ePD model of EGFR inhibition has some non-Boolean discrete properties:

1.) The model output was classified into three discrete categories: resistant, partial remission, or complete remission, depending on the tumor size (Table 2.3).

2.) Rasgap and Rkip have three states: under-active, normal, and over-active (Table 2.2)

Therefore, a multi-state discrete model may better reproduce the results of this model. Because the model output was assigned three categories: resistant, partial remission, or complete remission, we can assign the “Proliferation” node the same three states: 1 (remission), 0 (partial remission), and 1 (resistant). Similarly, the states of Rasgap and miR221 could be assigned three states depending on activity: 1 (under-active), 0 (normal), and 1 (over-active). The states of EGFR and Rkip can remain as Boolean (two-state) variables.

We configure the truth tables so that they preserve or refine the Boolean relationships indicated on Table 2.1. For example, the Boolean rule for Kras indicates that the presence of EGFR or the absence (OFF state) of Rasgap can lead to the activation of Kras. The refined truth table will have three states for Rasgap and Kras. We assume that the presence of EGFR combined with a low or intermediate state of Rasgap leads to a high activity (state 1) of Kras, as does the under-activity (state -1) or Rasgap. We assume that the absence of EGFR combined with over-activity of Rasgap leads to low activity (state -1) of Kras. The remaining two conditions are assumed to lead to an intermediate Kras activity (see Figure 2-2). The resulting model recapitulates all but two findings of the ePD model, or almost 89% of ePD model findings (Table 2.3). In fact, the R0,K+,M+ case is on the border between the partial and complete remission categories and would be consistent if a slight change were made in the tumor mass threshold between these two categories. Furthermore, we can stratify our results into three
categorical outcomes instead of just two in the Boolean model, which is substantially more informative. Lack of intermediate categories like "partial remission" is one of the major limitations of Boolean models. Using multi-state models, intermediate states can be recapitulated.

<table>
<thead>
<tr>
<th>Egfr</th>
<th>Rkip</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Raf1</th>
<th>miR221</th>
<th>Proliferation*</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>-1</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>0</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 2-2. Truth tables for the multi-state discrete model of Egfr inhibition on tumor growth. The Egfr and Rkip nodes have two states. Rasgap, Kras, Raf1, miR221, and Proliferation have 3 states. The symbol “*” marking a node’s name indicates the future state of the node.

2.4 Conclusions

As an example of how discrete dynamic modeling can be used in systems pharmacology, we translated a pre-existing ePD model of an EGFR inhibitor and its effect on tumor growth in the context of various tumor mutations. The results demonstrated that even in the absence of rigorous experimental detail (e.g. kinetics and concentration information) that was required to create the ePD model, our discrete dynamic model was able to recapitulate almost 90% of the ePD model results. This is a testament to not only the utility of discrete models but the importance of the structure of a biological network over the kinetic details of the individual processes in determining the network’s function.

In the post-genomic era, we have largely realized that there are major genomic differences among humans and that these individual differences are likely very important contributors to
disease and disease susceptibility. The major issue now is how we use this information to make smarter decisions about who we treat and how we treat them. Systems pharmacology offers a framework to approach these issues. Discrete dynamic modeling is a critical means to understand the network effects of a drug, how and why it might work in the context of differences in individual networks and which drugs might work best in different individuals. The era of personalized medicine is upon us. Discrete dynamic modeling is and will continue to be a powerful tool to systems pharmacology and personalized medicine in this new age of pharmaceutical intervention.

2.5 Acknowledgments

This work was partially supported by grants NIH Ruth L. Kirschstein National Research Service Award (F30 DK093234) and NSF PHY-1205840. This work was adapted from a book chapter: “Steinway SN, Wang RS, Albert R. Discrete dynamic modeling: A network approach for systems pharmacology. Systems Pharmacology and Pharmacodynamics. In Press.”

2.6 References

CHAPTER 3.
Network Modeling of TGFβ Signaling in Hepatocellular Carcinoma Epithelial-to-Mesenchymal Transition Reveals Joint Sonic Hedgehog and Wnt Pathway Activation

3.1 Abstract
Epithelial-to-mesenchymal transition (EMT) is a developmental process hijacked by cancer cells to leave the primary tumor site, invade surrounding tissue, and establish distant metastases. A hallmark of EMT is the loss of E-cadherin expression, and one major signal for the induction of EMT is transforming growth factor beta (TGFβ), which is dysregulated in up to 40% of hepatocellular carcinoma (HCC). We have constructed an EMT network of 70 nodes and 135 edges by integrating the signaling pathways involved in developmental EMT and known dysregulations in invasive HCC. We then used discrete dynamic modeling to understand the dynamics of the EMT network driven by TGFβ. Our network model recapitulates known dysregulations during the induction of EMT and predicts the activation of the Wnt and Sonic hedgehog (SHH) signaling pathways during this process. We show, across multiple murine (P2E and P2M) and human HCC cell lines (Huh7, PLC/PRF/5, HLE, and HLF), that the TGFβ signaling axis is a conserved driver of mesenchymal phenotype HCC and confirm that Wnt and SHH signaling are induced in these cell lines. Furthermore, we identify by network analysis eight regulatory feedback motifs that stabilize the EMT process and show that these motifs involve cross-talk among multiple major pathways. Our model will be useful in identifying potential therapeutic targets for the suppression of EMT, invasion and metastasis in HCC.

3.2 Quick Guide to Equations and Assumptions
3.2.1 Major Assumptions of the Model
The Boolean modeling approach is an abstraction that aims to explain the qualitative characteristics of dynamic systems. The activation of a biological node is represented as a switch in a Boolean model; the node can be either ON or OFF and there are no intermediate levels of activation. This is a reasonable assumption when a biological node exhibits two qualitatively different states and the transition between these two states is switch-like.

The regulatory relationships among nodes can be described with the Boolean operators AND, OR, and NOT. If a node has a single regulator then the regulated node’s state is solely
determined by its upstream node’s state. OR represents the combined effect of independent and individually effective upstream regulators on a downstream node, whereas AND indicates the joint requirement for multiple upstream regulators to achieve a downstream effect. NOT represents the effect of inhibitory regulators and can be combined with other regulation by using either OR or AND. Our baseline assumption is that multiple regulators of the same target node act independently of each other. However, if there is evidence of biological synergy (e.g. two proteins form a complex to perform a biological function), then an AND operator was used. We describe three Boolean equations in the section.

We do not know the kinetics of individual events, although we do know that signal transduction events occur substantially faster (seconds or faster) than transcriptional events (timescale of minutes) (31). Thus, we employ a stochastic asynchronous updating scheme with a ranking system. We randomly select a single node and update its state, meaning that we re-calculate its state according to its Boolean equation. In order to account for signal transduction events occurring faster than transcriptional events, we update nodes regulated by signal transduction events with a greater probability than nodes regulated by transcriptional events. The results are robust to changes in the value of this probability (see Figure 3-1). By performing a large number (10,000) of simulations with a stochastic update sequence, we identify the behaviors that do not depend on the detailed kinetics of the individual reactions. We define a time step as the average number of updates needed to update a transcriptionally regulated node.

Dynamic Boolean models do not require kinetic parameters, which makes them more computationally feasible than quantitative models, especially for biological systems with more than a dozen components. Boolean models assume that the structure of the biological networks they describe is more important than the kinetics of individual reactions and acquire their richness through the large number of interactions that are included in the networks they describe.

3.2.2 Key Equations
Figure 3-1. The ratio between the timescale of the slow and fast processes does not matter as long as it is significantly larger than one. For the ranked general asynchronous updating scheme used, a range of node update probabilities for fast (signal transduction) versus slow (transcription) biological events were tested. Our results are robust to changes in update probabilities as long as the ratio of update probabilities for signal transduction to transcriptional events is significantly greater than one. An update probability ratio of 5 means that signal transduction events are five times as likely as transcription level events (left panel). An update probability ratio of 50 means that signal transduction events are fifty times as likely as transcription level events (right panel).

An explanation of the Boolean regulatory functions of two key nodes is included below. Each equation indicates the future state (shown by the use of the asterisk) of a regulated node as a function of the current states of the nodes that regulate it (i.e. nodes that are the starting points of edges incident on the regulated node). For simplicity we denote the node state by the node name, thus “node X” in a rule means “the state of node X”, and “not node X” means “the state that is the opposite of the state of node X”.

We start by a hypothetical example describing the transcriptional regulation of a target gene by a transcriptional activator (A) and a transcriptional inhibitor (I). We denote the transcript of the target gene as T.

\[ T^* = A \text{ and not } I \]  \hspace{1cm} (Eq. 1)

The Boolean equation indicates the condition under which the future state of the target gene’s transcript is ON. The right hand side of the equation specifies that this condition is that the transcriptional activator A is ON and simultaneously the transcriptional inhibitor I is OFF. If this condition is not satisfied, then T decays (gets turned OFF), even if it was ON before.

\[ \beta\text{-catenin\textunderscore nuc}\text{*} = \text{not Dest\_compl and not } \beta\text{-catenin\textunderscore memb and (not SUFU or not E-cadherin)} \]  \hspace{1cm} (Eq. 2)

\[ \beta\text{-catenin\textunderscore memb}\text{*} = \text{E-cadherin and not } \beta\text{-catenin\textunderscore nuc} \]  \hspace{1cm} (Eq. 3)
β-catenin_memb is a node representing membrane-bound β-catenin (CTNNB1). β-catenin is degraded due to an active “destruction complex” (represented by the Dest_compl node) formed by APC, AXIN1 or AXIN2, and GSK3β. The destruction complex targets free (cytoplasmic and nuclear) β-catenin for degradation (50); thus, the destruction complex can suppress nuclear β-catenin (Eq.2) (51). The pool of nuclear (β-catenin_nuc node) and membrane bound β-catenin (β-catenin_memb node) is dynamically controlled by the Wnt signal: active Wnt signaling leads to β-catenin localization to the nucleus. Hence, we assume that β-catenin cannot be simultaneously localized both to the membrane and to the nucleus and implement this assumption as a mutually inhibitory relationship between β-catenin_memb and β-catenin_nuc (52). We consider an alternative rule for β-catenin_nuc in which there is no inhibition from β-catenin_memb to β-catenin_nuc in Figure 3-2. Suppressor of fused homolog (SUFU) binds β-catenin, exports it from the nucleus, and thereby represses nuclear β-catenin (Eq. 2) (50). E-cadherin has been shown to suppress β-catenin nuclear localization (Eq. 2) (51) and to stabilize membrane-bound β-catenin (Eq. 3) (52). We assume that predominant nuclear localization of β-catenin is possible if either SUFU or E-cadherin is absent, if simultaneously the destruction complex is also absent.

3.3 Introduction

Hepatocellular carcinoma (HCC) is the third most common cancer type worldwide (1). The only available cure for HCC is surgical resection, which is only an option in early stage disease. Death from HCC is primarily due to late stage disease, which is characterized by invasion, intran- and extra-hepatic metastasis, and post-surgical recurrence (2). The high mortality and frequency of recurrence in HCC make the need for novel therapeutic approaches paramount. HCC invasion is dependent upon a cellular process called epithelial-to-mesenchymal transition (EMT) (3). Therefore, targeting invasion and the EMT process could reduce HCC mortality substantially.

EMT is a crucial developmental event in embryonic tissue formation and regeneration (4). Pathological forms of EMT are hypothesized to occur during cancer progression and tissue fibrosis (5). During EMT epithelial cells lose their adhesive properties and transform into mesenchymal cells with a migratory phenotype. This provides a critical mechanism in the initial steps of metastasis, allowing cancer cells to leave the primary tumor. Loss of E-cadherin
CDH1, a cell-cell adhesion protein responsible for intercellular attachment, is widely considered the hallmark of EMT (6).

Transforming growth factor-β (TGFβ1, hereafter referred to as TGFβ) signaling has a dual role in tumor progression; it is tumor suppressive in normal epithelia and early stage tumor cells, contrasted by its stimulation of EMT as cancers progress (7). It is thought that genetic alterations
to the cell contribute to the role switch during tumor progression (8). Previous studies have demonstrated that TGFβ levels are elevated in 40% of HCC patients (9), while receptor levels are variable (10, 11).

TGFβ binds to TGFβRII, which recruits TGFβRI, leading to receptor phosphorylation and phosphorylation of SMAD2 and SMAD3. SMAD2 and SMAD3 form a complex with SMAD4 (in the canonical pathway), which translocates to the nucleus, where the complex down-regulates the transcription of epithelial junction proteins while increasing the expression of mesenchyme-related proteins (6). TGFβ can also activate PI3K/AKT, MAPK and other established oncogenic signals (non-canonical pathways) (12).

EMT and TGFβ signaling are complex biological processes that have been extensively studied using reductionist approaches. Network modeling provides a set of tools to study complex processes holistically, using an integration of individual and pairwise evidence (13). A network is a representation of a complex system; the nodes represent entities (proteins, mRNA) or concepts (outcomes) and the edges represent the relationships between nodes. Edges can be assigned a direction, representing information or signal flow from the source (upstream) node to the target (downstream) node. Furthermore, edges can be characterized as positive (activating relationship) or negative (inhibitory relationship). An additional layer of network analysis is the dynamic model, which characterizes each network node with a state variable. The biological relationships incorporated in the network can be translated into mathematical equations that summarize the regulation of each node. The dynamic model describes how the nodes’ state variables change in time, for example how a signal activates signal transduction and leads to the transcription of target genes (14).
In this study, we explore TGFβ-driven EMT in HCC by an integrated experimental and computational approach (Figure 3-3). We found that TGFβ signaling is a conserved driver of EMT across multiple HCC models. We developed a network model of EMT signaling and then translated this network into a discrete dynamic model. The dynamic model recapitulates the

Figure 3-2. Exploration of alternative rules to describe nuclear localization of β-catenin. β-catenin is either localized predominantly at the membrane (β-catenin_mem node) or in the nucleus (β-catenin_nuc node), depending on Wnt signaling. One possibility to ensure this is to assume a mutual inhibition between β-catenin_mem and β-catenin_nuc (A). In this case β-catenin_nuc activity starts increasing and β-catenin_mem activity starts decreasing after four time steps. E-cadherin activity starts decreasing shortly after β-catenin_mem and decreases below 0.5 in seven time steps. A less stringent alternative is to assume that β-catenin_nuc inhibits β-catenin_mem, but not vice versa (B). In this case the decrease of the β-catenin_mem activity is still accompanied by the increase of the β-catenin_nuc activity, thus this rule also reflects the localization of β-catenin. The increase of the Wnt activity is identical in both cases since Wnt is upstream of β-catenin. In contrast, the inter-regulation among β-catenin and E-cadherin causes a more rapid decrease in β-catenin_mem and E-cadherin activity and a more rapid increase of β-catenin_nuc activity. E-cadherin activity now decreases below 0.5 in five time steps. We used the former assumption in the model.
known steps of TGFβ-driven EMT. The model predicts, and our experiments confirm, that joint activation of Wnt and SHH signaling by TGFβ is a conserved feature of TGFβ signaling in EMT. Our analysis suggests a putative reliance on these pathways for robust activation of the EMT program.
3.4 Materials & Methods

3.4.1 Cell culture

P2E and P2M cells were cultured in DMEM:F12 1:1 medium (Mediatech) supplemented with 10% fetal bovine serum (FBS) as described (15). Huh7 cells were provided by Dr. Harriet Isom (Penn State College of Medicine) and were cultured in DMEM + 10% FBS medium (16). Human HLE cells were acquired from Dr. Curtis Harris (National Cancer Institute) and cultured in DMEM + 10% FBS. HLF cells were acquired from Dr. Jorge Filmus (University of Toronto) and cultured in DMEM + 10% FBS (17). PLC/PRF/5 cells were acquired from the American Type Culture Collection and cultured in Eagle's Minimum Essential Medium + 10% FBS. TGFβR inhibitor treatments were completed with LY2157299 (Selleck Chemical).

3.4.2 Cell Line Authentication

Huh7, HLE, and HLF cell lines were authenticated using short tandem repeat DNA profiling (Genetica DNA laboratories) on January 20th, 2014. PLC/PRF/5 cells were received directly from ATCC (July, 2013) and cultured for under three months so no authentication was deemed necessary.

3.4.3 Quantitative Real-Time Polymerase Chain Reaction

Real-time PCR experiments were conducted as described previously (18). The housekeeping gene GAPDH was used for ΔΔCt calculations and relative expression was calculated using Taqman primer/probe sets (Life Technologies).

3.4.4 Western Blot Analysis

Lysates were harvested with lysis buffer as described previously (18). 50 µg of protein lysates were separated on a NuPAGE 4-12% Bis-Tris Gel (Invitrogen) and transferred to polyvinylidene difluoride (PVDF) membrane (Invitrogen). Primary antibodies for E-cadherin (Cell Signaling), pSMAD2 (Cell Signaling), pSMAD3 (Cell Signaling), SMAD2 (Cell Signaling), SMAD3 (Cell Signaling), β-actin (Cell Signaling), GLI2 (Santa Cruz), and rabbit and mouse secondary antibodies (Cell Signaling) were used for this study. Signals were detected using Clarity chemiluminescence substrate on a ChemiDoc XRS+ imaging system (Bio-Rad).
3.4.5 Microarray Gene Set Enrichment Analysis (GSEA)

Murine HCC P2E and P2M cell line microarray datasets were acquired from NCBI Gene Expression Omnibus (15). Gene set enrichment analysis was performed to determine gene sets representing pathways and phenotypes enriched in P2M relative to P2E cell lines using datasets in the Molecular Signatures Database 4.0 (19). GSEA compares gene expression in biological samples by ranking the genes that are differentially expressed between two or more phenotypes. A running score is calculated by moving down the ranked gene list and increasing the score when a gene in the ranked list is in the enriched gene set. The score is decreased when a gene is not in the list. The enrichment score represents the maximum deviation from zero encountered while walking the gene list. A positive enrichment score means that the phenotype is over-represented in the gene list and a negative enrichment score means the phenotype is under-represented in the gene list. Statistical significance is determined using an empirical phenotype-based procedure, which produces a p-value. A database of established gene sets exists in the Molecular Signatures Database 4.0 (http://www.broadinstitute.org/gsea/msigdb/index.jsp).

3.4.6 Patient Samples

Human HCC samples were acquired from liver biopsy at Penn State Hershey Medical Center. Patients were on no cancer treatment at acquisition and were diagnosed as primary HCC. Specimens were flash frozen, and protein was extracted by homogenization of 100 mg liver tissue in 1 ml extraction buffer (150 mM NaCl, 50 mM Tris/HCl pH 7.4, 1 mM EDTA, 1% Triton 100, 0.1% SDS, 1% sodium deoxycholate, 1 mM sodium pyrophosphate, 10 mM β-glycerolphosphate, 10 mM NaF, 0.5 mM NaVO4, 1 µM Microcysteine-LR, and Protease inhibitor Cocktail Tablet without EDTA (1 tablet in 10 ml lysis buffer). Samples were incubated on ice for 30 minutes and vortexed every 5 minutes. Samples were centrifuged at 15,000 rpm at 4°C for 15 minutes. Protein (supernatant) was collected and quantified using BCA Protein Assay (Thermo Scientific).

3.4.7 Network Construction

Our strategy was the following: we knew that TGFβ is a prominent signal driving epithelial-to-mesenchymal transition in hepatocellular carcinoma, so we started with synthesizing a network for this process from the literature. Then we added the other known inducers of EMT
and other processes and components that are known to be deregulated during EMT. This resulted in a network of 70 nodes and 135 edges, which we call the EMT network (Figure 3-7 and Figure 3-3). Table 3-1 indicates the edges of this network, giving the source (upstream) node, the target (downstream) node, whether the edge represents an activating (positive) or inhibitory (negative) relationship, whether the edge corresponds to a direct physical or chemical interaction, and references where the relationship was reported.

Table 3-1. Evidence for interactions among nodes in the EMT network

<table>
<thead>
<tr>
<th>Source (upstream) node</th>
<th>Target (downstream) node</th>
<th>Interaction</th>
<th>Direct Interaction</th>
<th>Ref. (PMID)</th>
<th>Relevant information</th>
</tr>
</thead>
<tbody>
<tr>
<td>AKT</td>
<td>GSK3β</td>
<td>Inhibitory</td>
<td>N</td>
<td>21837363</td>
<td></td>
</tr>
<tr>
<td>AKT</td>
<td>IKKα</td>
<td>Activating</td>
<td>N</td>
<td>16207722</td>
<td></td>
</tr>
<tr>
<td>Dest_Compl</td>
<td>GSK3β</td>
<td>Inhibitory</td>
<td>Y</td>
<td>19536106</td>
<td></td>
</tr>
<tr>
<td>Dest_Compl</td>
<td>β-catenin_nuc</td>
<td>Inhibitory</td>
<td>Y</td>
<td>22935447</td>
<td></td>
</tr>
<tr>
<td>AXIN2</td>
<td>AXIN2</td>
<td>Activating</td>
<td>Y</td>
<td>23892894</td>
<td></td>
</tr>
<tr>
<td>AXIN2</td>
<td>Dest_compl</td>
<td>Activating</td>
<td>Y</td>
<td>20736375</td>
<td>AXIN2 is part of the destruction complex (Dest_compl node), which is a protein complex containing APC, AXIN2, and GSK3β proteins. Membrane-bound β-catenin opposes the activity of nuclear β-catenin</td>
</tr>
<tr>
<td>β-catenin_memb</td>
<td>β-catenin_nuc</td>
<td>Inhibitory</td>
<td>Y</td>
<td>20736375</td>
<td></td>
</tr>
<tr>
<td>β-catenin_memb</td>
<td>E-cadherin</td>
<td>Activating</td>
<td>Y</td>
<td>22007144</td>
<td></td>
</tr>
<tr>
<td>β-catenin_nuc</td>
<td>Dest_compl</td>
<td>Activating</td>
<td>Y</td>
<td>20736375</td>
<td></td>
</tr>
<tr>
<td>β-catenin_nuc</td>
<td>TCF/LEF</td>
<td>Activating</td>
<td>Y</td>
<td>21539828</td>
<td></td>
</tr>
<tr>
<td>β-catenin_nuc</td>
<td>AXIN2</td>
<td>Activating</td>
<td>Y</td>
<td>22874762</td>
<td>Nuclear β-catenin opposes the activity of membrane-bound β-catenin</td>
</tr>
<tr>
<td>β-catenin_nuc</td>
<td>β-catenin_memb</td>
<td>Inhibitory</td>
<td>Y</td>
<td>21539828</td>
<td></td>
</tr>
<tr>
<td>β-catenin_nuc</td>
<td>SNAI2</td>
<td>Activating</td>
<td>Y</td>
<td>14623871</td>
<td></td>
</tr>
<tr>
<td>βTrCP</td>
<td>SNAI1</td>
<td>Inhibitory</td>
<td>Y</td>
<td>20305697</td>
<td></td>
</tr>
<tr>
<td>CD44</td>
<td>cMet</td>
<td>Activating</td>
<td>Y</td>
<td>10485493</td>
<td></td>
</tr>
<tr>
<td>CDC42</td>
<td>PAK1</td>
<td>Activating</td>
<td>N</td>
<td>18940914</td>
<td></td>
</tr>
<tr>
<td>c-fos</td>
<td>c-fos</td>
<td>Activating</td>
<td>Y</td>
<td>2839774</td>
<td></td>
</tr>
<tr>
<td>c-fos</td>
<td>EGR1</td>
<td>Activating</td>
<td>N</td>
<td>16170359</td>
<td></td>
</tr>
<tr>
<td>CHD1L</td>
<td>CDC42</td>
<td>Activating</td>
<td>Y</td>
<td>20335658</td>
<td></td>
</tr>
<tr>
<td>cMet</td>
<td>SOS/GRB2</td>
<td>Activating</td>
<td>Y</td>
<td>10593929</td>
<td></td>
</tr>
<tr>
<td>cMet</td>
<td>SRC</td>
<td>Activating</td>
<td>Y</td>
<td>21119598</td>
<td></td>
</tr>
<tr>
<td>Csl</td>
<td>HEY1</td>
<td>Activating</td>
<td>Y</td>
<td>16525728</td>
<td></td>
</tr>
<tr>
<td>Csl</td>
<td>SNAI1</td>
<td>Inhibitory</td>
<td>N</td>
<td>18698484</td>
<td></td>
</tr>
<tr>
<td>Csn</td>
<td>βTrCP</td>
<td>Inhibitory</td>
<td>N</td>
<td>19400951</td>
<td></td>
</tr>
</tbody>
</table>
Csn       GSK3β        Activating  Y  19411070  
DELTA    NOTCH      Inhibitory  Y  18758484  
DSH      GSK3β      Inhibitory  Y  12097378  
E-cadherin β-catenin_mem  Activating  Y  22007144  
E-cadherin β-catenin_nuc Inhibitory  Y  15662113  

The EMT node represents epithelial-to-mesenchymal transition and is the output of this model.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Node</th>
<th>Status</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-cadherin</td>
<td>EMT</td>
<td>Inhibitory</td>
<td>N</td>
</tr>
<tr>
<td>EGF</td>
<td>EGFR</td>
<td>Activating</td>
<td>Y</td>
</tr>
<tr>
<td>EGFR</td>
<td>SOS/GRB2</td>
<td>Activating</td>
<td>Y</td>
</tr>
<tr>
<td>ERK</td>
<td>SNAI2</td>
<td>Activating</td>
<td>N</td>
</tr>
<tr>
<td>ERK</td>
<td>c-fos</td>
<td>Activating</td>
<td>N</td>
</tr>
<tr>
<td>ERK</td>
<td>SOS/GRB2</td>
<td>Inhibitory</td>
<td>Y</td>
</tr>
<tr>
<td>ERK</td>
<td>GSK3β</td>
<td>Inhibitory</td>
<td>N</td>
</tr>
<tr>
<td>ERK</td>
<td>RKIP</td>
<td>Inhibitory</td>
<td>Y</td>
</tr>
<tr>
<td>EGR1</td>
<td>SNAI1</td>
<td>Activating</td>
<td>Y</td>
</tr>
<tr>
<td>FGFR</td>
<td>FGFR</td>
<td>Activating</td>
<td>N</td>
</tr>
<tr>
<td>FGFR</td>
<td>SOS/GRB2</td>
<td>Activating</td>
<td>N</td>
</tr>
<tr>
<td>FGFR</td>
<td>SRC</td>
<td>Activating</td>
<td>N</td>
</tr>
<tr>
<td>FOXC2</td>
<td>E-cadherin</td>
<td>Inhibitory</td>
<td>N</td>
</tr>
<tr>
<td>Frizzled</td>
<td>DSH</td>
<td>Activating</td>
<td>N</td>
</tr>
<tr>
<td>FUS</td>
<td>SUFU</td>
<td>Inhibitory</td>
<td>Y</td>
</tr>
<tr>
<td>GLI</td>
<td>SNAI1</td>
<td>Activating</td>
<td>Y</td>
</tr>
<tr>
<td>GLI</td>
<td>TGFβ</td>
<td>Activating</td>
<td>Y</td>
</tr>
<tr>
<td>GLI</td>
<td>Wnt</td>
<td>Activating</td>
<td>Y</td>
</tr>
<tr>
<td>Goosecoid</td>
<td>FOXC2</td>
<td>Activating</td>
<td>N</td>
</tr>
<tr>
<td>Goosecoid</td>
<td>ZEB2</td>
<td>Activating</td>
<td>N</td>
</tr>
<tr>
<td>Goosecoid</td>
<td>TGFβ</td>
<td>Activating</td>
<td>N</td>
</tr>
<tr>
<td>Goosecoid</td>
<td>TWIST1</td>
<td>Activating</td>
<td>N</td>
</tr>
<tr>
<td>Goosecoid</td>
<td>SNAI1</td>
<td>Activating</td>
<td>N</td>
</tr>
<tr>
<td>Goosecoid</td>
<td>ZEB1</td>
<td>Activating</td>
<td>N</td>
</tr>
<tr>
<td>GSK3β</td>
<td>SNAI1</td>
<td>Inhibitory</td>
<td>Y</td>
</tr>
<tr>
<td>GSK3β</td>
<td>Dest_Compl</td>
<td>Activating</td>
<td>Y</td>
</tr>
<tr>
<td>GSK3β</td>
<td>RAS</td>
<td>Inhibitory</td>
<td>Y</td>
</tr>
<tr>
<td>EGFR</td>
<td>SOS/GRB2</td>
<td>Activating</td>
<td>Y</td>
</tr>
<tr>
<td>EGFR</td>
<td>SRC</td>
<td>Activating</td>
<td>Y</td>
</tr>
<tr>
<td>HEY1</td>
<td>E-cadherin</td>
<td>Inhibitory</td>
<td>Y</td>
</tr>
</tbody>
</table>

GSK3β is part of the destruction complex (Dest_compl node), which is a protein complex containing APC, AXIN2, and GSK3β proteins.
HGF  cMet  Activating  Y  23624790
HIF1α  LOXL23  Activating  Y  15492792
HIF1α  TWIST1  Activating  Y  19279556
HIF1α  ZEB1   Activating  Y  23185316
HIF1α  ZEB2   Activating  Y  23185316
Hypoxia  HIF1α  Activating  Y  23784441
IGF1    IGFR1  Activating  Y  15229476
IGF1    SRC    Activating  Y  0
IKKα    NFkB   Activating  Y  22833419
ILK     AKT    Activating  Y  23783575
Jagged  NOTCH  Activating  Y  22307554
LIV1    SNAI1  Activating  Y  19724917
LOXL23  SNAI1  Activating  Y  16294032
MEK     ERK    Activating  Y  23749166
miR200  ZEB1   Inhibitory  Y  19020711
miR200  ZEB2   Inhibitory  Y  18829540
NFkB    CN    Activating  Y  19411070
NFkB    TWIST1 Activating  Y  17332324
NOTCH  NOTCH_ic  Activating  Y  23323858
NOTCH  NOTCH_ic  Activating  Y  18363556
PAK1    SNAI1  Activating  Y  22421159
Patched SMO    Inhibitory  Y  20643350
PDGF    PDGFR  Activating  Y  23141925
PDGFR   SOS/GRB2 Activating  Y  8041791
PDGFR   SRC    Activating  Y  22080864
PI3K    AKT    Activating  Y  23867821
RAF     MEK    Activating  Y  23667175
RAS     DELTA  Activating  Y  16489124
RAS     PI3K   Activating  Y  23684925
RAS     RAF    Activating  Y  23601922
RKIP    MEK    Inhibitory  Y  15151133
SNAI2   E-cadherin Inhibitory  Y  23906494
SNAI2   SNAI2  Activating  Y  16510505
SMAD    HEY1   Activating  Y  22421041
SMAD    ILK    Activating  Y  21150927
SMAD    Jagged Activating  Y  22585622
SMAD    SHH    Activating  N  18174246
SMAD    SNAI1  Activating  Y  20519943
SMAD    SNAI1  Activating  N  18411277
SMAD    SNAI1  Activating  N  18347095
SMAD    SNAI1  Activating  N  21394833
The “destruction complex” (represented by the Dest_compl node) is formed by APC, AXIN1 or AXIN2, and GSK3β. The APC protein, a constituent of the destruction complex, is not included as an individual node in the EMT network, as there is no compelling evidence of its differential regulation during the EMT process. This is equivalent with assuming that it is constitutively present. We use a single AXIN node, AXIN2, as there is strong evidence that AXIN1 and AXIN2 are functionally equivalent; however, they are differentially expressed and AXIN2 is transcriptionally induced by Wnt signaling through β-catenin and the TCF/LEF transcription factors (1).
3.4.8 Boolean Dynamic Modeling

In Boolean models, each network node can be described by one of two qualitative states: ON or OFF. The ON state means an above threshold concentration of a molecular regulator whereas the OFF state means the below-threshold or inactivated form. The biological relationships defined by the network can be translated into mathematical equations using Boolean operators (2). Accordingly, we described each component in the network by a regulatory function that uses the Boolean logical operators OR, AND, and NOT and that reflects the known combinatorial effects of the regulators. Specifically, all rules involving multiple regulators were constructed with OR qualifiers, assuming that all regulators act independently, unless there was evidence of synergy or conditionality, in which case the regulators were connected by an AND qualifier. Here we explain the rules of three key nodes (TGFβ, E-cadherin and Destr_compl) that are important for understanding the paper and to clarify our modeling methodology.

The TGFβ rule is as follows:

\[ TGF\beta^* = \text{Gooseco} \text{d or SNAI1 or TWIST1 or GLI} \]

When thinking about Boolean rules, it might be more intuitive if instead of ON or OFF, we consider each node as having the logic states TRUE or FALSE, respectively. According to this update rule, if all inputs are OFF (FALSE), then TGFβ will turn OFF (FALSE):

\[ TGF\beta^* = \text{FALSE or FALSE or FALSE or FALSE} = \text{FALSE} \]

Now we consider the case when a single regulator, SNAI1, is ON (TRUE) at a previous time step:

\[ TGF\beta^* = \text{FALSE or TRUE or FALSE or FALSE} = \text{TRUE} \]

Thus the TGFβ node will be ON (active) if any of its individual inputs is ON and it will be inactive if all of its inputs are OFF.

The E-cadherin rule is as follows:

\[ \text{E-cadherin}^* = \beta-\text{catenin}_\text{memb} \text{ and (not SNAI1 or not HEY1 or not ZEB1 or not ZEB2 or not FOXC2 or not TWIST1 or not SNAI2)} \]
According to this rule, E-cadherin will turn OFF if \( \beta \)-catenin_memb is turned OFF or all of the 7 negative regulators are turned ON. First, let’s demonstrate what happens when \( \beta \)-catenin_memb is OFF (FALSE).

\[
E\text{-cadherin}^* = \text{FALSE and (not SNAI1 or not HEY1 or not ZEB1 or not ZEB2 or not FOXC2 or not TWIST1 or not SNAI2)}
\]

“AND” rules logically require both states to be TRUE in order to produce a TRUE output. Thus, regardless of the state of the nodes in the parentheses, if \( \beta \)-catenin_memb is OFF (FALSE), E-cadherin will be OFF. E-cadherin will also be OFF, if all of the 7 negative afferents are ON, regardless of the state of \( \beta \)-catenin_memb. If SNAI1 is ON (TRUE), the expression “not SNAI1” is FALSE and similarly for all negative regulators of E-cadherin. In this case the E-cadherin rule is

\[
E\text{-cadherin}^* = \beta \text{-catenin_memb and (FALSE or FALSE or FALSE or FALSE or FALSE or FALSE or FALSE)}
\]

This logically evaluates to FALSE, meaning that E-cadherin turns OFF.

If all 7 negative regulators are OFF and \( \beta \)-catenin_memb is ON, then E-cadherin will be ON:

\[
E\text{-cadherin}^* = \text{TRUE and (TRUE or TRUE or TRUE or TRUE or TRUE or TRUE or TRUE)} = \text{TRUE}
\]

In fact, if any one of the negative regulators of E-cadherin is OFF, the parenthesis is equivalent with “FALSE or TRUE”, which evaluates to TRUE. Thus E-cadherin will be ON if \( \beta \)-catenin_memb is ON and any of the 7 negative afferents are OFF.

The Dest_compl (destruction complex) rule is as follows:

\[
\text{Dest_compl}^* = (\text{GSK3}\beta \text{ and AXIN2 and } \beta \text{-catenin_nuc}) \text{ or (GSK3}\beta \text{ and Dest_compl)}
\]

According to this rule, the initial formation (ON state) of the complex requires GSK3\(\beta\), AXIN2 and \(\beta\)-catenin_nuc. However if the Dest_compl is already present (ON), which it is in the epithelial state, then it will remain sufficiently stable even if AXIN2 and \(\beta\)-catenin_nuc are present at a below-threshold level (OFF). The only additional node that needs to be ON for the destruction complex to stay ON is GSK3\(\beta\).

3.4.9 Boolean Rule Over-Rides

Biological dysregulations are simulated by forcing a node to be ON (positive dysregulation) or OFF (negative dysregulation). Positive dysregulation is analogous to over-expression, over-
abundance, or mutational activation of a protein and negative dysregulation is analogous to a knockout, down-regulation, or mutational inactivation of a protein. To model the effect of an over-active TGFβ signal we use a positive over-ride in the simulation results shown in Figure 3-8, in the state transition network in Figure 3-9B, and to produce the motifs in Figure 3-12 and Figure 3-4. We model the inhibition of SMAD by a negative over-ride, in the presence of an over-active TGFβ signal, to produce the simulation results in Figure 3-12B and the network motifs in Figure 3-12C.

3.4.10 Asynchronous Updating Algorithm and Node Activity

To account for different timescales in signaling networks, asynchronous models may be used in which the state of each node is updated according to its own timescale. While we do not know the kinetics of individual events, we do know that signal transduction events occur substantially faster (with a timescale of seconds or faster) than transcriptional events (which have a timescale of minutes) (3). Thus, we apply a general asynchronous updating scheme with a ranking scheme. In the general asynchronous update, a single randomly chosen node is updated (i.e. its state is re-evaluated) at each time instant. It is possible that the same node chosen in two consecutive time instants.

In order to account for signal transduction events occurring faster than transcriptional events, we incorporate a ranking system. We update nodes regulated by signal transduction events with a greater probability than nodes regulated by transcriptional events. We tested probability values that differed by more than an order of magnitude and determined that the quantitative value of the ratio between the timescale of the slow and fast processes does not matter as long as it is significantly larger than one (Figure 3-1). We define a time step as the average number of updates needed to update a slow (transcriptionally regulated) node. This time step thus corresponds to several minutes in real time. Altering the timing does not affect the possible steady states, nor does it affect the initial conditions that lead to these steady states. It does, however, alter the probability of specific trajectories and thus the relative rates at which nodes are affected by a signal in our network.
Because the ranked general asynchronous update is a stochastic process, different simulations that start from the same initial state can reach the different outcomes (steady states). Thus, a large number (10,000) of replicate simulations are performed and the average state of each node is calculated. We call the average state of a node the node activity. If a node’s activity over a large number of simulations stabilizes at 1, it means that each simulation led to a sustained ON state of the node. Conversely, stabilized node activity of 0 means that the node stabilized at OFF in each simulation. Node activities between 0 and 1 represent the fraction of simulations in which a node was ON. When running computer simulations it is important to run as many simulations as possible; however there are time constraints that arise when increasing the

Figure 3-4. All stable motifs (feedback loops) stabilizing the TGFβ-driven EMT phenotype. Black background represents nodes that are OFF in the EMT steady state. White background indicates nodes that are ON in the EMT steady state.
simulation size. At the same time, there is a diminishing return for doing more simulations. 10,000 simulations and 20 time steps was deemed sufficient for most simulations explored in the EMT network, as averaged values over runs from these simulations produced consistent results with only small deviations.

3.4.11 State Space Analysis

Biological systems like the EMT network can be modeled as dynamic systems; these models can explore the full range of possibilities in terms of outcomes (e.g., steady states) the system can reach and ways in which these outcomes can be reached (trajectories). State space analysis entails exploring every possible state of a network of size N ($2^N$ states) and how these states can lead to other system states, as governed by the Boolean rules and updating scheme defining the system. All sequences of consecutive states will finally reach an attractor (a steady state or a small set of states) that the network cannot leave. The set of system states directly reaching an attractor are called the attractor’s basin of attraction.

3.4.12 Network Reduction.

Because the size of the state space in Boolean models grows exponentially ($2^N$, where N = the number of nodes in the network), it becomes impossible to map all the trajectories in large networks (e.g. $2^{70} > 10^{20}$ possible states in the EMT network). Network reduction techniques have been developed to reduce the size of the network without limiting its number of outcomes. We used a two-step network reduction method. The first step is to identify and eliminate stabilized nodes. Because the focus of our work is TGFβ induction of EMT, all other signals (e.g. EGF) were set as OFF. Signal nodes unaffected by crosstalk (e.g. EGF, HGF) maintain their OFF states, and the effect of this OFF state iteratively fixes the state of certain downstream nodes. These signals and their fixed-state downstream nodes were identified analytically and removed as their effect was not targeted by TGFβ signaling. Our second network reduction step is to iteratively collapse mediator nodes that have one regulator and in turn regulate a single node. This method conserves the attractors of the system (4-6). In these cases the node directly upstream of the removed node was connected to the node directly downstream of the reduced node, and the state of the upstream regulator was inserted in the downstream-regulated node’s Boolean update rule. Thus using these reduction techniques we can reduce the size of the
network from 70 nodes and 135 edges to 19 nodes and 70 edges and perform a state space analysis on a smaller network (Reduced EMT dynamic model, Figure 3-9 and Figure 3-3) that is representative of the full EMT network. This network reduction was shown to have no effect on the permitted dynamic behaviors (e.g. steady states) of a system (4).

3.4.13 Stable Motif Analysis

We used stable motif analysis to identify key feedback loops regulating EMT. A stable motif of a Boolean model consists of a set of nodes and their respective states which satisfy two characteristics: 1) the nodes form a strongly connected component (in which each node is reachable from any other node) in a directed network representation of the Boolean model, and 2) the specified states are steady states of the nodes of the stable motif, independent of the state of any node outside of the stable motif (7). These two characteristics have a topological and dynamical interpretation. Topologically, they imply that the nodes of the stable motif form a type of positive feedback loop. Dynamically, they imply that the states associated to the stable motif act as a checkpoint or point of no return in the dynamics of the network. Previous work has shown that an iterative algorithm which finds the stable motifs and uses the influence of these motifs on the rest of the network can be used to find all the attractors of asynchronous Boolean models (7).

3.4.14 Miscellaneous Computational Methods

The networks in Figure 3-7, Figure 3-9A, Figure 3-9C, Figure 3-12A, Figure 3-12C and Figure 3-4 were created using the yEd graph editor by yWorks (http://www.yworks.com/). The state transition network in Figure 3-9B was generated with Cytoscape (24) by using the Prefuse force directed layout.

3.5 Results

3.5.1 TGFβ Signaling is a Conserved Driver of EMT in Multiple Liver Cancer Models

Our initial studies involved development of a novel murine model of epithelial-to-mesenchymal transition. This involved derivation of “sister” epithelial (P2E) and mesenchymal/metastatic (P2M) cell lines from a PTEN−/− mouse (15). Consistent with other carcinoma cell lines, treatment of the P2E cell line with TGFβ induced TGFβ signaling (Figure
47

3-5A) and led to induction of EMT markers (Figure 3-5B). These results demonstrate that activation of TGFβ signaling recapitulates the mesenchymal phenotype in a murine model of HCC. They also led us to hypothesize that the TGFβ pathway is constitutively active in P2M cells as compared to P2E cells.

We tested this hypothesis by gene set enrichment analysis (GSEA), a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two phenotypes, in this case P2E and P2M cell lines (19). The primary result of GSEA is the enrichment score, which shows the degree to which a gene set is over-represented at the top or bottom of a list of genes ranked by their association with a phenotype (see Methods). An EMT gene set representing genes up-regulated during EMT induced in a mammary epithelium cell line transformed by HRAS (EpH4 line)(25) was used to probe for the EMT phenotype in P2M versus P2E cells. The use of a breast cancer reference set is reasonable because EMT is a conserved process over multiple tissue types (26). A TGFβ gene set representing genes up-regulated by TGFβ1 in MCF10A immortalized non-tumorigenic mammary gland cells (27) was used to probe for TGFβ pathway activation in P2M versus P2E cell lines. The top portion of the plot shows the running enrichment score and a maximum enrichment score of 0.78 for the EMT gene set and 0.76 for the TGFβ gene set (Figure 3-5C). In the bottom portion of the plot, P2M cells are represented on the left of the figure and P2E on the right. The position of the members of the gene set in the ranked list of genes demonstrates distinct genes associated with EMT and TGFβ sets. These data support up-regulation of EMT and TGFβ pathway activation gene sets in the P2M cell line relative to P2E (Figure 3-5C).

In order to explore the prevalence of dysregulated TGFβ signaling in HCC, we tested established human HCC cell lines for TGFβ pathway and EMT markers. Human HCC cell lines Huh7, PLC/PRF/5, HLE, and HLF cell lines were tested for epithelial and mesenchymal markers to characterize their phenotype. Huh7 and PLC/PRF/5 cells have higher E-cadherin and lower pSMAD3 compared to HLE and HLF cell lines by western blot analysis (Figure 3-5D). TGFβ-treated PLC/PRF/5 cells, HLE and HLF HCC cell lines have lower E-cadherin, higher vimentin (VIM), and higher TWIST1 compared to PLC/PRF/5 cells as indicated by qRT-PCR (Figure 3-5E). Compared to PLC/PRF/5 cells, TGFβ treated PLC/PRF/5 cells and untreated HLF cells have higher N-cadherin expression (Figure 3-5E). These results indicate that the Huh7 and
Figure 3-5. Active TGFβ signaling correlates with mesenchymal phenotype in murine and human models of HCC. (A) Treatment of a mouse epithelial HCC cell line (P2E) with TGFβ at 5 ng/ml for 48 hours produces a molecular phenotype similar to mouse mesenchymal HCC cell line (P2M). (B) Up-regulation of mRNA levels of EMT markers in mouse epithelial HCC with TGFβ treatment (5 ng/ml for 48 hours). (C) Gene set enrichment analysis (GSEA) supports P2M cells are enriched for mesenchymal markers relative to P2E cells (left panel). GSEA reveals that TGFβ signaling is enriched in P2M versus P2E cells, suggesting that TGFβ may be driving the EMT phenotype in this mouse model of HCC (right panel). This gene set contains genes up-regulated by TGFβ1 in MCF10A cells. (FDR = 0.031). (D) Human mesenchymal HCC cell lines HLE and HLF have low E-cadherin expression and high pSMAD3, whereas epithelial-like HCC cell lines Huh7 and PLC/PRF/5 (PLC) express high E-cadherin and low pSMAD3. (E) Relative to the PLC/PRF/5 cell line, PLC/PRF/5 treated with TGFβ (10 ng/ml, 48 hrs), HLE, and HLF cell lines exhibited lower E-cadherin and higher vimentin, and TWIST1 expression. FDR = False discovery rate. n =3 technical replicates were used.
PLC/PRF/5 cell lines have an epithelial-like phenotype, the HLE and HLF cell lines have a mesenchymal phenotype, and that TGFβ treatment induces a mesenchymal-like phenotype.

In order to study the effect of TGFβ pathway activity on the EMT phenotype of the mesenchymal-like HLE and HLF cell lines, these cell lines were treated with TGFβRI inhibitor LY2157299 and EMT markers were assessed by western blot and qRT-PCR. TGFβRI inhibition with LY2157299 reduced SNAI1, TWIST1, N-cadherin (CDH2), and vimentin mRNA expression, while increasing E-cadherin expression in HLE cells (Figure 3-6). These data support the conclusion that the mesenchymal phenotype of HLE and HLF cell lines is due to constitutive TGFβ signaling. Overall, our data demonstrate that TGFβ axis dysregulation is a conserved driver of mesenchymal phenotype in multiple models of HCC.

3.5.2 Constructing the Signaling Network that Drives Epithelial-to-Mesenchymal Transition

In order to explore the systems-level dynamics of TGFβ signaling and EMT, a network modeling approach was used. Evidence supports the concept that a conserved EMT mechanism, driven by multiple signals, occurs in development and in cancer. These signals appear as a normal part of development, playing a crucial role in formation of the body plan and in differentiation of multiple tissues and organs (6). During carcinoma progression, EMT signals (e.g. TGFβ) are aberrantly produced and hijack this developmental process (28). In order to construct our network, we performed an extensive literature search and synthesized pre-existing information representing the conserved regulation of EMT. We focused on interactions from HCC EMT; however information about EMT in other tissue types was also used when no HCC-specific information

Figure 3-6. TGFβR inhibition suppresses mesenchymal features in human HCC cell lines. TGFβ-driven mesenchymal phenotype HLE HCC cell line reveals constitutively active TGFβ signaling and mesenchymal features, which are lost upon treatment with TGFβRI/II inhibitor (LY2157299) at 48 hours as demonstrated by qRT-PCR. n =3 technical replicates were used.
was available. The resulting network (Figure 3-7) spans heterogeneous molecular processes induced by growth factors, signal transduction pathways, and transcriptional regulators. This network includes the dysregulation of the TGFβ/TGFβR axis as a driver of EMT, as this is a commonly mutated pathway in HCC (9).

Figure 3-7. The 70 node, 135 edge network model representing the epithelial-to-mesenchymal transition (EMT) signaling network in HCC. Nodes represent molecular entities (proteins, small molecules, mRNAs) and edges represent activating and inhibitory relationships between nodes. Upstream signals (black text, dark gray fill) regulate transcriptional regulators (black text, light gray fill) through signal transduction pathways, which all converge on the regulation of E-cadherin. The output of the network, “EMT”, is indicated by a symbol with black background and white text. The full names of the nodes are indicated in Table 3-2. References for each edge in the network are indicated in Table 3-1. The Boolean rules describing the dynamic EMT network model are given in Table 3-4.
<table>
<thead>
<tr>
<th>Name used in network</th>
<th>Gene symbol</th>
<th>Gene name/official name</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>AKT</td>
<td>AKT1</td>
<td>v-akt murine thymoma viral oncogene homolog 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AKT2</td>
<td>v-akt murine thymoma viral oncogene homolog 2</td>
<td></td>
</tr>
<tr>
<td>AKT</td>
<td>APC</td>
<td>adenomatous polyposis coli</td>
<td></td>
</tr>
<tr>
<td>AXIN2</td>
<td>AXIN2</td>
<td>axin 2</td>
<td></td>
</tr>
<tr>
<td>AXIN2</td>
<td>GSK3β</td>
<td>glycogen synthase kinase 3 beta</td>
<td></td>
</tr>
<tr>
<td>Dest_compl</td>
<td>N/A</td>
<td>Destruction complex</td>
<td></td>
</tr>
<tr>
<td>AXIN2</td>
<td>CTNNB1</td>
<td>β-catenin (cadherin-associated protein), beta 1, 88kDa</td>
<td>*NOTE: “β-catenin_memb” refers to membrane-bound β-catenin.</td>
</tr>
<tr>
<td>β-catenin_memb</td>
<td>CTNNB1</td>
<td>β-catenin (cadherin-associated protein), beta 1, 88kDa</td>
<td>*NOTE: “β-catenin_nuc” refers to nuclear β-catenin.</td>
</tr>
<tr>
<td>β-catenin_nuc</td>
<td>CTNNB1</td>
<td>β-catenin (cadherin-associated protein), beta 1, 88kDa</td>
<td></td>
</tr>
<tr>
<td>βTrCP</td>
<td>BTRC</td>
<td>beta-transducin repeat containing E3</td>
<td></td>
</tr>
<tr>
<td>CD44</td>
<td>CD44</td>
<td>CD44 molecule (Indian blood group)</td>
<td></td>
</tr>
<tr>
<td>CDC42</td>
<td>CDC42</td>
<td>cell division cycle 42</td>
<td></td>
</tr>
<tr>
<td>FOS</td>
<td>FOS</td>
<td>FBJ murine osteosarcoma viral</td>
<td></td>
</tr>
<tr>
<td>c-fos</td>
<td>CHD1L</td>
<td>chromodomain helicase DNA binding protein 1-like</td>
<td></td>
</tr>
<tr>
<td>CHD1L</td>
<td>MET</td>
<td>met proto-oncogene (hepatocyte growth factor receptor)</td>
<td></td>
</tr>
<tr>
<td>cMet</td>
<td>RBPJ</td>
<td>recombination signal binding protein for immunoglobulin kappa J region</td>
<td></td>
</tr>
<tr>
<td>Csl</td>
<td>COPS2</td>
<td>COP9 Constitutive Photomorphogenic Homolog Subunit 2</td>
<td>Complex (COP9 signalosome)</td>
</tr>
<tr>
<td></td>
<td>COPS3</td>
<td>COP9 Constitutive Photomorphogenic Homolog Subunit 3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>COPS4</td>
<td>COP9 Constitutive Photomorphogenic Homolog Subunit 4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>COPS5</td>
<td>COP9 Constitutive Photomorphogenic Homolog Subunit 5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>COPS6</td>
<td>COP9 Constitutive Photomorphogenic Homolog Subunit 6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>COPS7a</td>
<td>COP9 Constitutive Photomorphogenic Homolog Subunit 7a</td>
<td></td>
</tr>
<tr>
<td></td>
<td>COPS7b</td>
<td>COP9 Constitutive Photomorphogenic Homolog Subunit 7b</td>
<td></td>
</tr>
<tr>
<td></td>
<td>COPS8</td>
<td>COP9 Constitutive Photomorphogenic Homolog Subunit</td>
<td></td>
</tr>
<tr>
<td>Csn</td>
<td>DLL1</td>
<td>DELTA-Like ligand 1</td>
<td></td>
</tr>
<tr>
<td>DELTA</td>
<td>DLL1</td>
<td>DELTA-Like ligand 1</td>
<td></td>
</tr>
<tr>
<td>DSH</td>
<td>DVL1</td>
<td>dishevelled, dsh homolog 1</td>
<td></td>
</tr>
<tr>
<td>E-cadherin</td>
<td>CDH1</td>
<td>cadherin 1, type 1, E-cadherin</td>
<td></td>
</tr>
<tr>
<td>EGFR</td>
<td>EGFR</td>
<td>epidermal growth factor receptor</td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td>Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EGR1</td>
<td>early growth response 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMT</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ERK</td>
<td>mitogen-activated protein kinase 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FGFR</td>
<td>fibroblast growth factor receptor 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FOXC2</td>
<td>forkhead box C2 (MFH-1, mesenchyme forkhead 1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frizzled</td>
<td>frizzled family receptor 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FUS</td>
<td>serine/threonine kinase 36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLI1</td>
<td>GLI family zinc finger 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLI2</td>
<td>GLI family zinc finger 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSK3B</td>
<td>glycogen synthase kinase 3 beta</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HEY1</td>
<td>YRPW motif 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HIF1α</td>
<td>hypoxia inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IGF1R</td>
<td>insulin-like growth factor 1 receptor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ILK</td>
<td>integrin-linked kinase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jagged</td>
<td>jagged 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIV1</td>
<td>Solute Carrier Family 39 (Zinc Transporter), Member 6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOXL2</td>
<td>lysyl oxidase-like 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOXL3</td>
<td>lysyl oxidase-like 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAP2K1</td>
<td>mitogen-activated protein kinase kinase 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAP2K2</td>
<td>mitogen-activated protein kinase kinase 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEK</td>
<td>microRNA 200b</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NFκB</td>
<td>nuclear factor of kappa light polypeptide gene enhancer in B-cells 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOTCH</td>
<td>NOTCH (Drosophila) Homolog 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOTCH1</td>
<td>NOTCH (Drosophila) Homolog 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOTCH1</td>
<td>intracellular domain (NICD)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOTCH_1c</td>
<td>p21 protein (CDC42/Rac)-activated kinase 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PTCH1</td>
<td>PTCH Homolog 1 (Drosophila)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDGFR</td>
<td>platelet-derived growth factor receptor, alpha polypeptide</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDGFRB</td>
<td>platelet-derived growth factor receptor, beta polypeptide</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PI3K</td>
<td>phosphoinositide-3-kinase, catalytic, alpha polypeptide</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PIK3B</td>
<td>phosphoinositide-3-kinase, catalytic, beta polypeptide</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PIK3C</td>
<td>phosphoinositide-3-kinase, catalytic, gamma polypeptide</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAFT</td>
<td>v-raf-1 murine leukemia viral oncogene</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAS</td>
<td>v-Ha-ras Harvey rat sarcoma viral oncogene homolog</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PEBP4</td>
<td>phosphatidylethanolamine-binding protein 4; RAF1 kinase inhibitory protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RKIP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHH</td>
<td>Sonic hedgehog</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNAI2</td>
<td>snail homolog 2 (Drosophila)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMAD2</td>
<td>SMAD family member 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMAD3</td>
<td>SMAD family member 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMAD4</td>
<td>SMAD family member 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMO</td>
<td>smoothened, frizzled family receptor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNAI1</td>
<td>snail homolog 1 (Drosophila)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOS/GRB2</td>
<td>son of sevenless homolog 1 (Drosophila) Complex</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRB2</td>
<td>growth factor receptor-bound protein 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRC</td>
<td>viral oncogene homolog (avian)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STAT3</td>
<td>signal transducer and activator of transcription 3 (acute-phase response factor)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STAT1</td>
<td>signal transducer and activator of transcription 1, 91kDa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SUFU</td>
<td>suppressor of fused homolog (Drosophila)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCF4</td>
<td>transcription factor 4, a basic helix-loop-helix transcription factor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LEF1</td>
<td>lymphoid enhancer-binding factor 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TGFB1</td>
<td>transforming growth factor, beta 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TGFB1B</td>
<td>transforming growth factor, beta receptor 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TWIST1</td>
<td>twist basic helix-loop-helix transcription factor 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wnt1</td>
<td>wingless-type MMTV integration site family, member 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZEB1</td>
<td>zinc finger E-box binding homebox 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZEB2</td>
<td>zinc finger E-box binding homebox 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EGF</td>
<td>epidermal growth factor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FGF</td>
<td>fibroblast growth factor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HGF</td>
<td>hepatocyte growth factor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZEB1</td>
<td>zinc finger E-box binding homebox 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZEB2</td>
<td>zinc finger E-box binding homebox 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EGF</td>
<td>epidermal growth factor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FGF</td>
<td>fibroblast growth factor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HGF</td>
<td>hepatocyte growth factor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSC</td>
<td>goosecoid homebox</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IGF1</td>
<td>Insulin-like growth factor 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDGFB</td>
<td>platelet-derived growth factor beta</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDGF</td>
<td>polypeptide</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Several network nodes represent the union of a few proteins with similar roles. In these cases, a single entry in the first column corresponds to several entries in the second column.
The information used to construct the network is summarized in Table 3-1 and Table 3-2. Proteins, mRNAs, and microRNAs were represented as nodes (Table 3-1) and the pairwise relationships among nodes (Table 3-1) were represented as edges. We choose as the output node of the network the concept node ‘‘EMT’’, serving as the major indicator of cell fate in our model. E-cadherin plays a crucial and direct role in EMT at diverse developmental stages and during carcinoma progression, and its loss is widely considered as the hallmark of EMT(6). Furthermore, the regulation of E-cadherin expression is much better understood than that of other EMT markers(29). Thus we focus on E-cadherin as the negative regulator of the output node of the network, ‘‘EMT’’. This network contains 70 nodes and 135 edges (Figure 3-7).

3.5.3 Translating the EMT Signaling Network into a Dynamic Model

To understand the dynamics of signaling abnormalities in HCC progression, invasion, and metastasis, we translated the EMT signaling network into a dynamic model. A dynamic model is used to express the behavior of a system over time by characterizing each node by a state variable (representing, e.g., its concentration or activity) and a function that describes its regulation. Dynamic models can be categorized as continuous or discrete, according to the type of node state variable used. Continuous models use a set of differential equations; however, the paucity of known kinetic details for biological interactions makes these models difficult to implement.

We employed discrete dynamic modeling because of its computational feasibility and capacity to be constructed with qualitative biological data. Specifically, we constructed a Boolean model wherein each network node is described by one of two qualitative states: ON or OFF. The ON state means an above threshold concentration of a molecular regulator whereas the OFF state means the below-threshold or inactivated form. We formulated a regulatory function for each node using its upstream regulators encapsulated in the network, evidence regarding the conditionality among these regulators, and known biological outcomes. During the construction of the regulatory functions we verified that individual modules within the network reproduced known outcomes (Table 3-3). If modules did not reproduce known outcomes, we modified the rules in such a way that the known outcomes can be reproduced (30). See the Quick Guide to Equations and Assumptions for more detailed information and example rule implementations.
Table 3-3: Key experimental outcomes reproduced by the EMT network model

<table>
<thead>
<tr>
<th>Statement</th>
<th>Ref. (PMID)</th>
<th>Does the model capture this?</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGFβ signaling leads to SMAD complex formation, MAP kinase signaling, and AKT signaling.</td>
<td>20495575</td>
<td>Yes</td>
</tr>
<tr>
<td>Wnt signaling leads to nuclear localization of β-catenin, AXIN2 induction, and suppression of the destruction complex</td>
<td>11809808</td>
<td>Yes</td>
</tr>
<tr>
<td>SHH signaling leads to induction of GLI transcription factors</td>
<td>21789137</td>
<td>Yes</td>
</tr>
<tr>
<td>miR200 inhibits TGFβ–driven EMT</td>
<td>18411277</td>
<td>Yes</td>
</tr>
<tr>
<td>E-cadherin suppressing transcription factors</td>
<td>22945800,</td>
<td>Yes</td>
</tr>
<tr>
<td>SNAI1, SNAI2, ZEB1, ZEB2, TWIST1, FOXC2, and HEY1 induce EMT when acting together</td>
<td>20713713</td>
<td>Yes</td>
</tr>
<tr>
<td>Constitutive SNAI1 or TWIST1 activation induces the activation of SNAI2, ZEB1, ZEB2, HEY1, and FOXC2, and drives EMT</td>
<td>22945800, 20713713, 21199805, 21317430</td>
<td>Yes</td>
</tr>
</tbody>
</table>

As in the biological system, there is a time lag between the state change of the regulators and the state change of the targets. This time lag is implemented by using discrete update events, each update meaning that the state of a node is re-calculated according to its Boolean equation. We used an updating scheme that differentiates between the two major types of biological regulation in the model: transcriptional events and signal transduction events. Although we do not know the specific kinetics of individual events, we do know that signal transduction events occur substantially faster (in seconds or faster) than transcriptional events (timescale of minutes)(31). Thus, we used a ranked asynchronous updating algorithm that updates nodes regulated by signal transduction events with a greater probability than nodes regulated by transcriptional events. This allows us to stochastically explore differences in the kinetics among molecular interactions in our network while satisfying known kinetic relationships (14).

To reproduce how a potentially diverse population of cells responds to the same signal we performed multiple simulations with the same initial conditions but different updating orders (i.e. different timing). The output of the model is the fraction of simulations in which a certain node is in the ON state, which we denote the activity of the node. We choose as time unit (step) the average number of updates needed to update a transcriptionally regulated node. This time step thus corresponds to several minutes in real time. We empirically determined that 10,000
Simulations and 20 time steps in each simulation are sufficient for convergent and reproducible results.

3.5.4 The EMT Network Model Reproduces Known TGFβ-Driven Dysregulations During Epithelial-to-Mesenchymal Transition

We focus on TGFβ as the prototypical inducer of EMT as the TGFβ signaling axis is disrupted in 40% of HCC(9) and TGFβ is a well-established inducer of EMT in HCC and other carcinomas(8). First, we set the states of the nodes according to their known or expected states in a prototypical epithelial cell. Specifically, known signals/pathways that could induce EMT, including TGFβ FGF, IGF, EGF, PDGF, Wnt/β-catenin, SHH/GLI, and HGF/cMet, were set to OFF. Downstream signal transduction pathways known to drive EMT, such as MAPK, AKT, SMAD, and STAT signaling, were also turned OFF. E-cadherin and miR200 were assumed to be ON and the E-cadherin transcriptional repressors were assumed to be OFF. Our model indicates that without any stimulus or dysregulation, the cell would remain in this epithelial state (listed in Table 3-4).

**Table 3-4. Boolean update rules and initial state of the 68 node EMT network**

<table>
<thead>
<tr>
<th>Update rules</th>
<th>Epithelial initial condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AKT* = ILK or PI3K</td>
<td>AKT = OFF</td>
</tr>
<tr>
<td>Dest_compl* = (GSK3β and AXIN2 and β-catenin_nuc) or (GSK3β and Dest_compl)</td>
<td>Dest_compl = ON</td>
</tr>
<tr>
<td>AXIN2* = AXIN2 or TCF/LEF</td>
<td>AXIN2 = OFF</td>
</tr>
<tr>
<td>β-catenin_memb* = E-cadherin and not β-catenin_nuc</td>
<td>β-catenin_memb = ON</td>
</tr>
<tr>
<td>β-catenin_nuc* = not Dest_compl and not β-catenin_memb and (not SUFU or not E-cadherin)</td>
<td>β-catenin_nuc = OFF</td>
</tr>
<tr>
<td>βTrCP* = not Csn</td>
<td>βTrCP = ON</td>
</tr>
<tr>
<td>CD44* = TCF/LEF</td>
<td>CD44 = OFF</td>
</tr>
<tr>
<td>CDC42* = TGFβR or CHD1L</td>
<td>CDC42 = OFF</td>
</tr>
<tr>
<td>c-fos* = ERK</td>
<td>c-fos = OFF</td>
</tr>
<tr>
<td>cMet* = HGF or CD44</td>
<td>cMet = OFF</td>
</tr>
<tr>
<td>Csl* = NOTCH_ic</td>
<td>Csl = OFF</td>
</tr>
<tr>
<td>Csn* = NFkB</td>
<td>Csn = OFF</td>
</tr>
<tr>
<td>DELTA* = RAS</td>
<td>DELTA = OFF</td>
</tr>
<tr>
<td>DSH* = Frizzled</td>
<td>DSH = OFF</td>
</tr>
<tr>
<td>E-cadherin* = β-catenin_memb and (not SNAI1 or not HEY1 or not ZEB1 or not ZEB2 or not FOXC2 or not TWIST1 or not SNAI2)</td>
<td>E-cadherin = ON</td>
</tr>
<tr>
<td>EGFR* = EGF</td>
<td>EGFR = OFF</td>
</tr>
<tr>
<td>EGR1* = c-fos</td>
<td>EGR1 = OFF</td>
</tr>
<tr>
<td>EMT* = not E-cadherin or EMT</td>
<td>EMT = OFF</td>
</tr>
<tr>
<td>ERK* = MEK</td>
<td>ERK = OFF</td>
</tr>
<tr>
<td>FGFR* = FGF</td>
<td>FGFR = OFF</td>
</tr>
<tr>
<td>FOXC2* = Goosecoid or SNAI1 or TWIST1</td>
<td>FOXC2 = OFF</td>
</tr>
</tbody>
</table>
Frizzled* = Wnt
FUS* = SMO
GLI* = TCF/LEF or not SUFU
GSK3β* = not DSH and not AKT and (not Csn or not ERK or not Dest_compl)
HEY1* = Csl or SMAD
HIF1α* = Hypoxia
IGF1R* = IGF1
IKKα* = AKT
ILK* = SMAD
Jagged* = TCF/LEF or SMAD
LIV1* = STAT
LOXL23* = HIF1α
MEK* = RAF or not RKIP
miR200* = not SNAI1 and not ZEB1 and not ZEB2
NFkB* = IKKα
NOTCH* = DELTA or Jagged
NOTCH_ic* = NOTCH
PAK1* = CDC42
Patched* = not SHH
PDGFR* = PDGF
PI3K* = RAS
RAF* = RAS
RAS* = SOS/GRB2 or SRC or not GSK3β or TCF/LEF
RKIP* = not ERK or not SNAI1
SHH* = SMAD or GL1
SNAI2* = ERK or β-catenin_nuc or SNAI2 or TWIST1
SMAD* = (ERK or TGFβR) and (ZEB1 or not ZEB2)
SMO* = not Patched
SNAI1* = GLI or LOXL23 or SMAD or LIV1 or PAK1 or Csl or EGR1 or Goosecoid or not βTrCP or not GSK3β
SOS/GRB2* = (PDGFR or cMet or TGFβR or FGFR or IGF1R or EGFR) and not ERK
SRC* = PDGFR or EGFR or FGFR or cMet or IGF1R
STAT* = SRC
SUFU* = not FUS
TCF/LEF* = β-catenin_nuc
TGFβ* = Goosecoid or SNAI1 or TWIST1 or GL1
TGFB* = TGFβ
TWIST1* = NFkB or HIF1α or TCF/LEF or Goosecoid or SNAI1
Wnt* = GL1
ZEB1* = (HIF1α or SNAI1 or Goosecoid) and not miR200
ZEB2* = (HIF1α or SNAI1 or Goosecoid) and not miR200

Frizzled = OFF
FUS = OFF
GLI = OFF
GSK3β = ON
HEY1 = OFF
HIF1α = OFF
IGF1R = OFF
IKKα = OFF
ILK = OFF
Jagged = OFF
LIV1 = OFF
LOXL23 = OFF
MEK = OFF
miR200 = ON
NFkB = OFF
NOTCH = OFF
PAK1 = OFF
Patched = ON
PDGFR = OFF
PI3K = OFF
RAF = OFF
RAS = OFF
RKIP = ON
SHH = OFF
SNAI2 = OFF
SMAD = OFF
SMO = OFF
SNAI1 = OFF
SOS/GRB2 = OFF
SRC = OFF
STAT = OFF
SUFU = ON
TCF/LEF = OFF
TGFβ = OFF
TGFB = OFF
TWIST1 = OFF
Wnt = OFF
ZEB1 = OFF
ZEB2 = OFF
EFG = OFF
FGF = OFF
HGF = OFF
Goosecoid = OFF
Hypoxia = OFF
IGF1 = OFF
PDGF = OFF
Next, in order to recapitulate a sustained TGFβ signal, the state of TGFβ was fixed in the ON state at the beginning of every simulation, while the other nodes were in their epithelial initial condition. Turning TGFβ ON led to changes in the activity of nodes in the network over the time course of the simulation and ultimately led to a steady state in which the state of all nodes stabilized (A). If the state of a node stabilized at ON at the end of the simulation, even though it was OFF at the beginning, we infer that it is activated in the steady state. If the state of a node was OFF even though it was in the ON state at the beginning of the simulation, we consider it inactivated in the steady state. Using this interpretation, the obtained steady state agrees with known markers of a mesenchymal state. The heatmap in Figure 3-8A and the time courses shown on Figure 3-8B qualitatively reflect experimental outcomes seen when starting from an epithelial condition and inducing EMT. Specifically, they reflect how TGFβ signals through TGFβRI/II and induces SMAD complex formation and non-canonical pathways including MAPK and AKT. This leads to induction of E-cadherin transcriptional repressors and ultimately to the loss of E-cadherin (Figure 3-8A and Figure 3-8B). The agreement of our network model with known experimental outcomes of TGFβ-driven EMT supports the validity of our model.

In order to explore the possibility of other steady states besides the previously described EMT steady state we performed a state space analysis of the model. The state space can be represented as a network (the state transition network), whose nodes represent the system’s states and whose edges represent every state transition that is possible for the given system. The state transition network summarizes all the possible trajectories or ways that a network state can lead to another network state. Steady states of the system are states that only have transitions leading to them but not leading out of them. The state space grows exponentially with the size of the network, making it impossible to fully map for large networks. To decrease the state space, we performed network simplification of the 70 node, 135 edge EMT network in such a way that the network remains consistent with causal experimental observations (32). This network reduction has been shown to have no effect on the permitted dynamic behaviors of a system (33). This led to a network with 19 nodes and 70 edges (Figure 3-9A and Table 3-5).
Figure 3-8. Network modeling reproduces established signaling abnormalities seen in TGFβ-driven epithelial-to-mesenchymal transition and predicts novel cross-talk mechanisms. A) Heat map representing the activity (averaged state) of all nodes according to the model (10,000 simulations, asynchronous update). The initial state corresponds to a prototypical epithelial state; the final state indicates a TGFβ-driven EMT steady state. Nodes whose state did not change during the simulation are not included. Node activity level of 0 (black square) indicates that the node is inactive (OFF) in all the simulations. Node activity level of 1 (white square) indicates that the node is active (ON) in all the simulations. B) Time-lapse dynamic simulation (10,000 simulations, asynchronous update) reveals induction of SNAI1 expression and loss of E-cadherin expression during induction of EMT from an epithelial initial condition. C) Time-lapse dynamic simulations (10,000 simulations, asynchronous update) reveal induction of Sonic Hedgehog signaling and downstream GLI expression, Wnt signaling and AXIN2 expression with induction of EMT from an epithelial initial condition.
Figure 3-9. Network reduction and state space analysis of the EMT network. A) Multiple network reduction techniques were applied to the 70 node, 135 edge EMT network to reduce it in size to a 19 node, 70 edge EMT network. This reduction maintains the network consistent with causal experimental observations, while permitting state space analysis that was not feasible on a larger network. The symbol shadings are the same as on Figure 3-7. The Boolean rules for the reduced EMT Network are included in Table 3-5. B) In the reduced 19 node network, TGFβ was fixed as ON and the partial state transition network was explored from the initial epithelial condition, using a ranked asynchronous updating algorithm. These findings demonstrated that a single EMT steady state is realized through multiple trajectories, and that this steady state results from constitutive TGFβ signaling without dysregulations of any other nodes in the EMT network. This partial state transition network has 3,489 nodes (states) and 16,434 edges. Nodes are colored according to the number of stable motifs they have activated, with blue representing no stable motifs and yellow representing all of the stable motifs. The large blue node represents the initial epithelial condition just before TGFβ is turned ON, and the large yellow node represents the EMT steady state. The most probable trajectory in the state transition network is highlighted with large red arrows. All possible trajectories lead to the EMT steady state. C) The four most common trajectories in the state transition network, starting from the epithelial state (large blue node on the left) and terminating at the mesenchymal steady state (large yellow node at the right). Each node represents a network state. Each edge represents a transition between two states. Since a general asynchronous updating algorithm is implemented, a single node turns ON or OFF during each state transition; this node and its corresponding new state are indicated above the edge. For example, the second transition of all four trajectories corresponds to SMAD turning ON from an initially OFF state in the epithelial initial condition. The transitions corresponding to each of the four trajectories are identified by different line types (solid red, dashed, dash-dotted, dotted).
### Table 3-5: Boolean rules and initial state for the 19 node reduced EMT network

<table>
<thead>
<tr>
<th>Update rules</th>
<th>Epithelial initial condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AKT* = SMAD or SOS/GRB2 or not GSK3β or β-catenin_nuc</td>
<td>AKT = OFF</td>
</tr>
<tr>
<td>AXIN2* = AXIN2 or β-catenin_nuc</td>
<td>AXIN2 = OFF</td>
</tr>
<tr>
<td>β-catenin_mem** = E-cadherin and not β-catenin_nuc</td>
<td>β-catenin_mem** = ON</td>
</tr>
<tr>
<td>β-catenin_nuc* = not Dest_compl and not β-catenin_mem** and (SMAD or GLI or not E-cadherin)</td>
<td>β-catenin_nuc* = OFF</td>
</tr>
<tr>
<td>Dest_compl* = (GSK3β and AXIN2 and β-catenin_nuc) or (GSK3β and Dest_compl)</td>
<td>Dest_compl* = ON</td>
</tr>
<tr>
<td>E-cadherin* = β-catenin_mem** and (not SNAI1 or (not NOTCH and not SMAD) or not ZEB1 or not ZEB2 or not TWIST1 or not SNAI2)</td>
<td>E-cadherin* = ON</td>
</tr>
<tr>
<td>GLI* = β-catenin_nuc or SMAD or GLI</td>
<td>GLI* = OFF</td>
</tr>
<tr>
<td>GSK3β* = not GLI and not AKT</td>
<td>GSK3β* = ON</td>
</tr>
<tr>
<td>MEK* = SOS/GRB2 or not GSK3β or β-catenin_nuc or not (not MEK or not SNAI1)</td>
<td>MEK* = OFF</td>
</tr>
<tr>
<td>NOTCH* = SOS/GRB2 or not GSK3β or β-catenin_nuc or SMAD</td>
<td>NOTCH* = OFF</td>
</tr>
<tr>
<td>SMAD* = (MEK or TGFβR) and (ZEB1 or not ZEB2)</td>
<td>SMAD* = OFF</td>
</tr>
<tr>
<td>SNAI1* = GLI or SMAD or β-catenin_nuc or TGFβR or NOTCH or MEK or AKT or not GSK3β</td>
<td>SNAI1* = OFF</td>
</tr>
<tr>
<td>SNAI2* = MEK or β-catenin_nuc or SNAI2 or TWIST1</td>
<td>SNAI2* = OFF</td>
</tr>
<tr>
<td>SOS/GRB2* = (β-catenin_nuc or TGFβR) and not MEK</td>
<td>SOS/GRB2* = OFF</td>
</tr>
<tr>
<td>miR200* = not SNAI1 and not ZEB1 and not ZEB2</td>
<td>miR200* = ON</td>
</tr>
<tr>
<td>TGFβR* = SNAI1 or TWIST1 or GLI</td>
<td>TGFβR* = OFF</td>
</tr>
<tr>
<td>TWIST1* = AKT or β-catenin_nuc or SNAI1</td>
<td>TWIST1* = OFF</td>
</tr>
<tr>
<td>ZEB1* = SNAI1 and not miR200</td>
<td>ZEB1* = OFF</td>
</tr>
<tr>
<td>ZEB2* = SNAI1 and not miR200</td>
<td>ZEB2* = OFF</td>
</tr>
</tbody>
</table>

In the reduced 19 node network, TGFβ was fixed as ON, and the part of the state space reachable from the otherwise epithelial initial condition was determined, using a ranked asynchronous updating algorithm. The resulting partial state transition network has 3,489 nodes (states) and 16,434 edges (state transitions) (Figure 3-9B). Only one node of the state transition network has no outgoing edges, the node corresponding to the EMT steady state. Thus this steady state is the only possible outcome for an epithelial cell in the presence of sustained TGFβ signaling; no other steady states were discovered. This steady state can be reached through many thousands of trajectories, each corresponding to a different set of state transitions. For example, the four most common trajectories (Figure 3-9C) include E-cadherin loss prior to nuclear localization of β-catenin, or following it. As trajectories in the model reflect the order of events possible under different kinetics, their convergence into a single outcome suggests that the TGFβ-driven EMT outcome is robust across all reasonable values of kinetic parameters. We note that none of the scenarios to reach EMT encapsulated in the state transition network involve specific dysregulations (e.g. mutational gain or loss of function) of any nodes in the EMT
network. The presence of constitutive TGFβ signaling drives all dysregulations we see in our model.

### 3.5.5 Network Modeling Reveals that the Constitutive Presence of TGFβ Induces Wnt and SHH Signaling During Epithelial-to-Mesenchymal Transition in Hepatocellular Carcinoma

Interestingly, our dynamic model of signaling induced by TGFβ activation revealed joint activation of Wnt and Sonic hedgehog (SHH) signaling pathways during induction of EMT from an epithelial initial condition (Figure 3-8A and Figure 3-8C). In order to test the predictions from our model we explored molecular readouts of Wnt and SHH signaling. AXIN2 is a well-established transcriptional marker of Wnt/β-catenin/TCF signaling and is a negative regulator of the Wnt signaling pathway (34). Furthermore, members of the GLI family of transcription factors (represented by the node GLI in our model) are major effectors of the SHH signaling pathway, and their mRNA expression can be used as a readout of SHH signaling (35). GLI2 is the major transcriptional activator of the SHH signaling pathway, with GLI2 itself undergoing transcriptional up-regulation when SHH signaling is active (36). Most evidence supports that GLI1 is activated transcriptionally by GLI2 (37). Indeed in our model, simulations (10,000 simulations, asynchronous update) starting from an epithelial state in the sustained presence (ON state) of Wnt led to AXIN2 expression (Figure 3-10A) and the presence of SHH in the same network model. In the experimental literature GLI expression is an established marker of SHH signaling and AXIN2 expression is an established marker of canonical Wnt signaling. A) Activating SHH signaling, while keeping all other signals inactive, leads to induction of GLI in our model (1000 simulations, asynchronous update), thus supporting experimental findings. B) Activating Wnt in our Boolean model of EMT, while keeping all other signals inactive, produces AXIN2 (1000 simulations, asynchronous update).

![Figure 3-10](image)

Figure 3-10. AXIN2 and GLI are induced by Wnt and SHH signals in the epithelial-to-mesenchymal transition network model. In the experimental literature GLI expression is an established marker of SHH signaling and AXIN2 expression is an established marker of canonical Wnt signaling. A) Activating SHH signaling, while keeping all other signals inactive, leads to induction of GLI in our model (1000 simulations, asynchronous update), thus supporting experimental findings. B) Activating Wnt in our Boolean model of EMT, while keeping all other signals inactive, produces AXIN2 (1000 simulations, asynchronous update).
initial state drove GLI expression (Figure 3-10B). These results affirm that our model reproduces known molecular readouts of Wnt and SHH signaling.

As shown in Figure 3-8C, TGFβ signaling induces both AXIN2 and GLI (with a delay compared to SHH) in our model. In order to experimentally explore the activation of Wnt and SHH signaling pathways by TGFβ, we tested TGFβ-driven expression of pathway markers using epithelial and mesenchymal cell lines. Treatment of the P2E cell line (murine epithelial HCC cell line) with TGFβ leads to induction of GLI2 mRNA and protein expression and induction of AXIN2, approximating the levels constitutively observed in mesenchymal P2M cells (A and Figure 3-11B). TGFβ treatment drives GLI1 and GLI2 expression in human epithelial-like HCC cell lines Huh7 and PLC/PRF/5 cells. Furthermore, AXIN2 is up-regulated by TGFβ in Huh7 and PLC/PRF/5 cells after TGFβ treatment. Additionally, human mesenchymal HCC cell lines HLE and HLF exhibit up-regulated GLI expression relative to epithelial PLC/PRF/5 cells (Figure 3-11C). Inhibition of TGFβ signaling with a chemical inhibitor of TGFβRI/II diminished pSMAD3 and GLI2 protein expression in the mesenchymal HLE and HLF cell lines (Figure 3-11D). Thus our measurements validate the model prediction that Wnt and SHH signaling is induced by TGFβ during induction of EMT.

To further test the joint activation of SHH signaling pathways by TGFβ, we determined the correlation between TGFβ and GLI2 expression in human HCC patient samples. Indeed, SMAD3 phosphorylation correlated positively ($R^2 = 0.629$) with GLI2 expression in these samples (Figure 3-11E). This information supports the network model’s prediction that induction of GLI2 occurs in human tumors with active TGFβ signaling.
For a deeper analysis of the nodes associated with steady states in the EMT network, a stable motif analysis was performed. This analysis is based on finding feedback loops that stabilize in a fixed state (23). Our findings further demonstrate the existence of a single EMT steady state (i.e. the state shown on Figure 3-8A). There are eight motifs associated with this steady state (Figure 3-12A and Figure 3-4). These motifs represent interaction and node activity patterns whose sustained presence is sufficient to stabilize the mesenchymal state. The motif shown on the left panel of Figure 3-12A indicates that sustained Wnt and GLI signaling and a low level of GSK3β

Figure 3-11. Network model-directed testing validates activation of Sonic hedgehog and Wnt signaling pathways by TGFβ. A) GLI and AXIN2 transcripts are elevated in murine mesenchymal cell line P2M and by TGFβ1 treatment (5 ng/ml, 48 hours) of murine epithelial HCC cell line, P2E. B) GLI2 expression is elevated by TGFβ1 treatment (5 ng/ml, 48 hours) of epithelial P2E cells and in mesenchymal P2M cells relative to P2E cells. C) Human epithelial-like HCC cell line Huh7 has low expression of GLI1, GLI2, and AXIN2 and these transcripts are increased by TGFβ1 treatment (5 ng/ml, 48 hours). Epithelial-like PLC/PRF/5 cells treated with TGFβ1 (10 ng/ml, 48 hours) exhibit increased GLI1 and GLI2 expression. GLI1 and GLI2 expression were also elevated in mesenchymal phenotype cell lines with constitutively active TGFβ signaling, HLE and HLF. Treatment of PLC/PRF/5 cells with TGFβ1 (10 ng/ml, 48 hours) led to increased AXIN2 expression. D) Treatment of HLE and HLF cells with TGFβRI inhibitor, led to dose dependent loss of GLI2 levels. E) In protein extracted from nine human HCC patient biopsies, GLI2 expression correlated with pSMAD3 levels ($R^2 = 0.6293$).
can maintain a low level of E-cadherin, the hallmark of the EMT steady state. The motif in the middle panel of Figure 3-12A indicates that a low level of GSK3β can be sustained by cross talk between MAPK and SMAD signaling. The motif on the right side of Figure 3-12A indicates that SMAD signaling can be activated by the TGFβ signal in our model, which itself can be sustained by positive feedback. Read from right to left, these motifs are consistent with the propagation of the TGFβ signal that leads to SMAD and SNAI1 activation, then to the activation of the SHH and Wnt pathways, and then to the loss of E-cadherin. However, because each stable motif is independent of the others and is sufficient to maintain the EMT steady state, the results shown on Figure 3-12A suggest that other, potentially TGFβ independent, mechanisms that lead to the stabilization of a motif (e.g. MAPK signaling) can also drive EMT. Indeed, evidence for the induction of EMT through MAPK pathway activation of SNAI1 has been reported in other studies (38). In our model, MAPK signaling can drive EMT, independent of TGFβ (data not shown). Taken together, our studies demonstrate that activation of SHH and Wnt pathways are a key conserved feature of hepatocellular carcinoma EMT.

3.6 Discussion

We began our studies when we realized that TGFβ signaling mimics the molecular phenotype of a murine mesenchymal cell counterpart (P2M) to an epithelial HCC cell line (P2E). We next identified other HCC cell lines with active TGFβ signaling and demonstrated that TGFβ-active cell lines also had mesenchymal features. These results suggest TGFβ signaling is a conserved dysregulation and driver of EMT across multiple HCC models. In order to better understand the complex signaling events producing the TGFβ-driven EMT phenotype, we curated signaling pathways involved in developmental and carcinoma EMT programs in HCC and compiled them into an EMT signaling network. By formulating a Boolean dynamic model of this network, we were able to identify downstream signaling events involved in the TGFβ-specific EMT process and detect cross-talk among pathways. As EMT and TGFβ signaling are conserved features of many solid tumor types, this model (or modifications of it) may be used to study EMT in other tumor types.
Figure 3-12. Network motifs that stabilize in the EMT=ON steady state. The color of the node indicates the node’s stabilized state: Black node = OFF; White node = ON. A) The left-most motif represents a Wnt/β-catenin feedback loop in which cross-talk with GLI stabilizes low E-cadherin expression. The middle motif is a feedback loop consisting of SMAD/MAPK cross-talk, which suppresses GSK3β a negative regulator of Wnt signaling. The right-most motif is a TGFβ signaling feedback loop, which drives SMAD activity. B) Network analysis supports that SMAD-dependent and SMAD-independent mediators are important for TGFβ-driven EMT. SMAD knockout has a nominal effect on TGFβ induction of EMT (10,000 simulations, asynchronous update). C) Network motifs stabilizing the EMT steady state in the presence of SMAD knockout (when TGFβ=ON and SMAD=OFF). Targeting these motifs may lead to effective combinations with SMAD inhibition to disrupt TGFβ-driven EMT.

TGFβ plays a dual role as a tumor suppressor and oncogene; however it remains unclear why. We argue that in the cell, network-level emergent outcomes (e.g. tumor suppression versus tumor growth) are not attributable to the TGFβ signal alone, but are dependent on the status of the entire cellular network that the signal acts upon. Network models are increasingly used to understand the organization and emergent properties of biological systems (39, 40). Network-based discrete dynamic models are particularly useful in poorly characterized biological systems, where they can be used to generate testable hypotheses (41). Boolean network models have led to new insight into signal transduction and gene regulatory networks in numerous organisms.
Network modeling can explain why in certain differential contexts (e.g. for different states of a few nodes other than TGFβ), the TGFβ signal produces different outcomes, which can then be investigated experimentally.

Perturbations in signaling pathways that define a disease or a specific feature of a disease can result from deregulation of only a subset of these pathways. However, identifying such a subset is difficult using traditional biological approaches. In this study, we assessed this question in EMT, a mechanism that drives invasion and metastasis in solid tumors, by using a Boolean dynamic model. Through network simulations, we revealed that the TGFβ signal could percolate across multiple signaling pathways, leading to a network of perturbed proteins during the induction of EMT. Across multiple murine and human HCC models, the constitutive presence of TGFβ was shown to be sufficient to induce Wnt and SHH signaling pathways in this context. The model’s result regarding the order of induction of the Wnt and SHH pathways obtained from the model may depend on the broad classification used for the time lags (transcriptional versus signal transduction events). A future experimental determination of the order of pathway activation may support the model result or may contradict it, in which case further timing constraints could be incorporated to establish agreement. Imposing timing constraints would affect how fast the TGFβ signal propagates through the network but it would not affect the network’s steady states.

It should be noted that TGFβ, Wnt and SHH signaling pathways have individually been implicated in oncogenic processes, including EMT in carcinomas. For example, Wnt pathway induction of EMT has been demonstrated in multiple cancer types (42). Furthermore, SHH can induce EMT in brain and gastrointestinal cancer models, among others (43). It has recently been suggested that the critical interaction between TGFβ and Wnt signaling pathways deserves special attention (44). Here we present a systems level explanation of the interaction of TGFβ signaling with the Wnt and SHH signaling pathways. How these signaling pathways act as a network to integrate a signal for a specific biological process has not been realized by previous studies. Our stable motif analysis revealed that the SHH/GLI and Wnt components of the EMT network are critical nodes in feedback loops that maintain the EMT steady state (Figure 3-12A).

EMT is a process that occurs throughout development and is hijacked by carcinomas during solid tumor progression. Tumor metastasis is the leading cause of mortality in solid tumors. Our model provides a strong foundation for the modeling of HCC invasion. As hepatic tissue lacks a
membrane barrier, it has been speculated that EMT is the major mechanism for the entry of migratory HCC cells into systemic circulation (3). To describe carcinomas of epithelia in other tissues, which do have a basement membrane, our model can be expanded by adding the pathway(s) regulating the proteolytic destruction of the basement membrane by matrix metalloproteinase proteins (45). Another possible extension is an agent-based model in which each cell is an agent with a specific morphology. Such a model could incorporate our Boolean signal transduction model, the initial epithelial architecture, as well as the reorganization of the actin cytoskeleton of each cell to enable directional motility and dynamic cell elongation (46).

Understanding the systems-level dysregulations in EMT is important not only to elucidate tumor progression and the metastatic process but also to develop therapeutic interventions to halt/suppress metastatic disease. Our analysis suggests that single interventions may not be effective. For example, SMAD proteins are the canonical downstream targets of TGFβ signaling. However, our model predicts that inhibiting SMAD proteins alone cannot completely block TGFβ-driven EMT (Figure 3-12B). Indeed, among the motifs that can sustain an EMT=ON state there is a TGFβ feedback loop that does not include SMAD (Figure 3-4). Stable motif analysis with TGFβ=ON and SMAD=OFF (a TGFβ active network with no SMAD signaling) reveals multiple feedback loops that can maintain EMT in the absence of SMAD signaling (Figure 3-12C). This implies that combining SMAD inhibition with Wnt, SHH, AKT, or MAPK inhibition may more effectively suppress EMT. Similar computational analyses have been completed on smaller networks, supporting a role for feedback loops in the regulation of EMT (29), EMT plasticity (47), and tumor cell heterogeneity (non-genetic heritability) (48). The concept of combinatorial inhibition will be explored in Chapter 4.

Combinatorial therapeutics is an emerging discipline in the treatment of cancer patients. Recent work indicates that targeting multiple oncogenic signals through combination of cancer therapies is superior to single therapy treatment (49). The development of effective combined therapeutics is an important goal within this field. Developing dynamic models of cancer networks to simulate single and combinatorial interventions targeting specific nodes may prove to be a rational approach to more efficiently choose single and multi-drug therapeutic regimens.

In summary, we used network analysis and Boolean modeling to investigate the signaling abnormalities that arise during EMT in the context of TGFβ signaling and confirmed our network-based results in multiple experimental liver cancer models (Figure 3-3). Our systems
biology approach was able to maximize the use of the available pathway information and to identify putative activation of major signaling pathways during EMT and liver cancer invasion. This study confirms the ability to integrate normal and disease pathway information into a single model that is robust enough to reproduce a clinically relevant complex process yet is constructed with qualitative information.

Acknowledgments

This work was previously published as “Steinway SN, Zañudo JG, Ding W, Rountree CB, Feith DJ, Loughran TP Jr, Albert R. Network modeling of TGFβ signaling in hepatocellular carcinoma epithelial-to-mesenchymal transition reveals joint sonic hedgehog and Wnt pathway activation. Cancer Res. 2014.” Research reported in this publication was supported by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health under Award Number F30DK093234. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors thank Dr. Hien Dang (NIH/NCI) for training and technical expertise at the onset/during pilot studies. We would like to thank Dr. Xin Liu (Penn State College of Medicine) for antibodies. This work was partially supported by grants NIH Ruth L. Kirschstein National Research Service Award NIH F30DK093234 (Steinway), NSF IIS 1161001, NSF PHY 1205840.

References

CHAPTER 4.
Combinatorial Interventions Lend Insight into the Inhibition of TGFβ Driven Epithelial-to-Mesenchymal Transition and the Existence of Hybrid Epithelial-Mesenchymal Phenotypes

4.1 Abstract

Epithelial-to-mesenchymal transition (EMT) is a developmental process hijacked by cancer cells to leave the primary tumor site, invade surrounding tissue, and establish distant metastases. A hallmark of EMT is the loss of E-cadherin expression, and one major signal for the induction of EMT is transforming growth factor beta (TGFβ), which is dysregulated in up to 40% of hepatocellular carcinoma (HCC). We previously constructed and experimentally validated an EMT network of 69 nodes and 134 edges by integrating the signaling pathways involved in developmental EMT and known dysregulations in invasive HCC. Our previous investigation suggested that suppression of canonical TGFβ signaling through inhibition of the SMAD complex was insufficient to completely inhibit the TGFβ-driven EMT. We hypothesized that inhibiting the SMAD complex in combination with some other node may sufficiently inhibit EMT. Thus, we used this as motivation to systematically explore perturbations to individual and combinations of nodes in the EMT network that suppress TGFβ-driven EMT, with the ultimate goal of identifying therapeutic interventions which suppress tumor invasion. Our results predict that targeting a key subset of nodes can effectively suppress TGFβ-driven EMT and that suppressing the SMAD complex in combination with other nodes is more effective than the SMAD complex alone. Out of close to half a million perturbations, our model predicts that 13 targeting strategies robustly suppress TGFβ-driven EMT by disrupting feedback loops that drive the EMT process. We have demonstrated experimentally that expression of a majority of these nodes is enriched in mesenchymal relative to epithelial phenotype HCC cell lines. An siRNA screen targeting these combinations suggests that many of these targeting strategies suppress TGFβ-driven EMT. Additionally, we reveal that some perturbations produce steady states that differ substantially from previously identified epithelial and mesenchymal steady states in our model. Further analysis reveals that these perturbations lead to states that are intermediate to epithelial and mesenchymal phenotypes. Similar hybrid (partial) EMT states have been described
in the literature. Quantitative analysis of these attractors reveals that these hybrid states form a subset of steady states that are distinct from epithelial and mesenchymal steady states.

4.2 Introduction

Epithelial-to-mesenchymal transition (EMT) plays important roles in normal physiological processes, such as development and wound healing, and also in pathological processes, such as fibrosis and cancer metastasis (1, 2). During EMT, cells gain the ability to invade and migrate through a loss of epithelial characteristics, such as tight cellular junctions, and acquisition of mesenchymal attributes, such as loss of cellular adhesion and increased motility (1, 3). Loss of E-cadherin, a cell adhesion protein, is considered the hallmark of EMT. Because tumor invasion and metastasis are dependent upon EMT, targeting invasion through the EMT process could reduce solid tumor mortality substantially (4-6). In addition to the epithelial and mesenchymal cellular states, there exists evidence of an intermediate EMT phenotype, known as the hybrid or partial EMT state (7-9). The hybrid EMT state is characterized by some epithelial and some mesenchymal features (9, 10). Like the mesenchymal state, hybrid EMT states may be relevant to pathological processes, such as tumor invasion and metastasis (5, 11).

Transforming growth factor-β (TGFβ1, hereafter referred to as TGFβ) is one of numerous signals that can induce EMT (12), and is frequently dysregulated in HCC (13-15). TGFβ signals through a complex network of proteins during the induction of EMT. After binding to the TGFβR complex, TGFβ activates canonical (SMAD) and non-canonical (MAPK and Akt) signaling components. The TGFβ signal converges on a core transcriptional regulatory network (SNAI1, SLUG, ZEB1, ZEB2, HEY1, FOXC2, and TWIST), which directly modulates E-cadherin transcription. The TGFβ pathway interacts with many other intracellular proteins and other EMT signals, and it turns out that cross-pathway components are part of numerous feedback loops and regulatory motifs that drive EMT. Thus, the complexity of the EMT signaling network makes it difficult to fully understand using classical experimental dissection.

Efforts have been made to use mathematical models of various sizes to study the properties of TGFβ signaling and EMT. Two smaller models focusing on ZEB, SNAI1, miR-200, and miR-34, have helped elucidate the transcriptional mechanisms driving E-cadherin expression and support the existence of a hybrid EMT state (16, 17). The model by Lu et al., suggests that miR-200 and ZEB act as a ternary switch, allowing for epithelial, mesenchymal and hybrid cellular
phenotypes. In this model, levels of miR-200 and ZEB determine the cellular state (17). The model by Tian et al. proposed a system of cascading bistable switches. The SNAI1/miR-34 feedback loop in this model produces a reversible switch and regulates initiation of EMT, whereas the ZEB/miR-200 feedback loop is responsible for an irreversible switch that leads to the mesenchymal state. Because each of the motifs in the Tian model serves as a bistable switch, this allows the existence of an intermediate (hybrid) EMT state, with only the first switch turned ON at an intermediate stimulus strength and duration (16).

These models have led to exciting insights into EMT regulatory mechanisms; however, they take only a small number of regulatory components into consideration and thus could be missing important dynamics. Because of the large number of participants in the TGFβ/EMT network, larger systems level studies are needed to understand this complex process. We previously developed a systems level model of EMT, consisting of 69 nodes and 144 edges, which incorporates a comprehensive map of the growth factors, receptors, signal transductions proteins, and transcription factors that regulate E-cadherin expression. Our network model produced dynamics consistent with known TGFβ and EMT signaling. Furthermore, our model suggests that crosstalk with other signaling pathways is critical to the induction of EMT (18). Interestingly, our previous investigation suggested that suppression of canonical TGFβ signaling through inhibition of the SMAD complex was insufficient to completely inhibit the TGFβ-driven EMT. We hypothesized that inhibiting the SMAD complex in combination with some other nodes may sufficiently suppress EMT. Thus, we used this as motivation to systematically explore perturbations to individual and combinations of nodes in the EMT network that suppress TGFβ-driven EMT, with the ultimate goal of identifying therapeutic interventions which suppress tumor invasion.

Here we use the previously developed EMT network model to systematically explore close to half a million perturbations in the EMT network to identify individual targets and combinations of targets (up to four nodes) that putatively suppress EMT, with the goal of identifying optimal targets to suppress invasive features of cancer cells. We demonstrate that whereas such a screen might be very time consuming and expensive experimentally, it is quite feasible computationally. We identify, through perturbation analysis and application of the EMT network model, that of the large number of perturbations that were screened, only dozens are predicted to suppress EMT. Interestingly, our results suggest that many inhibitory combinations require suppressing
the SMAD complex in combination with other nodes. We use this smaller subset of predictions
to perform a multi-faceted *in vitro* experimental screen to identify their effect on EMT inhibition.
Furthermore, during the course of the perturbation analysis, we believe we have uncovered
steady states that fall into the EMT hybrid steady state category. We believe these steady states
may be an extension of hybrid steady states previously described both computationally and
experimentally.

4.3 Materials & Methods

4.3.1 Cell culture

Huh7 cells were provided by Dr. Harriet Isom (Penn State College of Medicine, Hershey, PA) (19). HepG2 and PLC/PRF/5 (Alexander) cells were acquired from the American Tissue Culture Collection. Human HLE cells were acquired from Dr. Curtis Harris (National Cancer Institute). HLF cells were acquired from Dr. Jorge Filmus (University of Toronto, Toronto, ON, Canada)(20). PLC/PRF/5 cells were acquired from the ATCC. All cells were cultured in DMEM medium (Life Technologies) + 10% FBS (Atlanta Biologicals). TGFβ treatments were completed with recombinant human TGFβ (Peprotech). Complete cell culture media was removed and cells were washed with phosphate buffered saline. Cells were serum starved by incubating them in serum free growth media for three hours, followed by addition of TGFβ to the serum free media at specified doses. Cells were maintained in a humidified 5% CO2 37°C incubator.

4.3.2 siRNA transfections

siRNA transfections were performed in Huh7 cells to screen the effect of *in silico* predicted
node knockdown combinations in TGFβ induced EMT compared to TGFβ untreated controls.
Invitrogen Silencer Select siRNAs (Life Technologies) to specific targets genes or a scrambled
siRNA negative control (#4390846) were used. 2 nM siRNA per siRNA and 5 µl Lipofectamine
RNAiMAX were added to 1 ml Opti-MEM I medium (Life Technologies) without serum, then
combined with 7.5x10^5 cells in 5 ml complete media. Cells were then plated for analysis of
mRNA and protein analysis and also cell migration.
4.3.3 Quantitative real-time polymerase chain reaction

The Ambion Cells-to-C<sub>T</sub> Kit (Life Technologies) was used to extract RNA and reverse transcribe RNA to complementary DNA for quantitative real-time polymerase chain reaction (qRT-PCR) without having to purify RNA prior to amplification as per manufacturer’s instructions. qRT-PCR experiments were conducted using a CFX384 Touch Real-Time PCR Detection System (Bio-rad) and TaqMan Universal PCR Master Mix (Life Technologies). The housekeeping genes GAPDH and β-actin were used for ΔΔCt calculations and relative expression was calculated using Taqman primer/probe sets (Life Technologies).

4.3.4 Immunoblot analysis

All cells were lysed in RIPA buffer (Sigma Aldrich) with 1:100 P8340 protease inhibitor (Sigma Aldrich) and phosphatase inhibitor cocktail 2 (Sigma Aldrich). Protein concentrations of lysates were determined using the BCA Protein Assay Kit (Thermo Scientific). 30 µg of protein lysates were separated on a NuPAGE 4-12% Bis-Tris Gel (Invitrogen) and transferred to polyvinylidene difluoride (PVDF) membrane (Bio-rad). Membranes were stained in Ponceau S solution (Sigma Aldrich) to confirm protein transfer. Primary antibodies for E-cadherin, Vimentin, RAS, NOTCH1, NOTCH2, NOTCH3, NOTCH4, CSL, SOS1, GRB2, and β-actin (Cell Signaling) were used for this study. Blots were blocked in 5% BSA for 1 hour before incubation overnight with the appropriate primary antibody. Chemiluminescent signal was detected using anti-mouse or anti-rabbit horseradish peroxidase–conjugated secondary antibody (Cell Signaling Technology) and Clarity ECL (BioRad) on the Chemidoc XRS<sup>+</sup> system (Bio-Rad). Protein bands were analyzed and quantified using the Image Lab software suite (Bio-Rad). Quantified Vimentin and E-cadherin bands for each siRNA knockdown combination (Figure 4-5B) were normalized to the β-actin band (protein loading control). Next, for each siRNA knockdown sample, relative E-cadherin and vimentin in TGFβ treatment samples relative to TGFβ untreated samples was calculated. Finally TGFβ treatment relative to TGFβ untreated samples was calculated relative to the scrambled control, giving the effect of node knockdown combination on TGFβ driven EMT, relative to the effect of TGFβ on EMT markers with a scrambled siRNA control. All protein bands were within the linear range as determined by Image Lab.
4.3.5 Detection of cell migration using automated cell imaging acquisition and analysis

The Oris Cell Migration Assay (Platypus Technologies) was used to assess the effect of in silico predicted node combinations on TGFβ-driven cell migration. The Oris Cell Migration Assay comes in 96 well plates with stoppers, which exclude cells from attaching to the centers of the well. Once the stoppers are removed, a cell-free “detection zone” is revealed into which cells can migrate. Following siRNA transfections (described above), Huh7 hepatocellular carcinoma cells were plated in wells of Oris Cell Migration Assay 96 well plates. Cells were plated at a density of 1.5x10⁴ per well, and placed at 37°C in a humidified 5% CO₂ incubator for 48 hours. At 48 hours, migration stoppers were removed from all the wells, and cells were stained with a whole cell fluorescent stain CellTracker Green Dye (2.5 mM; Life Technologies) as per the manufacturer’s recommendations. The whole cell fluorescent stain was used for fluorescent imaging of cells to quantify migration at multiple time points throughout the experiment. After staining, cells were serum starved for three hours. For wells in the TGFβ treatment group, TGFβ was added to wells at a final concentration of 5 ng/ml. Twelve replicate wells were used per transfection: six for TGFβ treatment and six for control cells untreated with TGFβ. At 0, 24, and 48 hours after stoppers were removed (48, 72, and 96 hours post transfection), cells were imaged with the Cytation 3 automated fluorescent cell imager (Biotek). Cells were imaged using a 4X objective and GFP filter. Four images were taken with auto exposure in two by two dimensions such that the entire detection zone and surrounding well area were captured. The four images were spliced together using Gen5 version 2.0.7 Software (Biotek) in order to construct a continuous image of the detection zone and surrounding area for image based quantification of the detection zone area using ImageJ.

4.3.6 Quantification of cell migration with ImageJ

Cell migration was assessed by measuring the area of the detection zone at the pre-migration and 24 and 48 hour migration time points using ImageJ 1.48v analysis software (21) by modifying a protocol as previously described (22). To measure the area of the detection zone from fluorescence images, the image scale was set using the known micron/pixel values (Analyze-->Set Scale). Next, the threshold was set to enhance the contrast between the detection zone and the cell monolayer (Image-->Adjust-->Threshold). Manual adjustment of the threshold was completed in selected cases where the software-chosen threshold did not correctly identify
the detection zone boundary. By selecting “Apply” in the threshold window, the thresholded image was converted to a binary image. The Wand (tracing) tool was clicked on the Detection Zone until the region was outlined (indicated by a yellow trace). Using the menu command Analyze--->Set Measurements, the “Area” and “Display Label” were selected. Finally, the menu command Analyze--->Measure was selected to calculate the detection zone area. The “Area” measurements for each well at each time point were used to calculate the percent migration:

\[
\text{area}_{t=0} - \text{area}_{t=\text{timepoint}} \times 100 \over \text{Area}_{t=0}
\]

The mean percent area closure was calculated from the six replicate wells. The effect of each siRNA knockdown combination on TGFβ-driven migration was determined by calculating the percent area closure in TGFβ treated siRNA knockdown combinations relative to cells with siRNA knockdown combination that were not treated with TGFβ.

TGFβ treated siRNA knockdown combination

siRNA knockdown combination untreated with TGFβ

The TGFβ treated relative to TGFβ untreated percent migration for each node knockdown combination was normalized to scrambled siRNA control TGFβ treated relative to untreated percent area closures. This normalization scheme allows us to know the effect of node knockdown combinations on TGFβ-driven migration relative to a scrambled siRNA control sample.

4.3.7 Network modeling framework

We use a Boolean framework in which each network node is described by one of two qualitative states: ON or OFF. The ON (logical 1) state means an above threshold protein activity or abundance whereas the OFF (logical 0) state means below-threshold activity or abundance. The biological relationships among nodes in the EMT network are expressed as mathematical equations using Boolean operators (18, 23).

4.3.8 Dynamic Analysis

Dynamic analysis is performed by applying the Boolean functions in succession until a steady state is reached. Boolean models and discrete dynamic models in general focus on state transitions instead of following the system in continuous time. Thus, time is an implicit variable
in these models. The network transitions from an initial condition until an attractor is reached. An attractor can be a fixed point (steady state) or a set of states that repeat indefinitely (a complex attractor). The basin of attraction refers to the initial conditions that lead the system to a specific attractor, and the state transition network describes all possible trajectories from initial conditions to steady states in the network. We made use of a ranked asynchronous (stochastic) updating scheme to simulate network dynamics. In this updating scheme, we update nodes regulated by signal transduction events with a greater probability than nodes regulated by transcriptional events in order to account for signal transduction events occurring faster than transcriptional events, as previously described (18). The employed updating scheme does not affect fixed points identified by our model, because fixed points are independent of the implementation of time; however, we previously found that this updating scheme more accurately represented real dynamics. The simulations of the model were performed using custom Java and Python code (BooleanNet Python library (24)).

4.3.9 Perturbation Analysis

To capture the effect of knocking out or inhibiting nodes in the network model, modification of the states/rules to describe knockout or overexpressed states were performed. These modifications were implemented by setting the corresponding nodes to either OFF (knockout) or ON (overabundance) and then removing the corresponding updating rules for these nodes for the simulations. By examining many such forced perturbations, we can identify potential therapeutic strategies, many of which may not be obvious or intuitive, particularly as network complexity increases.

4.3.10 Network Control Methodology

We employed a recently developed methodology that identifies specific nodes that when controlled, lead to a specific cellular state. This methodology is based on finding stable motifs, that is, components of a network that stabilize in specific steady states. The network control approach identifies nodes in the stable motif whose manipulation forces the cell into a specific attractor (Zañudo & Albert, 2015).
4.4 Results

4.4.1 Systematic exploration of perturbations in the EMT network reveal combinatorial inhibition of specific nodes can suppress the TGFβ-driven EMT

Epithelial and mesenchymal cellular phenotypes are stable states that remain as such, unless a perturbation is employed that drives the cells to another state. In previous work, we constructed a dynamic network model of this process and explore the mechanism by which TGFβ drives EMT, interestingly through activation of multiple other pathways (18). We additionally determined that in the EMT network model, suppressing the canonical downstream TGFβ signaling component, the SMAD complex, was insufficient to completely abrogate the induction EMT. In other words what kind of cellular perturbations might suppress the TGFβ-driven EMT (Figure 4-1A). We are particularly interested in how to inhibit the induction of EMT, with particular focus on interventions that suppress EMT in combination with the SMAD complex.

Such relationships can be effectively analyzed using a dynamic model, which is used to express the behavior of a system over time by characterizing each node by a state variable (representing, e.g., its concentration or activity) and a function that describes its regulation. Dynamic models can be categorized as continuous or discrete, according to the type of node state variable used. Continuous models use a set of differential equations; however, the paucity of known kinetic details for biological interactions makes these models difficult to implement, especially for large network models, like the EMT model. Discrete dynamic modeling is used to study signal transduction because of its computational feasibility and capacity to be constructed with qualitative biological data(25). In the simplest discrete dynamic models, called Boolean models, each network node can be described by one of two qualitative states: ON or OFF. The ON state means an above threshold concentration of a molecular regulator whereas the OFF state means the below-threshold or inactivated form. Regulatory relationships are connected using the Boolean operators OR (independent regulation), AND (synergistic regulation), and NOT (inhibitory regulation). As in the biological system, there is a time lag between the state change of the regulators and the state change of the targets (see Methods). Similarly, as there are means to perturb experimental systems (e.g. with drugs, siRNA) in order to identify the effect a protein has on a cell, networks can be perturbed with “overrides” to identify the effect of node perturbations. For example, a negative node override (i.e. a knockout) can be applied to identify the effect of having a node OFF, regardless of any regulatory relationships within the network.
that would change the node state. Positive overrides (i.e. overexpressions) can also be applied to simulate the effect of having a node active regardless of regulatory relationships that would otherwise inactivate it. Advantageously, the effect of node perturbations can be systematically and rapidly applied through a computational screen with essentially no cost, whereas experimental screens can be quite expensive and time consuming, particularly if combination

Figure 4-1. An in silico combinatorial knockout screen in the EMT network reveals specific node combinations can suppress the TGFβ-driven EMT. A previous constructed network model of EMT was used to identify nodes that when knocked out, blocked TGFβ-driven EMT. A) Schematic demonstrating that epithelial and mesenchymal cellular states are stable unless a perturbation (e.g. TGFβ) is employed. B) The effect of knocking out nodes individually and in combinations and their effect on EMT. C) A network control approach was employed to identify the minimal set of nodes, that when their states are controlled, lead to the epithelial steady state. D) All combination knockouts are contained within the epithelial control set (yellow background) or nodes directly upstream of nodes in the epithelial control set (blue shading).
Thus, we use a previously constructed and validated Boolean dynamic model of EMT (18) to systematically explore the effect of individual and combination node perturbations on TGFβ-driven EMT (Table 4-1) with the goal of using computationally identified targeting strategies from this screen for experimental validation. We previously demonstrated that knocking out the SMAD complex led to only minimal suppression of the TGFβ signal because other SMAD-independent feedback loops (network motifs) remained active (18). We sought to comprehensively explore ways to suppress the TGFβ signal and TGFβ-driven EMT, through systematic analysis of node perturbations in the EMT network. Systematic knockout of all nodes individually in the EMT network, reveals that only targeting the seven direct E-cadherin regulating transcriptional factors may robustly suppress TGFβ-driven EMT (Figure 4-1B), which is consistent with the known role of these transcription factors. It was intriguing that similarly to SMAD in our previous study, a majority of the nodes in the EMT network could not suppress the TGFβ signal, likely due to EMT driving network motifs that remained intact or new network motifs arising that can sustain the TGFβ signal in the absence of a single node. We hypothesized that a subset of these nodes could suppress the TGFβ signal in combinations. Thus, we next systematically explored knocking out combinations of two nodes that could not individually suppress TGFβ-driven EMT (Table 1). Of the 1,171 possible combinations of two nodes, we identified six node combinations that could suppress the TGFβ signal in the EMT network model. Interestingly, all these combinations require the SMAD complex as one node in the combination (Figure 4-1B). Of the additional network nodes that could not suppress EMT individually or in combinations of two, we explored combinations of three (Table 1).
Interestingly, of the 23,426 possible combinations of three nodes, none completely inhibited EMT, although some led to partial suppression of EMT (Figure 4-1B). We next explored combinations of four nodes (292,825 total combinations), and similar to the three node combinations, no combinations of four nodes completely suppressed EMT. Thus, our model suggests that a specific subset of nodes may individually suppress the TGFβ-driven EMT and these combinations appear exhausted after combinations of two.

### 4.4.2 Node combinations identified by systematic perturbation analysis are consistent with a network motif based control methodology applied to the dynamic network model

Stable motif analysis is based on the principle that a certain group of nodes forms a network motif with identifiable properties that stabilize into a single cellular state regardless of the rest of the network. A more detailed explanation of criteria for identifying stable motifs was described previously (26), and this methodology was used to identify feedback loops responsible for the mesenchymal steady states of the EMT network (18). As we are currently interested in the network structure that keeps cells epithelial, we applied network motif analysis to the epithelial steady state. Whereas multiple smaller network motifs were previously found for the mesenchymal steady states (18), we identify a single larger stable motif (66% of the network) that stabilizes the epithelial steady state (Figure 4-1D). More recently, network motif analysis has been extended to identify control sets, the sets of nodes and their respective states (i.e. ON or OFF in the Boolean context) that unambiguously lead to a specific steady state regardless of network state (Zañudio & Albert, 2015). Control sets can be used to force networks into a specific steady state, regardless of the current network state. In the context of EMT, the set of nodes and their states that force cells to become epithelial represent the epithelial control set (Figure 4-1C and Table 4-2). The epithelial control sets are summarized as the yellow colored background.

<table>
<thead>
<tr>
<th>No. of nodes in each knockout</th>
<th>No. of nodes (not including TGFβ or TGFβR nodes)</th>
<th>Total possible combinations</th>
<th>No. of combinations that completely block EMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>66</td>
<td>68</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>59</td>
<td>1711</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>53</td>
<td>23,426</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>53</td>
<td>292,825</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4-1. Results of the EMT network model perturbation screen
(Figure 4-1D). Of special interest, are the nodes of the epithelial control set that are able to specifically inhibit the TGFβ-driven EMT from the epithelial steady state. It has been found that generally, subsets of control sets are enough to either completely force all network states to go to the target state or, in many cases, subsets of the complete control set greatly increase the number of initial conditions that go to the target state (Zañudo & Albert, 2015). This is consistent for the EMT model, as the nodes that inhibit TGFβ-driven EMT (Figure 4-1B) are a subset of the epithelial control set (yellow background; Figure 4-1D) or nodes directly upstream of the epithelial control set (nodes colored in blue; Figure 4-1D).

**Table 4-2. Components of the epithelial control set.** A complete control set requires one component from each of the 5 columns. A “~” signifies that a node be in the OFF state. Otherwise, a node is in the ON state.

<table>
<thead>
<tr>
<th>Component A</th>
<th>Component B</th>
<th>Component C</th>
<th>Component D</th>
<th>Component E</th>
</tr>
</thead>
<tbody>
<tr>
<td>(E-cadherin, β-catenin_memb)</td>
<td>~RAS</td>
<td>(~SHH, ~FUS, ~SMO, SUFU, PTCH, ~GLI)</td>
<td>~TGFβR</td>
<td>(~MEK, ~ERK)</td>
</tr>
<tr>
<td>(E-cadherin, β-catenin_memb)</td>
<td>~RAS</td>
<td>(~SHH, ~FUS, ~SMO, SUFU, PTCH, ~GLI)</td>
<td>~SMAD</td>
<td>~SNAI1</td>
</tr>
<tr>
<td>(-Ecadherin, β-catenin_memb)</td>
<td>GSK3β</td>
<td>(~SHH, ~FUS, ~SMO, SUFU, PTCH, ~GLI)</td>
<td>~TGFβR</td>
<td>(~MEK, ~ERK)</td>
</tr>
<tr>
<td>(E-cadherin, β-catenin_memb)</td>
<td>GSK3β</td>
<td>(~SHH, ~FUS, ~SMO, SUFU, PTCH)</td>
<td>(~TGFβR, ~TGFβ)</td>
<td>~SNAI1</td>
</tr>
<tr>
<td>(E-cadherin, β-catenin_memb)</td>
<td>GSK3β</td>
<td>(~SHH, ~FUS, ~SMO, SUFU, PTCH)</td>
<td>(~TGFβR, ~TGFβ)</td>
<td>RKIP</td>
</tr>
<tr>
<td>β-catenin_memb</td>
<td>GSK3β</td>
<td>~GLI</td>
<td>(~TGFβR, ~TGFβ)</td>
<td>RKIP</td>
</tr>
<tr>
<td>β-catenin_memb</td>
<td>GSK3β</td>
<td>~GLI</td>
<td>(~TGFβR, ~TGFβ)</td>
<td>~SNAI1</td>
</tr>
</tbody>
</table>
4.4.3 Mesenchymal phenotype cells are enriched for nodes predicted by the EMT network model to inhibit TGFβ-driven EMT

We next sought to characterize predicted inhibitory combinations produced by network model analysis in experimental HCC models. We focused on the set of double combinations with the ultimate goal of knocking down these nodes to determine whether, as predicted by the EMT network model, they inhibit the TGFβ-driven EMT. Prior to this analysis, we sought to determine whether these nodes were in fact expressed at appreciable levels in liver cancer cell lines Huh7, HepG2, PLC/PRF/5 and HLF. Previous analysis by immunoblot and qRT-PCR analysis demonstrated that the human HCC cell lines Huh7 and PLC/PRF/5 were more epithelial-like and the HLF cell line had a mesenchymal phenotype (18). We confirmed by immunoblot and immunofluorescent staining that Huh7 cells have high E-cadherin and low vimentin expression, while HLF cells have low E-cadherin and high vimentin expression (Figure 4-2A and Figure 4-2B). We additionally characterized two other cell lines, PLC/PRF/5 and HepG2, which have intermediate levels of E-cadherin and vimentin (Figure 4-2B). We tested expression of nodes in these cell lines that were part of double combinations predicted by the EMT network model to inhibit TGFβ-driven EMT. By qRT-PCR and immunoblot a majority of the nodes were expressed. Additionally, by qRT-PCR and immunoblot, nodes of interest to suppress EMT were enriched in HLF cell lines compared to Huh7, HepG2, and PLC/PRF/5 cell lines (Figure 4-2C and Figure 4-2D). Furthermore treating epithelial-like Huh7 cells with TGFβ for 48 hours at a low (1 ng/ml) and high (5 ng/ml) dose of TGFβ led to induction of node of interest expression (Figure 4-2E). Interestingly, the delta-like proteins (DLL1 and DLL3; Delta node in the EMT model) were not detectable by qRT-PCR in the Huh7 cell line, and DLL4 was only detectable at very high amplification cycles after TGFβ treatment by qRT-PCR (Figure 4-2E), suggesting that the delta ligands are not appreciably expressed in the Huh7 cell line. Overall, these results suggest that a majority of the nodes of interest in the EMT network model are expressed in experimental models and also enriched in mesenchymal relative to epithelial phenotype cell lines.

4.4.4 A multi-faceted siRNA screen to test predicted node knockout combinations in vitro

Systematic analysis of knockout perturbations in the in silico EMT network suggests that targeting a specific subset of elements individually or in combinations may suppress TGFβ-
driven EMT (Figure 4-1B). The *in silico* perturbation analysis was done by simulating the effect of TGFβ on the epithelial initial condition in the presence of combination knockouts (e.g. TGFβ=ON while SMAD=OFF and CSL=OFF). In order to test these combinations experimentally, a siRNA screen was employed using an experimental approach that mimicked *in silico* analysis. E-cadherin expression and cell migration were used as experimental readouts for EMT. The experimental screen was performed in the epithelial-like Huh7 cell line and 2 nM siRNA was used for each target (4 nM siRNA total for a combination of 2 nodes). Two negative controls were used to represent ineffective inhibition of EMT (low E-cadherin expression and higher migratory capacity). The first was a scrambled siRNA transfection to control for effects of siRNA transfection. In the second, only SMAD4 siRNA was used, as network analysis suggested that SMAD4 inhibition alone minimally inhibits the TGFβ-driven EMT. Two positive controls were used to represent effective inhibition of EMT (high E-cadherin expression and lower migratory capacity). The first was siRNA targeting SNAI1 plus SMAD4 because literature evidence and network analysis suggest that SNAI1 inhibition robustly suppresses TGFβ-driven EMT. The second was siRNA targeting TGFβR1 plus SMAD4, as this would directly disrupt TGFβ signaling at its source.

In order to test the knockout combinations, siRNA to each target element was transfected into Huh7 cells in the combinations predicted by the network model or with a scrambled siRNA control. At 48 hours post transfection, cells were treated with TGFβ or with a vehicle control. At 96 hours post transfection, cells were harvested to analyze the effect of siRNA node knockdowns on the epithelial phenotype marker E-cadherin (Figure 4-4 and Figure 4-5B). Combinations of SMAD4 with K-RAS, H-RAS, N-RAS, NOTCH1, NOTCH3, and GRB2 were most similar to the positive controls representing high expression of E-cadherin. Combinations of SMAD4 with NOTCH2 and CSL were most similar to the negative controls representing low expression of E-cadherin.

At 48 hours post transfection, cells were treated with TGFβ or with a vehicle control. At 96 hours post transfection, cells were harvested to analyze the effect of siRNA node knockdowns on the epithelial phenotype marker E-cadherin. Simultaneously, at 48, 72, and 96 hours post transfection, automated image acquisition and analysis was used to quantify cell migration (Figure 4-5A). The effect of siRNA combinations on TGFβ treated Huh7 cells relative to vehicle treated control cells normalized to a scrambled siRNA control was identified experimentally via
quantitative immunoblotting (Figure 4-4 and Figure 4-5B) and analysis of migration (Figure 4-5C). Furthermore, the percent change in TGFβ driven migration for siRNA combinations compared to scrambled control was calculated (Figure 4-5C). These results suggest that a number of the predicted siRNA combinations suppress the TGFβ-driven EMT.

To assay for cell migration, automated image acquisition and analysis was used to quantify cell migration at 48, 72, and 96 hours post transfection (Figure 4-5A). The effect of siRNA combinations on TGFβ treated Huh7 cells relative to vehicle treated control cells normalized to a scrambled siRNA control was identified experimentally via quantitative immunoblotting and analysis of migration (Figure 4-5B and Figure 4-5C, see Methods). Furthermore, the percent change in TGFβ driven migration for siRNA combinations compared to scrambled control was calculated (Figure 4-5 C). Combinations of SMAD4 with K-RAS, NOTCH2, NOTCH3, and NOTCH4 were most similar to the positive controls representing lower migratory capacity. Combinations of SMAD4 with N-RAS, SOS1, and GRB2 appeared less effective at inhibiting EMT. These results indicate that the majority of the predicted siRNA combinations suppress the TGFβ-driven EMT.
Figure 4-2. Mesenchymal phenotype cells are enriched for nodes predicted by the EMT network model to inhibit TGFβ-driven EMT. A) Expression of epithelial (E-cadherin) and mesenchymal (vimentin) markers in HCC cell lines. B) Immunofluorescent staining of EMT markers E-cadherin (green), vimentin (red), and nuclear stain DAPI (blue) in HCC cell lines Huh7, PLC/PRF/5 (Alexander), HepG2, and HLF. mRNA expression (C) by qRT-PCR and protein expression by immunoblot (D) of nodes predicted by the EMT network model to inhibit TGFβ-driven EMT. E) Epithelial-like HCC cells Huh7 were treated with TGFβ in serum free media for 48 hours (1 ng/ml and 5 ng/ml), then mRNA expression of nodes predicted by the EMT network model to inhibit TGFβ-driven EMT were measured by qRT-PCR.
Figure 4-3. Node knockdown efficiency for siRNAs used. Knockdown efficiency was assessed relative to an siRNA control for each siRNA used in the siRNA screen. The percent knockdown of node mRNA expression by qRT-PCR is shown. The average knockdown efficiency across all knockdowns is 70%.

Figure 4-4. Western blots used to quantitate the effect of node knockdown combinations on E-cadherin expression. A) The effect of node combinations on vehicle treated and B) TGFβ treated Huh7 cells. 2 nM of each siRNA was used per transfection. 48 hours post transfection, cells were treated with either TGFβ (5ng/ml) or a vehicle control for 48 hours before harvesting.
Figure 4-5. A multi-faceted siRNA screen to test predicted node knockdown combinations *in vitro*. A) Schematic of the experimental design of the screen to test node knockdown combinations that are predicted to inhibit TGFβ-driven EMT. At 0 hours, epithelial-like Huh7 cells were transfected with siRNA combinations or a scrambled siRNA control (2 nM per siRNA), and then plated for harvesting of protein, mRNA, and cell migration. At 45 hours post transfection, cells were serum starved for 3 hours. At 48 hours, pre-TGFβ treatment cells were harvested, then at 96 hours, cells were treated with TGFβ (5 ng/ml) for an additional 48 hours. At 96 hours, cells treated with TGFβ were harvested for mRNA and protein expression, and analysis of cell migration was performed. B) The effect of model-predicted node combinations on E-cadherin expression in TGFβ-treated cells relative to vehicle control treated cells, as measured by quantitative immunoblotting. C) The effect of model-predicted node combinations on *in vitro* cell migration in TGFβ-treated cells relative to vehicle control treated cells. The percent reduction in TGFβ-driven migration (top) and the effect of TGFβ-driven EMT relative to a scrambled siRNA control 24 and 48 hours after TGFβ treatment.
4.4.5 Analysis of motifs that arise after SMAD inhibition in the EMT network reveal an attractor landscape that is distinct from the SMAD unperturbed network

Results from the systematic network perturbation analysis and subsequent experimental work suggest a pivotal role for the SMAD complex in TGFβ-driven EMT. Exploration of the network attractors achievable in the presence of TGFβ and in the absence of the SMAD complex reveals the same mesenchymal attractor as in the unperturbed (TGFβ=ON) network, and an epithelial attractor which is relatively different from the epithelial state of the unperturbed TGFβ=ON model (Figure 4-6A). These new epithelial steady states contain some features of the epithelial steady state and some features of the mesenchymal steady state described in the unperturbed model. The epithelial steady state in the TGFβ=ON, SMAD=OFF network contains a subset of the unperturbed epithelial motif that keeps E-cadherin=ON & Beat_Memb=ON while keeping the SHH, MAPK, and Wnt motifs maintained OFF (Figure 4-6A and Figure 4-6B). After realizing that there were steady states in the SMAD=OFF perturbed network that were quite different from the unperturbed network, we sought to explore if attractors that differed from the unperturbed EMT model existed in other perturbed networks and whether there was a quantitative way to classify these attractors.

4.4.6 Quantitative analysis of steady states in the EMT network reveals a putative spectrum of EMT phenotypes

Some of the attractors of the TGFβ=ON, SMAD=OFF network, appeared different from epithelial and mesenchymal attractors from the unperturbed EMT network. Additionally, thus far we described attractors as being “epithelial” or “mesenchymal” based on if their E-cadherin expression. We sought now to find a quantitative way to classify the attractors relative to the epithelial and mesenchymal attractors of the unperturbed EMT network. To do this, all of the attractors from applying single perturbations to the EMT network were collected. All of these attractors were projected onto the unperturbed epithelial and mesenchymal states to look at how close these attractors were to the two unperturbed attractors. This projection is represented in a mesenchymal/epithelial plane, where the normal mesenchymal state has coordinates (mesenchymal, epithelial) = (1,0), and the epithelial state would have coordinates (mesenchymal, epithelial) = (0,1). It appears that some attractors exist that are closer to the epithelial plane, some closer to the mesenchymal plate, and some exist on the diagonal formed between both
planes. This strongly supports that the attractors we found are truly in between a mesenchymal and epithelial state.

We were also curious whether epithelial and mesenchymal states could be classified in an unbiased manner (i.e. without having to specify the mesenchymal/epithelial classification \textit{a priori}). To do this, we performed principal components analysis using all the attractors obtained using single perturbations attractors and the unperturbed EMT network attractors. As with the epithelial mesenchymal projection (Figure 4-6C), a number of the network steady states cluster distinctly from the epithelial and mesenchymal clusters (Figure 4-7).

\section*{4.5 Discussion}

EMT plays a crucial role in cancer invasion and metastasis. TGFβ is a signal for induction of EMT and is a frequent dysregulation in HCC (13, 15). Here we set out to explore network perturbations that increase the efficacy of SMAD complex inhibition (i.e. canonical TGFβ signaling) in order to inhibit the induction of EMT. To identify ways in which SMAD might inhibit EMT and the broader question of how inhibit EMT, we systematically screen individual and combination node perturbations in the EMT network to identify nodes and/or combinations of nodes, that when knocked out, suppress the TGFβ-driven EMT (Table 1 and Figure 4-1B). To get a better understanding of the way in which these combinatorial interventions suppress EMT, we compare them with the intervention targets obtained from a recently developed network control approach (Zanudo and Albert, 2015). This network control approach identifies the feedback regulatory motifs through which the epithelial phenotype is maintained, which allows us to determine the mechanism through which the combinatorial interventions suppress EMT. We test these combinations experimentally and the results suggest that many of nodes that when knocked out are predicted to inhibit EMT, are enriched in mesenchymal phenotype HCC cell lines (Figure 4-2). Screening the combinations of two nodes reveal that many combinations suppress the TGFβ-driven EMT \textit{in vitro} (Figure 4-5). Finally, we analyze the SMAD-perturbed EMT network and find that some steady states are significantly different from both the epithelial and mesenchymal steady states of the unperturbed EMT network (Figure 4-6 and Figure 4-7). This finding led us to quantitatively explore the steady states of other perturbed networks. Our results reveal numerous steady states that are intermediate to epithelial and mesenchymal steady states, suggesting the existence of hybrid
(intermediate) states with the SMAD-perturbed steady states located in the middle of the spectrum. Further analysis revealed the regulatory motifs responsible for this hybrid epithelial-mesenchymal state and the existence of an epithelial-mesenchymal spectrum (Figure 4-6 and Figure 4-7).

Figure 4-6. Analysis of motifs that arise after SMAD inhibition in the EMT network reveal an attractor landscape that is distinct from the SMAD unperturbed network. A) In the TGFβ=ON; SMAD=OFF model, the same mesenchymal attractor as in the unperturbed (TGFβ=ON) network appears, as well as epithelial states which are relatively different from the epithelial state of the unperturbed TGFβ=ON model. B) These new epithelial steady states contain a piece of a motif (a subset of the motif) that keeps E-cadherin=ON & Bcat_Memb=ON, a Wnt feedback loop maintained OFF (blue background), the SHH motif (yellow background), and the MAPK motif. C) Single node knockouts were performed in the TGFβ=ON EMT network model. The steady states from each of these knockouts were projected onto the epithelial (1,0 in the X-Y plane) and mesenchymal (0,1 in the X-Y plane) steady states from the TGFβ=ON model without any other perturbation. Quantitative analysis of the spectrum of steady states in single node perturbation models revealed a spectrum of steady states, many of which appeared intermediate to the TGFβ=ON unperturbed epithelial and mesenchymal states.
One advantage of using a computational model is that we were able to screen hundreds of thousands of perturbations rapidly and with essentially no cost. The same screen experimentally would require extensive time, resources, and cost. This screen provides a much smaller subset of targets that can then be taken through experimental analysis. Seven nodes individually and six nodes in combinations of two suppressed EMT in our network model. The seven individual nodes (SNAI1, SNAI2, ZEB1, ZEB2, TWIST, FOXC2, and HEY1) represent the network of transcription factors that directly repress E-cadherin transcription. This is consistent with literature evidence for the role of these transcription factors in the induction of EMT (16, 27-30). Interestingly, SMAD complex inhibition in combination with inhibition of RAS, DELTA, NOTCH, or SOS1/GRB2 was able to suppress the TGFβ-driven EMT. This suggests that

Figure 4-7. Principal components analysis of single perturbations in the EMT network model supports the existence of an EMT spectrum. The epithelial steady states from the unperturbed EMT network model cluster in the bottom right corner of the plot and the mesenchymal steady states along with numerous other perturbed states cluster in the bottom left corner of this plot. Numerous other steady states exist as distinct and intermediate clusters between the epithelial and mesenchymal clusters.
suppression of non-canonical TGFβ signaling in addition to canonical (SMAD) signaling is required for robust inhibition of EMT. TGFβ can signal non-canonically through RAS to activate DELTA and NOTCH, and this network path appears to be critical to EMT signal transmission. Experimental results support that many of these combinations are superior to SMAD inhibition alone as measured by E-cadherin expression and also suppression of migration, a functional readout of EMT. Interestingly, these results suggest that the cross-talk between the TGFβ and NOTCH signaling pathways is particularly important for induction of EMT by TGFβ.

There exists previous evidence of TGFβ and NOTCH cross-talk during the induction of EMT. Initial studies demonstrated that TGFβ signaling can increase expression of the NOTCH signaling pathway target gene HEY1/HES, and that TGFβ induced EMT was blocked by suppressing HEY1 (29), which is consistent with our model (Figure 4-1B). Additional evidence suggests Notch signaling is necessary for epithelial growth arrest by TGFβ in multiple epithelial cell types (31). Other work in human kidney epithelial cells demonstrated that Jagged1 and HEY1 were induced by TGFβ in kidney epithelial cells (32). Our experimental and computational results suggest SMAD4 has some effect alone but that combined NOTCH and SMAD suppression has an increased effect on EMT inhibition, which has not previously been described. DELTA ligands (DLL1, DLL3, and DLL4) were undetected in our cells, yet Notch signaling is activated and has a functional role in the TGFβ-driven EMT. Interestingly, Jagged is also included in our model but was not predicted to inhibit the TGFβ-driven EMT, although it is activated during the induction of EMT. Studies in murine lung adenocarcinoma cells support that NOTCH ligand Jagged2 is important for the TGFβ-driven EMT and metastasis (33). This suggests that the Jagged ligand could play a role in the TGFβ-driven EMT in liver cancer cells and is a focus of future work. If Jagged turns out to have a role in TGFβ-driven EMT in liver cancer cells, then the model could be adjusted to account for this. This could be completed by expanding the model if there exists evidence that signal transduction mediators not currently in the EMT network model could potentially explain this phenomenon. One potential mechanism that could explain a critical role for Jagged in the TGFβ-driven EMT is a recent paper showing that NOTCH can activate HIF1a independent of hypoxia in multiple cell lines, which is not currently represented in the EMT network model (34).
Interestingly, our initial analysis of all possible cellular states when SMAD is knocked out and the TGFβ signal is turn ON, led us to realize that many of these states had features of both epithelial and mesenchymal states as identified in the unperturbed EMT model. We explored all the steady states for single node perturbations in the EMT network and quantitative comparison of these states suggests that many single knockouts in the network lead to steady states that are intermediate to epithelial and mesenchymal states. There exists evidence of hybrid EMT states in experimental and computational models (7-9, 16, 17); however, a quantitative characterization of an EMT spectrum has not previously been completed. For the hybrid EMT states described computationally, only a single hybrid state is described, likely due to the fact that only a small number of EMT regulators were included in these models (16, 17). Our model suggests that looking at a larger number of regulators could be useful in a quantitative description of the EMT spectrum. Interestingly, the ZEB1 knockout (i.e. TGFβ=ON, ZEB1=OFF) steady states appear to induce hybrid steady states in the EMT network model (Figure 4-6 and Figure 4-7), and this is consistent with previous findings (16, 27). The TGFβ=ON, miR-200=ON steady state represents another interesting hybrid EMT state. Interestingly, previous work by our group suggests that miR-200 overexpression had some effect on E-cadherin expression in a HCC cell line; however, combination with a DNA methyltransferase (DNMT) inhibitor led to further increase in E-cadherin expression (35). This suggests the DNMT inhibitor is targeting other nodes in the EMT network, leading to a more robust epithelial state.

In summary, we used network analysis and Boolean dynamic modeling to investigate the role of SMAD dynamics in the TGFβ-driven EMT, with the goal of identifying ways to inhibit EMT in HCC. Our systems biology approach was able to use available pathway information and to identify putative targets that suppress EMT and liver cancer invasion, and in vitro analysis suggests that many of these predicted targets suppress the TGFβ-driven EMT. Future work will focus on more in depth analysis experimental analysis of these targets including exploring their effect on invasion in vitro and in vivo.

4.6 Acknowledgments

I would like to thank Paul Michel for assistance with experimental work completed here and Jorge Gomez Tejeda Zañudo for contributions to computational analysis. Research reported in this publication was supported by the National Institute of Diabetes and Digestive and Kidney
4.7 References

CHAPTER 5.
Inference of Network Dynamics and Metabolic Interactions in the Gut Microbiome

5.1 Abstract
The bacteria that live in our gut (collectively the gut microbiome) are an enormous ecological community consisting of hundreds of species and trillions of interacting bacteria. The microbiome composition persists for years without significant change; however, when change comes it can have devastating consequences. For example, infection by opportunistic pathogens. How exactly the intestinal microbiota interact remains poorly understood, but an understanding of microbial interactions could lend insight into key organisms leading to stability and collapse of this ecological community. We have developed a novel methodology to construct a Boolean dynamic model from time series metagenomic genus abundance information, and integrate this modeling with genome scale metabolic network reconstructions to identify metabolic underpinnings for microbial interactions. We apply this in the context of a critical health issue: clindamycin antibiotic treatment and opportunistic Clostridium difficile (C. difficile) infection. Our model recapitulates known dynamics of clindamycin antibiotic treatment and C. difficile infection and predicts therapeutic probiotic interventions to suppress C. difficile infection. Genome scale metabolic network reconstructions reveal metabolic differences between community members, and highlight the role of metabolism in the observed microbial interactions.

5.2 Introduction
Human health is inseparably connected to the billions of microbes that live in and on us. Current research shows that our associations with microbes are, more often than not, essential for our health (1, 2). The microbes that live in and on us (collectively our “microbiome”) help us to digest our food, train our immune systems, and protect us from pathogens (3, 4). The gut microbiome is an enormous community, consisting of hundreds of species and trillions of individual interacting bacteria (5). Microbial community composition often persists for years without significant change (6).

When change comes, however, it can have unpredictable and sometimes fatal consequences. Acute and recurring infections by Clostridium difficile have been strongly linked to changes in
gut microbiota (7). The generally accepted paradigm is that antibiotic treatment (or some other perturbation) significantly disrupts the microbial community structure in the gut, which creates a void that C. difficile will subsequently fill (8-11). Such infections occur in roughly 600,000 people in the United States each year (this number is on the rise), with an associated mortality rate of 2.3% (12). Each year, healthcare costs associated with C. difficile infection are in excess of $3.2 billion (12). An altered gut flora has further been identified as a causal factor in obesity, diabetes, some cancers and behavioral disorders (13-18).

What promotes stability of a microbial community, or causes its collapse, is poorly understood. Until we know what promotes stability, we cannot design targeted treatments that prevent microbiome disruption, nor can we rebuild a disrupted microbiome. Studying the system level properties and dynamics of a large community is impossible using traditional microbiology approaches. However, network science is an emerging field which provides a powerful framework for the study of complex systems like the gut microbiome (19-22). Previous efforts to capture the essential dynamics of the gut have made heavy use of ordinary differential equation (ODE) models (23, 24). Such models require the estimation of many parameters. With so many degrees of freedom, it is possible to overfit the underlying data, and it is difficult to scale up to larger communities (25, 26). Boolean dynamic models, conversely, require far less parameterization. Such models capture the essential dynamics of a system, and scale to larger systems. Boolean models have been successfully applied at the molecular (27, 28), cellular (21), and community levels (29). Here we present the first Boolean dynamic model constructed from metagenome sequence information and the first application of Boolean modeling to microbial community analysis.

We analyze the dynamic nature of the gut microbiome, focusing on the effect of clindamycin antibiotic and C. difficile infection on gut microbial community structure. We generate a microbial interaction network and dynamical model based on time-series data from a population of mice. We present the results of a dynamic network analysis, including steady-state conditions, how those steady states are reached and maintained, how they relate to the health or disease status of the mice, and how targeted changes in the network can transition the community from a disease state to a healthy state. Furthermore, knowing how microbes positively or negatively impact each other—particularly for key microbes in the community—increases the therapeutic utility of the inferred interaction network. We produced genome-scale metabolic reconstructions
of the taxa represented in this community (30), and probed how metabolism does—and does not—contribute to the mechanistic underpinnings of the observed interactions.

5.3 Methods

5.3.1 Data Sources

Buffie et al. reported treating mice with clindamycin and tracking microbial abundance by 16S sequencing (31). Mice treated with clindamycin were more susceptible to C. difficile infection than controls. The collection of 16S sequences corresponding to these experiments was analyzed by Stein et al. (23). First, Stein et al. aggregated the data by quantifying microbial abundance at the genus level. Abundances of the ten most abundant genera and an “other” group were presented as operational taxonomic unit (OTU) counts per sample. We use the aggregated abundances from Stein et al. as the starting point for our modeling pipeline (Figure 5-1).

5.3.2 Interpolation of Missing Genus Abundance information

Natural cubic spline interpolation was used to estimate genus abundances at missing time points in some samples. A spline is an approximation of a function that is continuous through a given set of points. A cubic spline is constructed of piecewise third order (cubic) polynomials which pass through the known data points and has continuous first and second derivatives across all points in the dataset. The polynomial coefficients are chosen such that the function is continuous and passes through all given points. Natural cubic spline is a cubic spline that has a second derivative equal to zero at the end points of the dataset (32). Natural splines were interpolated such that all datasets had time points at single day intervals through the 23 day timepoint (Figure 5-2B).
Figure 5-1. Dynamic analysis workflow. Time course genus abundance information was acquired from metagenomic sequencing of mouse gastrointestinal tracts under varying experimental conditions. Missing time points from experimental data were estimated such that genus abundances existed at the same time points across all treatment groups. Next, genus abundances were binarized such that Boolean regulatory relationships could be inferred. A dynamic Boolean model was constructed to explore systems gut microbial ecology dynamics, therapeutic interventions, and chemical mediating bacterial regulatory relationships.

5.3.3 Network Modeling Framework

We use a Boolean framework in which each network node is described by one of two qualitative states: ON or OFF. The ON (logical 1) state means an above threshold abundance of a bacterial genus whereas the OFF (logical 0) state means below-threshold or genus absence. The
putative biological relationships among genera are expressed as mathematical equations using Boolean operators (33). We inferred putative Boolean regulatory functions for each node, which are able to best capture the trends in the bacterial abundances.

Figure 5-2. Bacterial genera abundances over time in response to clindamycin treatment and/or C. difficile inoculation. A) Genera abundance information for the nine samples. Population 1 received spores of C. difficile (at t=0 days), Population 2 received a single dose of clindamycin (at t=-1 days), and Population 3 received a single dose of clindamycin (at t=-1 days) and, on the following day, was inoculated with C. difficile spores (at t= 0 days). Genera abundances were measured at 0, 2, 3, 4, 5, 6, 7, 9, 12, 13, 16, and 23 days; however, not all samples had measurements at all the timepoints. B) Cubic spline interpolation of data points was performed such that all the same time point measurements of bacterial abundance occurred in all samples and that single day intervals were present in all datasets.

5.3.4 Binarization

To build a Boolean model that describes the relationships among gut microbiome genera, we first binarized the bacterial metagenomic abundance data. An approach called iterative k-means
binarization was used to convert continuous abundance data into a binarized format. Briefly, this approach uses k-means clustering with a depth of clustering $d$, and an initial number of clusters $k = 2^d$. In each iteration, data for a specific genus $G$ are clustered into $k$ unique clusters $C_1^G, \ldots, C_k^G$, then for each cluster, $C_n^G$, all the values are replaced by the mean value of $C_n^G$. For the next iteration, the value of $d$ is decreased and clustering is repeated. This methodology is repeated until $d = 1$. This approach, with $d = 3$ (called here as KM3 binarization) has previously been demonstrated as a superior binarization methodology to other binarization approaches for Boolean model construction because it conserves oscillatory behavior (34). These analyses were performed using custom Python code based on a previously written algorithm (34).

5.3.5 Inference of Boolean Rules from Time Series Genus Abundance Information

The Best-fit extension was applied to learn Boolean rules from the binarized time series genus abundance information (35). For each variable (genus) $X_i$ in the binarized time series genus abundance data, best-fit identifies the set of Boolean rules $X' \subseteq \{X_1, \ldots, X_n\}$ with $k \leq n$ variables (regulators) that explains the variable’s time pattern with the least error size. The algorithm uses partially defined Boolean functions pdBf(T,F), where the set of true (T) and false vectors (F) are defined as $T = \{X'(t) \in \{0,1\}^k : X_i(t+1)=1\}$ and $F = \{X'(t) \in \{0,1\}^k : X_i(t+1)=0\}$. Intuitively, the partial Boolean function summarizes the states of the putative regulators that correspond to a turning ON (T) or turning OFF (F) of the target variable. The error size $\varepsilon$ of pdBf(T,F) is defined as the minimum number of inconsistencies within $X'$ that best classifies the T and F values of the dataset. The Best-fit extension works by identifying smallest size $X'$ for $X_i$. For more detailed information refer to (35). In line with this, we consider the most parsimonious representation of the rules with the smallest $\varepsilon$. If multiple rules fit these criteria for a given $X_i$, it implies that they can independently represent the inferred regulatory relationships. In cases where the alternatives have the same value of (non-zero) $\varepsilon$, we explored combinations (such as appending them by an OR rule) and used the combination that best described the experimentally observed final (steady state) outcomes. We used the implementation of Best-Fit in the R package BoolNet (36).
5.3.6 Dynamic Analysis

Dynamic analysis is performed by applying the inferred Boolean functions in succession until a steady state is reached. Boolean models and discrete dynamic models in general focus on state transitions instead of following the system in continuous time. Thus, time is an implicit variable in these models. The network transitions from an initial condition (initial state of the bacterial community) until an attractor is reached. An attractor can be a fixed point (steady state) or a set of states that repeat indefinitely (a complex attractor). The basin of attraction refers to the initial conditions that lead the system to a specific attractor. For the network under consideration, the complete state space can be traversed by enumerating every possible combination of node states \(2^{12}\) and applying the inferred Boolean functions (or “update rules”) to determine paths linking those states. The state transition network describes all possible community trajectories from initial conditions to steady states, given the observed interactions between bacteria in the community.

We made use of two update schemes to simulate network dynamics: synchronous (deterministic) and asynchronous (stochastic). Synchronous models are the simplest update method: all nodes are updated at multiples of a common time step based on the previous state of the system. The synchronous model is deterministic in that the sequence of state transitions is definite for identical initial conditions of a model. In asynchronous models, the nodes are updated individually, depending on the timing information, or lack thereof, of individual biological events. In the general asynchronous model used here, a single node is randomly updated at each time step (37). The general asynchronous model is useful when there is heterogeneity in the timing of network events but when the specific timing is unknown. Due to the heterogeneous mechanisms by which bacteria interact, we made the assumption of time heterogeneity without specifically known time relationships. Synchronous and asynchronous Boolean models have the same fixed points, because fixed points are independent of the implementation of time. However, the basins of attraction that lead to the fixed points may differ between synchronous and asynchronous models (Table 5-2). The simulations of the gut microbiome model were performed using custom Python code built on top of the BooleanNet Python library, which facilitates Boolean simulations (38).
5.3.7 Perturbation Analysis

To capture the effect of removing (knockout; targeted removal) or adding (probiotic; over abundances) genera, modification of the states/rules to describe knockout or overexpressed states were performed. These modifications were implemented in BooleanNet by setting the corresponding nodes to either OFF (knockout) or ON (overabundance) and then removing the corresponding updating rules for these nodes for the simulations. By examining many such forced perturbations, we can identify potential therapeutic strategies, many of which may not be obvious or intuitive, particularly as network complexity increases.

5.3.8 Generating Genus-Level Genome-Scale Metabolic Reconstructions

To generate draft metabolic network reconstructions for each of the ten genera in the paper, we first collected genome sequences for representative species by searching the “Genomes” database of the National Center for Biotechnology Information (NCBI). Complete genomes for the first ten (or if less than ten, all) species within the appropriate genus were downloaded. During the process of reconstructing genus-level metabolic reconstructions, some genera were underrepresented (fewer than 10 species genomes) in the NCBI Genome database, including Akkermansia, Barnesiella and Coprobacillus (Table 5-3). The search result order is based on record update time, and so is quasi-random. Genomes were uploaded to the rapid annotations using subsystems technology (RAST) server for annotation (39). Draft metabolic network reconstructions were generated by providing the RAST annotations to the Model SEED service (40). Metabolic network reconstructions were downloaded in “.xls” format. Genus-level metabolic reconstructions were produced by taking the union of all species-level reconstructions corresponding to each genus, as has been done previously (41). The one exception was C. difficile, which was produced by taking the union of three strain-level reconstructions.

5.3.9 Subsystem Enrichment Analysis

Subsystems were defined as the Kyoto Encyclopedia of Genes and Genomes (KEGG) map with which each reaction was associated (42, 43). These associations were determined based on annotations in the Model SEED database (40). To quantify enrichment, the complete set of unique reactions from all genus-level reconstructions was pooled, and the subsystem annotations corresponding to those reactions were counted. To determine enrichment for a given subset of
the community (either a single genus-level reconstruction, or a set of reconstructions corresponding to a subnetwork), the subsystem occurrences were counted within the subset. The probability of a reconstruction containing \( N \) total subsystem annotations, with \( M \) or more occurrences of subsystem \( I \), was determined by taking the sum of a hypergeometric probability distribution function (PDF) from \( M \) to the total occurrences of subsystem \( I \) in the overall population. Enrichment analysis was performed in Matlab (44).

5.3.10 Identifying Seed Sets and Defining Metabolic Competition Score

To quantify metabolic interactions, we started by utilizing the seed set detection algorithm developed by Borenstein et al. (45, 46). The algorithm follows three steps:

1) The genome-scale metabolic network reconstruction is reduced into simple one-to-one edges, such that for each reaction, each substrate and product pair forms an edge (e.g. \( \text{A} + \text{B} \rightarrow \text{C} \) would become \( \text{A} \rightarrow \text{C} \) and \( \text{B} \rightarrow \text{C} \)).

2) The network is divided into strongly connected components, those groups of nodes for which two paths of opposite directions (e.g. \( \text{A} \rightarrow \text{B} \) and \( \text{B} \rightarrow \text{A} \)) exist between any two nodes in the group.

3) Nodes (and strongly connected components with five or fewer nodes) for which there are exclusively outgoing edges are defined as “inputs” to the model, or seed metabolites.

The rationale is that metabolites that feed into the network, but cannot be produced by any reactions within the network, must be obtained from the environment.

Competition metrics were generated following the process of Levy and Borenstein (45). For a given pair of genera, the competition score is defined as:

\[
\text{comp\_score\_i\_j} = \frac{|\text{seed\_set\_i} \cap \text{seed\_set\_j}|}{|\text{seed\_set\_i}|}
\]

Here \( \text{seed\_set\_i} \) is the set of obligatory input metabolites to the metabolic network reconstruction for genus \( i \), and \( |\text{seed\_set\_i}| \) is the number of metabolites contained in \( \text{seed\_set\_i} \). The competition score indicates the fractional overlap of inputs that genus \( i \) shares with genus \( j \), and so ranges between zero and one. The higher the score, the more similar the metabolic inputs to the two networks, making competition more likely.

For a given pair of genera, the mutualism score is defined as:
mut_score\_i\_j = |seed\_set\_i \cap \neg seed\_set\_j|/|seed\_set\_i|

Here \(\neg seed\_set\_j\) is the set of metabolites that can be produced by the metabolic network for species j (i.e. all non-seed metabolites). The mutualism score indicates the fractional overlap of inputs that genus \(i\) consumes which genus \(j\) can potentially provide. The mutualism score ranges between zero and one. The higher the score, the more potential for nutrient sharing between species. While the score does not measure “mutualism” per se (it cannot necessarily distinguish between other interactions such as commensalism or amenalism (47), for simplicity, we will refer to these scores as the competition and mutualism scores. Seed set generation was performed using custom Matlab scripts (44). Statistical tests were performed in R (48).

5.3.11 Co-culture and Spent Media Experiments

*Barnesiella intestinohominis* DSM 21032 and *Clostridium difficile* VPI 10463 were grown anaerobically in PRAS chopped meat medium (CMB) (Anaerobe Systems, Morgan Hill, CA) at 37°C. To prepare *B. intestinohominis* spent medium, *B. intestinohominis* was grown in CMB until stationary phase (44 hours). The saturated culture was centrifuged, and the supernatant was filter sterilized (0.22 µM pore size). Growth curves were obtained by inoculating batch cultures in 96-well plates and gathering optical density measurements (870 nm) using a small plate reader that fits in the anaerobic chamber (49). Single cultures were inoculated from overnight liquid culture to a starting density of 0.01. The co-cultures were started at a 1:1 ratio, for a total starting density of 0.02. Optical density was measured every 2 minutes for 24 hours, and the resulting growth curves were analyzed in Matlab (44). Maximum growth rates were calculated by fitting a smooth line to each growth curve, and finding the maximum growth rate from among the instantaneous growth rates over the whole time course: \([\log(OD_{t+1}) - \log(OD_t)] / [t_{t+1} - t_{t}].\) The achieved bacterial density—area under the growth curve (AUC)—in a culture was calculated by integrating over the growth curve in each experiment using the “trapz()” function in Matlab. It can be thought of as representing the total biomass produced over time. The simply additive null model was calculated by fitting a Lotka-Volterra model (23) to the single cultures for both *B. intestihominis* and *C. difficile*. The null model of co-culture (assuming zero interaction between species) was
simulated by using the parameters from single culture, and summing the predicted OD870 values. All scripts used to analyze the data are available at https://bitbucket.org/gutmicrobiomepaper/microbiomenetworkmodelpaper/wiki/Home.

5.4 Results

5.4.1 Construction of a Microbial Genus Abundance Dataset for Network Inference

We re-analyzed a data set from two previous studies (23, 31). The raw data was generated in the first study, which assessed cecal microbiome changes in response to clindamycin treatment and *C. difficile* inoculation (31). In the second study, the raw sequences were processed into a genus-level abundance time course (23). This processed dataset consisted of nine samples and three treatment groups (n=3 replicates per treatment group). The first treatment group (here called “Healthy”) received spores of *C. difficile* at t=0 days, and was used to determine the susceptibility of the native microbiota to invasion. The second treatment group (here called “clindamycin treated”) received a single dose of clindamycin at t=-1 days to assess the effect of the antibiotic alone, and the third treatment group (here called “clindamycin+ *C. difficile* treated”) received a single dose of clindamycin (at t=-1 days) and, on the following day, was inoculated with *C. difficile* spores (Figure 5-2A). Under the clindamycin+ *C. difficile* treatment group conditions, *C. difficile* could colonize the mice and produce colitis; however this was not possible under the first two treatment group conditions.

The gut bacterial genus abundance dataset included some variation in terms of time points in which genera were sampled. That is, genus abundances were measured between 0 to 23 days; however, not all samples had measurements at all the time points (Figure 5-2A). Particularly, the healthy population only included time points at 0, 2, 6, and 13 days and Sample 1 of clindamycin+ *C. difficile* treated population was missing the 9 day time point. Missing abundance values for these 4 points were estimated using an interpolation approach as described in Methods (32) (Figure 5-2B). For healthy samples, the 16 and 23 day time points could not be interpolated as the last experimentally identified time point for these samples is at 13 days. The assumption of the approximated polynomial for these samples is that extrapolated data points are linear using the slope of the interpolating curve at the nearest data point. Because genera abundances are fairly stable across time in this treatment group (i.e. the slope of most of the
genera abundances is approximately zero), extrapolating two time points was deemed reasonable. A principal component analysis was completed on the interpolated data (Figure 5-5A) and shows that the interpolated time series bacterial genus abundance data clusters by experimental treatment group in the first two principal components. Furthermore, the results of the binarization for the healthy population suggest that interpolation did not have any concerning effects on the 16 and 23 day time points (Figure 5-3).

Figure 5-3. Averaged binarized genera abundances using iterative k-means binarization. Iterative k-means binarization was completed on all the samples 1000 times and average binarization is shown for each genus at each time point in each of the nine samples. If a node (genera) is binarized as 0 (OFF) at a time step, then it is colored blue, and if a node (genera) is binarized as 1 (ON) at a time step, then it is colored yellow. This figure represents the average of 1000 replicates of IKM binarization. Intermediate cell colors represent cases where a genus abundance at a time point was binarized to 1 (ON) in a fraction of the replicates.
Figure 5-4. Averaged binarized genera abundances using iterative k-means binarization were rounded to the most probable binarized state. The most probable binarized state of each genus at each time point. If the average genus abundance binarization (Figure 5-3) was greater than 0.5 (ON in over 500 of 1000 replicates), then that genus abundance was assumed to be 1 (ON) for downstream analysis. If the average genus abundance binarization was less than 0.5 (ON in less than 500 of 1000 replicates) then that genus abundance was assumed to be 0 (OFF) for downstream analysis.
5.4.2 Binarization of Gut Microbiome Genus Abundance Information for Construction of a Dynamic Network Model

A network is a representation of a complex system; the nodes of the network represent the elements (e.g. bacteria) of the system and the edges represent the relationships between nodes. The edges can be assigned a direction, representing information flow, i.e. effect from the source (upstream) node to the target (downstream) node. Furthermore, edges can be characterized as positive (growth promoting) or negative (growth suppressing). An additional layer of network analysis is the dynamic model, which is used to express the behavior of a system over time by characterizing each node by a state variable (e.g., abundance) and a function that describes its regulation. Dynamic models can be categorized as continuous or discrete, according to the type of node state variable used. Continuous models use a set of differential equations; however, the paucity of known kinetic details for inter-genus and/or inter-species interactions makes these models difficult to implement.

Discrete dynamic modeling is used to study biological systems because of its computational feasibility and capacity to be constructed with minimal and qualitative biological data (33). In the simplest discrete dynamic models, called Boolean models, each network node can be described by one of two qualitative states: ON or OFF. The ON state means an above-threshold abundance of a bacterial genus whereas the OFF state indicates the abundance is below threshold or the genus is absent. Regulatory relationships are described using the Boolean operators OR (independent regulation), AND (synergistic regulation), and NOT (inhibitory regulation).

To capture the dynamics of inter-genus interactions in the intestinal tract we employed a novel pipeline (Figure 5-1), which translates metagenomic genus abundance information into a dynamic Boolean model. This approach involves three steps: 1) discretization (binarization) of genus abundances, 2) learning Boolean relationships among genera, and 3) translation of genus associations into a Boolean dynamic model.

Genus abundance data was binarized (converted to a presence-absence dataset) to enable inference of Boolean relationships for modeling applications (see Methods). We adapted a previously developed approach called iterative k-means binarization with a clustering depth of 3 (KM3) for this purpose (34). This approach was employed because binarized data is able to maintain complex oscillatory behavior in Boolean models constructed from this data, whereas other binarization approaches fail to maintain these features(34). Because KM3 binarization has a stochastic component (the initial grouping of binarization clusters), we employed KM3
binarization on the entire bacterial genus abundance time series dataset 1000 times. The average binarization for each sample (Figure 5-3) was used to determine the most probable binarized state of each genus in each sample at each time point (Figure 5-4). A principal component analysis of the most probable binarized genus abundances for each sample demonstrates that as with the continuous time series abundances (Figure 5-5A), binarized bacterial genus abundance data cluster by experimental treatment group (Figure 5-5B). For inference of Boolean rules from the binarized genus abundances (Figure 5-4), the consensus of two of three samples for each treatment population was used as the binarized state of each genus at each time point in each sample (Figure 5-5C).

5.4.3 Construction of a Dynamic Network Model from Binarized Time Series Microbial Genus Abundance Information

Boolean rules (Table 5-1) were inferred from the time series binarized genus abundances using an implementation of the Best-fit extension (35) in the R Boolean network inference package BoolNet (36) (see Methods). A network of 12 nodes and 33 edges was inferred (Figure 5-5D). The inferred interaction network has a clustered structure: the cluster (subnetwork) containing the two Lachnospiraceae nodes and Barnesiella is strongly influenced by clindamycin whereas the other subnetwork is largely independent of the first, except for the single edge between Barnesiella and C. difficile (Figure 5-5D). In fact, Lachnospiraceae nodes, Barnesiella and the group of “Other” genera form a strongly connected component; that is, every node is reachable from every other node. Most nodes of the second subnetwork are positively influenced by C. difficile, with the exception of Coprobacillus, for which no regulation by other nodes was inferred, and Akkermansia, which is inferred to be only regulated by Coprobacillus. These latter two genera are transiently present (around day 5) in the clindamycin treatment group, but they do not appear in the final states of any of the treatment groups (see Figure 5-2).

We applied dynamic analysis (see Methods) to determine the possible time courses of the inferred network when started from the binarized state at the first time point. Exploration of the steady states of this network reveals 23 possible fixed point attractors (Figure 5-6). Three of the identified attractors (Figure 5-7A) are in exact agreement with the experimentally identified terminal time points of binarized genus abundances (Figure 5-5C). These attractors make up a small subset of the entire microbiome network state space (Table 5-2).
Figure 5-5. Construction of a network model of the gut microbiome from time course metagenomic genus abundance information. Principal component analysis was completed to identify characteristics of the interpolated genus abundances and the binarized interpolated genus abundances. Principal component analysis coefficients associated with each sample in the metagenomic genus abundance dataset was completed for A) interpolated genus abundances and B) binarized interpolated genus abundances. '*' = Healthy; '^' = Clindamycin treated; '#' = Clindamycin+ C. difficile treated. C) Consensus binarization of genus abundance information. Each heatmap represents the consensus binarization for each treatment group. The horizontal axis represents the day of the experiment that the sample came from. The vertical axis represents the specific genera being modeled. Each genus was binarized to a 1 (ON; above activity threshold) or 0 (OFF; below activity threshold). D) Interactions rules were inferred from the binarized data. The interaction rules were simplified for visualization (compound rules were broken into simple one-to-one edges).
The attractor landscape can be divided into six groups based on abundance patterns they share (Figure 5-6). Group 1 is made up of a single attractor wherein all genera are absent (OFF). The second group attractor consists of the experimentally defined healthy state (Attractor 2) and genera in the *C. difficile* subnetwork which can be abundant (ON) independent of the clindamycin subnetwork. The third grouping has the clindamycin treated steady state (Attractor 7) and genera in the *C. difficile* subnetwork that can survive in the presence of the clindamycin. Group 4 contains the clindamycin plus *C. difficile* steady state (Attractor 12) and its subsets in which one or both of the source nodes *Mollicutes* and *Enterobacteriaceae* are absent. Group 5 contains attractors in which clindamycin is absent and *C. difficile* is present. Even if clindamycin is absent, our model suggests that *C. difficile* can thrive if *Lachnospiraceae* and *Barnesiella* are absent, i.e. these states represent a clindamycin-independent loss of *Lachnospiraceae* and *Barnesiella*. Lastly, group 6 attractors have both clindamycin and *C. difficile* as OFF. *Blautia* and *Enterococcus* are always abundant in these attractors. Indeed, because of the mutual activation between *Blautia* and *Enterococcus* they always appear together. Attractors in this group may also include the abundance (ON state) of the source nodes *Mollicutes* and *Enterobacteriaceae*.

### Table 5-1. Boolean update rules for the gut microbiome network

<table>
<thead>
<tr>
<th>Node</th>
<th>Boolean Update Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akkermansia</td>
<td>Coprobacillus</td>
</tr>
<tr>
<td>Barnesiella</td>
<td>(Other or Lachnospiraceae_other or Lachnospiraceae) and not Clindamycin</td>
</tr>
<tr>
<td>Blautia</td>
<td>(Enterococcus or Blautia) or [(not Blautia and Coprobacillus) or (Blautia and not Coprobacillus)]</td>
</tr>
<tr>
<td>Clindamycin</td>
<td>Clindamycin</td>
</tr>
<tr>
<td>Clostridium_difficile</td>
<td>Clostridium_difficile and not Barnesiella</td>
</tr>
<tr>
<td>Coprobacillus</td>
<td>OFF</td>
</tr>
<tr>
<td>Enterobacteriaceae</td>
<td>Enterobacteriaceae</td>
</tr>
<tr>
<td>Enterococcus</td>
<td>Mollicutes or Enterobacteriaceae or not Coprobacillus or Clostridium_difficile or Blautia</td>
</tr>
<tr>
<td>Lachnospiraceae</td>
<td>(Other or Lachnospiraceae_other or Lachnospiraceae) and not Clindamycin</td>
</tr>
<tr>
<td>Lachnospiraceae_other</td>
<td>(Other or Lachnospiraceae_other or Lachnospiraceae) and not Clindamycin</td>
</tr>
<tr>
<td>Mollicutes</td>
<td>Mollicutes</td>
</tr>
<tr>
<td>Other</td>
<td>(Other or Lachnospiraceae_other or Lachnospiraceae) and not Clindamycin</td>
</tr>
</tbody>
</table>
5.4.4 Perturbation Analysis

We next explored the perturbation of genera in the gut microbiome network model. We considered the clinically relevant question of which perturbations might alter the microbiome steady states produced by clindamycin or clindamycin+\textit{C. difficile} treatment after clindamycin treatment was removed. Thus, we considered the clindamycin-treated steady state (Attractor 7 in Figure 5-6) and the clindamycin+\textit{C. difficile} treated steady state (Attractor 12) as initial conditions and assumed that clindamycin treatment was stopped. Our model shows that for both initial conditions, only the state of clindamycin changes after the treatment is stopped; these steady states become Attractor 1 and Attractor 19, respectively (Figure 5-6). In other words, the steady states remain identical in the absence of clindamycin. We next explored the effect of addition (overabundance; Figure 5-7B, left column) and removal (knockout; Figure 5-7B, right column) of individual genera, simultaneously with the stopping of clindamycin treatment, on the model predicted steady states. For the perturbation analysis, the model was initialized from the clindamycin treated steady state (Figure 5-7B, top row) or the clindamycin+ \textit{C. difficile} steady state (Figure 5-7B, bottom row). From the clindamycin treated state, inducing overabundance of \textit{Lachnospiraceae} or “Other” nodes restores the healthy steady state; however, no knockouts restore the healthy steady state (Figure 5-7B). From the clindamycin+ \textit{C. difficile} state, overabundance of \textit{Barnesiella}, \textit{Lachnospiraceae}, or “Other” nodes lead to a shift toward the healthy steady state (suppression of \textit{C. difficile}).

Table 5-2. Basin size as % of total state space (unique basin size) for experimentally realized network attractors

<table>
<thead>
<tr>
<th>Attractor</th>
<th>Synchronous</th>
<th>Asynchronous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy</td>
<td>1.563</td>
<td>4.688 (1.563)</td>
</tr>
<tr>
<td>Clindamycin treated</td>
<td>0.147</td>
<td>0.439 (0.098)</td>
</tr>
<tr>
<td>Clindamycin+ \textit{C. Difficile}</td>
<td>1.367</td>
<td>0.293 (0.098)</td>
</tr>
</tbody>
</table>

5.4.5 Generating Genus-Level Metabolic Reconstructions

Species-level reconstructions from the genus \textit{Enterobacteriaceae} contained the most reactions on average (1335), while those from \textit{Mollicutes} contained the least (485) (Table 5-3). The \textit{Barnesiella} and \textit{Enterococcus} reconstructions contained the most unique reactions (Table 5-4) and, interestingly, also displayed more overlap in reaction content between each other (503 reactions) than was observed between any other pair of reconstructions (Table 5-5).
*Lachnospiraceae* and *Barnesiella* had the next highest degree of overlap (424 reactions). *Mollicutes* and *Coprobacillus* had the least degree of overlap (363 reactions) (Table 5-5).

### 5.4.6 Subsystem Enrichment Analysis

Enrichment analysis was performed for the 99 unique subsystem annotations that were observed in the community (Figure 5-8). 22 subsystems displayed interesting enrichment patterns with respect to the structure of the interaction network (Figure 5-5). The subsystems for glycolysis/gluconeogenesis and nucleotide sugars metabolism are enriched in all taxa, highlighting the fact that all taxa contain relatively full complements of reactions within those subsystems. Interestingly, *C. difficile* is highly enriched for reactions in cyanamino acid metabolism compared to all other genera. Lipopolysaccharide (LPS) biosynthesis and cyanoamino acid metabolism subsystems are differentially enriched between *Barnesiella* and *Lachnospiraceae*, and *C. difficile*. Between *Barnesiella* and *Enterococcus*, *Barnesiella* is more highly enriched for d-glutamine and d-glutamate metabolism, pantothenate and CoA biosynthesis, LPS biosynthesis. With respect to *Enterococcus*, *Barnesiella* is less highly enriched in pyrimidine metabolism, and phenylalanine, tyrosine, and tryptophan biosynthesis.
<table>
<thead>
<tr>
<th>Genus</th>
<th>Average Model Size (average of number of reactions in member-species reconstructions)</th>
<th>Number of species-level models included in each genus</th>
<th>Average Network Overlap Within genus (average number of shared reactions, normalized by average network size)</th>
<th>Average Fraction of Unique Reactions (what fraction of a given species’ network is unique within its genus, normalized by average network size)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akkermansia</td>
<td>818</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Barnesiella</td>
<td>898</td>
<td>2</td>
<td>0.9621</td>
<td>0.0379</td>
</tr>
<tr>
<td>Blautia</td>
<td>980.9</td>
<td>10</td>
<td>0.8751</td>
<td>0.0093</td>
</tr>
<tr>
<td>Clostridium_difficile</td>
<td>999.5</td>
<td>2</td>
<td>0.9735</td>
<td>0.0265</td>
</tr>
<tr>
<td>Coprobacillus</td>
<td>910.3</td>
<td>3</td>
<td>0.9821</td>
<td>0.0092</td>
</tr>
<tr>
<td>Enterobacteriaceae</td>
<td>1335.5</td>
<td>10</td>
<td>0.8738</td>
<td>0.0054</td>
</tr>
<tr>
<td>Enterococcus</td>
<td>916.4</td>
<td>10</td>
<td>0.8865</td>
<td>0.0143</td>
</tr>
<tr>
<td>Lachnospiraceae</td>
<td>966</td>
<td>10</td>
<td>0.85</td>
<td>0.0175</td>
</tr>
<tr>
<td>Mollicutes</td>
<td>485.2</td>
<td>10</td>
<td>0.8233</td>
<td>0.0268</td>
</tr>
</tbody>
</table>
### Table 5-4. Unique reactions within genera. Each row has n reactions that each column does not have.

<table>
<thead>
<tr>
<th></th>
<th>Akkermansia</th>
<th>Barnesiella</th>
<th>Blautia</th>
<th>Clostridium_difficile</th>
<th>Coprobacillus</th>
<th>Enterobacteriaceae</th>
<th>Enterococcus</th>
<th>Lachnospiraceae</th>
<th>Mollicutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akkermansia</td>
<td>0</td>
<td>80</td>
<td>80</td>
<td>95</td>
<td>86</td>
<td>66</td>
<td>76</td>
<td>103</td>
<td>113</td>
</tr>
<tr>
<td>Barnesiella</td>
<td>163</td>
<td>0</td>
<td>155</td>
<td>167</td>
<td>167</td>
<td>148</td>
<td>65</td>
<td>144</td>
<td>155</td>
</tr>
<tr>
<td>Blautia</td>
<td>65</td>
<td>57</td>
<td>0</td>
<td>48</td>
<td>85</td>
<td>65</td>
<td>52</td>
<td>61</td>
<td>92</td>
</tr>
<tr>
<td>Clostridium_difficile</td>
<td>62</td>
<td>51</td>
<td>30</td>
<td>0</td>
<td>70</td>
<td>53</td>
<td>45</td>
<td>40</td>
<td>84</td>
</tr>
<tr>
<td>Coprobacillus</td>
<td>65</td>
<td>63</td>
<td>79</td>
<td>82</td>
<td>0</td>
<td>49</td>
<td>71</td>
<td>71</td>
<td>101</td>
</tr>
<tr>
<td>Enterobacteriaceae</td>
<td>73</td>
<td>72</td>
<td>87</td>
<td>93</td>
<td>77</td>
<td>0</td>
<td>74</td>
<td>95</td>
<td>109</td>
</tr>
<tr>
<td>Enterococcus</td>
<td>130</td>
<td>36</td>
<td>121</td>
<td>132</td>
<td>146</td>
<td>121</td>
<td>0</td>
<td>134</td>
<td>153</td>
</tr>
<tr>
<td>Lachnospiraceae</td>
<td>90</td>
<td>48</td>
<td>63</td>
<td>60</td>
<td>79</td>
<td>75</td>
<td>67</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>Mollicutes</td>
<td>90</td>
<td>49</td>
<td>84</td>
<td>94</td>
<td>99</td>
<td>79</td>
<td>76</td>
<td>70</td>
<td>0</td>
</tr>
</tbody>
</table>

### Table 5-5. Reaction overlap between genera

<table>
<thead>
<tr>
<th></th>
<th>Akkermansia</th>
<th>Barnesiella</th>
<th>Blautia</th>
<th>Clostridium_difficile</th>
<th>Coprobacillus</th>
<th>Enterobacteriaceae</th>
<th>Enterococcus</th>
<th>Lachnospiraceae</th>
<th>Mollicutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akkermansia</td>
<td>0</td>
<td>405</td>
<td>405</td>
<td>390</td>
<td>399</td>
<td>419</td>
<td>409</td>
<td>382</td>
<td>372</td>
</tr>
<tr>
<td>Barnesiella</td>
<td>0</td>
<td>0</td>
<td>413</td>
<td>401</td>
<td>401</td>
<td>420</td>
<td>503</td>
<td>424</td>
<td>413</td>
</tr>
<tr>
<td>Blautia</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>422</td>
<td>385</td>
<td>405</td>
<td>418</td>
<td>409</td>
<td>378</td>
</tr>
<tr>
<td>Clostridium_difficile</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>382</td>
<td>399</td>
<td>407</td>
<td>412</td>
<td>368</td>
</tr>
<tr>
<td>Coprobacillus</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>415</td>
<td>393</td>
<td>393</td>
<td>363</td>
</tr>
<tr>
<td>Enterobacteriaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>418</td>
<td>397</td>
<td>383</td>
<td>386</td>
</tr>
<tr>
<td>Enterococcus</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>405</td>
<td>392</td>
<td>0</td>
</tr>
<tr>
<td>Lachnospiraceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mollicutes</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Attractor Groupings

Figure 5-6. All possible steady states of the Boolean model of the gut microbiome. There are 23 predicted steady states in the Boolean model of the gut microbiome. Each attractor is a column in the heatmap and is made up of the state of each genus in the network model (rows). Each genus can be present above an activity threshold (yellow; ON) or below an activity threshold (blue; OFF). The steady states in the model are grouped based on their similarities to other steady states in the same group. The first steady state of group 2 (attractor 2) is the healthy steady state, the first steady state of group 3 (attractor 7) is the clindamycin treated steady state, and the first steady state of group 4 (attractor 12) is the clindamycin + C. difficile steady state. These states are directly corroborated by experimental metagenomic data.
Figure 5-7. Steady states and node perturbations in the gut microbiome model. A) Heatmap of the three steady states in the gut microbiome model. These steady states are identical to steady states identified in the three experimental groups. B) The effect of node perturbations represented by four heatmaps. On the Y-axis of each of the four heatmaps are nodes (genera) in each steady state. On the x-axis of each of the four heatmaps are the steady states found under normal model conditions (i.e. no node perturbations) and also the specific perturbation of a single network node. The two heatmaps in the left column of the figure demonstrate the effect of overabundance of individual genera, and the two heatmaps in the right column of the figure demonstrate the effect of knockouts of individual genera. The top row heatmaps show the effect of node perturbations on the Clindamycin treated group and the bottom row heatmaps show the effect of node perturbations on the Clindamycin +C. difficile treatment group.
Figure 5-8. Subsystem enrichment analysis highlights metabolic differences between taxa. The p-values from the enrichment analysis are log-transformed and negated, such that darker regions indicate greater enrichment. The enrichment analysis quantifies the likelihood that a given subsystem (row) would be as highly abundant as observed within a given metabolic reconstruction (column) by chance alone. A subset of 22 interesting subsystems is shown here. Subsystems of interest include those for which all taxa are enriched, such as glycolysis, and nucleotide sugars metabolism, highlighting the fact that all taxa contain relatively full compliments of reactions within those subsystems. Similarly, subsystems for which a single genus differs from the remaining genera are interesting, such as cyanoamino acid metabolism, where *C. difficile* is highly enriched for reactions in that subsystem. Some subsystems are differentially enriched between *Barnesiella* and *Lachnospiraceae*, and *C. difficile* such as lipopolysaccharide biosynthesis and cyanoamino acid metabolism.
5.4.7 Generating Metabolic Competition Scores

The two Lachnospiraceae genera were treated as metabolically identical, and the “Other”
group was excluded. Competition scores were generated for all ordered pairwise relationships between the genera considered in the model (self-edges were excluded). We divided the pairs of genera into five groups based on being connected by a positive or negative edge, a negative or positive path (meaning an indirect relationship), or no path. A positive relationship was found between competition score and edge type in the dynamic network (i.e. positive edges tend to have a higher competition score), which was not statistically significant, perhaps due to the small sample size (p-value = 0.058 by one-sided Wilcoxon rank sum test) (Figure 5-10A). The mutualism score did not display any obvious trends with respect to the network structure (Figure 5-10B). All pairs with inferred edges exhibited relatively high competition scores and low mutualism scores (Figure 5-10C). *Barnesiella*, a key inhibitor of *C. difficile* in the interaction network, holds the second smallest competition score with *C. difficile* (see Figure 5-10A). *Barnesiella* and *C. difficile* also have the highest mutualism score among all interacting pairs in the network (Figure 5-10C).

Figure 5-10. Competition and mutualism scores by edge and path type in Boolean network. A) Competition score values for all classes of path through the network, including direct edges, directed paths, and no directed path. A positive relationship was found between competition score and direct edge type in the dynamic network (self-edges were excluded), which was not statistically significant, perhaps due to the small sample size (p-value = 0.058 by one-sided Wilcoxon rank sum test), but is worthy of note. B) Mutualism score values for all classes of path through the network, including direct edges, directed paths, and no directed path. C) Competition and mutualism score plot for the interaction edges in the network. All the interactions reflect moderate to high competition scores and relatively low mutualism scores. All the interactions have a higher competition score than mutualism score. The two negative interactions (red circles) do not have higher competition scores, nor lower mutualism scores, than the positive interactions. In fact, the negative interaction between *Barnesiella* and *C. difficile* corresponds to the highest mutualism score.
5.5 Discussion

Here we have developed a novel strategy for generating a dynamic model of gut microbiota composition by inferring relationships from time series metagenomic data (Figure 5-1). To our knowledge, this is the first Boolean dynamic model of a microbial interaction network and the first Boolean model inferred from metagenomic sequence information. Metagenomic sequencing is a powerful tool that tells us the consequences of microbial interaction - changes in bacterial abundance. Bacterial interactions are, in fact, mediated by the many chemicals and metabolites the bacteria use and produce. In a network sense these relationships are a bipartite graph; bacterial genera produce chemicals/metabolites/factors, which have an effect on other bacteria. Because there is no comprehensive source for the bacterial metabolites and their effect on other bacterial genera, we infer the effects of genera on each other from the relative abundances of genera in a set of microbiome samples, and we employ genome scale metabolic reconstructions to gain insight into these relationships (Figure 5-11B). Binarization of the microbial abundances clarifies these relationships and is the starting point for the construction of a dynamic network model of the gut microbiome. Interestingly, principal component analysis demonstrates that the time series data clusters by experimental treatment group, suggesting that our initial assumption of binary relationships does not lead to significant information loss (Figure 5-5A and Figure 5-5B).

We analyze the topological and dynamic nature of the gut microbiome, focusing on the effect of clindamycin antibiotic and C. difficile infection on gut microbial community structure. We generate a microbial interaction network and dynamic model based on time-series data from a population of mice. We validate a key edge in this interaction network between Barnesiella and C. difficile through an in vitro experiment. Consistent with the literature, our model affirms that solely inoculating a healthy microbiome with C. difficile is insufficient to disrupt the healthy intestinal tract microbiome. Additionally, our results demonstrate that clindamycin treatment has a tremendous effect on the microbiome, greatly reducing many microbial genera, and that if during that time C. difficile is present, a certain subset of bacteria come to dominate the microbiome (Figure 5-2, Figure 5-3, and Figure 5-5C).

Our dynamic network model reveals the steady state conditions attainable by this microbial system, how those steady states are reached and maintained, how they relate to the health or disease status of the mice, and how targeted changes in the network can transition the community
from a disease state to a healthy state. Furthermore, we examine genome-scale metabolic reconstructions of the taxa represented in this community, examine broad metabolic differences between the taxa in the community, and probe how metabolism could—and could not—contribute to the mechanistic underpinnings of the observed interactions.

5.5.1 Network Structure

The first feature that stands out in the inferred interaction network is its clustered structure. Clindamycin has a strong influence on the subnetwork containing the two Lachnospiraceae nodes and Barnesiella. The other subnetwork contains C. difficile and other genera that become abundant during C. difficile infection (Figure 5-5D). This initial structure is consistent with published data in which the dominant Firmicutes (Lachnospiraceae) and Bacteroidetes (Barnesiella) are devastated by antibiotic administration (50, 51). Furthermore, the clustered structure (Figure 5-5D) supports the established mechanism of C. difficile colitis: loss of normal gut flora, which normally suppresses opportunistic infection (clindamycin cluster) and the presence of C. difficile at a minimum inoculum (C. difficile cluster) (52) (11). The network clusters have a single route of interaction between Barnesiella and C. difficile.

Also worth noticing are the two contradicting edges in the network, between Coprobacillus and Blautia, and the self-edges for Blautia (Figure 5-5D). These arise from rules in the Boolean model that are context-dependent. Such context-dependent rules can manifest as opposite edge types, depending on the state of other nodes in the network. Context-dependent interactions have been demonstrated in many microbial pairings, and nutritional environments can even be designed to induce specific interaction types (53). It is possible that subtle environmental changes over the course of the experiment altered conditions in a way that flipped the Coprobacillus-Blautia interaction.

The negative influence of Barnesiella on C. difficile is in agreement with recently published findings in which Barnesiella was strongly correlated with C. difficile clearance (54). The role of Barnesiella as an inhibitor of another pathogen (vancomycin-resistant Enterococci (VRE)) has been shown in mice (55), which is also visible in the network model as an indirect relationship between Barnesiella and Enterococcus (Figure 5-5D). Related species of Bacteroidetes have been shown to play vital roles in protection from C. difficile infection in mice (56). Furthermore, the network structure shows that Lachnospiraceae positively interacts with Barnesiella, leading
to an indirect suppression of *C. difficile*. Interestingly, the two *Lachnospiraceae* nodes and the “Other” node form a strongly connected component, suggesting a similar role in the network, particularly in promoting growth of *Barnesiella*, which directly suppresses *C. difficile*. In support of this finding, *Lachnospiraceae* has been shown to protect mice against *C. difficile* colonization (51, 57). Therefore, the structure of the network is both a parsimonious representation of the current data set, and is supported by literature evidence.

Because the interaction network is derived from time-series data, it is possible to estimate causality, and therefore, derive a directed graph. A directed network with clear, causative interactions can be used to study community dynamics. This is in contrast with association networks, which are often derived from independent samples, and cannot determine direction of causality (47, 58-60). Such networks are more limited in utility because they cannot be used to predict system behavior over time, or system responses to perturbations (23, 61).

### 5.5.2 Experimental Validation of *Barnesiella* Inhibition of *C. difficile*

We experimentally validated a key edge in the interaction network, and showed that *Barnesiella* can in fact slow *C. difficile* growth. *C. difficile* was grown alone, in co-culture with *B. intestinihominis*, and in *B. intestinihominis* spent media. *C. difficile* grew more slowly in both co-culture and spent-media conditions. Though moderate, the effect was statistically significant (Figures 5-9B). The fact that *C. difficile* growth rate was inhibited under spent-media conditions indicates that *B. intestinihominis*-mediated inhibition does not require *B. intestinihominis* to “sense” the presence of *C. difficile*. Further, *C. difficile* growth on *B. intestinihominis* spent media demonstrates that the two species have different nutrient requirements. Whether the reduction in growth rate is a result of nutritional limitations (e.g. *C. difficile* resists to a less preferred carbon source) is unknown, but unlikely given the AUC data.

The AUC—a summation of the OD over the entire time course—is a measure of the total bacterial density achieved over the course of the experiment. It can be thought of as representing the total biomass produced over time. Examining the AUC for all conditions showed that *C. difficile* AUC did not significantly change between fresh media and spent media (Figure 5-9C). Thus, *C. difficile* was able to produce comparable overall biomass despite a reduction in growth rate, further demonstrating that nutrient availability was sufficient in the spent media condition. The AUC for the co-culture was much lower than expected in a simulated null model (Figure 5-
9C). Apparently, in co-culture, the total community biomass production capacity is reduced from what would be expected in a scenario without species interaction. Thus, there is a measurable negative interaction between \textit{B. intestinohominis} and \textit{C. difficile} in co-culture that impacts biomass production. This can be observed over the full time-course of the co-culture, where the overall density is consistently lower than what would be expected in a null model (Figure 5-9D).

5.5.3 Network Dynamics and Perturbation Analysis

Computational perturbation analysis showed that forced overabundance of \textit{Barnesiella} led to a shift from the “disease” state (clindamycin+ \textit{C. difficile} treatment group) to a state highly similar to the original healthy state (loss of \textit{C. difficile}). This result is particularly interesting from a therapeutic design standpoint. In this case, the model results indicate that Barnesiella may serve as an effective probiotic. Model-driven analysis can be used to identify candidate organisms for probiotic treatments. Recent work by Buffie \textit{et al.} performed a proof-of-concept study in which they used statistical models to identify candidate probiotic organisms, which were then tested on a murine model of \textit{C. difficile} infection (54). This model-driven approach can be favorably contrasted with the brute-force experimental approach in which successive combinations of microbes are tested until a curative set is found (56). The model-driven approach requires far fewer experiments, and saves time and resources. While the computational model presented here differs from that used by Buffie \textit{et al.}, the integration of computational models in probiotic design has been shown to be a feasible, effective approach. Improved tools, such as the perturbation analysis presented here, will surely accelerate the probiotic design process and shorten the path to the clinic.

5.5.4 Metabolic Competition Scores Point Towards a Non-Metabolic Interaction Mechanism

Genome-scale metabolic network reconstructions can be used to estimate the interactions between microbes in a complex community based purely on genome sequence data. Our use of genus-level metabolic reconstructions (a union of several species-level reconstructions) may not reflect the unique, species-level interactions and heterogeneity within a community. This higher-level model will only capture broad trends and the possible extent of metabolic capacity within a genus. The positive relationship between edge type and competition score suggests that more metabolic similarity between genera tends to foster positive interaction. The converse is also
true, where less metabolic similarity tends to foster negative interactions (Figure 5-10A). Here, “positive/negative interaction” is derived from the Boolean model, where a positive edge between species A and B indicates that if A is ON at time t, then B is likely to turn ON at t+1. The approach demonstrated here could potentially be used to quickly establish a rough expectation (notice the spread of competition scores for the species pairs not connected by a path through the network) for community structure—based solely on genomic information—that can then be tested experimentally.

Interestingly, the fact that the relationship is positive (higher competition score leads to more positive interactions) relates with previous work that demonstrates that higher competition scores were associated with habitat co-occurrence (45). In this same work, the authors suggest that this effect is due to habitat filtering; that is, microbes with similar metabolic capabilities tend to thrive in similar environments. It has been shown experimentally that microorganisms from the same environment tend to lose net productivity in batch co-culture, indicating similar metabolic requirements (62). Thus, metabolically similar organisms tend to co-locate to similar niches, and over evolutionary time, co-localized organisms tend to develop positive relationships with each other.

Understanding this relationship between competition score and interaction type leads to the conclusion that negative interactions are probably not caused by metabolic competition. Of all the genus competition scores with C. difficile, Barnesiella showed the second lowest (Figure 5-9A). In other words, Barnesiella is among the least likely to share a similar metabolic niche with C. difficile, which fits with the broad trend mentioned above. The fact that the competition score between C. difficile and Barnesiella is so low strongly suggests that the negative interaction between them is due, not to competition for scarce resources (although it does not exclude the possibility), but rather to some non-metabolic mechanism. The similarity in reaction content between Barnesiella and Enterococcus indicates similar network structure (Table 5-5), and yet, Enterococcus does not inhibit C. difficile in the inferred interaction network (Figure 5-5D). Either the differences that are present between Barnesiella (65 unique reactions) and Enterococcus (36 unique reactions) are hints at the mechanism of interaction, or metabolism does not play a significant role in C. difficile inhibition in the environment of the gut. For example, enrichment analysis showed that Barnesiella is more highly enriched for d-glutamine and d-glutamate metabolism, pantothenate and CoA biosynthesis, LPS biosynthesis. With
respect to *Enterococcus*, *Barnesiella* is less enriched in pyrimidine metabolism, and phenylalanine, tyrosine, and tryptophan biosynthesis. The possible role of LPS is discussed further on. The possible roles of these other metabolic pathways in *C. difficile* inhibition is unclear.

There is experimental evidence that *Barnesiella* (and other normal flora) may combat pathogen overgrowth through non-metabolic mechanisms. As a first step, it has been shown that VRE can grow in sterile murine cecal contents—indicating the presence of sufficient nutrition to support VRE—but is inhibited in saline-treated cecal contents—indicating that live flora are needed to suppress VRE growth, and that this suppression is not through nutrient sequestration (63). Further, the presence of *B. intestinihominis* has been demonstrated to prevent and cure VRE infection in mice (55), and is strongly correlated with resistance to *C. difficile* infection in mice (54). Clearly, *Barnesiella* plays a key role in pathogen inhibition, and pathogen inhibition can be caused by mechanisms other than nutrient competition.

This non-metabolic mechanism may be direct or indirect (Figure 5-11A). We demonstrated *in vitro* that *B. intestinihominis* can inhibit *C. difficile* growth rate (Figures 5-9C and 5-9D). The fact that *C. difficile* grows on *B. intestinihominis* spent media at all indicates that the metabolic requirements of the two species are different, which is consistent with our computational results supporting the hypothesis that *C. difficile* and *Barnesiella* do not compete metabolically (Figure 5-9B). Further, *C. difficile* is moderately inhibited both in co-culture with *B. intestinihominis* and in *B. intestinihominis*-spent media, indicating a direct mechanism of inhibition. In further support of a direct mechanism, it has been shown that *Clostridium scindens* inhibits growth of *C. difficile* through the production of secondary bile acids (54). Perhaps *Barnesiella* works through an analogous mechanism *in vivo*, enhancing the moderate inhibition observed *in vitro*.

In support of an additional indirect mechanism of bacterial interaction, Buffie and Pamer, in a recent review, hypothesized that the normal flora (of which *Barnesiella* is a member) may prevent pathogen overgrowth by stimulation of a host antimicrobial response (64) (Figure 5-11A). Specifically, they point out that *Barnesiella* can activate host toll-like receptor (TLR) signaling, which activates host antimicrobial peptide production. For example, LPS and flagellin have been shown to stimulate the host innate immune response through toll-like receptor (TLR) signaling and production of bactericidal lectins (65, 66). *Barnesiella* shows enrichment for LPS biosynthesis pathways (Figure 5-8). However, this mechanism did not seem to be responsible for
inhibition of VRE by *Barnesiella* (55). An indirect, host-mediated mechanism is further supported by the fact that members of the normal gut flora can interact differently with pathogens depending on the host organism (54). Regardless, any indirect mechanism is in addition to the direct inhibitory mechanism observed *in vitro*. Both direct and indirect mechanisms may play a role *in vivo*, and further work is needed to clearly discern the underlying process that allows *Barnesiella* to play this protective role.

5.6 Conclusions

We demonstrate that dynamic Boolean models capture key microbial interactions and dynamics from time-series abundance data in a murine microbiome. We show that this computational approach enables exhaustive *in silico* perturbation, which leads to fast candidate selection for probiotic design. We further describe the use of genome-scale metabolic network reconstructions to explore the metabolic potential attributed to community members, and to estimate metabolic competition and cooperation between members of the microbiome community. Analysis of genome-scale metabolic reconstructions indicates that *Barnesiella* inhibits *C. difficile* through non-metabolic mechanism. We present empirical *in vitro* evidence that *B. intestiniihominis* does in fact inhibit *C. difficile* growth, likely by a non-metabolic mechanism, and our findings are in good agreement with published results from multiple research groups. We present this work as a demonstration of the use of dynamic Boolean models and genome-scale metabolic reconstructions to explore the structure, dynamics, and mechanistic underpinnings of complex microbial communities.

5.7 Acknowledgments

This work was collaborative effort with Matthew Biggs (PhD student in Jason Papin’s group at UVA). Matthew Biggs did metabolic analysis described in the figures/text. The manuscript describing this work is currently under review. We thank the Jefferson Trust and University of Virginia Data Science Institute Big Data Fellowship for supporting this endeavor. We thank Dr. David J. Feith (University of Virginia) for helpful comments/suggestions. We further thank Dr. Glynis Kolling (University of Virginia) for help obtaining bacterial isolates and carrying out *in vitro* experiments. Research reported in this publication was supported by the National Institutes of Health under Award Number R01 GM108501 (NIGMS) and F30 DK093234 (NIDDK). The
content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

5.8 References


CHAPTER 6.
Conclusions & Outlook

The beauty of network theory is that real world systems as diverse as the World Wide Web, human social networks, electrical circuits, and gene regulatory networks, can all be represented in a unifying framework – as a network. Because of this, the study of networks has led to generalizable physical properties for how complex systems work. For example, the identification that many real world networks have scale free degree distributions (1) and specific control profiles (2). This makes network theory such a powerful tool because studying a specific network can lead to identification of generalized principals and physical properties, a potentially far-reaching impact. The more that is learned about biological systems, the more we realize we do not know about them. The complexity of biological systems is extraordinary, and network theory provides the tools to understand such complex systems. Furthermore, biological networks are not static systems, but dynamically change in response to environmental or even structural (e.g. genetic) changes over time. A network can be translated into a dynamic model, a mathematical representation of the system, to study the time course evolution of a system. Throughout my dissertation, we apply network theory and dynamic modeling to understand three biological phenomena that are critical to human health: two models of signal transduction networks in cancer and a model of an ecological community of intestinal bacteria. The overarching goal of my work has been to translate network-directed insight into actionable biomedical findings that lead to better understanding of human disease and improved patient care. Here we summarize these results and explore future work for each system considered.

In chapter two we describe a model of EGFR signaling and its effect on tumor growth. We use this as a bridge between the introduction and the research chapters because it provides an opportunity to see how discrete (Boolean and multi-state) dynamic models behave compared to another commonly used dynamic modeling approach, ordinary differential equation (ODE) modeling. ODE models are generally considered superior to discrete models at describing the quantitative properties of a system when specific information (e.g. reaction rates and substrate concentrations) is available to parameterize these models. However, in biology this kind of information is frequently unavailable. Parameters for the simplistic ODE model of EGFR signaling that we use to create a discrete model are well known so construction of an ODE model
is feasible (3). Our discrete dynamic model of EGFR signaling is able to capture a majority of the results of ODE models with no required parameters. The parsimonious nature of these models is an eloquent example of how the structure of a network is more important than the specific kinetic information about the system. Additionally, EGFR and MAP kinase signaling are some of the most well studied signal transduction pathways, with known kinetic parameters; however, there are many pathways and systems where this kind of information does not exist. Discrete modeling would be a powerful tool in those systems. In terms of future work in this arena, this model of EGFR signaling is, in fact, quite simplistic. There exist many other proteins downstream of EGFR. A Boolean or multi-state discrete model of a larger EGFR signaling network could be constructed in the future to more comprehensively study the effect of EGFR signaling, particularly with the inclusion of nodes that are frequently mutated in cancers where EGFR signaling is important (e.g. non-small cell lung cancer). A more comprehensive model of EGFR signaling in lung cancer could be used to determine a priori whether a treatment will shrink the tumor and by how much. Additionally, the model could be used to predict putative secondary targets that would shrink a tumor when the tumor has a mutation that makes the cancer less susceptible to EGFR inhibitor therapy.

In chapters three and four we describe a model of a critical mechanism to the invasive and metastatic features of cancers, called epithelial-to-mesenchymal transition (EMT). Invasive and metastatic disease is what makes cancers (specifically solid tumors) harmful and thus identifying ways to suppress this process is an important medical problem. Additionally, the known EMT regulatory network is quite complex. We constructed this model and used it to identify putative targets to suppress EMT and invasive disease, and we were able to validate some of these targets in a human liver cancer cell line. It would be important to explore these targets in further cell line models and through more extensive analysis of invasive behavior. This would include studying more protein markers of the EMT phenotype, testing invasiveness through invasion assays (e.g. Boyden chamber assay), and in vivo invasion and metastasis through orthotopic mouse models. Additionally, an interesting future avenue would be personalization of the EMT network model in the context of known mutations in the tumor sample or cell line in which EMT is being studied. This issue is really important in cancer because the cancers networks that exist are not identical between patients because of mutations and epigenetic changes that effect which nodes are expressed and also the edges in the network. In diseases where the network structure is static
(i.e. genome and epigenome is stable), such as cardiovascular disease (4) and immune system regulation, this is likely less of a concern. Nonetheless, the EMT network we have constructed can be considered a “theme” network, representing a framework for which personalization to what is known about individual patients or cell lines can be employed as an additional layer. Employing genomic aberrations would likely improve the accuracy of model predictions.

Incorporating data driven approaches into a priori network construction would also be an interesting extension of this work to build onto the model and could potentially aid in model personalization. For example, time course TGFβ treatments in an epithelial cell line undergoing EMT could be completed by multiple assays for cellular changes in the signaling network by gene expression (e.g. RNA-seq), protein expression (e.g. multiplex phospho-protein assays or mass spectrometry-based approaches), and transcription factor binding information (e.g. ChIP-seq). One of the attractive features of network models is that they provide a unifying representation of heterogeneous information (i.e. you can represent multiple levels of biological information in a single model). Constructing a signaling network using this information from the signal (TGFβ) through the signal transduction pathway to the transcription factor binding and transcriptional target information could yield novel insight into how cells undergo EMT. Development of computational approaches to extract and agglomerate these diverse biological information sources is not a trivial pursuit but the task is certainly worthwhile.

TGFβ was the focus of this work; however, there exist numerous signals that can drive EMT, many of which exist in our constructed network model. Exploring other signals and also combinations of signals would be an important future direction, as there are likely multiple different signals in the tumor microenvironment that are driving EMT. Lastly, tumors evolve over time due to genomic instability. Thus, incorporation of some kind of evolutionary mechanism into the EMT model (or any network model of cancer) should be both feasible and a worthwhile pursuit. The network could mutate and evolve over time, through changes to nodes and edges. This could be done in a few ways: either by randomly mutating all nodes in the model or by assigning mutation probabilities based on known mutations/mutation rates in the specific cancer being studied. Additionally, some kind of fitness function could be employed to determine which networks are most likely to survive, proliferate, and/or metastasize. This could be used to more accurately model the heterogeneity seen in tumors.
In chapter five, we developed a novel methodology to construct a Boolean dynamic model from time series metagenomic genus abundance information and integrated this modeling with genome scale metabolic network reconstructions to identify metabolic underpinnings for microbial interactions. We applied this in the context of a critical health issue: clindamycin antibiotic treatment and opportunistic *C. difficile* infection. Our model recapitulates known dynamics of clindamycin antibiotic treatment and *C. difficile* infection and predicts therapeutic probiotic interventions to suppress *C. difficile* infection. Genome scale metabolic network reconstructions reveal metabolic differences between community members and highlight the role of metabolism in the observed microbial interactions. We experimentally tested the prediction that *Barnesiella* inhibits *C. difficile* growth and it appeared to inhibit *C. difficile* growth *in vitro*. This should be tested in *in vivo* preclinical (e.g. murine) models of *C. difficile* infection, and then by clinical trials as a probiotic to suppress *C. difficile* colitis, a major health burden. Interestingly, there is evidence accumulating that *Barnesiella* suppress vancomycin-resistant enterococcus (VRE), another critical health issue (5). A recent study blindly screened numerous bacterial species to determine their ability to suppress *C. difficile* infection and their results suggest that *Barnesiella* suppresses *C. Difficile* infection (6). The cost of this screen and pre-clinical/clinical research is not trivial and this is one advantage of using a dynamic model. Our inferred network model supports these findings and could be used to reduce overall amount of experimental testing and thus experimental costs. Another interesting predicted intervention to test would be the effect of *Lachnospiraceae* on *C. Difficile* growth. Because experiments used to infer this model were only sampling a subset of the possible microbiome state space, we would be less confident making inferences using the model outside of the context of clindamycin and *C. difficile* dynamics. We would speculate that incorporating more replicates and more test treatment interventions (e.g. other antibiotics, chemical perturbations, or diseases that affect the microbiome) into the inference pipeline, would likely lead to better and more generalizable rules for the model.

Other avenues for future investigation include construction of a model with more genera than the eleven most abundant genera we included, as there could be less abundant bacterial genera that have important effects on *C. difficile* and clindamycin dynamics. Going beyond that, a species level model would be even more interesting, as there exists evidence that not all species within the same genera play the same ecological roles. Albeit this would be more difficult to
construct and likely more noisy. Furthermore, specifically identifying the reasons for the edges in the network model is an important future goal. We began to explore metabolic relationships that could explain commensal or mutualistic relationships between bacterial genera; however, it is possible that non-metabolic mechanisms exist (e.g. secreted anti-microbial peptides). Experimental metabolomics and proteomic analysis would be useful in identifying specific biomolecules responsible for inferred relationships.

Classically, biological systems have been studied from a reductionist viewpoint. However, the emergence of network theory provides a methodology for holistic investigation of complex biological systems. The body of work presented here, demonstrates just how network theory, particularly network dynamics, can inform us about biological systems. We particularly focus on how to translate network-directed insight into actionable biomedical findings that lead to improved understanding of human disease and novel therapeutic targets for diseases with poor treatment options.

VITA
Steven Nathaniel Steinway

**Education**

<table>
<thead>
<tr>
<th>Year</th>
<th>Degree</th>
<th>Institution</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>2017</td>
<td>Doctor of Medicine (M.D.)</td>
<td>The Pennsylvania State University, Hershey, PA</td>
<td>USA</td>
</tr>
<tr>
<td>2015</td>
<td>Doctor of Philosophy (Ph.D.) in Molecular Medicine</td>
<td>The Pennsylvania State University, Hershey, PA</td>
<td>USA</td>
</tr>
<tr>
<td>2009</td>
<td>Bachelor of Science (B.S.) in Biology</td>
<td>The College of New Jersey, Ewing, NJ</td>
<td>USA</td>
</tr>
</tbody>
</table>

**Grants/Awards**

<table>
<thead>
<tr>
<th>Year</th>
<th>Grant/Award</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>Data Science Institute/Jefferson Trust, Big Data Collaborative Research</td>
<td>Fellowship, University of Virginia; <a href="http://gradstudies.virginia.edu/node/315">http://gradstudies.virginia.edu/node/315</a></td>
</tr>
<tr>
<td>2014</td>
<td>Data Science Institute, Big Data Travel Award</td>
<td>University of Virginia</td>
</tr>
<tr>
<td>2011-</td>
<td>Present National Institute of Health National Research Service Award F30</td>
<td>National Institute of Diabetes, Digestive, and Kidney Diseases, NIH</td>
</tr>
<tr>
<td>2011</td>
<td>Institute for Cyberscience Summer Student Award</td>
<td>Penn State University</td>
</tr>
<tr>
<td>2010</td>
<td>North American Society for Pediatric Gastroenterology, Hepatology and</td>
<td>Nutrition (NASPGHAN) Summer Student Award</td>
</tr>
<tr>
<td>2010</td>
<td>Department of Biology Alumni Award</td>
<td>Honorable Mention, The College of New Jersey</td>
</tr>
<tr>
<td>2009</td>
<td>Phi Kappa Phi Faculty-Student Research Award</td>
<td>The College of New Jersey</td>
</tr>
<tr>
<td>2008</td>
<td>NIH/NSF Bioinformatics Summer Institute Fellowship</td>
<td>University of Minnesota</td>
</tr>
<tr>
<td>2007</td>
<td>NSF Research Experience for Undergraduates (REU) in Bioinformatics</td>
<td>Fellowship, Loyola University in Chicago</td>
</tr>
</tbody>
</table>

**Selected Publications**


Steinway SN, Zañudo JGT, Michel P, Feith DJ, Loughran TP Jr., Albert R. Combinatorial Interventions Lend Insight into the Inhibition of TGFβ Driven Epithelial-to-Mesenchymal Transition and the Existence of Hybrid Epithelial-Mesenchymal Phenotypes. *In Preparation*.

Steinway SN, Dang H, You H, Rountree CB, Ding W. The EGFR/ErbB3 Pathway Acts as a Compensatory Survival Mechanism upon c-Met Inhibition in Human c-Met+ Hepatocellular Carcinoma. *(Under Revision at PLOS ONE)*.

Dang H, Steinway SN, Ding H, Rountree CB. Induction of tumor initiation is dependent on CD44s in c-Met+ hepatocellular carcinoma. *Accepted. BMC Cancer*.


