5-31-2021

Modeling and analysis of intracellular signaling networks and cellular decisions

Mustafa Ozen
New Jersey Institute of Technology

Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

MODELING AND ANALYSIS OF INTRACELLULAR SIGNALING NETWORKS AND CELLULAR DECISIONS

by
Mustafa Ozen

Developing molecular network analysis methods is important due to their applications on complex biological systems such as target discovery, development of drugs, discovering drug effects, and finding treatments for many complex diseases, e.g., cancer, autoimmune, and mental disorders. An example of analysis techniques is the fault diagnosis analysis, in which the purpose is to quantify how much vulnerable the entire network is to dysfunction of one or multiple molecules. Such analysis can be done after proper network models are implemented, trained, and tested against the experimental data. In this dissertation, a Boolean modeling framework is implemented and methods to train the models against data are presented on multiple networks. Furthermore, a mathematical framework for executing single and multi-fault vulnerability analysis of a given molecular network using the trained network models is provided. In addition, the worst possible signaling failures in molecular networks is examined by comparing the maximum vulnerability level, i.e., the highest probability of network failure, versus the number of faulty molecules to understand how the network functionality is affected in the presence of one or more dysfunctional molecules, for which an efficient algorithm is developed. Moreover, another algorithm is proposed that outputs the maximum number of time points needed for computing the vulnerability level of molecules in a Boolean domain. The methods are applied to the experimentally verified ERBB and T cell signaling networks. The results reveal that as the number of faulty molecules increases,

the maximum vulnerability values do not necessarily increase, which means that a few faulty molecules can cause the most detrimental network damages and an increase in the number of faulty molecules does not deteriorate the network function. Such a group of molecules whose dysfunction causes the worst signaling failure may contribute to the development of the disorder and can suggest some therapeutic strategies.

Abnormality of a highly vulnerable molecule or a group of molecules results in incorrect network responses, which may cause the entire cell to make wrong decisions on the received signals and hence may initiate bigger events causing complex diseases. Therefore, characterization of decision-making in cells in response to received signals is of importance for understanding how cell fate is determined in the absence and presence of such abnormalities. Considering the cellular heterogeneity and dynamics of biochemical processes, the problem becomes multi-faceted and complex. This dissertation reveals a unified set of decision-theoretic, machine learning, and statistical signal processing methods and metrics to model the precision of signaling decisions in the presence of uncertainty, using single-cell data. This is done by presenting an optimal decision strategy minimizing the total decision error probability. Later, the framework is extended to incorporate the dynamics of biochemical processes and reactions in a cell, using multi-time point measurements and multidimensional outcome analysis and decision-making algorithms. Furthermore, the developed binary outcome analysis and decision-making approach is extended to more than two possible outcomes. As an example, and to show how the introduced methods can be used in practice, they are applied to single-cell data of PTEN, an important intracellular regulatory molecule in a p53 system, in wild-type and abnormal cells.

MODELING AND ANALYSIS OF INTRACELLULAR SIGNALING NETWORKS AND CELLULAR DECISIONS

by
Mustafa Ozen

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Electrical Engineering

Helen and John C. Hartmann Department of
Electrical and Computer Engineering

May 2021

APPROVAL PAGE

MODELING AND ANALYSIS OF INTRACELLULAR SIGNALING NETWORKS AND CELLULAR DECISIONS

Mustafa Ozen

Dr. Ali Abdi, Dissertation Advisor — Date
Professor of Electrical and Computer Engineering, New Jersey Institute of Technology

---

Dr. Alexander Haimovich, Committee Member — Date
Distinguished Professor of Electrical and Computer Engineering, New Jersey Institute of Technology

---

Dr. Reka Albert, Committee Member — Date
Distinguished Professor of Physics and Biology, Pennsylvania State University

---

Dr. Joerg Kliewer, Committee Member — Date
Professor of Electrical and Computer Engineering, New Jersey Institute of Technology

---

Dr. Hongya Ge, Committee Member — Date
Associate Professor of Electrical and Computer Engineering, New Jersey Institute of Technology

BIOGRAPHICAL SKETCH

Author: Mustafa Ozen

Degree: Doctor of Philosophy

Date: May 2021

Undergraduate and Graduate Education:

- Doctor of Philosophy in Electrical Engineering,
  New Jersey Institute of Technology, Newark, NJ, 2021

- Master of Science in Mathematics,
  Georgia Southern University, Statesboro, GA, 2015

- Bachelor of Science in Mathematics and Computer Science,
  Cankaya University, Ankara, Turkey, 2014

- Bachelor of Science in Electronic and Communication Engineering,
  Cankaya University, Ankara, Turkey, 2014

Major: Electrical Engineering

Presentations and Publications:

Ozen, M., Emamian, E. S., & Abdi, A. (2021). Exploring extreme signaling failures in intracellular molecular networks. *Submitted.*

Ozen, M., Emamian, E. S., & Abdi, A. (2021). Learning molecular network models with feedbacks using integer linear programming. *Submitted.*

Ozen, M., Lipniacki, T., Levchenko, A., Emamian, E. S., & Abdi, A. (2020). Modeling and measurement of signaling outcomes affecting decision making in noisy intracellular networks using machine learning methods. *Integrative Biology*, *12*(5), 122–138. https://doi.org/10.1093/intbio/zyaa009

Ozen, M., Lesaja, G., & Wang, H. (2020). Globally optimal dense and sparse spanning trees, and their applications. *Statistics, Optimization & Information Computing*, *8*(2), 328-345. https://doi.org/10.19139/soic-2310-5070-855

Ozen, M., Wang, H., Wang, K., & Yalman, D. (2016). An edge-swap heuristic for finding dense spanning trees. *Theory and Applications of Graphs*, *3*(1), Article 1. https://doi.org/10.20429/tag.2016.030101

Lesaja, G. & Ozen, M. (2016). Improved full-Newton-step infeasible interior-point method for linear complementarity problems. *Croatian Operations Research Review: CRORR*, *7*(1), 1-18. https://doi.org/10.17535/crorr.2016.0001

Ozen, M., Wang, H., & Yalman, D. (2016). Note on Leech-type questions of trees. *Integers, 16*, Article 21.

Ozen, M. (2015, March 13-14). *Improved full-Newton-step interior point methods for linear optimization (LO) and linear complementarity problems (LCP)* [Conference session]. Mathematical Association of America Southeastern Section Spring 2015 Meeting, Wilmington, NC, United States.

To my family and to all who supported me
(Can yoldaşım Demet, babam Sabri, annem Zeliha ve kardeşlerim
başta olmak üzere beni destekleyen herkese…)

What really matters is not what you have achieved so far;

it is how far you are away from your goals.

ACKNOWLEDGMENT

First, I would like to express my special appreciation and thanks to my advisor Dr. Ali Abdi for enlightening me. Without his guidance, inspirations, and dedicated involvement in every step throughout my PhD studies, this dissertation would have never been accomplished. Also, I would like to thank to our collaborator Dr. Effat S. Emamian for her guidance and patience, who made this dissertation possible by supporting us in every step.

Second, I would like to thank to my committee members Dr. Alexander Haimovich, Dr. Reka Albert, Dr. Joerg Kliewer, and Dr. Hongya Ge for their valuable time and insightful comments on my dissertation.

Third, I would like to thank to all faculty and staff members of the Department of Electrical and Computer Engineering at NJIT for their help and support. Special thanks to the members of Center for Wireless Information Processing (CWiP), especially Ms. Kathleen Bosco for her help and making the center feel like a home to all of us.

Fourth, I would like to thank to my family and friends for their support and encouragement. Many thanks to my parents Sabri Ozen and Zeliha Ozen who raised me with love and supported me with all the sacrifices in all my pursuits.

Most of all for my loving, caring, supportive, and encouraging wife, Demet Yalman Ozen. Thank you for always being there and for your endless, unrequited support in this adventure.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Overview

Molecular networks are the networks representing interactions between the molecules. They can be portrayed as a graph in which nodes represent biological molecules, e.g., proteins, RNA, genes, and edges represent physical or biochemical interactions such as regulatory relationships between the molecules (Hasty et al., 2001; Jeong et al., 2000; Levine & Davidson, 2005; Maslov & Sneppen, 2002). The type of the edges indicates the type of regulation such as activatory or inhibitory interaction (see Figure 1.1 for a toy example). There are different types of molecular networks such as protein-protein interaction (PPI) networks in which the nodes are proteins and edges are the physical interaction between them (Camargo et al., 2007), gene regulatory networks (GRNs) in which the nodes are transcription factors and target genes, and the edges are transcription regulation (Emmert-Streib et al., 2014), and cell signaling networks (Eungdamrong & Iyengar, 2004). These networks have various functionalities and used for different applications such as discovering and developing drugs and analyzing their effects as presented by Mitsos et al. (2009), developing fault diagnosis methods (Abdi et al., 2008; Habibi et al., 2014a, 2014b), understanding cell decision-making processes (Habibi et al., 2017; Hat et al., 2016; Ozen et al., 2020), and many other applications to model and understand complex human diseases. Thus, constructing and analyzing molecular networks and network models emerged in the past decades.

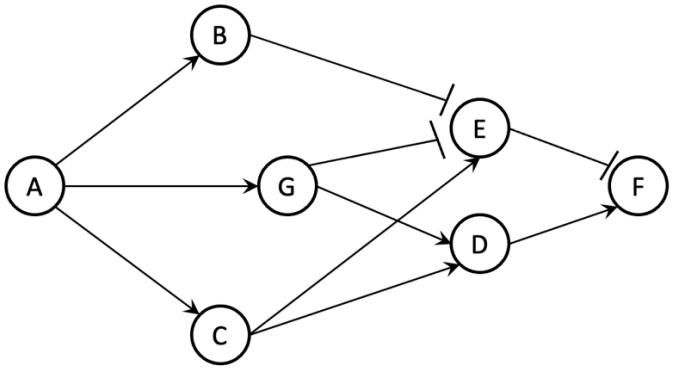


Figure 1.1 A graphical model of molecular networks.
Note: The normal arrows (→) represent an "activatory" relationship and blunt arrows (—|) represent an "inhibitory" relationship. The node at the starting point of an edge stands for input molecule and the node at the endpoint of the edge stands for a product (output). For instance, we say that node A activates node B and node B inhibits node E. A set of input nodes and a product together constitute a "reaction". To illustrate, nodes B, C, G, and E together represent a reaction in which B, C, and G are the input molecules and E is the output molecule (product). Consequently, it can be said that molecular networks are sets of reactions comprising inputs and an output.

To study molecular networks, one needs to convert molecular network graphs into numerable models so that they can be analyzed, and biologically relevant results can be inferred. There exist continuous models, discrete models, and hybrid models such as discrete models with continuous parameters (e.g., a model with logic-based ordinary differential equations (ODEs); Eduati et al., 2020) that are being studied. The continuous models convert molecular networks into a mathematical form by building a system of differential equations (Hat et al., 2016; Raue et al., 2013; Wittmann et al., 2009). They allow continuous tuning of the model parameters and hence provide more detailed network models. Although they provide detailed information, a drawback is that the knowledge of mechanistic details and kinetic parameters is very limited for continuous models, specifically in large networks with many molecules and interactions. They require free parameter estimation which is very challenging due to the lack of experimental data. In such scenarios, discrete models such as Boolean models are useful as they do not need

detailed kinetic information and still provide certain biologically relevant insight, as discussed by Abdi et al. (2008), Habibi et al. (2014a & 2014b), and in several review articles (e.g., Chaouiya et al., 2012; Chaouiya & Remy, 2013; Handorf & Klipp, 2012; Morris et al., 2010; Saadatpour & Albert, 2012, 2013; Samaga & Klamt, 2013; Stoll et al., 2012; Wang et al., 2012; Wynn et al., 2012).

Typically, the generated models for literature-curated molecular networks do not adequately fit experimental data due to the incompleteness of resources, databases, and literature used to construct the networks. The constructed network might be missing molecules/interactions that are biologically supposed to exist, or there might be molecules/interactions in the network that are irrelevant and should not be there. Another reason for the mismatch could be the model itself which may need to be tuned for higher accuracy. Thus, it is of interest to constitute tools to train the network models so that the networks with tuned parameters can efficiently represent the experimental data (Guziolowski et al., 2013; Saez-Rodriguez et al., 2009; Sharan & Karp, 2013). One way of training the network models is fixing the model rules and manipulating the network topology by systematically removing (adding) some interactions/molecules from (to) the network. In other words, one can seek a subnetwork of the initial network so that the subnetwork with fixed model rules has the optimal fit to the experimental data (Melas et al., 2013; Mitsos et al., 2009; Saez-Rodriguez et al., 2009). Another way of constructing biologically relevant models with high accuracy is by inferring the network directly from the experimental data. That is, the model itself can be learned from the data and the network can be constructed accordingly (Ideker et al., 2000; Saez-Rodriguez et al., 2009; Sharan & Karp, 2013; Videla et al., 2012). More specifically, one can fix the network topology with

its molecules and interactions then, systematically learn the functions or models of each molecule while optimizing the mismatch between the model predictions and the experimental data. Lastly, both removing (adding) interactions/molecules and learning functions of molecules can be done synchronously to obtain a network model reflecting the experimental data (Saez-Rodriguez et al., 2009; Sharan & Karp, 2013).

An important application of molecular network analysis methods is target discovery and drug development. This can be achieved by doing fault diagnosis analysis for which some computational and system biology techniques have been developed (e.g., Abdi et al., 2008; Abdi & Emamian, 2010; Habibi et al., 2014a, 2014b). The main purpose of such methods and many other approaches is to understand how vulnerable the entire network is to the dysfunction of each molecule. The dysfunction state of a molecule can be defined as a failure to respond correctly to its input signals, which may further induce incorrect responses at the output of the network. We define the vulnerability level of a molecule as the probability of having incorrect network responses when the molecule is dysfunctional. Vulnerability analysis can be performed for the dysfunction of a single molecule, as well as a group of molecules. The importance of the latter can be attributed to the widely known observations that many complex disorders such as schizophrenia are reported to be associated with the dysfunction of multiple molecules (Emamian, 2012). This contrasts with some diseases where only one molecule is known to cause the pathology (Emamian et al., 2003). Therefore, herein, we present methods for multi-fault vulnerability analysis in addition to single faults. By computing the vulnerability level of a molecule or a group of molecules, one can identify and rank the key components of the network that may

contribute to the development of the disorder. As one possible treatment strategy, such molecules can be targeted by certain specific therapeutic drugs.

The molecular networks have major roles in the characterization of cell fate. These networks generally have some specific outputs that initiate important biochemical processes. For example, depending on the received signals and dynamics of the network, a possible cell fate could be surviving or initiating apoptosis or moving in a certain direction, and so on. When a molecule is faulty, the entire network may fail, which may affect such important processes. Therefore, characterization of decision makings in cells in response to received signals is important for understanding how cell fate is determined in the absence and presence of such faulty molecules causing incorrect network responses.

Understanding how cells make decisions in response to input signals is an important challenge in molecular and cell biology. Emergence of single-cell data and methods has made it possible to study and model the behavior of each cell individually (Cheong et al., 2011; Habibi et al., 2017; Kolitz & Lauffenburger, 2012). An important factor that affects cell decisions is biological noise in various organisms (Balazsi et al., 2011), which can cause cells to exhibit different behaviors when receiving the same input signal. For example, under the same stimuli, some cells may decide to survive, whereas others may undergo apoptosis. Signaling outcomes can be affected by genetic and epigenetic regulation and misregulation, leading to errors in signaling outcomes and ensuing cell decisions. Given the probabilistic nature of cellular decisions (Cheong et al., 2011; Habibi et al., 2017), it is of interest to have a unified set of statistical metrics and methods to systematically study and characterize the signaling outcomes that may inform them, and determine probabilities associated with different outcomes.

In this dissertation, we first construct a Boolean modeling framework for molecular networks and show how the network models can be trained against data to optimize the discrepancy between the model predictions and data by developing training tools and techniques (Chapter 2). To do so, we provide an integer linear programming formulation that allows us to systematically remove edges and find a subnetwork of the initial network by minimizing the mismatch between the model and data. Next, we fix the network topology and show how the Boolean functions of each molecule can be learned using data. In Chapter 3, we show how the networks can be analyzed by performing vulnerability analysis using the derived mathematical equations. More specifically, we examine the worst possible signaling failures in molecular networks by comparing the maximum vulnerability level, i.e., the highest probability of network failure, versus the number of faulty molecules to understand how the network functionality is affected in the presence of one or more dysfunctional molecules. To do so, an efficient algorithm is developed. Furthermore, another algorithm is proposed that outputs the maximum number of time points needed for computing the vulnerability level of molecules in a Boolean domain. The methods are applied to the experimentally verified ERBB and T cell signaling networks. In Chapters 4 and 5, we show how the statistical decision-theoretic framework proposed by Habibi et al. (2017) can be used to study other molecular systems and signaling outcomes assuming a single decision variable is used. Then, we extend it such that one can model and analyze multidimensional signaling outcome processes using multi-time point measurements for both wild-type and abnormal cells, in Chapter 6. This allows us to incorporate signaling dynamics into decision making analysis. Moreover, we introduce the application of receiver operating characteristic curve as a graphical tool to visualize

decisions and outcomes under normal and abnormal conditions. In Chapter 7, we discuss beyond binary cell decisions such as ternary decision-making processes. To present the concepts, metrics, and algorithms related to decision making and outcome analysis, we use the tumor suppressor p53 system, as an example. Finally, we provide our concluding remarks on the proposed methods in Chapter 8.

1.2 Background Information

The goal of systems biology is construction of models of biological systems from systematic measurements. It focuses on the interactions between biomolecules at the system level. Earlier examples and use of the term "systems biology" started to appear in Ideker et al. (2001) and Kitano (2002). Afterward, thanks to the progress in molecular biology and advances in technology enabling to measure/generate comprehensive gene and protein data, systems biology has become very popular in the last two decades.

1.2.1 Molecular Networks and Network Construction

The availability of molecular data emerged scientists to develop proper methods and modeling techniques to integrate the data into the context of biology. Studying biological networks became a key for understanding complex biological activities. Different types of molecular networks such as protein-protein interaction (PPI) networks, gene regulatory networks (GRNs), and cell signaling networks have been introduced and studied. For instance, Camargo et al. (2007) studied a PPI network to identify protein-binding partners of the DISC1 protein in the human fetal brain, Emmert-Streib et al. (2014) provided a discussion on possible application domains of GRNs, and Eungdamrong and Iyengar

(2004) reviewed theoretical approaches to understand cell signaling networks using heterotrimeric G protein pathways as an example.

The importance of biological network studies led to the establishment of several databases so that one can build a network of interest and develop a theory on it. To illustrate, von Mering et al. (2003) constructed STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database containing known and predicted protein interactions. MINT (Molecular INTeraction) is another database designed to collect experimentally verified protein-protein interaction information offered by Chatr-aryamontri et al. (2007). Another example is TissueNet that is a tissue-specific interaction database containing tissue-specific data of 40 human tissues (Barshir et al., 2013). As reviewed by Miryala et al. (2018), there exist several such databases providing both predicted and experimentally confirmed interaction information allowing to construct a network of interest curated from these databases.

Another way of constructing molecular networks is by inferring the network itself from the data using reverse engineering techniques. Ideker et al. (2000) provided an earlier example of such approaches (discussed in detail below) in which there exist two methods so-called “predictor” and “chooser” that work interactively to infer the genetic network from gene expression measurements. Later, Husmeier (2003) applied Bayesian networks to infer genetic regulatory interactions from microarray gene expression data. Another Bayesian network approach was studied by Sachs, et al. (2005). They used machine learning for the automated derivation of causal interactions in cellular signaling networks, which relies on the simultaneous measurements of phosphorylated protein and phospholipid components in several primary human immune system cells. In 2006, the first

DREAM, Dialogue on Reverse Engineering Assessment and Methods, conference has been organized by Stolovitzky, et al. (2007), which aimed to understand the limitations and to enhance the reverse engineering methods of pathway inference from high-throughput data with the efforts of computational and experimental biologists. Furthermore, DREAM challenges became a non-profit organization examining questions in biology and medicine. In line with their purposes, a network inference challenge has been organized, so-called HPN-DREAM network inference challenge, in which several research groups put their efforts on developing methods and techniques to learn causal influences in signaling networks from phosphoprotein data obtained from cancer cell lines as well as *in silico* data from a nonlinear dynamical model (Hill et al., 2016). New techniques and methods are still being developed by many researchers as there are still many uncertainties in complex systems and new data becomes available day by day.

1.2.2 Molecular Network Modeling

Once the network is constructed, one needs to develop biological models to integrate the available data into the network so that the network reflects the biology, and methods and analysis techniques can be developed to draw biologically relevant conclusions. One way of modeling molecular networks is using continuous models that describe the system's development over time using mass-action kinetics for the rates of consumption and production of molecular species. Modeling with a system of ODEs is a simple way of building continuous network models if sufficient data of kinetic parameters is available. An earlier example of such models was introduced by Goodwin (1963). To illustrate further, Chen et al. (2004) proposed an ODE-based mathematical model of the cell-cycle regulatory network in budding yeasts, successfully explaining the phenotypes of mutants

with a few inconsistencies between model and experiments. Later on, several other ODE-based models have been proposed and reviewed in many articles (e.g., Aldridge et al., 2006; Karlebach & Shamir, 2008; Le Novere, 2015). In ODE-based models, generally, the concentration of the molecules is a function of time, and variation in space is not considered. To take variation in the space under consideration, partial differential equation (PDE) based models have been developed (e.g., De Jong, 2002; Smith et al., 2002).

The continuous models require knowledge of biological mechanisms and kinetic parameters such as rate constants, which is very limited and makes these models limited to well-characterized networks only. Furthermore, the need for prior information increases drastically when the network gets larger and more complex. On the other hand, discrete models such as Boolean and ternary models, Petri nets do not require detailed kinetic information and can still sufficiently model the dynamic behavior of the system. Therefore, discrete models are widely studied, and a lot of works exist in the literature. The appearance of Boolean modeling of molecular networks goes back to the late 1960s. Kauffman (1969) tested the hypothesis that contemporary organisms are randomly constructed molecular automata by modeling the gene as a binary device. Thanks to the availability of several databases and information in the literature, large and complex networks could be studied in the last two decades. Therefore, as opposed to continuous models, discrete modeling became much more popular due to its applicability and efficiency on large networks. For instance, Albert and Othmer (2003) proposed and analyzed a Boolean model of Drosophila melanogaster (a fruit fly) segment polarity gene expression network to test whether the steady states are determined by the topology of the network and the regulatory interactions' type. Saez-Rodriguez et al. (2007) proposed a large-scale Boolean model to analyze the

complex signaling network governing the activation of T cells. Albert et al. (2008) provided an overview of concepts in Boolean network simulations and provided a software library that can perform the simulations. Abdi et al. (2008, 2009) employed the Boolean modeling approach and proposed an error propagation probability method to perform fault diagnosis analysis of molecular networks and identify the most vulnerable molecules, which is further elaborated and applied to different systems in Abdi and Emamian (2010).

In addition to Boolean models, multi-valued logic models have been developed and analyzed. For example, Aldridge et al. (2009) proposed a fuzzy logic framework to analyze cell signaling downstream of TNF, EGF, and insulin receptors in human colon carcinoma cells and discover a relationship between MK2 and ERK pathways. Similarly, Morris et al. (2011) offered a constrained fuzzy logic approach for modeling and training pathway maps on dedicated experimental measurements. Moreover, Habibi et al. (2014a) developed a ternary logic-based fault diagnosis analysis method to identify the most vulnerable molecules in a given network. Aside from the abovementioned logic-based modeling techniques, Petri nets, defining a graphical and mathematical formalism capable of modeling and analysis of discrete event dynamic systems, are another way of modeling the molecular networks (Murata, 1989), that is further elaborated in Chaouiya (2007), presenting the basics of how Petri nets can be used to model complex biological networks. Lastly, there exist hybrid models that incorporate discrete and continuous models such as logic-based ODEs (e.g., Wittmann et al., 2009) in addition to pure continuous or discrete models. To exemplify, Eduati et al. (2020), recently proposed a logic-based ODE modeling framework to generate patient-specific dynamic models of extrinsic and intrinsic apoptosis signaling pathways. The modeling and analysis techniques are not limited to the examined

resources above. For other examples of molecular network modeling techniques, we refer to the following research and review articles: Albert (2007), Chaouiya et al., (2012), Chaouiya and Remy (2013), Handorf and Klipp (2012), Hecker et al. (2009), Machado et al. (2011), Morris et al. (2010), Saadatpour and Albert (2012, 2013), Samaga and Klamt (2013), Schlitt and Brazma (2007), Stoll et al. (2012), Wang et al. (2012). Wilkinson (2006), Wynn et al. (2012).

1.2.3 Training Molecular Network Models

After constructing the network and network models, an emergent issue is that the generated models' predictions may not agree with the experimental findings which might be due to the incompleteness of resources, databases, and literature used to construct the networks as well as the model itself. Therefore, many training methods have been proposed to improve the fitness of the models to the experimental data or to learn a new model from the data in recent years. An earlier example of learning (inferring) networks from experimental data was revealed by Ideker et al. (2000). They provided two methods so-called "predictor" and "chooser" that work interactively to infer genetic network from gene expression measurements. The predictor method determines the set of Boolean networks consistent with the data, while the chooser method uses an entropy-based approach to propose an additional perturbation experiment to reduce the number of Boolean networks found by the predictor. The proposed approach is very useful if there exist data for all of the molecules/genes in the network or if we only want to learn the network of the molecules in the dataset. However, in most cases, the experimental data is very limited especially for large networks with hundreds of components due to lack of technology to observe the data, lack of information in the literature, the complexity of the experiments, and so on, which

makes this approach not applicable for all networks. Another example of learning network models can be found in Videla et al. (2012) in which training of logic models using high-throughput phospho-proteomics data is reduced to a combinatorial optimization problem that is solved using Answer Set Programming (ASP) approach. ASP is a declarative problem-solving paradigm in which a problem is encoded as a logical program such that its answer sets represent solutions to the problem. On the other hand, Sharan and Karp (2013) proposed an algorithm that reduces this problem into an Integer Linear Programming (ILP) problem. Their algorithm does not require information on the interaction (the edge signs) as well as an initial model to start.

There exist other studies, in which the model training starts with an initial network (curated from literature), systematically improve the model prediction accuracy by adding/removing interactions. For instance, Saez-Rodriguez et al. (2009) developed the "CellNetOptimizer" (CNO) algorithm that starts with an initial model and trains it by using heuristic genetic algorithm against the experimental measurements to learn a more compact representation of the model that fits the data well. Moreover, Mitsos et al. (2009) presented an ILP formulation of the problem that allows learning of a subnetwork of the initial network that will provide an optimal fit to the observed data. Similarly, Melas et al. (2013) presented an ILP-based algorithm in which they perform four operations to detect and remove inconsistencies between experimental measurements and predicted behavior that are (i) finding/constructing an initial network relevant to the nodes measured in experiments; (ii) determining a set of nodes that have inconsistencies with the measurements and need to be corrected; (iii) determining the optimal subnetwork of the initial network which has the best fitness to the measurements; (iv) finding possible edges

to be added to the optimal subnetwork to further improve the accuracy with respect to the experimental measurements.

1.2.4 Molecular Network Analysis

The main purpose of collecting the experimental data, constructing the underlying network and network models, and training the models against experimental data to obtain better models with high fitness percentage is to eventually analyze the network and discover novel insights into complex biological systems. One of the purposes of this dissertation is to provide a mathematical framework to perform fault diagnosis analysis of molecular networks, which has many applications such as target discovery and drug developments (Csermely et al., 2013). A version of the vulnerability analysis technique we employ in this dissertation was initially proposed by Abdi et al. (2008), in which they proposed an error propagation probability method to perform fault diagnosis analysis of molecular networks. The method enumerates the probability of network failure at the output of the network in the presence of a single is faulty (dysfunctional) molecule at a time. They applied the method to caspase 3 network, p53 network, and CREB network to show the utility of the approach. Furthermore, they showed that this method is capable of reproducing known results as well as discovering novel vulnerable molecules as experimentally confirmed in the CREB network. Later on, Habibi et al. (2014a) expanded this method in different ways. First, they performed single fault vulnerability analysis by applying different levels of input combinations. Then, they examined the effect of different fault probability levels for each molecule on their vulnerability values. Lastly, they extended the fault diagnosis technique by considering three activity levels of molecules in addition to the Boolean model used to introduce their methods.

1.2.5 Single Cell Decision Making and Signaling Outcome Analysis

The molecular networks have some specific outputs that initiate important biochemical processes. For instance, depending on the received signals and dynamics of the networks, cells may make different decisions such as proliferating or initiating apoptosis. These decisions leading to well-defined macroscopic patterns, tissues, and organs are driven by chemical and mechanical signals, and are organized in space and time (Arias & Stewart, 2002). Balázsi et al. (2011) reviewed examples of cellular decision-making in several organisms such as viruses, yeast, bacteria, and mammals, and showed the role of molecular noise in cell decision-making. Due to the biological noise, identical cells may exhibit different behaviors, when they receive the same input signal.

For several decades, balls rolling down a slanted landscape with bifurcating valleys of Waddington and Kacser (1957), also known as "Waddington's epigenetic landscape", have been widely used to illustrate the differentiation of multicellular development, although it was unclear what the valleys and peaks are. Nowadays, thanks to single-cell measurement technologies (Svensson et al., 2017; Ziegenhain et al., 2017) which can simultaneously measure the expression of many genes in several single cells, it became possible to compute Waddington's landscape that can serve as a theoretical framework for cellular decision-making. The availability of single-cell data made it possible to understand developmental pathways and cell fate decisions and paved the way to examine complex disorders. Several computational and experimental studies have been conducted. For instance, Narula et al. (2016), investigated sporulating Bacillus subtilis (bacteria) single-cell data and concluded that cells sensing growth rates indirectly detect starvation without the need for evaluating specific stress signals. Moreover, Hat et al. (2016) investigated how

the complex p53 system is involved in cell fate decision-making. They constructed a mathematical model of the p53 network to understand the dynamics of cancerous and cancer-free cells. Similarly, Tudelska et al. (2017) combined experimentation with mathematical modeling to understand how TNF concentration reflects the binary decision cell makes that is the translocation of NF-κB. In addition, Habibi et al. (2017) provided a mathematical framework to quantify cell decision error probabilities assuming a univariate decision-making scheme. Recently, we extended this framework to a multivariate decision-making scheme to incorporate signaling dynamics into decision-making analysis (Ozen et al., 2020) as elaborated in Chapters 4, 5, 6, and 7 of this dissertation. Several other examples of single-cell studies exist (e.g., Garcia-Ojalvo & Martinez Arias, 2012; Griffiths et al., 2018; Guo et al., 2019; Mohammed et al., 2017; Moris et al., 2016; Sagar & Grün, 2020; Zernicka-Goetz et al., 2009; Zhang et al., 2019). Such single-cell analysis methods are being studied to better understand the transition from physiological to pathological conditions such as inflammation, various cancers, and autoimmune diseases.

CHAPTER 2

MODELING MOLECULAR NETWORKS AND TRAINING NETWORK MODELS

Analyzing molecular networks is essential to understand complex biological dynamics and shed light on complex diseases. Therefore, molecular network graphs need to be converted into numerable models so that one can perform different analyses and eventually observe biologically valuable results. One way of modeling molecular networks is building a system of differential equations that provide detailed information on network dynamics. Such models are in need of the knowledge of mechanistic details and kinetic parameters, which is very limited in general. Thus, they are not practical for large molecular networks with several components. On the other hand, discrete models such as Boolean models do not require detailed kinetic information and still provide relevant biological insights into complex systems. Furthermore, they are easier to understand and computationally simpler than continuous models. In this chapter, we provide examples of continuous and Boolean modeling frameworks.

After building a biologically relevant network model, a typical issue is that the model predictions generally do not reflect the experimental data for literature-curated networks, which might be due to the incompleteness of resources used to construct the network or the model itself. Since any observation made on a model with large discrepancy between biological evidence cannot be trusted, the network models need to be trained against data before doing further analyses. In this chapter, we provide two training approaches that are training by removing edges and training by learning Boolean functions of the molecules, explained in detail in the subsequent sections.

2.1 Continuous Modeling of Molecular Networks

The continuous models incorporate the continuous time and space-varying behavior of the molecular components of the network and hence provide a detailed representation of the underlying biological mechanism. These models convert molecular networks into a mathematical form by building a system of differential equations. More specifically, a system of ODEs is a widely used representation of biological networks whose general form is given below (Sontag, 1998, 2005):

$$\begin{aligned}
\frac{\mathrm{d}x_1(t)}{\mathrm{d}t} &= f_1\big(x_1(t), x_2(t), \dots, x_n(t), I_1(t), I_2(t), \dots, I_m(t)\big) \\
\frac{\mathrm{d}x_2(t)}{\mathrm{d}t} &= f_2\big(x_1(t), x_2(t), \dots, x_n(t), I_1(t), I_2(t), \dots, I_m(t)\big) \\
&\vdots \\
\frac{\mathrm{d}x_n(t)}{\mathrm{d}t} &= f_n\big(x_1(t), x_2(t), \dots, x_n(t), I_1(t), I_2(t), \dots, I_m(t)\big)
\end{aligned}$$

where $\boldsymbol{x}(t) = \big(x_1(t), x_2(t), \dots, x_n(t)\big)$ are the concentration of n molecular components at time t, $\boldsymbol{I}(t) = \big(I_1(t), I_2(t), \dots, I_m(t)\big)$ are m external inputs to the cellular system, and $f_1, f_2, \dots, f_n$ are the functions of $n + m$ variables indicating the relationship between each components.

Molecular networks can also be modeled by the well-known chemical master equation which describes the time evolution of the probability of a system jumping from one state to another in a continuous time (Kampen, 1981). It is a stochastic approach based on the law of mass action, which states that the rate of a chemical reaction is proportional to the product of the active masses of the reacting substances (Erdi & Toth, 1989).

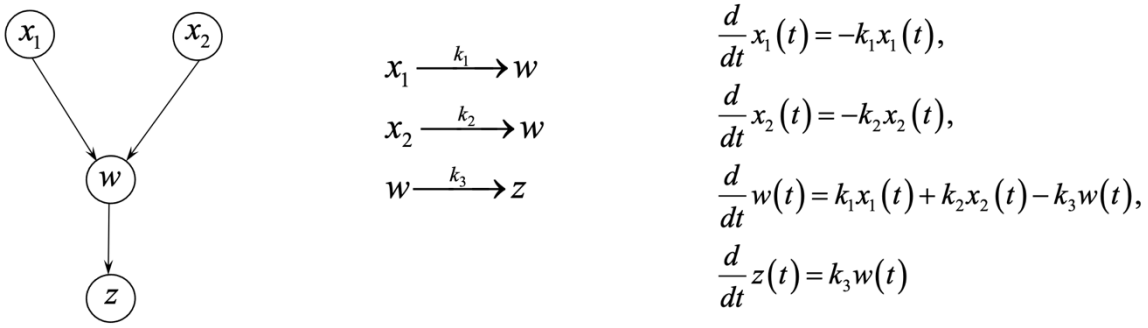


Figure 2.1 An example ODE-based continuous model of a toy network with four molecules.

Figure 2.1 is an example system of ODEs for the small network with four molecules, i.e., X_1, X_2, W, Z, and three interactions with activation rates k_1, k_2, k_3. As seen in this simple example, continuous models are costly by virtue of numerous free parameters such as rate constants that have to be estimated from limited amount of data, especially in large networks. When the size of the network increases to hundreds of molecules and interactions, the system contains more and more free parameters to be estimated which is very challenging in the absence of prior knowledge and data. Therefore, discrete modeling, e.g., Boolean models, becomes a useful tool to model and analyze large biological systems, which is explained in detail in the next section.

2.2 Boolean Modeling of Molecular Networks

Boolean models are one of the simplest yet very useful way of modeling molecular networks and capturing their dynamics. They assume two activity states of the molecules, i.e., ON (active) and OFF (inactive). The advantage of this type of modeling is that they do not need detailed kinetic information and still provide certain biologically relevant

predictions. In general, Boolean network models consist of the following three components (DasGupta & Liang, 2016):

- A Boolean state vector $\vec{s} = (s_1, s_2, \dots, s_n) \in \{0,1\}^n$.
- A global activation function $\vec{f} = (f_1, f_2, \dots, f_n)$, in which f_i is a Boolean function, i.e., $f_i: \{0,1\}^n \rightarrow \{0,1\}$.
- An update rule that incorporates the dynamic behavior of the network.

Boolean models allow representing molecular networks as digital circuits using digital gates such as AND, OR, and NOT. Thus, many analysis techniques developed for digital circuits such as reliability analysis (Han et al., 2011) and fault detection techniques (Kohavi & Kohavi, 1972) can be applied for molecular networks with Boolean models as well. In the following subsections, we provide examples of Boolean models for molecular networks.

2.2.1 Model 1: Increase in Activity "1", Decrease/No Change in Activity "0"

In a typical biological reaction, the activity level of the output (product of the reaction) can increase, decrease, or remain between some thresholds, depending on its upstream molecules. In this model, when a stimulus is applied to the network, increase in the activity level of a molecule is represented by binary 1 and decrease or no change in the activity level of a molecule is represented by binary 0. Assume there exists a reaction with multiple activators and inhibitors. Then, this model incorporates two update rules to specify the output molecule's state:

- Rule 1: The output is 1 if none of the inhibitors is 1 and at least one of the activators is 1.
- Rule 2: The output is 0 if at least one of the inhibitors is 1.

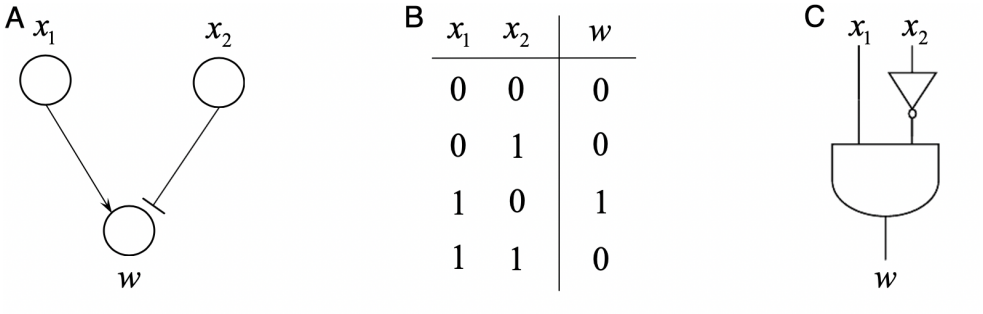


Figure 2.2 An example for Boolean Model 1. (A) A two-input one-output network. (B) Truth table of the network based on the model rules. (C) Logic circuit representation of the network.

Figure 2.2 exemplifies this model for a reaction in which x_1 and x_2 are activatory and inhibitory inputs of w (Figure 2.2A), respectively. Using these model rules, one can fill the associated truth table in Figure 2.2B which further helps to create the logic circuit representation of the network in Figure 2.2C. Furthermore, the rules lead to generate Boolean equation of the output, that is $w = x_1 \times (\sim x_2)$, where "×" represents the AND operation while "∼" represents the NOT operation. Using this approach, a network of hundreds of molecular components can be implemented as a digital circuit with many AND (×), OR (+), and NOT (∼) gates.

2.2.2 Model 2: Change in Activity "1", No Change in Activity "0"

In this model, when an input combination is applied, any change in molecule activity (activity increase (decrease) above (below) a certain threshold) is represented by binary 1 while no change in the activity (remaining between the thresholds) is represented by binary 0. For a given reaction with multiple input molecules, the two update rules specifying the output molecule's state of this model are:

- Rule 1: The output is 1 if at least one of the inputs is 1.
- Rule 2: The output is 0 if all inputs are 0.

Figure 2.3 presents application of this model on the small network used in Model 1 above. Using the rules, one can represent the network using an OR gate only (Figure 2.3C) and generate the Boolean equation $w = x_1 + x_2$.

There exist several different Boolean models studied in the literature. The model rules and their representations depend on the biological system being studied. Therefore, to incorporate biological properties of the network under interest, one can implement different model rules than the ones provided here, which may lead to a different logic circuit and equation representation of the molecules.

2.3 Training Boolean Network Models

After building biologically acceptable models for molecular networks, a typical issue is that the generated models' predictions do not agree with the experimental data for draft literature-based networks, which might be due to the incompleteness of resources, databases, and literature used to construct the networks. Therefore, the network models

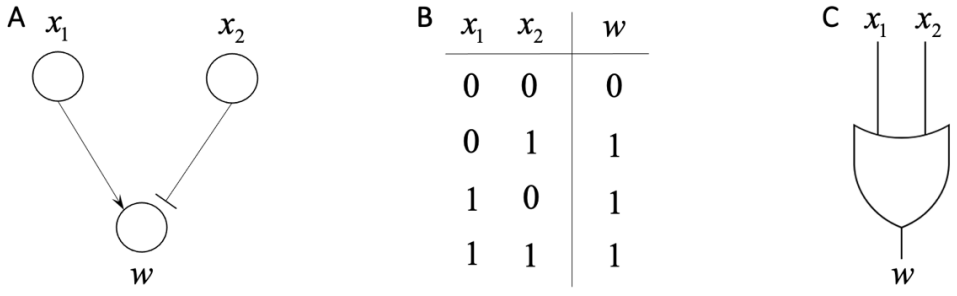


Figure 2.3 An example for Boolean Model 2. (A) A two-input one-output network. (B) Truth table of the network based on the model rules. (C) Logic circuit representation of the network.

need to be trained against data before doing further analyses. Many methods have been proposed to improve the fitness of the models to the experimental data or to learn a new model from the data in recent years as summarized in Chapter 1. In this section, we provide two methods to train the Boolean network models using data, that are elaborated in the following subsections:

2.3.1 Training by Edge Removing via Integer Linear Programming

One way of training network models is fixing the model rules and manipulating the network topology by adding or removing interactions between the existing molecules in the network. Due to the incompleteness of the resources used to construct the network, some of the edges (interactions) in the network may need to be removed (spurious interactions) or some new edges may need to be added so that the resulting network can reflect natural collective behaviors of the molecules, i.e., models that fit the experimental data. Herein, we prefer to remove edges and find a subnetwork of the initial network since adding new edges requires experimental evidence which is costly and time consuming to acquire. One simple way to do this is to conduct an optimization to remove edges one by one and check the number of mismatches between model predictions and experimental data. However, for large networks, this is computationally very complex and does not help as removing one edge at a time most often does not change model predictions. For this reason, we convert this problem into an integer linear programming (ILP) problem in which multiple edges can be removed systematically, i.e., the optimal solution to the ILP problem is a subnetwork of the initial network that fits the data. A similar approach was studied by other groups (e.g., Mitsos et al., 2009) on a network that does not include feedbacks. In

this dissertation, we provide a new formulation for a different Boolean model and show how we apply it when the network contains some feedback interactions.

The goal is to minimize the mismatch between model's responses and experimental data. The data is generally obtained by applying an input to the network and measuring the abundance of some of the intermediate as well as the output molecules in each experiment. Then, these experiments are repeated multiple times for different input perturbations.

Let n_E be the number of experiments and each experiment be indexed by the superscript $k = 1, \dots, n_E$. In the network, there exist n_R reactions which are indexed by the subscript $i = 1, \dots, n_R$. Each reaction i has the corresponding index set $I_i = A_i \cup H_i$ for its input molecules, in which A_i and H_i are the index set of activators and inhibitors, respectively. Lastly, let M be the index set of molecules for which we have experimental data. Then, in the general form of the proposed ILP formulation, we define all the other variables as shown below. For each reaction i, we have:

- x_j^k: model's predicted value of the j^{th} input node in the k^{th} experiment, for all $j \in I_i$.
- $x_j^{k,m}$: experimental value of the j^{th} node in the k^{th} experiment, for all $j \in M$.
- y_j: decision variable, for all $j \in I_i$. $y_j = 1$ means that j^{th} edge in the reaction i should be preserved in the network whereas $y_j = 0$ means that j^{th} edge in the reaction i should be removed from the network.
- z_j^k: transition variable, for all $j \in I_i$. It transits the input value x_j^k associated with the j^{th} edge to the output of reaction i if $y_j = 1$. Otherwise, $z_j^k = 0$.
- w_i^k: output value of the reaction i.

The objective function to be minimized is the summation of the number of mismatches between the experimental data and the model's prediction over all experiments (the absolute error). Thereby, the objective function is:

$$\sum_{j,k} \left| x_j^k - x_j^{k,m} \right|, \quad \forall j \in M, k = 1, \dots, n_E, \tag{2.1}$$

where x_j^k is the models' prediction and $x_j^{k,m}$ is the experimental value of j^{th} molecule in the experiment k. For Boolean x_j^k and $x_j^{k,m}$ values, Equation (2.1) can be linearized as:

$$\sum_{j,k} x_j^{k,m} + \left(1 - 2x_j^{k,m}\right) x_j^k, \quad \forall j \in M, k = 1, \dots, n_E.$$

For training the Boolean Model 1 introduced in Section 2.2.1, using all definitions given above, the constrained ILP formulation can be written as shown in Equation (2.2), in which the constraints (i), (ii), and (iii) are introduced for edge removal. More precisely, these three constraints assure that if the j^{th} interaction in reaction i is removed, i.e., $y_j = 0$, then the transition variable $z_j^k = 0$ so that the input molecule associated with the j^{th} interaction does not affect the value of the output molecule w_i^k. If the j^{th} interaction needs to stay, i.e., $y_j = 1$, then these constraints guarantees that the transition variable $z_j^k = x_j^k$. The constraints (iv), (v), and (vi) implement the rules of Boolean Model 1. To elaborate, depending on the constraints (i), (ii), and (iii), the transition variable z_j^k will be equal to either 0 or x_j^k. Then, if none of the inhibitors and at least one of the activators is 1, the constraints (iv) and (v) guarantee that the output $w_i^k = 1$ (Rule 1). Similarly, if at least one

$$\min_{y} \sum_{j,k} x_j^{k,m} + \left(1 - 2x_j^{k,m}\right) x_j^k, \quad \forall j \in M, k = 1, \dots, n_E$$

Subject to $\forall i = 1, \dots, n_R, \forall k = 1, \dots, n_E$

(i) $z_j^k \geq y_j + x_j^k - 1, \ \forall j \in I_i$

(ii) $z_j^k \leq x_j^k, \ \forall j \in I_i$

(iii) $z_j^k \leq y_j, \ \forall j \in I_i$

(iv) $w_i^k \leq 1 - \sum_{j \in H_i} \frac{z_j^k}{|H_i| + 1},$ (2.2)

(v) $w_i^k \geq \sum_{j \in A_i} \frac{z_j^k}{|A_i| + 1} - \sum_{j \in H_i} z_j^k,$

(vi) $w_i^k \leq \sum_{j \in I_i} z_j^k,$

(vii) $0 \leq x_j^k, z_j^k, y_j, w_i^k \leq 1, \ \ x_j^k, z_j^k, y_j, w_i^k \in \mathbb{Z}.$

of the inhibitors is 1, i.e., $\exists j \in H_i$ such that $z_j^k = 1$, then, the constraints (iv) and (v) make sure that the output $w_i^k = 0$ (Rule 2). The constraint (vi) is necessary to guarantee that the output $w_i^k = 0$ if all incoming edges are removed or the input values of the remaining edges are 0s. Lastly, the constraint (vii) is needed to guarantee that all variables are integers and they are either 0 or 1. Figure 2.4 further explains how the formulation works on an example.

A similar formulation can be adapted for training the Boolean Model 2 introduced in Section 2.2.2. This is done by removing the constraint (iv) and replacing constraint (v) of Equation (2.2) by $w_i^k \geq z_j^k, \forall j \in I_i$, which becomes to Equation (2.3) given below.

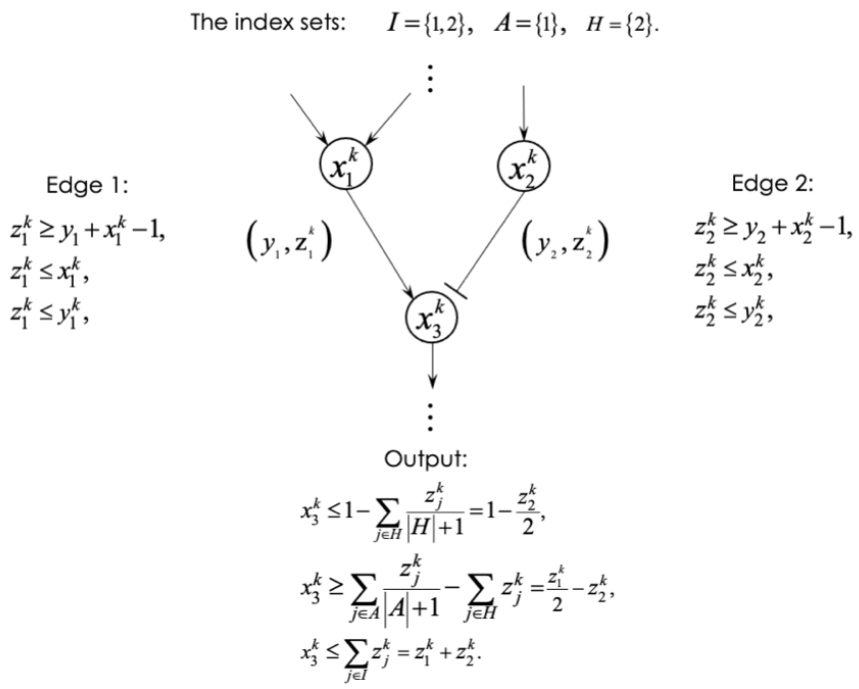


Figure 2.4 An example of the ILP formulation for Boolean Model 1.
Note: This is a hypothetical reaction in which the inputs are x_1 and x_2, and the output is x_3.

$$\min_{y} \sum_{j,k} x_j^{k,m} + \left(1 - 2x_j^{k,m}\right) x_j^k, \quad \forall j \in M, k = 1, \dots, n_E$$

Subject to $\forall i = 1, \dots, n_R, \forall k = 1, \dots, n_E$

(i) $z_j^k \geq y_j + x_j^k - 1, \ \forall j \in I_i$

(ii) $z_j^k \leq x_j^k, \ \forall j \in I_i$

(iii) $z_j^k \leq y_j, \ \forall j \in I_i$ (2.3)

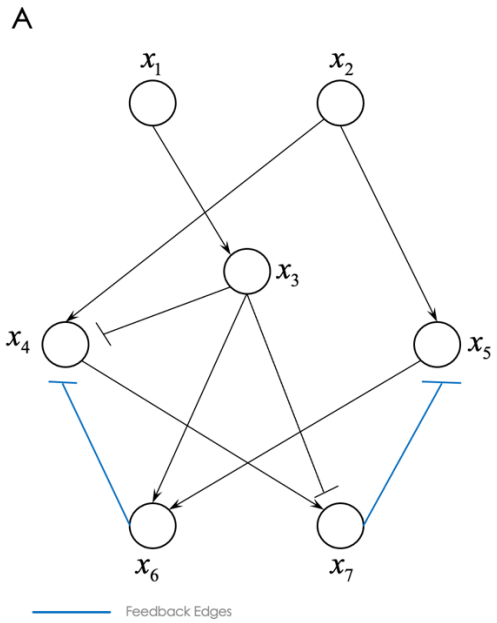
(iv) $w_i^k \geq z_j^k, \ \forall j \in I_i$

(v) $w_i^k \leq \sum_{j\in I_i} z_j^k,$

(vi) $0 \leq x_j^k, z_j^k, y_j, w_i^k \leq 1, \ \ x_j^k, z_j^k, y_j, w_i^k \in \mathbb{Z}.$

The ILP formulations (2.2) and (2.3) search for a vector $\boldsymbol{y} = [y_j]$, that is the vector of indices of edges in the network, minimizing the number of mismatches between predictions and the data. To elaborate, a network can be represented by a vector $\boldsymbol{y}$ that is vector of 1s with the length of the total number of interactions in the network. Thus, a subnetwork of the initial network can be represented by the same $\boldsymbol{y}$ vector with entries 0s and 1s. If the j^{th} entry of $\boldsymbol{y}$ is 0, then this means that the j^{th} interaction is removed in the subnetwork. As a result, by solving the ILP formulations (2.2) and (2.3), one can find the best $\boldsymbol{y}$ vector(s), i.e., the subnetwork(s), that has the optimal fit to the data for the given Boolean models 1 and 2.

Now we apply the ILP formulation in (2.2) to a toy network with hypothetical experimental data. Suppose that we are given the toy network in Figure 2.5A. Then, the



B

Inputs: x_1, x_2.

Intermediate Molecules:

$$x_3 = x_1,$$
$$x_4 = x_2 \times \sim (x_3 + x_6),$$
$$x_5 = x_2 \times (\sim x_7).$$

Outputs:

$$x_6 = x_3 + x_5,$$
$$x_7 = x_4 \times (\sim x_3).$$

Figure 2.5 A toy network and the associated Boolean equations based on Model 1. (A) The toy network with feedback. (B) The associated Boolean equations of each node created based on Boolean Model 1's rules.

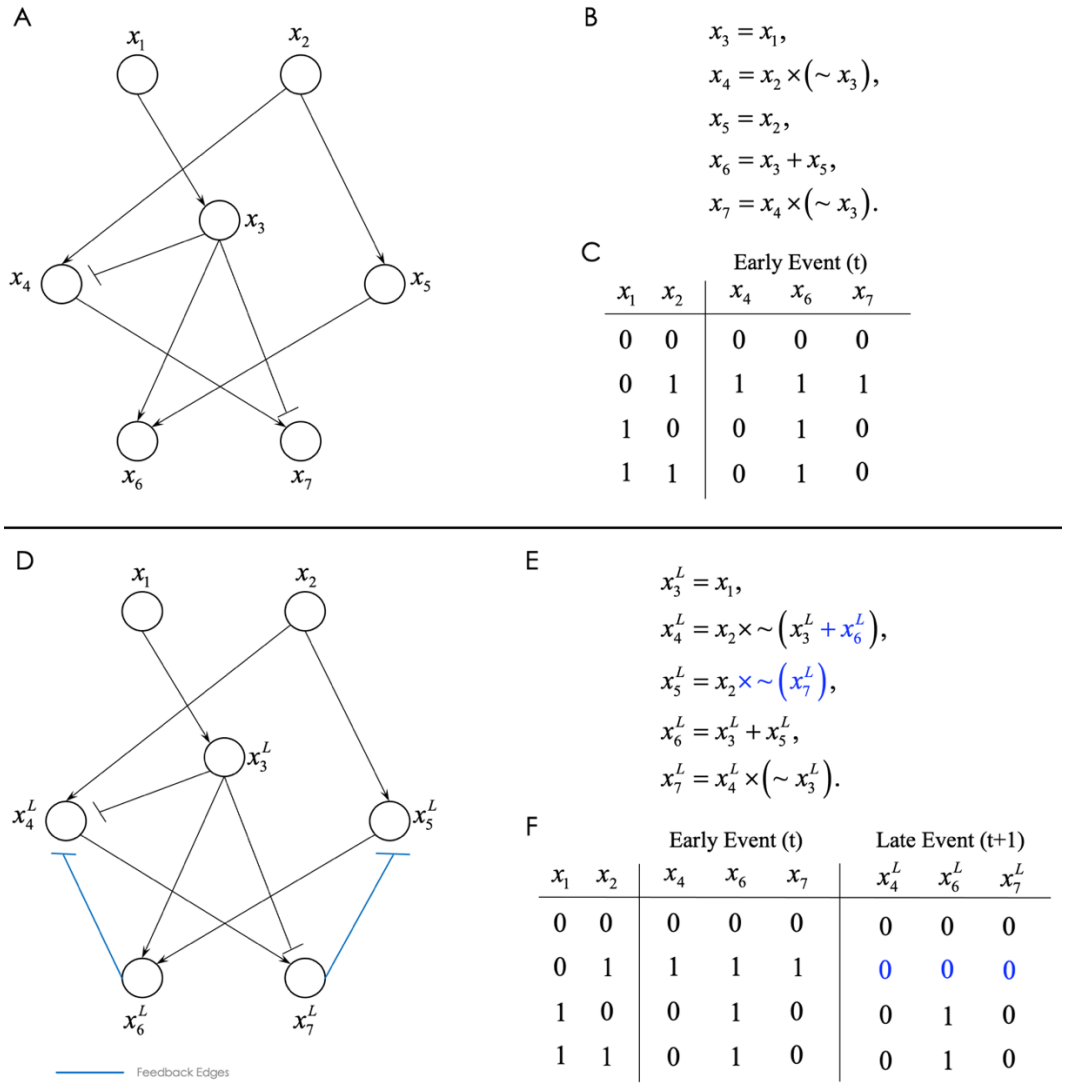


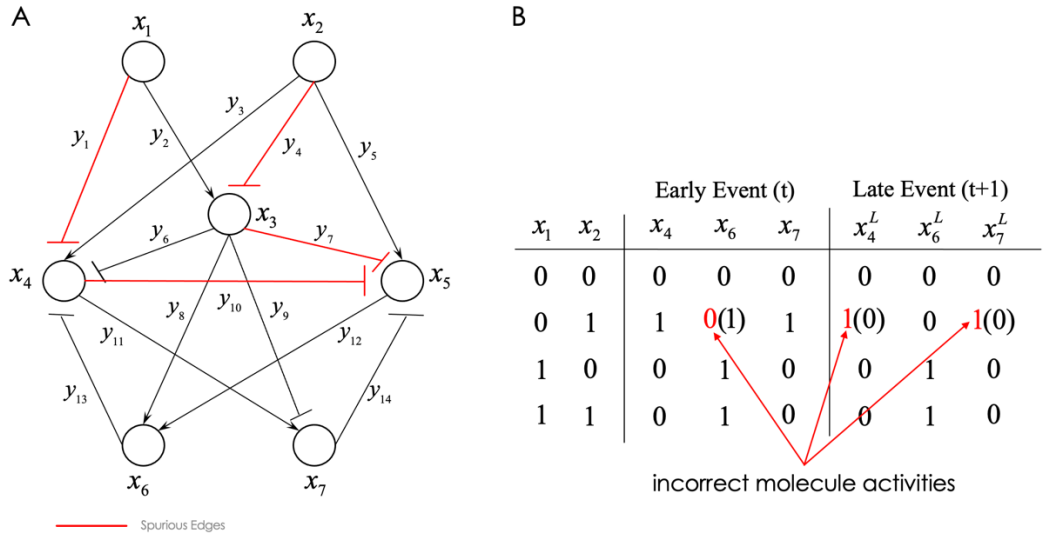
Figure 2.6 The early event and the late event components of the toy network. (A) The EE network. (B) The Boolean equations of the nodes in the EE network. (C) The truth table of the EE network. (D) The LE network. (E) The Boolean equations of the nodes in the LE network. (F) The truth table of the LE network.

associated Boolean equations (based on Model 1's rules) for each node can be written as shown in Figure 2.5B. Because of the presence of feedback interactions (the blue edges in Figure 2.5A), this network can be represented in two pieces: the early event (EE) network and the late event (LE) network as shown in Figure 2.6A and Figure 2.6D. When there is

feedback(s) in the network, the network response might be different at different time instances because of the dynamics and latency in the reactions caused by the feedbacks. Using the Boolean equations in Figures 2.6B and 2.6E, the EE and LE truth tables can be created as shown in Figures 2.6C and 2.6F.

Assume that the network in Figure 2.6D is the ground truth network and Figure 2.6F is the hypothetical experimental data measured in experiments and quantized properly. Then, to test the ILP formulation (2.2), we manipulate this network by adding new spurious interactions and generate a new network with a new truth table as shown in Figure 2.7.

Suppose that the new network with the spurious interactions is the initial network constructed from literature, and our initial model has some mismatches (red values in Figure 2.7B) compared to the experimental data in Figure 2.6F. Our purpose is to train this



x_1	x_2	x_4	x_6	x_7	x_4^L	x_6^L	x_7^L
0	0	0	0	0	0	0	0
0	1	1	0(1)	1	1(0)	0	1(0)
1	0	0	1	0	0	1	0
1	1	0	1	0	0	1	0

Figure 2.7 The extended toy network with spurious edges. (A) The extended toy network that hypothetically represents the literature-curated network. (B) The truth table of the untrained network model's predictions. The red entries of the table are the mismatches compared to the hypothetical data.

network by building and solving the ILP formulation in Equation (2.2). In other words, the purpose is to find a subnetwork of the initial network that has the optimal fit to the data. The expectation after solving the ILP is that the desired network in Figure 2.6D, which can be represented by the vector $\boldsymbol{y} = [0\ 1\ 1\ 0\ 1\ 1\ 0\ 1\ 1\ 0\ 1\ 1\ 1\ 1]$ is among the found solutions.

In the training, care should be taken while considering the feedbacks. Since the network may behave differently at different time instances (which is the case as seen in Figure 2.6F) implementing the constraints in Equation (2.2) is not trivial for the nodes having incoming feedback inputs. In fact, it is very challenging to mathematically formulate such nodes in one step because the feedback nodes need to be initialized and then updated when the LE data is considered. To solve this issue, we simply duplicate the EE network and connect these two identical networks using the feedback edges. Furthermore, we treat the nodes in both copies as a new node as shown in Figure 2.8. For instance, x_4 and x_4^L represent EE and LE values, respectively, where x_4^L has the feedback input initiated from x_6. Note that the molecule equations are the same if a node does not have any feedback input (e.g., $x_3 = x_3^L$). Moreover, the edges in the identical copies are labeled by the same decision variable y_i so that if $y_i = 0$, then both edges are removed from the network. For instance, the edges x_1 —| x_4 and x_1 —| x_4^L are labeled by y_1 (Figure 2.8) so that if $y_1 = 0$, then both edges are removed.

The ILP formulations are implemented using OPL (Optimization Programming Language), a high-level programming language, and are solved using the IBM ILOG CPLEX optimization studio (IBM, n.d.), a commercial software that solves optimization problems. CPLEX found twelve optimal solutions, i.e., twelve $\boldsymbol{y}$ vectors, with the objective value of 0, and $\boldsymbol{y} = [0\ 1\ 1\ 0\ 1\ 1\ 0\ 1\ 1\ 0\ 1\ 1\ 1\ 1]$ is among one of them, which confirms the

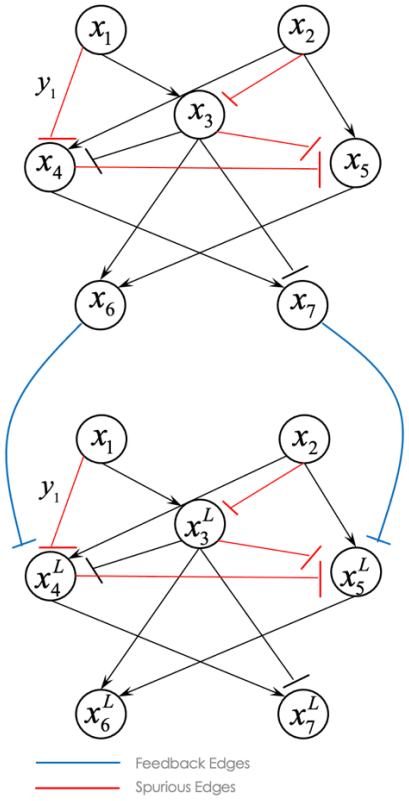


Figure 2.8 The duplicated network to handle the feedbacks while training via ILP.

ability of the proposed formulation by finding the desired subnetwork with the best fitness to the data while preserving the initial model rules.

The proposed training approach via ILP formulation and the strategy to handle feedbacks is applicable to very large networks and capable of finding the exact optimal subnetworks with the best fitness percentage to the data. However, a challenge of this approach is that the solutions to the ILP formulations may not be unique and multiple solutions might be obtained. Even so, it has been observed that the results are usually very correlated, indicating that the solutions are very similar. Moreover, the resulting subnetworks might be missing a lot of interactions that exist in the initial network. This

issue can be solved by adding a penalty term to the objective function in Equation (2.1) such as:

$$\sum_{j,k} x_j^{k,m} + \left(1 - 2x_j^{k,m}\right)x_j^k + \beta \sum_{i=1}^{|y|} (1 - y_i),$$

where β is a tunable penalty parameter that penalize the objective function for each removed edge. The higher values of β may result in worse fitness to the data while keeping more edges in the subnetwork. Therefore, there might be a tradeoff between the number of removed edges and the fitness percentage. Finally, some of the removed interactions to get the optimal solution might be well-known interactions that are experimentally reproduced by several groups, which is an undesired outcome. Therefore, searching for a subnetwork of the initial network may not be always the best training approach, although it usually performs well in terms of data fitness. Thus, in addition to training via ILP, we propose another training approach in which the network topology is fixed, and the Boolean functions are learned from data, explained in detail in the next subsection.

2.3.2 Learning Boolean Functions of Molecules

The training via ILP may result in losing so many interactions while optimizing the model predictions. However, one may want to keep all interactions in the network and analyze the network as it is without losing any interactions. In such scenarios, the mismatch between the model's predictions and the experimental data can be minimized by tuning the model against data. In other words, the model for each molecule can be inferred from the

data. For this purpose, we propose a method that learns Boolean equations of each molecule while minimizing the mismatch between the model and data in this subsection.

Recall the Boolean Model 1 in Section 2.2.1, whose Rules 1 and 2 can be simply represented by the following piecewise function:

$$\text{Output} = \begin{cases} 0, & \text{OR(inhibitors)} = 1, \\ \text{OR(activators)}, & \text{Otherwise.} \end{cases}$$

The fixed model above may not perform well in terms of model prediction when tested against the experimental data if the network topology is fixed. Therefore, instead of using a fixed AND and OR functions for each rule, we propose to learn f and h functions that are called activatory and inhibitory functions, respectively, for each molecule by minimizing the mismatch between model predictions and the experimental data. In other words, we learn new rules for each molecule by implicitly assuming that multiple inhibitors and multiple activators may need to work together to change the output molecule's state. A general representation of this approach is given in Equation (2.4):

$$\text{Output} = \begin{cases} 0, & h(\text{inhibitors}) = 1, \\ f(\text{activators}), & \text{Otherwise.} \end{cases} \tag{2.4}$$

where h(inhibitors) and f(activators) can contain combinations of AND and OR operators.

The learning is done by converting the problem into an optimization problem. In the learning, the purpose is finding the best gate (AND/OR) combinations for f and h functions of each molecules so that the network with the new models has the best fit to data. Similar to the ILP approach, a network can be represented by a binary vector

$\boldsymbol{g} = [g_i]$, i.e., $g_i \in \{0,1\}$, where $g_i = 0$ means that the i^{th} gate in the network is an AND gate whereas $g_i = 1$ means that the i^{th} gate in the network is an OR gate. Note that, for the initial model, $\boldsymbol{g}$ is a vector of 1s meaning that all Boolean operators of f and h functions are ORs. Consequently, the problem of finding the best gate combinations for each molecule is scaled to finding a vector $\boldsymbol{g}$ of 0s and 1s, that contains all gates in the network, so that a network constructed using $\boldsymbol{g}$ has the best fit percentage to the data. Equation (2.5) below is the general formulation of the optimization problem, in which E is the experimental dataset, $N(\boldsymbol{g})$ is the network constructed using the vector $\boldsymbol{g}$**.** The objective function basically counts the number of mismatches between the network constructed by $\boldsymbol{g}$**,** i.e., $N(\boldsymbol{g})$, and the experimental data, i.e., E (the absolute error). The biological constrains can make sure that activatory function f cannot consist of only the AND gates if it is assumed that there is no need for all activators of a molecule to be active at a time, or a similar constraint can be constructed for the inhibitory function h, etc. This can be done by constructing inequality constraints. For instance, suppose a molecule M has three activatory inputs I_1, I_2, and I_3. Thus, the activatory function of M, i.e., f, will include two gates, i.e., g_1 and g_2. If biologically it is not necessary or it is wrong to have all three inputs active at the same time to activate the molecule M, then g_1 and g_2 cannot be 0 at the

$$\min_{\boldsymbol{g}} \sum_{j=1}^{|E|} 1\left(x_j^{N(\boldsymbol{g})} \neq x_j^E\right)$$

Subject to

Biological Constraints, (2.5)

$g_i \in \{0,1\}\ \forall i$

same time, which can be ensured by the constraint $g_1 + g_2 \geq 1$ forcing at least one of the gates to be an OR gate. Similar constraints can be implemented to incorporate biological restrictions and processes.

Equation (2.5) is a nonconvex global optimization problem that is NP-hard. Therefore, we solve (2.5) by using Genetic Algorithm (GA), that is a well-known metaheuristic algorithm which attempts to find a global minimum or at least its good approximation (Mitchell, 1998).

Figure 2.9 further explains how this approach works on a simple toy network. Assume we have a very simple network with only activators with the untrained fixed Boolean Model 1 rules represented by the function string $\boldsymbol{g} = [1\ 1\ 1\ 1\ 1\ 1]$ (Figure 2.9A). Then, one possible learned network could be the one in Figure 2.9B in which two AND

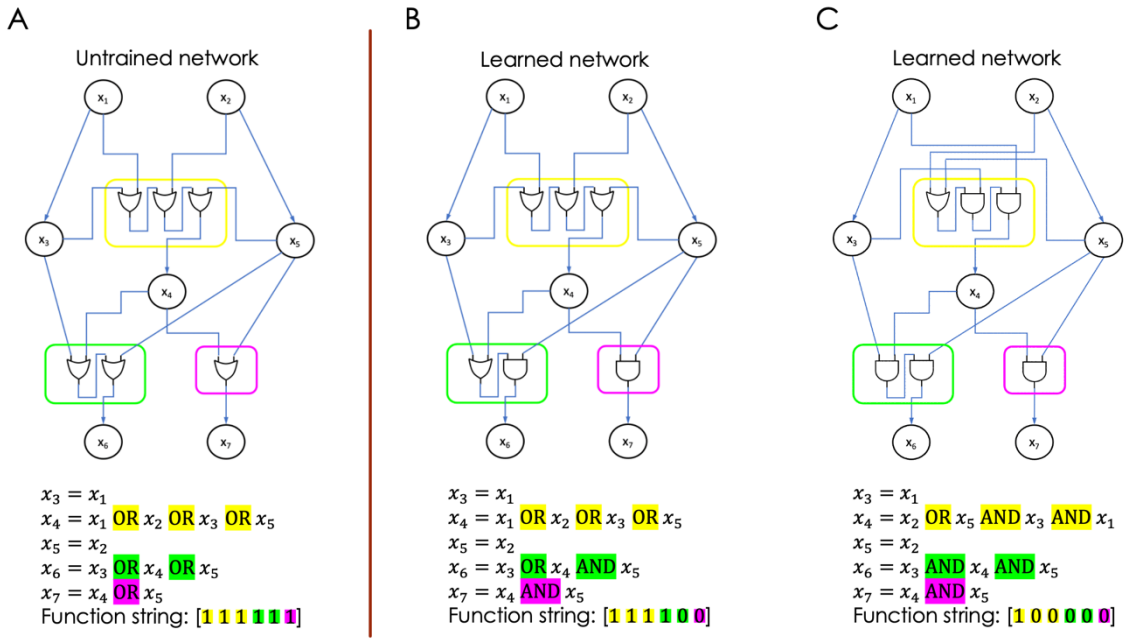
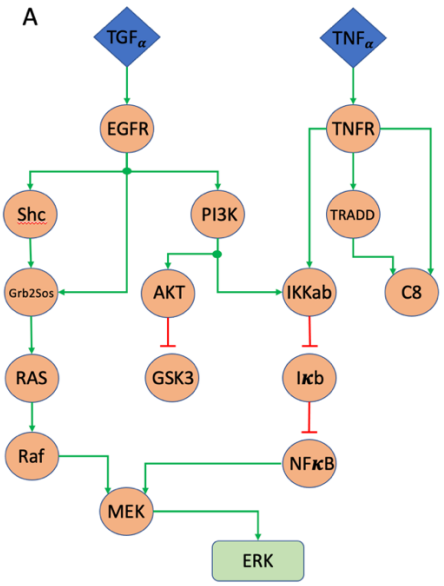


Figure 2.9 Examples of the trained networks with the learned Boolean functions. (A) The untrained network and associated Boolean functions based on Boolean Model 1. (B) A possible learned network with the gates having fixed input molecules. (C) A possible learned network with the gates that may have any possible input combination in its domain.

gates are learned instead of all OR gates, i.e., $\boldsymbol{g} = [1\ 1\ 1\ 1\ 0\ 0]$, that hypothetically results in better fitness. In this example, the learned model implies that x_4 and x_5 needs to be active at the same time to activate x_7 whereas it was enough to have either x_4 or x_5 to be active to get x_7 activated in the untrained model. Note that the inputs of the gates are fixed in this example. Namely, each gate g_i has always the same two input molecules. However, to make the learned functions more flexible, and hence to expand the search space with higher chance of getting better fitness percentage, the input edges of the gates should be floating so that the g_i can have any possible two inputs available in its domain (inputs of the molecule where g_i is located). To do so, we introduce another vector $\boldsymbol{p} = [p_i]$ that contains the indices of possible order of input molecules. To elaborate, if a molecule has three input molecules m_1, m_2, and m_3, then there are 6 possible arrangements of the input molecule orders ([m_1, m_2, m_3], [m_1, m_3, m_2], …, [m_3, m_2, m_1]). So, $p_i \in \{1,2,3,4,5,6\}$ picks the best order that gives the best fit with the selected gates. Therefore, the problem becomes searching for a bigger vector $\boldsymbol{G} = [\boldsymbol{g}, \boldsymbol{p}]$, and Equation (2.5) still applies to this problem. Figure 2.9C exemplifies a possible learned network in which the input edges are floating.

To test how well the proposed method works, we construct Equation (2.5) on a reasonable but imaginary toy network with a set of synthetic binary data. The toy network, previously studied by Saez-Rodriguez et al. (2009), consists of intracellular signaling proteins that are known to be activated by epidermal growth factor TNF (Tumor Necrosis Factor) receptors in mammalian cells. A directed graphical model of the network as well as the synthetic data is given in Figure 2.10. The synthetic binary data (Figure 2.10B) was obtained by using a reference model. The main goal is to learn a new model for each node that has the optimal fit to data. After constructing and solving the Equation (2.5) without

any biological constraints for this toy network, we learned the functions listed in Table 2.1. Note that we learn the function only for the molecules having at least two incoming inputs, that are in this case Grb2Sos, IKKab, C8, and MEK. The networks constructed with the functions given in Table 2.1 have 100% fitness to the synthetic data in Figure 2.10B. In Saez-Rodriguez et al. (2009), the training was done after the initial network was compressed. Therefore, the learned networks in Table 2.1 are not directly comparable to them. However, when the same compression is done here, the learned networks lead to the ones reported in Figure 1 of Saez-Rodriguez et al. (2009). This means that, with the proposed approach here, one can directly train the network without manipulating its initial topology and yet obtain the same results in terms of data fitness.



B

Inputs						
TGF_α	+	-	+	+	+	+
TNF_α	-	+	+	-	+	-
Inhibitors						
PI3K	-	-	-	+	+	-
Raf	-	-	-	-	-	+
Measurements						
NFκB	0	0	1	0	0	0
ERK	1	0	1	1	1	0
C8	0	1	1	0	1	0
AKT	1	0	1	0	0	1

Figure 2.10 The imaginary but reasonable toy network of intracellular signaling proteins. (A) The imaginary toy network of TNF downstream. The green normal arrows represent activatory interaction whereas the red blunt edges represent inhibitory interactions. (B) The synthetic readouts obtained from a reference model. "+" means the associated molecule is active (binary 1) and "-" means that the molecule is inactive (binary 0).

Source: This figure was reproduced from Saez-Rodriguez et al. (2009).

Table 2.1: Learned Boolean Functions for each Molecule

The Learned Functions 1	The Learned Functions 2	The Learned Functions 3	The Learned Functions 4
AKT = PI3K	AKT = PI3K	AKT = PI3K	AKT = PI3K
C8 = TRADD x TNFR	C8 = TRADD + TNFR	C8 = TRADD x TNFR	C8 = TRADD + TNFR
EGFR = TGFα	EGFR = TGFα	EGFR = TGFα	EGFR = TGFα
ERK = MEK	ERK = MEK	ERK = MEK	ERK = MEK
Grb2Sos = Shc x EGFR	Grb2Sos = Shc x EGFR	Grb2Sos = Shc + EGFR	Grb2Sos = Shc + EGFR
GSK3 = ~AKT	GSK3 = ~AKT	GSK3 = ~AKT	GSK3 = ~AKT
IKKab = TNFR x PI3K	IKKab = TNFR x PI3K	IKKab = TNFR x PI3K	IKKab = TNFR x PI3K
Iκb = ~IKKab	Iκb = ~IKKab	Iκb = ~IKKab	Iκb = ~IKKab
MEK = Raf + NFκB	MEK = Raf + NFκB	MEK = Raf + NFκB	MEK = Raf + NFκB
NFκB = ~ Iκb	NFκB = ~ Iκb	NFκB = ~ Iκb	NFκB = ~ Iκb
PI3K = EGFR	PI3K = EGFR	PI3K = EGFR	PI3K = EGFR
Raf = RAS	Raf = RAS	Raf = RAS	Raf = RAS
Ras = Grb2Sos	Ras = Grb2Sos	Ras = Grb2Sos	Ras = Grb2Sos
Shc = EGFR	Shc = EGFR	Shc = EGFR	Shc = EGFR
TNFR = TNFα	TNFR = TNFα	TNFR = TNFα	TNFR = TNFα
TRADD = TNFR	TRADD = TNFR	TRADD = TNFR	TRADD = TNFR

Training by learning Boolean functions of the molecules, i.e., the model, is useful if the initial network needs to be preserved while connecting it to the biological evidence, i.e., the experimental data. This approach is specifically practical if some analysis such as fault diagnosis analysis (explained in detail in the next chapter) is going to be conducted after training because it will preserve all molecules and their connections in the network, which prevents losing information on the molecules as well as the effect of their interactions that might be removed in other training approaches. In addition, after training, generally there is not a unique solution. In fact, if the size of the network increases, then the likelihood of observing more networks with the same objective value increases as well, which might be an issue in network analyses. However, the learned networks are usually correlated. Thus, the solution set can be clustered into a reasonable number of clusters, and then the cluster centroids can be analyzed.

CHAPTER 3

VULNERABILITY ANALYSIS OF MOLECULAR NETWORKS

Modeling the molecular networks and training network models are done to obtain biologically consistent models with high accuracy in experimental data prediction so that one can develop methods and frameworks to perform different analyses. Such analyses may result in important outcomes that may pave the way for a possible treatment to a complex disorder. In this chapter, we provide insights into one of the analysis techniques that is the vulnerability analysis (fault diagnosis analysis) studied to rank the molecules in terms of their disruptive effect on the network functionality when they are faulty (dysfunctional). Such analysis is useful for target discovery that leads to drug development.

The main purpose of fault diagnosis analysis methods is to understand how vulnerable the entire network is to the dysfunction of one or multiple molecules. A molecule may become dysfunctional due to intrinsic or extrinsic reasons. The dysfunction of a molecule can be defined as a failure to respond correctly to its input signals, which may further cause incorrect responses at the output(s) of the network. Therefore, one can define the vulnerability level of a molecule as the probability of having incorrect network responses when the molecule is dysfunctional. We first introduce some possible fault models that can be studied for modeling the dysfunctional state of a molecule in the next section. Then, we provide a mathematical framework to compute the single and multi-fault vulnerability levels of molecules, exemplified on a toy network. Next, the worst possible signaling failures in molecular networks is examined by comparing the maximum vulnerability level, i.e., the highest probability of network failure, versus the number of

faulty molecules to understand how the network functionality is affected in the presence of one or more dysfunctional molecules, for which an efficient algorithm is developed. Moreover, another algorithm is developed to understand how many time points might be needed to calculate vulnerability level of a molecule or group of molecules in a Boolean modeling framework. The methods are applied to the experimentally verified ERBB and T cell signaling networks. All of these studies and our observations are elaborated in the subsequent sections.

3.1 Molecular Fault Models

The molecules in a cell may become faulty because of different extrinsic or intrinsic reasons. If the faults cannot be repaired by cellular mechanisms, then they may initiate the formation of serious diseases such as cancer, autoimmune diseases, and mental disorders. Since we model the network using a Boolean modeling framework as introduced in Chapter 2, we also model the molecular faults in Boolean domain. One advantage of using Boolean models in the molecular network modeling is that the networks can be represented as digital circuits in which each clock cycle may model different time responses of the molecular network. This allows one to apply methods developed for digital circuits in the context of biological networks, especially for the fault models.

The faults in a molecule can be temporary or permanent. The temporary faults can be modeled by transient errors (also called soft error) that occurs in one clock cycle and then disappear in the other clock cycles in a Boolean model. In other words, the molecule state might incorrectly be binary 0 (or binary 1) in a single clock cycle, and then turns back to its nominal state afterwards. This incorrect transient error may or may not propagate to

the network output. The probability of propagation of the transient error to the output of the network for each molecule can be defined as the molecular vulnerability level whose formulation and computation is provided by Abdi et al. (2008).

Another way of modeling the molecular faults is using 1-bit inversion called von Neumann error that is studied in the reliability analysis concept of digital circuits (Han et al., 2011). In this error model, it can be assumed that the faulty molecule outputs the inverse of its nominal state. For instance, in a reaction, if a molecule's nominal response is binary 0 to its input molecules, then the same molecule response erroneously becomes binary 1 to the same inputs or vice versa if it is modeled using von Neumann error. Despite the unusuality of usage of this type of fault model in the concept of molecular faults, still it can be used to assess the effect of temporary faults in the molecular networks.

The most common fault type used to model molecular faults is the stuck-at permanent faults (e.g., Habibi et al., 2014a). In these fault models, the molecule's state stuck at a value permanently through all clock cycles regardless of what its inputs are. In the Boolean modeling framework, these faults can be stuck-at-0 (SA0) or stuck-at-1 (SA1). The SA0 fault mean that the molecule is always inactive, i.e., its state is binary 0 all the time, no matter what the state of its input molecules are. This type of fault model is practical for understanding effects of hypoactivity of the molecules. Similarly, the SA1 fault mean that the molecule is always active, i.e., its state is always binary 1 independently from its input molecules. The SA1 fault model can be used to analyze the effects of hyperactivity of the molecules. If there is not any specific preference on the stuck-at faults, one can study equiprobable SA0 and SA1 fault models while computing their vulnerability levels, which is reasonable if no prior knowledge exists about the molecules and network.

In our vulnerability analyses that are elaborated in the next section, we model the dysfunction of the molecules using stuck-at faults because they are biologically more realistic and they occur in complex diseases (e.g., attenuated PTEN levels exist in human breast cancer (Geva-Zatorsky et al., 2006) and elevated Wip1 levels exist in multiple human cancer types such as breast, lung, and pancreas cancers; Bulavin et al., 2002; Castellino et al., 2007; Li et al., 2002; Saito-Ohara et al., 2003).

3.2 Equations for Computing Vulnerability Levels

Computing the vulnerability level of a molecule or a group of molecules in a network can help with identifying and ranking the key components of the network, to discover appropriate therapeutics targets. Vulnerability of a molecule can be defined as the probability of having incorrect network responses when the molecule is dysfunctional. The dysfunction state of a molecule can be defined as a failure to respond correctly to its input signals. In this section, we consider stuck-at-0 (SA0) and stuck-at-1 (SA1) fault models to model the dysfunction of molecules, that are constantly 0, inactive, or 1, active, regardless of what the input signals of the molecule are.

To compute the vulnerability level of a molecule, one needs to first introduce the sample space associated with the correct and incorrect network responses at the network output. Suppose K is the number of intermediate molecules in the network. Let N be the number of molecules that are simultaneously faulty, i.e., dysfunctional, I be the number of the network input combinations, CC be the number of clock cycles (time points) for which the network response is computed (an algorithm for determining the required CC is given in Section 3.4.1), and finally, let l be the subscript ranging from 1 to $C(K, N)$, indexing

faulty molecules or groups of faulty molecules. Then, for each input combination stimulating the network in the presence of N faulty molecules, there will be CC number of output responses that may be correct, c, or erroneous, e, in each clock cycle. Therefore, the sample space S can be defined as the set of all possible output sequences of c and e responses over CC clock cycles, that is:

$$S = \{c, e\}^{\mathrm{CC}}. \tag{3.1}$$

Moreover, for the l-th faulty molecule or the l-th group of faulty molecules, we define the event S_l, a subset of S, as the set of all output sequences of c and e responses over CC clock cycles observed for all input combinations applied, that is:

$$S_l = \{v_1, v_2, \dots, v_I\}. \tag{3.2}$$

Note that $v_i \in S$, $i = 1, \dots, I$, is a sequence of c and e of length CC, in which having an e in the t-th element of v_i means that an erroneous response is observed in the t-th clock cycle. Depending on the possible network responses and that which molecule or group of molecules is faulty, v_is may have the same or different probabilities. Also note that some v_is may be identical, therefore, I is indeed the maximum number of elements of S_l. For $\mathrm{CC} = 2$ and $I = 2$, for example, we have $S_l = \{v_1, v_2\}$ where $v_1, v_2 \in S = \{c, e\}^2 = \{(c, c), (c, e), (e, c), (e, e)\}$.

In addition, we define CC number of events as follows: $E_1 =$ the event of having an erroneous network response at the output in the 1st clock cycle, …, and $E_{\mathrm{CC}} =$ the event of having an erroneous network response at the output in the CCth clock cycle. Note that

E_1 is the set of those v elements in S_l in Equation (3.2) that have an e as the 1st entry, E_2 is the set of those v elements in S_l that have an e as the 2nd entry, and so on. We define the vulnerability level of the l-th molecule M_l, $\text{Vul}(M_l)$, as the probability of having an erroneous network response in the 1st clock cycle, …, or in the CCth clock cycle, when M_l is dysfunctional. Therefore, $\text{Vul}(M_l)$ can be written as:

$$\text{Vul}(M_l) = P\left(\bigcup_{t=1}^{\text{CC}} E_t \;\&\; M_l \text{ is dysfunctional}\right). \tag{3.3}$$

Since we consider SA0 and SA1 as the fault models, Equation (3.3) can be expanded as follows:

$$\begin{aligned}\text{Vul}(M_l) = {} & P\left(\bigcup_{t=1}^{\text{CC}} E_t \;\&\; M_l \text{ is SA0}\right) P(M_l \text{ is SA0}) \\ & + P\left(\bigcup_{t=1}^{\text{CC}} E_t \;\&\; M_l \text{ is SA1}\right) P(M_l \text{ is SA1}).\end{aligned} \tag{3.4}$$

While we assume equi-probable SA0 and SA1 faults for each molecule, i.e., $P(M_l \text{ is SA0}) = P(M_l \text{ is SA1}) = 0.5$ in our computations, Equations (3.3) and (3.4) can be extended to other fault models and fault probabilities.

Equation (3.4) is provided for computing the vulnerability level of a single molecule. However, it is also of interest to study the abnormal network responses when

multiple molecules are faulty at the same time, for which Equation (3.4) can be extended as follows:

$$\mathrm{Vul}(M_1, \dots, M_N) = \sum_{k=1}^{2^N} P\left(\bigcup_{t=1}^{\mathrm{CC}} E_t \middle| (M_1, \dots, M_N)_k\right) P((M_1, \dots, M_N)_k), \tag{3.5}$$

where $(M_1, \dots, M_N)_k \in \{\mathrm{SA0}, \mathrm{SA1}\}^N$ is the k-th fault vector for the group of N dysfunctional molecules $M_1, \dots, M_N$.

To illustrate, assume both I and $\mathrm{CC} = 2$. When $N = 1$, a single faulty molecule, there are two events, i.e., $S_{l,\mathrm{SA0}} = \{v_1, v_2\}$ and $S_{l,\mathrm{SA1}} = \{v_1', v_2'\}$, associated with the l-th faulty molecule M_l being SA0 or SA1. Moreover, assume hypothetically that we have $S_{l,\mathrm{SA0}} = \{v_1, v_2\} = \{(c, c), (e, e)\}$ with probabilities $P(v_1) = 0.5$ and $P(v_2) = 0.5$, and $S_{l,\mathrm{SA1}} = \{v_1'\} = \{(c, e)\}$ with the probability of $P(v_1') = 1$. Based on the definition of E_t, it can be shown that $E_1 = E_2 = \{(e, e)\}$ when M_l is SA0 whereas $E_1 = \emptyset$ and $E_2 = \{(c, e)\}$ when M_l is SA1. Using Equation (3.4), vulnerability level of M_l can be computed for equi-probable SA0 and SA1 faults as follows:

$$\begin{aligned}
\mathrm{Vul}(M_l) &= P(E_1 \cup E_2 | M_l \text{ is SA0}) P(M_l \text{ is SA0}) \\
&\quad + P(E_1 \cup E_2 | M_l \text{ is SA1}) P(M_l \text{ is SA1}), \\
&= P\big((e, e)\big)0.5 + P\big((c, e)\big)0.5, \\
&= P(v_2)0.5 + P(v_1')0.5, \\
&= 0.5{\times}0.5 + 1{\times}0.5 = 0.75.
\end{aligned} \tag{3.6}$$

3.2.1 Single-fault Vulnerability Analysis

Single-fault vulnerability analysis is done by assuming one molecule is faulty at a time and then checking whether the network functionality changes or not at the output of the network. Figure 3.1 exemplifies computation of the vulnerability on a toy network ($I = 4$, $\mathrm{CC} = 1$) that is modeled using the Boolean Model 1 introduced in Chapter 2. Suppose x_3 in Figure 3.1A is SA0. Then, as seen in Figure 3.1C, three out of four responses are erroneous (e). More specifically, we have $S_{l,\mathrm{SA0}} = \{v_1, v_2, v_3, v_4\} = \{(c), (e), (e), (e)\} = \{(c), (e)\}$ in which $P((c)) = 0.25$ and $P((e)) = 0.75$. Therefore, using Equation (3.3), one can compute vulnerability of x_3 as ¾ = 0.75. Similarly, if we assume that x_3 is SA1, then there is only one erroneous response at the output, which results in the vulnerability level of ¼ = 0.25. In the case of equiprobable SA0 and SA1 fault model, Equation (3.4) can be used, which results in 0.75 × 0.5 + 0.25 × 0.5 = 0.5.

3.2.2 Double-fault Vulnerability Analysis

In this section, the computation of double-fault vulnerability levels is exemplified on a toy network. In the double-fault vulnerability analysis, it is assumed that two molecules are faulty synchronously at a time. It has been shown that multiple molecules are involved in the formation of complex diseases such as schizophrenia (Emamian, 2012). Therefore, performing double and multiple-fault vulnerability analysis is important and may help to discover a group of molecules that might be driver of some complex disorders.

Suppose that we have the toy network in Figure 3.2 that contains feedback interactions and modeled using Boolean Model 1. Due to the feedback interactions, the network may present different responses at different time points. For this example, we

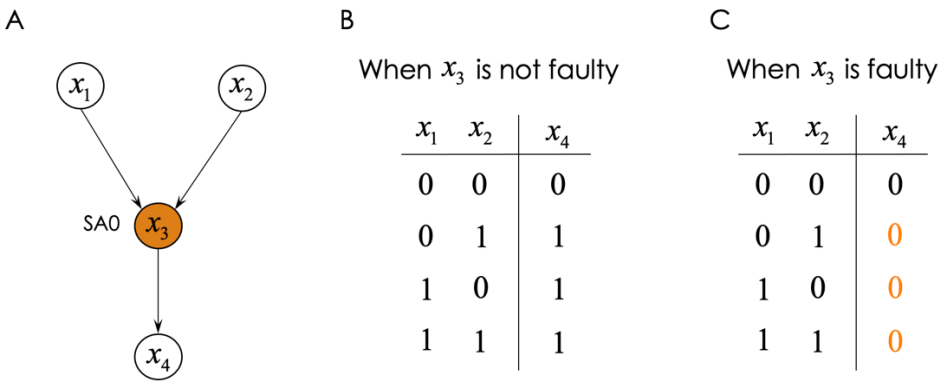


Figure 3.1 An example for vulnerability computation. (A) A toy network. (B) The normal network truth table. (C) The abnormal network truth table, in which x_3 is SA0.

consider only two clock cycles (CC = 2). First, we compute the single-fault vulnerability levels of each nodes in the toy network using Equation (3.4) for equiprobable SA0 and SA1 fault models. Later on, we compute the double-fault vulnerability levels for each possible faulty pair of molecules assuming equiprobable SA0 and SA1 fault models, i.e., the nodes can be SA0-SA0, SA0-SA1, SA1-SA0, and SA1-SA1, using Equation (3.5). The results are compared in Figure 3.3. A major observation is that a molecule with high single-fault vulnerability level is usually a component of the pairs with high double-fault vulnerability levels. To illustrate, the node x_3 has single-fault vulnerability level of 0.5 and it is one of the components of all pairs having a vulnerability level of 0.5 (Figure 3.3B). Another noteworthy observation is that some molecules with relatively low individual vulnerability levels may have high double-fault vulnerability when they are faulty at the same time (e.g., x_5-x_6 pair in Figure 3.3B). Such observations are not obvious without performing the vulnerability analysis presented in this chapter.

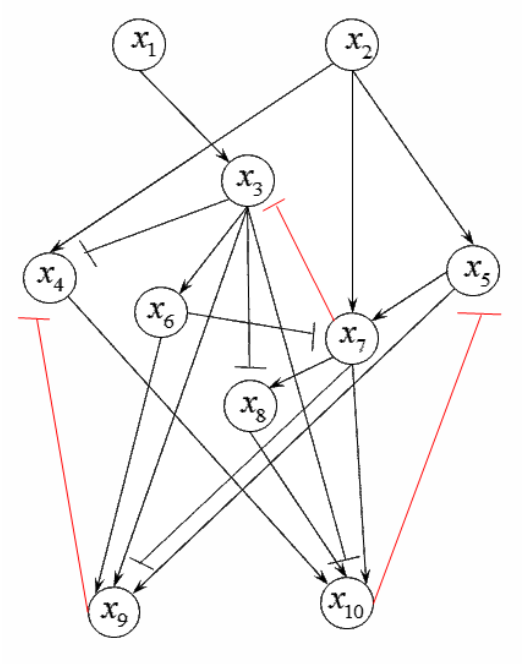


Figure 3.2 A toy network containing feedback interactions.
Note: The interactions represented by red edges are the feedbacks.

On a very large network with hundreds of molecules and interactions, one may end up with several vulnerability levels obtained for several faulty combinations of molecules (especially when the number of faulty molecules gets higher). This becomes an issue if one needs to test some of the observations in laboratory experiments. Since it is not convenient to test every vulnerability level for each combination of faulty molecules due to the lack of resources, the best candidates need to be systematically selected. One way to achieve this is filtering the results. To do so, one may use a prior knowledge existing in the literature to eliminate irrelevant results. For instance, assume it has been experimentally verified in the previous works that x_3 is an important molecule and its deficiency disrupts the network's functionality. Then, it would be a reasonable choice to keep the results of molecule groups whose one component is x_3. Also, some faulty molecules may not be technically tested in

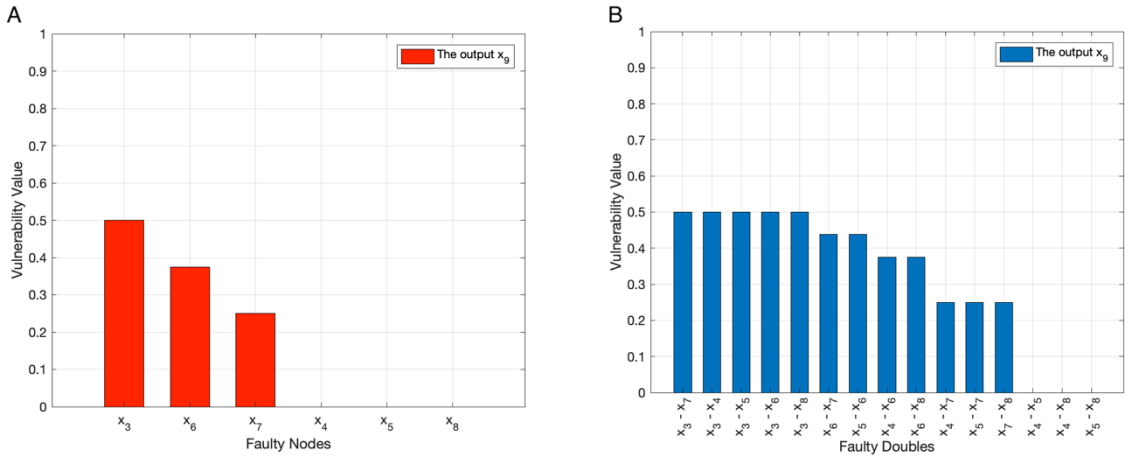


Figure 3.3 Single and double-fault vulnerability results of the toy network. (A) Single-fault vulnerability levels. (B) Double-fault vulnerability levels.

the experiments because the knowledge how to perturb the molecule in the lab experiments does not exist yet. Thus, the results might be filtered further out by eliminating the ones containing such molecules. In addition, one can select the groups of faulty molecules whose vulnerability levels are not obvious. To elaborate, the combinations containing one or more molecules with high single-fault vulnerability level can be classified as obvious. On the other hand, if each component of the faulty molecule group has a low single-fault vulnerability level, but when they all are dysfunctional simultaneously, they have a high vulnerability level, then such a combination would be a noteworthy observation and plausible to test in experiments. Lastly, some topological analyses can be conducted using graph theory methods and techniques to eliminate obvious vulnerability levels. More precisely, in a group of faulty molecules, the topological closeness (Equation 3.12) of the molecules may reveal whether they are closely connected or not, which can be considered as an important feature while classifying the obvious and nonobvious observations.

3.3 The Worst Possible Signaling Failures in Intracellular Signaling Networks

Cellular functions are partly controlled by signaling events within networks of molecules in a cell (Helikar et al., 2008; Saadatpour & Albert, 2012; Saez-Rodriguez et al., 2007). Signals are transmitted from the cell membrane to the nucleus via intracellular signaling networks, to regulate some target molecules and control the cell function. Failures in molecules within the signaling networks may cause them to generate erroneous signals whose propagation to the output of the network may disrupt the network functionality and eventually may cause some diseases (Abdi et al., 2008; Emamian, 2012).

In this section, our goal is to develop a systematic method to analyze and identify the worst possible signaling failures in intracellular signaling networks. We define the worst possible signaling failure as a pathological phenomenon that results in the highest probability of network failure, i.e., the maximum vulnerability level, where the network failure is defined as departure of the network response from its normal level. The said pathological phenomenon is characterized to be emerged from the presence of one or more dysfunctional molecules in the network. It is conceivable that different individual dysfunctional molecules can result in different network failure probabilities. It is not clear, however, what happens if two or more molecules are concurrently dysfunctional, and if the network failure probability increases with the number of simultaneously dysfunctional molecules or not. It is also of interest to have an efficient algorithm to determine the maximum possible network failure probability, over the large number of all possible groups of dysfunctional molecules. The computational complexity of an exhaustive search approach for a network with K molecules is extremely high, in the order of $K^{K/2}$, which is unmanageable as K increases. Here we introduce a computationally efficient algorithm

with a much less running time in the order of K^3, for identifying the worst possible signaling failures, considering multiple dysfunctional molecules. Then, using the algorithm, we analyze two experimentally verified signaling networks in the following subsections: a small signaling network that regulates the transmembrane tyrosine kinase ERBB (Sahin et al., 2009), a therapeutic target in breast cancer, and a large T cell signaling network (Saez-Rodriguez et al., 2007).

3.3.1 Algorithm for the Worst Possible Signaling Failure Analysis

In this section, we provide a detailed explanation of the proposed algorithm. The worst possible signaling failure analysis can be performed by an exhaustive search. However, the time needed by the exhaustive search grows exponentially as the size of the network increases, as presented in Section 3.3.4. To avoid this high computational complexity, we propose the main algorithm with the following four steps:

I. First, we compute an upper bound on the number of clock cycles needed for computing the vulnerability levels (Section 3.4.1), so that we prevent running network simulations longer than what is needed.

II. Next, we use Equation (3.5) to compute $\mathrm{Vul}(M_{l_1}, M_{l_2}, \dots, M_{l_N})$ for $N = 1, 2,$ and 3. This is motivated by the observation (Habibi et al., 2014a) that typically a molecule with high vulnerability appears in larger groups of molecules with high vulnerabilities, and based on our experiments, $N \geq 4$ is large enough and provides good accuracy, as described in Section 3.3.4. Thus far $\max_{l_1} \mathrm{Vul}(M_{l_1})$, $\max_{l_1,l_2} \mathrm{Vul}(M_{l_1},M_{l_2})$, and $\max_{l_1,l_2,l_3} \mathrm{Vul}(M_{l_1},M_{l_2},M_{l_3})$ represent the worst possible signaling failures when there are single, double and triple faults, respectively.

III. To determine the worst possible signaling failure when there are four simultaneously faulty molecules, $N = 4$, we pick the molecular triplet, group of $N - 1$ faulty molecules, having the highest vulnerability value, e.g., (a, b, c). Then we compute the vulnerabilities only for those $K - (N - 1)$ quadruplets, groups of N faulty molecules with $N = 4$, that include (a, b, c),

i.e., (a, b, c, M_l). This results in $\max_l \mathrm{Vul}(a, b, c, M_l)$ as the worst possible signaling failure when there are $N = 4$ simultaneous faults.

IV. Then, we repeat Step III for $N = 5, \ldots, K$, to complete the worst possible signaling failure analysis.

Note that this algorithm is not limited to a specific molecular network. Furthermore, in addition to the vulnerability parameter introduced in Section 3.2, other parameters that quantify and rank the importance of a molecule or a group of molecules can be used in the algorithm as well.

3.3.2 ERBB Signaling Network Worst Failure Study and Results

The ERBB network (Figure 3.4) has one input and one output. The input molecule is the epidermal growth factor (EGF), whereas the output molecule is the retinoblastoma protein (pRB). This network is studied in the context of breast cancer and understanding some drug effects (Sahin et al., 2009). The Boolean equations that specify how the activity of each molecule is regulated by its inputs are listed in Table A.1 (see Appendix). To model a dysfunctional molecule, we assume its activity state is either stuck-at-0, SA0, or stuck-at-1, SA1, each with a probability of 1/2 (Abdi et al., 2008).

Let N be the number of molecules that are simultaneously faulty, i.e., dysfunctional, in the network. According to the developed vulnerability analysis equations (Section 3.2) and using the proposed worst failure analysis algorithm (Section 3.3.1), the network maximum vulnerability is computed for each N (Figure 3.5). Here N varies from 1 to 18, since the number of intermediate molecules in the network (Figure 3.4) is 18. For any N, the network maximum vulnerability is the highest probability of network failure, where the network failure is defined as departure of the network response from its normal levels. More precisely, the network maximum vulnerability for a given N is a parameter that

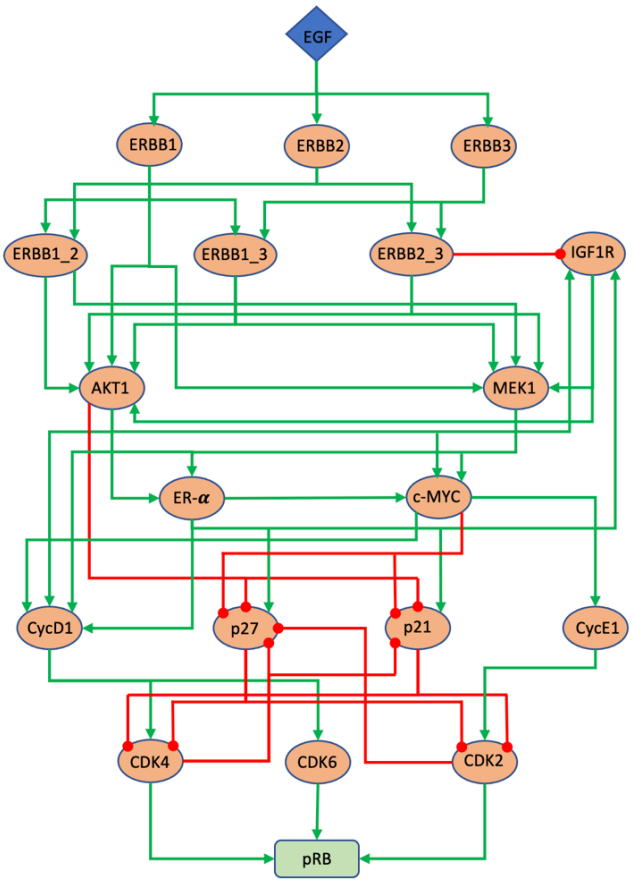


Figure 3.4 The experimentally verified ERBB signaling network.
Note: The green arrows represent activatory interactions and the red circle-ended edges represent inhibitory interactions. The input and output nodes represent EGF and pRB, respectively.

Source: This figure was reproduced from Sahin et al. (2009).

quantifies the worst possible signaling failure when there are N faulty molecules in the network.

A noteworthy observation is that as the number of faulty molecules N increases, maximum vulnerability values do not necessarily increase (Figure 3.5). While we see a maximum vulnerability increase going from single faults to double faults, $N = 1$ and 2, respectively, the maximum vulnerability does not increase further afterwards. Another

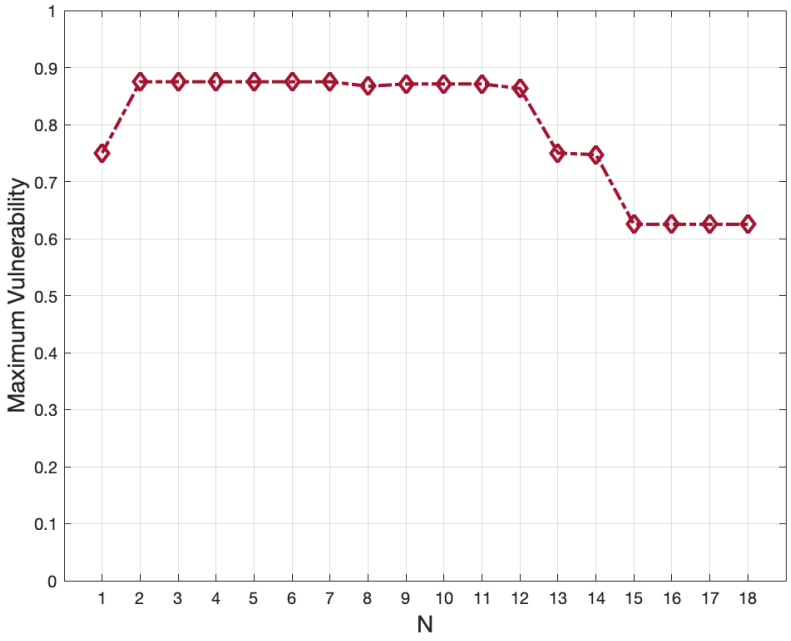


Figure 3.5 The ERBB signaling network maximum vulnerability levels, when there are N dysfunctional molecules in the network, computed using the proposed algorithm to study the worst possible signaling failures.

interesting observation is that the smallest N for which we see the highest maximum vulnerability in this network is $N = 2$, i.e., double faults. This means there are some pairs of faulty molecules that cause the most detrimental network damage, and an increase in the number of faulty molecules does not deteriorate the network function.

3.3.3 T Cell Signaling Network Worst Failure Study and Results

The T Cell network (Figure 3.6) has three inputs and fourteen outputs. The input molecules are cd28, cd4 and tcrlig, that stand for cluster of differentiation 28, cluster of differentiation 4, and ligand-bound T-cell receptor, respectively (Saez-Rodriguez et al., 2007). For the output molecules we have shp2 (Src homology region 2 domain containing phosphatase-2, bclxl (B-cell lymphoma-extra-large), p70s, ap1 (activator protein 1), sre (serum response

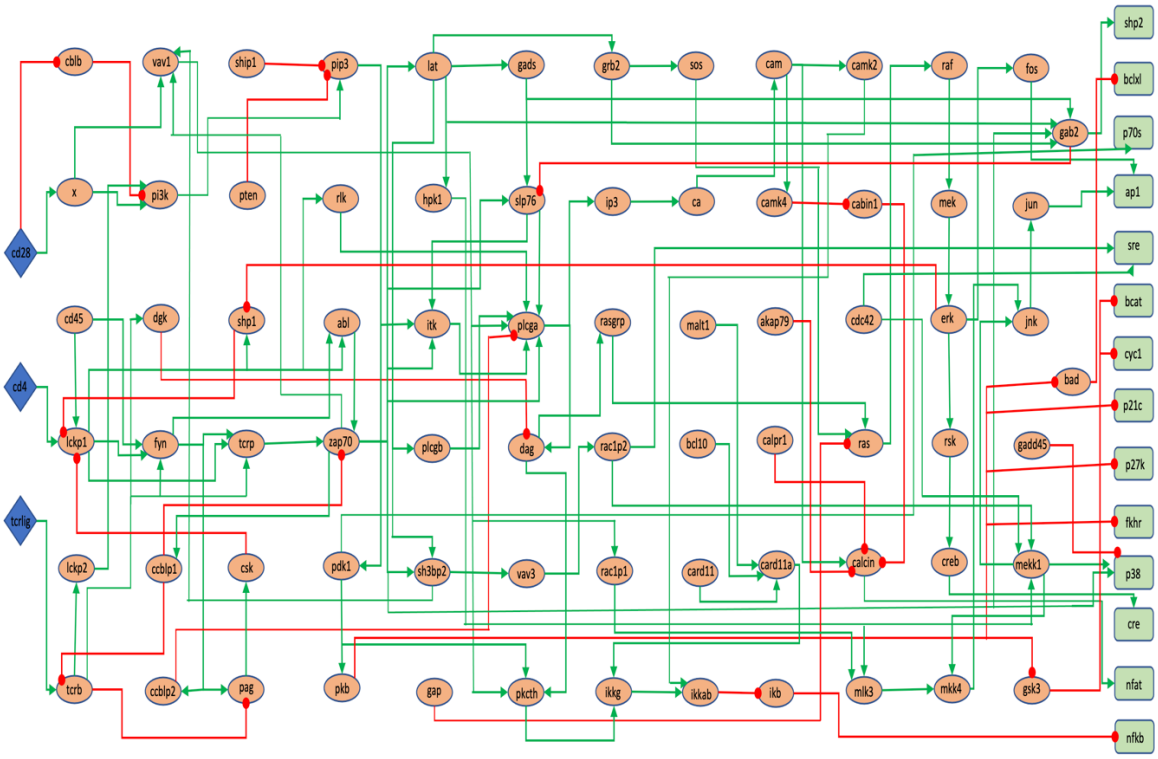


Figure 3.6 The experimentally verified T cell signaling network.
Note: The green arrows are activatory interactions and the red circle-ended edges are inhibitory interactions. The input nodes represent cd28, cd4, and tcrlig whereas the output nodes stand for shp2, bclxl, p70s, ap1, sre, bcat, cyc1, p21c, p27k, fkhr, p38, cre, nfat, and nfkb.

Source: This figure was reproduced from Saez-Rodriguez et al. (2007).

element), bcat (branched-chain amino acid transaminase), cyc1 (cytochrome c1), p21c, p27k, fkhr (forkhead transcription factor Foxo1), p38, cre (cAMP, cyclic adenosine monophosphate, response elements), nfat (nuclear factor of activated T-cells), and nfkb (nuclear factor kappa-light-chain-enhancer of activated B cells) (Saez-Rodriguez et al., 2007). The Boolean equations that specify how the activity of each molecule is regulated by its inputs are listed in Table A.2 (see Appendix).

Using the developed vulnerability analysis equations (Section 3.2) and the proposed worst failure analysis algorithm (Section 3.3.1), the network maximum

vulnerability is computed for each N (Figure 3.7), where N is the number of molecules that are simultaneously faulty. For any N, the network maximum vulnerability is the highest probability of network failure. We notice that similar to the ERBB network and for all outputs, as the number of faulty molecules N increases, maximum vulnerability values do not necessarily increase (Figure 3.7). Additionally, for the network outputs ap1, bcat, and p70s, while we see a maximum vulnerability increase going from single faults to double

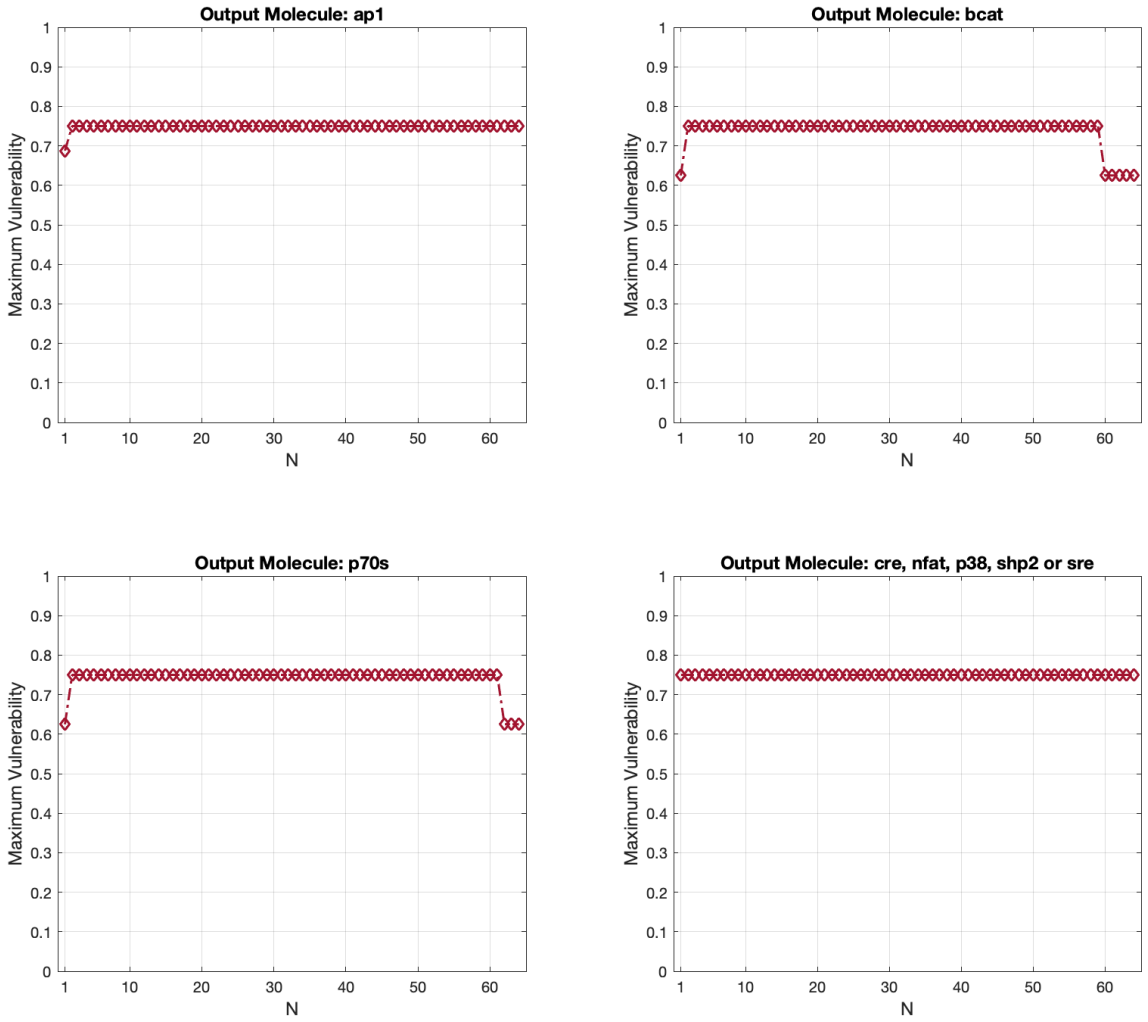


Figure 3.7 The T cell signaling network maximum vulnerability levels for the network outputs ap1, bcat, cre, nfat, p38, p70s, shp2 and sre, when there are N dysfunctional molecules in the network. The results are computed using the proposed algorithm to study the worst possible signaling failures.

faults, $N = 1$ and 2, respectively, the maximum vulnerability does not increase further afterwards. For these outputs, the smallest N for which we see the highest maximum vulnerability in this network is $N = 2$, i.e., double faults. This means that there are some pairs of faulty molecules that cause the most detrimental network damage for these outputs, and an increase in the number of faulty molecules does not make things worse. For the network outputs cre, nfat, p38, shp2 and sre, this behavior changes, i.e., the highest maximum vulnerability occurs when $N = 1$. This implies that there are some single faulty molecules that cause the worst possible network failures at these outputs.

3.3.4 Computational Complexity of the Worst Signaling Failure Analysis Algorithm

In this section, we determine the computational complexity of the proposed algorithm and compare it with the running time of exhaustive search. The worst possible signaling failure analysis can be performed via an exhaustive search. This means that if it is of interest to find the maximum network vulnerability when there are N faulty molecules in the network, all possible groups of N faulty molecules have to be considered one by one, and the vulnerability value for each group needs to be individually computed. For example, consider a network with $K = 50$ molecules. When $N = 2$, the total number of pairs of faulty molecules that the exhaustive search has to examine can be shown to be 1,225 (see Equation (3.7)). For $N = 5$, however, the total number of groups of five faulty molecules that the exhaustive search needs to consider increases to 2,118,760. This computational complexity becomes highly prohibitive as the network size K increases. In what follows, we show that the proposed worst signaling failure analysis algorithm is much less complex than the exhaustive search.

To determine and compare the computational complexities, let K be the number of molecules in a network, and N be the number of molecules that are simultaneously faulty in the network. The total number of groups of N faulty molecules out of K molecules, $1 \leq N \leq K$, is given by the following equation, in which $C(K, N)$ represents the number of possible combinations:

$$C(K, N) = \frac{K!}{(K-N)!\,N!} = \frac{K \times (K-1) \times \cdots \times (K-(N-1))}{N!} = \mathrm{O}(K^N). \qquad (3.7)$$

Here the O-notation stands for the asymptotic upper bound (Cormen et al., 2009), with K being large. The term $\mathrm{O}(K^N)$ represents the computational complexity of $C(K, N)$ as a function of K and N.

The computational complexity of the exhaustive search $\alpha(K)$ is the overall number of all possible groups of N faulty molecules for which vulnerabilities have to be computed, $N = 1, \ldots, K$, i.e., $\alpha(K) = C(K, 1) + \cdots + C(K, N)$. To simplify the notation and without loss of generality, assume K is even. We note that $C(K, N)$ has a maximum at $N = K/2$ (Cormen et al., 2009), it is symmetric, i.e., $C(K, N) = C(K, K-N)$, $N = 1, \ldots, (K/2) - 1$, and $C(K, K) = 1$. Therefore, the computational complexity of the exhaustive search simplifies to:

$$\begin{aligned} \alpha(K) &= C(K, 1) + \cdots + C(K, K/2) + \cdots + C(K, K) \\ &= 2 \sum_{N=1}^{(K/2)-1} C(K, N) + C(K, K/2) + 1. \end{aligned} \qquad (3.8)$$

With $C(K, K/2)$ being the dominant term in Equation (3.8) when K is large, and also using Equation (3.7), the exhaustive search computational complexity can be finally written as:

$$\alpha(K) = \mathrm{O}\left(K^{K/2}\right). \tag{3.9}$$

To determine the computational complexity of the proposed worst failure analysis algorithm (Section 3.3.1), we note that initially all groups of one, two and three faulty molecules are considered, $N = 1, 2, 3$, and for the rest, $N = 4, \ldots, K$, only $K - (N - 1)$ vulnerabilities are computed. This is inspired by the observation (Habibi et al., 2014a) that typically a molecule with a high vulnerability appears in larger groups of molecules with high vulnerabilities, and based on our experiments, $N \geq 4$ is large enough and provides good accuracy (for further details, see Section 3.3.1). Therefore, the computational complexity $\beta(K)$ of the algorithm can be written as:

$$\beta(K) = \sum_{N=1}^{3} C(K, N) + \sum_{N=4}^{K} (K - (N - 1)). \tag{3.10}$$

It can be verified that $C(K, 3)$ in the above expression is the dominant term, when K is large. This simplifies the proposed algorithm computational complexity to:

$$\beta(K) = \mathrm{O}(K^3). \tag{3.11}$$

Upon comparing Equations (3.9) and (3.11), we note that since $\mathrm{O}(K^3) \ll \mathrm{O}\left(K^{K/2}\right)$, the proposed algorithm is much simpler and therefore much faster than the exhaustive

search. For example, for a network with $K = 50$ molecules, the proposed algorithm complexity is in the order of $50^3 \approx 1.3\times10^5$, which is much smaller than $50^{25} \approx 3\times10^{42}$, the exhaustive search complexity. With regard to the accuracy, we have observed that the differences between the results of the algorithm and the exhaustive search are 0.5% and 0%, for the ERBB and T cell networks, respectively, over all N values for which it was not impractical to perform the exhaustive search.

To practically compare the execution times of the proposed algorithm and the exhaustive search, we ran them on a computer with Intel Core i7 CPU, 3.4 GHz and 32 GB RAM. For the small ERBB signaling network (Figure 3.4), the exhaustive search took about 10 days for $N = 1, \dots, 18$. In contrast and again for $N = 1, \dots, 18$ (Figure 3.5), the proposed algorithm took only about 20 seconds, with an accuracy of 99.5%, compared to the exhaustive search results.

3.4 The Number of Clock Cycles Needed to Compute Vulnerability Levels

Modeling and analysis of molecular networks become more challenging if there are positive or negative feedback paths. Due to the feedback mechanisms, network responses may change over time because of some internal compensatory or regulatory mechanisms (Azpeitia et al., 2017; Somogyi & Greller, 2001). Feedbacks can cause delays in propagation of signals to the network outputs, while passing through the feedback paths. Therefore, analysis of the effects of feedback in computing the vulnerability levels of the network molecules is of interest. More precisely, in this section we are interested in determining how many clock cycles are needed to compute the vulnerability level of a molecule or a group of molecules, when there are feedbacks in the network. For this

purpose, in this section we propose an algorithm that computes an upper bound on the number of clock cycles needed to generate the network response, to calculate its molecular vulnerabilities. Using this algorithm, one can specify how many times the network needs to be simulated for a normal or abnormal signal to complete its propagation to the network output. This is needed in the proposed main algorithm for the worst signaling failure analysis (Section 3.3.1), to minimize the overall simulation time.

The feedback paths in a network can be modeled by unit-delay memory elements called flip-flops (Abdi et al., 2008). In a network, if there is only one feedback path, then we intuitively need at most two clock cycles to see the full effect of an error, i.e., the effects of an incorrect signal value of a faulty molecule on possibly other molecules and pathways, that collectively determine the network output response. This is because of the delayed response of the flip-flop in the feedback path. In fact, if after the 1st clock cycle there exists an erroneous signal value of a faulty molecule at the input of the feedback flip-flop, then the 2nd clock cycle may be needed for that error to show its full effect at the network output. This is because the feedback-delayed erroneous signal of the faulty molecule may affect some other molecules and pathways in the 2nd clock cycle, which may increase the probability of incorrect responses at the network output. In general, if there exist F feedback paths in the network, then we need to simulate the network for at most $F + 1$ clock cycles, for an error to show its full effect at the network output. This maximum number of clock cycles is required, if these two conditions hold: (i) all the feedback paths are in the same pathway, connected in series and exhibiting F feedback flip-flops; and (ii) after the 1st clock cycle, an error appears at the input of the first feedback flip-flop.

The upper bound of $F + 1$ clock cycles can be tightened, by finding the pathways within the network containing the highest number of feedback paths in series. For instance, if there exist F feedback paths in the network, with $L \leq F$ being the largest number of feedback paths in series on the same pathway, then the maximum clock cycles needed is bounded above by $L + 1$, i.e., the number of clock cycles needed for an error to show its full effect is less than or equal to $L + 1$. Therefore, to determine the maximum number of required clock cycles, it is sufficient to examine the connections between the network feedback paths and determine how many of them are serially connected in a single pathway. To do this, we define a graph theory topological metric called *closeness* (CL). The $0 \leq CL(X, Y) \leq 1$ parameter for quantifying the closeness of two molecules X and Y in a network is the inverse of the distance $d(X, Y)$ between the two molecules, defined as the length of the shortest path between the two molecules in the network graph:

$$CL(X, Y) = 1/d(X, Y). \tag{3.12}$$

Some examples of how to calculate the closeness parameter are presented in Figure 3.8.

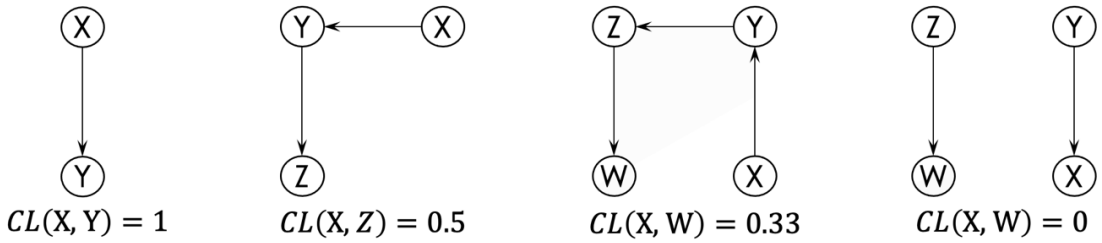


Figure 3.8 Numerical examples of the closeness parameter CL between two molecules.

3.4.1 Algorithm for Determining the Number of Required Clock Cycles to Compute Vulnerability Levels

In this section, to determine the maximum number of clock cycles needed for computing the vulnerability levels in a network, we propose the following algorithm with four steps:

I. Identify the feedback paths and the nodes that initiate the feedbacks, in the network. If they are not known a priori, identify them by finding the loops in the network graph, using a graph algorithm such as the depth-first search algorithm (Cormen et al., 2009).

II. Assume there exist F feedback paths in the network. Arbitrarily label the feedback initiating nodes as f_i, $i = 1, \dots, F$. Then start with $i = 1$, by calculating the closeness of f_1 with respect to the other feedback initiating nodes f_js, $j \neq i$ and $j = 1, \dots, F$. If $CL(f_1, f_j) = 0$ for all j values, it means f_1 is on no other pathway with other feedback initiating nodes, and now f_2 has to be examined similarly, i.e., $i = 2$, $j \neq i$ and $j = 1, \dots, F$, and so forth. However, if $CL(f_1, f_j) \neq 0$ for $j = j_0$, then this indicates that there is a path between f_1 and f_{j_0}, i.e., the feedback initiating nodes f_1 and f_{j_0} are on the same pathway. This finding needs to be pursued in the next step.

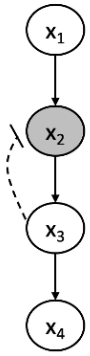
III. For fixed i and j values, e.g., $i = 1$ and $j = j_0$, calculate $CL(f_j, f_k)$ for all other feedback initiating nodes f_ks, $k \neq i, j$, and $k = 1, \dots, F$. Depending on the CL being non-zero or zero, it can be identified if f_k is on the same pathway that includes both f_i and f_j feedback initiating nodes or not.

IV. Repeat Step III until all feedback initiating nodes are examined.

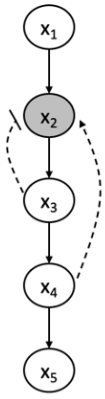
Using the information obtained from executing the above steps, the algorithm finds the pathway that contains the largest number of feedback initiating nodes on it, in series. With this specific number being called L, then at most $L + 1$ clock cycles are needed for computing the vulnerability levels.

In Figure 3.9, we compute vulnerability levels in two toy networks that have different number of feedback paths, to describe the relation between *F* and vulnerabilities. Note that since each network has a single pathway, we have $L = F$ in both networks.

A



C



B

$x_1(t)$ $t=1,2,3,4$	$x_2(t)$ $t=1,2,3,4$	$x_4(t)$ $t=1$	$t=2$	$t=3$	$t=4$
0	Fault free	0	0	0	0
1		1	0	1	0
0	Faulty (SA0)	0	0	0	0
1		0	0	0	0
0	Faulty (SA1)	1	1	1	1
1		1	1	1	1
Vulnerability of x_2 (CC = 1)		0.5			
Vulnerability of x_2 (CC = 2)		0.75			
Vulnerability of x_2 (CC = 3)		0.75			
Vulnerability of x_2 (CC = 4)		0.75			

D

$x_1(t)$ $t=1,2,3,4$	$x_2(t)$ $t=1,2,3,4$	$x_5(t)$ $t=1$	$t=2$	$t=3$	$t=4$
0	Fault free	0	0	0	0
1		1	0	1	0
0	Faulty (SA0)	0	0	0	0
1		0	0	0	0
0	Faulty (SA1)	1	1	1	1
1		1	1	1	1
Vulnerability of x_2 (CC = 1)		0.5			
Vulnerability of x_2 (CC = 2)		0.75			
Vulnerability of x_2 (CC = 3)		0.75			
Vulnerability of x_2 (CC = 4)		0.75			

Figure 3.9 Toy networks illustrating the number of clock cycles needed for the erroneous signal of a dysfunctional molecule to show its full effect at the network output. (A) Toy network with one feedback path. (B) Output truth table for fault-free and faulty x_2. (C) Toy network with two feedback paths. (D) Output truth table for fault-free and faulty x_2.
Note: The dashed lines ending in an arrowhead or a blunt line represent positive and negative feedback, respectively.

The first toy network (Figure 3.9A) has four nodes: $x_1(t)$ is the input node (molecule), the intermediate nodes are $x_2(t) = x_1(t) \times (\sim x_3(t-1))$ and $x_3(t) = x_2(t)$, and $x_4(t) = x_3(t)$ is the output node, where "×" is used for the AND operation and "~" is used for the NOT operation, and $x_3(0) = 0$. Herein, the node x_3 initiates a negative feedback to the node x_2. Since there is only one feedback path in the network, $F = 1$, at most two clock cycles are enough, $F + 1 = 2$, to observe the full error effect at the output.

To demonstrate this, we compute the vulnerability level of the node x_2 (Figure 3.9B), for different number of clock cycles CC $= 1, 2, 3, 4$, using Equation (3.4) in Section 3.2. We observe that the vulnerability of x_2 with 1 clock cycle is 0.5 and it becomes 0.75 with 2 clock cycles, and then remains at 0.75 with 3 and 4 clock cycles. These indicate that the full vulnerability of x_2 is 0.75 that is determined by analyzing the network having feedback for two clock cycles $(F + 1 = 2)$. In other words, two clock cycles are needed for the erroneous x_2 signal values to show their full effects at the network output x_4. Additionally, more clock cycles are not needed.

The second toy network (Figure 3.9C) has five nodes: $x_1(t)$ is the input node, $x_2(t) = (x_1(t) + x_4(t-1)) \times (\sim x_3(t-1))$, $x_3(t) = x_2(t)$ and $x_4(t) = x_3(t)$ are the intermediate nodes, and $x_5(t)$ is the output node, where "+" is used for the OR operation, and $x_3(0) = x_4(0) = 0$. Here the nodes x_3 and x_4 initiate a negative feedback and a positive feedback to the node x_2, respectively. Since there are two feedback paths in the network, $F = 2$, at most three clock cycles are needed, $F + 1 = 3$, to observe the full error effect of x_2 at the output. The computed vulnerability level of x_2 (Figure 3.9D) for different number of clock cycles corroborates what we stated earlier in this section, that is, $F + 1$ is indeed an upper bound and a smaller number of clock cycles may be needed for computing vulnerabilities in a network with feedbacks. In fact, we observe that the full vulnerability of x_2 is 0.75, obtained using 2 clock cycles only, and analyzing the network for the upper bound of $F + 1 = 3$ clock cycles is not needed (Figure 3.9D). In other words, 2 clock cycles are enough for errors originated from x_2 to show their complete effects at the output x_5.

3.4.2 ERBB Signaling Network – Vulnerability and the Number of Clock Cycles

In this section, we apply the proposed algorithm in Methods Section C, to the ERBB signaling network (Figure 3.4). We start by identifying feedbacks in the network. Given the small size of the network, visual inspection of the network reveals five loops, which topologically may contain the indicators of feedback interactions. The five loops are (i) IGF1R → AKT1 → IGF1R, (ii) IGF1R → MEK1 → ER-α → IGF1R, (iii) CDK4 —● p27 —● CDK4, (iv) CDK4 —● p21 —● CDK4, and (v) CDK2 —● p27 —● CDK2. Despite the existence of five loops, there are only four feedback initiating nodes, AKT1, ER-α, p21 and p27, because p27 is common between the two loops, i.e., the loops (iii) and (iv). Note that in the absence of prior information about the feedbacks, the feedback initiators may not be uniquely determined within the identified loops. For example, based on these five loops, one can identify IGF1R, MEK1, CDK4, and CDK2 as feedback initiators as well. Nevertheless, different choices for the feedback initiating molecules within these loops do not affect the algorithm that aims at finding the pathway that contains the largest number of feedback initiating nodes on it, in series. This is because if the feedback initiating nodes f_i and f_j chosen from two loops are connected through a pathway, i.e., $CL(f_i, f_j) \neq 0$, then other choices of the feedback initiating nodes f_i' and f_j' from the said two loops will be connected through a pathway as well, i.e., $CL(f_i', f_j') \neq 0$. Thus, the algorithm to determine L is independent of the choice of the feedback initiating nodes.

Using the identified feedback initiators, the algorithm outputs the upper bound of $L + 1 = 5$ clock cycles that may be needed for computing the vulnerability level of a molecule. This is because the algorithm identifies a pathway that contains all the feedback initiating molecules on it, in series. For instance, AKT1 → ER-α → p27 —● CDK4 —● p21

—● CDK2 → pRB is a pathway that contains all the feedback initiating molecules on it. Through this specific pathway, erroneous signals originated from an upstream molecule of AKT1 may require five clock cycles to show their full effect on the output molecule pRB, due to the signal propagation delays introduced by the feedback paths connected in series on the same pathway. Computed using Equation (3.4), Figure 3.10 presents the single-fault vulnerability levels versus the number of clock cycles CC for some molecules in the ERBB signaling network. We observe that the vulnerability levels of the molecules can be computed in less than five clock cycles, which confirms that it is sufficient to simulate and analyze the network for at most five clock cycles, as specified by the algorithm.

3.4.3 T Cell Signaling Network – Vulnerability and the Number of Clock Cycles

In this section, we apply the proposed algorithm in Methods, Section C to the T cell signaling network (Figure 3.6). In the network, first we identify four feedback initiating

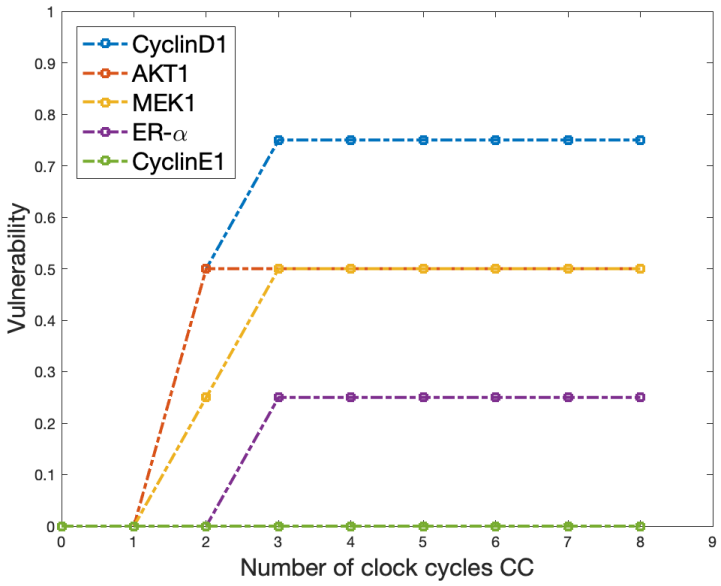


Figure 3.10 Vulnerability versus the number of clock cycles CC for some molecules in the ERBB signaling network.

nodes that are shp1, ccblp1, pag, and gab2, using the time indices in Table A.2 (see Appendix) and also by visual inspection of Figure 3.6. After using the proposed algorithm, we obtain the upper bound of $L + 1 = 5$ clock cycles, because there is one pathway that contains all the four feedback initiating nodes in series. Next, we compute the single-fault vulnerability levels of some molecules versus the number of clock cycles CC (Figure 3.11A), with cre considered as the output molecule, and using Equation (3.4). We observe that the vulnerability levels of the molecules can be computed in less than five clock cycles, which confirms that it is sufficient to simulate and analyze the network for at most five clock cycles, as determined by the algorithm.

Furthermore, we compute the double-fault vulnerability values of some pairs of molecules versus the number of clock cycles CC (Figure 3.11B), with cre considered as the output molecule, and using Equation (3.4). As anticipated by the algorithm, the vulnerability levels of the molecular pairs can be computed in less than five clock cycles, i.e., it is enough to simulate and analyze the network for at most five clock cycles, irrespective of considering single faults or double faults.

A noteworthy point is that when two molecules are faulty at the same time, more clock cycles may be needed to observe the aggregated full effects of the two erroneous signals at the network output, compared to single faults. However, the upper bound found by the algorithm still works for both scenarios. This is because the upper bound depends only on the topological positions of the feedback initiating nodes and the connections among them, and not on how many nodes are grouped together, to represent a group of faulty molecules. As some numerical examples, consider the scenario of zap70 and slp76 being individually faulty (Figure 3.11A), where three and one clock cycles are needed,

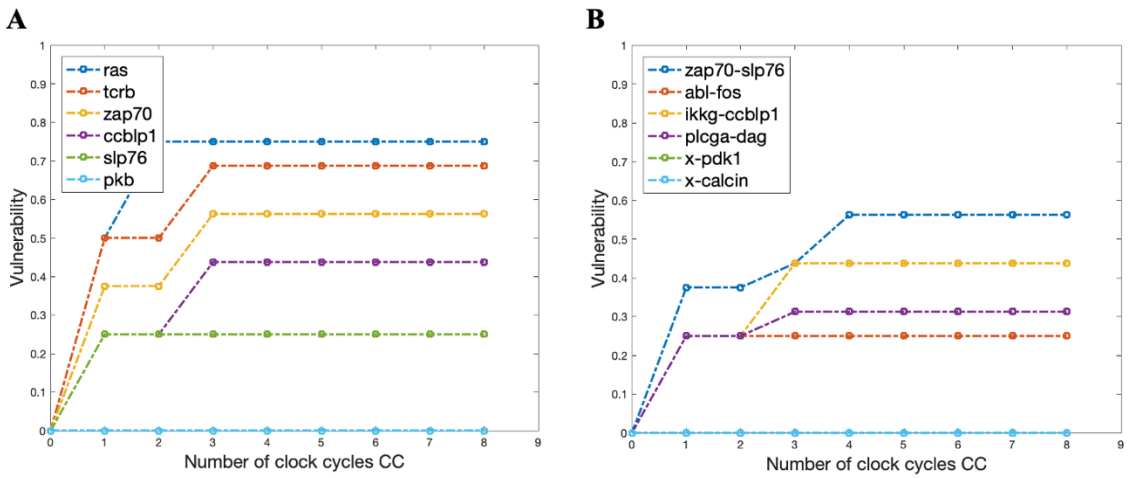


Figure 3.11 Vulnerability versus the number of clock cycles CC for some single and pairs of molecules in the T cell signaling network, with "cre" considered as the output molecule. (A) Single-fault vulnerability levels. (B) Double-fault vulnerability levels.

respectively, to compute their full single vulnerabilities of 0.56 and 0.25, respectively. On the other hand, when they are simultaneously faulty (Figure 3.11B), four, i.e., more clock cycles are needed to compute their full double vulnerability of 0.56. This indicates that if multiple molecules are faulty concurrently, then there may be further delays in observing the entire effects of multiple erroneous signals, propagating via various pathways towards the network output. Additionally, we observe that the upper bound of $L + 1 = 5$ clock cycles hold true for computing both single and double vulnerabilities.

CHAPTER 4

MODELING AND MEASUREMENT OF SIGNALING OUTCOMES AFFECTING CELL DECISION MAKING

In the previous chapters, we showed methods for training molecular network models against experimental data and for doing vulnerability analysis to calculate probability of an entire network being failed when one or multiple molecules are faulty (abnormal). Typically, the networks have major roles in characterization of cell fate. For instance, depending on the received signals and dynamics of the network of interest, a possible cell fate could be transcribing RNA and initiate a certain process or proliferating or initiating apoptosis or moving in a certain direction, and so on. Therefore, characterization of decision makings in cells in response to received signals is important for understanding how cell fate is determined in the absence and presence of such abnormalities causing incorrect network responses. In this chapter, we initiate methods and mathematical frameworks to calculate the probability of such decisions by modeling signaling outcomes of the cells using *in silico* single cell data (Ozen et al., 2020).

4.1 A Case Study: Signaling Outcomes and Decisions in the p53 System When DNA Damage Occurs

The transcription factor p53 has a significant role in DNA repair, cell cycle suppression, regulation of cell growth, and initiation of apoptosis (Elmore, 2007; Kastan et al., 1991; Levine, 1997; Vousden & Prives, 2009). It becomes active in response to DNA damage that may occur when the cell is exposed to ionizing radiation (IR), ultraviolet (UV) radiation, heat shock, etc. (Siliciano et al., 1997; Vogelstein et al., 2000; Vousden & Prives,

2009). In particular, exposure to IR results in DNA double strand breaks (DSBs), which are the most serious DNA lesion. When DSB is not repaired, it can cause cell death or DNA mutations which can propagate to new cell generations (Lliakis, 1991; Rothkamm et al., 2003; Vilenchik & Knudson, 2003). When DNA damage occurs, p53 can assume two phosphorylation states: $p53_{Arrester}$ and $p53_{Killer}$. Afterwards, the p53 system can take two actions: it either suppresses the cell cycle until DNA is repaired, if the damage is low and repair is possible; or it can trigger apoptosis if the damage is high and repair is not possible (Elmore, 2007; Rothkamm et al., 2003; Alberts et al., 2002). Herein, we intend to compute decision thresholds and incorrect decision rates when the DNA damages caused by various IR doses occur in a cell. With this goal in mind, we conduct stochastic simulations of cells exposed to different IR doses as shown in Hat et al. (2016) to obtain *in silico* single cell data.

Consider the p53 system model of Hat et al. (2016) shown in Figure 4.1. The p53 system is activated due to a DNA damage induced by IR. Initially the protein kinase ataxia-telangiectasia mutated (ATM) is activated by the DNA damage (Bakkenist & Kastan, 2003; Saito et al., 2002). The active ATM phosphorylates Mdm2, which is a p53 inhibitor (Maya et al., 2001). The ATM also activates p53 by phosphorylating it to one of its active phosphoforms: $p53_{Arrester}$ which further phosphorylates p53 to the $p53_{Killer}$ form (Banin et al., 1998; Canman et al., 1998; Shieh et al., 1999). Moreover, the $p53_{Arrester}$ activates the Mdm2 (Barak et al., 1993) and wild-type p53-induced phosphatase 1 (Wip1) (Choi et al., 2002; Fiscella et al., 1997). The active Wip1 inhibits the ATM (Shreeram et al., 2006) and dephosphorylates the $p53_{Killer}$ to the $p53_{Arrester}$ form (Takekawa et al., 2000). The $p53_{Killer}$

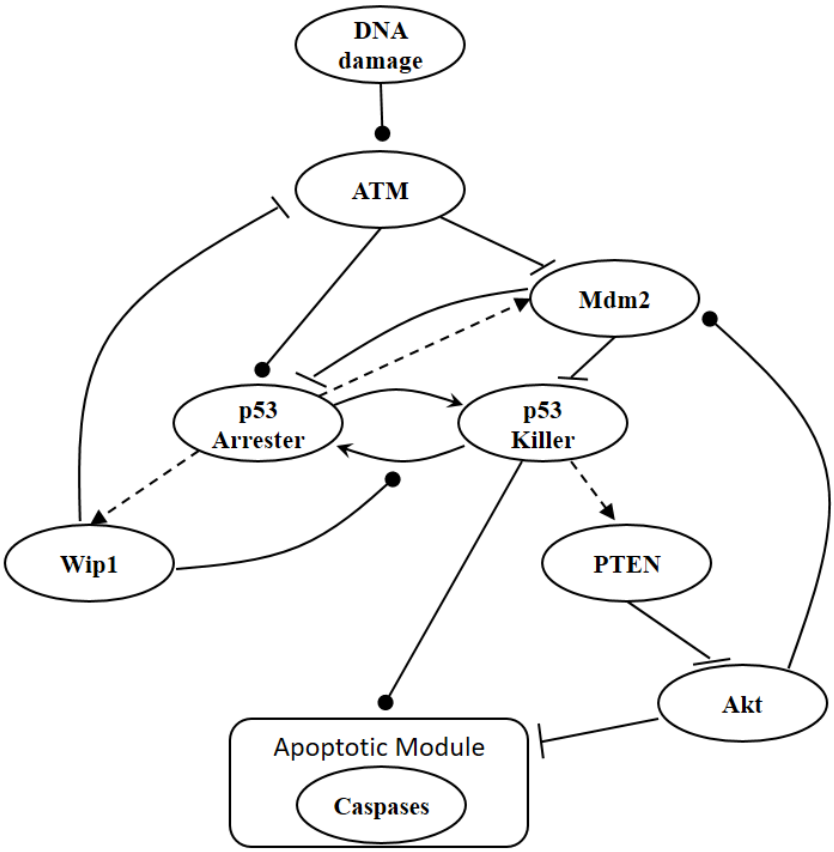


Figure 4.1 A p53 system model.
Note: The figure is generated based on Figure 1 of Hat et al. (2016). Arrow-headed dashed lines represent positive transcriptional regulations, arrow-headed solid lines stand for protein transformations, circle-headed solid lines are activatory regulations, and hammer-headed solid lines represent inhibitory regulations. All molecules and the interactions between them are described in the main body of the chapter.

regulates another phosphatase, phosphatase and tensin homolog (PTEN), which initiates a slow positive feedback loop stabilizing the level of p53 (Stambolic et al., 2001). If DNA damage is large and its repair takes longer time, PTEN accumulates to high levels and inhibits AKT, which may no longer phosphorylate Mdm2. Unphosphorylated Mdm2 remains in cytoplasm and may not target nuclear p53 for degradation. Thus, accumulation of PTEN results in disconnection of the negative feedback loop between p53 and Mdm2. The slow positive feedback loop acts as a clock giving cells time to repair DNA and

initiating apoptosis if DNA repair takes too long. The apoptotic module, where transcription of pro-apoptotic proteins is induced, is controlled by $p53_{Killer}$ and Akt that suppresses the apoptosis. When Akt is inhibited by increased level of PTEN, it will no longer suppress the apoptotic module. Thus, the $p53_{Killer}$ will initiate activation of cysteine-aspartic proteases (Caspases), enzymes having essential roles in cell death (Figure 4.1). Since we are interested in the analysis of the signaling outcomes which affect whether the cell survives or triggers apoptosis, we do not consider the cell cycle arrest module (regulated by $p53_{Arrester}$) and focus on the apoptotic module. Simulation files can be found in Hat et al. (2016) and more detailed information about the p53 system and each component and interaction there can be found in Hat et al. (2016) and Bogdał et al. (2013). More specifically, interested readers can refer to the Supporting Information S1 Text of Hat et al. (2016), which includes a summary of mathematical models of the p53 system, a detailed description of the model, a notation guide, and lists of parameters and reactions.

4.2 Decision Making and Outcome Analysis: Hypothesis Testing on Input Signals and Optimal Decisions with Minimum Errors

When cells are exposed to radiation, each cell may respond differently due to noise or some other factors. One may decide to survive, whereas another may trigger apoptosis, both under the same IR dose. Given the probabilistic nature of such decisions (Habibi et al., 2017), we can formulate p53-based decision making as a binary hypothesis testing problem, where the decision-making system is going to test which of the following two hypotheses is true regarding the applied IR dose, to trigger an action accordingly:

$$\begin{aligned} &\mathrm{H}_0\text{: IR dose is low,} \\ &\mathrm{H}_1\text{: IR dose is high.} \end{aligned} \tag{4.1}$$

Binary hypothesis testing is observed in other systems, e.g., the TNF/NF-κB system presented in Habibi et al. (2017).

In response to an IR dose, two types of incorrect decisions can be made. One is deciding that the input IR level is high, whereas in fact it is low (deciding H_1 when H_0 is true), which may falsely trigger apoptosis. The other one is deciding that the input IR level is low, whereas in fact it is high (deciding H_0 when H_1 is true), which may result in missing apoptosis. These two erroneous decisions can be called as *false alarm* and *miss event*, respectively, and their probabilities can be defined as:

$$\begin{aligned} P_{FA} &= P(\text{deciding } \mathrm{H}_1|\mathrm{H}_0), \\ P_M &= P(\text{deciding } \mathrm{H}_0|\mathrm{H}_1). \end{aligned} \tag{4.2}$$

The overall error probability P_E of making decisions is a combination of P_{FA} and P_M:

$$P_E = P(\mathrm{H}_0)P_{FA} + P(\mathrm{H}_1)P_M, \tag{4.3}$$

where $P(\mathrm{H}_0)$ and $P(\mathrm{H}_1)$ are prior probabilities of H_0 and H_1, respectively. Note that as mentioned in the Introduction section, IR causes DNA damage. Therefore, one can instead formulate the p53-based decision-making process as a binary hypothesis testing on DNA

damage being low or high and define the associated false alarm and miss events probabilities accordingly.

The optimal decision-making system which minimizes the above P_E is the one that compares probabilities of observed data under the hypotheses H_0 and H_1 (Kay, 1998). More precisely, suppose that x is the observation and $p(x|H_0)$ and $p(x|H_1)$ are the conditional probability density functions (PDFs) of x under H_0 and H_1, respectively. Also consider equi-probable hypotheses, i.e., $P(H_0) = P(H_1) = 0.5$, which is a reasonable assumption in the absence of prior information on the possibilities of H_0 and H_1. Then, the optimal system decides H_1 if $p(x|H_1) > p(x|H_0)$, otherwise, it decides H_0. This means that the hypothesis with the highest likelihood is decided. This decision is called the maximum likelihood decision (Kay, 1998).

4.3 Single Cell Data of the p53 System Exposed to Ionizing Radiation

To calculate the error probabilities in Equation (4.2), we use PTEN level as the decision variable because when unrepairable DNA damage occurs, the activated p53 triggers pro-apoptotic phosphatase PTEN (Stambolic et al., 2001), and PTEN initiates apoptosis (Hlobilkova et al., 2006). It has also been shown by Hat et al. (2016) that PTEN is a decent predictor of cell fate. After specifying the decision variable, we use the stochastic simulator of Hat et al. (2016) to generate 5000 cells for each IR dose. The stochastic simulation has three phases. The first phase is the "equilibrium phase" where we simulate 2 weeks of cell behavior when no IR dose is applied. The second phase is called "irradiation phase" in which 10 minutes of IR dose is applied. The last phase is called "relaxation phase" in which we simulate 72 hours of cell behavior after it is exposed to 10 minutes of IR. When IR dose

increases, apoptotic cell percentage increases as well as shown in Figure 4.2 (Hat et al., 2016). For more details on the simulation phases, see the supporting files of Hat et al. (2016). In order to decide whether a cell is apoptotic or not, we check the active caspase level in 72 hours after the irradiation phase and compare it with the threshold of 0.5×10^5 suggested in Hat et al. (2016). Cells with the level of active caspase higher (or lower) than the threshold of 0.5×10^5 are considered to be apoptotic (or surviving).

The data of normal cells includes eight sets of PTEN levels in 5000 cells, which correspond to eight doses of IR = 1, 2, 3, 4, 5, 6, 7, and 8 Gy. Here Gy stands for Gray, the unit of radiation dose, and 1 Gy is 1 Joule of energy absorbed by 1 kg of tissue. We focus our analysis on low IR versus high IR hypothesis testing, to see how accurately it can be decided whether the applied radiation level is low or high. We consider IR = 1 Gy as the low dose, whereas the higher dose can be IR = 2, 3, 4, 5, 6, 7, or 8 Gy. More specifically, scenarios in which signaling outcomes are analyzed are 1 vs. 2 Gy, 1 vs. 3 Gy, 1 vs. 4 Gy, 1 vs. 5 Gy, 1 vs. 6 Gy, 1 vs. 7 Gy, and 1 vs. 8 Gy. We quantitatively study in which of these scenarios more erroneous decisions are made. We also determine to what extent the decision between responses to low and high IR levels depends on the input IR separation. We conduct these studies by computing the optimal decision threshold in each scenario using the PTEN data, following the maximum likelihood principle that provides the best decision, i.e., the smallest decision error probabilities. We also compute numerical values of the decision error probabilities using the PTEN data.

In addition to the analysis of erroneous decision making and incorrect signaling outcomes in normal cells mentioned above, we analyze them in abnormal cells as well, where there is a dysfunctional molecule in the p53 system. Wip1 is one of the key

regulatory pro-survival phosphatases (Fiscella et al., 1997) in the p53 system (Figure 4.1). If the DNA damage can be repaired, then Wip1 expression returns the cell to the pre-stress state from cell-cycle arrest (Fiscella et al., 1997; Lu et al., 2007). It has been observed that elevated Wip1 level exists in multiple human cancer types such as breast, lung, pancreas, bladder, and liver cancer (Bulavin et al., 2002; Castellino et al., 2007; Hirasawa et al., 2003; Li et al., 2002; Rauta et al., 2006; Saito-Ohara et al., 2003). Therefore, to obtain abnormal cells, we generate cells with increased Wip1 synthesis rate. In normal cells, Wip1 synthesis rate is about 0.1 (Hat et al., 2016), and here we increase it to 0.15, a 50% increase, to reproduce abnormality. This increase in the Wip1 synthesis rate causes a significant decrease in the cell death percentage (Figure 4.2), which can be considered as an abnormal cell state. In addition to Wip1, we analyze abnormal cellular state caused by PTEN abnormalities. It has been observed that attenuated PTEN levels exist in MCF-7, a non-invasive form of human breast cancer cells (Geva-Zatorsky et al., 2006). Therefore, it is of interest to see how the abnormal PTEN level affects signaling outcomes in the p53 system. To study this, we generate abnormal cells by decreasing PTEN synthesis rate. In healthy cells, the PTEN synthesis rate is about 0.03 (Hat et al, 2016). Here we decrease it to 0.015, a 50% decrease, to reproduce abnormality. We observe a considerable decrease in the cell death percentage (Figure 4.2), representing an abnormal cellular state.

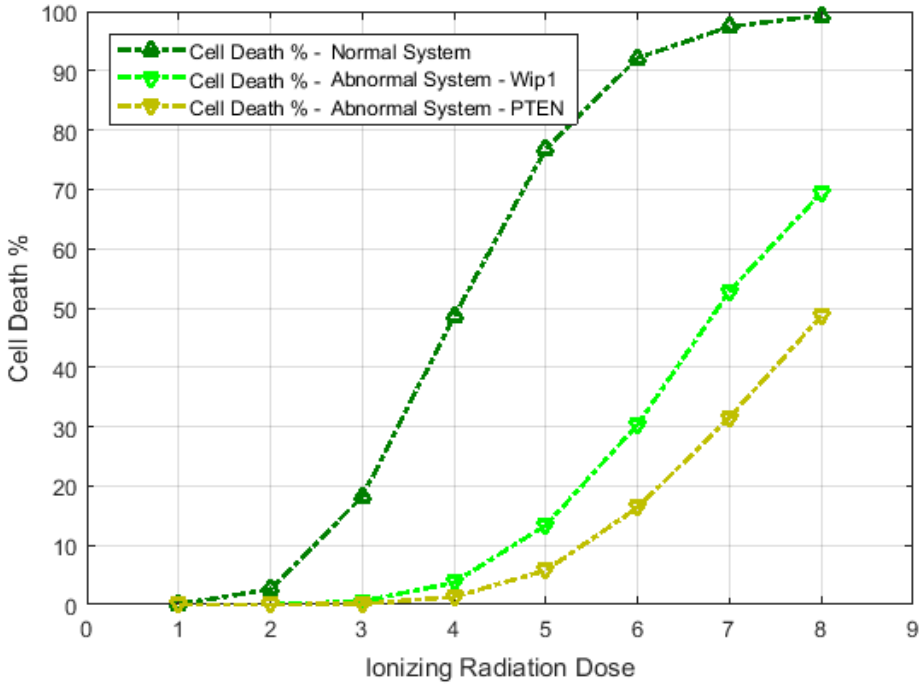


Figure 4.2 Cell death percentage versus ionizing radiation (IR) dose in both normal and abnormal p53 systems.

Note: The dark green curve at the top represents a normal p53 system with no perturbation, whereas the other two curves correspond to p53 systems behaving abnormally due to Wip1 or PTEN perturbations.

CHAPTER 5

UNIVARIATE CELL DECISION MAKING ANALYSIS

In Chapter 4, we introduced the p53 network, the hypotheses H_0 and H_1, and explained how the *in silico* single cell data was obtained and why PTEN is chosen as the decision variable. In what follows, we introduce methods for computing decision error probabilities using "single" time point measurements in the wild-type cells (Ozen et al., 2020).

5.1 Methods for Computing Decision Thresholds and Decision Error Rates Using Single Time Point Measurements in Individual Wild-type Cells

In this section, we analyze PTEN levels of 5000 cells measured in 72 hours after the irradiation phase. It has been observed that PTEN levels of both apoptotic and surviving cells become very distinct in 72 hours after 10 minutes of IR application (Hat et al., 2016) (decision analysis based on PTEN levels at other time instants, as well as multiple time instants are presented in other sections).

Histograms of natural logarithm, ln, of PTEN levels for IR = 1, 2 Gy data sets and IR = 1, 8 Gy data sets are shown in Figure 5.1A and Figure 5.1C, respectively. As presented in Figure 5.1B and Figure 5.1D, Gaussian PDFs whose means and variances are estimated from the data, reasonably represent the histograms. This indicates that the PTEN data can be reasonably approximated by lognormal PDF. Due to the mathematical convenience of working with Gaussian PDFs and variables, especially for the multivariate analysis of multiple time point data discussed later, we continue working with the logarithm of the PTEN data.

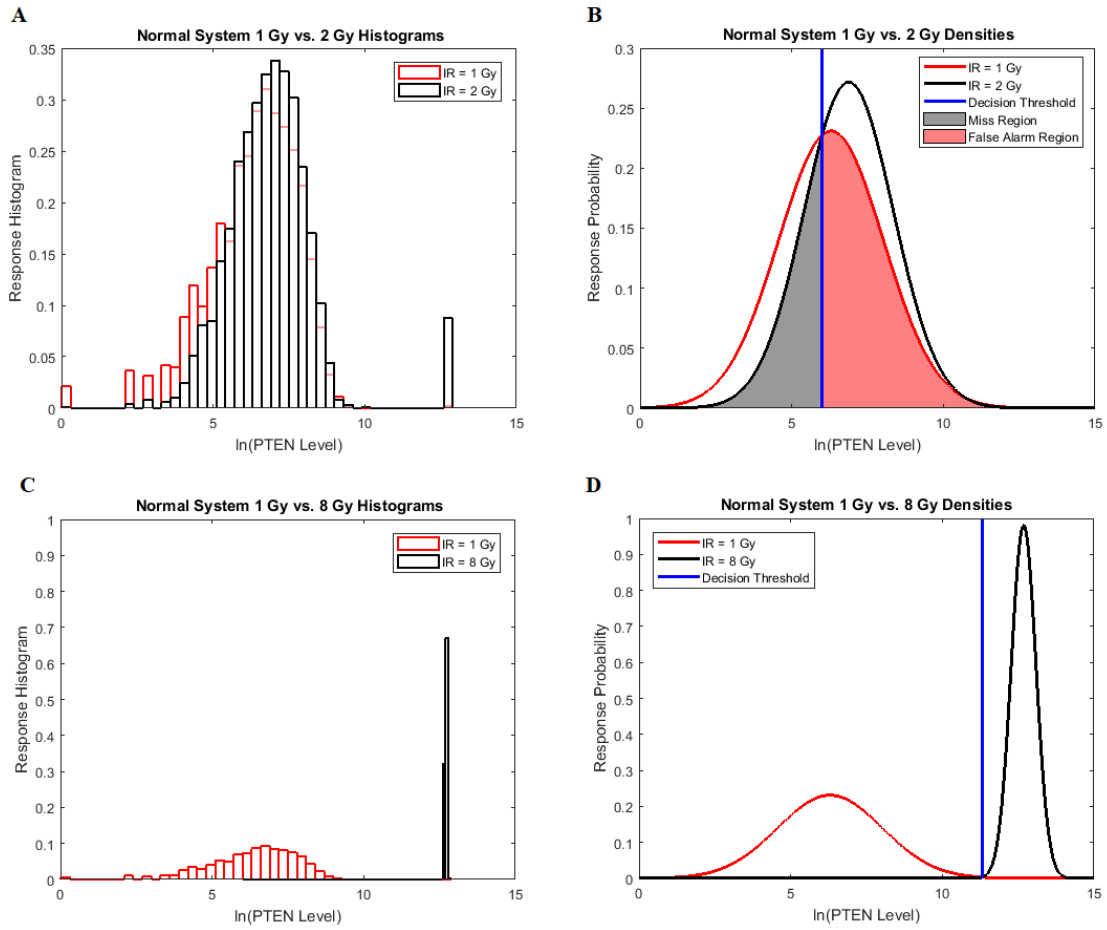


Figure 5.1 Univariate decision making and signaling outcome analysis in the normal p53 system based on PTEN response distributions. (A) Histograms of PTEN levels of cells under IR = 1 Gy and IR = 2 Gy doses. (B) Gaussian probability density functions (PDFs) for PTEN levels of cells under IR = 1 Gy and IR = 2 Gy doses, together with the optimal maximum likelihood decision threshold which minimizes the total decision error probability. (C) Histograms of PTEN levels of cells under IR = 1 Gy and IR = 8 Gy doses. (D) Gaussian PDFs for PTEN levels of cells under IR = 1 Gy and IR = 8 Gy doses, together with the optimal maximum likelihood decision threshold which minimizes the total decision error probability.

Let $x = \ln(\text{PTEN})$ be the Gaussian variable of interest with mean μ and variance σ^2, i.e., $x \sim \mathcal{N}(\mu, \sigma^2)$ where $\mathcal{N}$ stands for the following normal or Gaussian PDF:

$$p(x) = (2\pi\sigma^2)^{-1/2} \exp[-(x-\mu)^2/(2\sigma^2)\,].$$

The Gaussian PDFs shown in Figure 5.1 are indeed the conditional PDFs $p(x|\mathrm{H}_0)$ and $p(x|\mathrm{H}_1)$ under the hypotheses H_0 and H_1 defined earlier in Equation (4.1). For example, in Figure 5.1B, H_0 and H_1 correspond to IR = 1 Gy and IR = 2 Gy doses, respectively, and the red and black curves in there are the conditional PDFs $p(x|\mathrm{H}_0)$ and $p(x|\mathrm{H}_1)$, respectively.

5.1.1 The Optimal Maximum Likelihood Decision Making System, the Optimal Decision Threshold, and the Decision Error Probabilities

Recall our two hypotheses previously defined in Equation (4.1). The optimal decision maker, which minimizes the overall error probability P_E in Equation (4.3), compares the conditional likelihood ratio $L(x) = p(x|\mathrm{H}_1)/p(x|\mathrm{H}_0)$ with the ratio $\gamma = P(\mathrm{H}_0)/P(\mathrm{H}_1)$ (Habibi et al., 2017). The system decides H_1 if $L(x) > \gamma$. If the hypotheses are equi-probable, i.e., $P(\mathrm{H}_0) = P(\mathrm{H}_1) = 0.5$, then the optimal system decides H_1 if $p(x|\mathrm{H}_1) > p(x|\mathrm{H}_0)$.

To find the optimal decision threshold, we need to solve the equation $L(x) = \gamma$, i.e., $P(\mathrm{H}_1)p(x|\mathrm{H}_1) = P(\mathrm{H}_0)p(x|\mathrm{H}_0)$, for x. When H_0 and H_1 re equi-probable, the threshold equation to be solved simplifies to $L(x) = 1$, i.e., $p(x|\mathrm{H}_1) = p(x|\mathrm{H}_0)$. Once the optimal decision threshold is determined, it can be used to compute false alarm and miss decision error probabilities, by integrating the conditional PDFs of data over error regions. More specifically, using the conditional PDFs $p(x|\mathrm{H}_0)$ and $p(x|\mathrm{H}_1)$ representing the response probabilities of the ln of PTEN levels under the two hypotheses, Equation (4.2) can be written as (Habibi et al., 2017):

$$P_{FA} = \int_{x \in \text{false alarm region}} p(x|\mathrm{H}_0)dx, \tag{5.1}$$

$$P_M = \int_{x \in \text{miss region}} p(x|\mathrm{H}_1)dx. \tag{5.2}$$

The false alarm region in Equation (5.1) is defined by $\{x: p(x|\mathrm{H}_1) > p(x|\mathrm{H}_0)\}$ when H_0 is true, whereas the miss region in Equation (5.2) is defined by $\{x: p(x|\mathrm{H}_0) > p(x|\mathrm{H}_1)\}$ when H_1 is true. By substituting P_{FA} and P_M in Equation (4.3), the overall error probability P_E can be obtained.

5.1.2 Gaussian Data Model to Compute the Optimal Decision Threshold and Decision Error Probabilities

Here we focus on Figure 5.1B as an example, where two Gaussian PDFs are shown for $x = \ln(\text{PTEN})$, the natural logarithm of PTEN levels in the two data sets of IR = 1 Gy and IR = 2 Gy, with each data set consisting of 5000 cells. Let $\mathcal{N}(\mu_0, \sigma_0^2)$ and $\mathcal{N}(\mu_1, \sigma_1^2)$ represent the Gaussian PDFs that correspond to the IR = 1 Gy and IR = 2 Gy data sets, respectively, where (μ_0, σ_0^2) and (μ_1, σ_1^2) are mean/variance pairs estimated from their associated data sets. The optimal maximum likelihood decision threshold in Figure 5.1B is at the intersection of the two PDFs, and can be computed by solving the equation $p(x|\mathrm{H}_1) = p(x|\mathrm{H}_0)$ written below:

$$(2\pi\sigma_0^2)^{-1/2} \exp[-(x-\mu_0)^2/(2\sigma_0^2)] = (2\pi\sigma_1^2)^{-1/2} \exp[-(x-\mu_1)^2/(2\sigma_1^2)] \tag{5.3}$$

By multiplying both sides by $(2\pi\sigma_0^2)^{1/2}\exp[(x-\mu_1)^2/(2\sigma_1^2)]$ and then taking natural logarithm of both sides, Equation (5.3) can be written in the following quadratic equation form (Habibi et al., 2017):

$$(\sigma_0^2-\sigma_1^2)x^2+2(\sigma_1^2\mu_0-\sigma_0^2\mu_1)x+\sigma_0^2\mu_1^2-\sigma_1^2\mu_0^2-2\sigma_0^2\sigma_1^2\ln\left(\frac{\sigma_0}{\sigma_1}\right)=0 \qquad (5.4)$$

Equation (5.4) is derived assuming our hypotheses are equi-probable, i.e., $P(\mathrm{H}_0)=P(\mathrm{H}_1)=0.5$, as mentioned before. The solution of Equation (5.4) gives the optimal decision threshold $PTEN_{\mathrm{th}}$, located at the intersection of the two PDFs for IR = 1 Gy and IR = 2 Gy doses in Figure 5.1B (the italic style is adopted to clarify that the threshold is related to the logarithm of PTEN data). Interestingly, for equal variances, solution of Equation (5.4) for the optimal decision threshold simplifies to the average of the means, i.e., $(\mu_0+\mu_1)/2$, which intuitively makes sense. For other prior probabilities and PDF models, the optimal threshold can be obtained similarly, by solving the equation $P(\mathrm{H}_0)p(x|\mathrm{H}_0)=P(\mathrm{H}_1)p(x|\mathrm{H}_1)$, for x.

Using the $PTEN_{\mathrm{th}}$ obtained by solving Equation (5.4) and using the Gaussian PDFs, Equations (5.1) and (5.2) for the false alarm and miss error probabilities can be written as:

$$P_{FA}=\int_{PTEN_{\mathrm{th}}}^{\infty}p(x|\mathrm{H}_0)dx=Q\left(\frac{PTEN_{\mathrm{th}}-\mu_0}{\sigma_0}\right), \qquad (5.5)$$

$$P_M=\int_{-\infty}^{PTEN_{\mathrm{th}}}p(x|\mathrm{H}_1)dx=Q\left(\frac{\mu_1-PTEN_{\mathrm{th}}}{\sigma_1}\right), \qquad (5.6)$$

where $\mathcal{Q}(\eta)$ is tail probability of the standard Gaussian PDF $\mathcal{N}(0,1)$:

$$\mathcal{Q}(\eta) = (2\pi)^{-1/2} \int_{\eta}^{\infty} \exp\left(-\frac{u^2}{2}\right) du.$$

Equation (5.5) represents area of the pink region in Figure 5.1B under the tail of the IR = 1 Gy PDF, beyond the $PTEN_{th}$ threshold. In this region of $x > PTEN_{th}$ we have $p(x|\mathrm{H}_1) > p(x|\mathrm{H}_0)$, while H_0 is true. This is the *false alarm* region for which we have computed $P_{FA} = 0.57$ in Figure 5.1B. On the other hand, Equation (5.6) represents area of the gray region in Figure 5.1B under the tail of the IR = 2 Gy PDF, below the $PTEN_{th}$ threshold. In this region of $x < PTEN_{th}$ we have $p(x|\mathrm{H}_0) > p(x|\mathrm{H}_1)$, while H_1 is true. This is the *miss* region for which we have computed $P_M = 0.28$ in Figure 5.1B. After computing P_{FA} and P_M, we can now compute the overall error probability P_E using Equation (4.3), which results in $P_E = (P_{FA} + P_M)/2 = 0.43$. Similarly, by computing Equations (5.5) and (5.6) for the 1 vs. 8 Gy scenario we obtain $P_E = 0.001$ (Figure 5.1D). Based on the results of 1 vs. 2 Gy and 1 vs. 8 Gy decision scenarios, it can be concluded that when the difference between the two applied IR doses increases, the overall decision error probability P_E decreases. This is mainly because the two response PDFs become more distinct with less overlap, as the difference between the two applied IR doses increases.

5.1.3 Mixture of Gaussian Data Model to Compute the Optimal Decision Threshold and Decision Error Probabilities for Some Low vs High IR Cases

For some cases such as 1 vs. 3, 4, 5 and 6 Gy IR doses, some data sets need to be modeled by a mixture of Gaussian PDFs due to the bistable behavior of p53 system and hence cells'

bimodal histograms. Still the same underlying theory and the proposed framework hold. Nevertheless, in what follows we explain how to determine the optimal decision thresholds and how to compute the decision error probabilities when using a mixture model for the 1 vs. 4 Gy scenario.

Histograms of natural logarithm of PTEN levels for IR = 1, 4 Gy data sets are shown in Figure 5.2A. We notice that while the 1 Gy data histogram is unimodal, histogram of 4 Gy data is bimodal. Therefore, for the 1 Gy data we use a single Gaussian PDF as before, whereas for the 4 Gy data we utilize a mixture of two Gaussian PDFs. More specifically, we consider $\mathcal{N}(\mu_0, \sigma_0^2)$ for H_0 to represent the single Gaussian PDF that corresponds to the IR = 1 Gy data, whereas we use $\xi\mathcal{N}(\mu_{11}, \sigma_{11}^2) + (1-\xi)\mathcal{N}(\mu_{12}, \sigma_{12}^2)$ for H_1, with $0 \leq \xi \leq 1$ being the mixing parameter, to represent the mixture of two Gaussian PDFs which correspond to the IR = 4 Gy data set. The mean and variance (μ_0, σ_0^2) are estimated from the 1 Gy data and the associated single Gaussian PDF is shown in Figure 5.2B. Furthermore, the means and variances $(\mu_{11}, \sigma_{11}^2)$ and $(\mu_{12}, \sigma_{12}^2)$ and the mixing parameter ξ are estimated from the 4 Gy data using the MATLAB command "fitgmdist" which implements the iterative Expectation-Maximization (EM) algorithm. The resulting mixture of two Gaussian PDFs is shown in Figure 5.2B.

Similar to the previous scenarios, the optimal maximum likelihood decision thresholds shown in Figure 5.2B for equi-probable hypotheses are at the intersections of the conditional PDFs $p(x|H_0)$ and $p(x|H_1)$, the latter being a Gaussian mixture for the 4 Gy data. Note that here solving the equation $p(x|H_1) = p(x|H_0)$ results in four solutions for x, that is why there are four decision thresholds, $PTEN_{thi}$, i = 1,2,3,4 in Figure 5.2B

(Note that each decision threshold $PTEN_{\text{thi}}$ is listed as "Decision Threshold i" in Figure 5.2).

To compute the decision error probabilities, the false alarm and miss probabilities P_{FA} and P_M need to be calculated using Equation (5.1) and Equation (5.2), respectively. Since there are four decision thresholds in this case, integration has to be performed over multiple regions, which results in lengthy expressions. However, note that as can be seen in Figure 5.2B and its zoomed-in view in Figure 5.2C, PDFs for low dose (red) and the lower Gaussian mode for the high dose (black) assume very small values as they reach the third threshold. Therefore, their contributions to possible error events around the third and fourth thresholds are negligible (later this is shown numerically). Similarly, given the very small variance of the higher Gaussian mode of the PDF for the high dose, this PDF is substantially different from zero only between the third and fourth thresholds. Consequently, the contribution of the PDF of this mode to possible errors around the third and fourth thresholds is negligible as well. Overall, as just explained, the optimal decision when $PTEN_{\text{th3}} < x < PTEN_{\text{th4}}$ is $\mathrm{H_1}$ with no decision error, whereas for $x < PTEN_{\text{th1}}$, $PTEN_{\text{th1}} < x < PTEN_{\text{th2}}$ and $PTEN_{\text{th2}} < x < PTEN_{\text{th3}}$, the optimal decisions are $\mathrm{H_0}$, $\mathrm{H_1}$ and $\mathrm{H_0}$, respectively, with the following decision error probabilities:

$$P_{FA} = Q\left(\frac{PTEN_{\text{th1}} - \mu_0}{\sigma_0}\right) - Q\left(\frac{PTEN_{\text{th2}} - \mu_0}{\sigma_0}\right),$$

$$P_M = \xi\left[Q\left(\frac{\mu_{11} - PTEN_{\text{th1}}}{\sigma_{11}}\right) + Q\left(\frac{PTEN_{\text{th2}} - \mu_{11}}{\sigma_{11}}\right)\right].$$

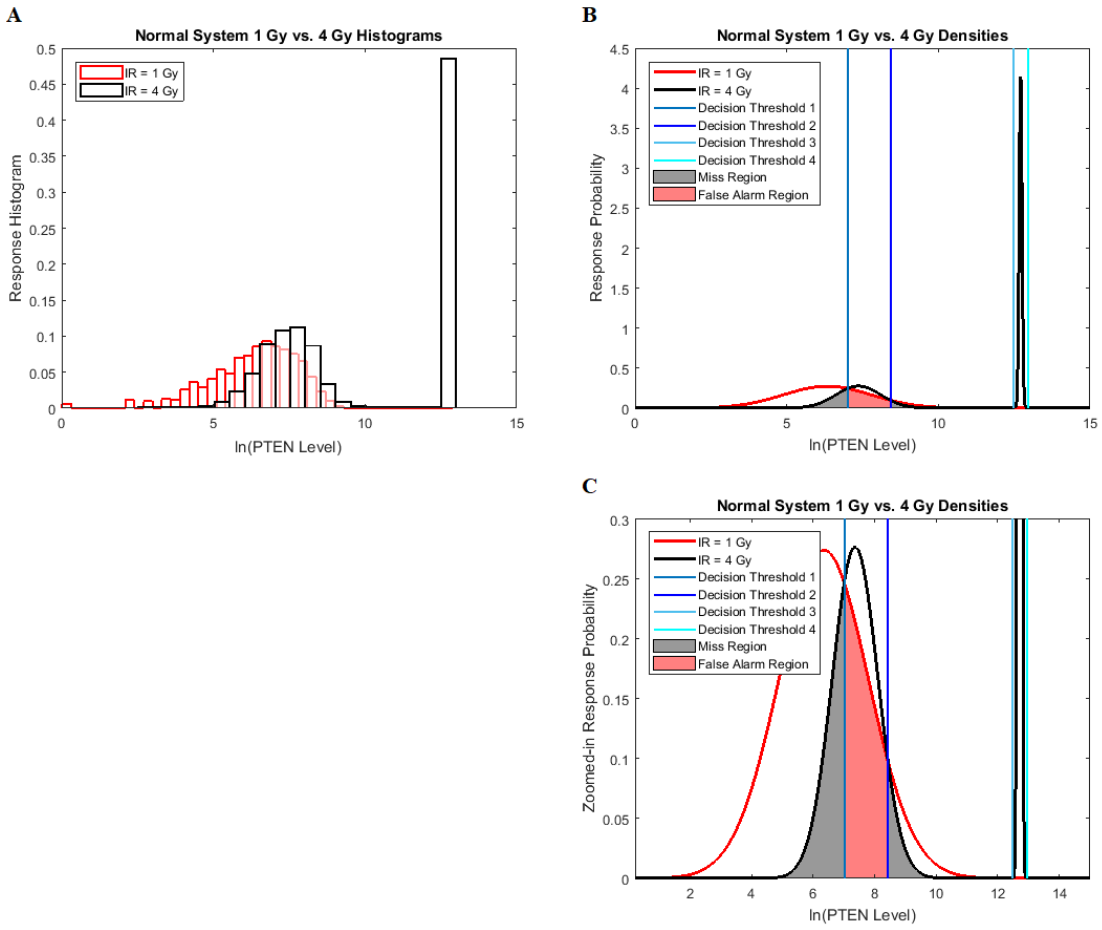


Figure 5.2 Univariate decision making and signaling outcome analysis in the normal p53 system when a PTEN response distribution is bimodal. (A) Histograms of PTEN levels of cells under IR = 1 Gy and IR = 4 Gy doses. (B) A Gaussian probability density function (PDF) for PTEN levels of cells under IR = 1 Gy and a mixture of two Gaussian PDFs for PTEN levels of cells under IR = 4 Gy doses, together with the optimal maximum likelihood decision thresholds which minimize the total decision error probability. (C) Zoomed-in view of panel B.

The P_{FA} expression corresponds to the pink region in Figure 5.2C, whereas the two Q functions in the P_M expression correspond to the two gray regions in Figure 5.2C, respectively. Using the data, computed numerical values are $\xi = 0.51$, $P_{FA} = 0.28 - 0.06 \approx 0.22$, $P_M = 0.51[0.41 + 0.07] \approx 0.25$ and $P_E \approx 0.24$, the last one being calculated using Equation (4.3).

As an example of a negligible decision error probability around the third and fourth thresholds mentioned earlier, consider the area under the red Gaussian PDF $p(x|\mathrm{H}_0)$ in Figure 5.2C for $PTEN_{\mathrm{th3}} < x < PTEN_{\mathrm{th4}}$. While not visible due to being very small, it can be understood that the aforementioned area is a false alarm probability of deciding H_1, although H_0 is true. Numerical value of this false alarm probability is $Q\big((PTEN_{\mathrm{th3}} - \mu_0)/\sigma_0\big) - Q\big((PTEN_{\mathrm{th4}} - \mu_0)/\sigma_0\big) = 1.2 \times 10^{-5} - 2.8 \times 10^{-6} \approx 0$, which is negligible compared to $P_{FA} \approx 0.22$ calculated in the previous paragraph.

5.2 Decision Making Analysis in the Abnormal p53 System

To see how an abnormality in the p53 system affects the decision making and signaling outcomes, we calculate P_E values when Wip1 synthesis rate is elevated by 50% from 0.1 to 0.15 (Figure 5.3), as mentioned previously. As suggested by Habibi et al. (2017), the decision thresholds are modeled to be those of the normal cells. This implies that abnormal cells are not aware of the abnormality, and therefore erroneously use the previous threshold. As we see later, this increases decision error probabilities, a behavior that can be anticipated from abnormal cells. Using Equations (5.5), (5.6), and (4.3), P_{FA}, P_M, and P_E are computed: $P_E = 0.44$ is obtained for 1 vs. 2 Gy scenario (Figure 5.3A), and $P_E = 0.16$ is obtained for 1 vs. 8 Gy scenario (Figure 5.3B). Compared to the normal system results, the overall error probability is significantly higher for the abnormal system under the 1 vs. 8 Gy scenario (we observe that $P_E = 0.001$ of normal cells markedly increases to $P_E = 0.16$ in abnormal cells). The reason is that when the Wip1 synthesis rate is increased, the two response PDF curves significantly overlap (notice the overlap between the left side

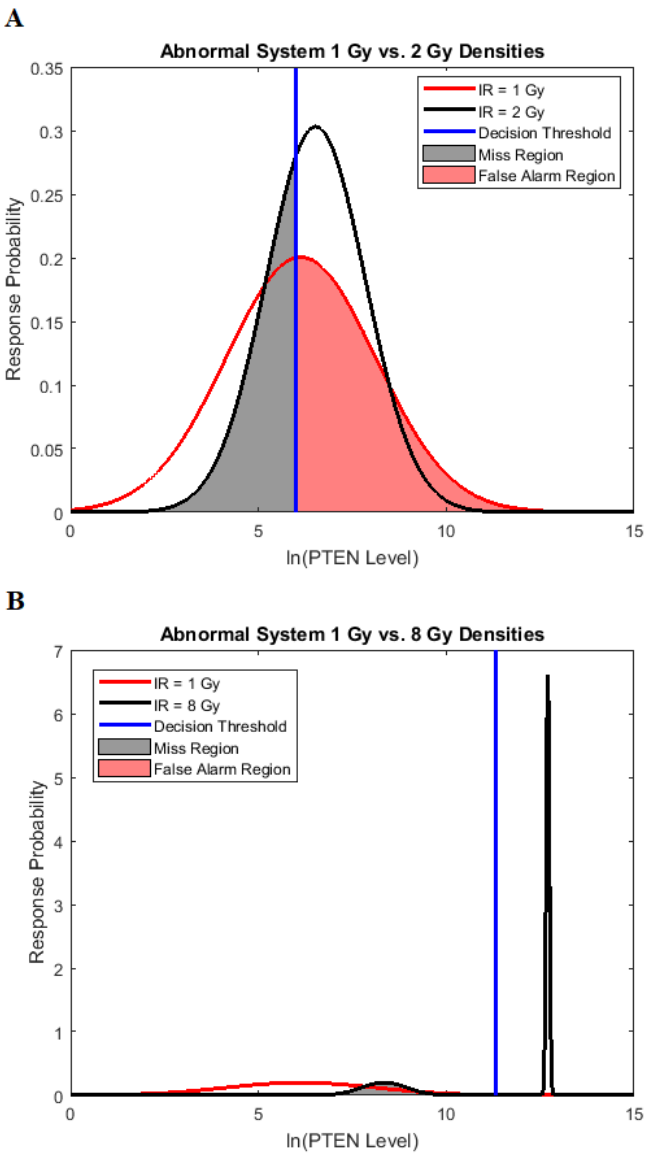


Figure 5.3 Univariate decision making and signaling outcome analysis in an abnormal p53 system, with increased Wip1 synthesis rate, based on PTEN response distributions. (A) Gaussian probability density functions (PDFs) for PTEN levels of abnormal cells under IR = 1 Gy and IR = 2 Gy doses, together with the decision threshold of normal cells. This implies that in abnormal cells the previous decision threshold is erroneously used (Habibi et al., 2017). As discussed later, this increases decision error probabilities, a behavior that can be anticipated from abnormal cells. (B) A Gaussian PDF for PTEN levels of abnormal cells under IR = 1 Gy dose and a mixture of two Gaussian PDFs for PTEN levels of abnormal cells under IR = 8 Gy dose, together with the decision threshold of normal cells.

component of the IR = 8 Gy PDF with the IR = 1 Gy PDF in Figure 5.3B). This is while in normal cells they had almost no overlap (Figure 5.1D).

Similarly, we compute error probabilities for the other abnormal p53 system we mentioned previously, generated by the PTEN synthesis rate reduced from 0.03 to 0.015 (50% reduction). Error probabilities for this abnormality for all different radiation exposure scenarios of 1 vs. 2 Gy up to 1 vs. 8 Gy are shown in Figure 5.4. For comparison, error probabilities for the Wip1-perturbed abnormal p53 system and the normal p53 system are provided in Figure 5.4 as well. We observe that as the difference between the two applied IR doses increases, the decision error probability in normal cells drops significantly. This is while in abnormal cells, decision error probabilities remain high. These signaling outcomes might be correlated with the observation that cell death percentages in abnormal systems are considerably lower than the normal system, even when the radiation dose increases (Figure 4.2). This could indicate that abnormal cells do not respond to IR levels properly and hence, decisions and signaling outcomes affecting apoptosis and survival become more erroneous. Care should be taken that these specific observations are based on the low versus high IR, e.g., d_0 vs d_1 IR hypothesis testing formulation where the low IR dose is fixed to 1 Gy (d_0 = 1 Gy) and the high IR dose is ranging from 2 Gy up to 8 Gy (d_1 = 2, 3, …, 8 Gy) in the p53 system, that is considered as an example in this dissertation. These observations may not be generalized to other selections of the low d_0 and high d_1 IR doses or other hypothesis testing formulations, case studies, or signaling networks. However, the proposed framework and its analytical tools, whose introduction has been the main goal of this study, can be similarly used.

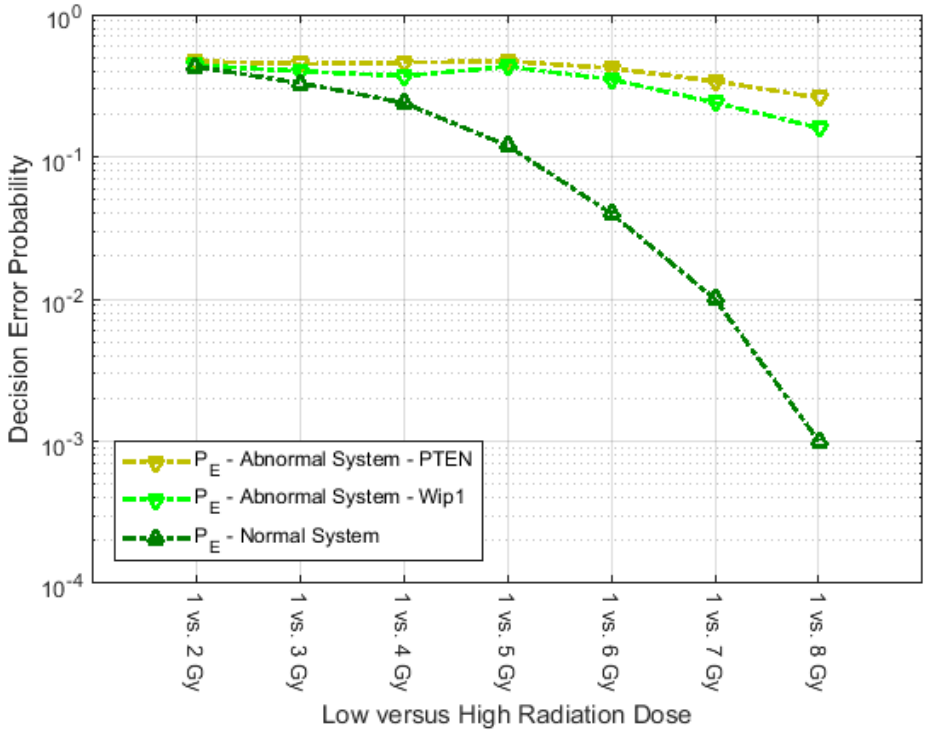


Figure 5.4 Decision error probabilities for several low IR versus high IR scenarios.
Note: The "Abnormal System – PTEN" legend refers to a p53 system whose PTEN synthesis rate is decreased by 50%, compared to its nominal value. The "Abnormal System – Wip1" legend refers to a p53 system whose Wip1 synthesis rate is increased by 50%, compared to its nominal value. Smaller decision error probabilities in the normal system are noteworthy.

5.3 Decision and Signaling Outcome Analysis Using Receiver Operating Characteristic (ROC) Curves

In this subsection, we show how to analyze the performance of a decision maker using receiver operating characteristic (ROC) curves. The ROC curve is developed to visualize the performance of decision-making systems (Kay, 1998; Van Trees et al., 2013), and is a graph of probability of detection, $P_D = 1 - P_M$, versus the probability of false alarm, P_{FA}. In Figure 5.5 we present ROC curves for both the normal p53 system (Figure 5.5A) and the abnormal p53 system (Figure 5.5B) whose Wip1 synthesis rate is elevated, for these two low vs. high IR decision making scenarios: 1 vs. 2 Gy and 1 vs. 8 Gy. The theoretical ROC curves in Figure 5.5 are graphed using the false alarm and miss decision error

probability formulas in Equations (5.5) and (5.6), respectively, with the parameters μ, σ, the thresholds estimated from the data. The empirical ROC curves in Figure 5.5 are graphed by using the data sets directly, using the MATLAB command "perfcurve". We observe that the theoretical and empirical ROCs are nearly the same. Therefore, in what follows, we focus on the theoretical ROC curves to explain concepts and results.

A ROC curve is above a 45° diagonal line (Kay, 1998), the gray dashed line in Figure 5.5. In our study it represents the worst possible decision maker, i.e., a decision-making system that does not use the data and instead randomly decides if the applied IR dose is low or high, by just flipping a coin. The 45° line is indeed a reference to judge the performance of a decision-making system. A ROC curve far away from the 45° reference line indicates a good decision maker. Each point on a ROC curve represents a (P_{FA},P_D) pair that corresponds to a certain decision threshold. Other properties of ROC curves can be found in Van Trees et al. (2013). The "×" marks in Figure 5.5A show the optimal (P_{FA},P_D) points that correspond to the optimal decision thresholds shown in Figure 5.1B and Figure 5.1D, previously computed using Equation (5.4) for the 1 vs. 2 Gy and 1 vs. 8 Gy scenarios, respectively.

Based on the normal p53 system ROC curves in Figure 5.5A, we observe that decisions are made better under the 1 vs. 8 Gy scenario, because of its ROC curve being very far from the 45° reference line, compared to the 1 vs. 2 Gy case whose ROC curve is much closer to the 45° reference line. This finding supports our results presented in Figure 5.4, showing the smaller decision error probability of 0.001 for 1 vs. 8 Gy, compared to the larger decision error probability of 0.43 for 1 vs. 2 Gy. ROC curves also show that

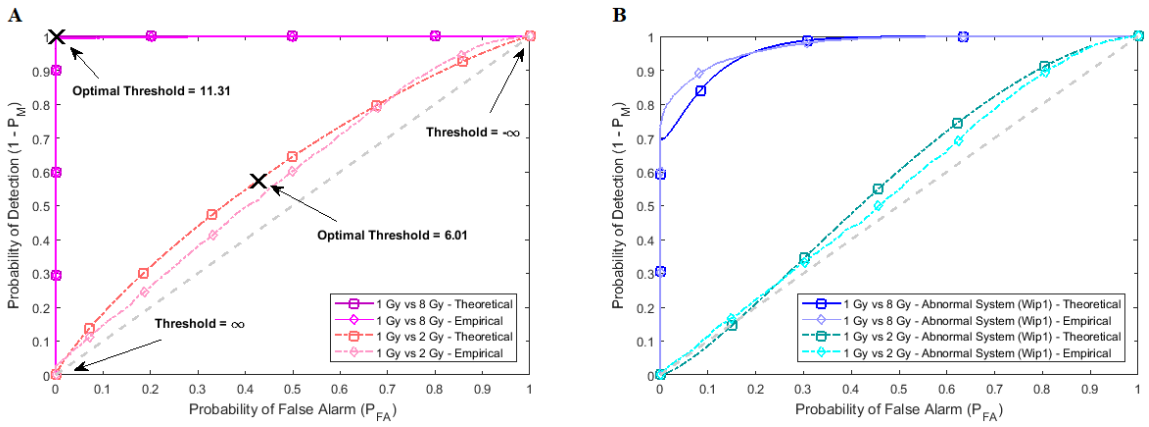


Figure 5.5 Empirical and theoretical receiver operating characteristic (ROC) curves for both normal and abnormal p53 systems. (A) ROC curves of 1 vs. 2 Gy and 1 vs. 8 Gy radiation scenarios for the normal system. (B) ROC curves of 1 vs. 2 Gy and 1 vs. 8 Gy radiation scenarios for the Wip1-perturbed abnormal system.
Note: The theoretical ROC curves labeled by □ are obtained from the Gaussian and mixture of Gaussians data models and formulas whose parameters are estimated from the data, whereas the empirical ROC curves labeled by ◊ are obtained directly from the data. We observe that the theoretical and empirical ROCs are nearly the same. Note that Threshold = ln(PTEN Level) in the figures.

abnormalities in the p53 system can cause decision precision loss. Comparing the normal (Figure 5.5A) and abnormal system ROC curves (Figure 5.5B), we observe that the abnormal system ROC curves are closer to the 45° reference line, meaning that more erroneous decisions are made, when there is an abnormality in the system.

CHAPTER 6

MULTIVARIATE CELL DECISION MAKING ANALYSIS

In this chapter, we extend the decision modeling framework introduced in Chapter 5 such that one can model and analyze multidimensional signaling outcome processes using multi-time point measurements. This allows to incorporate signaling dynamics into decision making analysis. To explain the concept, we start with a bivariate decision-making analysis and then generalize it to a multivariate decision-making framework (Ozen et al., 2020).

6.1 Methods for Computing Decision Thresholds and Decision Error Rates Using Two Time Point Measurements in Individual Cells

In this section, we analyze PTEN levels of 5000 cells measured in one hour and 30 hours after the irradiation phase. Using two variables instead of one allows to study the effect of temporal dynamical changes on decision making and signaling outcomes, and paves the way for analyzing decisions based on multiple time point data. Suppose x and y represent the ln(PTEN) levels in one hour and 30 hours, respectively, after radiation. Joint Gaussian PDF for x and y can be written as (Papoulis, 1991)

$$p(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2}-\frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}+\frac{(y-\mu_y)^2}{\sigma_y^2}\right]\right), \quad (6.1)$$

where (μ_x,σ_x^2) and (μ_y,σ_y^2) are means and variances of x and y, and ρ is correlation coefficient between x and y. Bivariate conditional likelihood ratio is given by $L(x,y)=p(x,y|\mathrm{H}_1)/p(x,y|\mathrm{H}_0)$, and the optimal decision maker which minimizes the overall error

probability P_E compares $L(x, y)$ with the ratio $\gamma = P(\mathrm{H}_0)/P(\mathrm{H}_1)$. The system decides H_1 if $L(x, y) > \gamma$. If the hypotheses are equi-probable, i.e., $P(\mathrm{H}_0) = P(\mathrm{H}_1) = 0.5$, then the optimal system decides H_1 if $p(x, y|\mathrm{H}_1) > p(x, y|\mathrm{H}_0)$. To find the optimal decision threshold curve, we need to solve the equation $L(x) = \gamma$, i.e., $P(\mathrm{H}_1)p(x, y|\mathrm{H}_1) = P(\mathrm{H}_0)p(x, y|\mathrm{H}_0)$, for x and y. When H_0 and H_1 re equi-probable, the threshold equation to be solved simplifies to $L(x, y) = 1$, i.e., $p(x, y|\mathrm{H}_1) = p(x, y|\mathrm{H}_0)$. To find false alarm and miss probabilities, Equations (5.1) and (5.2) can be extended to two variables as follows:

$$P_{FA} = \iint_{(x,y) \in \text{false alarm region}} p(x, y|\mathrm{H}_0)dx\,dy, \tag{6.2}$$

$$P_M = \iint_{(x,y) \in \text{miss region}} p(x, y|\mathrm{H}_1)dx\,dy, \tag{6.3}$$

where $\{x, y: p(x, y|\mathrm{H}_1) > p(x, y|\mathrm{H}_0)\}$ defines the false alarm region when H_0 is true, and $\{x, y: p(x, y|\mathrm{H}_0) > p(x, y|\mathrm{H}_1)\}$ specifies the miss region when H_1 is true. After computing P_{FA} and P_M, the overall decision error probability P_E can be calculated using Equation (4.3).

As an example, here we focus on Figure 6.1A, where two bivariate Gaussian PDFs are shown for $x = \ln(\text{PTEN at } 1^{\text{st}} \text{ hour})$ and $y = \ln(\text{PTEN at } 30^{\text{th}} \text{ hour})$, logarithms of PTEN levels in the two data sets of IR = 1 Gy and IR = 2 Gy, with each data set consisting of 5000 cells. The mean and variance parameters of each bivariate response PDF are estimated from the associated data set. The overlap between the two bivariate PDFs in

response to IR = 1 Gy and IR = 2 Gy can be better seen in the top view shown in Figure 6.1B. This figure also demonstrates that the decision threshold between the two PDFs is going to be a curve in the *x*-*y* plane, where the two PDFs intersect. Equation for this optimal threshold curve which minimizes the total decision error probability is given by $L(x, y) = 1$, where L is the bivariate conditional likelihood ratio defined previously. This decision threshold curve $curve_{\text{th}}$ is shown together with contour plots of the two bivariate PDFs in Figure 6.1C. To compute the decision error probabilities using the decision threshold $curve_{\text{th}}$, Equations (6.2) and (6.3) for the false alarm and miss error probabilities can be written as

$$P_{FA} = \int_{x=-\infty}^{\infty} \int_{y=curve_{\text{th}}}^{\infty} p(x, y|\mathrm{H}_0)\, dy\, dx, \tag{6.4}$$

$$P_M = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{curve_{\text{th}}} p(x, y|\mathrm{H}_1)\, dy\, dx. \tag{6.5}$$

After computing the integrals in Equations (6.4) and (6.5) numerically, we obtain $P_{FA} = 0.24$ and $P_M = 0.26$. Upon their substitution in Equation (4.3) and with equi-probable hypotheses, we obtain $P_E = 0.25$.

To compare the above two time point decision with individual one time point decisions, we compute the decision error probabilities based on the 1st hour data and the 30th hour data, individually, for the IR = 1 vs. 2 Gy scenario. We obtain $P_E = 0.5$ and $P_E = 0.27$ for individual univariate decisions in one hour and 30 hours after the radiation, respectively. We observe that the bivariate decision offers significant improvement over

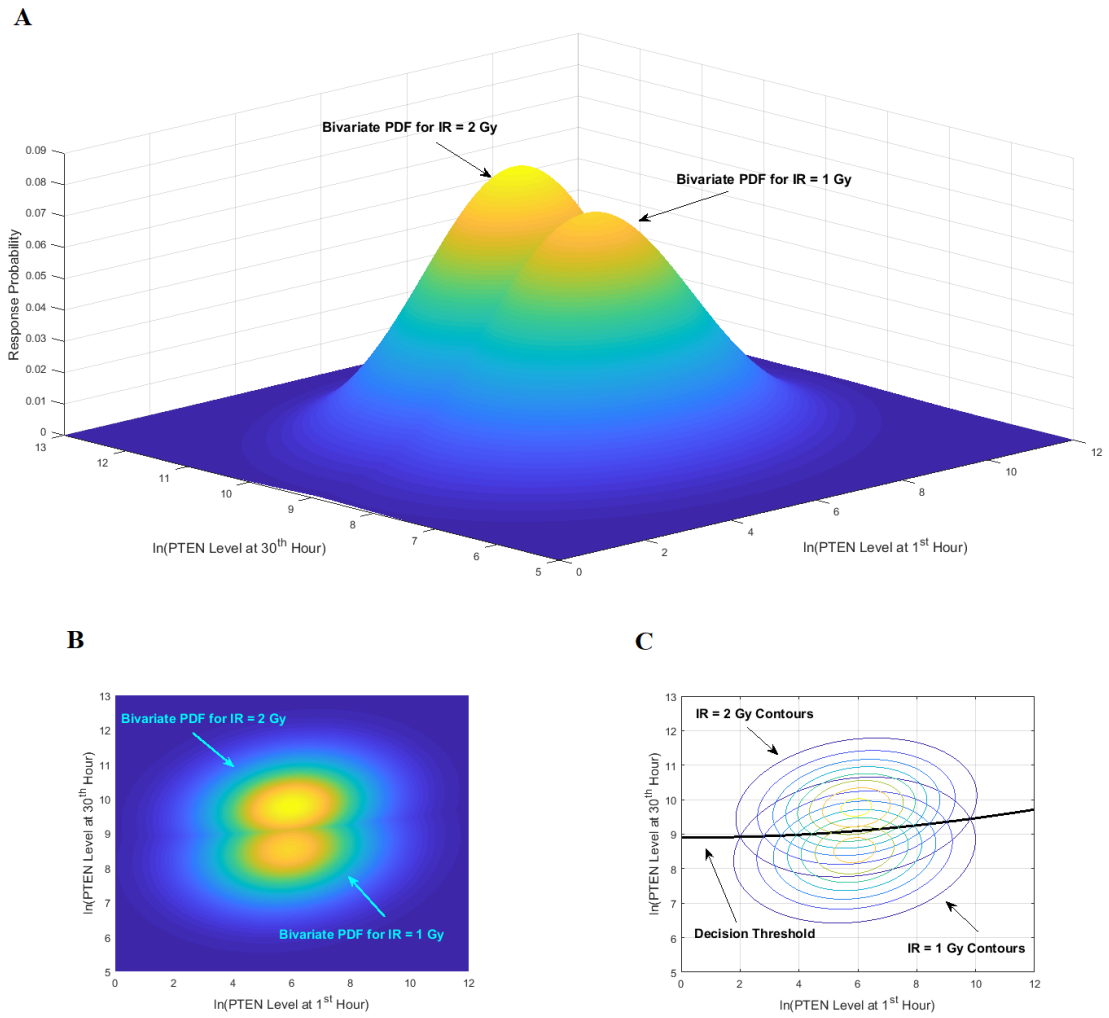


Figure 6.1 Bivariate decision making and signaling outcome analysis in the normal p53 system based on PTEN response distributions. (A) Bivariate Gaussian probability density functions (PDFs) for PTEN levels of cells at the 1st hour and the 30th hour, under IR = 1 Gy and IR = 2 Gy doses. (B) Top view of the two bivariate Gaussian PDFs. (C) Top contour view of the two bivariate Gaussian PDFs, together with the optimal maximum likelihood decision threshold curve which minimizes the total decision error probability.

the one-hour decision and slight improvement over the 30-hour decision. Univariate decision error probabilities at different time points are discussed in the next section, as well as how multivariate decision error probability changes, as the data of more time points are added to the decision process in a sequential manner.

6.2 Methods for Computing Decision Thresholds and Decision Error Rates Using Multiple Time Point Measurements in Individual Cells

In this section, we further study the effect of system dynamics on decision making and signaling outcomes, by considering multiple time point data. More specifically, we consider PTEN levels of 5000 cells measured in 1, 10, 20, 30, 40, 50, 60, and 70 hours after the irradiation phase. Let $\boldsymbol{\omega}$ be an $\mathrm{N} \times 1$ column vector that represents the ln(PTEN) levels at a subset or all of the aforementioned time instants. Joint Gaussian PDF for all decision variables in $\boldsymbol{\omega}$ can be written as (Duda, 2001; Van Trees et al., 2013):

$$p(\boldsymbol{\omega}) = \frac{1}{(2\pi)^{\mathrm{N}/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{\omega}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{\omega}-\boldsymbol{\mu})\right], \tag{6.6}$$

where $\boldsymbol{\mu}$ is the $\mathrm{N} \times 1$ mean vector, Σ is the $\mathrm{N} \times \mathrm{N}$ covariance matrix, $|\Sigma|$ and Σ^{-1} denote the determinant and inverse of Σ, respectively, and $^{\mathrm{T}}$ represents the matrix transpose. This multivariate Gaussian or normal PDF for the decision vector $\boldsymbol{\omega}$ can be symbolically shown by $\boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$. For $\mathrm{N} = 2$, Equation (6.6) simplifies to the bivariate PDF in Equation (6.1), such that:

$$\boldsymbol{\omega} = \begin{bmatrix} x \\ y \end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}.$$

Computation of the decision error probabilities using multiple decision variables can be accomplished using discriminant functions (Duda, 2001; Van Trees et al., 2013):

$$g_i(\boldsymbol{\omega}) = \ln p(\boldsymbol{\omega}|\mathrm{H}_i) + \ln P(\mathrm{H}_i)\,, i = 0{,}1, \tag{6.7}$$

where $p(\boldsymbol{\omega}|\mathrm{H}_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$ and i is index of the discriminant function associated with the hypothesis H_i. In our case, we have $i = 0{,}1$, referring to our two hypotheses in Equation (4.1). For any hypothesis H_i, substitution of Equation (6.6) in Equation (6.7) simplifies its discriminant function to:

$$g_i(\boldsymbol{\omega}) = -\frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\mu}_i)^{\mathrm{T}}\Sigma_i^{-1}(\boldsymbol{\omega} - \boldsymbol{\mu}_i) - \frac{\mathrm{N}}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_i| + \ln P(\mathrm{H}_i)\,, i = 0{,}1. \tag{6.8}$$

Using the discriminant functions in Equation (6.8) and for a given $\boldsymbol{\omega}$, the optimal decision-making system decides H_0 if $g_0(\boldsymbol{\omega}) > g_1(\boldsymbol{\omega})$, and decides H_1 if $g_1(\boldsymbol{\omega}) > g_0(\boldsymbol{\omega})$. The false alarm probability P_{FA} is the probability of deciding H_1, i.e., $g_1(\boldsymbol{\omega}) > g_0(\boldsymbol{\omega})$, whereas in fact H_0 is true. On the other hand, the miss probability P_M is the probability of deciding H_0, i.e., $g_0(\boldsymbol{\omega}) > g_1(\boldsymbol{\omega})$, although indeed H_1 is true. Computing P_{FA} and P_M using multivariate PDFs directly entails multivariate integrations over regions defined by decision surfaces. Given the complexity of such computations, as a simpler alternative we calculate P_{FA} and P_M using the data directly, by counting the number of times that false alarm and miss event occur, respectively, after comparing the discriminant function values $g_1(\boldsymbol{\omega})$ and $g_0(\boldsymbol{\omega})$ for each $\boldsymbol{\omega}$, and then divide them by the total number of data points. The overall decision error probability P_E can be calculated using Equation (4.3). Another method for computing P_{FA} and P_M relies on characteristic functions (Fukunaga, 1990).

6.2.1 Single and Multivariate Decision Making and Signaling Outcome Analysis as Time Evolves

To understand how decision making and signaling outcomes may change over time, first we look at the decision error probability P_E using PTEN levels measured at individual consecutive time instants (Figure 6.2A), for the 1 vs. 2 Gy scenario. A noteworthy observation is that the decision error exhibits a minimum value. The minimum occurs in 20 hours after radiation. This can be visually explained by the amount of overlap of PTEN histograms at each individual time point. For instance, we provide histograms of PTEN levels at the 20th and the 70th hours in Figure 6.3, for IR = 1 and 2 Gy doses. We observe that the 20th hour histograms have less overlap than the 70th hour histograms, shown in Figure 6.3A and Figure 6.3B, respectively, which results in the smaller P_E at the 20th hour in Figure 6.2A.

Now we focus on studying how decision-making works, if data of N time instants are utilized, such that N = 1, 2, …, 8 (Figure 6.2B). In the figure, N = 1 means the PTEN data of the 1st hour, N = 2 refers to the PTEN data of the 1st and the 10th hours, N = 3 indicates the PTEN data of the 1st, the 10th, and the 20th hours, etc. This assumes at any given time, the decision is made based on the data of that given time, plus the data of the previous time instants, which means progressively accumulating the data to make decisions. It is observed in Figure 6.2B that P_E first decreases, and after a certain point, it remains nearly constant. To understand this behavior, we note that if the data collected at various time instants are independent, then the error probability of a decision-making system that performs sequential hypothesis testing decreases as the number of observations N increases (Fukunaga, 1990). This property of a multivariate sequential decision maker is intuitively appealing. However, if the data collected at various time instants are correlated,

performance of the multivariate sequential decision maker can significantly degrade, and its error probability does not necessarily decrease as N increases (Fukunaga, 1990).

To examine possible temporal correlations among the data that the suggested sequential decision strategy employs, we compute condition numbers of Σ_0 and Σ_1, the $N \times N$ covariance matrices of the data for the two hypotheses H_0 and H_1, for IR = 1 and 2 Gy, respectively, as N increases from 2 to 8 (Figure 6.2C). The condition number of a matrix is the ratio of its largest singular value to its smallest. A large condition number indicates that the matrix is nearly singular. On the other hand, a near singular covariance matrix of several random variables means that some of the random variables are highly correlated. Therefore, a large condition number for a covariance matrix implies large correlations among some of its random variables. We observe in Figure 6.2C that as N increases, condition numbers of both of the covariance matrices Σ_0 and Σ_1 increase. This means as time evolves after a certain point, the suggested sequential decision maker incorporates a new observation that is correlated with the previously used observations. The correlation does not allow the decision error probability to decrease beyond a certain point, although N constantly increases (Figure 6.2B).

6.2.2 Multivariate Decision Making and Signaling Outcome Analysis of Two or More Molecules

So far, we have focused on multivariate decision making and signaling outcome analysis for one molecule at different time instants. However, the introduced methods and algorithms are not limited to the outcome analyses for just one molecule, and they can be applied to various other scenarios and studies. In fact, they can be used to analyze and compute decision error rates based on the concentration levels of two or more molecules,

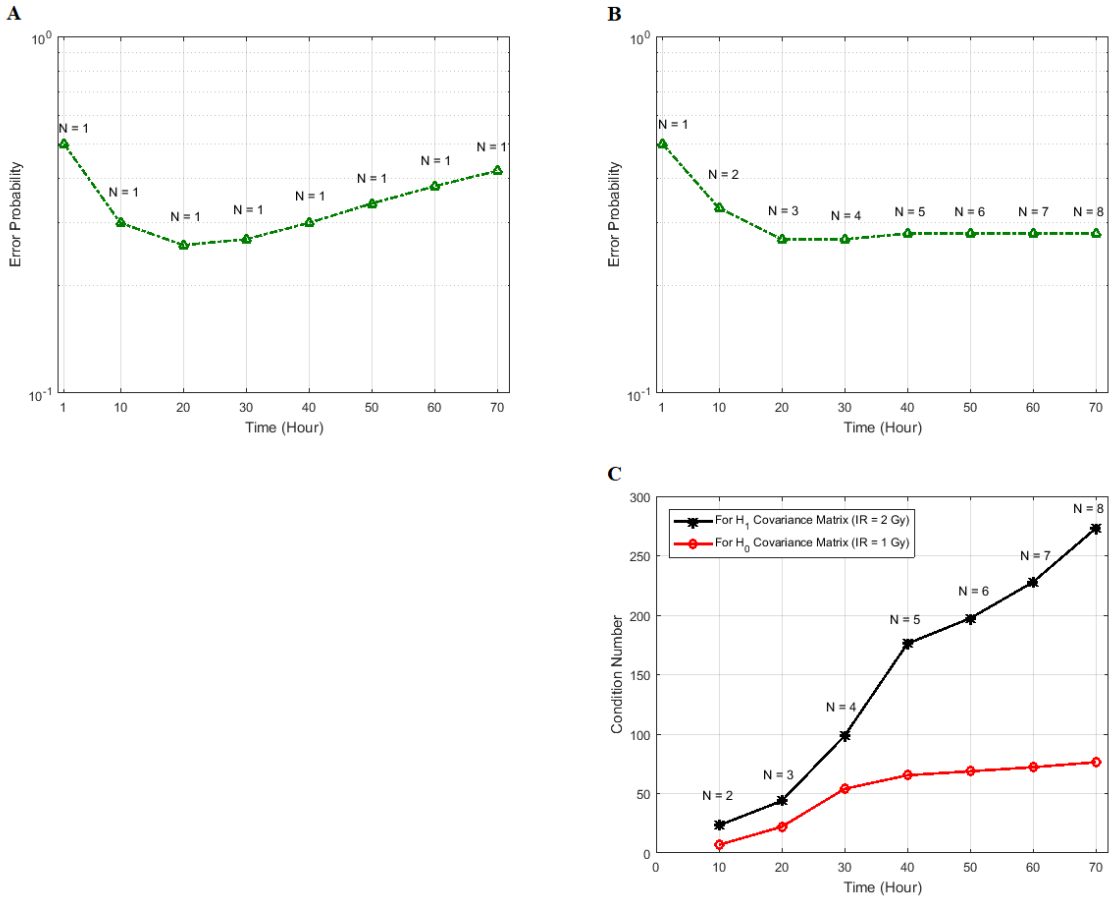


Figure 6.2 Decision error probabilities versus time in the normal p53 system: A single versus multiple time point study. (A) P_E as a function of time for the 1 vs. 2 Gy radiation scenario, computed using only the PTEN data of a single, $N = 1$, individual time instant. This assumes at any given time, decision is made based on the data of that time only. Having a minimum error probability at the 20th hour is noteworthy. (B) P_E as a function of time for the 1 vs. 2 Gy radiation scenario, computed using the PTEN data of N time instants, $N = 1, 2, \ldots, 8$ ($N = 1$ means the PTEN data of the 1st hour, $N = 2$ refers to the PTEN data of the 1st and the 10th hours, $N = 3$ indicates the PTEN data of the 1st, the 10th, and the 20th hours, etc.). This assumes at any given time, decision is made based on the data of that time, plus the data of the previous time instants, which means accumulating the data to make a decision. It is observed that P_E first decreases, and after a certain point, it remains nearly constant. (C) Condition numbers of Σ_0 and Σ_1, the $N \times N$ covariance matrices of the data for the two hypotheses H_0 and H_1, for IR = 1 and 2 Gy, respectively, as N increases from 2 to 8. When N increases, condition numbers of both of the covariance matrices Σ_0 and Σ_1 increase. On the other hand, a large condition number for a covariance matrix implies large correlations among some of its random variables. Therefore, as time evolves after a certain point, the suggested sequential decision maker incorporates a new observation that is correlated with the previously used observations. The correlation does not allow the decision error probability P_E to decrease beyond a certain point, although N constantly increases.

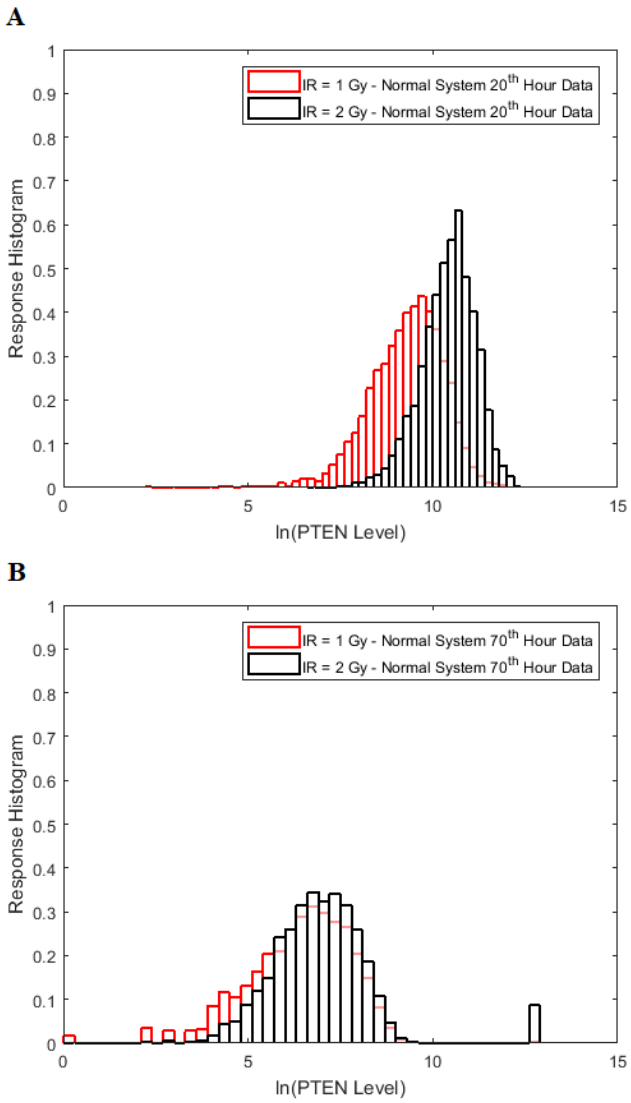


Figure 6.3 Comparison of the histograms of cells' PTEN levels at the 20th and the 70th hours under IR = 1 Gy and 2 Gy doses in the normal p53 system. (A) Histograms of the 20th hour PTEN data under IR = 1 and 2 Gy doses, which show less overlap. (B) Histograms of the 70th hour PTEN data under IR = 1 and 2 Gy doses, which show more overlap.

measured simultaneously or even at different time instants. For example, if decision and outcome analysis are going to be conducted based on simultaneous concentration level measurements of two molecules labeled by x and y, then Equations (6.1) – (6.5) can be used to find the maximum likelihood bivariate decision strategy and its minimum error

probability. As a more elaborate example, suppose concentration levels of molecule A measured at time instants t_1 and t_2 are labeled as variables x and y, respectively, concentration levels of molecule B measured at t_1 and t_2 are labeled as variables v and w, respectively, and finally concentration levels of molecule C measured at t_1 and t_2 are labeled as variables ψ and ζ, respectively. The 6×1 decision vector $\boldsymbol{\omega}$ including all these six decision variables can be defined as $\boldsymbol{\omega} = [x\ y\ v\ w\ \psi\ \zeta]^{\mathrm{T}}$, where $^{\mathrm{T}}$ stands for transpose. Now, Equations (6.6) – (6.8) can be used to find the maximum likelihood six-variate decision strategy and its minimum decision error probability.

CHAPTER 7

BEYOND BINARY CELL DECISIONS

In the previous three chapters, we focused on the binary hypothesis testing for the case where there exist only two possible outcomes that are cell death and cell survival. In this chapter, we discuss the case where there exist more than two possible outcomes on a hypothetical scenario. Furthermore, we show how heterogeneity of initial values and reaction rates affect the overall cell response. Lastly, we discuss the cost of correct and incorrect decisions (Ozen et al., 2020).

7.1 Ternary Decisions and Signaling Outcomes and Ternary Error Probabilities

While the focus of the previous three chapters is on binary hypothesis testing, it is possible to develop a multiple hypothesis testing model for outcome analysis, where there exist more than two possible outcomes. This entails more erroneous decisions than false alarm and miss events. Optimal decision thresholds and error probabilities for all incorrect decisions can be similarly computed. For example, assume there are three different signaling outcomes depending on concentration level of a hypothetical molecule called MOL, whose level can fall within one of three regions, which results in the following three possible hypotheses:

$$\begin{aligned} &H_0\text{: MOL level is low,} \\ &H_1\text{: MOL level is medium,} \\ &H_2\text{: MOL level is high.} \end{aligned} \tag{7.1}$$

Let us assume under each condition, PDF of the MOL level represented by x is normal or Gaussian, i.e., $x \sim \mathcal{N}(\mu_i, \sigma^2)$ such that $\mu_0 < \mu_1 < \mu_2$, where variances are assumed to be equal, to simplify the notation. These PDFs are shown in Figure 7.1, with $\mu_0 = 5, \mu_0 = 10, \mu_0 = 15$, and $\sigma^2 = 2.25$. By extending the binary decision errors presented earlier in Equations (5.5) and (5.6), ternary decision errors for the three hypotheses can be written as

$$P_{E,\mathrm{H}_0} = \int_{a}^{\infty} p(x|\mathrm{H}_0)dx = Q\left(\frac{a-\mu_0}{\sigma}\right), \tag{7.2}$$

$$P_{E,\mathrm{H}_1} = \int_{-\infty}^{a} p(x|\mathrm{H}_1)dx + \int_{b}^{\infty} p(x|\mathrm{H}_1)dx = Q\left(\frac{\mu_1-a}{\sigma}\right) + Q\left(\frac{b-\mu_1}{\sigma}\right), \tag{7.3}$$

$$P_{E,\mathrm{H}_2} = \int_{-\infty}^{b} p(x|\mathrm{H}_2)dx = Q\left(\frac{\mu_2-b}{\sigma}\right). \tag{7.4}$$

In the above equations, a and b are thresholds to decide between H_0 and H_1, and between H_1 and H_2, respectively. This means the decision regions for the three hypotheses can be written as:

$$\begin{aligned} &\mathrm{H}_0: x < a, \\ &\mathrm{H}_1: a < x < b, \\ &\mathrm{H}_2: b < x. \end{aligned} \tag{7.5}$$

For equi-probable hypotheses and similarly to the derivation that lead to Equation (5.4), optimal decision thresholds which minimize the total decision error probability can be shown to be:

$$a = \frac{\mu_0 + \mu_1}{2}, \quad b = \frac{\mu_1 + \mu_2}{2}. \tag{7.6}$$

Upon substituting Equation (7.6) in Equations (7.2), (7.3), and (7.4), the total error probability in making ternary decisions can be written as:

$$\begin{aligned} P_E &= \frac{1}{3}P_{E,\mathrm{H}_0} + \frac{1}{3}P_{E,\mathrm{H}_1} + \frac{1}{3}P_{E,\mathrm{H}_2} \\ &= \frac{1}{3}Q\left(\frac{\mu_1 - \mu_0}{2\sigma}\right) + \frac{1}{3}\left[Q\left(\frac{\mu_1 - \mu_0}{2\sigma}\right) + Q\left(\frac{\mu_2 - \mu_1}{2\sigma}\right)\right] + \frac{1}{3}Q\left(\frac{\mu_2 - \mu_1}{2\sigma}\right) \\ &= \frac{2}{3}Q\left(\frac{\mu_1 - \mu_0}{2\sigma}\right) + \frac{2}{3}Q\left(\frac{\mu_2 - \mu_1}{2\sigma}\right). \end{aligned} \tag{7.7}$$

As a reference, for the binary decision-making problem and outcome analysis studied in the earlier chapters and using Equations (5.5) and (5.6), the total error probability in making binary decisions with equal variances simplifies to:

$$P_E = \frac{1}{2}Q\left(\frac{\mu_1 - \mu_0}{2\sigma}\right) + \frac{1}{2}Q\left(\frac{\mu_1 - \mu_0}{2\sigma}\right) = Q\left(\frac{\mu_1 - \mu_0}{2\sigma}\right). \tag{7.8}$$

To compare ternary and binary error probabilities, let us assume $\mu_2 - \mu_1 = \mu_1 - \mu_0 = \gamma$, which reduces Equations (7.7) and (7.8) to $(4/3)Q\big(\gamma/(2\sigma)\big)$ and $Q\big(\gamma/(2\sigma)\big)$, respectively.

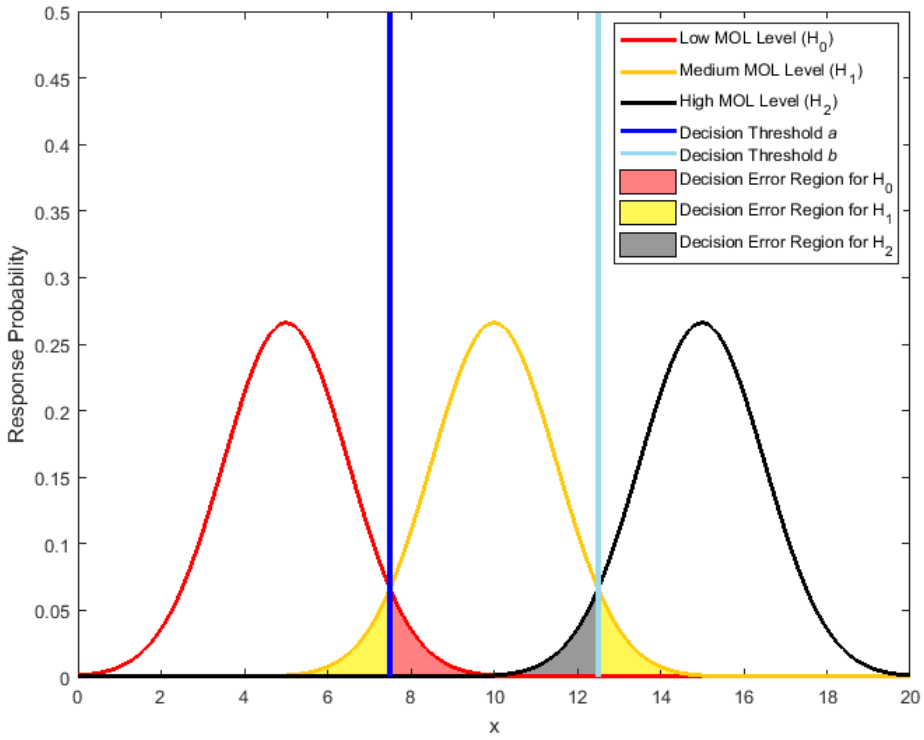


Figure 7.1 Response probability density functions of a hypothetical molecule called MOL whose level entails a ternary decision-making process with three signaling outcomes.
Note: Shaded tail areas with the same color represent decision error regions associated with each specific hypothesis. Assuming equi-probable hypotheses, the optimal maximum likelihood decision thresholds which minimize the total decision error probability are shown by vertical blue lines at the points of intersection of the probability density functions.

This indicates that the ternary decision error rate can be higher than the binary decision error rate, under the assumed conditions.

7.2 Effect of Heterogeneity of Initial Values and Reaction Rates on Cell Response Histograms

In addition to stochasticity in dynamic processes, it is natural to consider that the initial level of each protein is not the same in a cell population (heterogeneity of cells). Additionally, there are pseudo-first order dephosphorylation reactions for which reaction rate coefficients depend on the levels of implicit phosphatases. Thus, it is also natural to

assume that the reaction rate coefficients corresponding to pseudo-first order dephosphorylations may vary cell to cell. To see the effect of heterogeneous initial values and parameters on cell response distributions, we generated new PTEN data for IR = 2 Gy in 5000 cells, assuming that the initial values and parameters of the p53 system are coming from lognormal distributions with means equal to their default values, and ran simulations for different standard deviations, i.e., $\sigma = 0.2$, 0.5 and 1 (Grabowski et al., 2019). Cell response histograms of the new data for different σ values are shown in Figure 7.2, and compared against the 2 Gy data of homogenous cells ($\sigma = 0$). In this system and example, we observe that PTEN histograms undergo some change as σ increases, i.e., more cell heterogeneity, which may result in some changes in decision error probabilities. Nevertheless, one can still use the exact same methods and algorithms introduced in the Chapters 4 to 6, to conduct signaling outcome analyses of interest for inhomogeneous cells.

7.3 On the Cost of Correct and Incorrect Decisions

In decision theory, there can be some costs associated with correct or incorrect decisions. Let C_{ij} be the cost of deciding H_i when H_j is true. To minimize the expected cost, $C_{00}P(\mathrm{H}_0) + C_{01}P(\mathrm{H}_1)P_M + C_{10}P(\mathrm{H}_0)P_{FA} + C_{11}P(\mathrm{H}_1)$, the decision-making system decides H_1 if (Kay, 1998):

$$L(x) = \frac{p(x|\mathrm{H}_1)}{p(x|\mathrm{H}_0)} > \frac{(C_{10} - C_{00})P(\mathrm{H}_0)}{(C_{01} - C_{11})P(\mathrm{H}_1)} = \gamma, \tag{7.9}$$

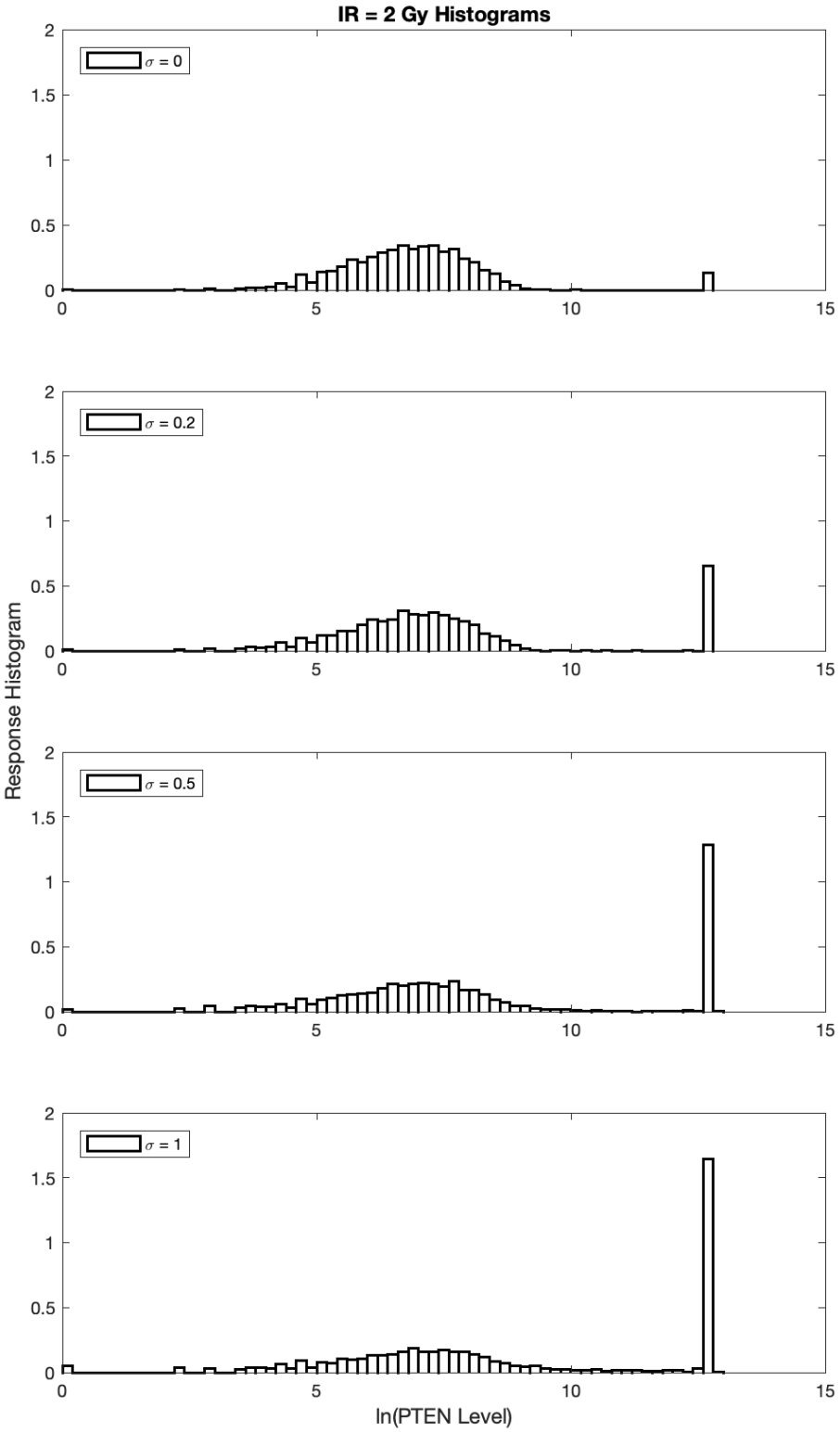


Figure 7.2 Effect of heterogeneity of initial values and pseudo-first order dephosphorylation reaction rates on PTEN histograms. Histograms of PTEN levels of cells under IR = 2 Gy dose, with σ = 0, 0.2, 0.5 and 1.

where $C_{10} > C_{00}$ and $C_{01} > C_{11}$. Usually, the costs associated with correct decisions are zero, i.e., $C_{00} = C_{11} = 0$. Additionally, if there is no preference in assigning different costs to different incorrect decisions, one can choose $C_{10} = C_{01}$. This is what we would consider as well, since we do not have a knowledge of the costs of incorrect decisions in the studied cellular system. Upon substituting $C_{00} = C_{11} = 0$ and $C_{10} = C_{01}$ in Equation (7.9), it simplifies to the following equation, which is the optimal maximum likelihood decision rule presented in Chapter 5:

$$L(x) = \frac{p(x|\mathrm{H}_1)}{p(x|\mathrm{H}_0)} > \frac{P(\mathrm{H}_0)}{P(\mathrm{H}_1)} = \gamma. \tag{7.10}$$

CHAPTER 8

CONCLUSION

In this chapter, we provide our concluding remarks on the developed methods and our observations on different case studies.

8.1 Modeling and Training Molecular Networks

Computational studies of molecular networks are essential and a necessity to understand complex molecular dynamics and to offer novel insights into whole systematic functionalities. Such studies can provide fast and reliable outcomes especially if they are tailored to experimental evidence. Molecular networks can be computationally analyzed by converting them into numerable models (Chapter 2). One way of modeling the networks is converting them into a system of continuous differential equations that captures temporal or spatial (or both) features of the system and provides detailed information on how the whole system behaves (e.g., Section 2.1). However, they are impractical especially for large networks because they require the knowledge of mechanistic details and kinetic parameters, and such knowledge is generally very limited. Thus, discreate models such as Boolean models becomes advantageous since they do not require detailed kinetic information and still provide valuable biological insight. Specifically, the Boolean models are very simple compared to the continuous models, and very helpful to understand system dynamics by correctly predicting the active/inactive states of the molecules (e.g., Section 2.2). Therefore, in this dissertation, we focus on developing Boolean model-based methods and frameworks for network analysis due to their applicability to large and complex

molecular networks and their efficiency in providing novel results as presented here and discussed in several research articles.

Developing biologically relevant network models is not always enough because they usually do not agree with the experimental measurements, specifically for the literature-curated networks. The disagreement between model predictions and the experimental data might be due to the incompleteness of resources, databases, and literature used to construct the networks. Analysis results of a network model with low accuracy on predicting the experimental evidence cannot be trusted. For this reason, developing tools to train the network models is significant so that more reliable models can be analyzed and the likelihood of confirming computational observations in laboratory experiments can be increased. Therefore, we introduce two training approaches (Section 2.3), in this dissertation. In the first training approach (Section 2.3.1), we train the network model by removing interactions from the network assuming that the initial network contains spurious interactions. To do so, we propose constrained integer linear programming (ILP) formulations (Equations (2.2) and (2.3)) that efficiently find subnetworks of the initial network by minimizing the number of mismatches between model predictions and the experimental data while preserving the model rules. Furthermore, we present a simple and effective strategy that overcomes feedback interactions in the training process (Figure 2.8), which allows using the measured data of a molecule at multiple time instances.

In the second training approach (Section 2.3.2), we fix the network topology and learn the Boolean equations of each molecules using the experimental data. In other words, we learn the model for each molecule by assuming multiple inhibitors and multiple

activators may work cooperatively to change the output molecule's state (Equation (2.4)). This is done by converting the learning process into an optimization problem whose general form is provided in Equation (2.5) and solved using Genetic Algorithm, a well-known metaheuristic algorithm. Learning Boolean functions of the molecules is useful if the initial network topology needs to be preserved in some analyses such as fault diagnosis analysis. It will preserve physical existence of all molecules in the network, which prevents losing information on the molecules as well as the effect of the interactions that might be removed in other training approaches. The developed models and methods are elaborated and exemplified in Chapter 2.

8.2 Vulnerability Analysis of Molecular Networks

Developing intracellular signaling network analysis methods is important for understanding complicated biological processes that underlie complex diseases such as mental disorders, autoimmune diseases, and cancer. One fundamental goal of such methods is to determine how much vulnerable a signaling network is to the dysfunction of one or multiple molecules, where the dysfunction of each molecule can be defined as responding incorrectly to its input signals. In this study, we define the vulnerability level of a molecule as the probability of having incorrect network responses when the molecule is dysfunctional, in which dysfunctionality of the molecules are modeled by stuck-at-0 (SA0) and stuck-at-1 (SA1) fault models (Section 3.1) that are constantly 0, inactive, or 1, active, regardless of what the input signals of the molecule are, respectively. The vulnerability levels associated with the dysfunction of molecules or groups of molecules can be measured by computing probabilities of having incorrect network responses in the presence

of molecular dysfunctionalities, using the equations introduced and developed in Section 3.2.

Using the developed mathematical framework for vulnerability computations, we try to understand in a given network, what molecule or group of molecules results in the most detrimental network failure. To answer this question, here we propose a systematic method to identify the worst possible signaling failures in signaling networks (Section 3.3). The worst possible signaling failure is defined as the pathological phenomenon that results in the highest probability of network failure, i.e., the maximum vulnerability level. The said pathological phenomenon is characterized to be emerged from the presence of one or more dysfunctional molecules in the network. While it is conceivable that different individual dysfunctional molecules may have different vulnerability levels, it is not clear what happens to the vulnerability levels, if two or more molecules are concurrently dysfunctional.

The worst signaling failure analysis is initially conducted on the ERBB signaling network (Figure 3.4). We observe that the maximum vulnerability values do not necessarily increase as the number of concurrently faulty molecules N increases (Figure 3.5). More precisely, we see a maximum vulnerability increase going from single faults to double faults, $N = 1$ and 2, respectively, and then the maximum vulnerability does not increase further afterwards. Moreover, we observe that the smallest N for which we see the highest maximum vulnerability in this network is $N = 2$, i.e., the double faults. This teaches that there are some pairs of faulty molecules that cause the most detrimental network damage, and an increase in the number of simultaneously faulty molecules does not induce a worse network failure.

Next, we perform the worst signaling failure analysis on the T cell signaling network (Figure 3.6). Similar to the ERBB network results, we notice that the maximum vulnerability values do not necessarily increase as the number of simultaneously faulty molecules N increases (Figure 3.7), for all network outputs. Furthermore, for the network outputs ap1, bcat and p70s, we see a maximum vulnerability increase going from single faults to double faults, $N = 1$ and 2, respectively, and do not observe further increase afterwards. This means that some pairs of faulty molecules can damage the network as seriously as many more simultaneously faulty molecules. For the network outputs cre, nfat, p38, shp2 and sre, this behavior changes, i.e., the highest maximum vulnerability occurs when $N = 1$. This implies that there are some single faulty molecules that cause the worst possible network failures at these outputs.

We also prove that the computational complexity, i.e., the running time of the proposed worst signaling failure analysis algorithm (Section 3.3.4) is $O(K^3)$, where K is the number of intermediate molecules in the network. This efficient algorithm is in contrast with an exhaustive search having an exponential running time, $O\left(K^{K/2}\right)$, that quickly becomes impractical to implement as K increases. For example, for a network with $K = 50$ molecules, the proposed algorithm complexity is in the order of $50^3 \approx 1.3\times10^5$, which is much smaller than $50^{25} \approx 3\times10^{42}$, the exhaustive search complexity.

The proposed worst signaling failure analysis algorithm makes use of another proposed algorithm (Section 3.4.1) that properly incorporates the effects of signaling feedbacks in the worst failure analysis. Essentially it determines the number of time points (clock cycles) needed for network analysis and simulation while computing the vulnerability levels to find the worst failures, so that we prevent performing network

simulations longer than what is needed. Usefulness of this algorithm is demonstrated by computing the required number of clock cycles for vulnerability analysis of the ERBB and the T cell signaling networks (Sections 3.4.2 and 3.4.3, respectively).

Overall, the proposed algorithms have the potential to uncover certain aspects of abnormal signaling network behaviors that can contribute to the development of the pathology and may suggest some therapeutic strategies. This study is particularly important in the context of complex disorders with unknown molecular sources, where more than one molecule is observed to be involved in the pathology.

8.3 Modeling and Measurement of Signaling Outcomes Affecting Cell Decision Making

The molecular networks have major roles in the characterization of cell fate. These networks have some specific outputs that initiate important biochemical processes and eventually lead cells to specify their fate. For example, depending on the received signals and dynamics of the network, a possible cell fate could be surviving or initiating apoptosis or moving in a certain direction, and so on. When a molecule is faulty, the entire network may fail, which may affect such important processes. Therefore, characterization of decision-making in cells in response to received signals is important for understanding how cell fate is determined in the absence and presence of such faulty molecules causing incorrect network responses. In this dissertation, we provide a set of decision-theoretic, statistical signal processing and machine learning methods and metrics for modeling and measurement of decision-making processes and signaling outcomes under normal and abnormal conditions, and in the presence of noise and other uncertainties (Ozen et al., 2020).

Due to the noise, signaling malfunctions, or other factors, cells may respond differently to the same input signal. Some of these responses can be erroneous and unexpected. Here we present univariate (Chapter 5) and multivariate (Chapter 6) models and methods for decision making processes and signaling outcome analyses, and as an example, apply them to an important system that is involved in cell survival and death, i.e., the p53 system (Chapter 4) shown in Figure 4.1. The p53 system becomes active due to DNA damage caused by ionizing radiation (IR), and as a result, cell can take two different actions: it can either survive by repairing the DNA or trigger apoptosis. In this context, we model the decisions and signaling outcomes triggered by the p53 system as a binary hypothesis testing problem, where two hypotheses are introduced in Equation (4.1). Regarding these two hypotheses, our approach identifies that there can be two types of incorrect decisions: *false alarm* and *miss*. To compute the likelihood of these decisions, we employ the simulator of Hat et al. (2016) and obtain single cell data of the p53 system by exposing the cells to different radiation doses. We consider PTEN levels in cells as the decision variable, since it was shown as a good predictor of cell fate (Hat et al., 2016). Our analysis focuses on low radiation dose versus high radiation dose scenarios, where we fix the low IR dose at 1 Gy, whereas we set the high IR dose at 2 Gy, 3 Gy, 4 Gy, 5 Gy, 6 Gy, 7 Gy and 8 Gy. We also analyze decision making events and signaling outcomes when an abnormality is present in the p53 system.

The incorrect decision probabilities provided in Equation (4.2) and the overall decision error probability in Equation (4.3) are computed after determining an optimal decision threshold. We obtain this decision threshold using the maximum likelihood principle, which states that the best decision can be made by selecting the hypothesis that

has the maximum probability of occurrence. We compute the decision threshold and error probabilities using single time point data of PTEN levels in both normal and abnormal p53 systems. For 1 Gy vs. 2 Gy and 1 Gy vs. 8 Gy case studies, we present histograms, response distributions, decision thresholds, and false alarm and miss decision regions in normal and abnormal p53 systems in Figures 5.1 and 5.3, respectively. Our decision analysis reveals and quantifies that more erroneous decisions are made when deciding between two nearly the same radiation doses in the normal p53 system (Figure 5.4). On the other hand, the difference between responses is easily identifiable for very low versus very high IR doses. This feature seems not be present in the abnormal p53 systems (Figure 5.4), according to our decision modeling approach. Our decision and outcome analyses and observations are further visualized and confirmed by using the receiver operating characteristic (ROC) curves (Figure 5.5), which are useful graphical tools to study the performance of decision-making systems. We would like to note that these observations are specifically made based on the low versus high IR case studies, e.g., d_0 vs d_1 IRs introduced in Chapter 4 for the p53 system, as an example of a signaling network, in which the low IR dose is fixed to 1 Gy (d_0 = 1 Gy) and the high IR dose is ranging from 2 Gy up to 8 Gy (d_1 = 2, 3, …, 8 Gy). Such conclusions may not be generalized to other biological hypotheses and systems, while the proposed framework and its analytical tools, whose introduction has been the main goal of this study, can be similarly used.

In addition to the above univariate single time point analysis, we extend our signaling outcome modeling framework to dynamical multi-time point measurements and multidimensional decision-making algorithms, to see how the number of decision variables affects the decisions and signaling outcomes over time (Chapter 6). To introduce the

concepts, first we conduct a bivariate analysis, in which bivariate response distributions of cells' PTEN levels measured at two different time instants are shown in Figure 6.1, as well as the optimal maximum likelihood decision boundary. Then we introduce a multivariate dynamic decision modeling framework, for the general scenario where there are more than two decision variables over time. This allows to model and understand how decision error probability changes over time, if at any time the decision is made based on the current observation, together with the previous observations. We observe in Figure 6.2B that as the decision-making strategy incorporates more and more PTEN data at various time instants into its decisions, for the p53 system exposed to two radiation doses of 1 and 2 Gy, the decision error probability reaches its smallest value at a certain time instant. However, adding more data afterwards does not necessarily improve the decision precision, i.e., the decision error probability does not necessarily decrease as N increases with time (Figure 6.2B). We show that this behavior can be related to the correlations that exist among the PTEN levels measured at different times (Figure 6.2C).

Although we focus on multivariate decision making and signaling outcome analysis for one molecule at different time instants, the introduced methods and algorithms are not limited to the outcome analyses for just one molecule. They can be applied to various other scenarios and studies. For instance, they can be used to analyze decision strategies and compute decision error rates based on concentration levels of two or more molecules, measured simultaneously or even at different time instants as shown in the Subsection 6.2.2.

We finally show how the introduced binary decision making and signaling outcome analysis models can be extended to more than two decisions, i.e., more than two hypotheses

(Chapter 7). A ternary scenario with three signaling outcomes is analyzed as an example, and it is shown that under certain conditions, the ternary decision error probability can be higher than the binary one.

The methods and models presented here can be expanded to describe the performance and precision of more complex systems and networks such as the ones whose inputs are multiple ligands or secondary messengers and whose outputs are several transcription factors involved in certain cellular functions. Analyzing the concentration levels of these transcription factors over time using the proposed approaches can model various decisions and signaling outcomes, and their probabilities, in the presence of noise or some cellular abnormalities, and in response to the input signals. Furthermore, the methods and formalism developed in this study are applicable to a wide variety of signaling outcome analyses, decision makings and signal transduction processes where there are two or more possible outcomes. For example, in the context of E-coli chemotaxis, binary decisions (influencing all chemotaxis processes) are either to continue motion in the same direction or to change the flagellum operation mode from run, counterclockwise, to tumble, clockwise, resulting in random direction changes (Watari & Larson, 2010). Based on the network or system of interest and the available data, the hypotheses in Equation (4.1) can be revised, and subsequently the same mathematical framework and algorithms and methods can be applied, using the underlying probability distributions of data. Overall, these decision-theoretic models and signaling outcome analysis methods can be beneficial for better understanding of transition from physiological to pathological conditions such as inflammatory diseases, various cancers, and autoimmune diseases.

8.4 Future Directions

The developed methods and techniques in this dissertation are not limited to only the systems used to elaborate and illustrate them. They can be used to analyze other complex molecular systems with complex experimental data. Thus, one possible future work might be to apply these techniques to other systems to test their robustness as well as to reveal novel insights into other complex biological processes.

Despite the efficiency of the Boolean models (Chapter 2) by not requiring detailed kinetic information to be used and their analysis techniques (Chapter 3) to understand molecular processes, this type of models assume only binary behavior of the system such as being active or inactive. To incorporate more dynamic behavior of these complex systems, one possible future action would be efficiently extending the Boolean models into multi-level models with high data prediction accuracy, which may provide more information about how the whole system behaves.

The proposed univariate and multivariate single cell decision-making and outcome analysis techniques (Chapters 4 - 7) can be studied for different complex systems for different purposes. One can apply these techniques to model cell fates of other systems using different types of measurements such as gene transcription data and single-cell RNA sequencing (scRNA-seq) data. Moreover, one can extend and modify the proposed methods to distinguish different cell types. For instance, it is a common phenomenon that tumor heterogeneity occurs both within and between tumors. This heterogeneity cannot be discerned from conventional bulk transcriptomic studies. On the other hand, one can extend the proposed framework and model scRNA-seq data to distinguish different tumor cell types in a bulk so that more effective and targeted treatments can be administered.

APPENDIX

BOOLEAN EQUATIONS FOR THE ERBB AND T CELL SIGNALING NETWORKS

Tables A.1 and A.2 present Boolean equations of the ERBB and T cell signaling networks in Chapter 3, which are provided by Sahin et al. (2009) and Saez-Rodriguez et al. (2007), respectively. In the equations, "×" is used for the AND operation, "+" is used for the OR operation, and "~" is used for the NOT operation. In Table A.2, the symbol "t" represents the current time whereas "t+1" stands for the next time instant.

Table A.1 Boolean Equations for the ERBB Signaling Network

Molecules	Boolean Equations
AKT1	AKT1 = ERBB1 + ERBB1_2 + ERBB1_3 + ERBB2_3 + IGF1R
c-MYC	c-MYC = AKT1 + MEK1 + ER-α
CDK2	CDK2 = CyclinE1 × (~ p21) × (~ p27)
CDK4	CDK4 = CyclinD1 × (~ p21) × (~ p27)
CDK6	CDK6 = CyclinD1
CyclinD1	CyclinD1 = ER-α × c-MYC × (AKT1 + MEK1)
CyclinE1	CyclinE1 = c-MYC
EGF	EGF: Input
ER-α	ER-α = AKT1 + MEK1
ERBB1	ERBB1 = EGF
ERBB1_2	ERBB1_2 = ERBB1 × ERBB2
ERBB1_3	ERBB1_3 = ERBB1 × ERBB3
ERBB2	ERBB2 = EGF
ERBB2_3	ERBB2_3 = ERBB2 × ERBB3
ERBB3	ERBB3 = EGF
IGF1R	IGF1R = (ER-α + AKT1) × (~ ERBB2_3)
MEK1	MEK1 = ERBB1 + ERBB1_2 + ERBB1_3 + ERBB2_3 + IGF1R
p21	p21 = ER-α × (~ CDK4) × (~ AKT1) × (~ c-MYC)
p27	p27 = ER-α × (~ CDK4) × (~ CDK2) × (~ AKT1) × (~ c-MYC)
pRB	pRB = (CDK4 × CDK6) + (CDK4 × CDK6 × CDK2)

Table A.2 Boolean Equations for the T Cell Signaling Network

Molecules	Boolean Equations
abl(t)	abl(t) = lckp1(t) + fyn(t)
akap79	akap79 = 0
ap1(t)	ap1(t) = fos(t) × jun(t)
bad(t)	bad(t) = ~ pkb(t)
bcat(t)	bcat(t) = ~ gsk3(t)
bcl10	bcl10 = 1
bclx1(t)	bclx1 = ~ bad(t)
ca(t)	ca(t) = ip3(t)
cabin1(t)	cabin1(t) = ~ camk4(t)
calcin(t)	calcin(t) = (~ cabin1(t)) × (~ akap79) × (~ calpr1) × cam(t)
calpr1	calpr1 = 0
cam(t)	cam(t) = ca(t)
camk2(t)	camk2(t) = cam(t)
camk4(t)	camk4(t) = cam(t)
card11	card11 = 1
card11a(t)	card11a(t) = card11 × bcl10 × malt1
cblc(t+1)	cblb(t+1) = ~ cd28
ccblp1(t+1)	ccblp1(t+1) = zap70(t)
ccblp2(t+1)	ccblp2(t+1) = fyn(t)
cd28	Input
cd4	Input
cd45	cd45 = 1
cdc42	cdc42 = 0
cre(t)	cre(t) = creb(t)
creb(t)	creb(t) = rsk(t)
csk(t)	csk(t) = pag(t)
cyc1(t)	cyc1(t) = ~ gsk3(t)
dag(t)	dag(t) = (~ dgk(t)) × plcga(t)
dgk(t+1)	dgk(t+1) = tcrb(t)
erk(t)	erk(t) = mek(t)
fkhr(t)	fkhr(t) = ~ pkb(t)
fos(t)	fos(t) = erk(t)

Molecules	Boolean Equations
fyn(t)	fyn(t) = tcrb(t) + (lckp1(t) × cd45)
gab2(t+1)	gab2(t+1) = lat(t) × zap70(t) × (gads(t) + grb2(t))
gadd45	gadd45 = 1
gads(t)	gads(t) = lat(t)
Gap	gap = 0
grb2(t)	grb2(t) = lat(t)
gsk3(t)	gsk3(t) = ~ pkb(t)
hpk1(t)	hpk1(t) = lat(t)
ikb(t)	ikb(t) = ~ ikkab(t)
ikkab(t)	ikkab(t) = ikkg(t) × camk2(t)
ikkg(t)	ikkg(t) = pkcth(t) × card11a(t)
ip3(t)	ip3(t) = plcga(t)
itk(t)	itk(t) = slp76(t) × zap70(t) × pip3(t)
jnk(t)	jnk(t) = mekk1(t) + mkk4(t)
jun(t)	jun(t) = jnk(t)
lat(t)	lat(t) = zap70(t)
lckp1(t)	lckp1(t) = (~ shp1(t)) × (~ csk(t)) × cd45 × cd4
lckp2(t)	lckp2(t) = tcrb(t)
malt1	malt1 = 1
mek(t)	mek(t) = raf(t)
mekk1(t)	mekk1(t) = hpk1(t) + cdc42 + rac1p2(t)
mkk4(t)	mkk4(t) = mlk3(t) + mekk1(t)
mlk3(t)	mlk3(t) = hpk1(t) + rac1p1(t)
nfat(t)	nfat(t) = calcin(t)
nfkb(t)	nfkb(t) = ~ ikb(t)
p21c(t)	p21c(t) = ~ pkb(t)
p27k(t)	p27k(t) = ~ pkb(t)
p38(t)	p38(t) = ((~ gadd45) × zap70(t)) + mekk1(t)
p70s(t)	p70s(t) = pdk1(t)
pag(t)	pag(t) = ~ tcrb(t)
pag(t+1)	pag(t+1) = fyn(t)
pdk1(t)	pdk1(t) = pip3(t)
pi3k(t)	pi3k(t) = ((~ cblb(t)) × X(t)) + ((~ cblb(t)) × lckp2(t))

Molecules	Boolean Equations
pip3(t)	pip3(t) = pi3k(t) × (~ ship1) × (~ pten)
pkb(t)	pkb(t) = pdk1(t)
pkcth(t)	pkcth(t) = pdk1(t) × dag(t) × vav1(t)
plcga(t)	plcga(t) = plcgb(t) × (~ ccblp2(t)) × slp76(t) × zap70(t) × vav1(t) × (itk(t) + rlk(t))
plcgb(t)	plcgb(t) = lat(t)
Pten	pten = 0
rac1p1(t)	rac1p1(t) = vav1(t)
rac1p2(t)	rac1p2(t) = vav3(t)
raf(t)	raf(t) = ras(t)
ras(t)	ras(t) = (~ gap) × rasgrp(t) × sos(t)
rasgrp(t)	rasgrp(t) = dag(t)
rlk(t)	rlk(t) = lckp1(t)
rsk(t)	rsk(t) = erk(t)
sh3bp2(t)	sh3bp2(t) = zap70(t) × lat(t)
ship1	ship1 = 0
shp1(t+1)	shp1(t+1) = (~ erk(t)) × lckp1(t)
shp2(t)	shp2(t) = gab2(t)
slp76(t)	slp76(t) = (~ gab2(t)) × zap70(t) × gads(t)
sos(t)	sos(t) = grb2(t)
sre(t)	sre(t) = rac1p2(t) + cdc42
tcrb(t)	tcrb(t) = (~ ccblp1(t)) × tcrlig
Tcrlig	Input
tcrp(t)	tcrp(t) = (tcrb(t) × lckp1(t)) + (tcrb(t) × fyn(t))
vav1(t)	vav1(t) = (sh3bp2(t) × zap70(t)) + X(t)
vav3(t)	vav3(t) = sh3bp2(t)
X(t)	X(t) = cd28
zap70(t)	zap70(t) = (~ ccblp1(t)) × abl(t) × tcrp(t)

REFERENCES

Abdi, A., & Emamian, E. S. (2010). Fault diagnosis engineering in molecular signaling networks: An overview and applications in target discovery. *Chemistry and Biodiversity*, *7*(5), 1111–1123. https://doi.org/10.1002/cbdv.200900315

Abdi, A., Tahoori, M. B., & Emamian, E. S. (2009). Identification of critical molecules via fault diagnosis engineering. *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC*, 4898–4901. https://doi.org/10.1109/IEMBS.2009.5332823

Abdi, A., Tahoori, M. B., & Emamian, E. S. (2008). Fault diagnosis engineering of digital circuits can identify vulnerable molecules in complex cellular pathways. *Science Signaling*, *1*(42). https://doi.org/10.1126/scisignal.2000008

Albert, I., Thakar, J., Li, S., Zhang, R., & Albert, R. (2008). Boolean network simulations for life scientists. *Source Code for Biology and Medicine*, *3*. https://doi.org/10.1186/1751-0473-3-16

Albert, R. (2007). Network inference, analysis, and modeling in systems biology. *Plant Cell*, *19*(11), 3327–3338. https://doi.org/10.1105/tpc.107.054700

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

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). *Molecular biology of the cell* (4th ed.). New York: Garland Science.

Aldridge, B. B., Burke, J. M., Lauffenburger, D. A., & Sorger, P. K. (2006). Physicochemical modelling of cell signalling pathways. *Nature Cell Biology*, *8*, 1195–1203. https://doi.org/10.1038/ncb1497

Aldridge, B. B., Saez-Rodriguez, J., Muhlich, J. L., Sorger, P. K., & Lauffenburger, D. A. (2009). Fuzzy logic analysis of kinase pathway crosstalk in TNF/EGF/Insulin-induced signaling. *PLoS Computational Biology*, *5*(4). https://doi.org/10.1371/journal.pcbi.1000340

Arias, A. M., & Stewart, A. (2002). *Molecular principles of animal development* (Vol. 1). Oxford University Press.

Azpeitia, E., Muñoz, S., González-Tokman, D., Martínez-Sánchez, M. E., Weinstein, N., Naldi, A., Álvarez-Buylla, E. R., Rosenblueth, D. A., & Mendoza, L. (2017). The combination of the functionalities of feedback circuits is determinant for the attractors' number and size in pathway-like Boolean networks. *Scientific Reports*, *7*. https://doi.org/10.1038/srep42023

Bakkenist, C. J., & Kastan, M. B. (2003). DNA damage activates ATM through intermolecular autophosphorylation and dimer dissociation. *Nature*, *421*, 499–506. https://doi.org/10.1038/nature01368

Balázsi, G., Van Oudenaarden, A., & Collins, J. J. (2011). Cellular decision making and biological noise: From microbes to mammals. *Cell*, *144*(6), 910–925. https://doi.org/10.1016/j.cell.2011.01.030

Banin, S., Moyal, L., Shieh, S. Y., Taya, Y., Anderson, C. W., Chessa, L., Smorodinsky, N. I., Prives, C., Reiss, Y., Shiloh, Y., & Ziv, Y. (1998). Enhanced phosphorylation of p53 by ATM in response to DNA damage. *Science*, *281*(5383), 1674–1677. https://doi.org/10.1126/science.281.5383.1674

Barak, Y., Juven, T., Haffner, R., & Oren, M. (1993). mdm2 expression is induced by wild type p53 activity. *EMBO Journal*, *12*(2), 461–468. https://doi.org/10.1002/j.1460-2075.1993.tb05678.x

Barshir, R., Basha, O., Eluk, A., Smoly, I. Y., Lan, A., & Yeger-Lotem, E. (2013). The TissueNet database of human tissue protein-protein interactions. *Nucleic Acids Research*, *41*, D841–D844. https://doi.org/10.1093/nar/gks1198

Bogdał, M. N., Hat, B., Kochańczyk, M., & Lipniacki, T. (2013). Levels of pro-apoptotic regulator Bad and anti-apoptotic regulator Bcl-xL determine the type of the apoptotic logic gate. *BMC Systems Biology*, *7*. https://doi.org/10.1186/1752-0509-7-67

Bulavin, D. V., Demidov, O. N., Saito, S., Kauraniemi, P., Phillips, C., Amundson, S. A., Ambrosino, C., Sauter, G., Nebreda, A. R., Anderson, C. W., Kallioniemi, A., Fornace, A. J., & Appella, E. (2002). Amplification of PPM1D in human tumors abrogates p53 tumor-suppressor activity. *Nature Genetics*, *31*(2), 210–215. https://doi.org/10.1038/ng894

Camargo, L. M., Collura, V., Rain, J. C., Mizuguchi, K., Hermjakob, H., Kerrien, S., Bonnert, T. P., Whiting, P. J., & Brandon, N. J. (2007). Disrupted in Schizophrenia 1 interactome: Evidence for the close connectivity of risk genes and a potential synaptic basis for schizophrenia. *Molecular Psychiatry*, *12*(1), 74–86. https://doi.org/10.1038/sj.mp.4001880

Canman, C. E., Lim, D. S., Cimprich, K. A., Taya, Y., Tamai, K., Sakaguchi, K., Appella, E., Kastan, M. B., & Siliciano, J. D. (1998). Activation of the ATM kinase by ionizing radiation and phosphorylation of p53. *Science*, *281*(5383), 1677–1679. https://doi.org/10.1126/science.281.5383.1677

Castellino, R. C., De Bortoli, M., Lu, X., Moon, S. H., Nguyen, T. A., Shepard, M. A., Rao, P. H., Donehower, L. A., & Kim, J. Y. H. (2008). Medulloblastomas overexpress the p53-inactivating oncogene WIP1/PPM1D. *Journal of Neuro-Oncology*, *86*(3), 245–256. https://doi.org/10.1007/s11060-007-9470-8

Chaouiya, C. (2007). Petri net modelling of biological networks. *Briefings in Bioinformatics*, *8*(4), 210–219. https://doi.org/10.1093/bib/bbm029

Chaouiya, C., Naldi, A., & Thieffry, D. (2012). Logical modelling of gene regulatory networks with GINsim. *Methods in Molecular Biology*, *804*, 463–479. https://doi.org/10.1007/978-1-61779-361-5_23

Chaouiya, C., & Remy, E. (2013). Logical modelling of regulatory networks, methods and applications. *Bulletin of Mathematical Biology*, *75*, 891–895. https://doi.org/10.1007/s11538-013-9863-0

Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L., & Cesareni, G. (2007). MINT: The molecular interaction database. *Nucleic Acids Research*, *35*, D572–D574. https://doi.org/10.1093/nar/gkl950

Chen, K. C., Calzone, L., Csikasz-Nagy, A., Cross, F. R., Novak, B., & Tyson, J. J. (2004). Integrative analysis of cell cycle control in budding yeast. *Molecular Biology of the Cell*, *15*(8), 3841–3862. https://doi.org/10.1091/mbc.E03-11-0794

Cheong, R., Rhee, A., Wang, C. J., Nemenman, I., & Levchenko, A. (2011). Information transduction capacity of noisy biochemical signaling networks. *Science*, *334*(6054), 354–358. https://doi.org/10.1126/science.1204553

Choi, J., Nannenga, B., Demidov, O. N., Bulavin, D. V., Cooney, A., Brayton, C., Zhang, Y., Mbawuike, I. N., Bradley, A., Appella, E., & Donehower, L. A. (2002). Mice deficient for the wild-type p53-induced phosphatase gene (Wip1) exhibit defects in reproductive organs, immune function, and cell cycle control. *Molecular and Cellular Biology*, *22*(4), 1094–1105. https://doi.org/10.1128/mcb.22.4.1094-1105.2002

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to algorithms* (3rd ed.). Massachusetts: The MIT Press.

Csermely, P., Korcsmáros, T., Kiss, H. J. M., London, G., & Nussinov, R. (2013). Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. In *Pharmacology and Therapeutics* (Vol. 138, Issue 3, pp. 333–408). https://doi.org/10.1016/j.pharmthera.2013.01.016

DasGupta, B., & Liang, J. (2016). Models and algorithms for biomolecules and molecular networks. In *Models and Algorithms for Biomolecules and Molecular Networks* (1st ed.). Wiley-IEEE Press. https://doi.org/10.1002/9781119162254

De Jong, H. (2002). Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, *9*(1), 67–103. https://doi.org/10.1089/10665270252833208

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). New York: John Wiley & Sons.

Eduati, F., Jaaks, P., Wappler, J., Cramer, T., Merten, C. A., Garnett, M. J., & Saez-Rodriguez, J. (2020). Patient-specific logic models of signaling pathways from screenings on cancer biopsies to prioritize personalized combination therapies. *Molecular Systems Biology*, *16*(2). https://doi.org/10.15252/msb.20188664

Elmore, S. (2007). Apoptosis: A review of programmed cell death. *Toxicologic Pathology*, *35*(4), 495–516. https://doi.org/10.1080/01926230701320337

Emamian, E. S. (2012). AKT/GSK3 signaling pathway and schizophrenia. *Frontiers in Molecular Neuroscience*, *5*. https://doi.org/10.3389/fnmol.2012.00033

Emamian, E. S., Kaytor, M. D., Duvick, L. A., Zu, T., Tousey, S. K., Zoghbi, H. Y., Clark, H. B., & Orr, H. T. (2003). Serine 776 of ataxin-1 is critical for polyglutamine-induced disease in SCA1 transgenic mice. *Neuron*, *38*(3), 375–387. https://doi.org/10.1016/S0896-6273(03)00258-7

Emmert-Streib, F., Dehmer, M., & Haibe-Kains, B. (2014). Gene regulatory networks and their applications: Understanding biological and medical problems in terms of networks. *Frontiers in Cell and Developmental Biology*, *2*. https://doi.org/10.3389/fcell.2014.00038

Erdi, P., & Toth, J. (1989). *Mathematical models of chemical reactions: Theory and applications of deterministic and stochastic models*. Manchester University Press.

Eungdamrong, N. J., & Iyengar, R. (2004). Modeling cell signaling networks. *Biology of the Cell*, *96*(5), 355–362. https://doi.org/10.1016/j.biolcel.2004.03.004

Fiscella, M., Zhang, H., Fan, S., Sakaguchi, K., Shen, S., Mercer, W. E., Vande Woude, G. F., O'Connor, P. M., & Appella, E. (1997). Wip1, a novel human protein phosphatase that is induced in response to ionizing radiation in a p53-dependent manner. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(12), 6048–6053. https://doi.org/10.1073/pnas.94.12.6048

Fukunaga, K. (1990). *Introduction to statistical pattern recognition* (2nd ed.). San Diego: Academic Press.

Garcia-Ojalvo, J., & Martinez Arias, A. (2012). Towards a statistical mechanics of cell fate decisions. In *Current Opinion in Genetics and Development* (Vol. 22, Issue 6, pp. 619–626). https://doi.org/10.1016/j.gde.2012.10.004

Geva-Zatorsky, N., Rosenfeld, N., Itzkovitz, S., Milo, R., Sigal, A., Dekel, E., Yarnitzky, T., Liron, Y., Polak, P., Lahav, G., & Alon, U. (2006). Oscillations and variability in the p53 system. *Molecular Systems Biology*, *2*. https://doi.org/10.1038/msb4100068

Goodwin, B. C. (1963). Temporal organization in cells; a dynamic theory of cellular control processes. In *Temporal organization in cells; a dynamic theory of cellular control processes.* Academic Press. https://doi.org/10.5962/bhl.title.6268

Grabowski, F., Czyż, P., Kochańczyk, M., & Lipniacki, T. (2019). Limits to the rate of information transmission through the MAPK pathway. *Journal of the Royal Society Interface*, *16*(152). https://doi.org/10.1098/rsif.2018.0792

Griffiths, J. A., Scialdone, A., & Marioni, J. C. (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. *Molecular Systems Biology*, *14*(4). https://doi.org/10.15252/msb.20178046

Guo, L., Lin, L., Wang, X., Gao, M., Cao, S., Mai, Y., Wu, F., Kuang, J., Liu, H., Yang, J., Chu, S., Song, H., Li, D., Liu, Y., Wu, K., Liu, J., Wang, J., Pan, G., Hutchins, A. P., … Chen, J. (2019). Resolving cell fate decisions during somatic cell reprogramming by single-cell RNA-Seq. *Molecular Cell*, *73*(4), 815–829. https://doi.org/10.1016/j.molcel.2019.01.042

Guziolowski, C., Videla, S., Eduati, F., Thiele, S., Cokelaer, T., Siegel, A., & Saez-Rodriguez, J. (2013). Exhaustively characterizing feasible logic models of a signaling network using Answer Set Programming. *Bioinformatics*, *29*(18), 2320–2326. https://doi.org/10.1093/bioinformatics/btt393

Habibi, I., Cheong, R., Lipniacki, T., Levchenko, A., Emamian, E. S., & Abdi, A. (2017). Computation and measurement of cell decision making errors using single cell data. *PLoS Computational Biology*, *13*(4). https://doi.org/10.1371/journal.pcbi.1005436

Habibi, I., Emamian, E. S., & Abdi, A. (2014a). Advanced fault diagnosis methods in molecular networks. *PLoS ONE*, *9*(10).

Habibi, I., Emamian, E. S., & Abdi, A. (2014b). Quantitative analysis of intracellular communication and signaling errors in signaling networks. *BMC Systems Biology*, *8*. https://doi.org/10.1186/s12918-014-0089-z

Han, J., Chen, H., Boykin, E., & Fortes, J. (2011). Reliability evaluation of logic circuits using probabilistic gate models. *Microelectronics Reliability*, *51*(2), 468–476. https://doi.org/10.1016/j.microrel.2010.07.154

Handorf, T., & Klipp, E. (2012). Modeling mechanistic biological networks: An advanced Boolean approach. *Bioinformatics*, *28*(4), 557–563. https://doi.org/10.1093/bioinformatics/btr697

Hasty, J., McMillen, D., Isaacs, F., & Collins, J. J. (2001). Computational studies of gene regulatory networks: In numero molecular biology. *Nature Reviews Genetics*, *2*, 268–279. https://doi.org/10.1038/35066056

Hat, B., Kochańczyk, M., Bogdał, M. N., & Lipniacki, T. (2016). Feedbacks, bifurcations, and cell fate decision-making in the p53 system. *PLoS Computational Biology*, *12*(2). https://doi.org/10.1371/journal.pcbi.1004787

Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., & Guthke, R. (2009). Gene regulatory network inference: Data integration in dynamic models-A review. *BioSystems*, *96*(1), 86–103. https://doi.org/10.1016/j.biosystems.2008.12.004

Helikar, T., Konvalina, J., Heidel, J., & Rogers, J. A. (2008). Emergent decision-making in biological signal transduction networks. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(6), 1913–1918. https://doi.org/10.1073/pnas.0705088105

Hill, S. M., Heiser, L. M., Cokelaer, T., Linger, M., Nesser, N. K., Carlin, D. E., Zhang, Y., Sokolov, A., Paull, E. O., Wong, C. K., Graim, K., Bivol, A., Wang, H., Zhu, F., Afsari, B., Danilova, L. V., Favorov, A. V., Lee, W. S., Taylor, D., … Zi, Z. (2016). Inferring causal molecular networks: Empirical assessment through a community-based effort. *Nature Methods*, *13*, 310–318. https://doi.org/10.1038/nmeth.3773

Hirasawa, A., Saito-Ohara, F., Inoue, J., Aoki, D., Susumu, N., Yokoyama, T., Nozawa, S., Inazawa, J., & Imoto, I. (2003). Association of 17q21-q24 gain in ovarian clear cell adenocarcinomas with poor prognosis and identification of PPM1D and APPBP2 as likely amplification targets. *Clinical Cancer Research*, *9*(6), 1995–2004.

Hlobilkova, A., Knillova, J., Svachova, M., Skypalova, P., Krystof, V., & Kolar, Z. (2006). Tumour suppressor PTEN regulates cell cycle and protein kinase B/Akt pathway in breast cancer cells. *Anticancer Research*, *26*(2A), 1015–1022.

Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, *19*(17), 2271–2282. https://doi.org/10.1093/bioinformatics/btg313

IBM. (n.d.). *IBM ILOG CPLEX Optimization Studio*. Retrieved February 14, 2021, from https://www.ibm.com/products/ilog-cplex-optimization-studio

Ideker, T. E., Thorsson, V., & Karp, R. M. (2000). Discovery of regulatory interactions through perturbation: inference and experimental design. *Pacific Symposium on Biocomputing*, 302–313. https://doi.org/10.1142/9789814447331_0029

Ideker, T., Galitski, T., & Hood, L. (2001). A new approach to decoding life: Systems biology. *Annual Review of Genomics and Human Genetics*, *2*, 343–372. https://doi.org/10.1146/annurev.genom.2.1.343

Jeong, H., Tombor, B., Albert, R., Oltval, Z. N., & Barabásl, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, *407*, 651–654. https://doi.org/10.1038/35036627

Kampen, V. N. G. (1981). Stochastic Processes in Physics and Chemistry. In *Stochastic Processes in Physics and Chemistry* (1st ed.). North Holland. https://doi.org/10.1016/B978-0-444-52965-7.X5000-4

Karlebach, G., & Shamir, R. (2008). Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, *9*, 770–780. https://doi.org/10.1038/nrm2503

Kastan, M. B., Onyekwere, O., Sidransky, D., Vogelstein, B., & Craig, R. W. (1991). Participation of p53 protein in the cellular response to DNA damage. *Cancer Research*, *51*(23), 6304–6311.

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

Kay, S. M. (1998). *Fundamentals of statistical signal processing: Detection theory*. New Jersey: Prentice-Hall PTR.

Kitano, H. (2002). Systems biology: A brief overview. In *Science* (Vol. 295, Issue 5560, pp. 1662–1664). https://doi.org/10.1126/science.1069492

Kohavi, I., & Kohavi, Z. (1972). Detection of multiple faults in combinational logic networks. *IEEE Transactions on Computers*, *C–21*(6), 556–568. https://doi.org/10.1109/TC.1972.5009008

Kolitz, S. E., & Lauffenburger, D. A. (2012). Measurement and modeling of signaling at the single-cell level. *Biochemistry*, *51*(38), 7433–7443. https://doi.org/10.1021/bi300846p

Le Novere, N. (2015). Quantitative and logic modelling of molecular and gene networks. *Nature Reviews Genetics*, *16*, 146–158. https://doi.org/10.1038/nrg3885

Levine, A. J. (1997). p53, the cellular gatekeeper for growth and division. *Cell*, *88*(3), 323–331. https://doi.org/10.1016/S0092-8674(00)81871-1

Levine, M., & Davidson, E. H. (2005). Gene regulatory networks for development. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(14), 4936–3942. https://doi.org/10.1073/pnas.0408031102

Li, J., Yang, Y., Peng, Y., Austin, R. J., Van Eyndhoven, W. G., Nguyen, K. C. Q., Gabriele, T., McCurrach, M. E., Marks, J. R., Hoey, T., Lowe, S. W., & Powers, S. (2002). Oncogenic properties of PPM1D located within a breast cancer amplification epicenter at 17q23. *Nature Genetics*, *31*(2), 133–134. https://doi.org/10.1038/ng888

Lliakis, G. (1991). The role of DNA double strand breaks in lonizing radiation-induced killing of eukaryotic cells. *BioEssays*, *13*(12), 641–648. https://doi.org/10.1002/bies.950131204

Lu, X., Ma, O., Nguyen, T. A., Jones, S. N., Oren, M., & Donehower, L. A. (2007). The Wip1 phosphatase acts as a gatekeeper in the p53-Mdm2 autoregulatory loop. *Cancer Cell*, *12*(4), 342–354. https://doi.org/10.1016/j.ccr.2007.08.033

Machado, D., Costa, R. S., Rocha, M., Ferreira, E. C., Tidor, B., & Rocha, I. (2011). Modeling formalisms in systems biology. *AMB Express*, 1–45. https://doi.org/10.1186/2191-0855-1-45

Maslov, S., & Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science*, *296*(5569), 910–913. https://doi.org/10.1126/science.1065103

Maya, R., Balass, M., Kim, S. T., Shkedy, D., Martinez Leal, J. F., Shifman, O., Moas, M., Buschmann, T., Ronai, Z., Shiloh, Y., Kastan, M. B., Katzir, E., & Oren, M. (2001). ATM-dependent phosphorylation of Mdm2 on serine 395: Role in p53 activation by DNA damage. *Genes and Development*, *15*(9), 1067–1077. https://doi.org/10.1101/gad.886901

Melas, I. N., Samaga, R., Alexopoulos, L. G., & Klamt, S. (2013). Detecting and removing inconsistencies between experimental data and signaling network topologies using integer linear programming on interaction graphs. *PLoS Computational Biology*, *9*(9). https://doi.org/10.1371/journal.pcbi.1003204

Miryala, S. K., Anbarasu, A., & Ramaiah, S. (2018). Discerning molecular interactions: A comprehensive review on biomolecular interaction databases and network analysis tools. *Gene*, *642*, 84–94. https://doi.org/10.1016/j.gene.2017.11.028

Mitchell, M. (1998). *An introduction to genetic algorithms*. Cambridge: MIT Press.

Mitsos, A., Melas, I. N., Siminelakis, P., Chairakaki, A. D., Saez-Rodriguez, J., & Alexopoulos, L. G. (2009). Identifying drug effects via pathway alterations using an integer linear programming optimization formulation on phosphoproteomic data. *PLoS Computational Biology*, *5*(12). https://doi.org/10.1371/journal.pcbi.1000591

Mohammed, H., Hernando-Herraez, I., Savino, A., Scialdone, A., Macaulay, I., Mulas, C., Chandra, T., Voet, T., Dean, W., Nichols, J., Marioni, J. C., & Reik, W. (2017). Single-cell landscape of transcriptional heterogeneity and cell fate decisions during mouse early gastrulation. *Cell Reports*, *20*(5), 1215–1228. https://doi.org/10.1016/j.celrep.2017.07.009

Moris, N., Pina, C., & Arias, A. M. (2016). Transition states and cell fate decisions in epigenetic landscapes. In *Nature Reviews Genetics* (Vol. 17, Issue 11, pp. 693–703). https://doi.org/10.1038/nrg.2016.98

Morris, M. K., Saez-Rodriguez, J., Clarke, D. C., Sorger, P. K., & Lauffenburger, D. A. (2011). Training signaling pathway maps to biochemical data with constrained fuzzy logic: Quantitative analysis of liver cell responses to inflammatory stimuli. *PLoS Computational Biology*, *7*(3). https://doi.org/10.1371/journal.pcbi.1001099

Morris, M. K., Saez-Rodriguez, J., Sorger, P. K., & Lauffenburger, D. A. (2010). Logic-based models for the analysis of cell signaling networks. *Biochemistry*, *49*(15), 3216–3224. https://doi.org/10.1021/bi902202q

Murata, T. (1989). Petri nets: Properties, analysis and applications. *Proceedings of the IEEE*, *77*(4), 541–580. https://doi.org/10.1109/5.24143

Narula, J., Kuchina, A., Zhang, F., Fujita, M., Süel, G. M., & Igoshin, O. A. (2016). Slowdown of growth controls cellular differentiation. *Molecular Systems Biology*, *12*(5). https://doi.org/10.15252/msb.20156691

Ozen, M., Lipniacki, T., Levchenko, A., Emamian, E. S., & Abdi, A. (2020). Modeling and measurement of signaling outcomes affecting decision making in noisy intracellular networks using machine learning methods. *Integrative Biology*, *12*(5), 122–138. https://doi.org/10.1093/intbio/zyaa009

Papoulis, A. (1991). *Probability, random variables, and stochastic processes* (3rd ed.). New York: McGraw-Hill.

Raue, A., Schilling, M., Bachmann, J., Matteson, A., Schelke, M., Kaschek, D., Hug, S., Kreutz, C., Harms, B. D., Theis, F. J., Klingmüller, U., & Timmer, J. (2013). Lessons learned from quantitative dynamical modeling in systems biology. *PLoS ONE*, *8*(9). https://doi.org/10.1371/journal.pone.0074335

Rauta, J., Alarmo, E. L., Kauraniemi, P., Karhu, R., Kuukasjärvi, T., & Kallioniemi, A. (2006). The serine-threonine protein phosphatase PPM1D is frequently activated through amplification in aggressive primary breast tumours. *Breast Cancer Research and Treatment*, *95*(3), 257–263. https://doi.org/10.1007/s10549-005-9017-7

Rothkamm, K., Krüger, I., Thompson, L. H., & Löbrich, M. (2003). Pathways of DNA double-strand break repair during the mammalian cell cycle. *Molecular and Cellular Biology*, *23*(16), 5706–5715. https://doi.org/10.1128/mcb.23.16.5706-5715.2003

Saadatpour, A., & Albert, R. (2012). Discrete dynamic modeling of signal transduction networks. In *Methods in Molecular Biology* (Vol. 880, pp. 255–272). https://doi.org/10.1007/978-1-61779-833-7_12

Saadatpour, A., & Albert, R. (2013). Boolean modeling of biological regulatory networks: A methodology tutorial. *Methods*, *62*(1), 3–12. https://doi.org/10.1016/j.ymeth.2012.10.012

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., & Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, *308*(5721), 523–529. https://doi.org/10.1126/science.1105809

Saez-Rodriguez, J., Alexopoulos, L. G., Epperlein, J., Samaga, R., Lauffenburger, D. A., Klamt, S., & Sorger, P. K. (2009). Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Molecular Systems Biology*, *5*. https://doi.org/10.1038/msb.2009.87

Saez-Rodriguez, J., Simeoni, L., Lindquist, J. A., Hemenway, R., Bommhardt, U., Arndt, B., Haus, U. U., Weismantel, R., Gilles, E. D., Klamt, S., & Schraven, B. (2007). A logical model provides insights into T cell receptor signaling. *PLoS Computational Biology*, *3*(8). https://doi.org/10.1371/journal.pcbi.0030163

Sagar, & Grün, D. (2020). Deciphering cell date decision by integrated single-cell sequencing analysis. *Annual Review of Biomedical Data Science*, *3*, 1–22. https://doi.org/10.1146/annurev-biodatasci-111419-091750

Sahin, Ö., Fröhlich, H., Löbke, C., Korf, U., Burmester, S., Majety, M., Mattern, J., Schupp, I., Chaouiya, C., Thieffry, D., Poustka, A., Wiemann, S., Beissbarth, T., & Arlt, D. (2009). Modeling ERBB receptor-regulated G1/S transition to find novel targets for de novo trastuzumab resistance. *BMC Systems Biology*, *3*(1). https://doi.org/10.1186/1752-0509-3-1

Saito-Ohara, F., Imoto, I., Inoue, J., Hosoi, H., Nakagawara, A., Sugimoto, T., & Inazawa, J. (2003). PPM1D is a potential target for 17q gain in neuroblastoma. *Cancer Research*, *63*(8), 1876–1883.

Saito, S., Goodarzi, A. A., Higashimoto, Y., Noda, Y., Lees-Miller, S. P., Appella, E., & Anderson, C. W. (2002). ATM mediates phosphorylation at multiple p53 sites, including Ser46, in response to ionizing radiation. *Journal of Biological Chemistry*, *277*(15), 12491–12494. https://doi.org/10.1074/jbc.C200093200

Samaga, R., & Klamt, S. (2013). Modeling approaches for qualitative and semi-quantitative analysis of cellular signaling networks. *Cell Communication and Signaling*, *11*. https://doi.org/10.1186/1478-811X-11-43

Schlitt, T., & Brazma, A. (2007). Current approaches to gene regulatory network modelling. In *BMC Bioinformatics* (Vol. 8). https://doi.org/10.1186/1471-2105-8-S6-S9

Sharan, R., & Karp, R. M. (2013). Reconstructing boolean models of signaling. *Journal of Computational Biology*, *20*(3), 249–257. https://doi.org/10.1089/cmb.2012.0241

Shieh, S. Y., Taya, Y., & Prives, C. (1999). DNA damage-inducible phosphorylation of p53 at N-terminal sites including a novel site, Ser20, requires tetramerization. *EMBO Journal*, *18*(7), 1815–1823. https://doi.org/10.1093/emboj/18.7.1815

Shreeram, S., Weng, K. H., Demidov, O. N., Kek, C., Yamaguchi, H., Fornace, A. J., Anderson, C. W., Appella, E., & Bulavin, D. V. (2006). Regulation of ATM/p53-dependent suppression of myc-induced lymphomas by Wip1 phosphatase. *Journal of Experimental Medicine*, *203*(13), 2793–2799. https://doi.org/10.1084/jem.20061563

Siliciano, J. D., Canman, C. E., Taya, Y., Sakaguchi, K., Appella, E., & Kastan, M. B. (1997). DNA damage induces phosphorylation of the amino terminus of p53. *Genes and Development*, *11*(24), 3471–3481. https://doi.org/10.1101/gad.11.24.3471

Smith, A. E., Slepchenko, B. M., Schaff, J. C., Loew, L. M., & Macara, I. G. (2002). Systems analysis of ran transport. *Science*, *295*(5554), 488–491. https://doi.org/10.1126/science.1064732

Somogyi, R., & Greller, L. D. (2001). The dynamics of molecular networks: Applications to therapeutic discovery. *Drug Discovery Today*, *6*(24), 1267–1277. https://doi.org/10.1016/S1359-6446(01)02096-7

Sontag, E. (1998). Mathematical control theory: Deterministic finite dimensional systems. In *Texts in applied mathematics: Vol. 2nd ed*. New York: Springer.

Sontag, E. D. (2005). Molecular systems biology and control. *European Journal of Control*, *11*(4–5), 396–435. https://doi.org/10.3166/ejc.11.396-435

Stambolic, V., MacPherson, D., Sas, D., Lin, Y., Snow, B., Jang, Y., Benchimol, S., & Mak, T. W. (2001). Regulation of PTEN transcription by p53. *Molecular Cell*, *8*(2), 317–325. https://doi.org/10.1016/S1097-2765(01)00323-9

Stoll, G., Viara, E., Barillot, E., & Calzone, L. (2012). Continuous time boolean modeling for biological signaling: application of Gillespie algorithm. *BMC Systems Biology*, *6*. https://doi.org/10.1186/1752-0509-6-116

Stolovitzky, G., Monroe, D., & Califano, A. (2007). Dialogue on reverse-engineering assessment and methods: The DREAM of high-throughput pathway inference. *Annals of the New York Academy of Sciences*, *1115*, 1–22. https://doi.org/10.1196/annals.1407.021

Svensson, V., Natarajan, K. N., Ly, L. H., Miragaia, R. J., Labalette, C., Macaulay, I. C., Cvejic, A., & Teichmann, S. A. (2017). Power analysis of single-cell rnA-sequencing experiments. *Nature Methods*, *14*(4), 381–387. https://doi.org/10.1038/nmeth.4220

Takekawa, M., Adachi, M., Nakahata, A., Nakayama, I., Itoh, F., Tsukuda, H., Taya, Y., & Imai, K. (2000). p53-inducible Wip1 phosphatase mediates a negative feedback regulation of p38 MAPK-p53 signaling in response to UV radiation. *EMBO Journal*, *19*(23), 6517–6526. https://doi.org/10.1093/emboj/19.23.6517

Tudelska, K., Markiewicz, J., Kochańczyk, M., Czerkies, M., Prus, W., Korwek, Z., Abdi, A., Błoński, S., Kaźmierczak, B., & Lipniacki, T. (2017). Information processing in the NF-κB pathway. *Scientific Reports*, *7*(1).

Van Trees, H. L., Bell, K. L., & Tian, Z. (2013). *Detection, estimation and modulation theory, part I: Detection, estimation, and filtering theory* (2nd ed.). New Jersey: John Wiley & Sons.

Videla, S., Guziolowski, C., Eduati, F., Thiele, S., Grabe, N., Saez-Rodriguez, J., & Siegel, A. (2012). Revisiting the training of logic models of protein signaling networks with a formal approach based on answer set programming. *CMSB - 10th Computational Methods in Systems Biology*, 342–361.

Vilenchik, M. M., & Knudson, A. G. (2003). Endogenous DNA double-strand breaks: Production, fidelity of repair, and induction of cancer. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(22), 12871–12876. https://doi.org/10.1073/pnas.2135498100

Vogelstein, B., Lane, D., & Levine, A. J. (2000). Surfing the p53 network. *Nature*, *408*(6810), 307–310. https://doi.org/10.1038/35042675

von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., & Snel, B. (2003). STRING: A database of predicted functional associations between proteins. *Nucleic Acids Research*, *31*(1), 258–261. https://doi.org/10.1093/nar/gkg034

Vousden, K. H., & Prives, C. (2009). Blinded by the light: The growing complexity of p53. *Cell*, *137*(3), 413–431. https://doi.org/10.1016/j.cell.2009.04.037

Waddington, C. H., & Kacser, H. (1957). *The strategy of the genes. A discussion of some aspects of theoretical biology* (Issue 3). Allen & Unwin.

Wang, R. S., Saadatpour, A., & Albert, R. (2012). Boolean modeling in systems biology: An overview of methodology and applications. *Physical Biology*, *9*(5). https://doi.org/10.1088/1478-3975/9/5/055001

Watari, N., & Larson, R. G. (2010). The hydrodynamics of a run-and-tumble bacterium propelled by polymorphic helical flagella. *Biophysical Journal*, *98*(1), 12–17. https://doi.org/10.1016/j.bpj.2009.09.044

Wilkinson, D. J. (2018). *Stochastic modeling for systems biology* (3rd ed.). New York: Chapman and Hall/CRC.

Wittmann, D. M., Krumsiek, J., Saez-Rodriguez, J., Lauffenburger, D. A., Klamt, S., & Theis, F. J. (2009). Transforming Boolean models to continuous models: Methodology and application to T-cell receptor signaling. *BMC Systems Biology*, *3*. https://doi.org/10.1186/1752-0509-3-98

Wynn, M. L., Consul, N., Merajver, S. D., & Schnell, S. (2012). Logic-based models in systems biology: A predictive and parameter-free network analysis method. *Integrative Biology*, *4*(11), 1323–1337. https://doi.org/10.1039/c2ib20193c

Zernicka-Goetz, M., Morris, S. A., & Bruce, A. W. (2009). Making a firm decision: Multifaceted regulation of cell fate in the early mouse embryo. In *Nature Reviews Genetics* (Vol. 10, Issue 7, pp. 467–477). https://doi.org/10.1038/nrg2564

Zhang, J., Nie, Q., & Zhou, T. (2019). Revealing dynamic mechanisms of cell fate decisions from single-cell transcriptomic data. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.01280

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., & Enard, W. (2017). Comparative analysis of single-cell RNA sequencing methods. *Molecular Cell*, *65*(4), 631–643. https://doi.org/10.1016/j.molcel.2017.01.023