Computational prediction of cell fates by systems modelling of cellular signalling pathways

Zhang, Fan

2016

Zhang, F. (2016). Computational prediction of cell fates by systems modelling of cellular signalling pathways. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/69208>

<https://doi.org/10.32657/10356/69208>

COMPUTATIONAL PREDICTION OF CELL FATES BY
SYSTEMS MODELLING OF CELLULAR
SIGNALLING PATHWAYS

ZHANG FAN

School of Computer Science and Engineering

2016

COMPUTATIONAL PREDICTION OF CELL FATES BY
SYSTEMS MODELLING OF CELLULAR
SIGNALLING PATHWAYS

ZHANG FAN

School of Computer Science and Engineering

A thesis submitted to the Nanyang Technological University
in fulfilment of the requirement for the degree of
Doctor of Philosophy

2016

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere appreciation and gratitude to Asst. Prof. Zheng Jie, for being such a great supervisor and mentor. He has given a most valuable technical guidance on my research, and led me through the unforgettable four years of PhD study. Without the encouragement and support from Asst. Prof. Zheng Jie, my research could not have been here.

I would like to show my gratitude to Dr. Wu Min, Dr. Liu Hui and Ms. Li Xuejuan for their advice and thoughtful assistance on my research project and on this report. I would also like to express my sincere gratitude to Assoc. Prof. Kwok Chee Keong and Asst. Prof. Francesc Xavier Roca Castella for sharing their knowledge and experiences. Many thanks Dr. Chen Haifen, Ms. Guo Jing, Dr. Zhang Xiaomeng and Mr. Shital Kumar Mishra for being amazing colleagues and friends.

Moreover, I am very grateful to the School of Computer Science and Engineering, Nanyang Technological University, Singapore, for providing me the opportunity to conduct my PhD research.

Finally, I want to thank my dear parents and wife for their unreserved support and love to me. I fail words to express my gratitude to them for all they have done for me.

AUTHOR'S PUBLICATIONS

JOURNAL PAPERS

- Xuejuan Li, Shital K Mishra, Min Wu, **Fan Zhang** and Jie Zheng. Syn-Lethality: An Integrative Knowledge Base of Synthetic Lethality towards Discovery of Selective Anticancer Therapies. *BioMed research international Journal*. 2014.
- Min Wu, Xuejuan Li, **Fan Zhang**, Xiaoli Li, Chee-Keong Kwoh and Jie Zheng. In Silico Prediction of Synthetic Lethality by Meta-Analysis of Genetic Interactions, Functions, and Pathways in Yeast and Human Cancer. *Cancer informatics*. 13 (Suppl 3), 71-80, 2014.
- **Fan Zhang**, Min Wu, Xuejuan Li, Xiaoli Li, Chee-Keong Kwoh and Jie Zheng. Predicting essential genes and synthetic lethality via influence propagation in signaling pathways of cancer cell fates. *Journal of Bioinformatics and Computational Biology*. Vol. 13, No. 03, 1541002, 2015.
- Shital K Mishra, Sourav S Bhowmick, Huey Eng Chua, **Fan Zhang** and Jie Zheng. Computational cell fate modelling for discovery of rewiring in apoptotic network for enhanced cancer drug sensitivity. *BMC Systems Biology*. 9 (Suppl 1), S4, 2015.
- **Fan Zhang**, Haoting Chen, Li Na Zhao, Hui Liu, Teresa M. Przytycka and Jie Zheng. Generalized logical model based on network topology to capture the dynamical trends of cellular signaling pathways. Accepted

by *BMC Systems Biology*.

- Hui Liu, **Fan Zhang**, Shital Kumar Mishra, Shuigene Zhou and Jie Zheng. Knowledge-guided fuzzy logic modeling to infer cellular signaling networks from proteomics data. Submitted to *Scientific Reports*.
- **Fan Zhang**, Chee Keong Kwoh, Min Wu and Jie Zheng. Nonlinear Modelling of cancer cell fates driven by signalling data to reveal drug effects. Accepted by *PLOS ONE*.

CONFERENCE PAPERS

- Min Wu, Xuejuan Li, **Fan Zhang**, Xiaoli Li, Chee-Keong Kwoh and Jie Zheng. Meta-analysis of Genomic and Proteomic Features to Predict Synthetic Lethality of Yeast and Human Cancer. *Proceedings of the 4th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. 384-391, 2013.
- **Fan Zhang**, Min Wu, Chee-Keong Kwoh and Jie Zheng. Data-driven prediction of cancer cell fates with a nonlinear model of signaling pathways. *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. 436-444, 2014.
- **Fan Zhang**, Runsheng Liu and Jie Zheng. Sig2GRN: A software tool linking signaling pathway with gene regulatory network for dynamic simulation. Accepted by *International Conference on Genome Informatics 2016*.

Contents

1	INTRODUCTION	1
1.1	BACKGROUND	1
1.1.1	COMPONENTS	2
1.1.2	INTERACTIONS	3
1.1.3	BIOLOGICAL NETWORKS	3
1.2	MOTIVATION AND OBJECTIVES	5
1.3	CONTRIBUTIONS	6
1.4	OUTLINE	7
2	RELATED WORKS	9
2.1	KNOWLEDGE-BASED MODELING OF SIGNALING PATHWAYS	9
2.1.1	LOGICAL MODELS	10
2.1.2	KINETIC MODELS BASED ON DIFFERENTIAL EQUATIONS	15
2.2	DATA-DRIVEN MODELING OF SIGNALING PATHWAYS	18
2.3	HYBRID MODELING OF SIGNALING PATHWAYS	22
2.4	CHAPTER SUMMARY	24
3	GENERALIZED LOGICAL MODEL OF CELLULAR SIGNALING PATHWAYS	26
3.1	INTRODUCTION	26
3.2	METHODS	28

3.2.1	COMPUTATIONAL MODEL FOR DYNAMICAL SIMULATION	28
3.2.2	NETWORK STRUCTURE	31
3.3	RESULTS	31
3.3.1	PERFORMANCE COMPARISON ON SIMULATING SIGNALING RESPONSES TO PERTURBATIONS . .	31
3.3.2	MODEL COMPARISON AND VALIDATION WITH REAL DATA	38
3.4	CHAPTER SUMMARY	43
4	NONLINEAR POWER-LAW MODELING OF CELLULAR SIGNALING PATHWAYS	45
4.1	INTRODUCTION	45
4.2	DATA AND METHODS	47
4.2.1	DATA	47
4.2.2	NONLINEAR POWER-LAW MODEL	49
4.2.3	CELLDESIGNER SIMULATION	53
4.2.4	EVALUATION METRICS OF MODEL PERFORMANCE	55
4.2.5	IDENTIFICATION OF DRUG EFFECTS	57
4.3	RESULTS	58
4.3.1	ANALYSIS OF SIMULATION DATA	58
4.3.2	RESULTS ON BREAST CANCER CELL LINES . . .	60
4.3.3	RESULTS ON HT-29 CELL LINE	69
4.3.4	RESULTS ON DREAM8 DATASET	72
4.4	CHAPTER SUMMARY	73
5	PREDICTING SYNTHETIC LETHALITY VIA MODELING OF CELLULAR SIGNALING PATHWAYS	75
5.1	INTRODUCTION	75
5.2	DATA AND METHODS	77
5.2.1	DATA	77
5.2.2	SIGNALING PATHWAYS	77
5.2.3	DATA-DRIVEN POWER-LAW MODEL	79

5.2.4	GENE KNOCK-DOWN SIMULATION	81
5.2.5	SYNTHETIC LETHALITY PREDICTION	83
5.3	RESULTS	84
5.3.1	SINGLE KNOCK-DOWN AND ESSENTIAL GENES PREDICTION	84
5.3.2	DOUBLE KNOCK-DOWN AND SYNTHETIC LETHAL- ITY PREDICTION	85
5.4	CHAPTER SUMMARY	87
6	SUMMARY AND FUTURE WORK	88
6.1	SUMMARY	88
6.2	FUTURE WORK	90
6.2.1	INTEGRATION OF SIGNALING PATHWAYS AND TRANSCRIPTIONAL REGULATORY NETWORKS	90
6.2.2	HYBRID MODELING OF CELLULAR SIGNALING PATHWAYS	91
6.2.3	MAPPING TIME SCALES OF COMPUTATIONAL SIMULATIONS TO REAL DATA	92

List of Figures

1.1	An example of signaling pathways.	4
2.1	An example of Boolean network.	11
2.2	An example of Fuzzy logic.	15
2.3	(a) k -means clustering. (b) Hierarchical clustering. (c) Self-organizing maps. (d) Principal Component Analysis.	20
2.4	Discrimination of cell types using principal components.	21
2.5	An example of hybrid modeling.	23
3.1	The workflow of simulation using the proposed model.	30
3.2	Signaling pathways constructed based on the dataset.	32
3.3	Comparison of simulation-based predictions using the proposed simulator (a–c), SimBoolNet (d–f) and GINsim (g–i).	34
3.4	(a) The influence of different initial values on simulations. (b) The influence of different edge weights to simulations.	37
3.5	The plots of the simulated and real data in the control group. The blue and green curves are the simulated data using the proposed model and SimBoolNet, respectively. The red dots are the real data.	41
4.1	(a) A signaling network for cell death regulation. (b) The simulation results of DNA Damage and Cell Death.	51

4.2	Cross-validation of the linear and the nonlinear models on synthetic data. For example, 999 out of the total 1000 predictions made by the nonlinear model have both the Spearman and the Pearson correlations with the testing data higher than 0.7. . .	60
4.3	Performance comparison of cell death prediction within each cell line. More outputs have high coefficients in the nonlinear model (a) than in the linear model (b). (c) Comparison between the two models by counting how many outputs have both coefficients higher than a threshold (e.g., 0.7, 0.8 and 0.9). . .	62
4.4	Comparison of cell death prediction by the linear and the nonlinear models when combining all three cell lines together. . .	64
4.5	Pathway information extracted from GeneGO MetaCore. The dash line denotes the predicted drug effects, <i>i.e.</i> , signal transduction is blocked here.	67
4.6	Performance of the linear and the nonlinear models at different time-staggered degrees.	69
4.7	Principal component analysis for cell line discrimination. (a) The nonlinear model. (b) The linear model.	70
4.8	Schematic representation of the signaling network induced by TNF, EGF, and insulin.	72
5.1	Pathway information extracted from GeneGO MetaCore. . . .	78

List of Tables

2.1	Comparison of various methods for modeling signaling pathways.	25
3.1	Comparison of the computational time required for SimBool-Net, GINsim and the proposed model.	35
3.2	Spearman and Pearson correlations between simulated and real data.	40
3.3	Inputs of the simulation in Section 3.3.2. The columns are the input nodes of simulation and the rows are various conditions indicating the treatments of biological experiments with drugs corresponding to wet-lab experiments. The symbols ‘+’ and ‘–’ represent the perturbation types, <i>i.e.</i> , activation and inhibition.	42
3.4	Goodness of fit of the simulations to the real experimental measurements under drug treatments.	43
4.1	Effects of parameters on cell death predictions.	54
4.2	Curve fitting results of Power function, Linear function, Gaussian function and Exponential function. The statistics are about EGFR and DNA Damage.	59
4.3	The mean and standard deviation of the correlation between model predictions and the real data.	61
4.4	Performance of the linear and the nonlinear models in predicting cell fate across cell lines.	65
4.5	Experimental design for identification of time-staggered input-output relationships.	68

4.6	Performance of the nonlinear and linear model on DREAM8 dataset. The values are the RMSE between the model predictions and testing data of activity of signaling proteins.	73
5.1	Rank of the single knock-down simulation.	84
5.2	Rank of the SL candidates.	86

ABSTRACT

In the last few decades, extensive computational studies have been conducted on signaling pathways with both logic modeling (knowledge-based) and data-driven modeling. The increasing availability of high-throughput molecular data and multiplex techniques for uncovering cellular systems have made predictive *in silico* modeling more accurate which is important for understanding and engineering cellular functions. Generally, signal transduction inside a cell involves various modifications, such as protein phosphorylation. Through a cascade of biochemical reactions, the signals are transmitted downward to the nucleus or other cellular organelles to regulate physiological functions and finally control the cellular phenotypes (*e.g.*, apoptosis, proliferation and cell cycle). It is widely believed that the dysregulation of signal transduction is one of the most important pathogeneses of many human diseases including cancer, which makes the study of signaling pathways crucial for discovery of new anticancer therapies. However, there remain several challenges for the systems modeling of signaling pathways, such as how to integrate prior knowledge with real data into a context-specific model. This thesis presents the studies I have carried out with my collaborators to address some of the challenges.

We first propose a generalized logical model that is able to simulate the graded responses to degradations and the effects of scheduled perturbations to the cells. A network is constructed for computational simulation based on the knowledge extracted from databases and literatures. In the network, the nodes and the edges represent the signaling proteins and the phosphorylation

interactions, respectively. Given the inputs, the activity level of each node is updated synchronously based on its own previous state and the incoming signals from the parent nodes. Different combinations of perturbations are applied to the network. The evaluations of the simulation results with real data demonstrate that our simulator has the ability to capture the main dynamical trends of signal transduction. Moreover, compared with existing simulators, our model achieves better performance in predicting the state transitions of signaling networks.

Secondly, to study how the cellular signals control the cell fates (*e.g.*, cancer cell death), a nonlinear power-law model is proposed to relate the signaling proteins to the cell death. In our nonlinear function, the independent and dependent variables denote the activities of signaling proteins and the probability of cell death, respectively. The parameters, which indicate the contributions of the signaling proteins to the cell death, are identified from the training dataset and subsequently used to predict the probability of cell death on the testing dataset. Compared with the linear model on three cancer datasets with phosphoproteomics and cell fate measurements, the proposed nonlinear model has superior performance in the cell death prediction. Moreover, our model can capture cell line specific information to distinguish one cell line from another in cell fates prediction. By *in silico* experiments of virtual protein knock-down, the proposed model is able to reveal the drug effects which can complement traditional approaches such as binding affinity analysis.

After linking the cellular signals with the cell fates, we further investigate how to increase the selectivity of the drugs to the cancer. It is believed that if two genes have Synthetic Lethality (SL) relationship with one of them being a cancer-specific mutated gene, the drug that targets the partner gene would be able to give rise to SL and kill the tumor cells selectively. Therefore, a hybrid model, which combines a data-driven method with prior knowledge of the underlying mechanisms (*e.g.*, signaling networks), is proposed to study the potential drug targets by predicting the change of the probability of cell death caused by the gene knock-down. The single gene knock-down and the

double genes knock-down are simulated to identify the human essential genes and the potential SL pairs of genes, respectively. A pair of genes is considered as an SL candidate if the double genes knock-down highly increases the predicted probability of cell death while the knock-down of either single gene does not. The promising performance of the hybrid model opens new directions for balancing the experimental data and the prior knowledge in future.

Chapter 1

INTRODUCTION

1.1 BACKGROUND

The growth of high-throughput technologies [1, 2, 3, 4] makes it possible to interrogate biological processes on a large scale, thus allows us to focus on biological systems beyond a small number of components or biochemical interactions. Such innovation will enable a systematic understanding of a biological system, which is crucial for studying diseases such as cancer.

Systems biology, instead of viewing a biological system in terms of its components such as a specific gene or protein, integrates all the components (*e.g.*, genes, mRNAs, proteins and metabolites) as a system to identify the mechanisms of how they influence each other and how they work together to respond to internal and external perturbations [5, 6].

Systems biology first identifies the list of components and the interactions between them, then describes their properties and functions using mathematical models and finally combines the models with real experimental data to reconstruct the model and predict outcomes after perturbations to the system with the help of a computer.

1.1.1 COMPONENTS

Generally, genes of living organisms are responsible for storing, utilizing and passing their hereditary information. A gene is encoded as a region of genomic sequences such as DNA (deoxyribonucleic acid) sequences or, for many types of viruses, RNA (ribonucleic acid) sequences. It is composed of regulatory regions which control if the mRNA (messenger RNA) should be produced, transcribed regions which encode the sequences of mRNA, and other functional sequence regions. The genes can be considered as units of inheritance [7]. And together with the non-coding sequences of the DNA/RNA, they formed the genome of an organism which contains a major part of hereditary information. The human genome contains approximate 25,000 genes [8]. For eukaryotic (such as human) genes, genetic information is transmitted from DNA to precursor mRNA (pre-mRNA) during transcription. Then post-transcriptional modifications take place, such as the alternative splicing process in which regions called introns are removed and other regions named exons are joined [9, 10, 11]. Thus, the pre-mRNA becomes mature mRNA which specifies the sequence of amino acids to produce proteins. About 95% of multiexonic genes in the human genome are subject to this post-transcriptional modification which is a way to control the number and categories of mRNA [12].

Proteins are first synthesized to polypeptides, which are sequences of amino acids, according to their mRNAs, then the polypeptides are folded into three dimensional shapes to gain biological functions [6]. There are 20 types of standard amino acids and the length of a polypeptide may range from tens to even thousands of amino acids which gives rise to the great abundance of proteins. Moreover, proteins are involved in post-translational modifications, such as phosphorylation, methylation and acetylation, which could further increase their orders of magnitude.

Metabolites, usually restricted to small molecules within organisms, are the intermediates and products of metabolism [6]. They have various functions, including biological structure construction, signal and energy trans-

duction, enzyme activity maintenance and so on.

1.1.2 INTERACTIONS

To determine if there will be an interaction between a set of components, information about biological reactions is needed [6]. A biochemical interaction such as enzymatic reactions is identified where substrates are catalyzed and converted into products. In kinetic modeling, a biochemical system is described using a set of equations. Each equation represents a biochemical reaction where the reactants are consumed and the products are generated [13, 14]. And gene regulatory interactions are identified if the translational product of one gene can control the expression level of another gene. A collection of regulatory genes that interact with each other and with other substances to control the gene expressions constitute the gene regulatory network [15, 16].

1.1.3 BIOLOGICAL NETWORKS

After identifying the components and interactions, they are integrated to construct a biological network. The basic connectivity description lists the associations between the components. More detailed causal methods can be used to describe the cause-effect relationships. As most of the reactions among components are biochemical interactions, the system can also be represented by an even more mechanistic model such as a stoichiometric matrix [6], where each row corresponds to a species, each column corresponds to a reaction, and the elements in the matrix appear as negative and positive values for reactants and products, respectively.

Based on the properties of the components and the relationships among them, it is a straightforward way to link them together as a network where the nodes and the edges denote the components and the interactions, respectively. A fundamental property of biological networks is that the biological networks can change with time or adapt to new environment which, in other words,

Figure 1.1: An example of signaling pathways.

is the ability to evolve [6]. Wiring diagram is introduced from electrical engineering to describe the relative position and arrangement of components and their relationships.

For a human body that comprises approximately 10^{14} cells that can be classified into more than 200 different cell types, only a subset of all the 25,000 genes are expressed in a specific cell type or under a particular condition. It means that tight and concerted regulations are needed to decide which genes are expressed and which are not so that the cell can respond appropriately to the environmental perturbations and developmental signals. This procedure relies on the transcriptional and translational regulatory networks. In regulatory networks, the localization and binding specificity of biological macromolecules play important roles.

The fundamental role of signaling networks is to transmit a variety of environmental signals (*i.e.*, the binding of a ligand to a cell surface receptor) into the cell. The subsequent signal transduction inside the cell involves

various modifications (*e.g.*, phosphorylation) on signaling proteins. The signals will be transmitted down to the nucleus or other cellular organelles to change biological functions (*i.e.*, control the expression level of a gene) or even influence the cell fates (*i.e.*, proliferation, differentiation, migration and apoptosis). Figure 1.1 [17] shows an example of signaling pathways which involves the regulation of cellular processes such as apoptosis and survival. The nodes (the dark blue ones have experimental measurements) represent signaling proteins and cell fates. Activation and inhibition interactions are denoted as green edges with an arrow head and red edges with a flat-head, respectively. For instance, as shown in Figure 1.1, P53 can be activated by the signals transmitted from ATM, CHK, P38 and JNK; on the contrary, the signals from AKT to P53 will inhibit the activation of P53. Although the transmission of signals relies on molecules that are constrained by the laws of conservation of mass and energy, the information itself is not because it can be amplified or decayed during transmission. Another key feature of signaling networks is that signals are processed fast. While regulatory processes or alterations of cell fate phenotypes can take several minutes to hours, signal transductions may finish in seconds [18].

1.2 MOTIVATION AND OBJECTIVES

It is known that the cell responds to the environment according to the signals transmitted by signaling pathways. The dysregulation of signal transduction is one of the most important pathogeneses of many human diseases including cancer. Therefore, many researches have been conducted to model the signaling pathways and understand the mechanisms of how the signal transduction regulates the cell phenotypes. However, challenges exist for computational methods, for example, the existing models are either purely knowledge-driven or data-driven thus hybrid modeling that combine the strengths of both are needed; moreover, there are insufficient context-dependent models which expects models that integrate the prior knowledge with the experimental data to identify context-specific information (*e.g.*, signaling network rewiring un-

der a certain condition). To facilitate researches in systems biology, the studies presented in this thesis include the development of computational methods that can reveal the process of degradation and the effects of scheduled perturbations to signaling networks and combine the prior knowledge with the experimental data to predict the cell fates. Hence, the aims of this thesis are (i) to dynamically simulate the trends (changes of the activities of the signaling proteins over time) of cellular signaling pathways; (ii) to predict the cell fates based on the cell signals; (iii) to predict the essential genes and synthetic lethality pairs of genes based on the different contributions of genes to the changes of cell fates.

1.3 CONTRIBUTIONS

The contributions of this thesis can be summarized as follows:

1. A generalized logical model based on network topology which is able to capture the dynamical trends of cellular signaling pathways. The proposed computational model is able to simulate the graded responses to degradations, the sigmoidal biological relationships between signaling molecules and the effects of scheduled perturbations to the cells. The simulation results are consistent with experimental data of protein phosphorylation, demonstrating that the proposed model can capture the main trends of protein activities during the process of signal transduction. Compared with existing simulators, our model has better performance in predicting the state transitions of biological systems. The proposed simulation tool provides a valuable resource for modeling signaling pathways using a knowledge-based method.
2. A nonlinear power-law model to relate the activities of signaling proteins to the probabilities of cell fate. In our experiments, we compare our nonlinear model with the linear model on three cancer datasets with phosphoproteins and cell responses measurements, which demonstrates that the presented nonlinear model has superior performance

on cell fate prediction. By *in silico* experiment of protein knock-down, the proposed model is able to reveal the drug effects which can facilitate traditional approaches such as binding affinity analysis. Moreover, our model is able to capture the cell line specific information thereby distinguishing one cell line from another in cell fate prediction. Our results show that the nonlinear data-driven model can perform better in cell fate prediction and reveal more insights of the biological system in cancer cells than the linear model.

3. A hybridized model to predict essential genes and Synthetic Lethality (SL) via influence propagation in signaling pathways of cancer cell fates. We combine a data-driven model with knowledge of signaling pathways to simulate the influence of single gene knock-down and double genes knock-down to cell death. A pair of genes is considered as an SL candidate when double knock-down increases the probability of cell death significantly, but single knock-down does not. The single gene knock-down is confirmed according to the human essential genes database. Our validation using literature shows that the predicted SL candidates agree well with wet-lab experiments. A few novel reliable SL candidates are also predicted by our model.

1.4 OUTLINE

The thesis is divided into six chapters including this first introductory chapter. From Chapter 3 to Chapter 5, each chapter presents a complete and independent work, while the contents are related with each other. Chapter 2 presents an extensive literature review on the state-of-the-art computational studies of signaling pathways with a focus on those using quantitative and logical modeling. Chapter 3 presents a knowledge-based model which can simulate the dynamics of signaling pathways. In Chapter 4, a data-driven model is proposed to describe the probability of cell fates as a nonlinear function of cell signals, and the model not only shows better performance

than the linear model on cell fates prediction but also reveals promising performance on drug targets and drug effects identification. In Chapter 5, a hybrid model, which combines the data-driven methods with knowledge of the signaling pathways, is proposed to predict essential genes and Synthetic Lethality pairs of genes in cancer cells. Finally, Chapter 6 summarizes the research presented in this thesis and provides recommendations for future work.

Chapter 2

RELATED WORKS

High-throughput technologies are making the biological data explode both in size and complexity. To acquire the information that can help us to understand the biological systems, computational methods are introduced to process the large-scale data. Computational models can be constructed based on the prior knowledge such as the molecular mechanisms and the network topology of the pathways. On the other hand, modeling approaches can focus on data. Moreover, hybrid methods which combine knowledge and data become more and more important in modeling systems biology.

This thesis mainly focuses on computational modeling of signaling pathways. In this section, we review previous studies that are closely related to my works.

2.1 KNOWLEDGE-BASED MODELING OF SIGNALING PATHWAYS

With the increasing availability of the experimental techniques, more and more details of the underlying mechanisms of the signaling pathways are uncovered, including the components of the pathways, the interactions between

the molecules, the parameters of the interactions and so on. Given the above knowledge, a straightforward way is to integrate them as a network where the nodes and the edges denote the components and the interactions, respectively. For example, given the information from the studies of PI3K [19], MAPK [20], JAK and STAT signalings [21], a model of EGFR signaling in tumour progression was constructed [22]. The conducted studies facilitate the description of the cascades of signal transduction. Subsequently, to make the model predictive, the network should be converted into a computable model by which the state of the signaling pathways can be calculated given a set of input conditions. One way to achieve this is to depict the signaling pathways as a logical model in which the relationships between the inputs and the outputs are specified by logical gates. When the numbers of the components and the interactions in the network are large, the logical model is suitable. On the other hand, if the detailed mechanisms (such as kinetic parameters) are available, an alternative approach which encodes the network as a set of differential equations shows more promising performance.

2.1.1 LOGICAL MODELS

Logic-based models are suitable formalisms for modeling relative large networks in which the detailed kinetic parameters are not fully available. The available knowledge for logical modeling should include the dependencies between components such as a signaling proteins A has a positive effect on the activation of C while B is able to block the activation of C . In logical models, the logic gates are employed to specify the dependencies, *i.e.*, to define the output state for each possible combination of input variables.

Boolean network [23], which is a simple logical model, has binary state (*i.e.*, either ON or OFF) for each variable and a Boolean function, associated with a truth table, assigned to the variable to determine its output for given inputs. The Boolean functions are determined by the connectivity (topology) on the variables, and the Boolean variables are represented as nodes in a network. The dynamics of the Boolean network is to update the state of

Figure 2.1: An example of Boolean network.

the network at time $t+1$ by evaluating all the boolean functions at time t . Figure 2.1(a) shows an example of Boolean network where node A is activated by itself and inhibited by node C, node B is activated by node A and node C, and node C is activated by node B only. The logical gates for node A and B are OR and AND, respectively (Figure 2.1(b)). Figure 2.1(c) shows the truth tables of the three nodes and Figure 2.1(d) describes the state transition graph of the network.

Boolean network has been successfully applied to the modeling of gene regulatory networks and signaling networks. Based on the pioneering studies of Kauffman [24] and Thomas [25], Mendoza *et al.* studied the dynamical properties of the Boolean network [26]. When the feedback loops are included, Boolean network predicted the same stable states as the differential equation models, demonstrating that Boolean network was able to reveal the dynamical properties of the regulatory network especially when there was a scarcity of detailed kinetic parameters. Albert and Othmer employed Boolean network to predict the gene expression pattern of *Drosophila*

melanogaster [27]. Their study suggested that the stable states of a system were determined mainly by the topology of the network and the interaction types (*i.e.*, activation or inhibition) between the components.

The time-dependent Boolean models have two updating schemes: synchronous and asynchronous. The synchronous updating scheme assumes that the time-scales of the interactions in the network are similar whereas the asynchronous pattern scheme assumes that the time-scales between different types of interactions (*e.g.*, signal transduction, translation and transcription) can vary widely. Thomas [25] first introduced asynchronism to study the stable states of a system, and Chaves *et al.* [28] further explored the effects of asynchronous updating scheme on the dynamics of a system.

For signaling networks, its input-output relationships, *i.e.*, the behaviors of signaling networks responde to the external perturbations, are also of interest in research besides its stable states [29]. Saez-Rodriguez *et al.* [29] constructed a Boolean model of T Cell Receptor (TCR) signaling pathways to study the global behaviors of the network under different knock-out perturbations. Their results revealed some key features of the network (*e.g.*, feedback loops and global dependencies) and predicted candidates for missing edges in the network. Calzone *et al.* [30] and Grieco *et al.* [31] analyzed the input-output relationships in signaling networks also using Boolean models. In their networks, not only the signaling proteins, but also the cell fates (*e.g.*, Apoptosis and Proliferation) were included as the nodes. They computationally simulated the dynamics of signaling networks to analyze how the cell fate decisions were made by the cellular signals.

Although the above works provided some nice frameworks for Boolean modeling of signaling pathways, the inference of Boolean models from experimental data has not been discussed. The logic gates were mainly extracted from the literature and described the generic relationships between the upstream and the downstream molecules, and thus cannot reveal the cell-type specific or treatment specific alterations in the signaling networks. To overcome this limitation, Saez-Rodriguez *et al.* [32] proposed a pipeline to reconstruct the Boolean model of signaling pathways against phosphoproteomics

data (*e.g.*, the time-series phosphorylation data of signaling proteins under various perturbations). Given the topology (*e.g.*, the connectivity between molecules and the interaction types) of the network from prior knowledge, a supergraph was first constructed by considering all the possible Boolean functions of each node. Then the superstructure of the Boolean network was optimized against the experimental data, using a heuristic genetic algorithm, by minimizing the discrepancy between model predictions and the measurements. The optimization process stopped when the criteria were fulfilled, such as tolerance from a perfect fit or achieving the pre-defined maximum number of generations. Compared with the generic network, the optimized network showed better performance on revealing the cue-signal-response relationships in signal transduction networks. Following this step, Mitsos *et al.* [33] proposed an approach to identifying the drug-induced rewirings in the signaling pathways. In this work, an Integer Linear Programming (ILP) optimization algorithm was employed for the reconstruction of the Boolean model of a generic signaling network. The objectives of the optimization process included the minimization of the error between the model predictions and the measurements, and the minimization of the size (*i.e.*, the number of the possible edges) of the network. After fitting the real data to the Boolean model, the identified network alterations successfully revealed the effects of the perturbations (*i.e.*, drugs). However, the aforementioned two methods have a shortcoming: they rely on the prior knowledge of the network topology (both connectivity and the interaction type), thereby they are able to identify only the interactions that were blocked by the drugs (*i.e.*, removing edges in the network) but are unable to identify missing interactions in the generic network (*i.e.*, adding new edges in the network) or the alterations of interaction types (*i.e.*, activation becomes inhibition and *vice versa*). Therefore, Sharan and Karp [34] provided a more generic approach to reconstructing the Boolean models of signaling pathways. In this work, the reconstruction process was also formulated as an ILP problem. The optimization process initially considered all the possible Boolean functions of each node regardless of the network topology and the interaction types, thus the inferred network was not limited by the prior knowledge. Although their method was generic,

they suggested that the model performance was better if the prior knowledge was available.

In spite of the aforementioned successful applications of Boolean network, its inability of encoding graded responses becomes a significant limitation since it is able to handle only binary values which is oversimplified compared with a real signaling network. Therefore, Zheng *et al.* [35, 36] proposed a generalized Boolean network model that was able to capture the general trends of the signaling pathways under perturbations. In their model, the state of each node varies continuously from 0 to 1 instead of two binary values (0 and 1). The state of a node at time point $t+1$ was updated based on its own previous state at time point t and the amounts of signals transmitted from its up-stream nodes at time point t . Their model was successfully applied to modeling a pro-survival signaling pathways and predicting the activities of the signaling proteins. However, it was limited on reflecting the graded responses to degradations and the effects of scheduled perturbations to the cells.

Another extension of Boolean models is fuzzy logic. In fuzzy logic models, fuzzy values (range between 0 and 1) are employed to describe the states of the variables. Each variable is associated with membership functions which are responsible for fuzzification and defuzzification of the fuzzy values. In Figure 2.2(a), the input value (from 0 to 1) will be fuzzified into degree of membership (DOM), which include states of Low (L), Medium (M) and High (H). And the fuzzy “IF-THEN” rules (Figure 2.2(c)) are employed for processing a given input to achieve a desired output. Aldridge *et al.* [37] applied a fuzzy logic framework to analysis the crosstalks among TNF-EGF-Insulin signaling pathways. The fuzzy logic gates were assembled manually and the parameters were obtained using nonlinear least squares regression method. They compared the model simulations of both fuzzy logic model and discrete model with the real data, and the results showed that the fuzzy logic model achieved a superior fit of data. Huang and Hahn [38] also studied fuzzy logic based modeling of signal transduction networks. Similar with [37], Huang and Hahn [38] manually constructed the fuzzy rules and used K-mean

Figure 2.2: An example of Fuzzy logic.

clustering method to identify the parameters of the membership functions. Their simulation results successfully predicted the dynamical behaviors of IL-6 signaling pathways. Then following the ideas in [32, 33, 34], Morris *et al.* [39] extended the fuzzy logic framework in [37] such that the fuzzy logic gates can be learned from experimental data. The model training process employed a genetic algorithm to minimize the mean squared error (MSE) with respect to the data. Due to the insufficiency of the data, a family of reconstructed fuzzy logic models were generated with equal goodness of fit to the experimental data but slightly different network topologies.

2.1.2 KINETIC MODELS BASED ON DIFFERENTIAL EQUATIONS

If the underlying biochemical mechanisms are known, biochemical kinetic modeling is a well-established strategy for describing the biology system using a set of mathematical equations. Each equation represents a biochemical reaction where the reactants are consumed and the products are generated. The parameters of the kinetic models include but are not limited to concentration, binding affinity and reaction rate, which represents the rate of production or consumption of a biomolecular species [13, 14]. The generic protocols followed by the construction of the kinetic models are the laws of conservation of mass and energy [40, 12].

Two commonly used approaches of biochemical kinetic modeling are ordinary and partial differential equations (ODEs and PDEs). ODEs are employed to describe the temporal evolution of each biomolecular species, *i.e.*, how its concentration changes over time. As shown in Equation (2.1), the change of the concentration of component C over time depends on the summation of the reaction rates producing and consuming C.

$$\frac{d[C]}{dt} = \sum Production - \sum Consumption \quad (2.1)$$

For example, given the interaction scheme in Equation (2.2),

the following ODEs (Equation (2.3)) of the concentration change over time of A, B and C can be derived according to the definition of mass action law.

$$\begin{aligned} \frac{d[A]}{dt} &= k_-[C] - k_+[A][B] \\ \frac{d[B]}{dt} &= k_-[C] - k_+[A][B] \\ \frac{d[C]}{dt} &= k_+[A][B] - k_-[C] \end{aligned} \quad (2.3)$$

PDEs extends ODEs by including spatial features, *i.e.*, modeling changes in species concentrations with respect to space. Typically there are three main steps for building an ODE (or PDE) model: (i) model construction which typically involves the identification of the molecular species and the interactions, and the translation of the prior knowledge to a set of mathematical equations; (ii) model calibration which is the process of parameter estimation; and (iii) model validation where the model predictions are eval-

uated against the wet-lab experimental data [13, 41, 14].

Michaelis and Menten [42], who are the pioneers of the kinetics modeling, studied the enzymatic reactions *in vitro* under controlled, well-mixed conditions. After nearly a century, “systems biologists” began to focus on the biological networks. Huang and Ferrell [43] studied the stimulus/response relationships in the mitogen-activated protein kinase (MAPK) cascade. The stimulus/response curves of the enzymes in the pathways followed an ultrasensitive pattern which was steeper than that of the hyperbolic Michaelis-Menten curve. Schoeberl *et al.* [44] constructed the epidermal growth factor receptor (EGFR) signaling pathways and explored the EGFR-induced MAPK cascade dynamics. By validating the model predictions against the experimental data, the results showed that the responses of the model were stable over a 100-fold variance of the stimuli concentration, and the initial velocity of the receptor activation was identified as the most important parameter in controlling the efficacy of the signal transduction. Novk and Tyson [45] simulated the physiological responses of mammalian cells to the control of cell cycle using a set of nonlinear differential equations. The dynamics of the pathways controlling cell apoptosis was then studied by Albeck *et al.* [46] and Neumann *et al.* [47]. These works described how the cell fates (*e.g.*, cell cycle and apoptosis) were regulated by the signaling pathways.

In spite of promising performance of the biochemical kinetic modeling of signaling pathways, limitations have also been noticed. First, although wet-lab experiments have uncovered more and more information of the signaling pathways, the detailed prior knowledge about the molecular mechanisms underlying the biochemical reactions and kinetic parameters, required by the kinetic models, are unfortunately not always available. Second, describing the multiple states of the biomolecular species is beyond the capability of the kinetic models [48]. For example, proteins with different covalent modifications, conformations and methylation states will lead to a combinatorial explosion of both variables and reactions in the mathematical formulation. Finally, although PDEs are introduced to deal with the situation of heterogeneous distribution of participants in volumes, the kinetic models still have

limited capability of modeling spatial dynamics [48].

2.2 DATA-DRIVEN MODELING OF SIGNALING PATHWAYS

To study the organisms from a systematic perspective, the responses of the biological systems under perturbations need to be measured, including but not limited to the expression levels of genes, the activity levels of signaling proteins and even the localizations of related molecules over a range of time scales and interventions. Together, all these data reflect the states of the biological systems.

The transcriptome is the complete set of transcripts in a cell under a specific environmental condition or developmental stage. Recently, it can be measured by the next-generation sequencing technologies (RNA-seq) [49, 50]. As a sequencing-based method, RNA-seq allows the entire transcriptome to be sequenced in an ultra-high-throughput manner. Nowadays, this sequencing technology is available on several platforms such as Illumina [51] and 454 Life Sciences [52]. RNAs are first fragmented into a library of cDNA fragments. Then the short reads, either single-end or paired-end reads, are obtained from each cDNA. After mapping the short reads to the reference genome or transcriptome, the amount of exonic reads (the reads aligned entirely inside an exon) and exon-exon junction reads can be counted, thus making it possible to infer the expression levels of genes or isoforms.

Similar to the transcriptome, the proteome is the collection of the entire set of proteins in the cell. However, to study signaling pathways which consist of signaling proteins, usually it is not the total abundance of proteins that we concern. Instead, activity levels of signaling proteins are more informative, *e.g.*, time-series phosphorylation level data measured by phosphorylation-specific antibody microarrays [53] are employed to measure the up- or down-regulations of signaling protein activities to monitor how signals are trans-

mitted through signaling pathways.

To describe systems biology data, a “data matrix” [54] is introduced in which rows represent observations and columns indicate variables (*e.g.*, gene or signaling proteins). By plotting data matrix into a 2-dimensional coordinate system, the changes in signals over time can be easily visualized. However, with the increase of the number of variables, the overlap and intersection of different plots will make it difficult to distinguish one from others. An alternative view is to give each variable its own dimension. The nodes in this multi-dimension space represent time points and their projections along a specific dimension are the observations of that variable at the corresponding time points.

There are three well known data-driven modeling approaches for processing the data: clustering, principal component analysis (PCA) and partial least squares regression (PLSR) [54, 55].

Clustering is used to organize the data matrix to reveal insights into biological measurements by grouping variables with similar characteristics, *e.g.*, genes with similar functions or proteins with similar behaviours. The first thing to do in clustering is to define what the “similarity” is. For a data matrix, it is convenient to use the vectors (either the row vectors or the column vectors) as the input of clustering since the “similarity” here can be easily explained by the “distance” between the vectors [56]. A wide range of clustering algorithms have been proposed to identify the patterns of biological data, such as hierarchical clustering [57], *k*-means clustering [58] and self-organizing maps (SOM) [59]. Figure 2.3(a) to (c) gives an example of how different clustering methods deal with a sample with 270 observations [60, 61]. Suppose each gene (represented as a dot) is plotted according to its expression levels under two experiments. Hierarchical clustering is done by iteratively grouping the closest pair of points into a single point and the final output is a tree. *k*-means clustering is initialized with partitioning the whole space into *k* subspaces. The centroid of each subspace is set to the average of the points within the same cluster. The genes are then rearranged to the cluster of the nearest centroid. The process is ended when no more genes change

Figure 2.3: (a) *k*-means clustering. (b) Hierarchical clustering. (c) Self-organizing maps. (d) Principal Component Analysis.

cluster. One shortcoming of *k*-means clustering is that the results rely on the initial positions of the *k* centroids. SOM also starts with a pre-defined number of cluster centroids. A gene is selected in each iteration and the closest centroid, as well as its neighborhood, is moved toward the gene. So the neighboring clusters of the final output show related expression patterns. The clusters become steady as the radius of neighborhood gradually shrinks. Clustering achieves success when a subset of the data carry prior biological knowledge and benefits user by narrowing down the dataset they are dealing with.

To reduce the number of input variables (dimensions), principal component analysis (PCA) can be employed. PCA can achieve dimensionality reduction while capturing as much information (*i.e.*, the variance in measurements over all observations) as possible (Figure 2.3(d)). The principal components are linear combinations of the original dimensions with the coefficients indicating the contributions of each variable to a principal component. The observations can be projected to these principal components for further analysis such as discrimination. For example, using signaling protein data,

Figure 2.4: Discrimination of cell types using principal components.

PCA can be applied to qualitatively discriminate cell types as shown in Figure 2.4 [61]. Studies also indicated that cell fates [62] and effects of drug treatments [61] can be distinguished by PCA.

Extraction of principal components is also a step in partial least squares regression (PLSR) which is used when the goal is to predict causal relations between the independent and the dependent variables. While the purpose of principal components extraction in PCA is to capture as much variance as possible, PLSR seeks to maximize the correlations between the principal components of the independent variables and those of the dependent variables. Therefore, the principal components extracted by PLSR emphasize the independent variables that have strong covariance with the dependent variables rather than just the ones that have large variances. After a PLSR model is constructed, it can be used to predict the values of dependent variables (*e.g.*, probability of cell fates [63, 64, 61]) given a set of measurements of independent variables (*e.g.*, activity levels of signaling proteins).

Data-driven modeling is powerful when: (i) the underlying mechanisms

of signal transduction are not clear, *i.e.*, the topology of the signaling pathways is incomplete; and (ii) the goal is to link the signaling pathways to cellular responses, and because no biochemical reaction can explicitly describe these processes, canonical edges in a network are not suitable for representing these kind of relationships. However, it is still a limitation of data-driven modeling that it is unable to incorporate prior knowledge about the underlying mechanisms. Therefore, hybrid modeling, which combines knowledge-based and data-driven methods, would be a promising approach.

2.3 HYBRID MODELING OF SIGNALING PATHWAYS

Hybrid modeling is an approach that combines the strength of both knowledge-based and data-driven methods. It is able to reveal the insights of the underlying mechanisms and includes the context-specific information from the data at the same time, thereby showing great potential for modeling signaling pathways. In hybrid models, data are not only used for calibration and validation of the knowledge-based models, but also provide evidence and methodologies that complement with the knowledge-based models. For example, in the most common paradigm of *perturbation-signal-response* [62] experiments, the knowledge-based models are applicable to the *perturbation-signal* partition when the prior knowledge of the signaling transduction (*e.g.*, the topology of signaling network) is likely to be available, and the data-driven models are applicable to modeling the *signal-response* relationship when the flows of the signals are not fully understand.

Following this strategy, Peng *et al.* [65] proposed a hybrid model to link the signal transduction with the gene expression regulatory responses. They used a set of differential equations to forward simulate the NF- κ B signaling pathways and Network Component Analysis [66, 67, 68], which is mainly a data-driven method based on matrix decomposition, to reversely engineer a gene regulatory network. Then they matched the forward and reverse

Figure 2.5: An example of hybrid modeling.

engineering results and distinguished the corresponding signaling profiles among three gene expression profiles with different perturbations. A similar study of Melas *et al.* [69] first employed a multi linear regression algorithm to identify correlation-based relationships between signaling proteins and cellular responses (cytokine releases) and connected them using “non-canonical” edges. Integrating the canonical network of the signaling pathways from prior knowledge, the whole network was then converted into a Boolean model. Next, they optimized the network against the experimental data using Integer Linear Programming as in [33] and identified the pathway activity that induced the diverse cellular responses including growth, death and cytokine secretion. Figure 2.5 gives an example of hybrid modeling of gene expression regulated by signal transduction. Given the prior knowledge of the signaling pathways, a logical model is used to capture the dynamical trends of signal transduction. The relationships between transcription factors and target genes relationships are learned from the gene expression data using a data-

driven method (*e.g.*, correlation-based methods). The two layers (signaling pathways and gene regulatory networks) can be linked through transcription factors. And gene expression patterns can be predicted by a thermodynamic-based model.

2.4 CHAPTER SUMMARY

A key factor in the choice of modeling methods is the type of knowledge/data available about the biological system. Each type of modeling approach is best suited to different situations and will provide different insights (Table 2.1).

Table 2.1: Comparison of various methods for modeling signaling pathways.

	Advantage	Limitation
Boolean network	Simple and easy to build; effectively explains the dynamic behavior of biological systems; large existing toolkit.	Only binary state for each node; cannot provide quantitative predictions.
Generalized logical model	Simple and easy to build; continuous state, provide quantitative predictions.	mechanisms are not revealed very precisely.
Fuzzy logic	Continuous state ranging from 0 to 1.	Hard to apply to large networks.
Kinetic models	Quantitative and precise; direct comparison with quantitative measurements.	Require quantitative knowledge of initial values and kinetic parameters.
Data-driven models	Not rely on prior knowledge; simple and usually unbiased.	Have limited capability of revealing underlying mechanisms.
Hybrid models	Reveal the insights of the underlying mechanisms and include the context-specific information from the data.	More complex; need to find the balance between prior knowledge and data.

Chapter 3

GENERALIZED LOGICAL MODEL OF CELLULAR SIGNALING PATHWAYS

3.1 INTRODUCTION

Signal transduction plays an essential role in the cellular processes in which cell responds to extracellular perturbations (*e.g.*, the exposure to drugs or ligands). According to the signals, the cell adjust its metabolism, shape, gene expression, *etc.*, to adapt to the environment. It is widely believed that the dysregulation of signal transduction is one of the most important pathogenesis of many human diseases including cancer. Although high-throughput experimental data show a great potential for uncovering unprecedented details of biological systems, it is still challenging to understand signaling networks at systems level. Therefore, computational simulation, which is a systems biology approach, is highly desirable for the analysis of the underlying mechanisms of how the signals are transmitted through signaling pathways.

Many existing models are able to simulate the process of signal transduction, such as Boolean network models, fuzzy logic models and kinetic

models based on ordinary differential equations (ODEs). Boolean network is a simple and promising framework for the modeling of protein-protein interactions and signaling pathways. It has been used with some success in identifying stable states of a system [30, 31], simulating the influence of deletion/knockout of important nodes in a network [70], predicting carcinogenesis and targeted therapy outcomes [71], reproducing the dynamics of the yeast MAPK pathways [72], modeling the mammalian cell cycle [73] and analyzing the behaviors of the apoptosis pathways [74, 75]. However, its inability of encoding graded responses and the typically sigmoidal biological relationships becomes a significant limitation since it is able to handle only binary values, *i.e.*, a simple on/off state which is over-simplified compared with a real signaling network. To overcome this limitation, fuzzy logic models, which generalize the on/off characteristic to a continuous range from 0 to 1, have been successfully applied to analyzing the crosstalk among the TNF/EGF/Insulin-induced signaling pathways [37] and the liver cell responses to inflammatory stimuli [39]. However, a large amount of prior knowledge is needed for the assembly of the membership functions and logical rules for the fuzzy logic models. On the other hand, ODEs have also been applied to modeling various biological processes, such as the simulation of physiological responses of mammalian cells to the control of cell cycle [45], mathematical modeling of the mechanisms for regulating the differentiation of hematopoietic stem cells [76], discovery of signaling pathway rewiring [77] and exploring the dynamics of the pathways controlling cell apoptosis [46, 47]. However, ODEs-based models require a relatively detailed knowledge of kinetic parameters which is hardly available for all the pathways. Previously, we proposed a simulation tool called SimBoolNet [35] which is based on an extended Boolean network model. Although the performance of SimBoolNet in predicting protein activities was promising [36], it has limited capability of dealing with blocking effects, degradations and sequenced perturbations.

Here, we present a generalized logical model, which is capable of revealing the process of degradation and the effects of scheduled perturbations to signaling networks. Compared with SimBoolNet [35] and GINsim [78]

(a Boolean network based simulation tool), the proposed simulator can not only predict the stable states of the signal transduction system but also dynamically simulate the effects induced by various timing and ordering of perturbations. The simulations are validated using experimental phosphoproteomics data of breast cancer cells perturbed by different combinations of drug additions [61]. The simulated time-series data of protein activity levels show significant correlations with the real time-course data, thereby demonstrating that the proposed model is able to capture the key features of the signaling pathways.

3.2 METHODS

3.2.1 COMPUTATIONAL MODEL FOR DYNAMICAL SIMULATION

Our model takes a directed graph as the input network to do simulation. In the network, each node denotes a molecular species (*e.g.*, a protein) and each directed edge (u, v) represents signal transduction from node u to node v . One variable with a nonnegative value $\in [0, 1]$ (0 means fully inhibited and 1 indicates fully activated) is associated with each node to represent the activity level of the protein. The edge weight is also a variable with the value $\in [0, 1]$ to denote the strength of the interaction and a sign ('+' or '-') to denote the type of the interaction (*i.e.*, positive means activation and negative blockage).

Users can select input nodes and a virtual node is added upstream of all the input nodes. This virtual node is the abstraction of the extracellular environment which is able to generate signals that can stimulate or inhibit the input nodes (*e.g.*, the receptors) of the signaling network. The effect of the virtual starting node to the input nodes over time is set to follow a sigmoidal distribution $\frac{1}{1+\exp(-p(t'-t'_0))}$. Here t' is the user-defined maximum number of simulation iterations (T) minus the current simulation iteration

(t), because the plateau of the sigmoidal curve is supposed to appear in the beginning of the simulation describing the effect of the inputs to the receptors which is strong at the early stage and decreases in a relatively long time scale. The parameters p and t'_0 are the steepness of the curve and the t' -value of the sigmoid's midpoint, respectively. Since a complete sigmoidal curve is required given any user-defined T , both p and t'_0 are defined as the functions of T thereby the only parameter in the sigmoidal function is T . In our experiments, p is set to the reciprocal of $\frac{T}{10}$ to give a gentle slope and t'_0 is $\frac{T}{3}$ so that the midpoint is reached at the one third of the simulation process. Both functions are editable by users.

Given random initial activities for all the nodes in the network, their states will be updated synchronously based on their own previous states and the incoming signals from their parent nodes, according to Equation (3.1). In this formula, X_t is the activity level of node x at iteration t , d is a pre-defined parameter denoting the degradation rate (value $\in [0, 1]$) of the activated x from iteration $t - 1$ to t , A_i (or B_j) is the signals (*i.e.*, the activity level times the edge weight) transmitted from the i th activating (or j th inhibiting) parent node upstream of x ; $[1 - \prod(1 - A_i)]$ is the overall activating effect generated by all the incoming activating signals and $\prod(1 - B_j)$ is the probability that the incoming inhibiting signals do not affect x . Altogether, they act on the inactivated form of x at the $(t - 1)$ th iteration (*i.e.*, $(1 - X_{t-1})$). The blocking effect acting on the activated form of x follows the similar logic. Overall, the updated state is defined by the nondegraded part and the newly activated part minus the inhibited part. Given a user-defined T , the discrete steps are employed to approximate the process that the activity levels of the nodes change over time. Figure 3.1 shows the workflow of the simulation using the proposed model.

$$\begin{aligned}
 X_t = & (1 - d)X_{t-1} + [1 - \prod(1 - A_i)] \prod(1 - B_j)(1 - X_{t-1}) \\
 & - \prod(1 - A_i)[1 - \prod(1 - B_j)]X_{t-1}
 \end{aligned} \tag{3.1}$$

Figure 3.1: The workflow of simulation using the proposed model.

Compared with simple kinetic models based on the law of mass action, which is a statement about equilibrium in chemistry that defines the rate of a chemical reaction is proportional to the product of the concentrations of the reactant [79], the proposed model also consists of terms that can increase or decrease the activity (or concentration) of a component. However, the main difference between them is whether the system is in dynamic equilibrium. The law of mass action explains and predicts behaviors of a system in the dynamic equilibrium, while our proposed method does not necessarily make that assumption because the cellular signals can be amplified through transduction. For example, if there is no inhibiting signal being transmitted, the activity of a component will not decrease although it can still keep sending signals to its downstream. In other words, the signals can be transmitted but not consumed.

It is suggested that cells respond to external perturbations through a time-dependent (*e.g.*, the schedule and duration of drug addition) process [61]. Wet-lab experiments have shown that different ordering and timing of drug additions have significantly different drug effects, such as inducing specific alterations of signaling pathways [33, 39, 34] and showing different efficiencies in killing cancer cells [61]. However, most existing simulation tools are not able to accommodate the time-staggered design of drug treatments

in biological experiments. Therefore, our model introduces time-staggered perturbations to explore the effects of not only dosage, but also the schedule and duration of the perturbations to cellular systems with a knowledge-based model. The timing and the order of drug additions can be specified by users as parameters. For example, the drug can start to affect its target at the m -th simulation iteration with a user-defined m . The target, input level, type of interaction (stimulation or inhibition) and schedule of the perturbations can all be specified according to user's design of experiment.

3.2.2 NETWORK STRUCTURE

A signaling network [80] (Figure 3.2) is constructed according to well-known pathway databases (GeneGO MetaCore [81] and KEGG [82]). The network comprises 35 nodes, indicating 32 signaling proteins or stimuli (round rectangles) and 3 cell fates (ellipses), and 57 edges (activation and inhibition interactions are denoted by green arrow and red flat-head edges, respectively.) denoting signal transduction from the source nodes to the target nodes. Dark blue nodes in the pathways represent 21 signaling proteins that have been experimentally measured in [61].

3.3 RESULTS

3.3.1 PERFORMANCE COMPARISON ON SIMULATING SIGNALING RESPONSES TO PERTURBATIONS

For comparison, we run simulations using our program based on SimBoolNet [35], GINsim [78] and the proposed model on the same network in Figure 3.2. Two different inputs are introduced: (1) the input levels of EGFR and TNFR are set to 0.5 and 0.8, respectively, (2) EGFR inhibitor is added

Figure 3.2: Signaling pathways constructed based on the dataset.

at the 10th iteration of simulation and TNFR is activated with input level 0.8 at the 20th iteration. The number of simulation iterations is set to 100. For the proposed model, the full inhibition is denoted as -1 and the perturbations can be executed at any iteration during the simulation. The degradation rate d is set to 0.2. For SimBoolNet and GINsim, the blocking effect is represented by setting the activity level of EGFR to 0 from the very first step of simulation which, to our understand, is unlikely to be a precise representation. There should be a process for the inhibitor to reduce the activity level of its target, especially when the inhibitor is not added at the beginning. The edge weights of activation and blockage are set to 0.7 and 0.8, respectively. GINsim simulation, on the other hand, does not accept parameters for edge weights and the number of iterations, and executes synchronously until the system reaches the stable state. GINsim also supports

the asynchronous mode, but it is a time-consuming task due to a much larger search space than with the synchronous mode. We did not get any result from running GINsim in asynchronous mode on our network (Figure 3.2) within endurable time using a desktop PC (Dell Precision T3600 workstation with Intel Xeon CPU E5-1620, 8GB RAM and Windows 7 Professional 64-bit operating system). We have also tried other different settings of input level, edge weight and degradation rate, and the results are shown in Section 3.3.2.

Figures 3.3(a), 3.3(b) and 3.3(c) show the simulation results of three proteins (*i.e.*, EGFR, TNFR and ERK) using the proposed model under the two different inputs. In Figures 3.3, the time points on the x-axis are the simulation iteration, the protein activities on the y-axis are normalized into 0 to 1, the values are non-dimensional. It can be seen that the trends of the activities of the input nodes follow an approximately sigmoidal function (the blue curves in Figures 3.3(a) and 3.3(b)). When the EGFR inhibitor is added at the 10th step, the EGFR activity drops sharply within a few steps (the red curve in Figure 3.3(a)). Consequently, the activity of its downstream node ERK (the red curve in Figure 3.3(c)) decreases with some time delay because it takes some time for signals to be transmitted from EGFR to ERK. Under input 2, the activity of TNFR first decreases from a random initial value with a degradation rate (here is 0.2) and increases to the maximum (the input level 0.8) almost immediately at the 20th iteration (the red curve in Figure 3.3(b)). In contrast, SimBoolNet has a limited capability of dealing with blocking effect, degradation or scheduled perturbations. We can see from the blue curves in Figures 3.3(d), 3.3(e) and 3.3(f) (which are outputs of SimBoolNet) that the activity levels increase monotonically from 0 to a maximum, considered as the stable state, which is unlikely to be precise in the biochemical reactions. Moreover, the inhibiting effect in SimBoolNet is represented by keeping the activity level of the target node to 0, which is unlikely to be realistic especially when the inhibiting effect should be produced in the middle of the simulation (red curves in Figures 3.3(d) and 3.3(f)). Given the initial states, GINsim is able to identify the stable states [31].

However, it has similar limitations for dealing with scheduled drug additions and degradations (Figures 3.3(g) to 3.3(i)).

Figure 3.3: Comparison of simulation-based predictions using the proposed simulator (a–c), SimBoolNet (d–f) and GINsim (g–i).

We then compared the computational time of SimBoolNet, GINsim (in asynchronous mode) and the proposed model. In addition to the network shown in Figure 3.2, a small network with 5 nodes and a large network with 500 nodes were constructed. The edges of these two networks were randomly generated. Totally, the numbers of the nodes/edges of the three networks are 5/13, 35/57 and 500/6124, respectively. The hardware being employed was Dell Precision T3600 workstation with Intel Xeon CPU E5–1620, 8GB RAM and Windows 7 Professional 64–bit operating system. Table 3.1 shows the computational time of SimBoolNet, GINsim and our model on the three networks. It can be seen that the simulation with our model is faster than SimBoolNet, whereas slightly slower than GINsim when the network is small. However, GINsim has its limitation for large networks. The simulation using GINsim (in synchronous mode) on the third network with 500 nodes did not produce any result within endurable time (hence marked “Not applicable”).

The influence of different initial values on the simulation results was

explored in the following mathematical proof. Reorganize the right side of the Equation (3.1) we have Equation (3.2):

$$X_t = [1 - d - (\prod(1 - A_i) + \prod(1 - B_j) - 2 \prod(1 - A_i) \prod(1 - B_j))]X_{t-1} + [1 - \prod(1 - A_i)] \prod(1 - B_j) \quad (3.2)$$

Introduce $A_{t-1} = \prod(1 - A_i) \in [0, 1]$ and $B_{t-1} = \prod(1 - B_j) \in [0, 1]$ into Equation (3.2):

$$X_t = [1 - d - (A_{t-1} + B_{t-1} - 2A_{t-1}B_{t-1})]X_{t-1} + (1 - A_{t-1})B_{t-1} \quad (3.3)$$

Introduce $P_k = 1 - d - (A_k + B_k - 2A_kB_k)$, $k = 0, 1, \dots, t-1$, we have:

$$\begin{aligned} X_t &= P_{t-1}X_{t-1} + (1 - A_{t-1})B_{t-1} \\ X_{t-1} &= P_{t-2}X_{t-2} + (1 - A_{t-2})B_{t-2} \\ &\dots \\ X_2 &= P_1X_1 + (1 - A_1)B_1 \\ X_1 &= P_0X_0 + (1 - A_0)B_0 \end{aligned} \quad (3.4)$$

Table 3.1: Comparison of the computational time required for SimBoolNet, GINsim and the proposed model.

Number of Nodes/edges	SimBoolNet	GINsim	Our model
5/13	9.51s	< 0.1s	0.09s
35/57	10.73s	< 0.1s	0.43s
500/6124	72.88s	Not applicable	10.68s

Substitute the last equation in Equation (3.4) into the second from the bottom:

$$\begin{aligned} X_2 &= P_1(P_0X_0 + (1 - A_0)B_0) + (1 - A_1)B_1 \\ &= P_1P_0X_0 + P_1(1 - A_0)B_0 + (1 - A_1)B_1 \end{aligned} \quad (3.5)$$

similarly we have:

$$\begin{aligned} X_3 &= P_2(P_1P_0X_0 + P_1(1 - A_0)B_0 + (1 - A_1)B_1) + (1 - A_2)B_2 \\ &= P_2P_1P_0X_0 + P_2P_1(1 - A_0)B_0 + P_2(1 - A_1)B_1 + (1 - A_2)B_2 \end{aligned} \quad (3.6)$$

then one by one recursively until the first equation in Equation (3.4), we have:

$$\begin{aligned} X_t &= P_{t-1}P_{t-2}\dots P_1P_0X_0 + P_{t-1}P_{t-2}\dots P_1(1 - A_0)B_0 \\ &\quad + P_{t-1}P_{t-2}\dots P_2(1 - A_1)B_1 + \dots + P_{t-1}(1 - A_{t-2})B_{t-2} \\ &\quad + (1 - A_{t-1})B_{t-1} \end{aligned} \quad (3.7)$$

Since the range of the bivariate function $z = A_k + B_k - 2A_kB_k$ with the domain $A_k, B_k \in [0, 1]$ is $z \in [0, 1]$, and $d \in [0, 1)$, the range of $P_k = 1 - d - (A_k + B_k - 2A_kB_k)$, $k = 0, 1, \dots, t-1$ is $P_k \in (-1, 1)$. When t is large enough, the product of $P_{t-1}P_{t-2}\dots P_1P_0$ which is the coefficient of the initial state X_0 towards zero making the series $x_t, x_{t+1}, \dots, x_{T-1}, x_T$ independent of the initial state X_0 . Therefore, the proposed model is robust to the variations of the initial states of the proteins.

Figure 3.4(a) shows the simulation curves of EGFR under the aforementioned input 1. Three different initial values of EGFR, namely 0.1, 0.5

and 0.9, were selected. Conclusions can be drawn from the plot that the initial value affects the time spent in reaching the stable state, but the overall trends of the state transitions and the levels of the stable states remain stable.

Figure 3.4: (a) The influence of different initial values on simulations. (b) The influence of different edge weights to simulations.

We went on to explore the robustness of the proposed model to the variations of edge weights. In principle, we randomly generated the edge weights to run the simulations, and then checked if the activity trend of each protein remained unchanged. For a specific input (*i.e.*, the input levels of EGFR and TNFR are both 0.5), we first randomized the weights of all the 57 edges and ran the simulation for 100 times as the background group. For each protein, the mean activity at each time point was regarded as the background trend over time. Next, we further generated 50 groups of simulations, each group consisting of 100 simulations of randomly generated edge weights. For each group, the mean activity trend of each protein was used to calculate the correlation with the background trend. Figure 3.4(b) gives the distribution of the 50 correlations between the simulated and the background trends for the 32 non-receptor nodes (ignoring the receptors EGFR, TNFR and DNA Damage because they have no incoming edges), and the proteins are ranked based on the median of the correlations. It can be seen that 21 out of 32 proteins (*i.e.*, from PI3K to AKT) have the medians of the correlations larger than 0.8; 10 out of 32 have the medians falling into the interval 0.5 to 0.8; and only one (*i.e.*, Proliferation) has the median which is lower than 0.5. Moreover, all the 32 signaling proteins show small ranges of the correlations between the simulated and the background trends, indicating that the proposed model is robust for capturing the dynamical trends of the signal transduction process under different settings of edge weights.

3.3.2 MODEL COMPARISON AND VALIDATION WITH REAL DATA

To evaluate the performance of the proposed model, the simulated results using the network in Figure 3.2 are compared with a real signaling dataset [61] containing the time-series phosphoproteomics data. In the dataset, perturbations (*i.e.*, inhibitor of EGFR or stimuli of DNA damage or both) were applied to cells of the breast cancer cell line BT20. For each perturbation, activity levels of 32 signaling proteins (21 out of 35 are included in the network

in Figure 3.2) were measured at 8 time points. For comparative purpose, activity levels of each protein are normalized to the same scale as the simulated data, *i.e.*, between 0 and 1. To simulate the perturbations, we use (i) half activation input signals (0.5) to represent the control situation where no stimuli or inhibitor is added; (ii) activation input signals (+1) to represent the addition of stimuli, *i.e.*, the targets are fully activated; and (iii) inhibition input signals (-1) to represent the effect of inhibitors, *i.e.*, the activity of the targets are suppressed.

We first simulated the dynamics of signaling without any drug addition, as a control dataset. The receptors EGFR and TNFR were selected as the input nodes with input levels both equal to 0.5. The edge weights of both activation and inhibition were set to 0.8. The simulation was executed for a number of iterations that is a multiple of 8 since there are 8 time points in the real dataset (here we chose 32 because most of the nodes reach the stable states after 30 steps). We then used Spearman and Pearson correlation coefficients to measure the goodness of fit between the simulated and real data to evaluate the performance of our model. Since calculating correlations requires the two vectors to have the same length, 8 time points of the simulated data (which is the same as the real data) were extracted from the simulated time-series with equal interval, *e.g.*, the 4th, the 8th, ... and the 32nd time points. Table 3.2 gives the Spearman and Pearson correlation coefficients for the 21 measured signaling proteins. We can see that for 16 out of 21 signaling proteins the Spearman correlation coefficients are relatively high (larger than 0.6) indicating that the simulation fits well to the real data. Similar conclusion can be drawn from Pearson correlations. For the proteins with relatively poor correlation coefficients, it might be because of over-simplification such as missing some cell line specific interactions in the network. It is noticed that for some proteins the two correlations have quite different values, for example, the two correlations of S6 and S6K have different signs. We then rechecked the data and one outlier has been identified in both S6 and S6K data. After removal of the outlier, the correlations of S6 and S6K are changed to the values in the parentheses which indicates that

our model performs well on S6 and S6K.

Table 3.2: Spearman and Pearson correlations between simulated and real data.

	SMAC	4EBP1	p53	ERK	S6	S6K	CABL
Spearman	0.95	0.88	0.69	0.71	0.31(0.96)	0.26(0.89)	-0.90
Pearson	0.94	0.82	0.95	0.43	-0.42(0.95)	-0.76(0.98)	-0.86
	Casp9	CDC25	CHK	p27	PUMA	AKT	JNK
Spearman	-0.60	0.62	0.21	-0.97	0.92	0.79	0.43
Pearson	-0.75	0.68	0.45	-0.96	0.98	0.73	0.44
	p38	BIM	BID	RIP1	CYCLIN	Casp8	EGFR
Spearman	0.67	0.93	0.43	0.76	0.86	0.90	0.62
Pearson	0.65	0.91	0.30	0.70	0.40	0.85	0.71

On the other hand, since the simulation is an approximately continuous process while the wet-lab experiment measured only at a few selected time points, we can scale the discrete dots in the real data to align with the 32 simulated steps by multiplying the index of each time point in the real data by 4 to match the index of each iteration in the simulation. For comparison, SimBoolNet was also employed to do the simulation under the same inputs, *i.e.*, the input levels of EGFR and TNFR are both 0.5 and the edge weights are all 0.8. Figure 3.5 shows the plots of the simulated (scatter plots each with a trend line) and real data (scatter plots) for EGFR, Caspase 8, p53 and JNK. The blue and green curves represent the simulated data using the proposed model and SimBoolNet, respectively, while the red dots represent the real data (normalized to the same scale as the simulated data). In Figure 3.5(a), the simulation curve of EGFR captures the trend of a quick drop from a plateau to the flat bottom although there is a time delay. The slow and small decrease of the activity of Caspase 8 is also captured by our simulation as shown in Figure 3.5(b). For p53 (Figure 3.5(c)), the simulation shows a high decreasing rate at the beginning and the activity of p53 reaches the stable

state quickly, which agrees with the trend of the real data. The gradient of the JNK simulation curve (Figure 3.5(d)) also fits the real data although the starting point of the dropping lags behind, which suggests that it is important to scale the simulated timing to accurately match the real time points. To address the issue of time scale would be an important future work. By contrast, the simulated trends of SimBoolNet are mainly monotonically increasing which cannot fit the real data well.

Figure 3.5: The plots of the simulated and real data in the control group. The blue and green curves are the simulated data using the proposed model and SimBoolNet, respectively. The red dots are the real data.

We then simulated the state transitions of signaling pathways under the perturbations in Table 3.3. For example, we gave half activating (+0.5) and full blocking (-1) signals to TNFR and EGFR, respectively, to simulate the addition of EGFR inhibitor. The edge weights of activation and inhibition were set to 0.7 and 0.8, respectively. The number of simulation iterations was set to 32, and the degradation rate 0.2.

It has been known that the *in vivo* drug effect on the signaling pathways is through the change of the activities of the proteins downstream of the drug targets [33, 60]. Therefore, for these downstream proteins, if the simulation data based on the perturbed inputs (row 3 to 5 in Table 3.3) fit

the experimental data better than the simulations using the control input (row 2 in Table 3.3), we believe our simulator captures the main influences of the drugs on the networks. For example, the second and the third columns of Table 3.4 show the correlation coefficients between the real data and the simulations using the control input and EGFR inhibitor input (rows 2 and 3 in Table 3.3), respectively. It can be seen that, for the nodes downstream of EGFR (such as AKT, S6, S6K and BIM), the fitness of simulations to the drugged data (biological measurements treated with drug which targets at EGFR) is significantly improved when the control input is replaced by the EGFR inhibitor input. A similar conclusion can be drawn for the proteins downstream of DNA damage, including p53, CABL and Caspase 8, when the simulations using the control input and DNA damage stimuli as input are employed to fit the real data (the forth and the fifth columns of Table 3.4). Moreover, it is believed that the treatment with both drugs have a higher efficiency on killing cancer cells than that with a single drug [61] which may be explained by the characteristics of SMAC and Caspase 8 (both upstream of Caspase 3), in that the improvement of fitness is mainly achieved under the treatment of both drugs (the last two columns of Table 3.4). There are proteins, such as ERK and CHK (downstream of EGFR and DNA damage, respectively), that do not follow the above patterns, probably due to our insufficient knowledge about signaling pathways.

Table 3.3: Inputs of the simulation in Section 3.3.2. The columns are the input nodes of simulation and the rows are various conditions indicating the treatments of biological experiments with drugs corresponding to wet-lab experiments. The symbols ‘+’ and ‘-’ represent the perturbation types, *i.e.*, activation and inhibition.

	EGFR	TNFR	DNA damage
control	+0.5	+0.5	random value
EGFR inhibitor	-1	+0.5	random value
DNA damage stimuli	random value	+0.5	+1
Both drugs	-1	+0.5	+1

3.4 CHAPTER SUMMARY

In this chapter, we present a model to dynamically simulate the process of intracellular signal transduction. According to a phosphoproteomics dataset [61], we constructed a network, which comprises 35 nodes (21 nodes have experimental measurements) and 57 edges, to do the simulation. The state of each node is calculated based on its own previous state with a degradation rate, and the activation and inhibition effects produced by the signals transmitted from its parent nodes. Different combinations of perturbations were applied to the network. The simulation results have been evaluated with the real data, demonstrating that our simulator has the ability to grasp the main dynamical trends of signal transduction. Compared with SimBoolNet [35] and GINsim [78], the proposed model shows promising performance in revealing the graded responses and the effects of scheduled perturbations to a signaling network. Moreover, by testing the proposed model with different values of

Table 3.4: Goodness of fit of the simulations to the real experimental measurements under drug treatments.

	control input	EGFR inhibitor	control input	DNA damage stimuli	control input	both drugs
AKT	0.64	0.81	-0.40	-0.54	0.75	0.93
ERK	0.32	0.40	-0.30	-0.39	0.19	0.23
S6	0.53	0.86	0.42	0.57	0.52	0.89
S6K	0.49	0.84	0.35	0.66	0.58	0.96
4EBP1	0.67	0.67	-0.33	-0.33	0.92	0.92
BIM	0.32	0.72	0.35	0.65	-0.15	-0.38
BID	0.30	0.37	0.03	0.03	0.16	0.16
JNK	0.02	-0.07	-0.47	-0.58	-0.06	0.01
p53	-0.04	-0.45	0.21	0.83	0.01	-0.09
CABL	-0.45	-0.52	0.04	0.55	0.21	0.78
CHK	-0.12	-0.18	-0.25	-0.28	-0.30	-0.69
CDC25	0.29	0.23	0.15	0.33	0.18	0.64
Casp8	0.06	-0.06	0.14	0.72	0.26	0.68
SMAC	-0.05	0.08	-0.71	-0.87	0.48	0.99

parameters (*e.g.*, the initial activities of the proteins and the edge weights), we have shown that our method performs robustly in revealing the dynamics of the signaling pathways when the prior knowledge of the network topology is reliable.

Studying the cell responses to extracellular perturbations is a major endeavor for biomedical research and pharmaceutical industry. With the development of high-throughput experiments, large-scale data are available to help uncover important biological mechanisms at systems level. However, most existing data-driven methods [60] (will discuss in Chapter 4) have limitations in revealing underlying molecular mechanisms. Therefore, computational simulation based on the integration of prior knowledge with data shows a great potential for revealing insights into the dynamical system of signal transduction, and thus would be a valuable complement to the data-driven methods. Although the proposed model is still limited in mapping the simulation steps to the experimental time points, we believe that the integration of both knowledge and data, such as learning the edge weights from experimental data, would be a powerful approach to understanding the signal transduction networks. In addition, generalization of the present model, which uses the synchronous updating scheme, such that it is able to deal with asynchronous dynamics, *e.g.*, updating a randomly selected node at each time point, would also be a valuable future direction.

Chapter 4

NONLINEAR POWER-LAW MODELING OF CELLULAR SIGNALING PATHWAYS

4.1 INTRODUCTION

The extracellular signals are captured and transmitted into the cells through signaling pathways. Generally, signal transduction involves various protein modifications such as phosphorylation. The signals are transmitted down to the nucleus or other cellular organelles to regulate their physiological functions and control the cell fates (*e.g.*, apoptosis, proliferation and cell cycle). Although wet-lab experimental evidence has suggested some proteins to be crucial to one of the cell fates (*e.g.*, engagement of TNF is able to trigger cell death [30]), it is still unclear how the signaling proteins are integrated to systematically decide the cell fates. Therefore, computational models are needed to study how the signaling “input” is related to the cell fate “output”.

Many statistical methods have been employed to model the relationship between the input and the output of a system, including the autoregressive with exogenous terms (ARX) model and the power-law statistical model.

In an ARX model [83, 84] the current value of each variable of interest is predicted based on its previous values in the time series and the past values of the exogenous series. The power-law distributions [85, 86], in which some quantity varies as a power of another quantity, have been investigated and applied in an extraordinarily diverse range of scientific areas, such as earth and planetary sciences [87] and social sciences [88]. In biology, studies have also shown that many complex phenomena of living systems scale with the biomass in a simple power-law fashion. For example, the metabolic rate scales as the $\frac{3}{4}$ power of mass [89] across nearly all life forms and the population density is inversely proportional to the individual body size with a power $-\frac{3}{4}$ [90]. However, to the best of our knowledge the power law has not been applied to studying the specific relationship between cellular signaling and cell fates (*e.g.*, apoptosis of cancer cells).

A straightforward way to model a signaling pathway is to assume a linear relationship between the inputs and the outputs. Janes *et al.* [64], Gaudet *et al.* [63] and Lee *et al.* [61] applied linear models to identify the effects of extracellular perturbations such as drug combinations on cancer cell death. They assumed that the phenotypic outputs are linearly related to the signaling inputs. By using the method of partial least squares regression (PLSR), different cellular responses between cell lines as well as various drug treatments were detected. However, signaling pathways are so complex in their interactions (*e.g.*, multiple levels of cross-talks and feedbacks) that their behaviours are expected to be nonlinear [91, 92]. Legewie *et al.* [93] found that the inhibition of pathways that control apoptosis results in a positive feedback leading to bistability. Eissing *et al.* [94] also revealed that a positive feedback loop plays an important role in signal-induced apoptosis. Callard *et al.* [95] summarized several nonlinear properties of biological systems, such as the difficulty in predicting the behaviors of the whole system based on parts, a small change in a specific component can have significant effect on the system thus making this component essential. They also studied typical cytokine dose-response curves and indicated that cells have no response if the concentration of cytokine is below a certain value; on the other hand,

the cellular response is approximately an exponential function of the signal quantity until the response reaches a plateau maximum.

In this chapter, we propose a power-law model to predict the probabilities of cell fates based on the activity levels (*e.g.*, the phosphorylation levels) of signaling proteins. Simulations based on generalized Boolean network and ordinary differential equations are exploited for model validation. The results show that the proposed model has better performance than the linear models on cell fate prediction. Moreover, the proposed model is also able to identify the signal transduction events that are blocked by the drugs, thereby revealing the drug-induced signaling pathway alterations. When our nonlinear model is applied to cell line discrimination, the cell lines are much better separated and more concentrated compared with the results of the linear model. Testing on 3 different cancer datasets [96, 63, 61] suggests that our nonlinear power-law model has some superiority over the linear model.

4.2 DATA AND METHODS

4.2.1 DATA

We downloaded 3 cancer datasets to evaluate the performance of the proposed model. The first dataset is from the published work of Lee *et al.* [61]. In this dataset, activity levels of 32 signaling proteins are measured at 8 time points. The measurements of 6 types of cell fates (*i.e.* apoptosis, proliferation and cell cycle phases including G1, S, G2 and M) refer to the normalized numbers of cells of the cell fates which are measured at 5 different time points using flow cytometry. 6 treatments designed for 3 breast cancer cell lines (*i.e.* BT20, MDA-MB-453 and MCF7) are performed on triplicate plates (*i.e.* 3 replicates). In each cell line, there are 18 observations of both signals and cell fates at each time point. In total, there are 432 (3 replicates $\times$ 8 time points $\times$ 6 treatments $\times$ 3 cell lines) measurements for each signaling protein and 270 (3 replicates $\times$ 5 time points $\times$ 6 treatments $\times$ 3 cell lines)

observations for each cell fate. The preprocessing of the data was as follows. First, we extracted the signaling “inputs” and cell fate “outputs” with the same experimental treatments and the cell lines. Second, we removed the observations with 0 or negative values or NaNs (not measured). Finally, there are 164 observations left (75 of BT20, 56 of MDA-MB-453 and 33 of MCF7). We further calculated the proportion of apoptotic cells under each observation, because our model aims to predict the cell apoptosis.

The second dataset [63] contains measurements of phosphorylation levels acquired from HT-29 cells (human colorectal adenocarcinoma cell line) under 12 different treatments (*i.e.* combination of 3 stimuli of TNFR, EGFR and Insulin). In each treatment, the phosphorylation levels of 19 proteins are measured at 13 time points. Four types of cell apoptotic responses (*i.e.* cleaved caspase-3, sub-G1, phosphatidylserine exposure and membrane permeabilization) are measured at 3 time points. For every experiment there are 3 replicates (except at time point 0 at which 6 replicates were made). As such, the second dataset consists of 504 (3 replicates $\times$ (13+1) time points $\times$ 12 treatments) measurements for each signaling protein and 108 (3 replicates $\times$ 3 time points $\times$ 12 treatments) measurements for each cell response. For HT-29 cell line data, the 19 signals at all 13 time points (247 dimensions), the instantaneous-derivative between each pair of adjacent time points for all the 19 proteins (247 dimensions), the maximum signal, the mean signal and the steady-state signal for each protein constitute a 570-dimensional space. For cell responses, a 12-dimensional space was extracted. All the data were normalized into the interval (0, 1) using the sigmoid function.

The third dataset is from the DREAM8 (Dialogue for Reverse Engineering Assessments and Methods) challenge in 2013 [96]. The DREAM8 dataset consists of 4 breast cancer cell lines, namely, BT20, BT549, MCF7 and UACC812, and each cell line is perturbed by the combination of 3 inhibitors and 8 ligand stimuli. These 3 inhibitors target at AKT, AKT with MEK, and FGFR1 with FGFR3, respectively. The 8 ligand stimuli include Serum, PBS, EGF, Insulin, FGF1, HGF, NRG1 and IGF1. In each cell line, a number of phosphoproteins are measured at 7 time points after a perturba-

tion, *e.g.*, 48 phosphoproteins in BT20, 45 in BT549, 41 in MCF7 and 45 in UACC812. Overall, there are 234, 252, 312 and 312 observations for cell lines BT20, BT549, MCF7 and UACC812, respectively. Note that this dataset contains only breast cancer proteomics data without any measurement of cellular response. The DREAM8 data released in 2013 are normalized using a procedure developed in [97]. We reversed the procedure to generate the raw data as follows. First, we multiplied each normalized linear value by the correction factor to get the median-centered ratio in linear value. Second, we converted median-centered ratios in a linear value to a median-centered $\log_2$ value. Then, we divided each median-centered data by its standard deviation for normalization.

4.2.2 NONLINEAR POWER-LAW MODEL

Suppose P is the score indicate the probability of cell death, x_i ($i = 1 \dots n$, where n is the number of measured signaling proteins) indicates the signaling activity (*e.g.*, phosphorylation level) of the i -th protein and α_i ($i = 1 \dots n$) represents the contribution of the i -th protein to the cell death. Then our goal is to construct a statistical model $P = \mathcal{F}(x_1, x_2, \dots, x_n)$ that relates the signaling “inputs” to the phenotypic “outputs”. The most commonly used linear model [61, 63] can be written as

$$P = \sum_{i=1}^n \alpha_i x_i + \varepsilon \quad (4.1)$$

where ε is a noise term representing the influence of the unmeasured signaling proteins to cell death.

If we regard the activities of signaling proteins as the independent variables and the cell death as the dependent variable, the coefficient α_i can be identified by regression methods. These parameters can be used to capture the covariance between independent and dependent variables, *i.e.*, to infer their causal relationship [54].

Despite promising results of using the linear model for signaling analysis [61], it is unlikely that the probabilities of cell fates are proportional to every molecular species with a constant ratio, *i.e.*, the linear function in Equation (4.1) may be unable to model the nonlinear relationship in a biological system. To search for suitable nonlinear models, we first designed simulation experiments to study how cell fates are related to signaling proteins one at a time, *i.e.*, construction of a function with a single variable. We first extracted a network (Figure. 4.1(a)), which describes cell death regulation, from the model suggested by Lee *et al.* [61]. In this network, cell death is enhanced by DNA damage while inhibited by EGFR. The arrow shape represents activation while a flat-head edge means inhibition. Pink nodes denote the species that have experimental measurements. Then SimBoolNet, an open source Cytoscape plugin which had been designed for dynamically simulating the process of signal transduction based on an extended Boolean network model [35], was employed to run the simulations. We selected EGFR and DNA Damage as the input signals. Since we have insufficient prior knowledge to determine the parameters here, the default settings of the software were employed: the edge weights (value $\in [0, 1]$) of both activation and blockage were set to 0.8; moreover, the input levels (value $\in [0, 1]$) of EGFR and DNA Damage were both set to 0.8. The simulation generated time-series data contain 100 time points. Figure. 4.1(b) shows the simulation results of DNA Damage and Cell Death. Considering the simulated time-series data as vectors with length of 100, the Curve Fitting Tool from MATLAB R2015b was used to analyze the data and identify the function that best describes the relationship between cell death and each node in the network. The results may provide a clue to help construct a function that relates cell death with all signaling proteins (*i.e.* a multivariate function).

For comparison, we employed four different models, *i.e.*, power function, linear function, exponential function and Gaussian function (Equation 4.2), to do the regression. In Equation 4.2, y and x represent the cell death and the activity level of a signaling protein at the same time point t ($t = 1 \dots 100$). Four statistical measures generated by the Curve Fitting Tool

Figure 4.1: (a) A signaling network for cell death regulation. (b) The simulation results of DNA Damage and Cell Death.

were used to evaluate the goodness of fit: sum of squares due to error (SSE), R-Square (value $\in [0, 1]$), degrees of freedom adjusted R-Square (value ≤ 1) and root mean squared error (RMSE). For SSE and RMSE whose values are always between 0 and 1, a value closer to 0 indicates that the model has a smaller random error component and that the fit is better. By contrast, for R-Square and adjusted R-Square, a value closer to 1 indicates a better fit.

$$\begin{aligned}
 \text{Power :} & & y &= a \times x^b \\
 \text{Linear :} & & y &= a \times x + b \\
 \text{Exponential :} & & y &= a \times \exp^{b \times x} \\
 \text{Gaussian :} & & y &= a \times \exp^{-\frac{(x-b)^2}{c^2}}
 \end{aligned}
 \tag{4.2}$$

To relate a dependent variable to more than one independent variables in a biological process, Savageau used the *Power-Law* formalism in biochemical systems theory such that one variable (*e.g.*, the rate of a biological pro-

cess) is represented as a power function of others (*e.g.*, the concentrations of biological moleculars) [98, 99]. Since many biological processes can be approximated by a straight line in a *log-log* plot over a wide range [100], we formulated the model as a product of power functions of the variables, given as

$$y = c \cdot \prod_{i=1}^n x_i^{\beta_i} \quad (4.3)$$

where y and x_i represent the components of the biochemical system, and c is a constant.

Therefore, we combined all the 6 nodes (*i.e.*, DNA Damage, EGFR, Oncogenic Signature, Casp8, Casp9 and Casp3) together as independent variables and fit the simulation data into Equation (4.1) and Equation (4.3), respectively. Least-squares regression was employed to estimate the coefficients of the regression functions, and root mean squared error (RMSE) was used to measure the goodness of fit.

Given the results in Section “Analysis of simulation data”, we propose a model to relate the cell death to the activities of all the measured signaling proteins, given as

$$P = e^{\beta_0} \cdot \prod_{i=1}^n x_i^{\beta_i} + \varepsilon \quad (4.4)$$

where the parameter $\beta_i (i = 1 \dots n)$ represents the contribution of the i th protein to the cell death and ε is a small constant term to keep the function from crossing the origin (*i.e.*, the value of a dependent variable is not required to be zero when the value of an independent variable equals zero). The proposed model is used to describe the system at any time point during a biological process, not necessarily the steady state.

The sensitivity of P to $\beta_i (i = 1 \dots n)$ is as follows: The Equation (4.5)

below shows the rate of the percentage change in the dependent variable P with respect to the percentage change in an independent variable β_i . The derivation shows that the rate is proportional to the value of the parameter β_i , demonstrating that the variance of β_i with large value leads to significant change of P . Furthermore, we tested the robustness of the parameters β_i ($i = 1 \dots n$, where $n = 32$, which is the number of measured proteins) by perturbing the values within $\pm 30\%$ [101]. Table 4.1 shows the results of the perturbation test for parameters. It can be seen that the proposed model is more sensitive to the parameters with larger absolute values which consists with the derivation.

$$\begin{aligned}
 S_{\beta_i}^P &= \frac{\% \Delta P}{\% \Delta \beta_i} \\
 &= \frac{\partial P}{\partial \beta_i} \cdot \frac{\beta_i}{P} \\
 &= \frac{\partial (e^{\beta_0} \cdot \prod_{j=1}^{i-1} x_j^{\beta_j} \cdot x_i^{\beta_i} \cdot \prod_{k=i+1}^n x_k^{\beta_k} + \varepsilon)}{\partial \beta_i} \cdot \frac{\beta_i}{P} \\
 &= (e^{\beta_0} \cdot \prod_{j=1}^{i-1} x_j^{\beta_j} \cdot x_i^{\beta_i} \cdot \ln x_i \cdot \prod_{k=i+1}^n x_k^{\beta_k}) \cdot \frac{\beta_i}{P} \\
 &= \ln x_i \cdot \frac{P - \varepsilon}{P} \cdot \beta_i
 \end{aligned} \tag{4.5}$$

4.2.3 CELLDESIGNER SIMULATION

In addition to the above simulation using Boolean Network (*i.e.*, SimBool-Net), we also compared the linear model with the nonlinear model on simulating data generated by solving ordinary differential equations (ODEs) which is a continuous model of biological systems with known biochemical kinetics. ODE models have been successfully applied to modeling numerous processes in living systems [45, 76]. CellDesigner is a process diagram editor for biochemical networks as well as an ODE-based simulator [102]. We chose CellDesigner to test if our nonlinear model can accurately describe the

Table 4.1: Effects of parameters on cell death predictions.

Selected parameters	+30% parameter value		-30% parameter value	
	Parameter value	Cell death (0.1779)	Parameter value	Cell death (0.1779)
β_1 (-0.0962)	-0.1250	0.1723(-3.17%)	-0.0673	0.1837(+3.28%)
β_2 (-0.2387)	-0.3103	0.1593(-10.4%)	-0.1617	0.1986(+11.6%)
β_3 (-0.0894)	-0.1162	0.1693(-4.85%)	-0.0626	0.1870(+5.09%)
β_4 (-0.4810)	-0.6253	0.1477(-16.9%)	-0.3367	0.2143(+20.4%)
β_5 (0.0092)	-0.0119	0.1782(+0.16%)	0.0064	0.1776(-0.16%)
β_6 (0.2105)	-0.2736	0.1853(+4.18%)	0.1473	0.1708(-4.01%)
β_7 (-0.1078)	-0.1401	0.1713(-3.70%)	-0.0755	0.1847(+3.84%)
β_8 (-1.0432)	-1.3562	0.1396(-21.53%)	-0.7303	0.2267(+27.44%)
β_9 (0.0856)	0.1113	0.1847(+3.81%)	0.0599	0.1714(-3.67%)
β_{10} (0.4249)	0.5523	0.1794(+0.83%)	0.2974	0.1764(-0.82%)
β_{11} (0.3188)	0.4144	0.1995(+12.16%)	0.2231	0.1586(-10.84%)
β_{12} (-0.2500)	-0.3249	0.1931(+8.54%)	-0.1750	0.1639(-7.87%)
β_{13} (-0.0955)	-0.1242	0.1734(-2.52%)	-0.0669	0.1825(+2.59%)
β_{14} (-0.1366)	-0.1775	0.1705(-4.18%)	-0.0956	0.1857(+4.36%)
β_{15} (-0.2648)	-0.3442	0.1582(-11.09%)	-0.1853	0.2001(+12.47%)
β_{16} (-0.0024)	-0.0032	0.1777(-0.10%)	-0.0017	0.1781(+0.10%)
β_{17} (0.3805)	0.4946	0.2089(+17.42%)	0.2663	0.1515(-14.83%)
β_{18} (-0.6204)	-0.8066	0.1470(-17.38%)	-0.4343	0.2153(+21.04%)
β_{19} (0.1586)	0.2062	0.1827(+2.69%)	0.1110	0.1732(-2.62%)
β_{20} (0.3572)	0.4643	0.1968(+10.64%)	0.2500	0.1608(-9.62%)
β_{21} (0.0110)	0.0143	0.1781(+0.12%)	0.0077	0.1777(-0.12%)
β_{22} (-0.1089)	-0.1416	0.1747(-1.82%)	-0.0763	0.1812(+1.86%)
β_{23} (-0.2284)	-0.2969	0.1549(-12.93%)	-0.1599	0.2043(+14.58%)
β_{24} (-0.0775)	-0.1008	0.1696(-4.64%)	-0.0543	0.1866(+4.86%)
β_{25} (-0.2650)	-0.3445	0.1596(-10.31%)	-0.1855	0.1984(+11.50%)
β_{26} (0.1596)	0.2075	0.1855(+4.27%)	0.1117	0.1706(-4.09%)
β_{27} (0.2479)	0.3223	0.2062(+15.93%)	0.1735	0.1535(-13.74%)
β_{28} (0.2144)	0.2788	0.1910(+7.35%)	0.1501	0.1657(-6.85%)
β_{29} (0.4650)	0.6045	0.2054(+15.45%)	0.3255	0.1541(-13.38%)
β_{30} (0.1387)	0.1803	0.1841(+3.51%)	0.0971	0.1719(-3.39%)
β_{31} (0.1070)	0.1391	0.1863(+4.70%)	0.0749	0.1699(-4.48%)
β_{32} (-0.0687)	-0.0892	0.1735(-2.48%)	-0.0481	0.1824(+2.54%)

relationships generated with an ODE-based model.

We selected from the BioModels Database [103] an ODE-based model that was constructed to quantitatively analyze the pathways responsible for controlling extrinsic apoptosis in single cells [64]. The network is composed of 112 nodes. Simulations were conducted using the simulator of CellDesigner with the default setting of the parameters, which spanned 20,000 time points. We then selected data every 200 time points and collected 100 observations. Before fitting the simulation data, the molecular species whose concentrations remained 0 during all the time points of the simulation were removed and as such 58 nodes were retained. Apoptosis was chosen as the dependent variable and all the other 57 nodes were regarded as the independent variables. PLSR was employed to preform the regression. Five-fold cross-validation was carried out, and the parameters in Equation (4.1) and Equation (4.4) were learned based on the training dataset. The prediction was performed on the testing dataset. The above process was repeated for 200 times to generate 1000 outputs.

Both the Spearman and Pearson correlation coefficients between the model predictions and the testing data were calculated to assess the performances of Equation (4.1) and Equation (4.4). The Spearman correlation coefficient benchmarks monotonic relationship and the Pearson correlation coefficient is a measure of the linear correlation between two variables. If both coefficients are high (*e.g.*, higher than 0.9), the model predictions are considered to have high accuracy.

4.2.4 EVALUATION METRICS OF MODEL PERFORMANCE

In addition to simulation studies, we also carried out cross-validation on the 3 real cancer datasets and employed PLSR to estimate the parameters. PLSR is used when the goal is to predict causal relations between independent and dependent variables. It seeks to maximize the correlations between

the principal components of independent variables and those of dependent variables. Therefore, PLSR emphasizes the independent variables that have strong covariance with the dependent variables.

For the nonlinear model, we substituted the cell death data and signaling protein data of the training dataset into Equation (4.6) which is the logarithmic deformation of Equation (4.4), given as

$$\ln(P - \varepsilon) = \beta_0 + \sum_{i=1}^n \beta_i \ln(x_i), \quad (4.6)$$

and employed PLSR to estimate the parameter β_i (where $i = 0 \dots n$, and n is the total number of measured signaling proteins). For the linear model, we directly use the original PLSR to estimate all the parameters from the training dataset. Hence, we were able to make predictions using both the linear and the nonlinear models on the testing dataset.

Besides the Spearman correlation and the Pearson correlation, Kullback-Leibler divergence was employed as the loss function to assess the discrepancy between the predicted and the actual probabilities (*e.g.*, the proportion of dead cells among all the cells in the real data is considered as the cell death score). In information theory, the Kullback-Leibler divergence is used to quantify the difference between two distributions [104]. Suppose $\hat{p}$ is the predicted distribution and p is the actual distribution, the loss function is written as

$$L(p, \hat{p}) = \sum_{i=1}^m p_i \log_2(p_i / \hat{p}_i), \quad (4.7)$$

where m is the total number of predictions. The value of Kullback-Leibler divergence is non-negative. If the distributions of predicted and actual probabilities are perfectly matched, this value will be zero.

4.2.5 IDENTIFICATION OF DRUG EFFECTS

The identification of drug effects plays an important role in biomedical research and pharmaceutical applications. One primary objective is to selectively target a signaling pathway while making the others unaffected. The *in vivo* drug effects on the signaling pathways can be estimated according to the change of the signals downstream the drug target [33]. For example, in Figure. 4.1, if Casp9 is the drug target, the activity of Casp3 will decrease (since the signal transduction from Casp9 to Casp3 is blocked) even though the signal transduction to Casp9 is not blocked.

According to the different treatments, we divided the real dataset [61] into two groups: control group and drugged group. Using the control data, we learned the parameters in Equation (4.4) and Equation (4.8). Equation (4.8) describes the relationship between one signaling protein with all the other measured proteins. The parameters in Equation (4.8) form a matrix M with each element λ_{ij} representing the influence of the j -th signaling protein on the i -th protein [80].

$$x_i = e^{\lambda_{i0}} \cdot \prod_{j=1}^n x_j^{\lambda_{ij}} + \varepsilon_i \quad (4.8)$$

where $\lambda_{ij} = 0$, if $i = j$, otherwise $\lambda_{ij} = 0$ will be learnt from the data.

We iteratively selected the k -th signaling protein as the blocked protein. Every entry of the k -th column of the matrix M is thus set to zero, meaning there is no influence of this protein on the other proteins any more. And β_k in Equation (4.4) is also set to zero to remove the contribution of the k -th protein to cell death. Subsequently, the new parameter β'_j ($j = 1 \dots n$, and if $j = k$, $\beta'_j = 0$) is calculated as $\beta'_j = \sum_{i=1}^n \lambda_{ij} \beta_i$ to modify the influence of the j -th protein to the cell death after the knock-down of the k -th protein. The term $\lambda_{ij} \beta_i$ denotes that the contribution of the j -th protein to cell death is made by influencing the activity of the i -th protein as an intermediary. Then using the signaling data of the drugged group, the predicted proba-

bilities of cell death were calculated and root mean squared error (RMSE) was employed to measure the goodness of fit to the real data. If the fit is good, it can be inferred that the drug blocks the signal transduction from the drug-targeted nodes to the k -th signaling protein. This *in silico* simulation of protein knock-down is thus able to predict drug effects.

4.3 RESULTS

4.3.1 ANALYSIS OF SIMULATION DATA

SINGLE VARIABLE SCENARIO. Table 4.2 shows the curve fitting results when the cell death is one of the following functions of EGFR (or DNA Damage), namely power function, linear function, exponential function and Gaussian function, respectively. We can see from the table that the power function has the best fit to the simulation data. Taking EGFR as an example, the SSE and the RMSE of the power function are closest to 0 and the R-Square and the adjusted R-Square of the power function are closest to 1. This means that using a power function is probably the best way to describe the relationship between cell death and EGFR among the four functions. A similar conclusion can be drawn for Oncogenic Signature, Casp8, Casp9 and Casp3 (data not shown). We then tried many other settings (*e.g.*, different edge weights and input levels) to run SimBoolNet which showed that the power function was also the best fit to the simulation data (data not shown).

MULTIPLE VARIABLES SCENARIO. Then all the 6 nodes (*i.e.*, DNA Damage, EGFR, Oncogenic Signature, Casp8, Casp9 and Casp3) were combined together as independent variables and fit the simulation data into Equation (4.1) and Equation (4.3). The RMSE values are 6.19×10^{-5} and 4.13×10^{-5} when the simulation data were fitted into Equation (4.1) and Equation (4.3), respectively. The Pearson's correlations coefficients between the model predictions and the synthetic data were also calculated. The cor-

relation coefficient between our model predictions and the synthetic data was 0.89, while the coefficient was 0.74 for the linear model prediction. The result shows that Equation (4.3) fits the simulation data better than Equation (4.1).

CELLDESIGNER SIMULATION. From Figure. 4.2, it is clear that the number of outputs with both coefficients higher than a threshold (*e.g.*, 0.9, 0.8 and 0.7) from the nonlinear model is larger than that from the linear model. As the threshold increases from 0.7 to 0.9, the superiority of the nonlinear model over the linear model becomes more obvious. The regression results demonstrate that the present nonlinear model fits the ODE-based system simulation data better than the linear model.

Table 4.2: Curve fitting results of Power function, Linear function, Gaussian function and Exponential function. The statistics are about EGFR and DNA Damage.

Node	Function	SSE	R-Square	Adjusted R-Square	RMSE
EGFR	Power	$1.3 \cdot 10^{-3}$	0.99	0.99	$3.6 \cdot 10^{-3}$
	Linear	$2.1 \cdot 10^{-2}$	0.97	0.97	$1.5 \cdot 10^{-2}$
	Exponential	$8.0 \cdot 10^{-2}$	0.88	0.88	$2.9 \cdot 10^{-2}$
	Gaussian	$1.5 \cdot 10^{-2}$	0.98	0.97	$1.2 \cdot 10^{-2}$
DNA Damage	Power	$1.3 \cdot 10^{-3}$	0.99	0.99	$3.7 \cdot 10^{-3}$
	Linear	$2.0 \cdot 10^{-2}$	0.96	0.96	$1.4 \cdot 10^{-2}$
	Exponential	$7.9 \cdot 10^{-2}$	0.88	0.87	$2.8 \cdot 10^{-2}$
	Gaussian	$1.4 \cdot 10^{-2}$	0.97	0.97	$1.2 \cdot 10^{-2}$

Figure 4.2: Cross-validation of the linear and the nonlinear models on synthetic data. For example, 999 out of the total 1000 predictions made by the nonlinear model have both the Spearman and the Pearson correlations with the testing data higher than 0.7.

4.3.2 RESULTS ON BREAST CANCER CELL LINES

PREDICTION OF CELL FATE. Based on the phosphorylation levels of 32 signaling proteins at a time point, we intended to predict the cell death at the corresponding time point. Both the linear (Equation (4.1)) and our nonlinear (Equation (4.4)) models were employed to relate the cell death to the activities of proteins. Cross-validation was used to estimate the performance of both models on cell death prediction.

First we carried out 5-fold cross-validation on the dataset within the same cell line. Because the number of observations of cell line MCF7 is smaller than the number of parameters, we ignored this cell line to avoid

overfitting. Altogether, we did 100 times of validations and generated 1000 outputs for the linear model and 1000 outputs for the nonlinear model (Figure. 4.3).

Figure. 4.3(a) and Figure. 4.3(b) show that the distribution of the outputs generated by the nonlinear model is more concentrated than the outputs of the linear model in the area where both the Spearman correlation coefficients and the Pearson correlation coefficients are high. Also, it is clear that the number of outputs with both coefficients higher than a threshold (*e.g.*, 0.9, 0.8 and 0.7) from the nonlinear model is larger than that from the linear model (Figure. 4.3(c)). When the threshold is set to 0.7, satisfactory results of the nonlinear model is 96.2%, compared with 65.5% from the linear model. The difference increases from 30.7% to 42.0% when the threshold is 0.8. A remarkable difference (32.4%) is also observed if the threshold is set to 0.9. Moreover, one-way analysis of variance (one-way ANOVA) was used to compare the correlations of model predictions with the real data of cell fate over time points between our model and the linear model. Table 4.3 shows the means and standard deviations of the correlation coefficients, and the p-value of the one-way ANOVA analysis. We observed that the correlations between our model predictions and the real data are significantly better than the linear model, suggesting that the nonlinear model performs better than the liner model of cell fate prediction with statistical significance.

Table 4.3: The mean and standard deviation of the correlation between model predictions and the real data.

	Our model		The linear model		<i>p</i> -value from one-way ANOVA
	Mean	SD	Mean	SD	
Spearman correlations	0.90	0.06	0.74	0.17	2.6E-145
Pearson correlations	0.89	0.08	0.82	0.13	2.3E-41

Among 1000 outputs, both the linear and the nonlinear models make 13,100 predictions. Using Equation (4.7), we calculated Kullback-Leibler

Figure 4.3: Performance comparison of cell death prediction within each cell line. More outputs have high coefficients in the nonlinear model (a) than in the linear model (b). (c) Comparison between the two models by counting how many outputs have both coefficients higher than a threshold (e.g., 0.7, 0.8 and 0.9).

divergence for the predictions of both models. 67.2 is obtained by the nonlinear model which is smaller than 277.7 by the linear model. This further

demonstrates that the probabilities predicted by our nonlinear model are more accurate.

We then combined 164 observations from all the 3 cell lines as a whole dataset to do 5-fold cross-validation. This experiment was designed for the situation when the data have no cell line information. Cross-validation was run for 200 times and 1000 outputs were generated.

Figure. 4.4 displays the performance comparison between the linear and the nonlinear models. It is clear that for all the 3 thresholds, the proposed nonlinear model has superiority over the linear model on cell death prediction. A consistent conclusion can be drawn from the Kullback-Leibler divergence, since for the 32,800 predictions of the linear and the nonlinear models the values are 1,117.5 and 450.5, respectively. Moreover, Akaike information criterion (AIC) was calculated based on Equation (4.9) where $L(p, \hat{p})$ is the loss function from Equation (4.7), k is the number of parameters and m is the number of predictions. Since the numbers of parameters in the linear and the power-law models are the same (*i.e.*, 33), the values of AIC for the linear model and nonlinear model are 10.1 and 8.8, respectively, demonstrating that it has a better fit than the linear model has a better fit.

$$AIC = \ln(L(p, \hat{p})) + \frac{2k}{m}, \quad (4.9)$$

Then, we tested both the linear and our nonlinear models by training and predicting in alternative cell lines to see which model can perform better. We chose the cell lines of BT20 and MDA-MB-453 but ignored MCF7 because the number of observations in MCF7 is smaller than the number of parameters. Table 4.4 shows the result of the prediction across cell lines. With regard to the absolute value of the correlation coefficient, the closer it is to 1, the better is the correlation between the predicted and real data. When BT20 is used for training and MDA-MB-453 is used for testing, the nonlinear predictions are more monotonically correlated with the testing data and have less loss of information than the linear predictions, but the linear predictions have better linear correlation with the real data. Alternatively,

Figure 4.4: Comparison of cell death prediction by the linear and the non-linear models when combining all three cell lines together.

when MDA-MB-453 is employed as the training dataset, the proposed non-linear model outperforms the linear model on both correlations. Also, the Kullback-Leibler divergence indicates that the nonlinear model can capture more information.

Overall, when predicting cell death probabilities from the activities of signaling proteins, our nonlinear model has some superiority over the linear model.

DRUG EFFECTS IDENTIFICATION BY *IN SILICO* SIMULATION OF PROTEIN KNOCK-DOWN. Suppose node u is the upstream signaling protein of node v and node u is a drug target. As such, the drug will block the signal flow from u to v . The signals transmitted to v are decreased while those transmitted to u should remain unchanged.

The interactions between v and other proteins are thus weakened. In the data-driven methods, the correlations between v and other nodes in the network (including other proteins and the cell death) are reduced. However, this does not necessarily happen to u . Hence, in the dataset of drug treatment, the contributions of v to other proteins should be less significant than in the control data, and the impact of the knock-down of v on cell responses should be relatively small. Therefore, by setting both β_v in Equation (4.4) and λ_{iv} in Equation (4.8) to zero (*i.e.*, remove the influences of v on the other molecules), the predicted probabilities of cell death should have smaller discrepancies (*e.g.*, smaller RMSE) with the real data in the drugged group.

For each of the 32 signaling proteins [61], we simulated the knock-down and ranked them according to the RMSE. STAT3, p27, p53, ERK and HSP27 are the top 5 proteins with the smallest RMSE. Therefore, we assumed that the drug target should be the common upstream node of these top-ranked proteins. Next we extracted pathway information from the GeneGO MetaCore database [81]. Figure. 4.5 displays the pathways that contain the top-ranked proteins as well as some related proteins. We can see that EGFR is the common upstream node of all the 5 proteins thus it is inferred as the drug target. A red edge denotes that the reaction should be removed from the pathway, since the signal transduction from EGFR to ERK is blocked. In [61], the drug (erlotinib) used is indeed known as an EGFR inhibitor. Erlotinib is a small molecule that is able to block the signal transduction

Table 4.4: Performance of the linear and the nonlinear models in predicting cell fate across cell lines.

Train	Test	Model	Spearman	Pearson	Kullback
BT20	MDA	nonlinear	0.64	0.53	0.19
		linear	0.60	0.68	0.23
MDA	BT20	nonlinear	0.70	0.57	0.30
		linear	-0.36	-0.25	1.0

downstream of EGFR, such as the PI3K-AKT or the RAS-RAF-MAPK-MEK-ERK pathway [105]. In our experiment, however, AKT was ranked the fifth from the bottom, meaning the signals transmitted to AKT were not blocked, which suggests that the PI3K-AKT pathway in Figure 4.5 cannot be removed although the drug targets EGFR. This interpretation was also consistent with that in [33], the authors of which identified erlotinib-induced pathway alterations using Integer Linear Programming and predicted that the RAS-RAF-MAPK-MEK-ERK cascade was removed while the PI3K-AKT network was retained. This is because the PI3K-AKT pathway was used by other pathways which cannot be blocked by erlotinib. Therefore, knowing that the drug is well designed to hit certain molecules is not sufficient for identifying the drug effects. The experiment shows that the proposed model is able to identify not only the drug target but also the drug effects by rewiring signaling pathways *in silico*, which shows a great potential to complement the analysis based on the drug's biochemical activity (*e.g.*, binding affinities). In addition, the *in silico* simulation of protein knock-down was executed using the linear model to test if the linear model is able to identify the drug targets and drug effects. Following the same process aforementioned, the top 5 proteins with the relatively small RMSE between the linear model predictions and the real data were Caspase 9, BID, RIP, JNK and S6. However, based on GeneGO MetaCore database [81], Caspase 9, BID, JNK and S6 are all downstream of PI3K-AKT pathway which should not be blocked by an EGFR inhibitor since PI3K-AKT pathway is used by other pathways (*e.g.*, TNFR-PI3K-AKT) [33]. There was no evidence of RIP being downstream of EGFR [81]. Therefore, the linear model showed limited capability of identifying drug targets and drug effects.

IDENTIFYING TIME-STAGGERED INPUT-OUTPUT RELATIONSHIP. The process of phosphorylation of signaling proteins and the transduction of the signals to the downstream pathways are accomplished within minutes. On the other hand, it may take hours for the cells to adjust its phenotypes in response to the input signals. Therefore, the identification

Figure 4.5: Pathway information extracted from GeneGO MetaCore. The dash line denotes the predicted drug effects, *i.e.*, signal transduction is blocked here.

of time-staggered input-output relationships is very important. We extracted 10 time points (*i.e.*, 0, 0.1, 0.25, 0.5, 1, 2, 4, 6, 8 and 12 hours) from the dataset in [61]. The phosphoproteomics data cover time points from 0 to 7 and the cell fate data are measured at time points 0 and from 6 to 9. So we designed experiments (Table 4.5) to see what time-staggered degree can reveal the most of the input-output relationship (Figure. 4.6). Take the time-staggered degree equal to 2 as an example, we assumed that the “output” at time point t is the response to the “input” at time point “ $t - 2$ ”. The “output” at time point 0 will be related to the “input” also at time point 0.

We constructed four datasets corresponding to the four time-staggered degrees (Table 4.5) to predict the cell death. Figure. 4.6 shows the numbers of outputs that have relatively high accuracy. The bars filled with slash and horizontal line indicate the numbers of validations which have both Spearman

and Pearson correlation coefficients higher than 0.8 and 0.9, respectively. The results generated by the proposed nonlinear model show that it is most reliable that the “output” at time point t is the response to the “input” at time point “ $t - 3$ ”. The same conclusion can be drawn for the linear model.

Table 4.5: Experimental design for identification of time-staggered input-output relationships.

time-staggered degree		Time Point								
	Input	0	1	2	3	4	5	6	7	
2	Output	0				6	7	8	9	
3	Output	0			6	7	8	9		
4	Output	0		6	7	8	9			
5	Output	0	6	7	8	9				

CELL LINE DISCRIMINATION. To model a biological system statistically, it is important to capture the common features while keeping the specific characteristics between different cell lines. Therefore, we designed an experiment to verify the performance of our nonlinear model in cell line discrimination. Signaling and apoptosis data from all the three cell lines [61] were combined together. Principal components were extracted using PLSR (for the nonlinear model, PLSR was used after log transformation) and observations of signaling proteins were projected against the first two principal components (Figure. 4.7(a)). Blue asterisks, red plus signs and black circles represent observations from the BT20, MCF7 and MDA-MB-453 cell lines, respectively. The same process was done using the linear model and the result is shown in Figure. 4.7(b). It is clear that the observations are highly cell line dependent. BT20 (red cluster) and MCF7 (green cluster) have similar features on the first principal component while MDA-MB-453 has a significantly different behavior. On the other hand, the second principal component captures variance that can be used to distinguish BT20 and

Figure 4.6: Performance of the linear and the nonlinear models at different time-staggered degrees.

MDA-MB-453 from MCF7. Based on the linear model, a significant overlap can be found between the MCF7 and MDA-MB-453 clusters and the BT20 cluster is slightly scattered. By contrast, the 3 clusters generated by our nonlinear model are more separated and concentrated. Specifically, over ten data points are misclassified (*e.g.*, an observation from MCF7 is classified into the BT20 cluster) by the linear model while there is only one misclassified data point by the nonlinear model.

4.3.3 RESULTS ON HT-29 CELL LINE

In the above experiments, the activity level of each signaling protein is considered as an independent variable and the dimension of the independent variables is the number of signaling proteins. For HT-29 cell line data, we con-

Figure 4.7: Principal component analysis for cell line discrimination. (a) The nonlinear model. (b) The linear model.

structured a relatively high dimensional space for independent variables [63, 64] besides the dimensions of the signaling proteins. In particular, the independent variables comprise the 19 signals at all 13 time points (247 independent variables), the instantaneous-derivative between each pair of adjacent time points for all the 19 proteins (247 independent variables), the maximum

signal, the mean signal and the steady-state signal for each protein, etc. All together, a 570-dimensional space was constructed and the top 20 most informative dimensions identified in [64] were all included. For dependent variables, a 12-dimensional space was extracted (4 types of cellular responses at 3 time points) and all the data were normalized into the interval $(0, 1)$ using the sigmoid function.

We then compared the performance of the model proposed in [63] and our proposed model in the prediction of cell death using a leave-one-out cross-validation. RMSE between the model predictions and the testing data are 2.11 and 1.77 for the model of Gaudet *et al.* [63] and our model, respectively. Our model achieves a lower RMSE, indicating that it could better predict the cell apoptotic responses.

We also designed experiments following the same procedure as in Section “Drug effects identification by *in silico* simulation of protein knock-down” to identify the effect of the perturbations on the pathways. Caspase-3, which plays a crucial role in the execution-phase of cell apoptosis, was selected as the dependent variable. All the remaining 18 proteins were treated as independent variables. The control group of the dataset in [63] was used for training the parameters in Equation (4.4) and Equation (4.8). The observations treated by one of the 3 stimuli (*i.e.*, TNFR, EGFR and Insulin) were first extracted as the perturbed group. The *in silico* protein knock-down was simulated and the proteins which induce relative large discrepancy with the real data after knock-down are identified as the enhanced signals by the stimuli. The typical targets of the stimuli are all ranked very high as expected, such as JNK and IKK (ranked 3rd and 4th) for TNFR; EGFR and AKT (1st and 3rd) for EGFR; ERK, MEK and AKT (1st, 4th and 5th) for Insulin [63]. However, when we looked into the combination of TNFR and EGFR, the signal flows of TNFR-JNK and TNFR-IKK were significantly weakened (11th and 18th), and the signal of AKT was enhanced (2nd). On the other hand, significant reduction of the AKT signal (ranked 16th) was detected under the the combination of TNFR and Insulin. This suggests that the synergistic effect of TNFR and EGFR could enhance the signal flow PI3K-AKT,

while the cascades of TNFR-PI3K-AKT and Insulin-IRS1-PI3K-AKT could conflict with each other when both TNFR and Insulin are present, as shown in Figure. 4.8, where the measured signals are highlighted in pink.

Figure 4.8: Schematic representation of the signaling network induced by TNF, EGF, and insulin.

4.3.4 RESULTS ON DREAM8 DATASET

Since no cell response were directly measured in the dataset, we iteratively selected one signaling protein as the dependent variable and the remaining signaling proteins were regarded as independent variables.

Within each cell line, the dataset was divided into two subsets, *i.e.*, two thirds were used for training and one third for testing. On the training part, leave-one-out cross-validation was employed to learn the parameters in Equation (4.4). With the mean values of the parameters from the cross-validation, we were able to make predictions on the testing data. We repeated the above procedure for 100 times. Table 4.6 shows the RMSE which estimates the accuracy of the model predictions for each cell line. It is clear that our nonlinear model predicts the biological system with a higher accuracy than the linear model.

Table 4.6: Performance of the nonlinear and linear model on DREAM8 dataset. The values are the RMSE between the model predictions and testing data of activity of signaling proteins.

Cell lines	Nonlinear model	Linear model
BT20	0.41	0.71
BT549	0.46	0.69
MCF7	0.33	0.39
UACC812	0.57	1.46

4.4 CHAPTER SUMMARY

Data-driven models are able to provide new biological insights by analysing a dataset itself. They are particularly useful when the underlying molecular mechanisms are unclear. In this chapter, we proposed a nonlinear power-law model to describe the relationship between cell fates and cell signals. Simulations based on an extended model of Boolean network and ordinary differential equations (ODEs) provided hints about the form of the nonlinear function as well as how to validate the proposed model. By predicting the cell responses, we compared the performance of our nonlinear model with the linear model on 3 real datasets, demonstrating that the proposed nonlinear model has higher accuracy than the linear model. Given network topology, the proposed model performs well on drug target identification and is able to reveal the drug effects by rewiring the signaling pathways *in silico*. Then our nonlinear model was used for time-staggered input-output relationships identification and cell line discrimination.

In spite of the promising performance of our proposed model, limitations have also been noticed. First, as a data-driven method, it is unable to incorporate prior biological knowledge to take into account the underlying mechanisms. Second, our model has not been tested for the situation when more than one signaling proteins are inhibited or the corresponding genes

are knocked out at the same time. Moreover, when we compared the performance of the linear and the nonlinear models in predicting cell fate across cell lines, the superiority of the proposed model may not significant enough. In future, we will include network topology to improve the model so that it can deal with synergistic effect of multiple perturbations.

Chapter 5

PREDICTING SYNTHETIC LETHALITY VIA MODELING OF CELLULAR SIGNALING PATHWAYS

5.1 INTRODUCTION

A major challenge to establish treatments for human cancer is how to kill cancer cells specifically and spare healthy cells. Among the developed anti-cancer therapies, the lack of drug selectivity is likely to lead to the elimination of both the tumour and the normal cells, thus causing various side-effects. To improve the classic cytotoxic therapies, a novel anti-cancer strategy named synthetic lethality (SL) [106] has been introduced and has shown great potential to treat cancers. Two genes have SL relationship when the combination of mutations is able to kill a cell, whereas the mutation by only one gene can not. Since a lot more genetic mutations exist in tumour cells than in normal cells, the identification of the SL pair of genes could play an important role in pharmaceutical industry when one of them is a cancer-specific mutated

gene. The drug that targets the partner gene is thus able to give rise to SL and kill the tumour cells selectively.

However, the majority of the established SL identification techniques, such as RNA interference (RNAi) [107], are faced with multiple issues like the design of an effective small interfering RNA (siRNA) sequence and the stability across different platforms or cancer subtypes. Moreover, while the underlying mechanisms that lead to SL are crucial for developing reliable anti-cancer therapies, they can hardly be reflected by the screening-based methods. Wu *et al* [108] proposed a meta-analysis data-mining method that was able to predict SL based on genomic and proteomic features, Güell *et al* [109] also proposed a computational method to screen synthetic lethal reaction pairs from the perspective of metabolism, however, their results mainly focused on yeast. Another computational method was developed by Heiskanen and Aittokallio [110], which infers human SL from yeast SL based on the conserved features between the two species. However, key features that are specific to human were likely to be lost.

In this chapter, we propose a computational method to predict SL pairs of genes by combining a data-driven model [54, 60], which is discussed in Chapter 4, with prior knowledge of the signaling pathways. The data-driven model is used to construct a function that describes the relationship between the signaling proteins and the cell death. Then the single gene knock-down and the double genes knock-down are simulated by integrating the pathway information. A predicted probability of cell death is calculated after gene knock-down. A pair of genes is considered as an SL candidate when the predicted probability of cell death after double genes knock-down is highly increased compared with real data, while the predictions after the knock-down of either single gene are not. We use human essential genes to validate the simulation of single gene knock-down. The Syn-Lethality database [111], which consists of 113 SL pairs manually collected from literatures of wet-lab experiments, is employed to evaluate the SL predictions of our model. The two SL pairs of genes that coexist in the Syn-Lethality database and in the employed dataset are successfully identified. Explanations of the mechanisms

of the novel predicted SL pairs of genes are also presented.

5.2 DATA AND METHODS

5.2.1 DATA

We downloaded the time-series signaling data from the work of Lee *et al.* [61]. We extracted the phosphorylation levels of 32 signaling proteins at 5 time points, and the measured cell fates (*e.g.*, apoptosis, proliferation) using flow cytometry at the corresponding time points. The signaling data contain 6 treatments (*i.e.*, 6 groups), which were designed on breast cancer cell line BT20. We chose the control group data since the other five groups had been treated by different drugs which may introduce unpredictable biases to our model. For each signaling protein, the values of its phosphorylation levels from all the biological experiments were normalized into (0, 1). Suppose max and min represent the maximum and minimum measurements of a protein, respectively, then for every biological measurement (denoted as mea) of this protein, the normalized value is $(mea - min * 0.99) / (max * 1.01 - min * 0.99)$. Finally, we approximately estimated the probability of cell death as the proportion of dead cells under each treatment.

5.2.2 SIGNALING PATHWAYS

Pathway information was extracted from GeneGO MetaCore [81]. Classic signaling pathways about the regulation of cell fates (*e.g.*, apoptosis and proliferation) were extracted and combined into a network as shown in Figure 5.1, where signaling proteins with and without experimental measurements are represented as pink rectangular and blue circle, respectively, and cell fates are represented as yellow diamond. Activation and inhibition interactions are denoted as green and red edges, respectively. This generic pathway only contains the measured proteins in the dataset [61] and their

Figure 5.1: Pathway information extracted from GeneGO MetaCore.

direct interaction neighbors. The network has 59 signaling proteins and 3 cell fates. Among the 59 proteins, 28 had measurements in the dataset we downloaded [61].

Algorithm 1. Learning parameters of Equation 5.1

INPUT: Time-series signaling protein data $X_{(t \times n)}$ and cell death data $P_{t \times 1}$, where t is the number of measurements and $n = 28$.

OUTPUT: The vector $A : (\alpha_0, \alpha_1, \dots, \alpha_n)$ with $n + 1$ elements.

//Take the logarithm of Equation (1).

$$\ln(P - \varepsilon) = \alpha_0 + \sum_{i=1}^n \alpha_i \cdot \ln(x_i);$$

Independent variables $\leftarrow$ signaling proteins;

Dependent variables $\leftarrow$ cell death;

Use partial least squares regression (PLSR) method (*e.g.*, the *plsregress* function in MATLAB) to do the regression and learn the parameters $(\alpha_0, \alpha_1, \dots, \alpha_n)$;

Return The vector A .

5.2.3 DATA-DRIVEN POWER-LAW MODEL

We first constructed a mathematical model according to the Equation (5.1) [60], as discussed in Chapter 4, to relate the cell signals to the cell fates. At each time point, P is the score indicate probability of cell death, x_i (where $i = 1, \dots, 28$) represents the activity (*e.g.*, phosphorylation level) of the i -th signaling protein, α_i indicates the influence of the corresponding protein to the cell death (as such, it's not propagation through the network, but rather direct regression to the target node, *i.e.*, apoptosis) and ε is a small constant representing random error.

$$P = e^{\alpha_0} \cdot \prod_{i=1}^n x_i^{\alpha_i} + \varepsilon \quad (5.1)$$

Next, we built the functions in Equation (5.2) [60] to describe the relationship between a signaling protein and all its measured upstream nodes in Figure 5.1. For the i -th protein, we selected it as the dependent variable in Equation (5.2) and all its measured upstream nodes as independent variables, with λ_{ij} representing the contribution of the j -th measured upstream node to the node i and all the λ parameters forming a matrix M . To identify the upstream nodes of the node i , we considered not only its direct parent nodes, but also the measured proteins that had at least one path leading to the node i . For example, from Figure 5.1, we can see that SMAC has two direct parent nodes (JNK and BAX, but only JNK has measurements), and since there is a path BIM-BAX-SMAC and BIM has measurements, both JNK and BIM will be considered as the measured upstream nodes of SMAC.

$$x_i = e^{\lambda_{i0}} \cdot \prod_j x_j^{\lambda_{ij}} + \varepsilon_i \quad (5.2)$$

Then the parameters in Equation (5.1) and (5.2) (*i.e.*, α_i and λ_{ij}) were learnt using partial least squares regression (PLSR) as discussed in Chapter 4 [60] (Algorithm 1 and 2).

Algorithm 2. Learning parameters of Equation 5.2

INPUT: Time-series signaling protein data $X_{(t \times n)}$; adjacency matrix $AD_{(N \times N)}$, where $N = 59$ is the number of nodes in the pathways and $N_{ij} = 1$ if there is an edge leading from node N_j to N_i (*i.e.*, N_j is a direct parent node of N_i).

OUTPUT: The matrix $M_{(n \times n)}$ where each element m_{ij} is the parameter λ_{ij} .

//Extract measured upstream nodes.

for each signaling protein i **do**

 //Extract direct parent nodes of protein i .

$rowAD(i) \leftarrow i$ -th row of matrix AD ;

for each parent node j ($j \neq i$) in $rowAD(i)$ **do**

if the j -th parent node has measurements and $m_{ij} = 0$ **do**

$m_{ij} \leftarrow 1$;

elseif the j -th parent node has no measurements and $m_{ij} = 0$

do

 //Extract direct parent nodes of protein j .

$rowAD(j) \leftarrow j$ -th row of matrix AD ;

$rowAD(i) \leftarrow \text{union}(rowAD(i), rowAD(j))$;

end for when the i -th row of M is conserved or reach the maximum

 steps of loops;

end for

//Calculate λ_{ij} .

for each signaling protein i **do**

 //Extract measured upstream nodes of protein i .

$rowM(i) \leftarrow i$ -th row of matrix M ;

 //Take the logarithm of Equation (2)

$\ln(x_i - \varepsilon_i) = \lambda_{i0} + \sum_j \lambda_{ij} \cdot \ln(x_j)$;

 Independent variables $\leftarrow$ proteins with nonzero values in $rowM(i)$;

 Dependent variables $\leftarrow$ protein i ;

 Use PLSR to learn the parameters λ_{ij} as in Algorithm 1;

$m_{ij} \leftarrow \lambda_{ij}$;

end for

Return The matrix M .

5.2.4 GENE KNOCK-DOWN SIMULATION

Suppose the signaling protein u is upstream of protein v , and u is knocked down (*e.g.*, mutated or blocked). The signals received by v are thus changed under the assumption that all the other upstream nodes of v remain unchanged. For example, v will be down-regulated or up-regulated after the removal of u when u activates or inhibits v , respectively. After u is knocked down, we next estimate the changes of its downstream nodes and the cell death, based on our data-driven models in Equations (5.1) and (5.2).

For the knock-down of a single gene (gene and protein are used interchangeably in this chapter), we removed not only its contribution to the cell death, but also its influence on all its downstream nodes. For example, if the j -th node in the pathway was knocked down, the effect of its knock-down on cell death was estimated as follows. First, we removed the influence of the j -th protein to all its measured downstream nodes by setting the j -th column in Matrix M to be 0 ($\lambda_{ij} = 0$, for all i) and recalculated the activity levels of its downstream nodes. Second, we set $\alpha_j = 0$ to remove the contribution of the j -th protein to cell death. Then we estimated the cell death by substituting the updated protein data and vector $(\alpha_0, \alpha_1, \dots, \alpha_n)$ into Equation (5.1). Forth, we employed the KullbackLeibler divergence (the relative entropy, given in Equation (5.3)) to estimate how much the cell death predictions differ from experimental measurements, where t is the number of predictions (or measurements). Algorithm 3 gives the pseudocode of the process. We defined the divergence as the “single knock-down score” of the j -th protein. If the single knock-down score was equal to zero, there was no increment of cell death due to knock-down event. Therefore, the bigger the single knock-down score is, the greater the cell death is promoted.

$$D_{KL}(Predicted||Measured) = \sum_t Predicted(t) \ln\left(\frac{Predicted(t)}{Measured(t)}\right) \quad (5.3)$$

Similarly, the effect of double genes knock-down on cell death can be

inferred. For each protein, we extracted the set of its downstream nodes. Given a pair of proteins to be knocked down, we first computed the union set of their downstream nodes and recalculated the activity levels for the nodes in this union set by modifying the parameters in Equation (5.2) to be 0. Then,

Algorithm 3. Single gene knock-down simulation

INPUT: Time-series signaling protein data $X_{(t \times n)}$ and cell death data $P_{t \times 1}$; the vector A and the matrix M .

OUTPUT: Single knock-down score for each gene.

for each signaling protein j **do**

$X' \leftarrow X$;

$M' \leftarrow M$;

//Extract measured downstream nodes of protein j .

$colM(j) \leftarrow$ the j -th column of matrix M ;

for each downstream node i in $colM(j)$ **do**

$m'_{ij} \leftarrow 0$;

//Extract measured upstream nodes of protein i .

$rowM'(i) \leftarrow$ the i -th row of the matrix M' ;

Recalculate the activity levels of protein i by substituting the matrix X

and $rowM'(i)$ into Equation (2);

Update the i -th column of X' using the recalculate activity levels of

protein i ;

end for

$A' \leftarrow A$;

$\alpha'_j \leftarrow 0$;

Recalculate the probability of cell death by substituting the matrix X' and

the vector A' into Equation (1);

Use Equation (3) to calculate the Kullback Leibler divergence between the

predicted and measured cell death as the knock-down score of protein j ;

end for

Return List of single knock-down scores for all genes.

we estimated the cell death by feeding the updated activity levels and vector $(\alpha_0, \alpha_1, \dots, \alpha_n)$ to Equation (5.1), and similarly used Equation (5.3) to define the “double knock-down score” to measure the effect of the knock-down of a pair of genes.

5.2.5 SYNTHETIC LETHALITY PREDICTION

For a protein pair (u, v) , let $d(u, v)$ denotes its double knock-down score, $s(u)$ and $s(v)$ are the single knock-down scores for u and v , respectively. Our assumption was that a protein pair (u, v) with large $d(u, v)$ and low $s(u)$ and $s(v)$ tended to be a synthetic lethality pair. Hence, we defined the SL score for the protein pair (u, v) as in Equation (5.4), to quantify their mutual dependence on knock-down events. For example, if u and v are independent, then knocking down u does not give any influence on knock-down of v and vice versa, so their SL score is zero. The log-ratio term $(\ln(\frac{d(u,v)}{s(u) \times s(v)}))$ helps to identify the pairs whose double knock-down score increases significantly compared with single knock-down scores. However, large log-ratio is insufficient for the discovery of an SL candidate since an SL candidate should have a large double knock-down score in the first place. Therefore, by timing the double knock-down score($d(u, v)$), the pairs with large log-ratio but small double knock-down score (such that the SL score is small) are filtered out to reduce the false positive. The larger the SL score is, the more likely the pair is considered as SL candidate.

$$SL(u, v) = d(u, v) \ln\left(\frac{d(u, v)}{s(u) \times s(v)}\right) \quad (5.4)$$

5.3 RESULTS

5.3.1 SINGLE KNOCK-DOWN AND ESSENTIAL GENES PREDICTION

To evaluate the performance of single gene knock-down simulation, 28 measured signaling proteins were ranked according to their single knock-down scores. Out of these 28 proteins, 15 proteins had single knock-down scores larger than 0. Table 5.1 shows the top-5 proteins (*i.e.*, AKT, p53, CHK1, S6K and CYCLIND1), whose removal brings the highest impact on the cell death. For example, AKT was in the top of the list since knock-down of AKT gave rise to the biggest increase of the cell death according to Equation (5.3).

Table 5.1: Rank of the single knock-down simulation.

Protein	Gene	Single Knock-down Score	Rank
AKT	ENSG00000142208	0.037	1
p53	ENSG00000141510	0.031	2
CHK1	ENSG00000149554	0.021	3
S6K	ENSG00000108443	0.016	4
CYCLIND1	ENSG00000110092	0.015	5

AKT, which is also known as Protein kinase B (PKB), plays an essential role in many regulation processes including proliferation, apoptosis, cell cycle and metabolism. From Figure 5.1, it can be seen that the activation of mTOR by AKT contributes to cell proliferation. AKT also regulates cell apoptosis via several pathways such as AKT-Caps9, and controls cell cycle by phosphorylating its substrates including GSK3 [112]. The tumor suppressor p53 is able to induce cell cycle arrest, initiate DNA repair process when DNA is damaged and activate cell apoptosis if DNA is irreparable. Therefore, it is crucial in multiple cellular mechanisms [113, 114, 115]. CHK1, which is also referred to as checkpoint kinase 1, responds to checkpoint-mediated cell cycle arrest and DNA damage [116]. S6K (Ribosomal protein S6 kinase beta-1), which responds to mTOR and growth factors, is responsible for regulating

protein synthesis, cell growth, and cell proliferation [117]. CYCLIND1, acting as the regulator of cyclin-dependent kinase (CDK), belongs to a family whose members have a significant periodicity in their abundance over cell cycle [118]. Moreover, all these five genes were identified as human essential genes in Online GENE Essentiality database (OGEE) [119].

5.3.2 DOUBLE KNOCK-DOWN AND SYNTHETIC LETHALITY PREDICTION

We performed double knock-down simulation over all the 378 protein pairs (*i.e.*, $\binom{28}{2}$) as we had 28 measured signaling proteins). 252 of them had double knock-down scores larger than 0, indicating that these double knock-down events tended to increase the cell death.

SL scores were calculated based on Equation (5.4). Table 5.2 gives the top 20 pairs of SL candidates after ranking based on the SL scores. For example, the pair of AKT and BID was ranked the first since it had the highest SL score based on Equation (5.4). A recent work has investigated the SL interactions in the signaling pathways related to DNA damage and checkpoint control [120]. Two pairs, $(p53, CHK1)$ and $(p53, Wee1)$ were reported as SL pairs in their work, and these two pairs were ranked the seventh and the eleventh by our method. As reported by Kaelin [107] and Le Meur [121], three types of mechanisms were proposed, from the perspectives of signaling pathways, that could be the candidate explanations of SL interactions in human cancers. First, two proteins that coexist in a linear essential pathway are likely to have SL interaction, such as (AKT, BID) in the cascade AKT-BID-BAX-Casp9 and (AKT, BIM) in the pathway AKT-FOXO3A-BIM-BAX-Casp9 (Figure 5.1). The mutation of the first protein (*e.g.*, AKT) would decrease the signals transmitted in the pathway whereas the mutation of both destroys the pathway. Second, two proteins from two parallel essential pathways could be SL partners when they are backups of each other. For example, $(p53, Cabl)$ is predicted since p53 and Cabl are in two parallel essential pathways, *i.e.*, ATM-CHK1-p53-PUMA-BAX and ATM-Cabl-BAX,

respectively. Another example is $(p53, Wee1)$, where p53 is in the pathway BRCA1-p53-CDK1 and Wee1 is in the pathway BRCA1-Wee1-CDK1. Third, two components have the same essential function or they both contribute to the construction of an essential protein complexes. Moreover, according to the Syn-Lethality database [111], both the two SL pairs of proteins that have measurements in the dataset [61], *i.e.*, $(p53, CHK1)$ and $(p53, Wee1)$ have been successfully predicted by our method.

Table 5.2: Rank of the SL candidates.

Protein Pairs		SL Score	Rank
AKT	BID	2.053	1
AKT	Cabl	2.029	2
AKT	BIM	2.021	3
AKT	CYCLIN	1.983	4
AKT	Casp9	1.978	5
p53	BID	1.976	6
p53	CHK1	1.971	7
AKT	Wee1	1.969	8
p53	BIM	1.954	9
p53	Cabl	1.933	10
p53	Wee1	1.914	11
p53	Casp8	1.907	12
AKT	Casp6	1.902	13
AKT	Casp8	1.899	14
p53	S6K	1.897	15
p53	Casp9	1.889	16
p53	CYCLIN	1.874	17
AKT	S6K	1.860	18
p53	Casp6	1.841	19
CHK1	CYCLIN	1.807	20

5.4 CHAPTER SUMMARY

Anti-cancer therapy based on synthetic lethality strategy shows great potential for its capability of treating the cancer and the normal cells differently. Finding SL pairs of genes provides great opportunities for drug target identification if one of the genes in the pair has been mutated in the cancer cells. However, due to the high cost of the screening technology, highly reliable results generated from wet-lab experiments are growing slowly. Therefore, we proposed a computational method which hybridizes a data-driven model [60] with knowledge of signaling pathways to predict potential SL pairs of genes. We first related the activities of signaling proteins to the probability of cell death using a mathematical function. Second, we identified the relationship between each protein and all its measured upstream nodes based on both the biological experimental data and the signaling pathways. Next, we simulated the single knock-down of each measured protein and the double knock-down of all possible pairs of proteins, to estimate the significance of knock-down events to cell death. We then defined the SL score to do virtual screening for the candidates of SL pairs of genes. The single knock-down simulations were validated according to the human essential genes. And the double knock-down simulations gave both wet-lab confirmed results and novel predictions which are suggested to be reliable by evidence from literatures.

In spite of promising performance of our proposed method, limitations have been noticed which point to our future work. First, due to the lack of information, underlying mechanisms are insufficiently considered. For example, the interaction types (*e.g.*, phosphorylation and transcription regulation) in Figure 5.1 are unclear. Also the interaction between two proteins could be activation in one cell type while inhibition under other circumstance due to different genetic contexts. Second, some well established network models such as Boolean network and ODE based model should be involved to gain better performance and mechanistic understanding. And we believe that by employing these models, our method will be improved.

Chapter 6

SUMMARY AND FUTURE WORK

6.1 SUMMARY

The research in this thesis consists of knowledge-based, data-driven and hybrid modeling of signaling pathways for the prediction of cellular signal flows and cell fates. The proposed methods and obtained results can be summarized as follows.

1. Chapter 3 proposes a generalized logical model based on network topology to capture the dynamical trends of cellular signaling pathways. The model is applied to simulating the dynamics of the signaling pathways under different perturbations. Given the inputs, the activity level of each node in the signaling network is updated synchronously based on their own previous states and the incoming signals from the upstream nodes. Compared with the existing simulators, our model is able to simulate the graded responses to degradations and the effects of scheduled perturbations to the cells. However, our model is still limited to simulating the signaling pathways only, thus not applicable to integrating the transcriptional regulatory networks and the control of the

cell fates. Moreover, the synchronous update scheme is also an oversimplification and the asynchronous scheme, which is more realistic in dynamics of the signaling pathways, should be considered in future.

2. Chapter 4 presents a nonlinear power-law modeling of cancer cell fates driven by signaling data to reveal drug effects. The statistical relations between the activities of signaling proteins and the probabilities of cell fates are detected via a data-driven modeling of signaling transduction. Testing on 3 different cancer datasets, the results suggest that our nonlinear model has some superiority over the linear model on cell fates prediction. Besides the prediction, the proposed model is also able to identify the signal flows that are blocked by the drugs, thereby revealing the drug-induced signaling pathway alterations. Moreover, when our nonlinear model is applied to cell line discrimination, the cell lines are much better separated and more concentrated compared with the results of the linear model. In spite of promising performance of the proposed nonlinear model, data-driven modeling has its limitation on revealing the biological insights into the underlying mechanisms. For example, the effectiveness of our nonlinear model on the cell line discrimination task suggests that the power-law model can capture some essential causal relations between the signaling pathways for cell fates decision and the identity of a cell line. However, why this relation can be better described by the power-law than a linear model remains to be elucidated.
3. Chapter 5 introduces a hybrid model, which combines the knowledge-based and the data-driven methods, to predict essential genes and Synthetic Lethality via influence propagation in signaling pathways of cancer cell fates. The prior knowledge of the signaling pathways were applied for guiding the construction of the data-driven model which links the signaling pathways to the cell fates. By simulating the single knock-down of each measured protein and the double knock-down of all possible pairs of proteins, we estimated the significance of knock-down events to cell death. The single knock-down simulations were verified

by the human essential genes. And the double knock-down simulations give both wet-lab confirmed results and novel predictions which are consistent with evidence from the literature. However, how to balance the prior knowledge and the data is still an open challenge for hybrid modeling.

6.2 FUTURE WORK

This section presents a preliminary study that has been carried out, which deserves further detailed studies, as well as some recommendations for future work.

6.2.1 INTEGRATION OF SIGNALING PATHWAYS AND TRANSCRIPTIONAL REGULATORY NETWORKS

McMurray *et al.* [122] studied the gene expression profiles controlled synergistically by the loss of functions of p53 and Ras, which suggests that gene expression being regulated by signaling pathways is a highly cooperative process. Following this study, we investigated the relationship between extracellular stimuli and the corresponding cellular responses, *e.g.*, the gene expression profile.

Since the activities of the transcriptional factors are controlled by the signals transmitted through signaling pathways and the transcriptional factors regulate the gene expression, signal transduction and transcriptional regulation can be linked via the transcriptional factors. In Chapter 3, we discussed the simulation of the dynamics of the cellular signaling pathways. The transcriptional factor activities perturbed by diverse environmental stimuli can be simulated. However, the underlying mechanisms of transcriptional regulation is different (*e.g.*, in time scale) from signal transduction,

and accordingly we use an alternative methods, named Network Component Analysis (NCA) [66, 67, 68], for modeling transcriptional regulatory networks. NCA is a data decomposition method that can reversely derives the activities of the transcription factors from the gene expression profiles. NCA considers the biochemical reactions of the network to deduce the activity profiles of transcription factors using temporal gene expression data and relationships between transcription factors and their respective targeted genes.

Given the forward simulated and reversely engineered transcriptional factor profiles, we plan to construct a mathematical model that integrate the models of signaling pathways and transcriptional regulatory networks for different applications (*e.g.*, to study combinatorial drug effects) in future.

6.2.2 HYBRID MODELING OF CELLULAR SIGNALING PATHWAYS

In Chapter 3, we proposed a computational model to dynamically simulate the process of intracellular signal transduction. The parameters in this model (*e.g.*, the edge weights) are defined by users according to their prior knowledge. We can extend the work in Chapter 3 by learning the parameters from real data. For example, we can iteratively update the parameters of a knowledge-based model by matching the model simulations with real data. Compared with the generic model, the trained model can provide more specificities, such as the rewiring of the signaling pathways. On the other hand, the learning process can also include some prior knowledge of the signaling pathways instead of using pure data-driven methods.

6.2.3 MAPPING TIME SCALES OF COMPUTATIONAL SIMULATIONS TO REAL DATA

When a computational model is designed to simulate the dynamics of cellular signaling pathways, a challenge is how to match the simulation iterations to the real time points. Compared with a time-series experimental dataset, the computational simulation iterations are unitless [14]. It is difficult to map one simulation iteration corresponds to a real time point. In Chapter 3, we mapped the real time points uniformly to the simulated time-series. It is interesting to investigate how to match the time scales between simulated and real time-series data.

Bibliography

- [1] Z. Wang, M. Gerstein, and M. Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.
- [2] Y.J. Chu and D.R. Corey. Rna sequencing: Platform selection, experimental design, and data interpretation. *Nucleic Acid Therapeutics*, 22(4):271274, 2012.
- [3] G.P. Alamancos, E. Agirre, and E. Eyras. Methods to study splicing from high-throughput rna sequencing data. *Methods in Molecular Biology*, 1126:357–397, 2014.
- [4] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D.R. Kelley, H. Pimentel, S.L. Salzberg, J.L. Rinn, and L. Pachter. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature Protocols*, 7(3):562–578, 2012.
- [5] L. Alberghina and H.V. Westerhoff. *Systems biology: definitions and perspectives*. Springer, 2005.
- [6] B.O. Palsson. *Systems Biology: Properties of Reconstructed Networks*. Cambridge University Press, 2006.
- [7] H. Pearson. Genetics: what is a gene? *Nature*, 441(7092):398–401, 2006.
- [8] E. Pennisi. Human genome. a low number wins the genesweep pool. *Science*, 300(5625):1484, 2003.

- [9] D.L. Black. Mechanisms of alternative pre-messenger rna splicing. *Annual Review of Biochemistry*, 72:291–336, 2003.
- [10] A.J. Matlin, F. Clark, and C.W. Smith. Understanding alternative splicing: towards a cellular code. *Nature Reviews Molecular Cell Biology*, 6(5):386–398, 2005.
- [11] Q. Pan, O. Shai, L.J. Lee, B.J. Frey, and B.J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40(12):1413–1415, 2008.
- [12] J.A. Papin, T. Hunter, B.O. Palsson, and S. Subramaniam. Reconstruction of cellular signalling networks and analysis of their properties. *Nature Reviews Molecular Cell Biology*, 6(2):99–111, 2005.
- [13] B.B. Aldridge, J.M. Burke, D.A. Lauffenburger, and P.K. Sorger. Physicochemical modelling of cell signalling pathways. *Nature Cell Biology*, 8(11):1195–1203, 2006.
- [14] C. Terfve and J. Saez-Rodriguez. Modeling signaling networks using high-throughput phospho-proteomics. *Advances in Systems Biology*, 736:19–57, 2012.
- [15] E.H. Davidson and D.H. Erwin. Gene regulatory networks and the evolution of animal body plans. *Science*, 311(5762):796–800, 2006.
- [16] E.N. Olson. Gene regulatory networks in the evolution and development of the heart. *Science*, 313(5795):1922–1927, 2006.
- [17] F. Zhang, H. Chen, L.N. Zhao, H. Liu, T.M. Przytycka, and J. Zheng. Generalized logical model based on network topology to capture the dynamical trends of cellular signaling pathways. *BMC Systems Biology*, 10(Suppl 1):7, 2016.
- [18] J.M. Lee, E.P. Gianchandani, J.A. Eddy, and J.A. Papin. Dynamic analysis of integrated signaling, metabolic, and regulatory networks. *PLOS Computational Biology*, 4(5), 2008.

- [19] A. Kiyatkin, E. Aksamitiene, N.I. Markevich, N.M. Borisov, J.B. Hoek, and B.N. Kholodenko. Scaffolding protein Grb2-associated binder 1 sustains epidermal growth factor-induced mitogenic and survival signaling by multiple positive feedback loops. *Journal of Biological Chemistry*, 281(29):19925–19938, 2006.
- [20] C.Y. Ung, H. Li, X.H. Ma, J. Jia, B.W. Li, B.C. Low, and Y.Z. Chen. Simulation of the regulation of EGFR endocytosis and EGFR-ERK signaling by endophilin-mediated RhoA-EGFR crosstalk. *FEBS Letters*, 582(15):2283–2290, 2008.
- [21] S. Yamada, S. Shiono, A. Joo, and A. Yoshimura. Control mechanism of JAK/STAT signal transduction pathway. *FEBS Letters*, 534(1):190–196, 2003.
- [22] G. Bidkhori, A. Moeini, and A. Masoudi-Nejad. Modeling of tumor progression in NSCLC and intrinsic resistance to TKI in loss of PTEN expression. *PLOS ONE*, 7(10):e48004, 2012.
- [23] S.A. Kauffman. Homeostasis and differentiation in random genetic control networks. *Nature*, 224(5215):177–178, 1969.
- [24] S.A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, 22(3):437–467, 1969.
- [25] R. Thomas. Boolean formalization of genetic control circuits. *Journal of Theoretical Biology*, 42(3):563–585, 1973.
- [26] L. Mendoza, D. Thieffry, and E.R. Alvarez-Buylla. Genetic control of flower morphogenesis in arabidopsis thaliana: a logical analysis. *Bioinformatics*, 15(7-8):593–606, 1999.
- [27] R. Albert and H.G. Othmer. 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, 2003.

- [28] M. Chaves, R. Albert, and E.D. Sontag. Robustness and fragility of boolean models for genetic regulatory networks. *Journal of Theoretical Biology*, 235(3):431–449, 2005.
- [29] J. Saez-Rodriguez, L. Simeoni, J.A. Lindquist, R. Hemenway, U. Bommhardt, B. Arndt, U.U. Haus, R. Weismantel, E.D. Gilles, S. Klamt, and B. Schraven. A logical model provides insights into t cell receptor signaling. *PLOS Computational Biology*, 3(8):e163, 2007.
- [30] L. Calzone, L. Tournier, S. Fourquet, D. Thieffry, B. Zhivotovsky, E. Barillot, and A. Zinovyev. Mathematical modelling of cell-fate decision in response to death receptor engagement. *PLOS Computational Biology*, 6(3):e1000702, 2010.
- [31] L. Grieco, L. Calzone, I. Bernard-Pierrot, F. Radvanyi, B. Kahn-Perles, and D. Thieffry. Integrative modelling of the influence of MAPK network on cancer cell fate decision. *PLOS Computational Biology*, 9(10):e1003286, 2013.
- [32] J. Saez-Rodriguez, L.G. Alexopoulos, J. Epperlein, R. Samaga, D.A. Lauffenburger, S. Klamt, and P.K. Sorger. Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Molecular Systems Biology*, 5(331), 2009.
- [33] A. Mitsos, I.N. Melas, P. Siminelakis, A.D. Chairakaki, J. Saez-Rodriguez, and L.G. Alexopoulos. Identifying drug effects via pathway alterations using an integer linear programming optimization formulation on phosphoproteomic data. *PLOS Computational Biology*, 5(12):e1000591, 2009.
- [34] R. Sharan and R.M. Karp. Reconstructing boolean models of signaling. *Journal of Computational Biology*, 20(3):249–257, 2013.
- [35] J. Zheng, D. Zhang, P.F. Przytycki, R. Zielinski, J. Capala, and T.M. Przytycka. Simboolnet—a cytoscape plugin for dynamic simulation of signaling networks. *Bioinformatics*, 26(1):141–142, 2010.

- [36] R. Zielinski, P.F. Przytycki, J. Zheng, D. Zhang, T.M. Przytycka, and J. Capala. The crosstalk between EGF, IGF, and INSULIN cell signaling pathways—computational and experimental analysis. *BMC Systems Biology*, 4(3), 2009.
- [37] B.B. Aldridge, J. Saez-Rodriguez, J.L. Muhlich, P.K. Sorger, and D.A. Lauffenburger. Fuzzy logic analysis of kinase pathway crosstalk in TNF/EGF/INSULIN-induced signaling. *PLOS Computational Biology*, 5(4):e1000340, 2009.
- [38] Z.Y. Huang and J. Hahn. Fuzzy modeling of signal transduction networks. *Chemical Engineering Science*, 64(9):20442056, 2009.
- [39] M.K. Morris, J. Saez-Rodriguez, D.C. Clarke, P.K. Sorger, and D.A. Lauffenburger. 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):e1001099, 2011.
- [40] D.B. Hyduke and B.O. Palsson. Towards genome-scale signalling network reconstructions. *Nature Reviews Genetics*, 11(4):297–307, 2010.
- [41] W.W. Chen, M. Niepel, and P.K. Sorger. Classic and contemporary approaches to modeling biochemical reactions. *Genes Development*, 24(17):1861–1875, 2010.
- [42] L. Michaelis and M.L. Menten. Die kinetik der invertinwirkung. *Biochem, Z*(49):333369, 1913.
- [43] C.Y. Huang and J.E. Jr. Ferrell. Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proceedings of the National Academy of Sciences USA*, 93(19):10078–10083, 1996.
- [44] B. Schoeberl, C. Eichler-Jonsson, E.D. Gilles, and G. Muller. Computational modeling of the dynamics of the map kinase cascade activated by surface and internalized EGF receptors. *Nature Biotechnology*, 20(4):370–375, 2002.

- [45] B. Novk and J.J. Tyson. A model for restriction point control of the mammalian cell cycle. *Journal of Theoretical Biology*, 230(4):567–579, 2004.
- [46] J.G. Albeck, J.M. Burke, B.B. Aldridge, M. Zhang, D.A. Lauffenburger, and P.K. Sorger. Quantitative analysis of pathways controlling extrinsic apoptosis in single cells. *Molecular Cell*, 30(1):11–25, 2008.
- [47] L. Neumann, C. Pforr, J. Beaudouin, A. Pappa, N. Fricker, P.H. Krammer, I.N. Lavrik, and R. Eils. Dynamics within the CD95 death-inducing signaling complex decide life and death of cells. *Molecular Systems Biology*, 6(352), 2010.
- [48] N. Le Novere. Quantitative and logic modelling of molecular and gene networks. *Nature Reviews Genetics*, 16(3):146–158, 2015.
- [49] J.C. Marioni, C.E. Mason, S.M. Mane, M. Stephens, and Y. Gilad. RNA-Seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–17, 2008.
- [50] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.
- [51] Illumina. <https://www.illumina.com/>.
- [52] J.C. Vera, C.W. Wheat, H.W. Fescemyer, M.J. Frilander, D.L. Crawford, I. Hanski, and J.H. Marden. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, 17(7):1636–47, 2008.
- [53] C.P. El-Haibi, R. Singh, P. Gupta, P.K. Sharma, K.N. Greenleaf, S. Singh, and J.W. Lillard. Antibody microarray analysis of signaling networks regulated by CXCL13 and CXCR5 in prostate cancer. *Journal of Proteomics Bioinformatics*, 5:177–184, 2012.
- [54] K.A. Janes and M.B. Yaffe. Data-driven modelling of signal-transduction networks. *Nature Reviews Molecular Cell Biology*, 7(11):820–828, 2006.

- [55] K. Jaqaman and G. Danuser. Linking data to models: data regression. *Nature Reviews Molecular Cell Biology*, 7(11):813–819, 2006.
- [56] P. D’haeseleer. How does gene expression clustering work? *Nature Biotechnology*, 23(12):1499–501, 2005.
- [57] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences USA*, 95(25):14863–14868, 1998.
- [58] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22(3):281–5, 1999.
- [59] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences USA*, 96(6):2907–12, 1999.
- [60] F. Zhang, C.K. Kwoh, M. Wu, and J. Zheng. Data-driven prediction of cancer cell fates with a nonlinear model of signaling pathways. *The 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 436–444, 2014.
- [61] M.J. Lee, A.S. Ye, A.K. Gardino, A.M. Heijink, P.K. Sorger, G. MacBeath, and M.B. Yaffe. Sequential application of anticancer drugs enhances cell death by rewiring apoptotic signaling networks. *Cell*, 149(4):780–94, 2012.
- [62] K.A. Janes, J.R. Kelly, S. Gaudet, J.G. Albeck, P.K. Sorger, and D.A. Lauffenburger. Cue-signal-response analysis of TNF-induced apoptosis by partial least squares regression of dynamic multivariate data. *Journal of Computational Biology*, 11(4):544–61, 2004.
- [63] S. Gaudet, K.A. Janes, J.G. Albeck, E.A. Pace, D.A. Lauffenburger, and P.K. Sorger. A compendium of signals and responses triggered

- by prodeath and prosurvival cytokines. *Molecular Cellular Proteomics*, 4(10):1569–1590, 2005.
- [64] K.A. Janes, J.G. Albeck, S. Gaudet, P.K. Sorger, D.A. Lauffenburger, and M.B. Yaffe. A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. *Science*, 310(5754):1646–1653, 2005.
- [65] S.C. Peng, D.S. Wong, K.C. Tung, Y.Y. Chen, C.C. Chao, C.H. Peng, Y.J. Chuang, and C.Y. Tang. Computational modeling with forward and reverse engineering links signaling network and genomic regulatory responses: NF-kB signaling-induced gene expression responses in inflammation. *BMC Bioinformatics*, 11:308, 2010.
- [66] C. Chang, Z. Ding, Y.S. Hung, and P.C. Fung. Fast network component analysis (fastnca) for gene regulatory network reconstruction from microarray data. *Proceedings of the National Academy of Sciences USA*, 24(11):1349–1358, 2008.
- [67] J.C. Liao, R. Boscolo, Y.L. Yang, L.M. Tran, C. Sabatti, and V.P. Roychowdhury. Network component analysis: reconstruction of regulatory signals in biological systems. *Proceedings of the National Academy of Sciences USA*, 100(26):15522–15527, 2003.
- [68] A. Noor, A. Ahmad, E. Serpedin, M. Nounou, and H. Nounou. ROB-NCA: robust network component analysis for recovering transcription factor activities. *Bioinformatics*, 29(19):2410–2418, 2013.
- [69] I.N. Melas, A. Mitsos, D.E. Messinis, T.S. Weiss, and L.G. Alexopoulos. Combined logical and data-driven models for linking signalling pathways to cellular response. *BMC Systems Biology*, 5:107, 2011.
- [70] S. Gupta, S.S. Bisht, R. Kukreti, S. Jain, and S.K. Brahmachari. Boolean network analysis of a neurotransmitter signaling pathway. *Journal of Theoretical Biology*, 244(3):463–469, 2007.

- [71] H.F. Fumi and M.L. Martins. Boolean network model for cancer pathways: predicting carcinogenesis and targeted therapy outcomes. *PLOS One*, 8(7):e69008, 2013.
- [72] T. Mori, M. Flttmann, M. Krantz, T. Akutsu, and E. Klipp. Stochastic simulation of boolean rxncon models: towards quantitative analysis of large signaling networks. *BMC Systems Biology*, 9(45), 2015.
- [73] A. Faur, A. Naldi, C. Chaouiya, and D. Thieffry. Dynamical analysis of a generic boolean model for the control of the mammalian cell cycle. *Bioinformatics*, 22(14):e124–131, 2006.
- [74] Z. Mai and H. Liu. Boolean network-based analysis of the apoptosis network: irreversible apoptosis and stable surviving. *Journal of Theoretical Biology*, 259(4):760–769, 2009.
- [75] R. Schlatter, K. Schmich, I. Avalos Vizcarra, P. Scheurich, T. Sauter, C. Borner, M. Ederer, I. Merfort, and O. Sawodny. On/off and beyond—a boolean model of apoptosis. *PLOS Computational Biology*, 5(12):e1000595, 2009.
- [76] T. Tian and K. Smith-Miles. Mathematical modeling of GATA-switching for regulating the differentiation of hematopoietic stem cell. *BMC Systems Biology*, 8, 2014.
- [77] S.K. Mishra, S.S. Bhowmick, H.E. Chua, F. Zhang, and J. Zheng. Computational cell fate modelling for discovery of rewiring in apoptotic network for enhanced cancer drug sensitivity. *BMC Systems Biology*, 9(Suppl 1):S4, 2015.
- [78] C. Chaouiya, A. Naldi, and D. Thieffry. Logical modelling of gene regulatory networks with ginsim. *Methods in Molecular Biology*, 804:463–479, 2012.
- [79] P. rdi and J. Tth. *Mathematical Models of Chemical Reactions: Theory and Applications of Deterministic and Stochastic Models*. Princeton University Press; 1st edition, 1989.

-
- [80] F. Zhang, M. Wu, X.J. Li, X.L. Li, C.K. Kwoh, and J. Zheng. Predicting essential genes and synthetic lethality via influence propagation in signaling pathways of cancer cell fates. *Journal of Bioinformatics and Computational Biology*, 11(1541002), 2015.
- [81] S. Ekins, Y. Nikolsky, A. Bugrim, E. Kirillov, and T. Nikolskaya. Pathway mapping tools for analysis of high content data. *Methods in Molecular Biology*, (356):319–350, 2007.
- [82] M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nature*, 1(28):27–30, 2000.
- [83] I.J. Leontaritis and S.A. Billings. Input-output parametric models for non-linear systems. part i: deterministic non-linear systems. *International Journal of Control*, 41:303–328, 1985.
- [84] I.J. Leontaritis and SA Billings. Input-output parametric models for non-linear systems. part ii: stochastic non-linear systems. *International Journal of Control*, 41:329–344, 1985.
- [85] A. Clauset, C.R. Shalizi, and M.E.J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661703, 2009.
- [86] M.E.J. Newman. Power laws, pareto distributions and Zipfs law. *Contemporary Physics*, 46:323–351, 2005.
- [87] E.T. Lu and R.J. Hamilton. Avalanches of the distribution of solar flares. *Astrophysical Journal*, 380:8992, 1991.
- [88] D.H. Zanette and S.C. Manrubia. Vertical transmission of culture and the distribution of family names. *Physica A*, 295:18, 2001.
- [89] M. Kleiber. Body size and metabolism. *Physica A*, 6:315–353, 1932.
- [90] B.J. Enquist and K.J. Niklas. Invariant scaling relations across tree-dominated communities. *Nature*, 410:655–660, 2001.
- [91] E. Bullinger. System analysis of a programmed cell death model. *IEEE Conference on Decision and Control*, pages 7994–7995, 2005.

- [92] JP Higgins. Nonlinear systems in medicine. *Yale Journal of Biology and Medicine*, 75(5-6):247–260, 2002.
- [93] S. Legewie, N. Blthgen, and H. Herzel. Mathematical modeling identifies inhibitors of apoptosis as mediators of positive feedback and bistability. *PLOS Computational Biology*, 2(9):1061–1073, 2006.
- [94] T. Eissing, H. Conzelmann, E.D. Gilles, F. Allgwer, E. Bullinger, and P. Scheurich. Bistability analyses of a caspase activation model for receptor-induced apoptosis. *Journal of Biological Chemistry*, 279(35):36892–36897, 2004.
- [95] R. Callard, A.J. George, and J. Stark. Cytokines, chaos, and complexity. *Immunity*, 11(5):507–513, 1999.
- [96] Hpn-dream breast cancer network inference challenge (accessed: 2015-04-07). <https://www.synapseorg/#!/Synapse:syn1720047>.
- [97] Mdanderson rppa resource (accessed: 2015-04-18). <http://www.mdanderson.org/education-and-research/resources-for-professionals/scientific-resources/core-facilities-and-services/functional-proteomics-rppa-core/faq/functional-proteomics-reverse-phase-protein-array-core-facility-faq.html>.
- [98] M.A. Savageau. Biochemical systems analysis i some mathematical properties of the rate law for the component enzymatic reactions. *Journal of Theoretical Biology*, 25:365–369, 1969.
- [99] M.A. Savageau. Biochemical analysis ii the steady-state solutions for an n-pool system using a power-law approximation. *Journal of Theoretical Biology*, 25:370–379, 1969.
- [100] H.W. Bode. *Network Analysis and Feedback Amplifier Design*. Van Nostrand, Princeton, NJ, 1945.

- [101] J.Y. Hong, G.H. Kim, J.W. Kim, S.S. Kwon, E.F. Sato K.H. Cho, and E.B. Shim. Computational modeling of apoptotic signaling pathways induced by cisplatin. *BMC Systems Biology*, 11(6):122, 2012.
- [102] A. Funahashi, N. Tanimura, M. Morohashi, and H. Kitano. Cellde-signer: a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO*, 1:159–162, 2003.
- [103] C. Li, M. Donizelli, N. Rodriguez, H. Dharuri, L. Endler, V. Chelliah, L. Li, E. He, A. Henry, M.I. Stefan, J.L. Snoep, M. Hucka, N. Le Novre, and C. Laibe. Biomodels database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Systems Biology*, 4(92), 2010.
- [104] S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(51):79–86, 1952.
- [105] S. Boeck, A. Jung, R.P. Laubender, J. Neumann, R. Egg, C. Goritschan, U. Vehling-Kaiser, C. Winkelmann, L. Fischer von Weikersthal, M.R. Clemens, T.C. Gauler, A. Mrten, S. Klein, G. Kojouharoff, M. Barner, M. Geissler, T.F. Greten, U. Mansmann, T. Kirchner, and V. Heinemann. EGFR pathway biomarkers in erlotinib-treated patients with advanced pancreatic cancer: translational results from the randomised, crossover phase 3 trial AIO-PK0104. *British Journal of Cancer*, 108(2):469–476, 2013.
- [106] W.G.Jr. Kaelin. Synthetic lethality: a framework for the development of wiser cancer therapeutics. *Genome Medicine*, 1(10):99.1–99.6, 2009.
- [107] W.G.Jr. Kaelin. The concept of synthetic lethality in the context of anticancer therapy. *Nature Reviews Cancer*, 5(9):689–698, 2005.
- [108] M. Wu, X. Li, F. Zhang, X.L. Li, C.K. Kwoh, and J. Zheng. In silico prediction of synthetic lethality by meta-analysis of genetic interactions, functions, and pathways in yeast and human cancer. *Cancer informatics*, 13(S3):71–80, 2014.

- [109] O. Güell, F. Sagués, and M. Ángeles Serrano. Essential plasticity and redundancy of metabolism unveiled by synthetic lethality analysis. *P-LOS Computational Biology*, 10(5):e1003637, 2014.
- [110] M.A. Heiskanen and T. Aittokallio. Mining high-throughput screens for cancer drug targets—lessons from yeast chemical-genomic profiling and synthetic lethality. *WIREs Data Mining Knowl Discovery*, 2(3):263–272, 2012.
- [111] X.J. Li, S.K. Mishra, M. Wu, F. Zhang, and J. Zheng. Syn-lethality: An integrative knowledge base of synthetic lethality towards discovery of selective anticancer therapies. *BioMed Research International*, 2014:196034, 2014.
- [112] A. Basu and U. Sivaprasad. Protein kinase cepsilon makes the life and death decision. *Cell Signal*, 19(9):1633–1642, 2007.
- [113] L.A. Carvajal, P.J. Hamard, C. Tonnessen, and J.J. Manfredi. E2F7, a novel target, is up-regulated by P53 and mediates dna damage-dependent transcriptional repression. *Genes Development*, 26(14):1533–1545, 2012.
- [114] S. Liu, A. Mirza, and L. Wang. Generation of P53 target database via integration of microarray and global P53 dna-binding site analysis. *Methods in Molecular Biology*, 281:33–54, 2004.
- [115] F.J. Reyes-Zurita, G. Pachn-Pea, D. Lizrraga, E.E. Rufino-Palomares, M. Cascante, and J.A. Lupiez. The natural triterpene maslinic acid induces apoptosis in HT29 colon cancer cells by a JNK-P53-dependent mechanism. *BMC Cancer*, 11(154):10.1186/1471-2407-11-154, 2011.
- [116] S. McNeely, R. Beckmann, and A.K. Bence Lin. CHEK again: revisiting the development of CHK1 inhibitors for cancer therapy. *Pharmacology Therapeutics*, 142(1):1–10, 2014.
- [117] J. Chung, T.C. Grammer, K.P. Lemon, A. Kazlauskas, and J. Blenis. PDGF- and INSULIN-dependent pp70S6k activation mediated by phosphatidylinositol-3-OH kinase. *Nature*, 370(6484):71–75, 1994.

- [118] C. Greider, A. Chattopadhyay, C. Parkhurst, and E. Yang. BCL-x(L) and BCL2 delay Myc-induced cell cycle entry through elevation of P27 and inhibition of G1 cyclin-dependent kinases. *Oncogene*, 21(51):7765–7775, 2002.
- [119] W.H. Chen, P. Minguez, M.J. Lercher, and P. Bork. OGEE: an on-line gene essentiality database. *Nucleic Acids Research*, 40(Database issue):D901–D906, 2012.
- [120] S. Morandell and M.B. Yaffe. Exploiting synthetic lethal interactions between dna damage signaling, checkpoint control, and P53 for targeted cancer therapy. *Progress in Molecular Biology and Translational Science*, 110:289–314, 2012.
- [121] N. Le Meur and R. Gentleman. Modeling synthetic lethality. *Genome Biology*, 9(9):R135, 2008.
- [122] H.R. McMurray, E.R. Sampson, G. Compitello, C. Kinsey, L. Newman, B. Smith, S.R. Chen, L. Klebanov, P. Salzman, A. Yakovlev, and H. Land. Synergistic response to oncogenic mutations defines gene class critical to cancer phenotype. *Nature*, 453(7198):1112–1116, 2008.