Machine learning methods for predicting cancer drug effects from signaling and cell fate data

Mishra Shital Kumar

2017

Mishra Shital Kumar. (2017). Machine learning methods for predicting cancer drug effects from signaling and cell fate data. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/70489>

<https://doi.org/10.32657/10356/70489>

MACHINE LEARNING METHODS FOR PREDICTING
CANCER DRUG EFFECTS FROM SIGNALING AND
CELL FATE DATA

A Thesis Submitted to the
School of Computer Science and Engineering
of the Nanyang Technological University

by

MISHRA SHITAL KUMAR

in Fulfilment of the Requirement
for the Degree of Doctor of Philosophy

2017

Acknowledgments

I was very new to the field of bioinformatics and cancer research when I joined NTU in 2012. However, with the continuous guidance of my supervisor, Asst. Prof. Zheng Jie, pursuing PhD was a wonderful journey. His knowledge, suggestions, and discussions helped me to become a capable researcher. His encouragement also helped me to overcome several challenges encountered in my research. It is a great honour for me work under his supervision as a PhD student. I have my deepest regard and special thanks for him for the invaluable supervision of my PhD research.

I express my deepest thanks Dr. Cheng Li and Dr. Sourav Saha Bhowmick for being part of my Thesis Advisory Committee and providing me invaluable guidance and encouragement in doing research.

I am highly indebted to Dr. Huey Eng Chua, Dr. Chen Haifen, Dr. Liu Hui, Dr. Chong Ket Hing, Dr. Wu Min, Dr. Rajesh Sharma, Mr. Tan Sing Yau, Mr. Fan Zhang, Ms. Li Xuejuan, and Ms. Guo Jing for their advice on my research projects, this report, and for exchanging various ideas with me.

I am very grateful to the members of our Biomedical Informatics Lab (BIL) team in NTU. It is a pleasure to collaborate with my team mates.

Moreover, I am very grateful to the SCSE, NTU for providing me the invaluable opportunity for becoming a graduate student at NTU and to pursue my doctoral research.

Author's Publications

- (i) **Mishra SK**, Sourav S. Bhowmick, Huey Eng Chua, and Jie Zheng*; “Predictive modeling of drug effects on signaling pathways in diverse cancer cell lines.” *ACM-BCB 2016, Seattle, WA, USA* .
- (ii) **Mishra SK**, 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, no. 1 (2015): 1.
- (iii) Xue-juan Li, **Mishra SK**, 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*, vol. 2014, Article ID 196034, 7 pages, 2014. doi:10.1155/2014/196034
- (iv) Haifen Chen, Jing Guo, **Mishra SK**, Paul Robson, Mahesan Niranjan, and Jie Zheng*; “Single-cell transcriptional analysis to uncover regulatory circuits driving cell fate decisions in early mouse development,” *Bioinformatics*, 31.7 (2015): 1060-1066.
- (v) Hui Liu, Fan Zhang, **Mishra SK**, Shuigeng Zhou, and Jie Zheng*; “Knowledge-guided fuzzy logic network modeling to infer cancer signaling pathways from time-series data,” *Scientific Reports*, 2016.

Contents

Acknowledgments	i
Publication	ii
List of Figures	vi
List of Tables	ix
List of Abbreviations	xi
List of Notation	1
Abstract	2
1 Introduction	4
1.1 Background	4
1.1.1 Components	5
1.1.2 Interactions	6
1.1.3 Integration and reconstruction	7
1.1.4 Biological networks	8
1.2 Motivation and objectives	8
1.3 Contributions	9
1.4 Outline	10
2 Related works	11
2.1 Signaling pathways	12
2.2 Omics data	12
2.3 Computational modeling approaches	14
2.3.1 Knowledge-driven dynamical modeling	14
2.3.2 Data-driven statistical modeling	15

2.3.3	Hybrid modeling	17
2.4	Estimation of rate constants and initial concentrations	19
2.5	Context-specific modeling	21
2.6	Validation	22
3	Discovery of network rewiring for enhanced cancer drug sensitivity	24
3.1	Introduction	24
3.2	Methods	25
3.2.1	Data	25
3.2.2	Apoptosis network construction	27
3.2.3	Partial least squares regression	30
3.2.4	Simulation with ODE	30
3.2.5	Genetic algorithm	31
3.2.6	Dynamic time warping	33
3.2.7	Sensitivity analysis	35
3.2.8	Simulated annealing	35
3.3	Results	36
3.3.1	PLSR for simulation data from N_1	36
3.3.2	Inferring the tumorigenic network N_2	37
3.3.3	Inferring the drug sensitive network N_3	38
3.3.4	Sensitivity analysis of the model	39
3.3.5	Performance of genetic algorithm with varying mutation probability	39
3.3.6	Comparison of GA and simulated annealing derived rewiring	40
3.4	Discussion and conclusion	41
4	Predicting drug effects in diverse cancer cell lines	47
4.1	Introduction	47
4.2	Methods	50
4.2.1	Single-cell data	50
4.2.2	Derivation of signaling network structure	50
4.2.3	Kinetic modeling	52

4.2.4	Parameter inference with ABC-SysBio	52
4.2.5	Context-specific modeling	53
4.3	Results	54
4.3.1	Simulations with random and inferred parameters	54
4.3.2	Prediction of apoptosis	55
4.3.3	Virtual knockdown of caspases	58
4.3.4	Predicting drug-induced pathway alterations	61
4.4	Conclusion and discussion	62
5	Summary and future work	66
5.1	Summary	66
5.2	Future directions	67
5.2.1	Integrating alterations in network structure with parameter inference	68
5.2.2	Cancer attractors	69
5.2.3	Synthetic lethality	73
5.2.4	Genomic mutations	74
	References	77

List of Figures

3.1	Basic apoptosis network and various rewiring events. Basic apoptotic network contained all edges except the broken edges (green and blue). Edges represented by thick red colour were deleted by Genetic Algorithm while inferring N_2 from N_1 . Broken edges represented by green colour were inserted by Genetic Algorithm while inferring N_2 from N_1 . Edges represented by magenta colour were deleted by Genetic Algorithm while inferring N_3 from N_2 . Broken edges represented by blue colour were inserted by Genetic Algorithm while inferring N_3 from N_2	28
3.2	PLSR plot for simulation data for apoptosis from network N_1. .	37
3.3	Simulation results for Genetic Algorithm: fitting experimental data with simulation data. (a) Simulation result for basic apoptotic network, (b) Simulation result for DMSO treatment (N_2 with DMSO treatment). (c) Comparing N_2 based simulation data with ERL-DOX treatments data by correlation of apoptosis rates over time points. (d) Simulation result for ERL-DOX treatments. The scores represents DTW score.	43
3.4	Evaluating performance of the Genetic Algorithm for different mutation rates. Comparison of computational apoptosis data with experimental apoptosis data from [12] for different mutations probabilities, <i>i.e.</i> , 0.1, 0.05 and 0.01 for deriving N_2 from N_1 . The results suggest that lower mutation probability is better for enhancing goodness of fit using Genetic Algorithm. Our findings suggest that a lower mutation probability can lead to convergence of GA as compared to a higher mutation probability.	44

3.5	Simulation results for Simulated Annealing (a) Simulation result for basic apoptotic network, (b) Simulation result for DMSO treatment. (c) Comparing N_2 based simulation data with ERL-DOX treatments data. (d) Simulation result for ERL-DOX treatments. The scores represent DTW score.	45
4.1	Flow diagram depicting the steps in our method.	49
4.2	Apoptosis signaling network. The network kinetic parameters were inferred from the phosphoproteomics data from [12]. Nodes in cyan color (total 19) represent proteins with phosphoproteomics data available in Yaffe's dataset. The black and red colored edges represent the activation and inhibition of the signals, respectively. The blue edges indicate the inhibitory drug effect induced by Erlotinib on EGFR signaling pathway, i.e., blockage of signal transduction in BT-20 cells as inferred from the drug-sensitive phosphorylation dataset in [12]. The same model, after calibration with cell line specific drug sensitive data from [12], was used to analyse drug effects on three breast cancer cell lines, i.e., BT-20, MCF7 and MDA-MB-453.	51
4.3	Distributions of Euclidean distances. Distributions of Euclidean distances (in blue) between the simulated and the Yaffe's data for each of the 19 signaling proteins and apoptosis (bottom-right panel) for 1000 sets of random parameters for the BT-20 cell line. The nabla symbol (in red) represents the Euclidean distance between the simulated and experimental data after model calibration. The x -axis represents the distributions for the Euclidean distances of the corresponding variable.	55
4.4	Inference of Parameters. Fitting of simulation data to experimental data for the 19 proteins for the BT-20 cell line from [12] based on inferred parameters for the signaling network.	57

4.5	Prediction of context-specific apoptosis. Prediction of apoptosis and comparison of cell line specific computational apoptosis data with experimental apoptosis data from [12] for three cell lines, i.e., BT-20, MCF7 and MDA-MB-453. The calibrated computational model in Figure 4.2 was able to predict the cell line specific apoptosis with high Pearson correlation coefficients, such as 0.994 in BT-20, 0.9569 in MCF7 and 0.904 in MDA-MB-453 cell lines. The Euclidian distance between simulated and experimental apoptosis data for the BT-20, MCF7, and MDA-MB-453 cell lines were 18.89, 58.98, and 10.25, respectively.	58
4.6	Prediction of context-specific CASP8 and CASP6 knockdown effects. Effects of CASP8 and CASP6 knockdown on apoptosis in trained models for the BT-20, MCF7, and MDA-MB-453 cell lines. The increase of apoptosis is weakened in response to the knockdown in BT-20 and MDA-MB-453 cells. Moreover, the combined effect of knockdown of CASP6 and CASP8 was higher than the individual knockdown. The impact of knockdowns in MDA-MB-453 cell line was higher and lower than the MCF7 and BT-20 cells, respectively. However, the caspase knockdown has little impact on apoptosis of MCF7 cell line as the apoptotic curves for CASP6 and CASP8 knockdown remained similar to the apoptosis before the knockdown.	60
5.1	Prediction of apoptosis after rewiring and the parameters inference in MCF7 cell line. Comparison of the topological rewiring induced changes in apoptosis with the experimental apoptosis data from [12] for MCF7 cells. The newly added edges in Figure 4.2 were a) p53-CASP8, b) Smac-CASP9, and c) both p53-CASP8 and SMAC-CASP9, simultaneously. After addition of new edges in Figure 4.2, the computational model was able to predict the apoptosis in MCF7 cell line with improved Euclidean distance as compared to Figure 4.5 (b), such as a) p53-CASP8 (55.65), b) SMAC-CASP9 (47.11), c) both p53-CASP8 and SMAC-CASP9 (42.09). The Euclidian distance after rewiring between simulated and experimental apoptosis data for MCF7 was significantly improved.	70

List of Tables

3.1	Differential equations for the basic apoptosis network N_1 (Figure 3.1)	42
3.2	Rate constants used for the rewiring in basic apoptosis network	46
4.1	Z-test details for the population size of 1000	56
4.2	Activation rate constants downstream of EGFR among different breast cancer cell lines. The DMSO treatment rate constants have been inferred from the phosphorylation level in the DMSO treated cancer cells. The Drug treatment by Erlotinib rate constants have been inferred from the phosphorylation level data in the drug sensitive cancer cells. There is no healthy data in the published dataset by [1]. All the data are recorded from cancer cells corresponding to the control treatment and drug treatment. The table uncovers the changes in the rate constants to explain the enhanced drug sensitivity.	59

List of Abbreviations

HGNO	Human Genome Nomenclature Organization
ODE	Ordinary Differential Equation
COSMIC	Catalogue Of Somatic Mutations In Cancer
DTW	Dynamic Time Warping
BCL-2	B-Cell Lymphoma 2
TNBC	Triple-Negative Breast Cancer
MCF7	Michigan Cancer Foundation-7
GA	Genetic Algorithm
SA	Simulated Annealing
SCT	Single Cell Technology
GEO	Gene Expression Omnibus
CCLE	Cancer Cell Line Encyclopedia
PCA	Principal Components Analysis
PLS	Partial Least Squares
VIP	Variable Importance in Projection
PLSR	Partial Least Squares Regression
SBML	Systems Biology Markup Language
CHER	Contextual Heterogeneity Enabled Regression
DMFS	Distant Metastasis Free Survival
TNF	Tumor Necrosis Factor
EGFR	Epidermal Growth Factor Receptor
FasR	FAS Receptor
FasL	FAS Ligand
MAPK	Mitogen-Activated Protein Kinase
TAR	Erlotinib
DOX	Doxorubicin
DT	Doxorubicin and Erlotinib together
D-T	Doxorubicin followed by Erlotinib
T-D	Erlotinib followed by Doxorubicin
DMSO	Dimethyl sulfoxide
ABC	Approximate Bayesian Computation
SMC	Sequential Monte Carlo
SL	Synthetic lethality
TCGA	The Cancer Genome Atlas
TNFR	Tumor Necrosis Factor Receptor

List of Notations

$ \mathbf{a} - \mathbf{b} ^2$	Euclidian distance between $\mathbf{a}$ and $\mathbf{b}$
$\mathbf{a} \cdot \mathbf{b}$	Elementary multiplication between $\mathbf{a}$ and $\mathbf{b}$

Abstract

An important area of research within the discipline of the Computational Systems Biology is to investigate the mechanisms of determination of cell fate. As the cellular development progresses, a cell goes through several phenotypes (*e.g.* Apoptosis, Proliferation, G1, S, G2 and M). Apoptosis, which is a process of programmed cell death in case of a healthy cell, is a special case of cell fate. For example, development of fingers in human embryo occurs due the apoptotic cell death of the cells between digits. Regulatory networks such as transcriptional regulatory network for regulating cell differentiation in the embryonic development, signaling networks for regulating activity level of proteins, etc., determine the course of cell fate and lineage commitment. Signaling pathways play important regulatory roles in apoptosis. However, in cancer cells signaling pathways are in an atypical forms that lead to continuous survival, growth and proliferation of tumor cells. A major objective in cancer research is to investigate the dynamics of signaling pathways that influence the apoptosis of tumor cells. The changes in signaling pathways supporting uncontrolled cell growth, termed as rewiring, can lead to dysregulation of cell fates. Hence comparative analysis of normal and oncogenic signal transduction pathways may provide insights into dynamics of cancer drug induced signaling changes and enable the discovery of novel and effective anti-cancer therapies. A better understanding of rewiring is essential for developing effective anti-cancer therapeutic strategies.

In this thesis, we first demonstrate how rewiring of signaling pathways regulate diverse signaling molecules in an apoptotic network. We use a simple yet intuitive differential equation based computational model to understand the roles played by certain essential regulatory proteins in the regulation of programmed cell death. Inspired by the success of dynamical modeling and data analysis in cancer biology, we propose a hybrid modeling approach combining computational models with experimental phosphoproteomics data.

We use a hybrid modeling approach based on ordinary differential equation (ODE) models and machine learning techniques to map network rewiring in the apoptotic pathways that may be responsible for the increase of drug sensitivity of tumour cells in triple-negative breast cancer. Our method employs Genetic Algorithm to search for the most likely network topologies by iteratively generating simulated protein phosphorylation data using ODEs and the rewired network and then fitting the simulated data with real data of cancer signaling and cell fate. Most of our predictions are consistent with experimental evidence from literature. Combining the strengths of knowledge-driven and data-driven approaches, our hybrid model can help uncover molecular mechanisms of cancer cell fate at the systems level.

Secondly, we investigate the context-specific drug effects on diverse cancer cell lines. We construct a knowledge-based model of ordinary differential equations (ODEs) for the apoptotic signaling network and subsequently infer model parameters (e.g. reaction rates) from real phosphoproteomics data for three breast tumor cell lines, *i.e.* BT-20, MCF7 and MDA-MB-453 using a Bayesian framework of inference. The model is used to predict apoptosis in response to various perturbations such as caspase knockdown for each of the three cell lines which can be validated using the experimental literature. The inferred changes of the parameters reveal drug effects on diverse cell lines under the treatment with the drug of Erlotinib.

Despite having a limited amount of real data, we believe that our *in silico* apoptosis model is able to capture several dynamic characteristics of drug sensitivity of tumor cells in a coherent way. We hope that as new and larger datasets become available in future, our proposed model can be used to gain more in-depth and quantitative understanding about the regulation of cancer cell death. Our studies offer a novel method of hybrid modeling for linking the computational model with biological data. It is demonstrated to be promising for explaining and predicting the impact of anti-cancer therapies on cancer cells at the systems level.

Chapter 1

Introduction

1.1 Background

In the last few decades humanity has witnessed unprecedented growth of science, engineering and technology. The development in technology has witnessed tremendous application in the field of computational biology. With the use of technology, now it is possible to understand the biological system from mathematical perspective. Technological innovations have now enabled mankind to study a larger and more complex biological systems. The phenomenal rise in the use of technology, such as simulation and modeling, has made it possible to look at a biological system in a holistic manner, where each and every part of the system has a role to play, and system as a whole is much more than just addition of its parts. This has enabled scientific community to study various types of diseases including cancers in a comprehensive and coherent way.

Systems biology, also termed as Systemics, uses the concept of the systems theory from the management discipline. In the field of biological research, it is still an emerging approach with tremendous potential for solving biological problems. Traditional reductionism based approach would usually try to explain behaviour of a cell based on one of its individual constituents, such as gene expression patterns. However, systems biology is a highly interdisciplinary approach based on the concept of holism. This concept is very different from the traditional concept of biology which is based on the concept of reductionism which focuses on individual constituents of a complex system. For example, to study a cell, the systems biology based computational model would integrate various constituents of the cell like genes, proteins, signaling pathways, -omics data etc.

in a holistic manner. This helps us in understanding the integrative complex system behaviour. Systems biology aims at discovering characteristics of cells, tissues, organs etc. as a system that is made of interconnected parts. Systems biology often relies on signal transduction networks to understand complex systems behaviour. It heavily utilises variety of computational models (*e.g.* Continuous and Boolean) to understand the systems behaviour [1].

1.1.1 Components

According to Human Genome Nomenclature Organization (HGNO), a gene is part of a DNA segment that is capable of contributing to different phenotypes and functions [2]. A gene can be termed as the molecular unit of heredity [3]. Genes are responsible for storing, utilising and providing template information for protein synthesis. For humans, the size of genes varies from a couple of hundred base-pairs to more than two million base pairs. Each individual human has two copies for each of the genes which are inherited from each parent. Each of the genes has regulatory regions (like promoters) which can regulate the expression pattern of corresponding gene and the protein synthesis. As per the findings of Human Genome Project, the numbers of genes in human are around 20 to 25 thousands genes [4].

Transcription and translation are two very important processes involved in the development of phenotype from genotypes. Transcription is the process of copying the nucleotide base pairs from the DNA segment into RNA. This activity is carried out by the enzyme RNA polymerases which produces a complementary RNA strand, also called primary transcript that further goes through the process of splicing inside the nucleus. Introns do not code for proteins and they are removed during the process of splicing. The exons, which are basically the coding segments, are joined together. This process converts primary transcript into a mature transcript which provides the template for protein synthesis. For most of the eukaryote genes, the primary transcript processing is a compulsory process for obtaining mature messenger mRNA for protein synthesis [5].

Once the mRNA is available, it is used for the protein synthesis via the process called translation. During translation, the mRNA transcript is used as template by cellular ribosomes to code for the amino acid sequence. A group of three nucleotide triplet

code for an amino acid. The base pair sequences (A, C, G, T) in the template strand eventually guides the generation of amino acids [6]. Amino acids are added at C-terminal for peptide. There are a total of twenty types of amino acids but a protein may contain even thousands of amino acids. These polypeptides are folded into three dimensional shapes to become active protein [7]. The process of chemical transformation within cells, termed as metabolism, produces usually small molecules called metabolites. Metabolites are responsible for stimulating and inhibiting the enzymes, directly or indirectly, thereby playing significant role in the growth, development and reproduction [8].

1.1.2 Interactions

The various components of biological systems interact with each other using biochemical reactions where an enzyme works on substrate and catalyses the substrate. In many reactions, the reaction of enzyme and substrates leads to the formation of enzyme-substrate complex. The substrates are eventually converted into products of the reaction. The enzyme is preserved in the reaction and once a reaction is completed, the enzyme is free to interact with other substrates. The reaction enzyme-substrate chemical reaction is mathematically encoded as:

where A, B, and C represent an enzyme, a substrate, and a product, respectively. The constants k_f , k_r , and k_{cat} represent the biochemical reaction rate constants [9].

The Gene Regulatory Networks (GRN) plays a very crucial role in determining cell cycle, cell differentiation, signal transduction, etc. The dynamics of GRN can help understand the mechanism of various diseases, such as cancer. The computational prediction of behaviours of GRN requires modeling and simulation of GRN. In gene regulatory networks, one gene can stimulate or inhibit the expression of another gene. This happens when the protein synthesised from one gene binds to the promoter of another gene. A gene product can also bind to the promoter of its own gene, and thereby inducing self-activation or self-inhibition. A GRN can be modelled using the Hill Equations [10]. For

example, the computational model for a GRN containing two mutually inhibiting genes x and y is given as [10] :

$$\frac{dx}{dt} = B_x + \frac{F_{yx} \cdot K_{dyx}^{n_H}}{K_{dyx}^{n_H} + y^{n_H}} + D_x \cdot x, \quad (1.2)$$

$$\frac{dy}{dt} = B_y + \frac{F_{xy} \cdot K_{dxy}^{n_H}}{K_{dxy}^{n_H} + x^{n_H}} + D_y \cdot y \quad (1.3)$$

where x and y represent the concentrations of the corresponding gene products, D_x and D_y are the degradation rate constants, B_x and B_y represent the basal expression rates, F_{yx} and F_{xy} are the rate constants, and K_{dyx} and K_{dxy} are the affinity constants for the mutual inhibition effect of gene y on x , and of gene x on y , respectively. The parameter n_H represents the Hill Coefficient. The mutually inhibitory interaction is ultra-sensitive for $n_H > 1$. Identifying such biochemical reactions are very crucial for the mathematical modeling of the biological systems. For example, the study in [10] used similar model to simulate bistability in a GRN.

1.1.3 Integration and reconstruction

Once a set of reactions for the biological system has been identified, the system can be modeled by integrating descriptions of all the reactions into one model. The modeling of the biological system requires identification of the various system components and their responses to the stimuli. The reaction sets are often described using ordinary or partial differential equations. The solutions to these equations provide insights into the system. The size of biological system may also vary considerably. In some cases we may need to have hundreds to thousands of biomolecules to represent a complex biological system. In addition to this, feedback loop, bifurcation, oscillation can make the modeling process more complicated [11].

The simulation of the model further requires a variety of information. For example, finding the reaction rate constants is often challenging owing to the lack of such information in the existing literatures. Sometimes the proposed models have not been validated using “omics” data (*e.g.*, genomics, transcriptomics). These data are usually measured using high-throughput technologies and it is often desirable to integrate biological data

with the computational models. For example, the study in [12] has measured six types of cell fate data for 35 signaling proteins. It would be desirable to construct a model that can integrate such experimental data and predict the cellular phenotypic responses [13].

1.1.4 Biological networks

Once the relation among various components of a biological network is identified, the relation can be used to establish the network based on cause-effect relationship. A directed edge between two network nodes can represent the stimulus or inhibitory relationship in addition to the interaction between them. Changes in the biological networks can be induced by the external perturbations, such as by drug effects. For example, the signaling pathways inside the cancer cells can be rewired to achieve less tumorigenic behaviour and such rewired tumour cells can achieve higher sensitivity to the cell death induced by DNA damage repair mechanism [12].

1.2 Motivation and objectives

Biological systems, in general, are very complex by nature and it is often challenging to directly perform biological experiments. Therefore, it is desirable to create mathematical models which resembles as much as possible with the corresponding real biological systems. But constructing such models requires substantial amount of knowledge like reaction rate constants, initial molecular concentrations etc. Also, often times just creating a theoretical model may not be enough to emulate the actual biological system. So a mathematical model which can integrate a biological data is much sought after for understanding the biological system compared to a pure theoretical model (which represents a highly simplified system).

Signal Transduction Pathways are very important in deciding the cell fate. They represent highly inter-connected sets of signaling molecules. For example, binding of an extracellular ligand (Epinephrine) on trans-membrane receptor (*e.g.* G-Protein) can activate the phosphorylation cascade. This induces signaling inside the cell that eventually reaches inside the nucleus or other target site. To release glucose from glycogen, ligand Epinephrine binds to the cell membrane receptor and induces a signaling cascade

inside the cell. This process eventually leads to the release of glucose from glycogen. As another example, binding of the the Tumor Necrosis Factor (TNF) to the the Tumor Necrosis Factor Receptor (TNFR), and binding of the FasL to the FasR can induce a signal transduction in a cell which may drive the cell to survive or its death by either necrosis or apoptosis [14].

To understand the mechanism of cell fate development and drug effects, gaining deeper insights into the functioning of signaling transduction pathways is very important. However, the details regarding signaling mechanisms are still far from clear. For example, the study in [12] conducted various experiments for three cancer cell lines to investigate and understand the roles played by 35 signaling molecules (independent variables) on six types of cell fate (dependent variables) decisions. They experimentally measured the data (activity levels of the proteins) and applied data-driven methods to investigate the functions of various signaling proteins in apoptosis. They used modeling techniques such as PCA, PLS and VIP scores to investigate the hidden patterns in the experimental data. They found that in some of the breast cancer cell lines, four signaling proteins can predict the apoptosis. In this thesis we further analyse the impact of signaling proteins on apoptosis using our novel computational model that is based on ordinary differential equations. By using our ODE based model, we analyse the context-specific interplay of various signaling molecules in the context of apoptosis. A major objective of this thesis is to combine data-driven and knowledge-driven methods, thereby proposing novel hybrid modeling approaches for understanding the impact of anti-cancer therapies at systems level.

1.3 Contributions

Some of the important contributions from this thesis are:

- We propose an intuitive differential equation based computational model to investigate the roles played by essential regulatory proteins in the event of programme cell death. We assess the correspondence between the experimental apoptotic data, and the *in silico* data derived from the computational model. We uncover rewiring

inside the tumour cell using the cell death data. We iteratively compare our simulation results with cell death data using genetic algorithm and improve our model based on comparison results. Our uncovered network rewiring, developed as per the interaction of various regulatory proteins involved in the Apoptosis, is supported by existing literature.

- We further propose a novel scalable computational model based on rate equations dynamics that can predict context-specific apoptosis based on model calibration with experimental data. We derive ODEs for the rate of change of signaling molecules based on the general modeling approach of phosphorylation and unphosphorylation driven by kinases and phosphatases, respectively. The model rate constants are calculated using Bayesian framework. For each of the three cell lines under study, virtual caspases knockdown experiment is performed and comprehensive analysis of the result is provided. The context-specific pathways alteration under the treatment with compound Erlotinib is also analysed.

1.4 Outline

There are totally five chapters in this thesis including this “Introduction” chapter. *Chapter 2* provides a review of some of the important works done in recent years in the direction of cell fate modeling. This includes a summary of both continuous and Boolean models for cell fate modeling. Some of methods for analysing the biological data is examined in the context of “data-driven” modeling. *Chapter 3* explains the continuous computational model to uncover rewiring inside the tumour cells based on biological data taken from high throughput laboratory experiments. *Chapter 4* explains an ordinary differential equation based model to perform context-specific analysis using signaling pathways. The mechanism of integrating data into the model is also explained. Finally, *Chapter 5* summarises the current work and provides an outline of future research.

Chapter 2

Related works

Cancer is the common name assigned to a group of more than a hundred devastating diseases [15]. Despite such a huge variety, each type of cancer develops due to untreated malignant cells. Among all types of cancer, malignant breast tumours are reportedly one of the most devastating carcinoma for women. There were 1.68 million cases and 522,000 deaths due to breast cancer in 2012 [16]. Therefore, it is desirable to investigate the cancer causing factors.

However, cancer is a complex disease that contains a hallmark of cancer. For example, uncontrolled growth of cells is a hall mark of cancer that can lead to severe diseases and death [17]. Although in the initial stage tumour cells are more stable to a particular location inside the body [15], mature cancer cells tend to move to adjacent tissues and other parts of the body, a process termed as metastasis [16]. The main cause of various types of cancer development is the changes in the genomic sequences. In the past few decades, several research breakthroughs have been achieved in understanding the cancer, with many discoveries being about finding mutations and the oncogenes. There are diverse groups of mutations responsible for the development of cancer [18]. These findings have shed light on the aetiology of cancer and inspired development of novel anti-cancer therapies [19]. However, our understanding about the biological events for cancer-causing mutations is still limited.

This chapter provides a literature survey of several important aspects of systems biology that are closely related to the computational methods in cancer systems biology proposed in this thesis.

2.1 Signaling pathways

Programmed cell death, also known as apoptosis, is regulated by very complex and non-linear interactions of signaling molecules. These interactions, *e.g.*, activation, inhibitions, complex formation etc., can be represented in the form of signaling pathways. Generally, in a signaling pathway, a node represents a gene or a protein, and the directed edge connecting two nodes represents the type of interaction (*e.g.*, activation). The signaling pathways, even inside a healthy cell, are usually very complex in nature. In the case of a cancer cell, which contains genomic mutations, the pathways tend to be more disturbed and in an atypical form. Construction of models for such biochemical pathways is usually very important and prerequisite for computer simulation of biological activities [20]. There are several databases that contain models of signaling pathways. These databases are instrumental for users to construct new pathway models, and edit the existing models. For example, GeneGO MetaCore database [21] is such a database where a valid user can specify the name of proteins, and construct the signaling pathways containing all the specified proteins. However, GeneGO does not allow users to encode computational models directly (*e.g.*, encoding ODEs). Also, context-specific model parameters (*e.g.*, reaction rates) are not available, and a platform to perform dynamic simulation is not yet available. The BioModels database is another repository of the curated and non-curated models of signaling pathways [22]. The database allows dynamic simulation, free download of the pathways (in SBML format), and provides curated models parameters. However, the overall number of curated models in the BioModels database is still small, though it has been increasing gradually. The downloaded SBML model from the BioModels can be simulated with several software, such as, CellDesigner [23]. KEGG Pathway Database is another source for curated signaling pathways [24]. The Pathway Commons database is another source of signaling information. The database provides several free facilities for investigating pathways, such as, searching, visualization, and download [25].

2.2 Omics data

With the advancement in technology it is possible to identify genomic alterations inside tumour cells on larger scale. For example, the Single Cell Technology (SCT) is capable

of uncovering the changes within a single cell. SCT can be used to study the mutations inside the tumour cells. For example, the study in [26] used the single cell data to construct the mutation tree. Their method (which is quite different from conventional phylogenetic tree) directly describes temporal relations between mutation sites while handling sequencing errors. The study inferred arrangements of the mutational sites based on single cell data which can be helpful in understanding relationships among diverse mutations and how they cause tumour growth.

However, investigating the mechanisms that govern the cancer cell fates requires understanding of different types of high throughput “omics” data. There are various online repositories to obtain the “omics” data. For example, Gene Expression Omnibus (GEO) is a public database for the genomic data. GEO provides several tools to perform online query and download curated genomic data [27]. The UniProt database is an invaluable resource for finding high-quality protein sequences and curated protein domain related functions [28]. The Cancer Genome Atlas (TCGA) provides high-throughput genomic mutation information based on techniques, such as gene expression profiling, which are causal for cancer [29]. The Catalogue Of Somatic Mutations In Cancer (COSMIC) database provides curated somatic (non-germ cells) mutation data for different types of human cancer (*e.g.*, Triple-Negative Breast Cancer (TNBC)) [30].

In addition to the aforementioned data resources, there have been several biological studies trying to demystify the complexity in the cell signaling pathways by measuring and investigating the high throughput data. For example, the study in [12] conducted biological experiments for the namely, BT-20, MDA-MB-453, and MCF7 cell lines and found that dose-dependent treatments of breast cancer cells can make the tumorigenic cell less tumorigenic. They performed experiments using various combinations of genotoxic drugs and signaling inhibitors. The study provided high throughput data for six types of treatments (DMSO, TAR, DOX, DT, T-D, and D-T). It was found that after time staggered treatments of the tumour cells by Erlotinib and Doxorubicin, the cancer cells became more susceptible for DNA damage induced apoptosis. This study suggests that the rewiring inside the tumour cell might be able to reverse the effect of some of the mutations (that initially may have led to the tumour growth). For our study, we used an experimental single-cell dataset, published by the group of Michael B. Yaffe [12].

2.3 Computational modeling approaches

The huge amount of biological data, for example the data generated by Cancer Genome Anatomy Project, requires carefully designed computational analysis [31]. Computational biologists have been actively investigating the biological data by modeling and simulation to understand the biological systems. There have been several studies for *in silico* modeling of the cancer, which are often focused on uncovering the complexity and crosstalk in the biological networks. Since signaling pathways in the cancer cells are much more complex due to inherent mutations, it is often difficult to analyse the regulatory mechanisms in the tumour cells. With the advancement of technology, there has been several whole-genome sequencing data for the tumour cells available, providing unprecedented opportunity for understating the tumour developmental processes [32]. However, despite all these efforts, the regulatory mechanisms of apoptosis (especially in cancer cells) remains elusive [33]. It is still difficult to model and simulate the dynamics of cellular signaling pathways partly due to insufficient understanding of the biochemical kinetics at systems level. To address these challenges, researchers in computational systems biology have developed *in silico* systems modeling approaches to gain deeper insights into these pathways. These modeling approaches can be broadly classified as knowledge-driven and data-driven modeling.

2.3.1 Knowledge-driven dynamical modeling

Numerous knowledge-driven dynamical models have been proposed to study cell fates. Through quantitative modeling, knowledge-driven methods can be used to investigate qualitative behaviors of complex biochemical networks. One important type of methods employed to study the interactions among molecules is discrete models. For example, the study in [14] focused on discrete modeling of an apoptosis network. They initially constructed a model for cell fate prediction with nearly 25 important regulatory genes (*e.g.*, CASP3, BCL2, XIAP etc.). The study combined the cell survival pathway with cell death (*i.e.*, necrosis and apoptosis) pathways. Based on the interaction patterns of various nodes, the study modeled three cell fates, namely, apoptosis, proliferation, and survival. Initially a network model was constructed, and then based on logical inference;

the overall number of nodes in the network was decreased from the original network. This enabled the analysis of the roles of molecules in determining cell fates. The study used GINsim [34] software for performing simulations, and analysed the importance of Cytokines (*e.g.*, TNF, FASL) in deciding the cell fates.

The study in [35] analysed the differential equation based continuous model of apoptosis triggered by the drug Cisplatin. The study initially constructed a network model of the apoptosis pathway using existing literature. The computational model was configured to respond to the external stimulus, *i.e.* the drug Cisplatin. To investigate crosstalks among various signaling pathways, the study integrated three pathways into one model (*i.e.* the apoptosis model included three main pathways), namely, death receptors (*e.g.*, FasL) induced pathway, mitochondrial pathway, and ER stress pathway. The study found that the apoptosis caused by Cisplatin was dependent on dose and time. The apoptosis level was almost stagnant for higher concentrations of the drug. They also observed mitochondrial pathways caused the strongest effect on apoptosis (out of the three pathways) and crosstalk contributed roughly 25% of overall apoptosis level. The model parameters were derived using the existing literature [35].

The study in [36] proposed a computational model consisting of 52 ordinary differential equations (ODEs), for estimating apoptosis based on the activation of the mitochondrial signaling network. In [37], another model of apoptosis was proposed to analyze the effects of fast degradation of CASP8 and CASP3 proteins on apoptosis. The study suggests that a rapid increase in the activity level of CASP3 can be sufficient for initiating apoptosis. The authors of [38] proposed a computational model combining the strengths of ODEs and Boolean models for analyzing the behaviors of the NF-kB pathways. The study in [39] used a Boolean model to analyse MAPK pathway. The authors of [40] modeled signaling pathways of TNF and EGFR proteins, which were connected to the Gene Regulatory Network (GRN) inside nucleus and the computational model was encoded using ODEs. However, downstream effects on caspases were not part of the model.

2.3.2 Data-driven statistical modeling

On the other hand, data-driven statistical modeling has allowed the quantitative modeling of the signaling pathways. The study in [41] has discussed three data-driven modeling approaches in the context of signal-transduction networks. These are:

- (i) Clustering - for organising the data.
- (ii) Principal Components Analysis (PCA) - for compressing the data.
- (iii) Partial Least Squares Regression (PLSR) - for mapping independent variables with dependent variables (*e.g.*, mapping the signaling molecules' activity levels to the phenotypic responses).

Both PCA and PLSR decompose a given data set into several dimensions as per the scores and loading vectors. In PCA, data is factored into several components where the first component captures the maximum possible variance in the data. However, one of the limitations of PCA is that it cannot be used to map the independent variables to dependent variables. PLSR takes into account the relationship between the independent variable (*e.g.*, signaling data at various time points) and the dependent variable (*e.g.*, cell fate data at various time points). Therefore, PLSR can be used to map the signaling data to the cell fate data.

Data-driven modeling of signaling pathways is a promising approach for uncovering regulatory mechanisms for cancer cell fate. For example, the study in [12] used PLSR to reveal anti-cancer drug effects on breast tumor cell lines. The authors in [12] investigated three cell lines of breast cancer, namely, BT-20, MDA-MB-453 and MCF7 for their responses to various combinations of exposure to several genotoxic drugs and signaling inhibitors. The authors found that it was the pre-treatment, rather than co-treatment or post-treatment, of a subset of TNBC cells with EGFR inhibitors that can increase sensitivity of tumour cells to the DNA-damaging chemotherapy. The study suggested that such treatments may cause rewiring in the oncogenic signaling network that can potentially make the cancer cells more susceptible to death. The rewiring involves alterations of signaling pathways, such as the addition or deletion of edges in the network, change in reaction rates, and change in the concentration of molecules.

In [42], basal and signaling profiles for different treatments in the panel of "NCI-ICBP-43" cell lines were measured. The study identified several proteins that can be potential biomarkers for enhanced drug effect on tumor cells. The authors also used the PLSR method to predict the sensitivities of cancer cells to 23 targeted therapeutics. The

significance of a prediction was assessed using correlation between predicted and measured GI50 values. In [43], a comprehensive growth factor response dataset was analyzed to understand the roles of AKT and ERK proteins in the growth of breast cancer. The authors used unsupervised k -means clustering (where $k = 4$) of the time-series based signaling data to identify characteristics of signaling that varied across the dataset with respect to cell lines, ligands and receptors. The study in [44] provided a Bayesian approach to combining gene expression data with phosphorylation data, each corresponding to 26 stimuli, to identify network components transmitting signals for a diverse set of stimuli in rat and human cells. These data were combined with a prior network of 220 nodes and 501 edges to identify network components with an inference algorithm using Markov chain Monte Carlo (MCMC) approach. The study in [45] used the data-driven modeling approach to investigate the rewiring program in melanoma. The study shows the role of rewiring in ERK-JNK signaling pathway in Melanoma. Due to rewiring in the ERK-JNK pathway, ERK increases transcription of c-Jun oncogene. The overexpression of c-Jun is also responsible for an elevated level of cyclin D1 expression, which is frequently overexpressed in human melanoma. The study suggest that the signaling pathway in cancer cells goes through a “rewiring program”. The study in [46] suggests that the oncoproteins can rewire the signaling networks (such as in Bcr-Abl-transformed cells) to support cancer growth by uncontrolled cell proliferation and differentiation. The study, based on the data-driven modeling, suggests that the phosphorylation of the Bcr protein can induce rewiring in the signaling network, thereby expanding the signaling range of Bcr-Abl. Such expanded Bcr-Abl signaling network can strongly contribute to the malignancy.

2.3.3 Hybrid modeling

All the aforementioned modeling based studies in previous two subsections are either data-driven or knowledge-driven computational models. There are many significant differences between the knowledge-driven and the data-driven approaches [47]. For example, the knowledge-driven approaches emphasise the logic as a tool for representing beliefs in the system. The data-driven approach relies on the observed data for making claims and inferences, and as such it doesn’t use logic as a modeling tool. In a knowledge-driven

modeling, believes about a system are represented in the form of mathematical formulas. A set of formula can constitute a knowledge-driven model. In a data-driven approach, the information is contained in the observed data that corresponds to the observations in the real world. A knowledge-driven method makes predictions about potentially observable world, where as a data-driven approach makes inferences based on data from the actual events that happened in the past. Formulas in a knowledge-driven model are mutually consistent. However, in a data-driven approach, the collected data are regarded as the examples from the real world and they may not be mutually consistent in a logical way due to dissonance in the data. For example, observed data for a same phenomenon are not always consistent [47].

Although both the knowledge-driven and data-driven methods are promising to yield insights into cancer cellular signaling, it is desirable to combine the strengths of the two types of modeling approaches. The study in [48] identified the network-attacking mutations (NAMs) that could possibly rewire the signaling network. The study developed a computational tool named ReKINect to identify NAMs in diverse cancer types. The study used hybrid modeling to investigate the the NAMs by combining exome next-generation sequencing (NGS) for several cancer cell lines with quantitative mass spectrometry (MS). The study in [49] has predicted synthetic lethal genes using hybrid model approach. They performed single and double knockdown experiments on model proteins to assess the knockdown impact on apoptosis. Our own study in [50] used a hybrid modeling approach to predict rewiring happening inside cancer cells that may be causal for enhanced drug sensitivity. The study in [51] has proposed a hybrid model to assess cell growth, cell death etc., based on treatments with cytokines and inhibitors. The study measured the signaling activity levels for normal and cancerous liver cells. Based on the data, a signaling pathway was constructed that correlated intracellular activities with cellular responses (such as cell death) using a Boolean model.

The hybrid modeling method has also been used in many different areas to improve healthcare. For example, the study in [52] combined data-driven and knowledge-driven modeling to assess heart failure risk factors. The study used a sparse regression model to predict the onset of heart failure. The study in [53] used the hybrid modeling approach to identify “clinical narratives”.

2.4 Estimation of rate constants and initial concentrations

Differential equation based dynamic modeling has been used in several scientific fields of studies such as biochemistry, biophysics, and engineering. A long-standing issue with the ODE-based modeling is about estimation of model parameters. Several methods have been proposed to estimate the model parameters [54, 55, 56]. However, different methods for estimating model parameters from experimental data come with their own pros and cons. For example, the study in [35] derived parameter values for the ODEs from the literature. These parameter values were chosen from the previous experimental studies on apoptotic pathways. When the parameters could not be obtained from the literature, they were adjusted manually while ensuring consistency with output and with other parameters derived from the literature. The initial concentration values of the molecules were set to 0 or 1. The study provides evidence for justifying the usefulness of the parameters by methods such as the sensitivity analyses of the model parameters. Therefore, the lack of parameters did not affect model predictions. The study in [57] derived model parameter values (*i.e.*, rate constants) by trial and error to obtain simulation results consistent with the various properties of the programmed cell death. All concentrations are expressed in arbitrary units (*i.e.*, they are dimensionless). Dimensionless concentrations are used for the lack of experimental data to estimate intracellular concentrations of all proteins in the model. In [58], values of several model parameters were assigned from previous studies. However, many parameters were simply adjusted to achieve consistent results. For example, parameters for the MAP kinase pathway were selected by trial and error. However, the study established that the model was able to emulate the resistance to CASP3 activation with a lower concentration of p53. The model shows bistability to apoptotic stimuli for a high level of p53. In [59], several model parameters were chosen to reflect the experimentally known results. The model analysis in the study reported bistability for Bax activation for both deterministic and random models, and the bistability was found to be robust to the variations in model parameters. The study in [60] investigated parameter values yielding bistable behaviour of the model. The range of the parameters was predefined, and within the range parameter vectors were randomly

selected while ensuring uniform distribution of the parameters. The model was tested for each parameter vectors for the presence of bistability.

The Bayesian framework has been widely used as a flexible approach to infer model parameters [61]. Often times, some information about the model parameters may be already available (such as from the previous studies) before conducting current experiments. On one hand, Bayesian framework provides a modeller an opportunity to include prior knowledge about the model parameters; on the other hand, Bayesian framework also provides an estimation of the uncertainty about the inferred parameters. The uncertainty in the unknown parameters after observing the data are estimated in the form of posterior distribution. Theoretically, the posterior distribution is expressed in the form of Bayes' formula [62] by combining the prior information with a likelihood function as:

$$\textit{Posterior Distribution} = \textit{Likelihood Function} \times \textit{Prior Knowledge} \quad (2.1)$$

Mathematically, the Bayes' formula is encoded as:

$$p(\theta|Data) = L(\theta|Data) \times p(\theta) \quad (2.2)$$

where θ represents the parameter vector, $Data$ is the observed data, and L is the likelihood function that provides probabilistic explanation for the $Data$ in a formal way.

Based on Bayesian framework, several software tools have been developed to infer the model parameters. For example, the study in [63] developed the BioBayes software for parameter estimation and model ranking. The software is based on Bayesian statistical framework. BioBayes estimates model parameters using the likelihood function. However, for several complex ODE models, deriving the likelihood function is not always possible. The study in [62] developed a novel parameter estimation software suite called ABC-SysBio. It is an open source python package which combines the strengths of the Bayesian framework and Sequential Monte Carlo (SMC) approaches for parameter estimation. Approximate Bayesian Computation (ABC) is a widely used technique for estimating model parameters from experimental data. A very important feature of the ABC-based technique for parameter estimation is that they are likelihood free. As such, ABC-based parameter inference can be used with problems with unknown likelihood

functions [64]. Moreover, ABC-SysBio is able to handle missing data. A critical issue for parameter inference is overfitting of model with the experimental data. However, parameter estimation methods based on Bayesian inferential techniques are promising to address the issue of overfitting [62].

2.5 Context-specific modeling

There have been phenomenal applications of “omics” data in the anti-cancer research. For example, the “omics” data have been used to develop models for predicting cell line specific drug responses in cancer patients. As the technologies, such as the next-generation sequencing have been advanced, the focus has become to provide “precision medicine” to the cancer patients [65]. There are several data sources that catalogue cell-line specific data. For example, the Cancer Cell Line Encyclopedia (CCLE) database provides a diverse set of “omics” data (such as gene expression profiles) for 947 human cancer cell lines [66]. As the new information (*e.g.*, data, knowledge, biological context) becomes available, computational models need to be re-constructed to accommodate the new context-specific information. By considering the genetic contexts of diverse cancer cells, better statistical estimation for cancer causing factors can be obtained. For example, the study in [67] has proposed CHER (Contextual Heterogeneity Enabled Regression) algorithm that can derive a context-specific drug sensitivity predictive model. CHER was designed to find the common genomic features across different cancer types. The study in [68] proposed a web-based context-specific application named Proteus that can build ODE models of molecular network. Proteus comprises of modular components which represent biological systems of different complexities. A user can choose some of the components and derive an ODE-based computational model.

However, the techniques of mapping context-specific genomic data (*e.g.*, cell lines specific mutation data) to the enhanced cancer drug sensitivity are far from complete. A drug that works for a particular mutation in one type of cancer may not work for the same mutation in another cancer type. For example, the oncogene RAF is reported to have V600E mutation in melanoma (skin cancer). The drug Vemurafenib (RG7204/PLX4732) inhibits RAF protein in melanoma [69]. However, the drug is not effective on RAF protein

with the mutation V600E in colon cancer [70]. Therefore, considering the context of the cancer cells becomes an important part of investigations pertaining to the cancer drug effects. The study in [66] reported that the model constructed using the mixed cell line data was not good at predicting drug sensitivity from a testing dataset that contained only melanoma derived cell lines data. However, a predictive model constructed using the melanoma-derived cell line specific data performed better. All these examples highlight the need for cell line specific studies that could lead us closer to the precision medicine. In this thesis, we investigate cell line specific roles played by the caspases, and demonstrate the context-specific variation in apoptosis corresponding the knockdown of key regulatory proteins.

2.6 Validation

Numerous approaches, such as data-driven modeling, knowledge-driven modeling, etc., have been proposed to make model predictions. The predictions from these models need to be validated in consistent and coherent ways. Validation of the model prediction is crucial for justifying the prediction accuracy, as well as for determining the efficacy of the model training process. Several methods have been proposed to validate the model predictions.

- (i) Resubstitution - this is the simplest method for validating the model prediction. In resubstitution the same data set is applied to train and test the model predictions. A drawback of the resubstitution is that it might lead to overfitting of the model, *i.e.*, the model performs very well on an available dataset, but might perform poorly on an unforeseen dataset [71].
- (ii) Hold-out validation - another technique for validating the model prediction while dealing with the issue of overfitting is called the Hold-out validation or Split-sample procedure. In the Hold-out validation, the entire data set is randomly split into two non-overlapped subsets, namely, testing and the training datasets [72]. However, the results of the predictions may be biased and dependent on the choice of the splitting of the training and testing datasets. Also, a dataset may be small, which makes splitting of the data not feasible.

- (iii) Cross validation - in cross-validation a dataset can be repeatedly split into n different groups and n iterations of testing and training are evaluated. For each round of the splitting, the training set may contain a larger portion (e.g., $n-1$ groups) of the data. In cross validation, the prediction accuracy is determined based on the prediction of a subset of data that was not part of the training. The overall accuracy of the prediction is decided based on average accuracy of numerous training-test partitions [73]. The study in [73] conducted model cross-validation using 8 breast cancer gene expression datasets with the focus on predicting Distant Metastasis Free Survival (DMFS) based on C-index for various pairwise combinations of the data.
- (iv) Experimental validation - all these aforementioned methods can be applied to validate the model predictions only if the experimental data are available. However, the availability of the genomic data is often incomplete, or with noise. Therefore, computational models predictions often have to be validated using novel experiments. For example, the study in [74] proposed a computational method to predict localization of proteins. The method, named ngLOC, predicted localization of proteins targeted to different types of cell organelles. The study in [75] validated the ngLOC method from [74] experimentally with the prediction accuracy of 89%, confirming the usability of the ngLOC method.
- (v) Literature validation - conducting new experiments may not always be feasible. In fact, model predictions are rarely validated experimentally [75]. Another way to deal with the lack of availability of the data is to establish predictions that are consistent with existing literature [35].

In this thesis, we have used cellular signaling response data to train the computational model. The cellular phenotypic response data, *i.e.* apoptosis, was not part of the model training. We predict the *in silico* apoptosis and assess the prediction using the Pearson correlation. We validate our findings with existing literature of experimental observations.

Chapter 3

Discovery of network rewiring for enhanced cancer drug sensitivity

3.1 Introduction

The anti-cancer therapeutics is targeted to enhance killing of cancer cells with minimum or no damage to the healthy cells. To this end, a solid understanding of the cell fate decisions (*e.g.*, apoptosis, proliferation) of different cells under various conditions would be required. The signaling pathways play very crucial roles for regulating the cancer cell fate [17]. However, it is challenging to understand the dynamics of signal transduction at systems level, due to non-linearity of the network dynamics, *e.g.*, feedback and crosstalk. In cancer cells, this becomes even more complicated due to various types of alterations (*e.g.*, DNA mutations, genome rearrangement, epigenetic changes, and pathway alterations). These alterations allow cancer cells to adapt to new conditions and evolve drug resistance. Therefore, to find effective anti-cancer therapies, cancer-specific alterations in the signaling pathways must be taken into account. Moreover, it is desirable to understand how cancer cells respond to different combinations of drugs, and how drug sensitivity can be enhanced. Genomic and proteomic data of cellular responses to drugs, synchronized with cell fate observations, would shed light on cancer drug effects at systems level. However, even if sufficient data are available, it is challenging to construct a model of cell signaling to explain the data and make accurate predictions. As more “omics” data about cancer are available recently, computational methods for modeling and discovery of cancer cell fate are becoming more important.

In this chapter, a hybrid modeling framework is proposed that integrates knowledge-driven and data-driven approaches for modeling cancer signaling pathways and cell fate decisions. From a generic apoptotic network, we simulate the time series data of signaling proteins using ODEs based on a network candidate modified from the generic network by *in silico* rewiring; then the simulated data are compared with the real data. This *in silico* rewiring, simulation and data fitting is repeated iteratively to improve the goodness of fit using Genetic Algorithm. Through such an optimization, we can detect the alterations in topology of the networks, which allows close fitting of the model to the real data. As such, network rewiring can be inferred from real data. Most of rewiring events predicted by our model are supported by existing literature.

3.2 Methods

Our work is broadly organised in three interrelated parts. We initially constructed a basic apoptosis network (say N_1) which is free of abnormal mutation and represents signaling pathways inside a healthy cell. We used both the control (DMSO) and drug sensitive data of cellular signaling and cell fate of (breast) cancer for our studies based on the phosphorylation levels of signaling proteins and phenotypic responses taken from the biological experiments conducted in [12]. Hereafter, we call this data set as “Yaffe’s data”. We compared the simulation data from this network with the Yaffe’s data using Genetic Algorithm and Dynamic Time Warping for DMSO treatment and derived a tumorigenic network (say N_2). Then we modified N_2 to fit with Yaffe’s data to derive a drug sensitive network (say N_3). N_1 was constructed directly from literature; N_2 consisted of edges added and deleted from N_1 by the Genetic Algorithm; N_3 consisted of edges added and deleted from N_2 by the Genetic Algorithm.

3.2.1 Data

For our study, we used an experimental single-cell dataset, published by the group of Michael B. Yaffe [12], to calibrate and test our computational model of signaling network (Figure 3.1). This dataset contains time-course data for both cell signaling activities (measured by phosphorylation levels) and its corresponding cell fates measured with flow

cytometry. These data are available for three cancer cell lines, i.e., BT-20, MCF7, and MDA-MB-453. Phosphorylation levels of 35 proteins were recorded at 8 time points (numbered as 0, 1, 2, ..., 7). The indices from 0 to 10 correspond to the time points of 0, 0.1, 0.25, 0.5, 1, 2, 4, 6, 8, 12 and 24 hours, respectively, after the induction of the drug to the cancer cells. Out of the 35 proteins, however, data of only 32 proteins were available in the published dataset. The cell fate data were measured for 6 cellular phenotypes (i.e., Apoptosis, Proliferation, G1, S, G2 and M) for the three breast cancer cell lines over 24 hours at various time points corresponding to a subset of the aforementioned experimental time points (i.e. 0, 6, 7, 8, 9 and 10). The cell signaling data were measured at 8 time points (numbered as 0 to 7), corresponding to the experimental time of 0, 0.1, 0.25, 0.5, 1, 2, 4, and 6 hours, whereas data for the five types of cell fates (i.e., Apoptosis, G1, S, G2 and M) were measured at five time points (numbered as 0, 6, 7, 8 and 9) corresponding to the experimental time of 0, 4, 6, 8 and 12 hours. The data for the last two time points (numbered as 8 and 9) for the experimental time of 8 and 12 hours are available only for the measurement of the cell fates, whereas three time points (numbered as 0, 6 and 7) corresponding to the experimental time of 0, 4, 6 hours are common for the measurement of both signaling and cell fate data. The data corresponding to the time point 10 was available for only one of the cell fate, i.e. cell proliferation. The measured signaling and cell fate data are synchronized with respect to the time, e.g., the aforementioned time point 6 refers to the same physical time of 4 hours after the drug induction for both signaling and cell fate measurements. Six types of drug treatments were applied to the cancer cells, i.e. Dimethyl sulfoxide (DMSO), Erlotinib (TAR), Doxorubicin (DOX), Doxorubicin and Erlotinib together (DT), Doxorubicin followed by Erlotinib (D-T), and Erlotinib followed by Doxorubicin (T-D); then the phosphorylation levels of signaling proteins and cell phenotypes in response to the treatments were measured. When the researchers want to treat cancer cells with a chemical (e.g. an anti-cancer drug), the chemical first needs to be dissolved in some solution. However, not all chemicals are soluble in the water. When a drug (e.g. Erlotinib) is applied to the cancer cells, a solvent called Dimethyl sulfoxide (DMSO) is used because the drugs, such as Erlotinib, are soluble in DMSO. However, the DMSO solution itself can affect the cancer cells and a bias can occur in the measured signaling and cell fate data. To avoid such biases, we

need to know the signaling and cell fate data for the DMSO treatment. The DMSO treatment is referred to as the control treatment. In the published data set by [12], the control treatment data was measured after induction of the Dimethyl sulfoxide (to avoid any a bias induced by the DMSO solvent). The DMSO treatment had no anti-cancer drug. The Drug treatment (by Erlotinib) data refers to the data in the drug sensitive cancer cells (after drug is applied). The study [12] reported that the level of apoptosis could increase by as much as 500% in response to the T-D treatment compared with the other treatments. Henceforth, this dataset is referred to as the “Yaffe’s dataset”. The study in [12] employs several mathematical modeling approaches to relate signaling data to cell phenotypes. Principal component analysis (PCA) was used to identify co-variation between signals, whereas partial least-squares regression (PLSR) was used to identify co-variation between molecular signals and corresponding cellular phenotypic responses. In both PCA and PLSR modeling, the input vectors contained quantitative measurements of the signaling proteins.

3.2.2 Apoptosis network construction

We constructed a generic apoptosis network consisting of 22 signaling proteins including 18 proteins from the biological experiment conducted in [12]. The authors of [12] provided Variable Importance in Projection (VIP) scores for different selected proteins to demonstrate how the selected proteins were useful for predicting apoptosis of different cell lines. For example, proteins with VIP score > 1 were important variables, proteins with VIP score < 0.5 were unimportant variables. The VIP score plot in the study by [12] has been given for drug treatment data. For example, after drug treatment Casp8, Casp6 etc. were reported to be the most important variable (with VIP score > 1) whereas EGFR, HER2 were among the least important variable (with VIP score < 0.5). However, considering that this trend could be different in the cancer cells (without any drug treatments), we selected a mix of 18 protein set that included Casp8, EGFR, etc. Since our goal is to test the goodness of fit of our model with Yaffe’s data, we selected proteins that have been measured in the experiment. Also, our signaling network includes molecules based on

Figure 3.1: Basic apoptosis network and various rewiring events. Basic apoptotic network contained all edges except the broken edges (green and blue). Edges represented by thick red colour were deleted by Genetic Algorithm while inferring N_2 from N_1 . Broken edges represented by green colour were inserted by Genetic Algorithm while inferring N_2 from N_1 . Edges represented by magenta colour were deleted by Genetic Algorithm while inferring N_3 from N_2 . Broken edges represented by blue colour were inserted by Genetic Algorithm while inferring N_3 from N_2 .

the prior knowledge from several pathways playing critical roles in inducing cell death, which include DNA damage response and cell death pathways, stress response pathways, etc.

The network diagram is shown in Figure 3.1. We constructed connections between various signaling proteins in our basic apoptosis network using the existing literature. HER2 is a cell transmembrane receptor which plays a significant role in breast cancer by suppression of tumour growth. It provides growth signals in PI3K-AKT and Ras-MAPK

pathways. HER2 also inhibits p53-induced cell death by stimulating MDM2 and AKT. p53 promotes cell death by sensing signals for DNA damage and it can stimulate PUMA and BCL2 proteins that are pro-apoptotic and inhibit IAPs (*e.g.*, survivin) [76]. PUMA is a pro-apoptotic protein and plays crucial roles in apoptosis. Tumour suppressor gene p53 regulates the expression of PUMA. Once activated, PUMA frees the apoptosis regulator Bax and Bak owing to its high affinity with BCL2 oncogenic group of proteins. The freed proteins further transduce the apoptotic signals to the mitochondria. This causes the Caspases to be activated and the activated caspases eventually lead to the cell death [77]. Activation of pro-apoptotic BAX/Bak facilitates heme protein cytochrome c being released from the inner mitochondria membrane [78][79]. The cytochrome c further stimulates CASP9, which then stimulates CASP3 and Casp7, which eventually lead to apoptotic cell death [80]. Up-regulation of DAPK induces apoptosis by enhancing response of tumour suppressor p53 whereas down-regulation of DAPK reduces the response of p53 to several oncogenes including c-Myc and E2F-1. The activated form of DAPK further activates the Beclin-1 by phosphorylation. Beclin-1 activation is crucial for Autophagy and also for its crosstalk with apoptosis.

Beclin-1 is an autophagy molecule that reacts with the BCL2 family signaling molecules to form BCL2:Beclin-1 complex that facilitates a switch from autophagy to apoptosis and vice-versa. The formation of the complex causes inhibition of Beclin-1 by BCL2 protein [81][82]. STAT3 is an important gene for regulating cell fate. Experimental results have shown that a decreased level of STAT3 leads to a decreased level of BCL2 [83]. Activation of STAT3 gene stimulates transcription of quite a few cell fate controlling genes, which include quite a few proteins from the BCL2 [84]. The Wee1 and Cdk1 (Cyclin-dependent kinase 1) proteins regulate the cell cycle. During mitosis, the size of a cell is controlled by the inhibition of Cdk1 by Wee1 which prevents the cell from going into mitosis. The suppression of Cdk1 by Wee1 is done by phosphorylation of Cdk1 [85]. Cdk1 in turn phosphorylates BCL2 and suppresses its activity [86]. The EGF receptor (EGFR) is a cell membrane protein receptor with high affinity for several specific proteins (ligands) including Epidermal Growth Factor (EGF). EGFR further activates many targets including Jun N-terminal kinases (JNKs). Mutation in EGFR may lead to its uncontrolled expression which can potentially cause uncontrolled cell growth, which is a hall mark of

cancer [87]. The BCL2-associated death promoter (BAD), which is highly pro-apoptotic protein, is inhibited by JNK as JNK can phosphorylate BAD protein. This leads to the inhibition of apoptosis [88]. Trimeric Tumor Necrosis Factor (TNF) activates multiple signaling pathways which in turn lead to apoptosis. TNF interacts with death domain (DD) protein TRADD. TRADD in turn has been shown to have high affinity for RIP [89]. CASP3 mediates cleavage of multiple proteins and causes DNA fragmentation which eventually leads to apoptosis [14].

3.2.3 Partial least squares regression

The Partial Least Squares Regression (PLSR) is a computational technique to estimate linear relation between the independent (e.g. cell signaling data) and depend variables (e.g. cell fates). The independent and independent variables are represented in the form of matrix and both the variables are projected to new spaces. PLSR is suitable when number of independent variables is more than the response variables [90]. The general formulation for PLSR modeling is given as

$$A = BC^T + D$$

$$E = FG^T + H$$

Where A is an $n \times m$ dimensional matrix containing independent variables. E is an $n \times p$ dimensional matrix containing dependent variables. B and F are the vectors of $n \times l$ dimensions representing projections of A and E . C and G are loading vectors of dimensions $m \times l$ and $p \times l$. The vectors D and H are randomly distributed error terms [91].

3.2.4 Simulation with ODE

Using the topology of the basic apoptosis network (Figure 3.1), constructed from the literature, we derived a computational model by generating rate equations for the network. The complete set of differential equations for the basic network N_1 is given in Table 3.1. Concentration of each molecule is represented by enclosing the molecule name with square brackets (e.g., [CASP8]). Reaction rate constants are represented as rc_1 , rc_2

etc. For example, based on the basic network N_1 in Figure 3.1, the differential equations for the change of concentration of AKT and p53 were derived as:

$$\frac{d}{dt}AKT = -[AKT] \cdot rc_{18} + [HER2] \cdot rc_{20}, \quad (3.1)$$

$$\frac{d}{dt}p53 = -[p53] \cdot rc_{10} + [AKT] \cdot rc_{18} - [p53] \cdot rc_{27}, \quad (3.2)$$

where rc_{18} is the binding rate constant from AKT to p53 and $[AKT]$ is the concentration of the AKT molecule. Also, rc_{20} represents the binding rate constant for HER2 to AKT. Similarly, for p53, variables in the corresponding equation incorporate information for the molecular concentrations and the binding rate constants for AKT and p53.

In our study, the ODEs were solved using the standard Matlab library function *ode45*. This function can interpret the differential equations in the form of rate reactions. The *ode45* accepts the initial values of the molecules and the time duration in the form of function parameters [23][92]. Using this procedure, we derived the rate equations and time course data for all twenty-two signaling proteins. The initial concentrations of the signaling molecules were set to dimensionless numeric value of one [57]. Some of the rate constants were taken directly from the literature. These include CASP9 & CASP3, CASP9 & XIAP, CASP8 & BID, CASP3 & XIAP, CASP3 & CASP8, CASP3 & Apop, p53 & PUMA, CASP3 & CASP6 and XIAP & SMAC [35][93][94]. The remaining parameters were generated randomly but we verified them by Partial Least Square Regression (PLSR) test. We also performed parameter sensitivity analysis by varying them to ensure overall model robustness for the small variations of parameters (see result section for PLSR test and sensitivity analysis).

3.2.5 Genetic algorithm

In order to uncover the rewiring inside the tumour cells, we applied the Genetic Algorithm to search for the network topologies that improves the goodness of fit with the real data. There are several optimisation techniques available in the literature which can be used to search for the network topologies. These techniques include, for example,

Simulated Annealing, Evolution Strategies, Genetic Algorithm, Evolutionary Programming etc. Simulated Annealing has population size of only one which represents current solution, thus the solution space is not as wide as GA. Similarly, Evolutionary Programming can only use mutation operator. By contrast, Genetic Algorithms have several solutions in the beginning and crossovers can be performed to generate the diversity in the population of network structures. While the probabilistic approach by GA helps us simulate the stochastic biochemical networks in better ways, Evolution Strategies uses deterministic selection [95][96]. After comparing these techniques, we decided to use Genetic Algorithm in our study. As a global search technique, GA is considered to be very robust as it makes very few assumptions about the problem under consideration [97].

The implementation of Genetic Algorithm assumes that the variables from problem can be represented in the form of binary strings typically known as Chromosomes. A possible solution is also encoded as a chromosome. The initial population for the Genetic Algorithm is a random set of chromosomes. These chromosomes are evaluated using an objective function. The chromosomes representing better solutions are given higher probability to reproduce offspring in next generation. Chromosomes with lower scores are removed from the solution set. The pseudocode for the Genetic algorithm used in our experiments is given in Algorithm 1.

The network topology is represented by an adjacency matrix. For each edge between two nodes in the network there is a Boolean value of ‘1’ in the matrix, and a ‘0’ entry represents the absence of connection between the corresponding nodes. The adjacency matrix can be modified in Genetic Algorithm as follows [97]. We initially derive the chromosome (vector of 1×484) by merging the rows of the adjacency matrix, corresponding to the signaling pathways, into one binary string. As such, the network topology of 22 nodes is represented as a chromosome i.e. a binary string of length 484. The topology of the basic network N_1 was changed randomly using mutation and crossover. The crossover points were randomly selected uniformly from the range of [1:484]. The networks to be crossed over are selected randomly. Generally, the mutation rate for any particular bit is less than 1% [97]. In our implementation the mutation operator was applied to each of the bit in a chromosome (after performing crossover) with a low probability value of 0.01. Once the process of selection of chromosome, recombination and mutation is finished, the population of next generation is evaluated [97].

3.2.6 Dynamic time warping

To compare the time course data, we used Dynamic Time Warping (DTW) [98][99]. The DTW is a technique for assessing optimal match between two given time series data. It performs non-linear mapping of two time course data while optimizing the goodness of fit. In addition to this, DTW also has the strength to capture the variation in data that may vary owing to different speed. Hence, DTW can be used to estimate the optimal match between two temporal data. The objective function of DTW, for comparing two time series datasets s and t is given as:

$$D[i, j] = cost + \min \begin{cases} D[i-1, j] \\ D[i, j-1] \\ D[i-1, j-1] \end{cases}, \quad (3.3)$$

where $cost = dist(s[i], t[j])$, D is a matrix of size $[0..n, 0..m]$, $D[0, 0] = 0$. D is an array that contains pairwise distances between two time series data. The entries of $D[0..n, 0..m]$ are computed based on the data contained in two time series variables called $s[1..n]$ and $t[1..m]$. Initially each entry of $D[i, 0]$, for $0 < i \leq n$, and $D[0, j]$, for $0 < j \leq m$, are set to infinity, *i.e.* a larger value which cannot be part of the data sequences. Here $s[1..n]$ and $t[1..m]$ are the two time series representing protein phosphorylation levels over time points to be compared, whereas $s[i]$ and $t[j]$ represents the data at time points i and j respectively. For any two numbers $s[m]$ and $t[n]$, $dist(s[m], t[n])$ represents the distance between the numbers, e.g., $dist(s[m], t[n]) = |s[m] - t[n]|$. Here the objective function accepts the simulation data and the experimental data as input and estimates the goodness of fit between them.

During the execution of genetic algorithm, we check for the network convergence which is basically a test criterion for the termination of the genetic algorithm. For example, when a new population representing a new network topology is obtained (using crossover and mutation), we calculate the DTW score for the new network and compare it with the DTW score from the previous network. If the two scores are similar (say difference < 0.001), then we output the solution set.

Algorithm 1 Genetic Algorithm for finding alterations in apoptotic network

INPUT: Objective function, Network Topology (NT) in the form of Adjacency matrix of size $n \times n$, reaction rate constants matrix of size $n \times n$, upper limit allowed for the number of generations *Max_num_gen* for the algorithm, biological data measurements.

OUTPUT: A matrix consisting topology of best uncovered network, score matrix *S*, for such uncovered networks.

```

1: procedure GENETICALGORITHM
2:    $NT_0 \leftarrow$  topology of network
3:    $RC_0 \leftarrow$  rate constants
4:    $Max\_num\_itr \leftarrow$  total number of iteration allowed
5:    $S_0 \leftarrow 0$ 
6:   top:
7:     Derive Population P (randomly generated networks)
8:   loop for max_num_gen
9:     Derive rate equations for each network in P
10:    Solve each of the Rate equations
11:    Derive numerical solutions (time series data)
12:    Compare the simulation results with Yaffe's data using DTW objective function
        (refer to the Section 3.2.6) and calculate the Score for each network in P

$$D[i, j] = cost + \min \begin{cases} D[i - 1, j] \\ D[i, j - 1] \\ D[i - 1, j - 1] \end{cases}, \quad (3.4)$$

13:     $S \leftarrow$  Select 50 networks with best scores
14:    Perform crossover among best selected networks to formulate next generation,
        total of 100 networks again.
15:    Perform mutation for each bit with the probability of 0.01, for each of the hundred
        networks.
16:    Set new population P = mutated network from the previous step.
17:    Check for the convergence.
18:    if network not converged then
19:      goto top.
20:    else
21:      Output solution set S.
22:    close.

```

3.2.7 Sensitivity analysis

Relying on the model parameters derived from the independent literature sources come with its own set of issues. For example, a model parameter can be sensitive to the environmental conditions, such as the ambient pH , temperature, etc. Therefore, it is desirable to infer the best model parameters that can help in predicting behaviour of the real biological systems. The sensitivity analysis (SA) of a model is the investigation of model behaviour and its predictions with respect to the changes in the model parameters. SA is one of the most widely used techniques available to the modellers to assess decision making the models. [100]. One of the common ways to perform SA is to induce perturbation in the model parameters and assess the impact of such perturbations in the model predictions. For example, the study in [35] performed SA for the rate constant for an ODEs based computational model. The model rate constants were perturbed by $\pm 30\%$ and the impact on model predictions were estimated. The study found that the model was sensitive to some of the parameters as the perturbations in corresponding parameters led to a higher level of change in the model predictions.

3.2.8 Simulated annealing

Simulated Annealing (SA) [101] is used to search for the best topological configurations that can generate simulated data fitting the experimental data with a minimum distance (or error). SA is considered to be a very robust optimization technique because its chances of failing with the complex functions are relatively low. In SA, next solution is determined from the current solution. If the newly derived solution is better than the previous one, the new solution is accepted. However, if the new solution is found to be worse than the previous solution, the solution is not rejected directly; rather, the acceptance of the newly found solution is decided based on a probability function. This allows SA to accept the sub-optimal solution as well. As such, SA is unlikely to be trapped in a local optimum. This increases the overall probability of SA finding the global optimum solution [102]. In SA, we allow the system to cool from a very high temperature to a low temperature. The probability of accepting a non-optimal solution is greater at a higher temperature. As the system temperature goes down, the probability of accepting worst solution gets reduced. The pseudocode for the SA is given in the Algorithm 2.

Algorithm 2 Algorithm for Simulated Annealing (SA)

```
1: State  $S \leftarrow S_0$ 
2:  $T \leftarrow Temperature$ ;
3:  $CR \leftarrow CoolingRate$ ;
4: while  $T > 1$  do
5:   State  $S_{new} \leftarrow Random\_neighbour(S)$ 
6:   if  $(P(S, S_{new}) > Rand(0, 1))$  then
7:      $S \leftarrow S_{new}$ 
8:    $T = T * (1 - CR)$ ;
9: Output  $S$ 
```

3.3 Results

3.3.1 PLSR for simulation data from N_1

We first used Partial Least Squares Regression (PLSR) to test if the simulation data generated using our basic network N_1 were reliable for performing the prediction of network rewiring. PLSR is used for mapping independent variable with dependent variables (e.g., mapping signaling molecules to phenotypic responses) [41]. In a PLSR plot, if a predictor variable is closely associated with the predicted variable, then such a predictor variable influences the predicted variable relatively more strongly than the predictor variable who is far away in the PLSR plot. For reaction parameters, we ensured that the selected parameter values enabled us to associate the important predictor variables like Caspases, XIAP, SMAC etc. closely with the apoptosis. The Partial least Square Regression (PLSR) plot for the simulation data from the basic network from N_1 , is given in Figure 3.2, which shows the close association between apoptosis and several apoptotic proteins. For example, as per the Apoptosis Curation Manual from the Gene Ontology (GO) [103], the apoptosis process is very closely regulated by two types of caspases, namely, the initiator caspases (*e.g.* Casp9) and the executioner caspases (*e.g.* Casp3, Casp6). The Gene Ontology terms GO: 0097190 and GO: 0097194 provide the detailed functions of the initiator and executioner caspases, respectively. For example, to initiate apoptotic cell death, initially only the initiator caspases are activated. The execution of apoptosis starts with the activation of the executioner caspases. This result is consistent with the experimental findings in [12].

Figure 3.2: PLSR plot for simulation data for apoptosis from network N_1 .

3.3.2 Inferring the tumorigenic network N_2

From network N_1 , we will infer the rewired network N_2 corresponding to cancer cell lines by fitting simulated data to the cancer specific dataset. To this end, we adopt Genetic Algorithm for multiple generations to search for the most likely network rewiring events and collected 50 rewired networks in each generation along with corresponding scores. The Genetic Algorithm performed rewiring on the basic apoptotic network of Figure 3.1. The comparison for simulated data from network N_1 and Yaffe's data for control treatments using Genetic Algorithm, before and after rewiring event in network N_1 , is shown in Figure 3.3(a) and Figure 3.3(b).

Due to inherent randomness in Genetic Algorithm, the edges added and deleted in N_1 (while inferring N_2) were different among different run of simulations. Some of the changes in the network were found to be more frequent compared to other. So, just few runs of simulation couldn't be relied upon to find the changes in the N_1 network causing cancer. Therefore, to find consistent (or conserved re-wirings) changes to the network, we performed the simulation for 150 times for inferring N_2 from N_1 and another 150 rounds of simulation for inferring N_3 from N_2 . The changes which occurred in the network with the highest frequency were selected to construct N_2 (and subsequently N_3). The most frequently removed edges in the N_1 network while inferring N_2 were SMAC-XIAP,

CASP8-CASP3 and p53-PUMA. For 150 rounds of simulations, each of these deletions occurred with a frequency of 146, 143 and 141 respectively. The network edges found to be inserted in the N_1 by the Genetic Algorithm with the highest frequencies were XIAP-CASP8 and TNFR-Stat3, with frequencies of 149 and 146.

We verified the various rewiring events using the existing literature and confirmed the validity for the above mentioned network changes. p53 is a tumour suppressor and controls the regulation of PUMA. PUMA induces apoptotic signals inside the cells. Less activity of PUMA leads to apoptosis deficiency which in turn leads to the increased risk of cancer [104]. We also realised that SMAC down-regulates XIAP and XIAP down-regulates CASP3, a pro-apoptotic protein. So if SMAC is working properly, there will be more cell death leading to less tumorigenic behaviour by cells. But if SMAC is not able to down-regulate XIAP, there will be less apoptosis and so more chances of cancer [14]. We also came across literature evidence highlighting the critical role of CASP8 for affecting breast cancer directly [12][105]. STAT3 interacts with various molecules involved in programmed cell death, regulating their functions including MOMP formation, that causes release of the cytochrome c [106]. XIAP inhibits the proteolytic processing of CASP8. In fact, proteolytic processing of several caspases (*e.g.*, CASP9 is not allowed in the presences of XIAP. The lack of processing of CASP8 and CASP9 allows them to interact with CASP3 and promote apoptosis [107]. TNFR activates the Stat3 signaling. The stimulation of TNFR leads the phosphorylation of the Stat3 and Stat5b and the phosphorylated molecules are trans-located into the nucleus [108].

3.3.3 Inferring the drug sensitive network N_3

After inferring N_2 from N_1 , our aim was to infer the drug sensitive network (N_3), by fitting modified topology of N_2 to the real data of drug treatment. Before predicting rewiring of N_2 , we plotted the N_2 based simulation data with the real dataset of signaling and cell fate of cancer cells after treatment with drugs. The plot is shown in Figure 3.3(c). We again used Genetic Algorithm to search for the network rewirings that allow better fitting of simulation to the real data. For each run of simulation, we compared the simulation data with the drug sensitive experimental apoptosis data for ERL-DOX treatments. The best network plot, uncovered using Genetic Algorithm, is shown in Figure 3.3(d). The edges

of N_2 that were most frequently removed from the network were BIM-CASP9 and EGFR-CASP8, deleted 141 and 137 times, respectively. The network edges inserted with highest frequencies by the Genetic Algorithm were CASP8-RIP and CASP8-MOMP, inserted 144 and 135 times, respectively. These rewiring events have been mentioned in related literature [14][12][87][109]. We also confirmed the validity of the newly added edges with other protein knowledgebase, such as Reactome [110], UniProt [28], and GeneGO MetaCore [21] databases.

3.3.4 Sensitivity analysis of the model

To analyse the behaviour of our model with respect to the fluctuations in model parameters, we verified our model by performing sensitivity analysis for the rate constants. We initially tested apoptosis level for each of the molecules by reducing the parameter by 30%. Then we increased the parameter by 30%. Among all the parameters, parameters for the edges “CASP3-CASP8” and “CASP3-Apop” were found to be most sensitive parameters as they induced maximum impact on apoptosis level. We found that the level of apoptosis for most of the parameters changed only slightly, within $\pm 5\%$, except for the most sensitive parameters, “CASP3-CASP8” and “CASP3-Apop”, which affected the level of apoptosis by $\pm 20\%$.

We also performed the perturbation test to assess the robustness of our model. We changed the initial concentration of each of the molecules and identified the most sensitive parameters of the model. The apoptosis level was predicted for each of the molecule whose initial concentration was reduced by 30%. Then, for each molecule we increased the initial concentration by 30%. We found that the level of apoptosis for most of the initial concentrations was stable, except for the concentration of the molecules from the edges “CASP3-CASP8” and “CASP3-Apop”. This suggests an overall robustness of our model.

3.3.5 Performance of genetic algorithm with varying mutation probability

The mutation operator in GA is applied after the crossover has already happened. In order to maximize the preservation of the effects of the crossover functions, the effects of

mutation should be minimized. This suggests that the mutation operator shouldn't be used. However, after the execution of the GA for several generations, the selection may place all the bits at certain fixed some positions. As such, the algorithm may prematurely converge. Moreover, without using a mutation operator, it's impossible to reintroduce a missing bit from the chromosome. In conclusion, we use the mutation operator, but we keep the mutation probability low. For example, the study in [97] suggest using mutation operator with a probability of 0.01. However, performance of the GA based optimization can depend on selection of mutation probability. So we evaluated the performance of the genetic algorithm with varying mutation probability for deriving N_2 from N_1 . The performance of GA with mutation probability as 0.01, 0.05, and 0.1 been shown in Figure 3.4. The DTW based goodness of fit score for the mutation probabilities as 0.1, 0.05, and 0.01 were 1.13, 0.43, and 0.024, respectively. The simulation results suggest that the lower mutation probability in the Genetic Algorithm can be preferred to assess the goodness of fit. The crossover function is basically a convergence function that tries to bring the population towards a maxima or minima. The mutation operator is divergence operation as it tries to disrupt the population from reaching maxima/minima. With a lower mutation probability the effect of crossover function is preserved in a better way as compared to a higher mutation probability. Our findings also suggest that the GA can converge at a low mutation probability.

3.3.6 Comparison of GA and simulated annealing derived rewiring

We used Simulated Annealing (SA) [101] derived rewiring to compare the GA based rewirings. In our study, the input for SA was a vector containing topology from the network N_1 . Within the SA, we used DTW to compare the simulation and real data. The output of SA was a vector containing rewired network that optimized the goodness of fit of simulation data with real data. We tried different combinations of the initial temperature and the cooling rate in SA to assess goodness of fit. We observed that GA based rewiring was more effective to uncover topological changes that can be causal for enhanced drug sensitivity. For example, we set the initial temperature to 1000. After each iteration of SA, the system was allowed to cool at a rate of 0.01. The lower cooling rate

represents a slow reduction in the probability of accepting worse solutions as SA executes to find the solution. Accepting worse solutions by SA allows for a more extensive search for the optimal solution. We followed the same model selection framework where we initially derived N_2 from N_1 , and subsequently N_3 from N_2 . The simulation results are shown in the Figure 3.5.

As shown in the Figure 3.5, the performance of SA was not as good as the performance of GA (Figure 3.3). For example, the DTW score for comparing simulation data from basic network with the experimental apoptosis data was 1.98 (Figure 3.3 a and 3.5 a). After the execution of SA based optimization, the fitness score was 1.64 (Figure 3.5 b), which is higher than the GA based optimisation score of 0.024 (Figure 3.3b). Similarly the performance of SA was again not good for deriving N_3 . The score for comparing N_2 based simulated data with drug sensitive apoptosis was 0.25 (Figure 3.5 d). However, the score for goodness of fit between N_3 based data and the drug sensitive data using GA based rewiring was 0.20. These findings suggest that the topological rewirings using Genetic Algorithm is relatively more promising approach to investigate the enhanced drug sensitivity.

3.4 Discussion and conclusion

In our proposed study, we used Genetic Algorithm to infer rewiring of the apoptotic network. The most frequent rewiring events are selected, and most of them were found to be consistent with existing literature. However, using Genetic Algorithm we may not find a globally optimal solution, as some of the network rewiring events might have been missed by the Genetic Algorithm based searching. So far we have included only 18 out of 32 proteins in the Yaffe's data. A network with more signaling data may help us understand the pathway changes in more comprehensive ways. In future, with expanded apoptosis network and more proteins, we would be able to detect more network alterations through comparisons between simulation data and experimental data. Also, beside Genetic Algorithm we plan to use other computational techniques (e.g., Simulated Annealing). In this study Yaffe's data set has been used. Our next goal is to construct a larger network model for cancer cell phenotypic responses which can simulate observations for some other datasets taken from multiple sources.

Table 3.1: Differential equations for the basic apoptosis network N_1 (Figure 3.1)

S.N.	Rate of Change	Differential equations
1	$\frac{d}{dt}$ CASP6	- [CASP6]·rc ₁₁ + [CASP9]·rc ₁₂
2	$\frac{d}{dt}$ CDK1	+ [Wee1]·rc ₁₃ - [CDK1]·rc ₁₄
3	$\frac{d}{dt}$ Wee1	- [Wee1]·rc ₁₃
4	$\frac{d}{dt}$ JNK	- [JNK]·rc ₁₅ + [EGFR]·rc ₁₇
5	$\frac{d}{dt}$ BID	- [BID]·rc ₁₆ - [BID]·rc ₂₅
6	$\frac{d}{dt}$ EGFR	- [EGFR]·rc ₁₇ - [EGFR]·rc ₃₂ ;
7	$\frac{d}{dt}$ PUMA	- [PUMA]·rc ₁₉ + [p53]·rc ₂₇
8	$\frac{d}{dt}$ CASP3	- [CASP3]·rc ₁ + [XIAP]·rc ₂ + [SMAC]·rc ₄ + [CASP8]·rc ₇ - [CASP3]·rc ₈ + [CASP9]·rc ₉ + [CASP6]·rc ₁₁ + [MOMP]·rc ₃₀
9	$\frac{d}{dt}$ CASP9	- [CASP9]·rc ₉ - [CASP9]·rc ₁₂ + [PUMA]·rc ₁₉ - [CASP9]·rc ₂₄ + [BCL2]·rc ₂₆
10	$\frac{d}{dt}$ HER2	- [HER2]·rc ₂₀ - [HER2]·rc ₃₁
11	$\frac{d}{dt}$ CASP8	- [CASP8]·rc ₃ + [RIP1]·rc ₆ - [CASP8]·rc ₇ + [CASP3]·rc ₈ + [BID]·rc ₂₅ + [EGFR]·rc ₃₂
12	$\frac{d}{dt}$ Apop	+ [CASP3]·rc ₁
13	$\frac{d}{dt}$ XIAP	- [XIAP]·rc ₂ + [SMAC]·rc ₂₁ + [CASP9]·rc ₂₄ + [HER2]·rc ₃₁
14	$\frac{d}{dt}$ SMAC	+ [CASP8]·rc ₃ - [SMAC]·rc ₄ + [BCL2]·rc ₅ - [SMAC]·rc ₂₁ + [MOMP]·rc ₂₉
15	$\frac{d}{dt}$ BCL2	- [BCL2]·rc ₅ + [p53]·rc ₁₀ + [CDK1]·rc ₁₄ + [JNK]·rc ₁₅ + [BID]·rc ₁₆ + [STAT3]·rc ₂₂ - [BCL2]·rc ₂₆
16	$\frac{d}{dt}$ RIP1	- [RIP1]·rc ₆ + [TNFR]·rc ₂₃
17	$\frac{d}{dt}$ p53	- [p53]·rc ₁₀ + [AKT]·rc ₁₈ - [p53]·rc ₂₇
18	$\frac{d}{dt}$ AKT	- [AKT]·rc ₁₈ + [HER2]·rc ₂₀
19	$\frac{d}{dt}$ STAT3	- [STAT3]·rc ₂₂
20	$\frac{d}{dt}$ TNFR	- [TNFR]·rc ₂₃
21	$\frac{d}{dt}$ DAPK1	- [DAPK1]·rc ₂₈
22	$\frac{d}{dt}$ MOMP	+ [DAPK1]·rc ₂₈ - [MOMP]·rc ₂₉ - [MOMP]·rc ₃₀

(a)

(b)

(c)

(d)

Figure 3.3: Simulation results for Genetic Algorithm: fitting experimental data with simulation data. (a) Simulation result for basic apoptotic network, (b) Simulation result for DMSO treatment (N_2 with DMSO treatment). (c) Comparing N_2 based simulation data with ERL-DOX treatments data by correlation of apoptosis rates over time points. (d) Simulation result for ERL-DOX treatments. The scores represents DTW score.

Figure 3.4: Evaluating performance of the Genetic Algorithm for different mutation rates. Comparison of computational apoptosis data with experimental apoptosis data from [12] for different mutations probabilities, *i.e.*, 0.1, 0.05 and 0.01 for deriving N_2 from N_1 . The results suggest that lower mutation probability is better for enhancing goodness of fit using Genetic Algorithm. Our findings suggest that a lower mutation probability can lead to convergence of GA as compared to a higher mutation probability.

(a)

(b)

(c)

(d)

Figure 3.5: Simulation results for Simulated Annealing (a) Simulation result for basic apoptotic network, (b) Simulation result for DMSO treatment. (c) Comparing N_2 based simulation data with ERL-DOX treatments data. (d) Simulation result for ERL-DOX treatments. The scores represent DTW score.

Table 3.2: Rate constants used for the rewiring in basic apoptosis network

Node	BID	BIM	cCASP8	cCasp6	cCasp9	Cyclin-D1	EGFR	HER2	MOMP	p53	pAKT	pDAPK1	pJNK	PUMA	pWee1	RIP	Stat3	SMAC	CASP3	TNFR	XIAP
BID	0.316	0.665	0.287	0.547	0.167	0.917	0.366	0.271	0.115	0.082	0.343	0.439	0.839	0.865	0.715	0.546	0.387	0.395	0.407	0.938	0.354
BIM	0.078	0.811	0.492	0.69	0.393	0.512	0.511	0.446	0.668	0.928	0.492	0.229	0.904	0.736	0.168	0.974	0.949	0.881	0.076	0.263	0.741
cCASP8	0.012	0.228	0.56	0.679	0.894	0.255	0.259	0.574	0.544	0.035	0.792	0.61	0.621	0.657	0.883	0.119	0.541	0.293	0.616	0.041	0.296
cCasp6	0.638	0.672	0.422	0.813	0.664	0.405	0.578	0.108	0.178	0.227	0.257	0.926	0.425	0.497	0.078	0.635	0.548	0.688	0.333	0.649	0.471
cCasp9	0.482	0.519	0.512	0.131	0.33	0.96	0.468	0.111	0.396	0.5	0.619	0.636	0.006	0.099	0.704	0.148	0.297	0.739	0.086	0.387	0.759
Cyclin-D1	0.047	0.242	0.396	0.17	0.825	0.658	0.732	0.016	0.493	0.143	0.997	0.055	0.971	0.263	0.67	0.215	0.75	0.904	0.752	0.705	0.541
EGFR	0.551	0.439	0.149	0.702	0.243	0.77	0.105	0.426	0.438	0.253	0.014	0.993	0.531	0.465	0.689	0.035	0.119	0.721	0.747	0.07	0.809
HER2	0.343	0.623	0.885	0.352	0.908	0.115	0.98	0.301	0.636	0.909	0.459	0.296	0.341	0.512	0.989	0.775	0.166	0.644	0.232	0.932	0.151
MOMP	0.021	0.664	0.355	0.63	0.292	0.264	0.416	0.272	0.153	0.538	0.521	0.241	0.592	0.597	0.061	0.258	0.992	0.678	0.315	0.255	0.201
p53	0.645	0.654	0.583	0.83	0.624	0.812	0.795	0.656	0.93	0.859	0.878	0.386	0.892	0.471	0.622	0.786	0.068	0.732	0.162	0.673	0.765
pAKT	0.55	0.92	0.589	0.258	0.807	0.759	0.431	0.767	0.871	0.524	0.152	0.262	0.454	0.414	0.487	0.61	0.759	0.899	0.043	0.253	0.298
pDAPK1	0.415	0.608	0.63	0.867	0.032	0.424	0.811	0.911	0.266	0.85	0.186	0.225	0.695	0.514	0.305	0.03	0.433	0.934	0.554	0.945	0.498
pJNK	0.805	0.057	0.605	0.952	0.88	0.69	0.485	0.48	0.51	0.239	0.93	0.369	0.915	0.157	0.147	0.91	0.142	0.442	0.672	0.327	0.631
PUMA	0.159	0.416	0.709	0.456	0.989	0.688	0.539	0.875	0.491	0.116	0.785	0.175	0.112	0.243	0.956	0.375	0.051	0.352	0.988	0.937	0.425
pWee1	0.753	0.474	0.722	0.327	0.65	0.183	0.011	0.084	0.042	0.648	0.16	0.769	0.224	0.196	0.961	0.414	0.966	0.931	0.899	0.262	0.194
RIP	0.488	0.283	0.413	0.158	0.379	0.013	0.606	0.675	0.405	0.108	0.937	0.584	0.489	0.7	0.874	0.925	0.229	0.954	0.149	0.843	0.069
Stat3	0.421	0.272	0.186	0.811	0.732	0.029	0.23	0.74	0.379	0.286	0.066	0.315	0.799	0.548	0.751	0.457	0.388	0.069	0.089	0.651	0.109
SMAC	0.682	0.584	0.024	0.807	0.92	0.102	0.691	0.385	0.167	0.099	0.381	0.905	0.442	0.102	0.38	0.968	0.148	0.342	0.353	0.635	0.439
CASP3	0.995	0.536	0.406	0.688	0.496	0.844	0.518	0.659	0.85	0.287	0.693	0.899	0.17	0.348	0.618	0.198	0.582	0.014	0.851	0.299	0.389
TNFR	0.984	0.36	0.866	0.224	0.181	0.104	0.197	0.02	0.653	0.913	0.928	0.435	0.884	0.884	0.711	0.786	0.433	0.397	0.704	0.654	0.252
XIAP	0.999	0.549	0.875	0.334	0.474	0.321	0.84	0.599	0.519	0.514	0.888	0.071	0.029	0.116	0.059	0.839	0.104	0.606	0.164	0.98	0.642

Chapter 4

Predicting drug effects in diverse cancer cell lines

4.1 Introduction

Programmed cell death, also known as apoptosis, is a well-coordinated cellular process that is essential for the homeostatic growth and functioning of a multicellular organism as it removes damaged cells to avoid harmful effects such as toxicity [33]. Apoptosis is regulated by several signaling pathways and crosstalks among the pathways which are highly complex. In malignant cells these signaling pathways are dysregulated which allows the abnormal cells to grow and proliferate beyond their expected lifetime leading to diseases such as cancer [12, 111]. Therefore, in-depth understanding of the apoptotic signaling pathways is very important for finding effective treatments to selectively kill tumor cells.

In our previous study [50] (Chapter 3), we proposed a hybrid model that used ODE-based simulations and Genetic Algorithm for model selection to predict the drug-induced rewiring in the apoptotic signaling network in cancer cells. We changed the network topology while fixing the model parameters to find rewiring in response to drug-induced perturbations. Since the signaling networks inside the cells are complex, containing several interconnected pathways, there could be multiple causal effects responsible for drug sensitivity. For instance, changes in the kinetic rates may affect the drug sensitivity of tumor cells [112]. Thus, deriving model parameters from experimentally measured phosphorylation levels of signaling proteins can help in understanding the network dynamics and drug effects.

In this Chapter, we propose a new hybrid model of cancer signaling network, which is used to conduct new prediction and analysis of anticancer drug effects. Compared with previous Chapter ([50]), this Chapter presents significant extensions in the following aspects. First, the signaling network in [50] has been expanded to a larger and more comprehensive network that contains more proteins including kinases and phosphatase for phosphorylation and dephosphorylation of signaling proteins, respectively. Secondly, instead of changing a network topology, we now focus on network parameters in a fixed network. So, instead of using genetic algorithm (which was primarily used for finding new network topology by rewiring) for heuristic optimization in model selection, here we use a well-studied software tool named ABC-SysBio [62]. ABC-SysBio is based on a framework of Bayesian learning that can infer the model parameters from real data. Thirdly, by integrative and comparative analysis of data from multiple cancer cell lines, we use our model to predict context-specific apoptosis. Moreover, we perform *in silico* caspase knockdown experiments for each of the aforementioned three cell lines and assess the regulatory impact of caspases on apoptosis. A major difference between the study in this Chapter and previous chapter is that, while we focused on the inference of network rewiring in [50], here we infer parameters (i.e. reaction rate constants) for a fixed network topology. The rationale is that most events of network rewiring induced by drug treatment are edge deletions, which can be represented by the change of rate constants associated with network edges. In a sense, such network rewiring events can be considered as a special type of changes of parameters. However, we are aware that not all network rewiring can be represented by a change of parameter, such as adding a new edge. This issue will be discussed at the end of this Chapter. Our pipeline of studies is illustrated with the flowchart in Figure 4.1.

Given a knowledge-based ODE model derived from the literature, our method can infer parameters from experimental data by iterative optimization and also predict cell line specific changes of apoptosis in response to drug effects. Our results demonstrate that the hybrid modeling approach, which combines the knowledge-driven dynamical modeling and data-driven statistical modeling, can predict signaling activities and cell fates (apoptosis in this study), more accurately than models without calibration to real data. The study, therefore, demonstrates the scalability of the proposed computational

model to diverse cancer cell lines. In several studies, such as in [42], where experimental cell signaling data are available, but experimental apoptosis data are missing, our model can be used to predict the apoptosis and the drug effects. Considering the inherent complexity of cancer and the emergence of large amount of new “omics” data from cancer research, such a hybrid modeling strategy would be useful for understanding cancer and discovery of novel anti-cancer therapeutics. Context-specific modeling that takes into account genetic variations would be a future work to extend our current model [65].

Figure 4.1: Flow diagram depicting the steps in our method. We derive the signaling network manually using GeneGO MetaCore database and encode the ODEs for the network using Michaelis-Menten kinetics. The ODE-based computational model is calibrated using cell line specific data. Subsequently, for each cell line, we predict the apoptosis and assess the impact of caspases on apoptosis using *in silico* knockdown experiments.

4.2 Methods

4.2.1 Single-cell data

As explained in the Chapter 3, for our studies, we used Yaffe’s data set contains six types of cellular phenotypic responses from three cancer cell lines, namely triple negative BT-20, Hormone sensitive MCF-7 and HER2 overexpressing MDA-MB-453. The six phenotypes are Apoptosis, Proliferation, G1, S, G2 and M. These phenotypic responses are linked with on 35 signaling proteins (only 32 available in the published data set). The cell fates and signaling data are available for all the three context-specific cell lines. The measurements of the signaling and cell fate data have been done for six treatments namely DMSO, TAR, DOX, DT, T-D and D-T. The apoptosis level is reportedly high for the T-D treatment, meaning the cells are now less tumorigenic for this treatment and thereby more susceptible to death induced by DNA damage response pathways.

4.2.2 Derivation of signaling network structure

Model construction of signaling network is a crucial step for subsequent simulation, prediction and data analysis. Our signaling network was derived using the GeneGO Meta-Core database [21]. The diagram of the constructed network is shown in Figure 4.2. The network comprises a total of 71 nodes and 131 edges, in which there are 19 proteins with phosphorylation data available in the Yaffe’s dataset [12]. The network includes kinases and phosphatases that catalyze the phosphorylation and dephosphorylation reactions, respectively. The types of reactions included in the network are phosphorylation, dephosphorylation, and cleavage. Since our goal is to study the dynamics of phosphorylation of proteins in a signaling network, we did not represent protein complexes or regulation of gene expression (e.g. transcription and translation) explicitly. Our network includes several important proteins from different apoptotic pathways, such as BID, p53, CASP9, CASP3, CASP8, CASP6. Although in this study we have focused on the study of responses of breast cancer cells to Erlotinib, this network model of signaling pathways could be adapted to study other aspects of apoptosis and the effects of other drugs.

4.2.3 Kinetic modeling

For the signaling network manually constructed (Figure 4.2), we encode the kinetics in ODEs using the CellDesigner software [23]. The ODEs for biochemical reactions in the network were formulated based on the Michaelis-Menten kinetics [9]. The computational model comprises 71 ODEs and 163 reaction rate constants. The numerical solution of the ODEs in the form of time-series data was obtained using the MATLAB ODE solver. We denote the activity level of protein X as $[X]$, the rate constant for the i th catalysed reaction as k_i and Michaelis-Menten rate constant by m_i . For example, the ODE for $pSTAT3$ with respect to time t is given as:

$$\begin{aligned} \frac{d}{dt}[pSTAT3] = & k_1 \cdot [STAT3] \cdot [pEGFR]/(m_1 + [STAT3]) \\ & + k_2 \cdot [STAT3] \cdot [pCHK1]/(m_2 + [STAT3]) \\ & + k_3 \cdot [STAT3] \cdot [pP38]/(m_3 + [STAT3]) \\ & - k_4 \cdot [pSTAT3] \cdot [pPH10]/(m_4 + [pSTAT3]), \end{aligned} \tag{4.1}$$

where $[pSTAT3]$ and $[STAT3]$ are the concentrations of the phosphorylated and unphosphorylated forms of the $STAT3$ protein. Here we assume that the overall concentration of each protein (i.e. both phosphorylated and unphosphorylated) in a cell is stable, so that the phosphorylation levels (measured as the percentage of copies of each protein in the phosphorylated form) can represent the cellular state. The steady state constraint is mathematically encoded as:

$$\frac{d}{dt}[pSTAT3] + \frac{d}{dt}[STAT3] = 0. \tag{4.2}$$

4.2.4 Parameter inference with ABC-SysBio

A long-standing challenge for dynamic modeling in systems biology is parameter estimation, for which various methods have been proposed [54, 55, 56]. Several software have been developed based on Bayesian inference for parameter estimation and model ranking using likelihood function [63]. However, for complex ODE models, deriving the likelihood function is not always possible. To address this issue, we use the software suite called ABC-SysBio [62]. Considering the limited amount of experimental data available

in our study, we estimated model parameters by using the software suite called ABC-SysBio which is an open source python package. The ABC-SysBio is suitable for fitting the computational models to the experimental observations. It can be used to estimate model parameters of biochemical reaction network where the nodes denote the proteins and the edges denote the biochemical reactions (e.g. phosphorylation). A very important feature of ABC-SysBio is that it is likelihood-free by using the framework of Approximate Bayesian Computation (ABC). Moreover, ABC-SysBio is able to handle missing data. A critical issue for parameter inference is overfitting of model with the experimental data. However, parameter estimation methods based on Bayesian inferential techniques are promising to address the issue of overfitting [62].

4.2.5 Context-specific modeling

Using the Bayesian inference based parameter estimation, we are able to calibrate the knowledge-derived network model to real data, and thereby make the model context-specific. To this end, we incorporate cell line specific data for our model calibration. A breast cancer subtype can be caused by mutations in thousands of genes [113]. To understand context-specific drug sensitivity of breast cancer, data from diverse breast cancer cell lines can be analysed together [114]. In this study we use a generic network topology in Figure 4.2 to predict drug effects in the three breast cancer cell lines, i.e., BT-20, MCF7 and MDA-MB-453. The model is calibrated with signaling data (i.e., independent variables) from each of the three cell lines. Then we predict the cell line specific apoptotic response (i.e., dependent variable) and assess the correlation between *in silico* and experimental apoptosis data. To study the impact of caspases on apoptosis, we perform *in silico* caspase knockdowns experiment on each of the three cell lines. The knockdown simulations show that different cell lines have different signaling dynamics which can be caused by a combination of overactivation and suppression of the network edges in the cancer signaling pathways.

4.3 Results

4.3.1 Simulations with random and inferred parameters

To demonstrate the importance of parameter inference from real data, we first simulated signaling and phenotypic responses using the network model with 1000 sets of randomly generated parameters, and compared the predictions with the experimental data in [12] for the BT-20 cell line. The initial phosphorylation level for each protein was fixed to one while the rate constants were varied with uniformly generated random numbers. We computed the sum of Euclidean distances between the simulated and experimental data for each of the 19 proteins. Euclidean distances for signaling proteins were calculated by comparing signaling levels of the 19 proteins in Yaffe’s dataset with simulated signaling data at 8 time points (numbered as 0, 1, 2, $\dots$, 7). Since the experimental data were missing for quite a few time points (*i.e.* 1, 2, 3, 4, 5), we compared the simulated and real time-series datasets only on the available time points. For apoptosis, the Euclidean distance was calculated by comparing the experimental apoptosis data with simulated apoptosis data at 5 time points (*i.e.* 0, 6, 7, 8 and 9). The distribution of the Euclidean distances for the signaling data of the 19 proteins and apoptosis, based on the 1000 sets of random parameters, are shown in Figure 4.3. Next, we will calibrate the model using the Yaffe’s phosphorylation dataset and show that, after model calibration, the goodness of fit between predicted and real apoptosis data can be improved.

To infer parameters from the real data, the input for the ABC-SysBio software was a vector containing phosphorylation levels of 19 proteins from the Yaffe’s dataset which was iteratively compared with a vector of simulated signaling data. The output of ABC-SysBio was a vector of estimated model parameters that optimized the goodness of fit between the simulated time-series data with the real data. The initial phosphorylation levels for the 19 proteins in the model were set to the initial levels in the Yaffe’s data at time point 0. For other proteins in the network, of which the initial phosphorylation levels were unknown, the initial phosphorylation level was set to the dimensionless numerical value of one [57]. Apoptosis data in [12] was not used for the parameter inference. There are totally 163 parameters in the network model, and all of them were inferred. The different rate constants for edges downstream of EGFR are shown in Table 4.2. The nabla

Figure 4.3: Distributions of Euclidean distances. Distributions of Euclidean distances (in blue) between the simulated and the Yaffe’s data for each of the 19 signaling proteins and apoptosis (bottom-right panel) for 1000 sets of random parameters for the BT-20 cell line. The nabla symbol (in red) represents the Euclidean distance between the simulated and experimental data after model calibration. The x -axis represents the distributions for the Euclidean distances of the corresponding variable.

symbols (in red) in Figure 4.3 represent the Euclidean distance between the simulated and experimental data after model calibration, which shows the improvement in goodness of fit for signaling and apoptosis data after model calibration. The plots of the predicted time-course phosphorylation data and apoptosis data (after parameter inference) along with corresponding real data are shown in Figure 4.4, and 4.5(a), respectively.

We performed Z -test to test the null hypothesis that the trained data follows the distribution for the given population mean (μ) and the standard deviation (σ), against the alternative hypothesis for the sample mean being less than the population mean. Table 4.1 shows the Z -test details based on Euclidean distance for the population size of 1000 and sample size of 1. The calculated p -values indicate that Z -test rejects the null hypothesis at the default 5% significance level, in favour of the alternative hypothesis that the sample mean is less than the population mean.

4.3.2 Prediction of apoptosis

Apoptosis in a cell depends on various molecular processes including signal transduction, gene regulation and metabolic activities. We aim to test if our model, after being trained

Table 4.1: Z-test details for the population size of 1000

Protein name	Population Variance (σ^2)	Population standard deviation (σ)	Population Mean (μ)	Sample Mean (M)	Change (%)	<i>p</i> -values
pSTAT3	0.00166862	0.040828315	65.48044006	51.91633402	20.71	0
BIM	0.000272388	0.01649593	4.513296266	0.401556521	91.10	0
pP38	0.000693108	0.02631377	28.93602503	10.24261052	64.60	0
pAKT	0.000315265	0.017746823	50.52707976	21.16214016	58.12	0
'EGFR'	0.000742087	0.027227653	1251.823036	97.12951999	92.24	0
pS6	0.000196062	0.013995197	9.997004952	1.836861554	81.63	0
pChk1317	0.000215795	0.014682615	0.880672938	0.820222384	6.86	2.89814E-09
pWee1	0.000166433	0.012894425	11.18653483	3.173290279	71.63	0
p53	0.001022602	0.031962155	114.6508178	47.10371723	58.92	0
pP53	0.000752577	0.027419419	263.8236052	162.502428	38.40	0
BID	0.000464189	0.021534277	1.707962759	0.397814926	76.71	0
pCabl	0.000277594	0.016652829	5.37316161	0.687712511	87.20	0
cCasp6	0.000934425	0.030553079	23.39689728	16.66715482	28.76	0
cC8	0.001221269	0.03492918	4.954032047	3.113688906	37.15	0
cCasp9	0.001116348	0.033395085	1.03898876	0.326151535	68.61	1.7438E-200
SMAC	1.22377E-08	0.000110569	1.113955891	0.508626443	54.34	0
pCDC25C	6.48046E-05	0.008046107	0.37684775	0.24085154	36.09	1.4185E-126
pDAPK1	0.000973158	0.031179878	7.550271328	5.183711251	31.34	0
Beclin1	4.19792E-07	0.000647589	0.334703914	0.157985115	52.80	0

Figure 4.4: Inference of Parameters. Fitting of simulation data to experimental data for the 19 proteins for the BT-20 cell line from [12] based on inferred parameters for the signaling network.

with only a limited amount of signaling data from [12], would be able to predict changes in apoptosis accurately when compared with observed *in vitro* measurements in drug sensitive cells. The experimental dataset in [12] contains both signaling and cell fate data. We have used only the signaling data (excluding the cell fate data) to train our model. Initially we predicted apoptosis using the model in Figure 4.2 with random parameters for 1000 times. The distribution of the Euclidean distances between the simulated data (using random parameters) and real data of apoptosis is shown in the bottom-right panel of Figure 4.3. We also predicted apoptosis using the calibrated model with parameters inferred from real signaling data (but not apoptosis data), and compared the predicted time-series data with the real data of apoptosis. The comparison result for the BT-20 cell line is shown in Figure 4.5(a). The Euclidean distance between Yaffe’s real data of apoptosis and the data predicted by the trained model was 18.9. Both the simulated and experimental apoptosis levels were found to be increasing with time. There is a strong correlation between two time series, with Pearson correlation coefficient 0.994. These observations indicate a close association between *in vitro* and *in silico* phenotypic observations of programmed cell death.

Figure 4.5: Prediction of context-specific apoptosis. Prediction of apoptosis and comparison of cell line specific computational apoptosis data with experimental apoptosis data from [12] for three cell lines, i.e., BT-20, MCF7 and MDA-MB-453. The calibrated computational model in Figure 4.2 was able to predict the cell line specific apoptosis with high Pearson correlation coefficients, such as 0.994 in BT-20, 0.9569 in MCF7 and 0.904 in MDA-MB-453 cell lines. The Euclidian distance between simulated and experimental apoptosis data for the BT-20, MCF7, and MDA-MB-453 cell lines were 18.89, 58.98, and 10.25, respectively.

We used a similar training procedure to calibrate the computational model in Figure 4.2 with the time-course drug sensitive data for the hormone-sensitive Michigan Cancer Foundation-7 (MCF7) cell line from [12]. The MCF7 cancer cell line is insensitive to the ERL/DOX combination of drug treatment, which otherwise, caused high apoptosis level in BT-20 cells. Our calibrated model was able to predict the apoptosis with the Pearson correlation coefficient of 0.9569. The predicted apoptosis is shown in Figure 4.5(b). We further trained the network model in Figure 4.2 with drug sensitive data for the MDA-MB-453 cell line. The predicted apoptosis for the MDA-MB-453 cell line is shown in Figure 4.5(c). The Pearson correlation coefficient is 0.90, indicating a very strong linear correlation between the *in silico* and experimental apoptosis data.

4.3.3 Virtual knockdown of caspases

The caspase proteins are key players in regulating apoptosis. The caspases can be categorized into initiator (*e.g.* CASP8, CASP9) and executioner caspases (*e.g.* CASP3, CASP6). Initiator caspases are responsible for activating executioner caspases via chain reactions. Once activated, executioner caspases degrade several cellular components in order to induce the morphological changes for apoptosis [115]. Enhanced cell death in

Table 4.2: Activation rate constants downstream of EGFR among different breast cancer cell lines. The DMSO treatment rate constants have been inferred from the phosphorylation level in the DMSO treated cancer cells. The Drug treatment by Erlotinib rate constants have been inferred from the phosphorylation level data in the drug sensitive cancer cells. There is no healthy data in the published dataset by [1]. All the data are recorded from cancer cells corresponding to the control treatment and drug treatment. The table uncovers the changes in the rate constants to explain the enhanced drug sensitivity.

Network edge	BT-20 (DMSO)	BT-20 (Drug)	MCF7 (DMSO)	MCF7 (Drug)	MDA-MB-453 (DMSO)	MDA-MB-453 (Drug)
EGFR-RAS	0.99994	6.53E-06	0.59745	6.63E-06	0.39985	9.64E-06
RAS-bRaf	0.79519	7.12E-06	0.39931	6.55E-06	0.54668	9.51E-06
EGFR-Stat3	1.1956	0.00089811	0.49992	0.19786	0.69579	9.99E-06
Stat3-PI3K	0.14384	7.73E-06	0.29652	9.37E-06	0.29368	9.96E-06
CASP6-CASP8	0.094878	0.20336	3.31E-06	6.89E-06	0.049838	0.09958
CASP8-CASP3	0.096599	0.50483	0.00097789	0.0072124	0.00073772	0.0099999
CASP8-CASP9	0.0090839	0.60083	3.91E-06	7.88E-06	0.00090093	9.85E-06
CASP9-CASP3	0.89505	1.5073	0.29154	0.69989	0.24888	0.49941

cancer cells is often driven by caspases after EGFR inhibition by anti-cancer drugs, such as Erlotinib. Pro-apoptotic molecules, such as CASP8 and CASP6, have been found to be important variables for regulating apoptosis in the BT-20 cell line [12]. To further understand the roles of caspases in the regulation of apoptosis, we performed *in silico* knockdown of CASP8 and CASP6. The predicted effects of knockdown in the BT-20 cell line are shown in Figure 4.6(a). We did three knockdown experiments, namely, CASP8 knockdown, CASP6 knockdown and the simultaneous knockdown of both CASP8 and CASP6. We set the activity level of CASP8 or CASP6 to zero, initially one by one, and then simultaneously activity level of both CASP8 and CASP6 were set to zero. All other model parameters were kept unchanged during the knockdown experiments. A decrease in apoptosis level was observed for each of the knockdown experiments. The decrease in the apoptosis level was higher for CASP8 knockdown than for CASP6 knockdown for the BT-20 cell line. The effect of combined CASP8 and CASP6 knockdown was higher than individual CASP8 or CASP6 knockdown. The simulation result suggests that the caspases were important molecules for regulating apoptosis. There is evidence of CASP8 contributing to enhanced cell death when EGFR is suppressed [116, 117].

Figure 4.6: Prediction of context-specific CASP8 and CASP6 knockdown effects.

Effects of CASP8 and CASP6 knockdown on apoptosis in trained models for the BT-20, MCF7, and MDA-MB-453 cell lines. The increase of apoptosis is weakened in response to the knockdown in BT-20 and MDA-MB-453 cells. Moreover, the combined effect of knockdown of CASP6 and CASP8 was higher than the individual knockdown. The impact of knockdowns in MDA-MB-453 cell line was higher and lower than the MCF7 and BT-20 cells, respectively. However, the caspase knockdown has little impact on apoptosis of MCF7 cell line as the apoptotic curves for CASP6 and CASP8 knockdown remained similar to the apoptosis before the knockdown.

We also simulated CASP8 and CASP6 knockdown in the MCF7 cells. We found that the CASP8 and CASP6 knockdown did not affect the apoptosis in the MCF7 drug sensitive cells. The apoptotic responses, before and after knockdowns, were similar at different time points indicating that these two knockdowns had almost no impact on the enhanced drug sensitivity. The effects of CASP8 and CASP6 knockdown on apoptosis in trained model for the MCF7 cell line are shown in Figure 4.6(b). Despite caspases knockdown, the apoptosis remained similar to the apoptosis before caspases knockdown. The rate constants (Table 4.2) in the trained model for the edges CASP6-CASP8, CASP8-CASP9 and CASP8-CASP3 were $6.889\text{e-}06$, $7.8777\text{e-}06$ and $7.2124\text{e-}03$, respectively. This indicates that, due to the blockage of signals, the CASP6 and CASP8 were not able to send pro-apoptotic signal to induce cell death, and therefore CASP6 and CASP8 were not significant drivers of apoptosis in the drug sensitive MCF7 cells. However, for CASP9-CASP3 the rate constant was 0.69989 , indicating that CASP9 may play an active role in stimulating apoptosis in the MCF7 cells. There are several pieces of experimental evidence indicating the down regulation of CASP8 in different breast cancer cells including MCF7, MB231, SKBR3 and HCC1937 [12, 118]. Promoter methylation has been reported to be one important reason for reduced CASP8 signaling profile in MCF7 cell

line [118]. The study in [12] found that CASP6 was not a strong driver of apoptosis in MCF7 cell line. The activation of the CASP9 in MCF7 cells following treatment with DNA damaging agents has been reported in [119].

Subsequently, we performed similar virtual knockdown of CASP6 and CASP8 for the MDA-MB-453 cell with results, shown in Figure 4.6(c). We found that CASP6 knockdown has slightly more impact on apoptosis than CASP8. The rate constants for the caspase edges (Table 4.2) CASP6-CASP8, CASP8-CASP9 and CASP8-CASP3 were 0.09958, 9.9999e-03 and 9.8506e-06, respectively. Since the rate for CASP6-CASP8 was higher than CASP8-CASP3 and CASP8-CASP9, CASP6 seems able to stimulate apoptosis more strongly than CASP8. The edge CASP9-CASP3 with rate constant of 0.49941 may also contribute to the apoptosis. In summary, our virtual knockdown experiments suggest that CASP6 appeared to be more effective in inducing apoptosis for the BT-20 and MDA-MB-453 cell lines, but not for the MCF7 cell line. By contrast, CASP8 seems not important for regulating apoptosis of the MCF7 and MDA-MB-453 cells. CASP9 is predicted to play an important role in stimulating apoptosis for all the three cell lines.

4.3.4 Predicting drug-induced pathway alterations

Although in this study we have focused on the parameter estimation on a network of fixed topology, the predicted changes of parameters (in Table 4.2) can be used further to predict network rewiring events. For example, if a rate constant in Table 4.2 is lower than a threshold of say 0.001, then we consider the corresponding edge deleted from the network. In Figure 4.2, we plotted several blue arrows to represent such edge deletion events, which represent predicted drug effects where signal transductions were inhibited in the downstream of EGFR, according to the predicted values of rate constants in Table 4.2. The activation rate constants in the blue edges were significantly low. For example, in BT-20 cells following the drug treatment, the rate constants for the edges EGFR-RAS, RAS-bRAF, EGFR-STAT3, STAT3-PI3K, CASP8-CASP9, and CASP9-CASP3 were 6.5294e-06, 7.1158e-06, 8.9811e-04, 7.7281e-06, 0.60083, and 1.5073, respectively (Table 4.2).

The inferred parameters show that in the DMSO treatment data (that basically represented cancer cells), several edges (*e.g.* EGFR-RAS) that are known to be causal for breast cancer, were transmitting signals at a higher rate, whereas the signaling from CASP6 and CASP8 were inhibited. By contrast, the signaling from CASP9 remained effective for different treatment. The model calibration predicted higher rate constants for several edges which could be activated by the inhibition of EGFR. Moreover, the inferred rate constants suggest that the signaling from the initiator (*e.g.* CASP8) to the executioner Caspase (*e.g.* CASP3) could be blocked in the cancer cells. However, in the drug sensitive cells, the signaling between the Initiator and executioner caspases can be activated. Our simulation-based findings about the blockage and activation of signaling pathways as drug effects may be used to explain some experimental observations reported in the literature [20, 120, 121].

4.4 Conclusion and discussion

In this study, we proposed a hybrid modeling approach which integrates dynamical ODE modeling with data-driven statistical learning to uncover the mechanism of anti-cancer drug effects on cancer cells. The ODE based computational model of signaling pathways was calibrated for three breast cancer cell lines, *i.e.*, BT-20, MCF7 and MDA-MB-453, by iteratively fitting to the cell-line specific phosphoproteomics data. The calibrated model was then used to simulate the cellular responses (*e.g.* cell death) to different drug-induced perturbations. We also identified the blockage in the pathways downstream of EGFR leading to enhanced drug sensitivity in the BT-20 cells. Our results of comparing model prediction with real data suggested that model calibration by Bayesian inference of parameters from real data can improve the accuracy (as compared to a model based random parameters) in predicting time-course signaling and apoptosis data. Our virtual experiments of caspase knockdown have also yielded some insights into the mechanisms of apoptotic regulation. The changes in activation rates inferred from the real data can be used to map network rewiring events induced by the anticancer drug of Erlotinib. In conclusion, our hybrid modeling method can take into account both prior knowledge

and real data, to capture the context-specific dynamics of cancer cell fate. The models constructed and calibrated as such can be used to do simulation, prediction, and data analysis. Therefore, they will be instrumental for the precision medicine, as more biomedical knowledge is accumulated and more patient-specific data become available.

Despite the promising performance and prospect of the hybrid modeling in cancer systems biology, a few challenges remain. Without attempting to be complete, we list some of the challenges here, as a discussion about future work. First, to calibrate a dynamic model against time-series real data, it is crucial to align time points between simulated and real time-course data. This is challenging for multiple reasons, such as, there may be missing time points in the real data; cellular processes tend to be multiscale by nature (e.g. changes in signal transduction occur much faster than gene expression or apoptosis). Secondly, the accurate inference of network models from limited amounts of data could be hindered by issues like overfitting, lack of identifiability, etc. Different models might fit the same dataset equally well, if the information from the real data is insufficient to distinguish the models. Hybrid modeling that takes into account of prior knowledge may help address this issue, but the prior knowledge itself might be incomplete or biased. Thirdly, the relation between network structures and parameters has long been a fundamental question in systems biology. In this study, we assumed that a fixed network structure with parameters being estimated from real data could capture the context-specific alterations in cancer cells. Moreover, we have attempted to reconcile the parameter changes with network rewiring by predicting edge deletions from the activation rate constants that were inferred to be low. However, this may not be sufficient to fully address the issue, as some network rewiring events like the addition of a new edge cannot be represented by the change of parameters of the existing edges. Fourthly, different proteins in a network could have very different ranges of signaling activities, thus an appropriate normalization method would be needed to avoid potential bias in the optimization process of parameter inference from real data. This is more crucial when there is noise in the measurement, especially for the single-cell data such as the Yaffe’s dataset used in this study. Another long-standing challenge is how to validate computational predictions. In this study, we have mainly used qualitative comparisons of real and predicted time-course data and cited evidence from the literature. Such types

of validation are indirect and may not be conclusive in some cases. Using ABC-SysBio we tried to fit the model with experimental data by improving the overall goodness of fit based on Euclidean distances. However, as the optimization is approximate, some proteins may fit the corresponding real data well while for some other proteins the goodness of fit may be poor. At the same time different fitting patterns may result in similar scores. However, overall there is an improvement in the goodness of fit for every protein as compared to the random model. In order to assess the statistical significance, we performed Z -test and calculated the p -values (Table 4.1). Last but not the least, for the context-specific modeling of cancer cell fate and drug effects, we need to integrate diverse types of data. Here we have mainly used the phosphoproteomics data. Other types of data, such as somatic mutations, copy number variations, epigenetic modifications, gene expression, etc., should also be taken into account. Indeed, the fusion of dynamic network models with genomic information would be particularly important for the applications of computational techniques in precision medicine.

Another crucial issue that remains open for further investigation is about the inference of the model parameters. As the modeling of biochemical systems has become widespread trend in computational systems biology, researchers are often faced with the issue of calibrating mathematical models against the experimentally measured data. ABC-SysBio, that is a Linux/Mac based python package, is one such software suit that enables model calibration against the experimentally measured data. ABC-SysBio uses Bayesian framework to infer model parameters. The software is suitable for investigating the fitness of complex stochastic models to the experimental observations. The package includes several examples based on imaginary biochemical reaction network where each node and edge denote a protein and a biochemical reaction between two proteins (e.g. phosphorylation), respectively. Although ABC-SysBio provides a state-of-the-art computational approach, some of its computational aspects still need to be investigated. For example, as the model size increases, the computational speed of ABC-SysBio gradually becomes bottleneck for the parameter inference. So the software program may be upgraded to enhance its computational efficiency. In particular, we would like to extend the package by improving computational aspects from two directions:

- (i) By identifying the bottlenecks of running time;

- (ii) To design strategies to speed up computational cost of ABC-SysBio for larger biochemical network (nearly 80 proteins).

The computational speed issues can be addressed by parallel computing, reducing space of random sampling and by use of C routines instead of Python routines.

Chapter 5

Summary and future work

5.1 Summary

The main aim of the studies in this thesis has been to develop novel hybrid modeling approaches that combine data-driven and knowledge-driven modeling in the systems biology of cell fate.

The main contributions of this thesis in the form of proposed methods and obtained results can be summarised as:

- (i) In Chapter 3 we combine experimental data with a computational model to make novel predictions. Based on the literature a basic apoptotic signaling network N_1 , which represents pathways inside a healthy cell, was constructed and the corresponding ODE based computational model was encoded. The Genetic Algorithm (GA) was used to change the network topology. There were two phases of the application of the genetic algorithm. Initially, we used GA to rewire the basic network N_1 to obtain the cancer network N_2 , and subsequently to rewire N_2 to obtain the drug sensitive network N_3 . Within the framework of GA, we used the dynamic Time Warping (DTW) for optimizing the score based on the distance between the experimental and simulation data. We predict network rewirings happening inside the cancer cell that can be causal for the enhanced drug sensitivity.
- (ii) In Chapter 4 we performed context-specific modeling to investigate enhanced drug sensitivity in three diverse cell lines, *e.g.*, BT-20, MCF7, and MDA-MB-453. A signaling pathway was manually constructed using GeneGO MetaCore database

that contained 71 nodes, each representing the activity level of a protein (which could be a phosphatase or kinase protein). We derived a computational model by encoding the ODEs for the signaling network and inferred the network kinetic parameters using the ABC-SysBio software using Bayesian framework of inference. We predicted the context-specific apoptosis and assess the comparison between *in silico* and experimental apoptosis data using the Pearson correlation coefficient. We performed the caspases knockdown experiments for each cell line to investigate and understand their impact on enhanced drug sensitivity. Our proposed model can be scalable to diverse set of cancer cell lines.

As mentioned in Chapter 2, there have been quite a few hybrid modelling framework proposed earlier. However, the studies proposed in this thesis looks into a different aspect of healthcare with a focus on cancer drug sensitivity and provides several novel contributions, such as:

- (i) We computationally infer cancer and drug sensitive network by applying Genetic Algorithm based optimization
- (ii) We infer model parameters using phosphorylation data by using Bayesian framework of inference
- (iii) We predict computational apoptosis and assess the correlation with experimental data
- (iv) We predict pathways alterations in diverse cancer cell lines.
- (v) We predict network rewirings leading to the enhanced drug sensitivity
- (vi) We compare changes in the reaction rate constants in diverse cancer cell lines.

5.2 Future directions

Despite the promising performance, the potential of the proposed models has not been fully exploited yet and we hope that the further research in this field will yield more

fruitful results. Some of the preliminary studies conducted and the future enhancements of the studies proposed in this thesis are discussed in this section.

Basically, in this thesis we have investigated apoptosis using the Yaffe’s data set. However, we aim at constructing a general model for cancer cell phenotypic responses that can simulate observations for a diverse set of “omics” data taken from different sources. In future we shall continue to investigate various mechanisms for integrating high throughput single-cell gene profiling data and phosphorylation data with computational models to understand apoptosis in a more holistic way. Some of the possible enhancements of the work proposed in this thesis are summarised in the following.

5.2.1 Integrating alterations in network structure with parameter inference

In Chapter 3, we presented a genetic algorithm based approach to identify topological alterations leading to the enhanced cancer drug sensitivity. In Chapter 4, we provided a parameter inference based approach to investigate cancer drug effects. It would be interesting to combine the work presented in Chapter 3 and 4 to investigate drug induced changes in cell fates. The changes in a network topology can be in two forms

- (i) Addition of network edges
- (ii) Deletion of network edges.

In the Chapter 4 (Section 4.3.4), we suggested several alterations in the network topology owing to the blockage of the network edges corresponding to the low values for the activation rate constants. These alterations, however, can only represent the deletion of the network edges. We wish to emphasize that it would be challenging to infer network parameters using ABC-SysBio for each of the network topology changes (such as the changes induced by genetic algorithm in Chapter 3). For example, we need to perform rewiring, and then we need to run ABC-SysBio for parameter inference for the rewired network. For each of the changes in network topology, we need to infer network parameter that can take 4-5 hours. ABC-SysBio can infer network parameter for only one network topology at a time in 4-5 hours. These issues put a limitation for automating the parameter process inference with ABC-SysBio.

We, however, can certainly test at least a few rewirings (with edge additions) to see how rewirings can affect cancer drug sensitivity. For example, the Euclidean distance for the apoptosis after model calibration for the MCF7 cells (Figure 4.5 b) was much higher (58.98) as compared to the BT-20 (18.89) and MDA-MB-453 (10.25) cells. This inspired us to perform some random rewirings to assess goodness of fit for the MCF7 cells under the randomly rewired network conditions. We performed following rewiring experiments:

- (i) Addition of a network edge p53-CASP8.
- (ii) Addition of a network edges SMAC-CASP9.
- (iii) Addition of both the network edges p53-CASP8 and SMAC-CASP9 simultaneously.

The predicted apoptosis for aforementioned rewiring conditions are shown in Figure 5.1. An increase in apoptosis level was observed for each of the rewirings. The combined effect of the two edges addition on apoptosis was higher than the individual rewiring. There are literature evidences for p53 induced activation of apoptosis pathways following caspases activation [122] in cancer cells. The study in [123] have found that p53 can be upstream regulator of CASP8 in cancer cells. We could also find evidences in MCF7 [124] where Caspase can be activated following Erlotinib drug treatment. The studies in [125][126] have reported the role of SMAC protein in CASP9 activation. SMAC promotes apoptosis by bindings to the inhibitor of apoptosis proteins (IAPs) and suppresses their inhibitory activities, which eventually activates CASP9.

Our experiments suggests that combining the rewiring events with the parameters inference techniques could be a promising approach to understand and investigate the enhanced cancer drug sensitivity. Although our findings are preliminary, such experiments can be conducted for various combinations of combined effects of rewiring events with parameter inference. Moreover, these *in silico* rewirings events further needs to be verified with *in vivo* experiments, or with the cell line specific data.

5.2.2 Cancer attractors

The idea of using system attractors (*i.e.* representing stable states in a dynamic system) to represent cell types was initially advocated by Stuart Kauffman [127]. For a large

Figure 5.1: Prediction of apoptosis after rewiring and the parameters inference in MCF7 cell line. Comparison of the topological rewiring induced changes in apoptosis with the experimental apoptosis data from [12] for MCF7 cells. The newly added edges in Figure 4.2 were **a)** p53-CASP8, **b)** Smac-CASP9, and **c)** both p53-CASP8 and SMAC-CASP9, simultaneously. After addition of new edges in Figure 4.2, the computational model was able to predict the apoptosis in MCF7 cell line with improved Euclidean distance as compared to Figure 4.5 (b), such as **a)** p53-CASP8 (55.65), **b)** SMAC-CASP9 (47.11), **c)** both p53-CASP8 and SMAC-CASP9 (42.09). The Euclidian distance after rewiring between simulated and experimental apoptosis data for MCF7 was significantly improved.

amount of perturbation, the system might move from one attractor state to another. The study in [32] investigated the concept of “cancer attractors” and proposed that the tumour cells are trapped inside abnormal attractors. The study explained and linked the process of cellular development with oncogenesis while highlighting the roles of genetic and non-genetic aberrations. As per the theory in [32], a lower quasi-potential means a higher stability.

The classical concept of Waddington’s epigenetic landscape, where the height of the landscape is represented by the quasi-potential, can be used for investigating the context-specific stability of cancer cells. We believe that another possible application for hybrid modeling approach could be to investigate cancer attractors and plot the 3-Dimensional landscape.

The Quasi-potential for the Yaffe’s dataset can be calculated using the Taylor’s expansion and other concepts from physics, as per the techniques described in [10] and in the book [128]. For example, suppose that in a dynamic system (such as two gene network) represented by two variables u and v , there is a steady state point (u', v') where the potential $V(u', v') = 0$. If we consider a slight disturbance for a given stable state,

the system will reach some nearby point $(u' + \Delta u, v' + \Delta v)$, where both Δu and Δv are extremely small quantities. Let's say the new potential is $V(u' + \Delta u, v' + \Delta v)$. The change in potential, between point (u', v') and point $(u' + \Delta u, v' + \Delta v)$, ΔV is given by

$$\Delta V = V(u' + \Delta u, v' + \Delta v) - V(u', v'). \quad (5.1)$$

$$\Delta V = V(u' + \Delta u, v' + \Delta v) - V(u', v'). \quad (5.2)$$

However, we have $V(u', v') = 0$. So we obtain

$$\Delta V = V(u' + \Delta u, v' + \Delta v). \quad (5.3)$$

Using the Taylor's two variable series expansion we get

$$\Delta V = V(u', v') + \Delta u \cdot \frac{\partial V}{\partial u} + \Delta v \cdot \frac{\partial V}{\partial v} + \text{higher order terms}. \quad (5.4)$$

The higher order terms, denoting quadratically small terms that involve Δu^2 , Δv^2 , $\Delta u \cdot \Delta v$, could be ignored. Also, we have $V(u', v') = 0$. So we obtain:

$$\Delta V = \Delta u \cdot \frac{\partial V}{\partial u} + \Delta v \cdot \frac{\partial V}{\partial v}. \quad (5.5)$$

However, when a particle is sliding down the walls of a potential well, the potential V is defined by [128]

$$\frac{\partial V}{\partial u} = -\frac{du}{dt}, \quad (5.6)$$

$$\frac{\partial V}{\partial v} = -\frac{dv}{dt}. \quad (5.7)$$

So, we obtain:

$$\Delta V = -\Delta u \cdot \frac{du}{dt} - \Delta v \cdot \frac{dv}{dt}, \quad (5.8)$$

Also,

$$\Delta u = \frac{du}{dt} \cdot dt, \quad (5.9)$$

$$\Delta v = \frac{dv}{dt} \cdot dt, \quad (5.10)$$

(i.e. distance = gradient $\times$ time). This gives us

$$\Delta V = -\left(\frac{du}{dt} \cdot dt\right)^2 - \left(\frac{dv}{dt} \cdot dt\right)^2. \quad (5.11)$$

Similarly, we can follow the Taylor's expansion for three variables to calculate the quasi-potential.

$$\Delta V = V(u' + \Delta u, v' + \Delta v, w' + \Delta w) - V(u', v', w'). \quad (5.12)$$

Again, $V(u', v', w') = 0$, and if we ignore the higher order terms, then we obtain

$$\Delta V = -\left(\frac{du}{dt} \cdot dt\right)^2 - \left(\frac{dv}{dt} \cdot dt\right)^2 - \left(\frac{dw}{dt} \cdot dt\right)^2. \quad (5.13)$$

The negative sign in the definition of quasi-potential (V) follows the standard convention in physics; it implies that the particle always moves “downhill” as the motion proceeds. Thus V decreases along trajectories, and so the particle always moves toward lower potential. The study in [10] calculated the quasi-potential for two gene network. However, this procedure can be generalized to a larger number of variables and we shall expand this procedure to calculate quasi-potential for a larger network. Since now we have the network (Figure 4.2) which was trained using both cancer and drug sensitive data, we want to compare the context-specific quasi-potential values for both the control and drug treatment datasets. Since we have data for six types of treatments (DMSO, TAR, DOX, DT, T-D and D-T), we will compare the corresponding quasi-potentials for each treatment.

For the drug sensitive data, it will be interesting to investigate how the quasi potential values fluctuate with time. It might be the case that the values are found to be decreasing with time (such as reaching the lowest at the 7th time point compared to the initial cell states). Since each cell fate can correspond to an attractor, we can say that, as the time progresses, the drug sensitive cells are reaching a stable (attractor) state that can be apoptosis.

5.2.3 Synthetic lethality

Synthetic lethal (SL) genetic interactions were first reported in *Drosophila* [129]. The SL interaction between a gene pair leads to the lethal phenotype when both genes are mutated simultaneously. However, mutation of single gene alone does not lead to the lethality. This concept of synthetic lethality can be used for anti-cancer targeted therapeutic effects by identifying the cancer-specific inactive gene and then inhibiting its SL partner.

Several quantitative approaches have been proposed that expanded the concept of SL by further integrating knowledge of genetic epistasis data [130], overexpression [131], mutations [129], etc. Several computational approaches have been proposed to investigate the SL genetic interactions [132][131][133]. For example, the study in [133] used data-driven computational approach to predict the SL interactions. However, in most of the studies, SL interactions are usually detected by evaluating correspondence between synthetic lethal interactions in yeast and their human orthologs. Though previous studies have enhanced our understanding of SL interactions, our understanding about SL interactions are far from complete. For example, previous studies have not incorporated signaling activity levels of the proteins (e.g. phosphorylation) and the phenotypic responses based interactions. We have not come across any paper using a large network that predicts SL interactions by combining phosphorylation and unphosphorylation data with ODEs based model. Hence, new computational methods are required to further investigate and complement the anti-cancer research with a focus on SL interactions.

One application of the hybrid modeling approach could be to investigate Synthetic Lethality. In our previous study [134], we constructed a synthetic lethal knowledge base. In another study [50], we constructed a signaling network. In our future study, we propose an integrative hybrid modeling approach for identifying SL gene pairs in different cancer cell lines. The ODE-based computational model can be calibrated against experimentally measured cellular signaling responses using Bayesian framework. We also aim to use the DMSO (control) dataset as the activity level of signaling protein for the control treatment data is closest to the activity level of proteins in the cancer cells. Next, we will perform single and double gene knock-down experiments to estimate the corresponding variations in the phenotypic responses. Then we can predict synthetic lethality based on the single

and double knockdown effects for all possible pairs for measured proteins. Based on the effect of knockdowns for different proteins, we will rank the SL gene pairs. The predicted SL gene pairs shall be validated using experimental evidences, existing literature and published databases. Our investigation aims at enhancing the interpretations of SL interactions that are directly linked with signaling pathways, and thereby extending the computational approach to model synthetic lethality for the cancer cells.

5.2.4 Genomic mutations

Alterations in the genomic sequences have long been held as the causal for tumorigenesis. There are several types of mutations (*e.g.* substitution-missense, deletion-frameshift, substitution-coding silent, whole gene deletion, insertion-frameshift, deletion-in frame, etc.) [30] that can stimulate the development of cancer [18]. In fact the same type of mutation may have different disease phenotypic responses. For example, if there are two individuals containing the same mutations, one may be affected by the disease, whereas other may be totally unaffected [135]. The data sources such as TCGA [29], COSMIC [30] etc. contain high-quality genomic sequencing and mutation data. However, only a few computational studies have modeled the roles of genomic mutations on cellular phenotypic responses using systems biology modeling. Especially, in the case of cancer that can result due to genetic heterogeneity, where more than one mutation can be responsible for the same cancer type, it would be interesting to investigate the roles of genomic mutations on network dynamics and cell fates. This inspires us to perform context-specific analysis of the cancer cells using hybrid modeling approach that can integrate high throughput “omics” data with computational models. We also aim to combine genomic mutation data with signaling data and predict the mutation-related effects on cell signaling and vice versa. The role of specific mutation on signaling network and the phenotypes needs to be investigated. For example, some mutations might create new interaction of the signaling molecules, whereas other mutation may lead to loss of interactions.

The study in [136] proposed a Multiscale statistical mechanical (MSM) framework. The framework established a quantitative relation between energy of a system state s ,

known as the Hamiltonian $H(s)$, with the probability of the occurrence $P(s)$ of the state s . In simple systems when the $H(s)$ is known, $P(s)$ can be calculated as

$$P(s) = \frac{1}{Z} \cdot e^{-\frac{1}{k \cdot T} \cdot H(s)} \quad (5.14)$$

where Z is a normalizing constant (the partition function), k is the Boltzmann constant, T the temperature of the system. However, in a complex systems such as signaling networks, it is difficult to compute $H(s)$ from the physical principles. In such system $H(s)$ can be computed using machine learning techniques where $H(s)$ can be inferred using the domain level binding data *e.g.* association and dissociation constants between SH2 domain and phosphotyrosine (pY) complexes. The learned $H(s)$ can be used to make predictions for the Protein-level quantities. The Protein-level quantities can be used to predict the Network level quantities. The approach could be to propagate the mutational changes through the following layers:

- (i) Residue level (amino acids) changes
- (ii) Domain-level
- (iii) Protein level
- (iv) Network level

The main idea is to link the genomic changes happening inside the nucleus (*e.g.* amino acids changes) to the cell surface signaling (*i.e.*, EGFR) level changes. The network can be constructed by integrating residue level (amino acids) changes in the model. So if there is any genomic mutation (such as, mutation in the amino acid sequences), the changes can be reflected by the alterations in network structure. Based on this framework, we can show the impact of changes in nucleus (such as new mutation) on the network structure and dynamics [136]. The study in [136] provides a comprehensive details along with the different types of data to predict the effects of the mutations on a signaling network structure. We can also integrate changes happening in the nucleus (*e.g.* gene expression data) with the changes happening in the cytoplasm (*e.g.* signaling data) using an ODEs based model. An ODE equation in such models can integrate dynamics of gene expression

with the changes in the signaling patterns. For example, the study in [40] provides an ODEs based model for the EGFR pathways to link the signaling proteins with the GRN inside nucleus. However, it's a knowledge-driven study, suggesting an opportunity to integrate the “omics” data with the model.

References

- [1] Z. Bu and D. Callaway, “Proteins move! Protein dynamics and long-range allostery in cell signaling,” *Adv Protein Chem Struct Biol*, vol. 83, pp. 163–221, 2011.
- [2] H. M. Wain, E. A. Bruford, R. C. Lovering, M. J. Lush, M. W. Wright, and S. Povey, “Guidelines for human gene nomenclature,” *Genomics*, vol. 79, no. 4, pp. 464–470, 2002.
- [3] H. Pearson, “Genetics: what is a gene?,” *Nature*, vol. 441, no. 7092, pp. 398–401, 2006.
- [4] E. Pennisi, “DNA study forces rethink of what it means to be a gene,” *Science*, vol. 316, no. 5831, pp. 1556–1557, 2007.
- [5] S. Clancy, “Rna splicing: introns, exons and spliceosome,” *Nature Education*, vol. 1, no. 1, p. 31, 2008.
- [6] C. Neill, “Biology,” *Fourth edition. The Benjamin/Cummings Publishing Company. p. 309,310*, 1996.
- [7] L. Stryer, “Biochemistry,” *Fifth edition. W. H. Freeman and Company. p. 826.*, 2002.
- [8] A. L. Demain, “Microbial production of primary metabolites,” *Naturwissenschaften*, vol. 67, no. 12, pp. 582–587, 1980.
- [9] J. M. Berg, J. L. Tymoczko, and L. Stryer, “Biochemistry. 5th,” 2002.
- [10] S. Bhattacharya, Q. Zhang, and M. E. Andersen, “A deterministic map of waddington’s epigenetic landscape for cell fate specification,” *BMC systems biology*, vol. 5, no. 1, p. 1, 2011.

REFERENCES

- [11] H. Resat, L. Petzold, and M. F. Pettigrew, “Kinetic modeling of biological systems,” *Computational Systems Biology*, pp. 311–335, 2009.
- [12] M. J. Lee, S. Y. Albert, 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*, vol. 149, no. 4, pp. 780–794, 2012.
- [13] N. Gehlenborg, S. I. O’Donoghue, N. S. Baliga, A. Goesmann, M. A. Hibbs, H. Kitano, O. Kohlbacher, H. Neuweger, R. Schneider, D. Tenenbaum, *et al.*, “Visualization of omics data for systems biology,” *Nature methods*, vol. 7, pp. S56–S68, 2010.
- [14] 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 Comput Biol*, vol. 6, no. 3, p. e1000702, 2010.
- [15] “Defining cancer,” *National Cancer Institute*. Retrieved 24 June 2016, 2014.
- [16] “World cancer report 2014,” *World Health Organization*, Chapter 1.1, 2014.
- [17] D. Hanahan and R. A. Weinberg, “The hallmarks of cancer,” *cell*, vol. 100, no. 1, pp. 57–70, 2000.
- [18] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, *et al.*, “Signatures of mutational processes in human cancer,” *Nature*, vol. 500, no. 7463, pp. 415–421, 2013.
- [19] M. R. Stratton, P. J. Campbell, and P. A. Futreal, “The cancer genome,” *Nature*, vol. 458, no. 7239, pp. 719–724, 2009.
- [20] 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 Comput Biol*, vol. 5, no. 12, p. e1000591, 2009.

- [21] S. Ekins, Y. Nikolsky, A. Bugrim, E. Kirillov, and T. Nikolskaya, “Pathway mapping tools for analysis of high content data,” *High content screening: A powerful approach to systems cell biology and drug discovery*, pp. 319–350, 2006.
- [22] N. Juty, R. Ali, M. Glont, S. Keating, N. Rodriguez, M. Swat, S. Wimalaratne, H. Hermjakob, N. Le Novère, C. Laibe, *et al.*, “BioModels: content, features, functionality, and use,” *CPT: pharmacometrics & systems pharmacology*, vol. 4, no. 2, pp. 55–68, 2015.
- [23] A. Funahashi, Y. Matsuoka, A. Jouraku, M. Morohashi, N. Kikuchi, and H. Kitano, “CellDesigner 3.5: a versatile modeling tool for biochemical networks,” *Proceedings of the IEEE*, vol. 96, no. 8, pp. 1254–1265, 2008.
- [24] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, “KEGG as a reference resource for gene and protein annotation,” *Nucleic acids research*, vol. 44, no. D1, pp. D457–D462, 2016.
- [25] E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, Ö. Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander, “Pathway Commons, a web resource for biological pathway data,” *Nucleic acids research*, vol. 39, no. suppl 1, pp. D685–D690, 2011.
- [26] K. I. Kim and R. Simon, “Using single cell sequencing data to model the evolutionary history of a tumor,” *BMC bioinformatics*, vol. 15, no. 1, p. 1, 2014.
- [27] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, *et al.*, “NCBI GEO: archive for functional genomics data setsupdate,” *Nucleic acids research*, vol. 41, no. D1, pp. D991–D995, 2013.
- [28] U. Consortium *et al.*, “Uniprot: a hub for protein information,” *Nucleic acids research*, p. gku989, 2014.
- [29] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, C. G. A. R. Network, *et al.*,

REFERENCES

- “The cancer genome atlas pan-cancer analysis project,” *Nature genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.
- [30] S. A. Forbes, D. Beare, P. Gunasekaran, K. Leung, N. Bindal, H. Boutselakis, M. Ding, S. Bamford, C. Cole, S. Ward, *et al.*, “COSMIC: exploring the world’s knowledge of somatic mutations in human cancer,” *Nucleic acids research*, vol. 43, no. D1, pp. D805–D811, 2015.
- [31] K. Buetow, R. Strausberg, E.-B. MR, and K. RD, “The cancer genome anatomy project: Building an annotated gene index,” *Trends Genet.* 16: 103106, 2000.
- [32] S. Huang, I. Ernberg, and S. Kauffman, “Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective,” in *Seminars in cell & developmental biology*, vol. 20, pp. 869–876, Elsevier, 2009.
- [33] S. Elmore, “Apoptosis: a review of programmed cell death,” *Toxicologic pathology*, vol. 35, no. 4, pp. 495–516, 2007.
- [34] A. Naldi, D. Berenguier, A. Fauré, F. Lopez, D. Thieffry, and C. Chaouiya, “Logical modelling of regulatory networks with GINsim 2.3,” *Biosystems*, vol. 97, no. 2, pp. 134–139, 2009.
- [35] 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 cis-platin,” *BMC systems biology*, vol. 6, no. 1, p. 1, 2012.
- [36] H. J. Huber, M. Rehm, M. Plchut, H. Düssmann, and J. H. Prehn, “APOPTO-CELL-a simulation tool and interactive database for analyzing cellular susceptibility to apoptosis,” *Bioinformatics*, vol. 23, no. 5, pp. 648–650, 2007.
- [37] C. Gu, J. Zhang, Y. Chen, and J. Lei, “A trigger model of apoptosis induced by tumor necrosis factor signaling,” *BMC systems biology*, vol. 5, no. 1, p. 1, 2011.
- [38] M. Chaves, T. Eissing, and F. Allgöwer, “Regulation of apoptosis via the NF κ B pathway: modeling and analysis,” in *Dynamics On and Of Complex Networks*, pp. 19–33, Springer, 2009.

- [39] 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 Comput Biol*, vol. 9, no. 10, p. e1003286, 2013.
- [40] M. Mol, M. S. Patole, and S. Singh, “Immune signal transduction in leishmaniasis from natural to artificial systems: role of feedback loop insertion,” *Biochimica et Biophysica Acta (BBA)-General Subjects*, vol. 1840, no. 1, pp. 71–79, 2014.
- [41] K. A. Janes and M. B. Yaffe, “Data-driven modelling of signal-transduction networks,” *Nature Reviews Molecular Cell Biology*, vol. 7, no. 11, pp. 820–828, 2006.
- [42] M. Niepel, M. Hafner, E. A. Pace, M. Chung, D. H. Chai, L. Zhou, B. Schoeberl, and P. K. Sorger, “Profiles of basal and stimulated receptor signaling networks predict drug response in breast cancer lines,” *Science signaling*, vol. 6, no. 294, 2013.
- [43] M. Niepel, M. Hafner, E. A. Pace, M. Chung, D. H. Chai, L. Zhou, J. L. Muhlich, B. Schoeberl, and P. K. Sorger, “Analysis of growth factor signaling in genetically diverse breast cancer lines,” *BMC biology*, vol. 12, no. 1, p. 20, 2014.
- [44] C. Cai, L. Chen, X. Jiang, and X. Lu, “Modeling signal transduction from protein phosphorylation to gene expression,” *Cancer Informatics*, vol. 13, no. Suppl 1, p. 59, 2014.
- [45] P. Lopez-Bergami, C. Huang, J. S. Goydos, D. Yip, M. Bar-Eli, M. Herlyn, K. S. Smalley, A. Mahale, A. Eroshkin, S. Aaronson, *et al.*, “Rewired erk-jnk signaling pathways in melanoma,” *Cancer cell*, vol. 11, no. 5, pp. 447–460, 2007.
- [46] T. Pawson and N. Warner, “Oncogenic re-wiring of cellular signaling pathways,” *Oncogene*, vol. 26, no. 9, pp. 1268–1275, 2007.
- [47] D. Dubois, P. Hájek, and H. Prade, “Knowledge-driven versus data-driven logics,” *Journal of logic, Language and information*, vol. 9, no. 1, pp. 65–89, 2000.

REFERENCES

- [48] P. Creixell, E. M. Schoof, C. D. Simpson, J. Longden, C. J. Miller, H. J. Lou, L. Perryman, T. R. Cox, N. Zivanovic, A. Palmeri, *et al.*, “Kinome-wide decoding of network-attacking mutations rewiring cancer signaling,” *Cell*, vol. 163, no. 1, pp. 202–217, 2015.
- [49] F. Zhang, M. Wu, X. J. Li, X. 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*, vol. 13, no. 03, p. 1541002, 2015. PMID: 25669329.
- [50] 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*, vol. 9, no. 1, p. 1, 2015.
- [51] 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*, vol. 5, no. 1, pp. 1–12, 2011.
- [52] J. Sun, J. Hu, D. Luo, M. Markatou, F. Wang, S. Ebadollahi, Z. Daar, and W. F. Stewart, “Combining knowledge and data driven insights for identifying risk factors using electronic health records,” in *AMIA*, vol. 2012, pp. 901–10, 2012.
- [53] A. Dehghan, A. Kovacevic, G. Karystianis, J. A. Keane, and G. Nenadic, “Combining knowledge- and data-driven methods for de-identification of clinical narratives,” *Journal of Biomedical Informatics*, vol. 58, Supplement, pp. S53 – S59, 2015. Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.
- [54] M. Peifer and J. Timmer, “Parameter estimation in ordinary differential equations for biochemical processes using the method of multiple shooting,” *Systems Biology, IET*, vol. 1, no. 2, pp. 78–88, 2007.
- [55] J. O. Ramsay, G. Hooker, D. Campbell, and J. Cao, “Parameter estimation for differential equations: a generalized smoothing approach,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 5, pp. 741–796, 2007.

REFERENCES

- [56] H. Liang and H. Wu, “Parameter estimation for differential equation models using a framework of measurement error in regression models,” *Journal of the American Statistical Association*, 2012.
- [57] T. Zhang, P. Brazhnik, and J. J. Tyson, “Computational analysis of dynamical responses to the intrinsic pathway of programmed cell death,” *Biophysical journal*, vol. 97, no. 2, pp. 415–434, 2009.
- [58] E. Bagci, Y. Vodovotz, T. Billiar, G. Ermentrout, and I. Bahar, “Bistability in apoptosis: roles of bax, bcl-2, and mitochondrial permeability transition pores,” *Biophysical journal*, vol. 90, no. 5, pp. 1546–1559, 2006.
- [59] C. Chen, J. Cui, H. Lu, R. Wang, S. Zhang, and P. Shen, “Modeling of the role of a Bax-activation switch in the mitochondrial apoptosis decision,” *Biophysical Journal*, vol. 92, no. 12, pp. 4304–4315, 2007.
- [60] T. Eißing, F. Allgöwer, and E. Bullinger, “Robustness properties of apoptosis models with respect to parameter variations and intrinsic noise,” *IEEE Proceedings-Systems Biology*, vol. 152, no. 4, pp. 221–228, 2005.
- [61] B. Efron, “Bayes’ theorem in the 21st century,” *Science*, vol. 340, no. 6137, pp. 1177–1178, 2013.
- [62] J. Liepe, P. Kirk, S. Filippi, T. Toni, C. P. Barnes, and M. P. Stumpf, “A framework for parameter estimation and model selection from experimental data in systems biology using approximate bayesian computation,” *Nature protocols*, vol. 9, no. 2, pp. 439–456, 2014.
- [63] V. Vyshemirsky and M. Girolami, “BioBayes: a software package for Bayesian inference in systems biology,” *Bioinformatics*, vol. 24, no. 17, pp. 1933–1934, 2008.
- [64] J.-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder, “Approximate Bayesian computational methods,” *Statistics and Computing*, vol. 22, no. 6, pp. 1167–1180, 2012.

REFERENCES

- [65] T. W. Kensler, A. Spira, J. E. Garber, E. Szabo, J. J. Lee, Z. Dong, A. J. Dannenberg, W. N. Hait, E. Blackburn, N. E. Davidson, *et al.*, “Transforming cancer prevention through precision medicine and immune-oncology,” *Cancer Prevention Research*, vol. 9, no. 1, pp. 2–10, 2016.
- [66] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, *et al.*, “The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity,” *Nature*, vol. 483, no. 7391, pp. 603–607, 2012.
- [67] B.-J. Chen, O. Litvin, L. Ungar, and D. Peer, “Context sensitive modeling of cancer drug sensitivity,” *PloS one*, vol. 10, no. 8, p. e0133850, 2015.
- [68] F. Gnad, J. Estrada, and J. Gunawardena, “Proteus: a web-based, context-specific modelling tool for molecular networks,” *Bioinformatics*, vol. 28, no. 9, pp. 1284–1286, 2012.
- [69] R. Fisher and J. Larkin, “Vemurafenib: a new treatment for braf-v600 mutated advanced melanoma,” *Cancer management and research*, vol. 4, p. 243, 2012.
- [70] D. Barras, “Braf mutation in colorectal cancer: an update,” *Biomarkers in cancer*, vol. 7, no. Suppl 1, p. 9, 2015.
- [71] U. Braga-Neto, R. Hashimoto, E. R. Dougherty, D. V. Nguyen, and R. J. Carroll, “Is cross-validation better than resubstitution for ranking genes?,” *Bioinformatics*, vol. 20, no. 2, pp. 253–258, 2004.
- [72] S. Baek, C.-A. Tsai, and J. J. Chen, “Development of biomarker classifiers from high-dimensional data,” *Briefings in bioinformatics*, vol. 10, no. 5, pp. 537–546, 2009.
- [73] C. Bernau, M. Riester, A.-L. Boulesteix, G. Parmigiani, C. Huttenhower, L. Waldron, and L. Trippa, “Cross-study validation for the assessment of prediction algorithms,” *Bioinformatics*, vol. 30, no. 12, pp. i105–i112, 2014.

REFERENCES

- [74] B. R. King, S. Vural, S. Pandey, A. Barteau, and C. Guda, “ngLOC: software and web server for predicting protein subcellular localization in prokaryotes and eukaryotes,” *BMC research notes*, vol. 5, no. 1, p. 1, 2012.
- [75] N. K. Chaturvedi, R. A. Mir, V. Band, S. S. Joshi, and C. Guda, “Experimental validation of predicted subcellular localizations of human proteins,” *BMC research notes*, vol. 7, no. 1, p. 1, 2014.
- [76] R. L. Carpenter and H.-W. Lo, “Regulation of apoptosis by HER2 in breast cancer,” *Journal of carcinogenesis & mutagenesis*, vol. 2013, no. Suppl 7, 2013.
- [77] J. Yu and L. Zhang, “PUMA, a potent killer with or without p53,” *Oncogene*, vol. 27, pp. S71–S83, 2008.
- [78] C. Weng, Y. Li, D. Xu, Y. Shi, and H. Tang, “Specific cleavage of Mcl-1 by caspase-3 in tumor necrosis factor-related apoptosis-inducing ligand (TRAIL)-induced apoptosis in Jurkat leukemia T cells,” *Journal of Biological Chemistry*, vol. 280, no. 11, pp. 10491–10500, 2005.
- [79] E. Buytaert, G. Callewaert, J. Vandenheede, and P. Agostinis, “Deficiency in apoptotic effectors Bax and Bak reveals an autophagic cell death pathway initiated by photodamage to the endoplasmic reticulum,” *Autophagy*, vol. 2, no. 3, pp. 238–240, 2006.
- [80] K. Lei, A. Nimnual, W.-X. Zong, N. J. Kennedy, R. A. Flavell, C. B. Thompson, D. Bar-Sagi, and R. J. Davis, “The Bax subfamily of Bcl2-related proteins is essential for apoptotic signal transduction by c-Jun NH2-terminal kinase,” *Molecular and cellular biology*, vol. 22, no. 13, pp. 4929–4942, 2002.
- [81] E. Zalckvar, H. Berissi, L. Mizrachy, Y. Idelchuk, I. Koren, M. Eisenstein, H. Sabanay, R. Pinkas-Kramarski, and A. Kimchi, “DAP-kinase-mediated phosphorylation on the BH3 domain of beclin 1 promotes dissociation of beclin 1 from Bcl-XL and induction of autophagy,” *EMBO reports*, vol. 10, no. 3, pp. 285–292, 2009.
- [82] R. T. Marquez and L. Xu, “Bcl2: Beclin 1 complex: multiple mechanisms regulating autophagy/apoptosis toggle switch,” 2012.

REFERENCES

- [83] S. Bhattacharya, R. M. Ray, and L. R. Johnson, “STAT3-mediated transcription of Bcl-2, Mcl-1 and c-IAP2 prevents apoptosis in polyamine-depleted cells,” *Biochemical Journal*, vol. 392, no. 2, pp. 335–344, 2005.
- [84] H.-J. Choi and J.-S. Han, “Overexpression of phospholipase D enhances Bcl-2 expression by activating STAT3 through independent activation of ERK and p38MAPK in HeLa cells,” *Biochim. Biophys. Acta.*, p. Pages 10821091., 2012.
- [85] G. Den Haese, N. Walworth, A. Carr, and K. Gould, “The Wee1 protein kinase regulates T14 phosphorylation of fission yeast Cdc2,” *Molecular biology of the cell*, vol. 6, no. 4, pp. 371–385, 1995.
- [86] N. Pathan, C. Aime-Sempe, S. Kitada, S. Haldar, and J. C. Reed, “Microtubule-targeting drugs induce bcl-2 phosphorylation and association with Pin1,” *Neoplasia*, vol. 3, no. 1, pp. 70–79, 2001.
- [87] N. Prenzel, O. Fischer, S. Streit, S. Hart, and A. Ullrich, “The epidermal growth factor receptor family as a central element for cellular signal transduction and diversification,” *Endocrine-related cancer*, vol. 8, no. 1, pp. 11–31, 2001.
- [88] C. Yu, Y. Minemoto, J. Zhang, J. Liu, F. Tang, T. N. Bui, J. Xiang, and A. Lin, “JNK suppresses apoptosis via phosphorylation of the proapoptotic Bcl-2 family protein BAD,” *Molecular cell*, vol. 13, no. 3, pp. 329–340, 2004.
- [89] Z. Jin and W. S. El-Deiry, “Distinct signaling pathways in trail-versus tumor necrosis factor-induced apoptosis,” *Molecular and cellular biology*, vol. 26, no. 21, pp. 8136–8148, 2006.
- [90] K. S. Ng, “A simple explanation of partial least squares,” 2013.
- [91] A. Monecke and F. Leisch, “sempls: structural equation modeling using partial least squares,” 2012.
- [92] Mathworks., “ode45:solve nonstiff differential equations,” *User’s Guide (r2014b)*, Retrieved June 28, 2016 from www.mathworks.com.

REFERENCES

- [93] S. Legewie, N. Blüthgen, and H. Herzel, “Mathematical modeling identifies inhibitors of apoptosis as mediators of positive feedback and bistability,” *PLoS Comput Biol*, vol. 2, no. 9, p. e120, 2006.
- [94] 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*, vol. 30, no. 1, pp. 11–25, 2008.
- [95] G. Jones, “Genetic and evolutionary algorithms,” *Encyclopedia of Computational Chemistry*, vol. 2, pp. 1127–1136, 1998.
- [96] J. Kohonen., “A brief comparison of simulated annealing and genetic algorithm approaches,” 1999, Retrieved June 27, 2016 from <http://www.cs.helsinki.fi/u/kohonen/papers/gasa.html>.
- [97] D. Whitley, “A genetic algorithm tutorial,” *Statistics and computing*, vol. 4, no. 2, pp. 65–85, 1994.
- [98] E. Keogh and C. A. Ratanamahatana, “Exact indexing of dynamic time warping,” *Knowledge and information systems*, vol. 7, no. 3, pp. 358–386, 2005.
- [99] G. Al-Naymat, S. Chawla, and J. Taheri, “SparseDTW: a novel approach to speed up dynamic time warping,” in *Proceedings of the Eighth Australasian Data Mining Conference-Volume 101*, pp. 117–127, Australian Computer Society, Inc., 2009.
- [100] D. J. Pannell, “Sensitivity analysis: strategies, methods, concepts, examples,” *Agric Econ*, vol. 16, pp. 139–152, 1997.
- [101] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, *et al.*, “Optimization by simulated annealing,” *science*, vol. 220, no. 4598, pp. 671–680, 1983.
- [102] W. L. Goffe, G. D. Ferrier, and J. Rogers, “Global optimization of statistical functions with simulated annealing,” *Journal of econometrics*, vol. 60, no. 1-2, pp. 65–99, 1994.

REFERENCES

- [103] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, *et al.*, “Gene ontology: tool for the unification of biology,” *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.
- [104] B. Vogelstein and K. W. Kinzler, “Cancer genes and the pathways they control,” *Nature medicine*, vol. 10, no. 8, pp. 789–799, 2004.
- [105] N. J. Camp, M. Parry, S. Knight, R. Abo, G. Elliott, S. H. Rigas, S. P. Balasubramanian, M. W. Reed, H. McBurney, A. Latif, *et al.*, “Fine-mapping casp8 risk variants in breast cancer,” *Cancer Epidemiology Biomarkers & Prevention*, vol. 21, no. 1, pp. 176–181, 2012.
- [106] K. Szczepanek, E. J. Lesnefsky, and A. C. Larner, “Multi-tasking: nuclear transcription factors with novel roles in the mitochondria,” *Trends in cell biology*, vol. 22, no. 8, pp. 429–437, 2012.
- [107] Q. L. Deveraux, N. Roy, H. R. Stennicke, T. Van Arsedale, Q. Zhou, S. M. Srinivasula, E. S. Alnemri, G. S. Salvesen, and J. C. Reed, “IAPs block apoptotic events induced by caspase-8 and cytochrome c by direct inhibition of distinct caspases,” *The EMBO journal*, vol. 17, no. 8, pp. 2215–2223, 1998.
- [108] S. Miscia, M. Marchisio, A. Grilli, V. Di Valerio, L. Centurione, G. Sabatino, F. Garaci, G. Zauli, E. Bonvini, and A. Di Baldassarre, “Tumor necrosis factor alpha (TNF-alpha) activates Jak1/Stat3-Stat5B signaling through TNFR-1 in human B cells,” *Cell growth & differentiation: the molecular biology journal of the American Association for Cancer Research*, vol. 13, no. 1, pp. 13–18, 2002.
- [109] T. d. L. M. Rouge, L. Galluzzi, K. A. Olaussen, Y. Zermati, E. Tasdemir, T. Robert, H. Ripoche, V. Lazar, P. Dessen, F. Harper, *et al.*, “A novel epidermal growth factor receptor inhibitor promotes apoptosis in non-small cell lung cancer cells resistant to erlotinib,” *Cancer research*, vol. 67, no. 13, pp. 6253–6262, 2007.
- [110] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, *et al.*, “The reactome pathway knowledgebase,” *Nucleic acids research*, vol. 42, no. D1, pp. D472–D477, 2014.

REFERENCES

- [111] D. Hanahan and R. A. Weinberg, “Hallmarks of cancer: the next generation,” *cell*, vol. 144, no. 5, pp. 646–674, 2011.
- [112] D. W. C. Li, J. P. Liu, Y. W. Mao, H. Xiang, J. Wang, W. Y. Ma, Z. Dong, H. M. Pike, R. E. Brown, and J. C. Reed, “Calcium-activated RAF/MEK/ERK signaling pathway mediates p53-dependent apoptosis and is abrogated by α B-crystallin through inhibition of RAS activation,” *Molecular biology of the cell*, vol. 16, no. 9, pp. 4437–4453, 2005.
- [113] R. M. Neve, K. Chin, J. Fridlyand, J. Yeh, F. L. Baehner, T. Fevr, L. Clark, N. Bayani, J. P. Coppe, F. Tong, *et al.*, “A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes,” *Cancer cell*, vol. 10, no. 6, pp. 515–527, 2006.
- [114] D. L. Holliday and V. Speirs, “Choosing the right cell line for breast cancer research,” *Breast Cancer Res*, vol. 13, no. 4, p. 215, 2011.
- [115] G. Sollberger, G. E. Strittmatter, M. Garstkiewicz, J. Sand, and H.-D. Beer, “Caspase-1: the inflammasome and beyond,” *Innate immunity*, vol. 20, no. 2, pp. 115–125, 2014.
- [116] N. Kang, J. H. Zhang, F. Qiu, S. i. Tashiro, S. Onodera, and T. Ikejima, “Inhibition of egfr signaling augments oridonin-induced apoptosis in human laryngeal cancer cells via enhancing oxidative stress coincident with activation of both the intrinsic and extrinsic apoptotic pathways,” *Cancer letters*, vol. 294, no. 2, pp. 147–158, 2010.
- [117] F. Morgillo, E. D’Aiuto, T. Troiani, E. Martinelli, T. Cascone, R. De Palma, M. Orditura, F. De Vita, and F. Ciardiello, “Antitumor activity of bortezomib in human cancer cells with acquired resistance to anti-epidermal growth factor receptor tyrosine kinase inhibitors,” *Lung Cancer*, vol. 71, no. 3, pp. 283–290, 2011.
- [118] Y. Wu, M. Alvarez, D. J. Slamon, P. Koeffler, and J. V. Vadgama, “Caspase 8 and maspin are downregulated in breast cancer cells due to cpg site promoter methylation,” *BMC cancer*, vol. 10, no. 1, p. 32, 2010.

REFERENCES

- [119] Y. Liang, C. Yan, N. F. Schor, *et al.*, “Apoptosis in the absence of caspase 3,” *Oncogene*, vol. 20, no. 45, pp. 6570–6578, 2001.
- [120] T. Iwai, Y. Moriya, M. Shirane, K. Fujimoto-Ouchi, and K. Mori, “Continuous inhibition of epidermal growth factor receptor phosphorylation by erlotinib enhances antitumor activity of chemotherapy in erlotinib-resistant tumor xenografts,” *Oncology reports*, vol. 27, no. 4, pp. 923–928, 2012.
- [121] R. Kim, “Recent advances in understanding the cell death pathways activated by anticancer therapy,” *Cancer*, vol. 103, no. 8, pp. 1551–1560, 2005.
- [122] M. Schuler, E. Bossy-Wetzel, J. C. Goldstein, P. Fitzgerald, and D. R. Green, “p53 induces apoptosis by caspase activation through mitochondrial cytochrome c release,” *Journal of Biological Chemistry*, vol. 275, no. 10, pp. 7337–7342, 2000.
- [123] J. Liu, H. Uematsu, N. Tsuchida, and M.-A. Ikeda, “Essential role of caspase-8 in p53/p73-dependent apoptosis induced by etoposide in head and neck carcinoma cells,” *Molecular cancer*, vol. 10, no. 1, p. 1, 2011.
- [124] Y. T. Li, X. J. Qian, Y. Yu, Z. H. Li, R. Y. Wu, J. Ji, L. Jiao, X. Li, P. F. Kong, W. D. Chen, *et al.*, “EGFR tyrosine kinase inhibitors promote pro-caspase-8 dimerization that sensitizes cancer cells to DNA-damaging therapy,” *Oncotarget*, vol. 6, no. 19, p. 17491, 2015.
- [125] C. Du, M. Fang, Y. Li, L. Li, and X. Wang, “Smac, a mitochondrial protein that promotes cytochrome c-dependent caspase activation by eliminating iap inhibition,” *Cell*, vol. 102, no. 1, pp. 33–42, 2000.
- [126] L. Jia, Y. Patwari, S. M. Kelsey, S. M. Srinivasula, S. G. Agrawal, E. S. Alnemri, and A. C. Newland, “Role of smac in human leukaemic cell apoptosis and proliferation,” *Oncogene*, vol. 22, no. 11, pp. 1589–1599, 2003.
- [127] S. Kauffman, “Differentiation of malignant to benign cells,” *Journal of theoretical biology*, vol. 31, no. 3, pp. 429–451, 1971.

REFERENCES

- [128] S. H. Strogatz, *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. Westview press, 2014.
- [129] T. Dobzhansky, “Genetics of natural populations. XIII. Recombination and variability in populations of *Drosophila pseudoobscura*,” *Genetics*, vol. 31, no. 3, p. 269, 1946.
- [130] C. Boone, H. Bussey, and B. J. Andrews, “Exploring genetic interactions and networks with yeast,” *Nature Reviews Genetics*, vol. 8, no. 6, pp. 437–449, 2007.
- [131] E. S. Kroll, K. M. Hyland, P. Hieter, and J. J. Li, “Establishing genetic interactions by a synthetic dosage lethality phenotype,” *Genetics*, vol. 143, no. 1, pp. 95–102, 1996.
- [132] A. Zinovyev, I. Kuperstein, E. Barillot, and W.-D. Heyer, “Synthetic lethality between gene defects affecting a single non-essential molecular pathway with reversible steps,” *PLoS Comput Biol*, vol. 9, no. 4, p. e1003016, 2013.
- [133] L. Jerby-Arnon, N. Pfetzer, Y. Y. Waldman, L. McGarry, D. James, E. Shanks, B. Seashore-Ludlow, A. Weinstock, T. Geiger, P. A. Clemons, *et al.*, “Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality,” *Cell*, vol. 158, no. 5, pp. 1199–1209, 2014.
- [134] 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*, vol. 2014, 2014.
- [135] I. Lobo, “Same genetic mutation, different genetic disease phenotype,” *Nature Education*, vol. 1, no. 1, p. 64, 2008.
- [136] M. AlQuraishi, G. Koytiger, A. Jenney, G. MacBeath, and P. K. Sorger, “A multi-scale statistical mechanical framework integrates biophysical and genomic data to assemble cancer networks,” *Nature genetics*, vol. 46, no. 12, pp. 1363–1371, 2014.