Investigation of the Regulatory Roles of MicroRNAs by Systems Biology Approaches

YANG YANG

(B.Eng, USTC, China)

A THESIS SUBMITTED

FOR THE DOCTOR OF PHILOSOPHY OF

DEPARTMENT OF ELECTRICAL AND COMPUTER

ENGINEERING

National University of Singapore

2011

To My Beloved Parents

&

My Dear Wife and Son

Abstract

Systems biology is a field of increasing importance in biology research. It aims to study the functioning of inter- and intra-cellular dynamic networks, using signal- and system-oriented approaches. In this thesis, we apply this idea to investigate the regulatory roles of microRNAs.

MicroRNAs are small non-coding RNAs, which inhibit the gene expression by binding to the target genes. Mounting evidence shows that microRNAs are involved in many crucial biological processes, including cancer. Among them, one critical process—p53-dependent apoptosis pathway—is selected to accommodate microRNA to conduct the study. During the investigation, we solve the core problem step by step.

First of all, the surrounding network about the well-known protein p53 is investigated. Ordinary differential equations are built to describe the underlying mechanisms. Based on the mathematical model, two novel phenomena are predicted to describe the stability change and frequency shift due to the varying levels of external stimulus. Experiment guidelines to validate these predictions are also provided accordingly.

Secondly, we employ a discrete formalism—Petri net—to model a large-scale network, p53-dependent apoptosis pathway. One challenge in systems biology is how to obtain an accurate and predictable computational model for the biomolecular networks under study. Therefore, to enhance the reliability, we propose two approaches to check the model's correctness, which are based on invariant analysis and reachability analysis, respectively. The case studies show good competency of those approaches.

Thirdly, we tackle the core problem about microRNA. The prediction of microRNAs'

targets presents a big obstacle in microRNA studies. Because bioinformatics tools offer enormous targets, most of which are believed to be false positive. Model checking based method is developed to address this issue. MicroRNA and its targets are put into p53-dependent apoptosis pathways. Then, the validity of the predicted targets is determined by the comparisons between models with and without considering microRNA's inhibition on respective targets. The experimental evidence provides the evaluation criteria. In case of lacking evidence, experimental design schemes are provided based on the desired specifications as well.

In summary, in this thesis, we illustrate the whole procedure to investigate the regulatory role of microRNAs by addressing the problem of microRNA target validation. In addition, the approach developed here may finally evolve into a formal method to comprehensively and rapidly validate target mRNAs for the microRNA, which may help us to understand cancer better and design new therapeutic strategies for cancer.

Acknowledgements

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Since DNA was deciphered, molecular biology has been experiencing a fast pace of evolvement. The biologists were constantly trying to make it clear that life is made of chemistry and physics. They believed that once we had found the smallest component of life, we would be able to have a thorough understanding of life. Accordingly, molecular biology had been developed to identify and characterize the individual gene or protein [2]. Unfortunately, the interactions among components are always neglected, which produces an incomplete picture and thus hinder the understanding of the organism as a whole.

The shortcomings of the aforementioned component-based research have led to a revival of holistic approaches. Moreover, the conventional experimentation is strongly dependent on the experience of the biologists and is usually performed in a trial-and-error manner, which is time-consuming, labour-intensive and inefficient. The biology society urges the revolution of methodologies and in recent years have seen more and more research activities in the field of systems biology.

1.1   Systems Biology

Systems biology is concerned with the dynamics of biochemical reaction networks within cells and in cell population, using signal- and system-oriented approaches, thereby

observing the behaviours at the system level [48]. Systems biology is an emerging area of research, which is truly inter-disciplinary, as it combines various disciplines and areas of research, such as life science, systems engineering, mathematical modelling and simulation, computer science, statistics, etc.

Systems biology in the biological revolution is closely associated with the fields of "genomics, transcriptomics, proteomics and metabolomics" [109]. With the emergence of these fields, molecular biology shifts its focus from molecular characterization to the understanding of functional activities. For example, in the past, single gene was studied, whereas with DNA microarray technology we can now measure the activity levels of thousands of genes simultaneously. Thus, it is possible to identify inter-relationships between groups of genes and analyse dynamic interactions among these genes.

As we can see, all the above interactions are the consequences of dynamics and controlled processes. Therefore, it is not surprising to apply systems theory to biological systems. However, this work is not a routine application of control methods and systems theory to an unconventional plant. The most appreciated work of systems biology is to be conducted by closely cooperating with the biologists. It is necessary to learn their demands and requirements before collecting the data from the collaborators and the biological literatures. Moreover, the engineering solutions must be meaningful to their implementations and understandable to the biologists, as opposed to some impractical ideas.

The development of experimental and measurement techniques makes systems biology urgent. New technologies, such as modern microscopy, laser tweezers, nanotechnology, as well as DNA microarrays, and mass spectrometry accelerate the generation of data. It becomes apparent that the methodological advances in data analysis are urgently required [47]. We must convert the newly available data into information and knowledge. Therefore, how to manage these data is as important as the technological development. Currently, we may apply a variety of systems biology methods to handle the complexity of biological systems. As the discoveries of molecular biology pave the

way for system structure, one motivation for systems biology is to bring these static diagrams to life by modelling and simulating the biochemical reactions that underlie cell functions, development and diseases.

Figure 1.1 shows an interactive workflow that we follow to apply systems biology methods. From the discoveries of science and initial experimentation, we are able to learn the basic mechanisms of the biological system. Together with the measured data, we can build some preliminary computational models to generate simulation results. To increase the reliability of the model, we ought to perform iterative model verifications. When the model is robust enough, we may use this model to propose hypotheses. This will help the biologists to judiciously design the experiments and further discover more insightful facts.
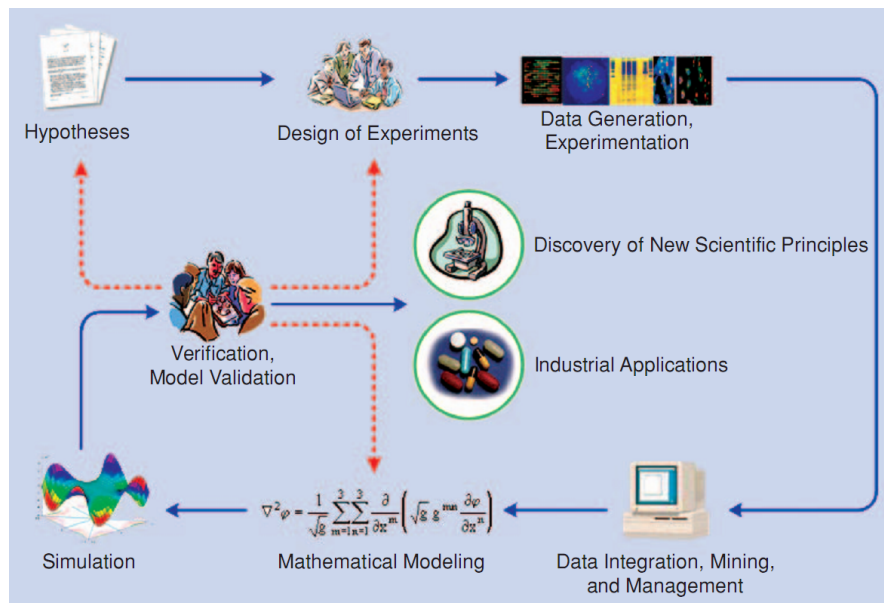
Figure 1.1: Systems Biology Workflow Diagram [108]

1.2 Motivation and Purpose

Bearing this philosophy in mind, we conduct the systems biology research in a problem-driven manner. The initiatives start from the potential facilities and solutions to the

problems in biology society. In this part, I will introduce our motivation and purpose of the research topics in this thesis.

MicroRNA, usually of length of 20 nt, can repress the genes' expression and accelerate or decelerate the formation of cancer cells. Therefore, the discovery of this tiny RNA is expected to bring revolutionary means to cancer treatment [9]. However, the regulatory roles of microRNAs remain to be elucidated. One main problem for microRNA is to identify the repressed targets [56]. However, one microRNA can regulate a great number of target genes, usually in the scale of hundreds. To make things more intricate, one gene could be regulated by several microRNAs. The bioinformatic tools could predict the targets based on base-pairing principles [87]. However, the screening of whole genome gives too many targets to be considered as true [56]. The only thing that biologists can do is to validate these targets in the way of trial-and-error one by one. With the validated targets, some well-known pathways where the roles of microRNAs are included could be re-evaluated accordingly. For example, in [1], the roles of E2F and Myc in cancer pathways are studied by including the regulation by mir-17-92.

Our motivation in this topic is to develop a formal method to predict the highly possible targets out of the results from bioinformatics tools. Then, the shortlisted candidates are provided to the biologists to save the time and labour-cost. Meanwhile, mounting evidence shows that microRNAs are involved in many crucial biological processes to regulate the tumorigenesis. Among them, we select one critical process—p53-dependent apoptosis pathway where the role of microRNAs is investigated. Then, we develop our methods for the core problem step by step. To the best of our knowledge, this is the first time to validate the targets of microRNAs in the context of dynamical pathways using the systems biology approach.

First of all, the networks surrounding the well known protein p53 is investigated. The protein p53 lies at the centre of several critical pathways in our body. It correlates with cell cycle, cell death, angiogenesis, etc [99]. When the cell is stressed by oncogenic stimuli, p53 will prevent the progression of malignancy in the involved pathways. The

main role of p53 is the transcriptional activation of the target genes which then join the downstream pathways to exert the corresponding repressive functions. Among the target genes, Mdm2 influences the p53 level negatively in return. Thus, the feedback makes the p53 level oscillate [7]. This phenomenon has been observed in the wet lab and drawn much attention of the modellers. To deeply investigate the p53-Mdm2 core regulation, we build our own model to reproduce the oscillation and the model itself facilitates further analyses and predictions. Many results have not been discovered before, and could provide potential evidence to explore this core regulation. For example, the level of the stimulating agent has two thresholds which govern the occurrence of oscillation. The first threshold is reported by the literatures, whereas the second one has not been reported before. Moreover, from the simulation results, we find the frequency drift with respect to the agent level. This phenomenon is also ignored by the previous work. We predict this frequency shifting and provide the practical operation guideline to assist the validation experimentation. The verification of these two new phenomena could be fed back to the modelling work and improve the credibility of our model.

Next, we move to a large-scale network—p53-dependent apoptosis pathway. In the aforementioned modelling work, it is built by non-linear ordinary differential equations in continuous-time domain. The core regulation contains a small number of players, thereby avoiding many troubles on the parameter identifications. Manual tuning is sufficient. However, in p53-dependent apoptosis pathway, there are dozens, even hundreds of players and the related parameters which generate an insurmountable gap [40]. To address this problem, we adopt the formalisms in discrete domain to emphasize more on the structure information and qualitative properties. In this work, we choose Petri net as one ideal candidate due to its natural affinity to represent the species and reactions under its framework. Accordingly, model validation is an immediate task to guarantee the model's correctness. We analyse the mathematical description of Petri net and provide two alternative ways to tackle this. Invariant and reachability analysis are obtained from the characteristics of state equations. The solutions both efficiently

verify the structure information of the model.

Succeeding the discrete modelling of p53-dependent apoptosis pathway, we continue the investigation of our core topic—microRNA. MicroRNAs exert their functions to inhibit the gene expression by binding to the target genes [9]. To deal with the target validation problem, we still work on the discrete model and borrow the model-checking technique from computer science to explore the solutions. Since the well-established model could be interpreted in many different perspectives, in this work, we put microRNA at the locations of its targets which are involved in the p53-dependent apoptosis pathways, and investigate the influences. The proposed method compares the behaviours of the model with the real evidence. Based on the differences, we may conclude the validity of these targets.

As can be seen, we start from the biological problems and formulate into a manageable mathematics and engineering problem. We develop our approaches from multiple disciplines, such as mathematics, engineering, computer science, physics, bioinformatics, etc. We hope that the provided solutions could finally benefit the biology society and facilitate their verification and subsequent discoveries.

1.3 Organization of Thesis

In the following three chapters, we discuss our main research topics of the p53-Mdm2 core regulations, apoptosis pathway modelling and validation and microRNA targets, respectively. Each chapter will begin with the introduction of biological objectives, then followed by the technical backgrounds. The methods are illustrated by solving the concerned problems or the representative examples. Discussion and conclusion sections are presented to summarize each chapter.

Finally, Chapter 5 concludes the whole thesis and propose the future directions which extend current work.

Besides, Appendix C is employed to introduce the background knowledge about biology and experimental methods. Basic biological concepts, such as cell, nucleus,

DNA, protein, central dogma of molecular biology, gene expression, are introduced to help understanding our research objectives. Moreover, some popular experimental tools are primarily surveyed. Both introductions will facilitate the interpretation of the results in my research work.

Chapter 2

P53-Mdm2 Core Regulation

As described in Section 1.2, we initiate our research by investigating the p53 protein. P53 is a well-known tumour suppressor with many anticancer mechanisms. The protein itself has drawn intensive attention in biology research. It can activate apoptosis, the programmed cell death in which microRNAs is also believed to influence the biomolecular regulations. Therefore, it is our first step to explore the activities around this protein.

2.1 Introduction

2.1.1 P53

The p53 tumour suppressor lies at the centre of cellular pathways that sense DNA damage, cellular stress and oncogenetic stimulation [99]. P53 integrates such signals and, in response, induces growth arrest, triggers apoptosis (programmed cell death), blocks angiogenesis or mediates DNA repair, etc. [35] The critical role of p53 is experimentally evidenced by the presence of mutations found in almost 50% human tumours. Therefore, studies of p53 have attracted attentions of many researchers in life science for decades [15].

P53 serves as a transcriptional activator to promote the target genes' expressions and the downstream products will repair the double-strand breaks (DSB) and ulti-

mately mitigate the DNA damage [27]. The structure of human p53 protein is shown in Figure 2.1. DNA binding site is used for targeting the genes. MDM2 TA site is bound by Mdm2, inducing the degradation of p53. When the cell is stressed by DNA damage signal or other stimuli, some agents will change the formation of p53 through phosphorylation and acetylation on the N and C-terminal, respectively. For instance, ATM's phosphorylation will enhance the binding ability of p53 to target gene, and meanwhile, it will weaken the binding ability of Mdm2 to p53 because ATM will add a phosphotate group at ser 15, which is inside the Mdm2 site.
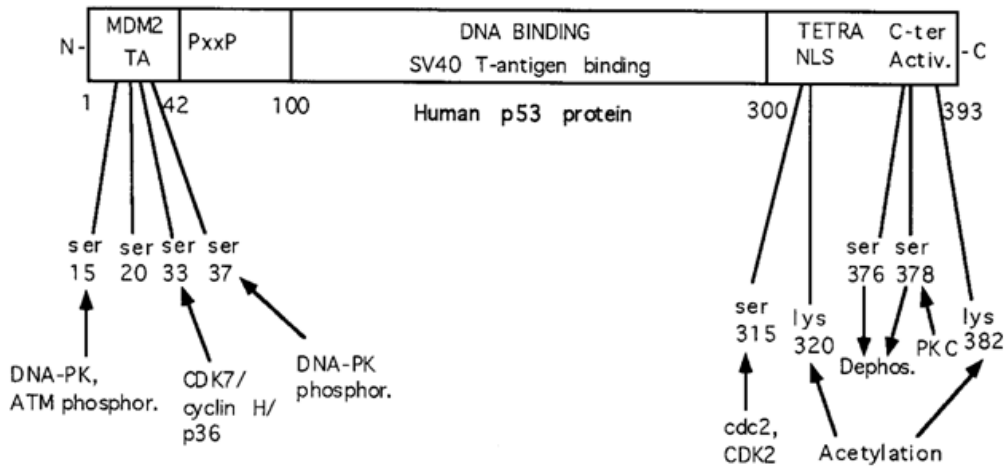
Figure 2.1: The structure of human p53 protein [27].

2.1.2 P53-Mdm2 Core Regulation

The p53 network is normally "off". In normal cells, p53 protein usually maintains at a low level and has a short half-life due to the degradation by ubiquitination and proteolysis. The inhibitor is Mdm2 protein which is a E3 ubiquitin ligase for p53 and also a target gene of p53 simultaneously. Apparently, there exists a negative feedback to maintain the low p53 level. The core regulation can be simply represented as p53 → Mdm2 ⊣ p53. Furthermore, the Mdm2-interacting region in p53 resides at the 1-42

amino acids within N-terminal region. On the other hand, when the cell is stressed by DNA damage signal, such as ultraviolet (UV), ionizing radiation (IR), ATM will add phosphate group to the serine 15 which leads to the poor binding ability of Mdm2 to p53. Thus the p53 level will be raised and activated to perform its major functions. Besides, ATM has another role to accelerate the transcription of target genes by phosphorylation of p53 [5]. All the above introductions can be summarized in Figure 2.2.
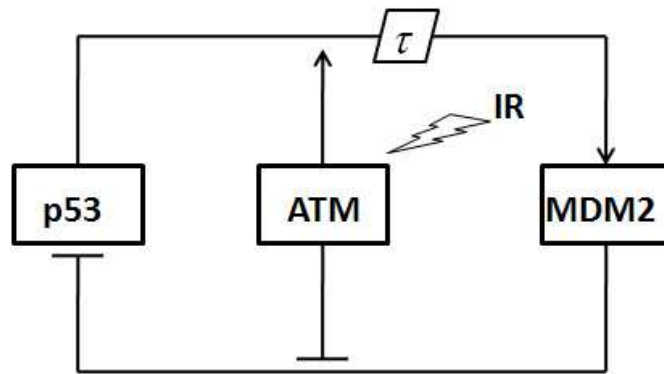
Figure 2.2: Schematic diagram to illustrate p53-Mdm2 core regulation. Arrow represents activation, while arrow-bar means inhibition. IR is short for ionizing radiation. τ is the assumed time lag from p53 to Mdm2's translation.

Recently, two research groups found the oscillation phenomena in p53-Mdm2 loop [7, 52]. The capture by western blot from [7] is shown in Figure 2.3. Damped oscillatory behaviours in population of cells and undamped oscillatory behaviours in individual cells were observed after the irradiation. Oscillatory expressions are actually observed in many other systems, such as Hes1 and NF-κB related networks [66, 70, 83]. Due to the lack of biological evidence and experimental data, the true mechanisms are not illustrated yet. Therefore, these oscillations motivate researchers' interest in the study of p53-Mdm2 core regulation; and many investigations have been devoted to build a reasonable model to qualitatively explain this oscillatory phenomenon.
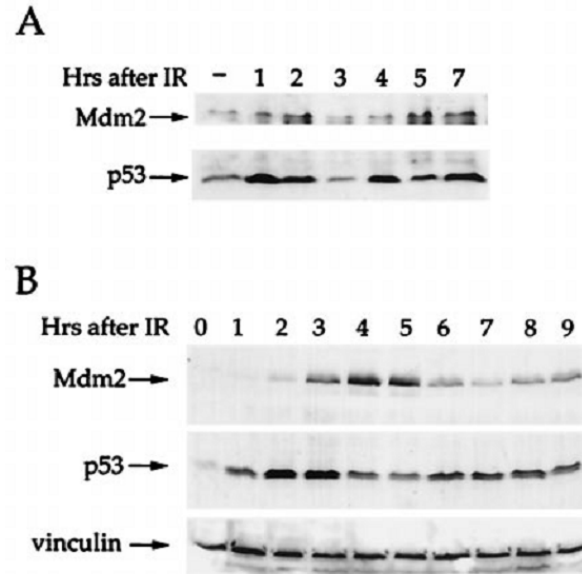
Figure 2.3: P53-Mdm2 oscillation observation after ionizing radiation from [7]. A is from Mouse fibroblasts NIH 3T3 cells and B is from Human breast cancer epithelial MCF-7 cells.

2.1.3 Objective

In this chapter, the main objective is to investigate the p53-Mdm2 regulation in both time and frequency domains so as to obtain more insights on the regulatory mechanisms and propose verifiable hypotheses. First of all, a new mathematical model, which falls into the category of delayed feedback, is proposed by taking ATM's dual role into account. ATM is involved to associate the DNA damage signal with this core regulation, which is expressed by a simple dynamics in the model. Next, using this converter, bifurcation analysis of p53 with respect to ionizing radiation is performed; consequently, a threshold mechanism of radiation dose, which has never been discussed before, is found. Moreover, variation of p53-Mdm2 oscillation frequency is usually ignored in the existing literature. Inspired by this, we investigate frequency shifting phenomenon by Fourier frequency analysis on the model. Accordingly, we facilitate the experiment design by an optimized guideline. Bifurcation and frequency analysis are both contributing to the experimental validation and design in practice.

The rest parts are organized as follows. The modelling principle by mass action law is firstly introduced in Section 2.2. In Section 2.3, mathematical expressions are derived one by one according to the biological bases and assumptions. Next, simulation results and bifurcation analyses are given to exploit the model. In Section 2.5, through Fourier frequency analysis, a design scheme is provided to help conducting the wet lab experiments. Discussion part is dedicated to advise experimental verifications for model predictions. Finally, this chapter ends with the conclusion part.

2.2 Mass Action Law Based Modelling

When we want to determine the reaction kinetics, we have to evaluate the reaction rates for a product or reactant in a particular reaction: the amount(in numbers or concentrations) per unit time that is formed or removed. It is named as "Mass Action Law" [105]. The basic assumption is collision theory, i.e. the reaction can be triggered by the collision of two reactants. And the reaction rate is proportional to the probability of collision of the reactants.

For example, there is a bimolecular reversible reaction.

$$S_1 + S_2 \underset{k2}{\overset{k1}{\rightleftharpoons}} 2P$$
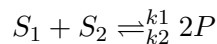
The reaction rate cab be calculated below.

$$r(t) = -\frac{d[S_1]}{dt} = -\frac{d[S_2]}{dt} = \frac{d[P]}{dt} = k_1[S_1][S_2] - k_2[P]^2$$

The bracketed variables $[S_1], [S_2] and [P]$ denote the time-varying amount of interest, concentrations or molecular numbers. k_1, k_2 are called rate constant, which depends on temperature, pH-value, pressure, etc., while time-independent. Thus, given the environment keeps invariant, the reaction rate will depend only on the concentration or numbers.

Most biochemical reactions will be catalysed by huge numbers of enzymes. So the catalytical reactions exist universally in cells. "Michaelis-Menten Kinetics" describes the rate of enzyme-mediated reactions based on mass action law.

The reaction equation is as follows:

$$E + S \underset{k_{-1}}{\overset{k1}{\rightleftharpoons}} C(ES) \xrightarrow{k_2} E + P$$
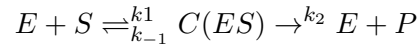
It has a reversible formation of the enzyme-substrate complex ES and an irreversible release of the product P from the complex.

Use mass action law to model each reaction.

$$
\begin{aligned}
\frac{d[S]}{dt} &= -k_1[E][S] + k_{-1}[ES] \\
\frac{d[ES]}{dt} &= k_1[E][S] - (k_{-1} + k_2)[ES] \\
\frac{d[E]}{dt} &= -k_1[E][S] + (k_{-1} + k_2)[ES] \\
\frac{d[P]}{dt} &= k_2[ES]
\end{aligned}
$$

A key assumption here is the quasi steady-state approximation, namely that the concentration of the complex change much more slowly than the substrate and product.

$$\frac{d[ES]}{dt} = 0 = k_1[E][S] - (k_{-1} + k_2)[ES]$$

,which can be rearranged to $[ES] = \frac{k_1[E][S]}{k_{-1}+k_2}$.

Define Michaelis constant as $K_m = \frac{k_{-1}+k_2}{k_1}$ to yield

$$[ES] = \frac{[E][S]}{K_m} \tag{2.1}$$

Based on conservation law, we will have

$$[E_0] = [E] + [ES] \tag{2.2}$$

Putting Eq. 2.1 and 2.2 together, we may obtain

$$[ES] = [E_0]\frac{1}{1 + \frac{K_m}{[S]}} \tag{2.3}$$

The reaction rate is:

$$V = \frac{d[P]}{dt} = k_2[ES] \tag{2.4}$$

From Eq. 2.3 and 2.4, it is easily to derived that

$$\frac{d[P]}{dt} = k_2[E_0]\frac{[S]}{K_m + [S]} = V_{max}\frac{[S]}{K_m + [S]}$$

As can be found, K_m is an indicator of the affinity that an enzyme has for a given substrate and hence the stability of the enzyme-substrate complex. Furthermore, when substrate concentration [S] is equal to K_m, the reaction rate is half of the maximal rate V_{max}, which is approximately got from the large enough [S]. At low [S], it is the availability of substrate that is the limiting factor. As more substrate is added, there will be a rapid increase in the initial rate of the reaction. The relationship between the reaction rate against substrate concentration [S] is shown in Figure 2.4.

2.3 Modelling and Simulation Results

The existing modelling works about p53-Mdm2 core regulation are all in continuous domain. It has been learned that oscillations can arise from negative feedback alone, which is composed of at least three components [97]. Hence, in Lev Bar-Or's work [7], they resorted to a putative intermediary in the negative feedback loop. They explored the dependence of oscillations on different parameters, such as k_{delay}, which represents the time lag from intermediary to Mdm2. This inspired other research efforts which considered this time lag as an explicit parameter in the transcriptional and translational process of Mdm2. One of the representative studies was done by Monk in [66], where he proposed a delayed feedback model and integrated all the time lags as one explicit term

Figure 2.4: Michaelis-Menten Kinetics. [S] axis is drawn in logarithmic-scale.

in the formation process of Mdm2. From then on, most researchers have adopted this idea for modelling the p53-Mdm2 regulation, such as [100, 14]. In particular, Wagner and his co-workers took a significant step in investigating the global dynamics under different parameter bifurcations in [100]. An alternative approach was suggested in [23] by Tyson and his colleagues via introducing a positive feedback mechanism besides the common negative feedback loop, without relying on the explicit time delay.

Another remarkable work from Alon's research group gave a long-term (up to 3 days) experimental data set in [33]. Moreover, they summarized six different model types for the dynamics of p53-Mdm2 network. They built a stochastic model concerning about the variability between cells as well. Other studies from a stochastic point of view were done in [62, 18]. Most recently, Ramalingam and his colleagues collected the experimental data using protein lysate microarrays [77]. Then based on the observations, they identified the parameters of the mathematical model adopted from [7, 62]. Subsequently, they knocked out p53 gene *in silico* by setting the production rate as zero. Finally, they made a good verification by the real experiment *in vivo*.

15

Our model relies on prevailing evidence and widely accepted assumptions. For the sake of simplicity, only the p53 and Mdm2 proteins are considered, rather than the messenger RNAs of them. The reliability of this simplification will be verified by the later simulation results. The delays happened in the transcription, translation and translocation processes are all merged as one delay term appearing explicitly in the Mdm2 dynamics. The selections of parameters are performed after scaling the original equations.

2.3.1   Model

First of all, p53 dynamics is evaluated as

$$\frac{\mathrm{d}p53}{\mathrm{d}t} = a_p - d_p \times p53 - deg(S(t)) \times \frac{p53}{p53 + K_p} \times Mdm2. \tag{2.5}$$

Here the first term a_p specifies the synthesis rate of the p53 protein; the second term reflects the Mdm2-independent p53 degradation, while d_p is the basal degradation rate; the last term describes the Mdm2-induced p53 degradation. Michaelis-Menten kinetics is applied to this process, consistent with an enzyme (Mdm2)-catalyzed degradation from a substrate (p53 protein). As for $deg(S(t))$, it is the degradation rate which is a function of ATM, denoted by $S(t)$.

The expression for $deg(S(t))$ is

$$deg(S(t)) = d_0 \times (1 - \frac{S^n}{S^n + K_1^n}), \tag{2.6}$$

where d_0 is the basal rate for Mdm2-dependent p53 degradation. As shown in [5] , when the cell is exposed to the ionizing radiation, ATM can weaken the binding ability of Mdm2 to p53. So this basal degradation rate will be reduced by the existence of ATM. It is assumed that the reduction follows a Hill function with order, also called cooperativity, n.

Secondly, the dynamics of Mdm2 is described in the following equation,

$$\frac{\mathrm{d}Mdm2}{\mathrm{d}t} = a_m - d_m \times Mdm2 + agg(S(t)) \times \frac{p53^4(t-\tau)}{p53^4(t-\tau) + K_m^4}, \tag{2.7}$$

where the coefficients a_m and d_m give the basal rate of synthesis and degradation for Mdm2, respectively. The last term represents the transcription activation of Mdm2 by p53. Here transcription product—Mdm2 messenger RNA—is replaced by Mdm2 protein and phosphorylated p53 is replaced by p53 protein. The two forms of p53 will not be discriminated in this model. The phosphorylation by ATM kinase is expressed in the coefficient function $agg(S(t))$. To account for p53's preference for tetramerisation [102], P2 promoter's dependence on p53 is modeled as a Hill function with cooperativity 4. Time lag τ is utilized to represent all the duration cost in this process.

The function $agg(S(t))$ is formulated as

$$agg(S(t)) = a_0 \times \frac{S^m}{S^m + K_2^m}, \tag{2.8}$$

Finally, the connection from stress signal to the core regulation via ATM's kinetics comprises the following two first-order dynamics.

$$\frac{\mathrm{d}S}{\mathrm{d}t} = k \times dam - d_s \times S. \tag{2.9}$$

$$\frac{\mathrm{d}dam}{\mathrm{d}t} = \frac{1}{T_1} \times (IR - dam). \tag{2.10}$$

Eq.(2.9) shows the ATM's dependence on the DNA damage denoted by dam, in which the second term describes the degradation of ATM. Eq.(2.10) describes damage generated due to ionizing radiation IR, i.e. the input to the whole system. On the other hand, when the stress signal is withdrawn, it is assumed that the repair of DNA damage will follow the process below.

$$\frac{\mathrm{d}dam}{\mathrm{d}t} = -\frac{1}{T_2} \times dam. \tag{2.11}$$

2.3.2 Selection of Parameters

In this subsection, we introduce the following new variables and scaling relationships.

$$\hat{p53} = \frac{d_m}{a_p}p53, \, \hat{Mdm2} = \frac{d_m}{a_m}Mdm2, \, \hat{S} = \frac{d_m}{k}S$$

$$\hat{t} = d_m t, \, \hat{\tau} = d_m \tau$$

$$\hat{d_p} = \frac{d_p}{d_m}, \, \hat{K_p} = \frac{d_m}{a_p}K_p, \, \hat{K_m} = \frac{d_m}{a_p}K_m$$

$$\hat{d_0} = \frac{a_m d_0}{d_m a_p}, \, \hat{a_0} = \frac{a_0}{a_m}, \, \hat{T_1} = \frac{1}{T_1 d_m}, \, \hat{T_2} = \frac{1}{T_2 d_m}$$

$$\hat{d_s} = \frac{d_s}{d_m}, \, \hat{K_1} = \frac{d_m}{k}K_1, \, \hat{K_2} = \frac{d_m}{k}K_2$$

Here, we use dimensionless scaling to help reducing the burden for selection of parameters, which is a common method in systems modelling [100, 1]. Thus, the rescaled dynamics is expressed by the new variables in the following form.

$$
\begin{aligned}
\frac{\mathrm{d}\hat{p53}}{\mathrm{d}\hat{t}} &= 1 - \hat{d_p} \times \hat{p53} - deg(\hat{S}(\hat{t})) \times \frac{\hat{p53}}{\hat{p53} + \hat{K_p}} \times \hat{Mdm2} \\
\frac{\mathrm{d}\hat{Mdm2}}{\mathrm{d}\hat{t}} &= 1 - \hat{Mdm2} + agg(\hat{S}(\hat{t})) \times \frac{\hat{p53}^4(\hat{t}-\hat{\tau})}{\hat{p53}^4(\hat{t}-\hat{\tau}) + \hat{K_m}^4} \\
\frac{\mathrm{d}\hat{S}}{\mathrm{d}\hat{t}} &= dam - \hat{d_s} \times \hat{S} \\
\frac{\mathrm{d}dam}{\mathrm{d}\hat{t}} &= \frac{1}{\hat{T_1}} \times (IR - dam) \\
\frac{\mathrm{d}dam}{\mathrm{d}\hat{t}} &= -\frac{1}{\hat{T_2}} \times dam, \text{ when stress signal is withdrawn} \\
deg(\hat{S}(\hat{t})) &= \hat{d_0} \times (1 - \frac{\hat{S}^n}{\hat{S}^n + \hat{K_1}^n}) \\
agg(\hat{S}(\hat{t})) &= \hat{a_0} \times \frac{\hat{S}^m}{\hat{S}^m + \hat{K_2}^m}
\end{aligned}
$$

To highlight the role of p53 in transcriptional activation, $\hat{a}_0$ should be selected much greater than 1, which is the unitized basal synthesis rate of Mdm2. The same selection criterion is applicable to the p53's degradation rates. Mdm2 makes the p53's proteolysis much faster compared with the basal degradation. Hence $\hat{d}_0$ is reasonably considered much greater than $\hat{d}_p$. In most existing literatures, the basal synthesis rate of Mdm2 and degradation rate of p53 are neglected. As for the Hill function's cooperativity, orders of 1 and 4 in Eq.(2.5) and Eq.(2.7) are selected according to the Michaelis-Menten kinetics and p53's tetramerisation. The orders of n and m used in Eq.(2.6) and Eq.(2.8) are determined by the sensitivity of the components. Moreover, the time delay τ is a key factor for the existence of oscillation [73]. For example, the values below a critical point $\tau_0 = 0.875$ will eliminate the oscillation when IR is set as 0.5 in this model. The rest selection of parameters are based on literatures [7, 66, 100] and manual tuning.

Summarizing above, all the parameters are listed in Table 2.1. In the following, we omit the hat accent from the variables and parameters in the scaled equations and use P and M as abbreviations of p53 and Mdm2 respectively, as these changes do not cause misunderstandings.

Table 2.1: Parameter list of dimensionless kinetics equations

Parameter	Value	Parameter	Value
$\hat{d}_p$	0.2	$\hat{d}_0$	2
$\hat{K}_p$	0.2	$\hat{K}_1$	0.3
		$\hat{a}_0$	4
$\hat{K}_m$	0.5	$\hat{K}_2$	0.2
m	2	n	2
$\hat{\tau}$	1	$\hat{d}_s$	1
$\hat{T}_1$	2	$\hat{T}_2$	100

2.3.3 Simulation Result

Our model exhibits sustained oscillation in response to increased radiation dose. As can be seen in Figure 2.5, during the interval $0 \leq t \leq 15$, the cell stays under normal condition without exposure to ionizing radiation ($IR = 0$). p53 is maintained at low level due to the spontaneous inhibition by Mdm2. After $t > 15$, the cell is exposed to ionizing radiation ($IR = 0.5$). The oscillation persists until ionizing radiation is withdrawn at $t = 100$. Then the p53 and Mdm2 both return to the original states through a transient process, which consists of damped oscillations. It will be seen that the levels of p53 and Mdm2 differ much, which is due to the scaling operation. However, we will focus on the qualitative behavior rather than the quantitatively accurate time and concentration information in this work.

Figure 2.5: Temporal performance of p53 and Mdm2. During $15 \leq t \leq 100$, $IR = 0.5$. In other durations, $IR = 0$. Other parameters are listed in Table 2.1.

The first peak of p53 is earlier than Mdm2 after onset of IR, and the lag is about 1.8. The periods for both variables are the same. These performances fit to the experimental

data in [7, 52] and previous simulation results. The difference resides on the scale, which is due to the parameters' selections. The evolvement also agrees with the observed experimental phenomenon.

After performing simulations under different dose levels, it is observed that the oscillation period is changing, although it is not very obvious in the time domain of simulation results using current parameter set. This interesting variation inspires the detailed frequency analysis discussed in Section 2.5.

As evidenced by the experimental data shown in Figure 6b of [7], weak damage signal will slow the rise of steady state and no observable oscillations exist within the time frame of the experiment. To verify this point, IR is reduced to 0.2, and the result depicted in Figure 2.6 shows that the oscillation disappears and settling time is elongated compared to the previous case.

Figure 2.6: Temporal performance of p53 and Mdm2. During $15 \leq t \leq 100$, $IR = 0.2$. In other durations, $IR = 0$. Other parameters are listed in Table 2.1.

2.4 Bifurcation Analysis

According to the simulation results above, during the sustained oscillation interval, when $IR = 0.2$, p53 stays at a stable steady state. When IR is raised to 0.5, p53 will oscillate, meaning that the original fixed point changes its stability. Moreover, if IR is considered as a parameter, it will induce the bifurcation of the nonlinear systems expressed by Eq.(2.5)–(2.10). Specifically, this is Hopf Bifurcation, i.e. the stable equilibrium point becomes unstable by a parameter change, and a limit cycle appears in the neighbourhood [94].

Eq.(2.9) and (2.10) show the independence of S and dam on the p53 and Mdm2 dynamics. The solutions of $S(t)$ and $dam(t)$ are shown as

$$
\begin{aligned}
S(t) &= e^{-d_s t} S(0) + k \int_0^t dam(\tau) e^{-d_s(t-\tau)} \mathrm{d}\tau \\
dam(t) &= e^{-t/T_1} dam(0) + IR \times T_1 \int_0^t e^{-T_1(t-\tau)} \mathrm{d}\tau
\end{aligned}
$$

The settling times depend on the parameter d_s and T_1.

Meanwhile, given $d_s = 1$ and $T_1 = 2$, let

$$
\begin{aligned}
\frac{\mathrm{d}S}{\mathrm{d}t} &= dam - d_s \times S^* = 0 \\
\frac{\mathrm{d}dam}{\mathrm{d}t} &= \frac{1}{T_1} \times (IR - dam^*) = 0
\end{aligned}
$$

then get the steady-state equality $S^* = dam^* = IR$, where asterisk denotes the steady state. Therefore, in the given parameter set, the variables S and dam get equal to input IR fast.

Thus, it is convenient to study the bifurcation of reduced systems, which is comprised of only p53 and Mdm2 kinetics. Let the right hand sides of scaled p53 and Mdm2 equations equal zero and replace S with $S^* = IR$.

$$
1 - d_p \times P^* - deg(IR) \times \frac{P^*}{P^* + K_p} \times M^* = 0 \tag{2.12}
$$

$$1 - d_m \times M^* + agg(IR) \times \frac{P^{*4}}{P^{*4} + K_m^4} = 0 \qquad (2.13)$$

After arrangement, by using the same parameter set above, an implicit function of P^* with the parameter IR is derived below.

$$1 - 0.2P^* - \frac{0.18}{IR^2 + 0.09} \times \frac{P^*}{P^* + 0.2} \left[1 + \frac{4IR^2}{IR^2 + 0.04} \times \frac{P^{*4}}{P^{*4} + 0.5^4}\right] = 0 \qquad (2.14)$$

According to Descarte's rule of sign [49], it is assured to have real positive solution. Because of the high order existing in the implicit function, it is hard to get a close-form solution of P^*. By sampling IR's range $[0.1, 0.3]$ by interval of 0.02, we perform symbolic solver in Matlab iteratively. Several pairs of coordinates can be obtained. The negative real roots and complex roots are filtered because of the reality consideration. When $IR = 0.2$, the steady state of p53 is approximately 0.277, the same as shown in Figure 2.6, which can also be seen in bifurcation diagram shown in Figure 2.7.
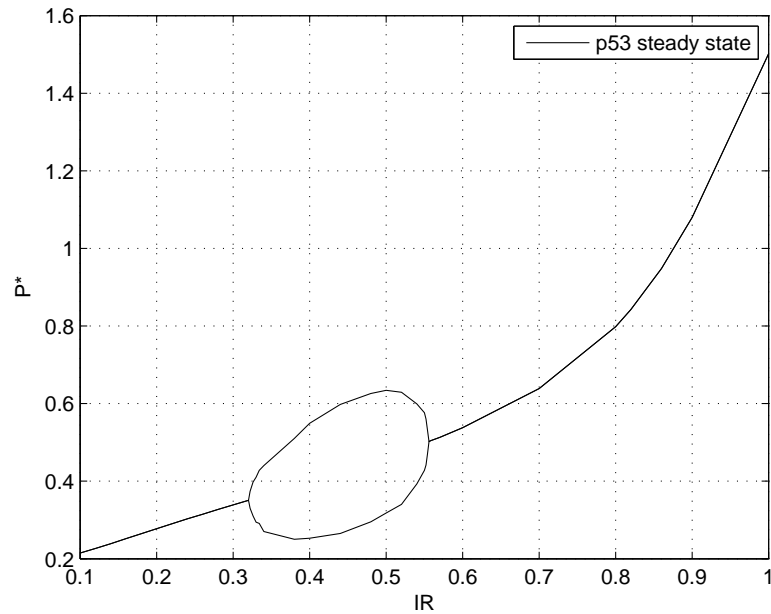
Figure 2.7: Bifurcation diagram of p53's steady state with respect to IR. For the stable limit cycle, the maxima and minima are drawn. When IR is greater than 20, the steady state converges to 5. Data are not shown here.

According to the bifurcation diagram, when $IR > 0.32$, the oscillation happens. When $IR > 0.56$, the oscillation disappears and returns to the unique steady state again. Theoretically, it is because ATM's level also becomes bigger when IR is sufficiently large. p53's degradation by Mdm2's ubiquitination is largely inhibited by ATM. That's to say, the third term of Eq.(2.5) can be neglected. Consequently, p53 level will be definitely raised, and Mdm2 is also aggregating due to the transcriptional activation by p53, leading the level higher than the basal level. Thus, p53 and Mdm2 will not be influenced by the ATM as much as in the oscillation region, and the core regulation is modified by the elimination of Mdm2's inhibition on p53. An example can be seen in Figure 2.8. So far, there are no experimental data showing the response of big dose of ionizing radiation. The analysis based on this model predicts the retrieval of stable steady states at higher level.
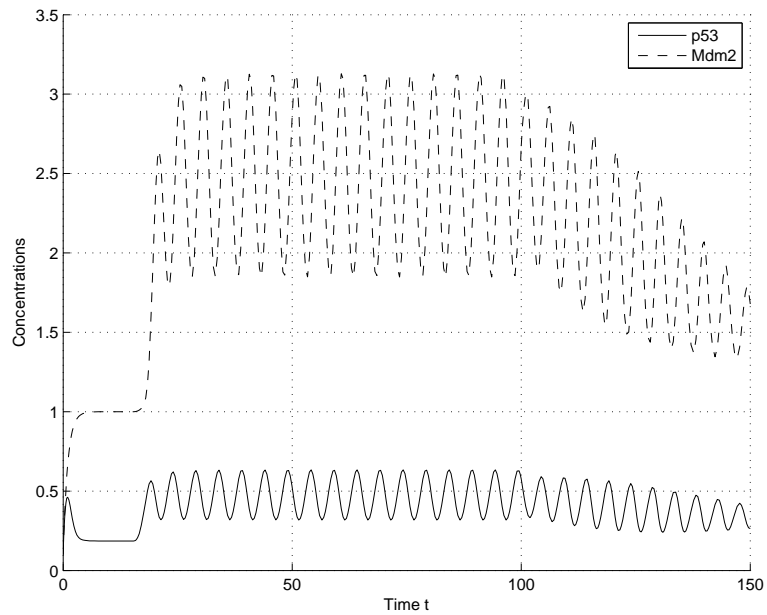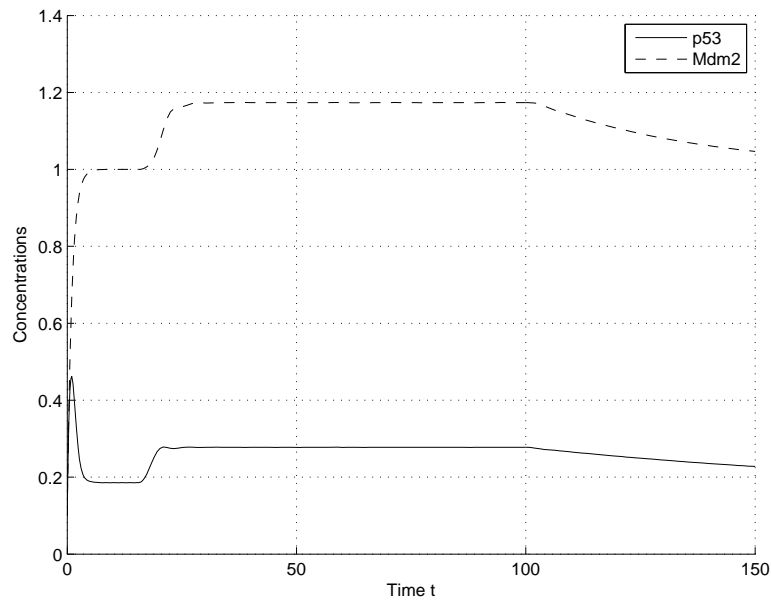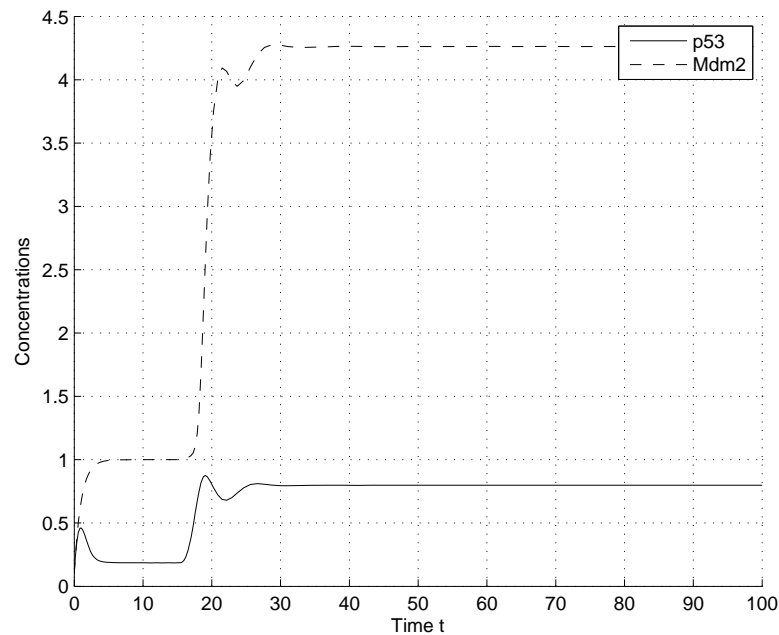

Figure 2.8: Temporal performance of p53 and Mdm2. During $15 \leq t \leq 80$, $IR = 0.8$. In other durations, $IR = 0$. Other parameters are listed in Table 2.1.

2.5 Frequency Analysis and Experiment Design

The numerical simulations suggest the changes of p53-Mdm2 oscillation periods with respect to the level of IR. This is a very interesting phenomenon and has never been discussed from simulations of the core regulation to the best of the authors' knowledge. In the time domain, the changes of periods are hard to be detected given the existence of noises, which motivate us to consider this issue in the frequency domain. In the frequency spectrum, the dominant frequencies due to oscillations will appear as pulses, which can be distinguished from noise. However, the direct validation of this predicted frequency shifting phenomenon requires accurate measurements of the p53 and Mdm2 concentrations at a very high sampling rate, say 10 times measurements per hour. This seems to be an unreasonable expectation for the current wet lab experiment techniques. To address this issue, we turn to frequency domain analysis, in particular Discrete Fourier Transform(DFT) and Fast Fourier Transform(FFT) [74]. The main purpose of frequency domain analysis is to determine the frequency at which p53 oscillates when the value of IR changes at a relatively lower requirement for the data measurements. Another advantage of dealing with the experiment design in the frequency domain is that the original frequency of the oscillation can be perfectly reconstructed in theory with limited sampled data. Based on Fourier analysis, the sampling frequency and total sample points required for proper design of experiment can be selected such that in practice, the original time series on p53 concentration can be reconstructed perfectly from the sampled data points.

2.5.1 Frequency Domain Analysis

Our first task here is to determine the frequency of the p53-Mdm2 oscillations under a specified IR level from our numerical simulation, which is called a predicted frequency. This is achieved through doing DFT on the simulated time series data and analysing it in the frequency domain.

To obtain the DFT of the time series of p53 concentration, Fast Fourier Transform

(FFT) is performed on the simulation result in Matlab. The IR value is set to be non-zero for the whole simulation interval as only the region where sustained oscillations occur is of interest. The time domain simulation result under this setting is given in Figure 2.9. Then, the first one-third of the time series obtained is truncated before the FFT so that only the oscillations with constant amplitude are considered. Besides, the solution given by numeric solver is not equally spaced in time, so it is needed to interpolate the solution and re-sample it at a regular interval before performing FFT. This will make sure that the signal is a valid input for the FFT routine [81].

Based on Nyquist sampling theorem, the sampling frequency is chosen as 1, which is much more than twice the different dominant frequencies (results shown later). Another parameter to be considered here is the number of sample points N. N is chosen to be as large as possible so that the frequency determined from DFT is more accurate.
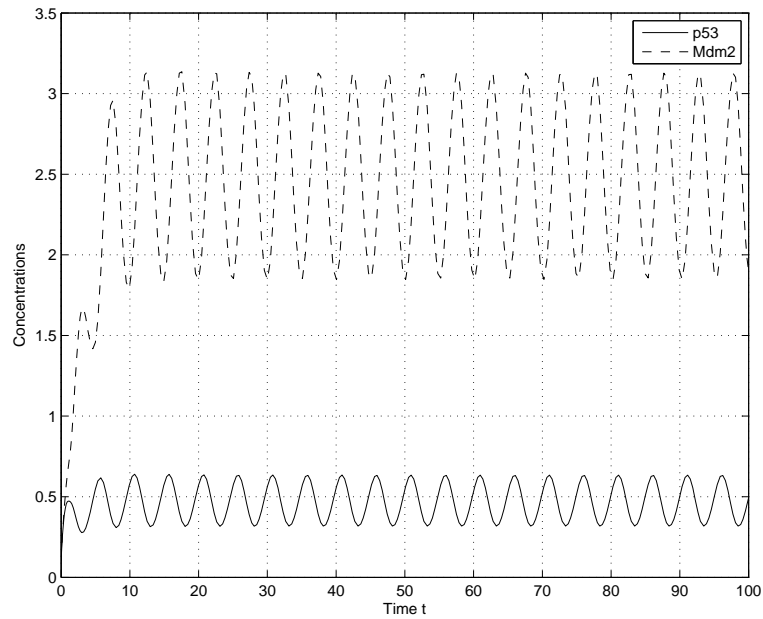
Figure 2.9: Time domain simulation result with oscillation for whole time interval, $IR = 0.5$

The amplitude spectrum of the DFT of the time series of p53 concentration when

IR is set to 0.5 is shown in Figure 2.10. There are two prominent peaks observed at $F = 0$ and $F = 0.1996$. The peak at $F = 0$ arises because there is a DC offset in the waveform. The peak at $F = 0.1996$ gives the dominant frequency of oscillation. The process was repeated for IR in oscillatory range [0.32, 0.56]. The frequencies of oscillation corresponding to different radiation doses are shown in Table 2.2.
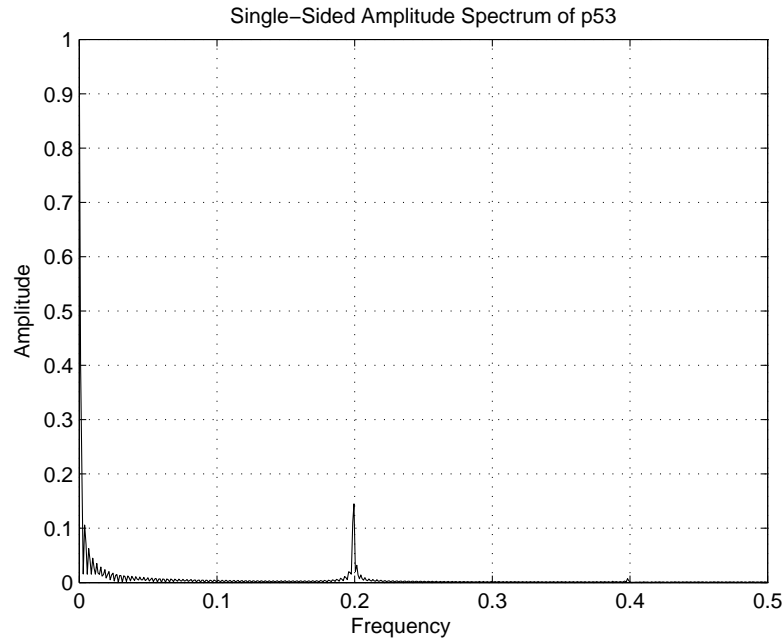
Figure 2.10: Amplitude spectrum of DFT of p53 concentration when $IR = 0.5$

Table 2.2: Normalized frequencies of oscillations for different IR values

IR	Frequency	IR	Frequency
0.33	0.2133	0.45	0.2045
0.36	0.2114	0.48	0.2016
0.39	0.2094	0.51	0.1986
0.42	0.2065	0.55	0.1937

From the spectral analysis, simulated results show that the maximum frequency of oscillation is 0.2133, which occurs at $IR = 0.33$. The minimum frequency of oscillation

is 0.1937, which occurs at $IR = 0.55$, around the upper bound. Beyond these boundary IR values, no dominant peak can be observed in the amplitude spectrum, reflecting the fact that there is no sustained oscillation. Within the oscillatory region, we observe a monotonical decrease of the frequency when IR level increases. This is consistent with the observation in the time domain simulation which shows that the period increases when IR value increases. To confirm this, the DDE-BIFTOOL [29] is used to obtain a plot of period of oscillation against IR values as shown in Figure 2.11, whose result is consistent with our frequency analysis. In [33], the authors stated that the period of oscillations did not appear to significantly depend on irradiation level. As investigated here, the period correlates to the IR dose. Therefore, more explorations need to be carried out.
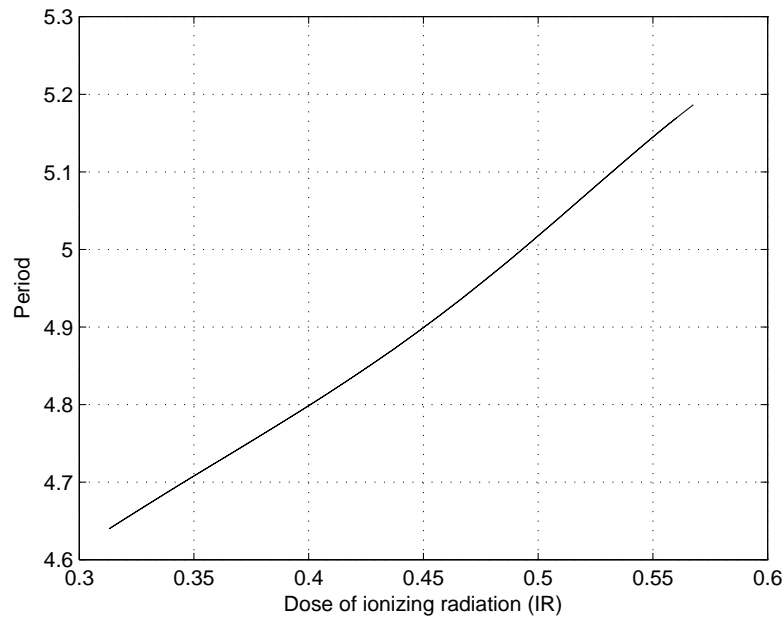
Figure 2.11: Period of oscillation against different IR values analyzed by DDE-BIFTOOL [29]

2.5.2 Experimental Design

Based on the above analysis, a frequency shifting phenomenon is predicted by the mathematical model. As stated earlier, one purpose of performing frequency analysis is that it will help to determine the number of samples (N) and sampling frequency (F_s) necessary in the practical experiment such that the original signal is not distorted. N determines the total number of data points needed to sample the concentration of p53 while F_s determines how frequent the measurements need to be made. By optimally selecting N and F_s, much cost and time can be saved for carrying out the verifying experimental design. Here, we do not intend to adopt the optimization formulation, such as linear or nonlinear programming, as introduced in [4].

For DFT operation, perfect reconstruction of the original time domain signal from the discrete time samples requires $k = N\frac{F_0}{F_s}$ to be an integer where F_0 denotes the dominant frequency in the input signal. This is to guarantee that the DFT result can represent the original signal perfectly in the frequency domain. Else, the frequency domain representation is only an approximation of the original signal. Bearing this in mind, the method proposed here is to fix F_s as an integer multiple of F_0, e.g. $F_s = 3F_0$. N can then be chosen as any multiple of 3 and k will always be an integer. In other words, as few as 3 samples are required to obtain the frequency domain representation of the p53 concentration theoretically. In fact, this is an ideal case because the Nyquist frequency merely gives the lower bound of the sampling frequency. In practice, we usually need higher frequency than that to prevent aliasing problem due to the existence of noise. However, the balance between the accuracy and the cost must be taken into account. Therefore, following the aforementioned design, the DFT condition can be satisfied while the time and cost of conducting the verifying experiment can be minimized.

With the previous simulation result, the predicted frequency F_0 with respect to different IR values can be obtained. Hence corresponding F_s can be calculated. Herein one thing to note is that F_0 obtained in the simulation may not be the same as the actual frequency of oscillation obtained in experiment. Therefore, further adjustments of F_s

and N might be needed in practice since the model presented here is more of qualitative nature. However, the method presented here can be employed directly as guidelines of designing experiments provided that a reasonably reliable mathematical model is built. Therefore, a suggested experiment procedure could be designed as follows.

1. Build a quantitative-reliably enough mathematical model from more available datasets

2. Select the stable oscillatory simulated data to be dealt with by Matlab routine FFT, and identify the dominant predicted frequency F_0

3. When performing the experiment, measure the concentrations of p53 and Mdm2 every T ($= \frac{1}{n \times F_0}$, where $n = 4, 5 \ldots$) time after entering the steady oscillation stage

4. Collect N ($= k \times n$, where $k = 2, 3, \ldots$) numbers of data points in total

5. Reconstruct the actual oscillation including amplitude and frequency information from the collected data points after filtering the noise signal

2.6   Discussion

The threshold mechanism discovered by the bifurcation analysis has not been verified by the current experimental data so far. Hence, a verifying experiment is suggested to be conducted. Once the experiment is done, two scenarios may occur. On one hand, if a higher steady state of p53 is observed, which coincides with the prediction from the model, the model is validated, revealing true mechanisms in the core regulation to some extent. On the other hand, if the observation shows that there are still sustained p53 oscillations or low level of stable steady state, then we have to refine or revise our current model. More aspects, such as the downstream activities of p53, which are involved in cell cycle, apoptosis and detailed DNA repair process [35], may have to be considered

to improve the model's reliability, or, the role of ATM should be re-evaluated to modify the interactions with the core regulation.

Regarding the frequency analysis, the predicted frequencies are normalized by the sampling frequency and converted from dimensionless time factor. According to Table 2.2, we can easily discriminate the differences in the frequency domain, and the practical measurement period usually takes several hours to days. Therefore, it is believed that the differences in the time domain are considerable in reality.

Furthermore, the real dominant frequency of p53 oscillations awaits more experimental data to be determined precisely. This real frequency will help to improve the model and design the experimental procedure. First, it can be utilized to refine parts of time-related parameters, such as degradation rate, repairing rate, etc. Secondly, with the knowledge of real frequency, it can replace the step 1 and 2 of the experimental procedure mentioned in Section 2.5.2. The following steps could get along directly without the help of mathematical model. Moreover, the concentration information is necessary for accurate reconstruction in practice. At this point, measurement techniques need to be improved to accomplish this.

Admittedly, with the existence of the inevitable noises and fluctuations under current technical status, the guideline provided is somewhat conceptual, still requiring further fine tuning and modifications with respect to practical considerations, such as noise filtering, statistical data processing, etc. Handling the noise is an open question in quantitative systems modelling. There are several ways for noise reduction besides the statistical inference of data. First of all, technical improvements help to make high-precision measurements [71]. Secondly, considering the variability between cells, stochastic modelling methods will comply with the natural essence more compared to deterministic modelling [59]. Thirdly, it is required to develop method to position the noise origin [6] and use specific mechanisms to lower the noise impact [75]. Therefore, it is a long way to fully develop quantitative computational biology.

In the current model, the variable *dam*, representing DNA damage, helps to resemble

the damped oscillation after irradiation. The simulation results show that the magnitude of oscillations will monotonically decrease after the impulsive stimulus of irradiation, which depends on the degradation time constant T_2 in Eq.(2.11). Without the variable *dam*, the ATM's influence will fast disappear through a decaying dynamics, i.e. the oscillation will not persist for a long duration, which violates the common observations. Besides, ATM is expressed as the independent variable of Hill function in Eq.(2.6) and Eq.(2.8). The selection of K_1 and K_2 plays an important role for the tuning effects of ATM, which need to give enough tuning range for ATM. Otherwise, ATM will become dispensable.

To emphasize again, our main focus here is to develop a predictive mathematical model such that it recaptures the observations and provides new insights. So far, the model is a scaled version and semi-quantitative, and the parameters are mostly estimated by approximations or trial-and-error. This is due to the lack of quantitative reliable experimental data, such as the real-time concentrations. Our emphasis is to determine which variables to be included and identify their relationships than the precise values [108]. Once new data become available, the model can be refined and make more accurate predictions.

2.7 Conclusion

In this chapter, a new mathematical model was proposed to explain the inherent mechanisms in p53-Mdm2 core regulation in response to DNA damage. The selections of the components and the dynamical relations among them were based on biological facts and reasonable approximations. Nevertheless, many factors or details were not included, such as DNA repair mechanisms, the volume ratio of cytoplasm and nucleus where proteins and messenger RNAs are exported in and out, etc.

The main structure was built in delayed differential equations. The simulation environment was in Matlab mainly using dde23 [88]. Selections of parameters and simplification assumptions were proven qualitatively appropriate by the good agreement

of simulation results with the experimental phenomena. Admittedly, parameters are scaled version and most of them were obtained from trial-and-error. They are far from the real values, which require to be refurnished in an iterative way. With the new-coming evidence, we are confident the model will become more capable.

In addition, a more detailed investigation was performed to analyse the bifurcation of p53's concentration with respect to the dose of ionizing radiation to predict a new threshold mechanism used to explore this core regulation. Meanwhile, the phenomenon of frequency shifting was observed from the simulation results. To help discriminating different frequencies, Fourier frequency analysis was applied to transform the oscillations in the time domain to distinct pulses in the frequency domain. Furthermore, based on the dominant frequency identified, an experiment procedure was provided to give suggestion on sampling frequency and data point number for the wet lab.

Chapter 3

Model Validation of Petri Net for Apoptosis Pathways

In this chapter, we enlarge our picture to build a model of p53-dependent apoptosis pathways, which establish the environment for the investigation of microRNA. To avoid the burdens to identify the numerous parameters in ordinary differential equations, we resort to discrete formalism—Petri net—to describe the inter-connections among the components. Furthermore, we propose two approaches to verify the discrete model's reliability. The work in this chapter paves the way for the next-step exploration.

3.1 Introduction

3.1.1 Classical Apoptosis Pathways

Apoptosis is a well-regulated cell death process which is required for development of normal tissues and suicide of threatening cells. A mostly used example is the resorption of the tadpole tail at the time of its metamorphosis into a frog. Malfunction of this process will result in severe diseases, such as cancer. There is another form of cell death called cell necrosis. Cells that die induced by acute injury will swell and burst, spilling their contents to the environment. This eruption triggers a potentially damag-

ing inflammatory response. Different with necrosis, apoptosis happens gently without damaging neighbours, expressing in the shrink and condense of cell content, collapse of cytoskeleton, engulfment by phagocytic cells [2].

This phenomenon was first studied in *Caenorhabditis elegans*. In this organism, 1090 somatic cells are generated in the formation of adult worms, of which 131 of these cells die at particular points, which is invariant between worms, showing that there must be an accurate and programmed control mechanism in nematode. So apoptosis is also called cell programmed death [40]. Apoptosis is conserved in higher organisms. Although it is studied for almost 30 years, the whole process remains to be elucidated. The huge number of molecules and interactions put difficulties in the way of understanding on this process.

As is known, the classical apoptosis pathways can be mainly divided into two types— extrinsic pathways and intrinsic pathways [28]. No matter which type is applied, the whole process will be mediated by the caspases, a family of proteases. Caspase will catalyse the hydrolytic breakdown of proteins, such as the lamin proteins, which form the nuclear lamina underlying the unclear envelope. The functions are performed in a cascaded style, i.e. the activated caspases will active other members of the family, resulting an amplifying cascade process. The initial procaspase, inactive precursors of caspase, is triggered by the apoptosis signal, e.g. DNA damage, γ radiation, UV light, survival signal withdrawal and signalling from the death receptors.

Extrinsic pathways involve transmembrane receptor mediated interactions. The death receptors, most of which are members of tumour necrosis factor (TNF) receptors, are bound by respective natural ligands, leading the death domain (DD) of the receptor transmitting the signal from the cell surface to the intracellular signalling pathway. Resorting to the adaptor protein, the death-inducing signalling complex (DISC) is formed. This complex consists of procaspase-8, procaspase-10, adaptor protein, e.g. FADD, and c-FLIP, which will inhibit the activities of caspase-8 and FADD. The formation of DISC will result in the autocatalytic activation of procaspase-8, which is followed by the

activation of caspase-3—the main executioner caspase, and the consequent proteolytic cascade.

Intrinsic pathways include a different way to trigger the caspase cascade[2][54]. Non-receptor stimuli, such as oncogene stress, DNA damage, hypoxia, will deliver the proapoptotic and antiapoptotic protein, all of which are members of BCL-2 family, to the mitochondria surface, positively or negatively resulting in an opening of the mitochondrial permeability transition (MPT) pore, allowing the release mitochondria protein into the cytosol. The group of mitochondria proteins consists of cytochrome c, Smac/DIABLO. Cytochrome c binds and activates Apaf-1 and procaspase-9 as well, forming an "apoptosome". Apoptsome will trigger a caspase cascade, initiating caspase-3, finally leading to apoptosis. Smac/DIABLO will promote apoptosis by reducing inhibitors of apoptosis protein (IAP) activity.

All the above introduction could refer to Figure 3.1.
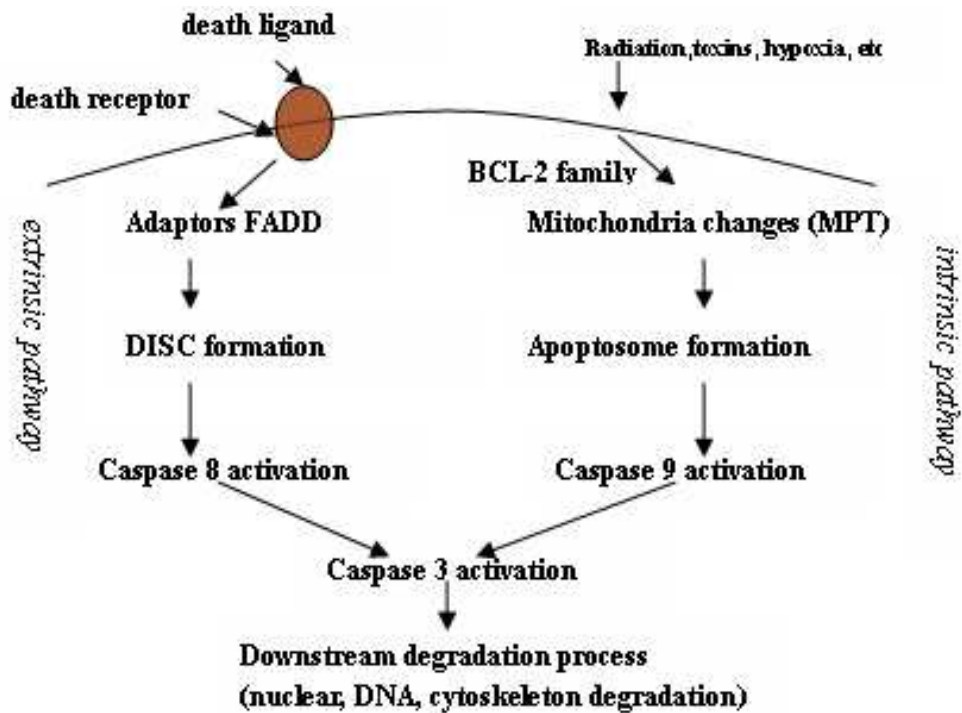
Figure 3.1: extrinsic and intrinsic apoptosis pathway

In the intrinsic pathways, p53 is the transcription factors of Bcl-2 family members, pro-apoptosis protein BAX, Puma, Noxa. As for the extrinsic pathway, Fas(CD95/Apo-1) is also the target. All these targets are crucial elements in this whole process. So the importance of p53 is obvious and p53's importance in apoptosis has been widely accepted. The pictorial description is shown in Figure 3.2.
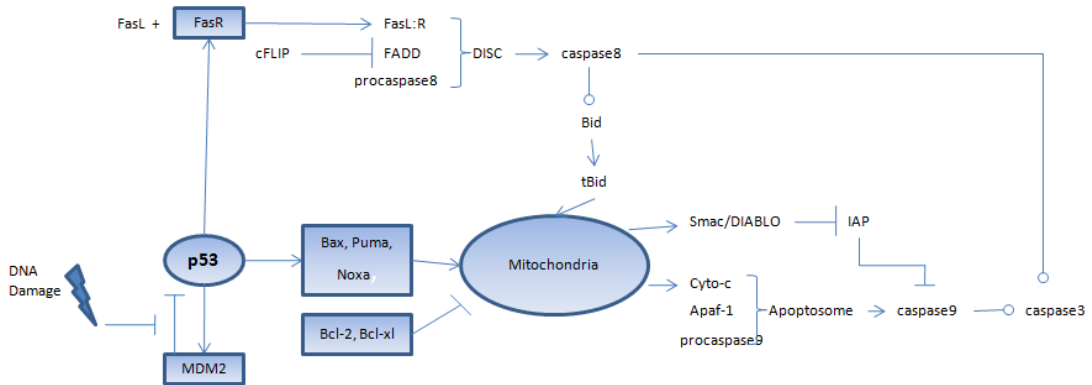
Figure 3.2: Block Diagram of p53-apoptosis pathway. Arrow represents activation; bar-headed line represents inhibition; circle-headed line represents catalysis.

There are also several survival pathways which influence the cell death activities, such as growth factor related pathways, NFκB related pathways, etc. More details about survival pathways could be found in Chapter 4.

3.1.2   Objective

Systems biology is a field of growing importance in biology research. It is concerned with the modelling, simulation and analysis of biological processes within a modelled biological system. By harnessing the vast calculative power of computer systems, a systems biological approach with a model at its centre can break the bottleneck of wet lab research to make new discoveries [48].

Naturally, a strong and accurate model is required before we can start to tap the calculative power of computers. There is thus a cyclic process in model development that requires the validation of the model, ensuring that the intended biological processes

that have been modelled have some forms of mathematical representations [107]. Due to limited experimental data, it is unavoidable to have incomplete or even wrong models. Therefore, it is a critical task in systems biology to check the model's correctness, which is called model validation problem. For example, in the genetic regulatory network, incomplete information of relations will lead to uncertainties of connections among genes and regulatory proteins. Sometimes, the model needs to tolerate the inconsistent information from different sets of experimental data. Under such circumstances, the models are required to be step-wisely developed and validated to increase the confidence to reflect the reality.

Model validation is commonly addressed by simulation and test method, which verifies the model performance by the agreement with the data points. For example, in [41], the models describing different binding conditions of Bcl-2 to pro-apototic proteins were tested by comparing the simulation results with the observed experiment data. In discrete domain, one sophisticated approach is model checking. The system is modelled by certain form of transition system. The algorithm will exhaustively search all the reachable states avoiding the miss of any single state. Due to the efficiency, it could be introduced to large-scale networks. Therefore, most discrete formalisms are applying model checking to do the validation tasks, such as the work done in [10, 16]. In our work done in Chapter 4, we also adopt the underlying idea to solve the problem of target validation of microRNAs. However, in this chapter, two solutions are provided to address model validation issue. Both are based on the analysis of basic state equations under Petri net framework, which are considered as most straightforward way to solve model validation problem. One is based on the invariant analysis; the other one is based on the reachability analysis.

In the sequel, the background of Petri net formalism is introduced in Section 3.2, and the necessary theoretical bases are followed immediately. In Section 3.3, the modelling of apoptosis pathways is presented followed by the invariant analysis result on it in Section 3.4. The reachability analysis in Section 3.5 includes two approaches to do the

validation and one artificial example and case study are used to illustrate the procedures accordingly. Finally, this chapter ends with the concluding words.

3.2 Petri Net

3.2.1 Petri Net Introduction

Petri nets have a graphical representation of discrete event system. Simple but pragmatic interface makes Petri nets user-friendly and popular. Since 1960s, Petri nets have been widely employed to model the systems of computer science, communication protocol and industrial management, etc [68].

Petri net is chosen as a modelling approach for biological system because of the natural affinity in three folds. 1) Biological systems are bipartite, consisting of the species/substrates and their interactions. 2) They are concurrent, allowing several interactions to take place simultaneously and independently of each other. 3) Finally, the interactions are stochastic and non-deterministic. Bearing these features, Petri net was firstly employed to model biochemical systems in [79]. From then on, Petri net and its extended variants are popular modelling frameworks for biological networks [22, 57].

A Petri net is a directed, bipartite multi-graph. Bipartite nodes consist of *places* and *transitions*. Graphically, places are denoted as circles, and transitions are rectangles or bars. Arcs are used to connect places with transitions. Same types of nodes have no arcs linking in between.

The meanings of the transitions and places depend on the modelling context. For example, in event/condition systems, a transition stands for the occurrence of an event, while the input places of the transition are the pre-conditions, all the necessary conditions needed to trigger the event; and the output places are post-conditions, all the conditions satisfied after the occurrence of the event. In biochemical reactions, transitions represent chemical reactions. The input places and output places are respectively the reactant compound and product compound in these reactions. Sometimes, one net

can have mixed types of places and transitions for different interpretations.

Petri nets are executable. The execution is based on the idea of *tokens*, which are graphically depicted by dots and reside in places. The distribution of tokens in places is called *marking*. Firing of transitions will change the distribution of tokens. Hence, Petri net is also dubbed "token game". The marking of all places comprises the state space of Petri net. One example is given in Figure 3.3.
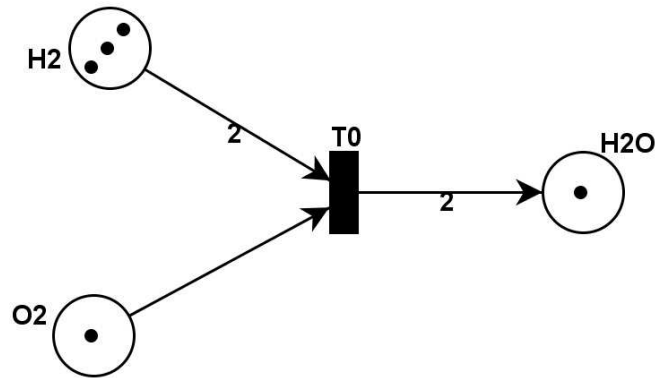
Figure 3.3: Petri net example showing the synthesis of water molecule. There are 3 places, H2, O2 and H2O, 1 transition T_0. The weights of arc connecting from place H2 to transition T_0 and T_0 to place H2O are 2. The weight of arc connecting from O2 to T_0 is one, usually omitted by default.

The definitions of Petri nets are mathematically expressed in the following forms, which are adapted from [68].

Definition A Petri net is a 5-tuple, $PN = (P, T, F, W, M_0)$ where,

$P = \{p_1, p_2, \cdots, p_m\}$ is a finite set of m places,

$T = \{t_1, t_2, \cdots, t_n\}$ is a finite set of n transitions,

$F \subseteq (P \times T) \cup (T \times P)$ is a set of arcs,

$W : F \to \{1, 2, 3, \cdots\}$ is a weight function,

$M_0 : P \to \{0, 1, 2, 3, \cdots\}$ is the initial marking,

$P \cap T = \emptyset$ and $P \cup T \neq \emptyset$.

The weight of an arc from place p_i to transition t_j, and from transition t_j to place p_i is represented as $w(p_i, t_j)$ and $w(t_j, p_i)$, respectively. The pre- and post-set of transitions

$t \in T$ are $\bullet t = \{p \mid w(p, t) > 0\}$ and $t\bullet = \{p \mid w(t, p) > 0\}$, respectively. Similarly, the pre- and post-set of places $p \in P$ are $\bullet p = \{t \mid w(t, p) > 0\}$ and $p\bullet = \{t \mid w(p, t) > 0\}$, respectively. The state of a Petri net is the marking $M : P \rightarrow \{0, 1, 2, \cdots\}$, a non-negative integer-valued vector. The i-th component of the marking vector, denoted by $M(p_i)$, represents the number of tokens at place p_i.

Definition A transition is said to be *enabled* in marking M, if all places in its pre-set have at least as many tokens as the weights of the arcs connecting those places to the transition, that is, if $\forall p \in \bullet t, M(p) \geq w(p, t)$, transition t is said to be *enabled* in the marking M.

Definition Firing an enabled transition changes the token distribution by removing as many tokens as the weights on the arcs leading in from all its pre-set places and adding as many tokens as the weights on the arcs leading out to all its post-set places, that is, by firing an enabled transition t, the change in token distribution from M to M' is given by

$$
M'(p) = \begin{cases}
M(p) - w(p, t) & \forall p \in \bullet t \\
M(p) + w(t, p) & \forall p \in t\bullet \\
M(p) & \forall p \notin \{\bullet t \cup t\bullet\}
\end{cases}
$$

In the example shown in Figure 3.3, the transition T_0 is enabled, and firing the transition makes the marking change from $M_0(\text{H2,O2,H2O}) = [3\ 1\ 1]^T$ to $M_1(\text{H2,O2,H2O}) = [1\ 0\ 3]^T$.

These changes in the state of the system can be represented in an algebraic form using the state equations given by

$$
M_k = M_{k-1} + C^T \mathbf{u}_k,
$$

where $C = [c_{ij}]$ is an $n \times m$ incidence matrix of integers; the superscript T stands for transpose; $\mathbf{u}_k$ is called firing or control vector, which is an $n \times 1$ column vector of $n - 1$ **0**'s and one nonzero entry, a **1** in the i-th position indicating that transition i fires at

the k-th firing. The typical entry of incidence matrix C is given by

$$c_{ij} = c_{ij}^+ - c_{ij}^-,$$

where $c_{ij}^+ = w(i, j)$ is the weight of the arc from transition i to its output place j and $c_{ij}^- = w(j, i)$ is the weight of the arc to transition i from its input place j.

Theorem 3.2.1 *(Necessary Reachability Condition)[68] Suppose that a destination marking M_d is reachable from initial marking M_0 through a firing sequence*

$$\{\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_d\}.$$

The state equations can be derived as follows.

$$M_d = M_0 + C^T \sum_{k=1}^{d} \boldsymbol{u}_k, \tag{3.1}$$

which can be rewritten as

$$C^T \boldsymbol{x} = \triangle M, \tag{3.2}$$

where $\triangle M = M_d - M_0$ and $\boldsymbol{x} = \sum_{k=1}^{d} \boldsymbol{u}_k$.

Here $\mathbf{x}$ is an $n \times 1$ column vector of non-negative integers and is called the firing count vector. The i-th entry of $\mathbf{x}$ denotes the number of times that transition i must fire to transform M_0 to M_d.

3.2.2 Invariant Analysis

- Transition invariant: Suppose X has non-negative integers as entries and $C^T X = 0$. Then X is called transition invariant, also called T-invariant. Stated differently, in any execution which corresponds to X, the resulting marking is the same the initial marking. The physical meaning is that there is no net change in the concentration levels and the system enters steady state.

- Place Invariant: Suppose Y has non-negative integers as entries and $CY = 0$. Then Y is called a place invariant, also called P-invariant. That's to say, the sum of all tokens multiplied by the weigh of the corresponding place is invariant. The Y-weighted count of any reachable marking M is equal to the Y-weighted count of M_0.

3.2.3   Reachability Analysis

By Theorem 3.2.1, the existence of a non-negative integer-valued solution of state equations (3.2) is a necessary condition for the reachability of Petri net, i.e. if the destination marking M_d is reachable from the initial marking M_0, there exists non-negative integer solution for state equations (3.2). The sufficiency of this condition is only restricted to acyclic Petri nets, a subclass of Petri nets. Acyclic Petri net, having no directed circuit, belongs to such class of Petri nets [43].

3.3   Modelling of Apoptosis Pathways

3.3.1   Model Structure

The apoptosis modelling by Petri net was discussed in the paper by Heiner et al [38]. T-invariant analysis was performed on their model as a means of analysis and model validation. It is of interest to see what a related analytical calculation, the P-invariant analysis, will reveal if it is performed on the model.

While the model developed by Heiner et al [38] is used as a starting point to build the model, for P-invariant analysis to be possible, it is necessary for a couple of structural differences to be put in place.

Firstly, input and output transitions must be excluded. Input transitions are transitions providing tokens to a place but do not consume tokens from other places. Conversely, output transitions remove tokens but do not provide tokens to any place. These "boundary" transitions are found in Heiner's model together with substrates that oth-

erwise did not have any means to gain or lose tokens.

Secondly, inhibitory interactions must be modelled because many such interactions take place within the apoptotic pathways, such as between Bcl-2 and Bax [90]. Such interactions usually involve the consumption of both the inhibiting substrate and the inhibited substrate into an inactive complex.
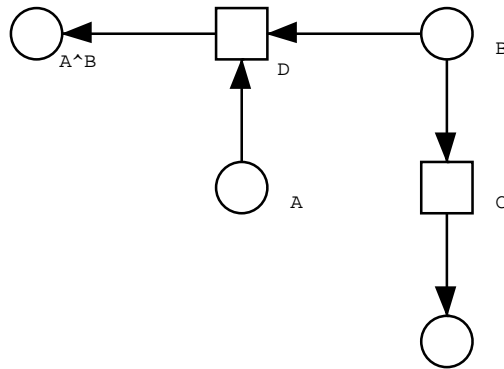
Figure 3.4: Inhibiting transition D between place A and B.

Figure 3.4 shows the representation of such an inhibiting interaction D between substrates A and B forming the inactive substrate complex $A^{\wedge}B$. This consumes tokens from B and down-regulates the occurrence of the "forward" reaction C. Although conceptually substrate A is the inhibiting species and B the inhibited species, A and B are considered mutually inhibiting since both are consumed in the inhibitory process.

These two structural features of the model are necessary for P-invariant analysis to be carried out, which will be explained later.

3.3.2 Petri Net Model

The basic mechanisms and processes could refer to the background introduction in Section 3.1.1. The modelling is also borrowing the evidence from the following additional sources [96, 28, 92, 46, 90, 42, 61, 30, 76, 89, 86, 11].

Extrinsic Pathway

Figure 3.5 shows the partial Petri net model of the extrinsic pathway as described by the above sources. The formation of the DISC (Death Inducing Signaling Complex) can be seen as a stepwise incorporation of such substrates as the Fas ligand (FasL), Fas receptor (Fas, activated by p53), Fas-associated protein with death domain (Fadd) and procaspase-8. DISC then releases the active caspase-8, which triggers the caspase cascade involving caspase-3. Caspase-8 itself is deregulated by cFLIP. All place names with a $^\wedge$ symbol represent inactivated species.

Intrinsic Pathway

Figure 3.6 shows the model of the intrinsic pathway as described by the sources. Up-regulation of NOXA and PUMA by p53 and their deactivation of Bcl-2 is modelled. Bax, which is up-regulated by p53 and down-regulated by Bcl-2, releases cytochrome c from the mitochondria. This stimulates the formation of the apoptosome, which cleaves and activates caspase-9 from its pro-form, initiating the caspase cascade just like in the extrinsic pathway.

Various inhibitory interactions are modelled, such as between SMAC/DIABLO and IAP, between IAP and executioner caspase-3 and procaspase-9, as well as between pro- and anti-apoptotic substrates Bax and Bcl-2.

It should be noted that for model simplicity and readability, several places in the model represent more than one substrate indicated by its place name. These substrates share similar functions and a fair extent of redundancy. Substrates that represent multiple similar substrates are shown in Table 3.1.

Roles of Bid

Finally, Figure 3.7 shows the partial model of the functions of Bid (tBid) as it is up-regulated by caspase-8. This provides a link from the extrinsic pathway to the intrinsic pathway. By releasing SMAC/DIABLO and cytochrome c, as well as inhibiting Bcl-2,
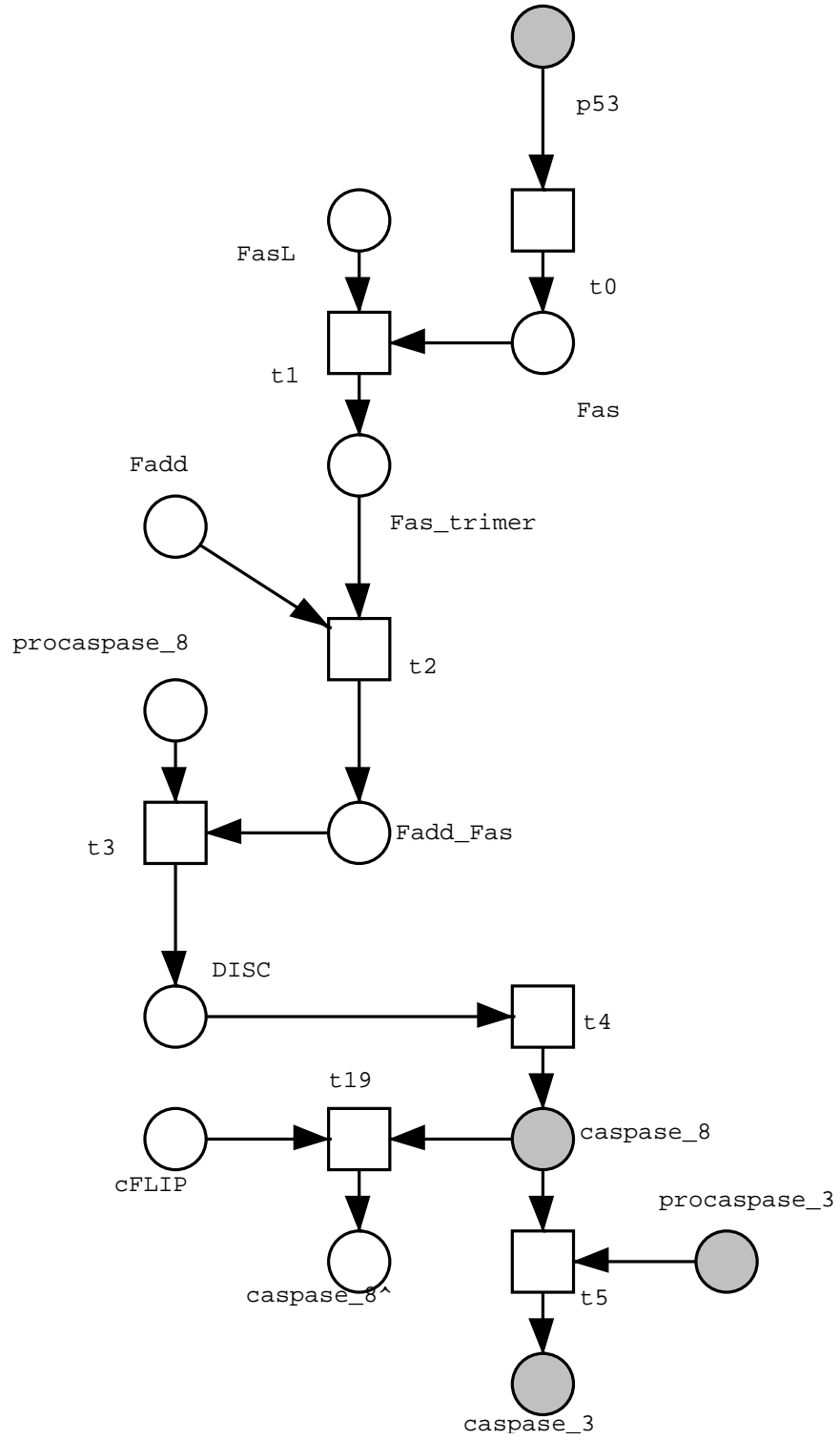
Figure 3.5: Modelling the extrinsic pathway
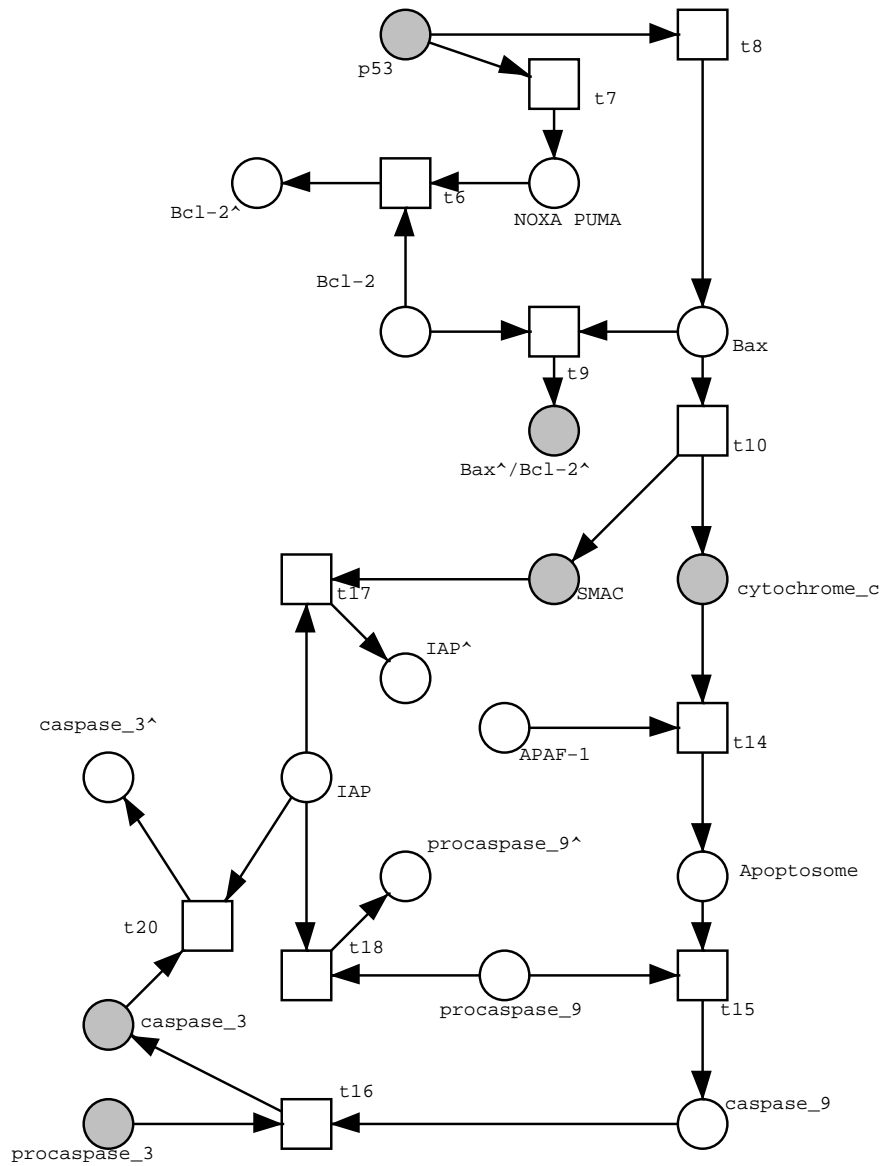
Figure 3.6: Modelling the intrinsic pathway

Table 3.1: Multiple substrates represented by single substrate

Name	Substrates Represented
Fas	Death receptors including Fas and TNFR1
FasL	Ligands of death receptors including FasL and TNF-α
Fas_trimer	Complex of 3 death receptors with ligands
FADD_Fas	Protein with death domain related to death receptor
procaspase_3	Procaspase-3, -6, -7
caspase_3	Active executioner caspases(-3, -6, -7)
Bax	Proapoptotic Bcl-2 family proteins (Bax, Bak, etc.)
Bcl-2	Antiapoptotic Bcl-2 family proteins (Bcl-2, Bcl-XL, etc.)

the role of tBid in the intrinsic pathway is similar to Bax. An intermediate caspase_8_i is introduced along with a transition t_x to characterize the enzymatic interaction which caspase-8 has on Bid. In this way, this "crossing-over" phenomenon will not cause a breakdown in the extrinsic pathway in which Bid is always competing for caspase_8 tokens, preventing the extrinsic caspase cascade.

These three net components are intentionally displayed in this work separately to promote readability and at the same time distinguish them as major pathways. The full net model can be obtained by joining together the three partial models at the place nodes indicated in grey.

3.4   P-invariant Analysis Result

Heiner et al. have established this ability with T-invariant analysis [38]. It would be of interest to see if P-invariant analysis could be used to the same purpose. In this section, P-invariant analysis will be performed on the above Petri net model, and ultimately established as a means of validating the model.
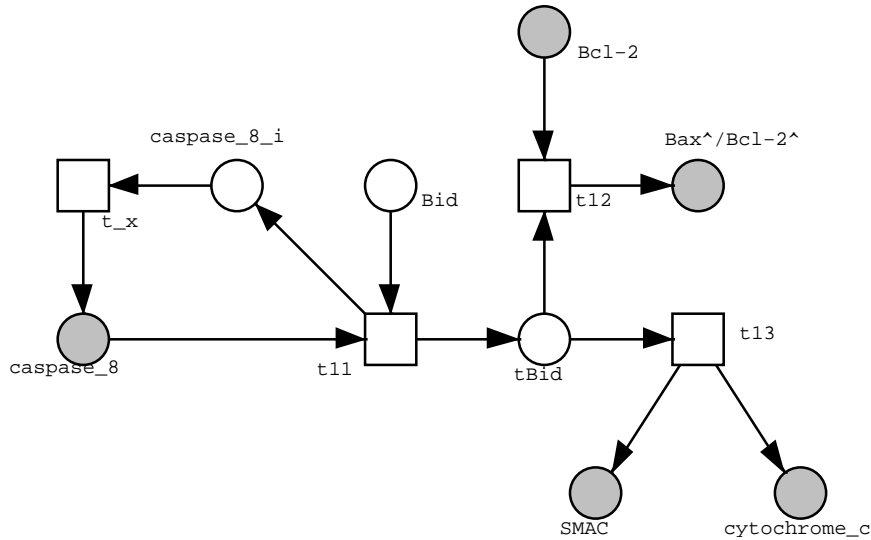
Figure 3.7: Modelling the roles of Bid

3.4.1 P-invariant of Model

P-invariant analysis allows the identification of conservatory networks that always maintain a certain conservation of tokens. Applied to the model, it is hoped that networks containing substrates (places) that are related by a certain signal transduction pathway would emerge as the invariants. The conservation of tokens can be interpreted as a conservation of the same biological signal within the network. Hence, places that are linked in a P-invariant are possibly linked by the same biological signal.

The structural changes as mentioned in Section 3.3.1 are necessary so that there will not be any unregulated flow of tokens into/out of the net via input and output transitions. Otherwise, this would make P-invariants impossible since weighted sum of tokens would then be uncontrolled. Inactive species are also included as a way to provide conservation of tokens for every transition, in order for their firings to be tracked by P-invariant analysis.

Using the Snoopy software [80] to construct the Petri net and its companion software Charlie [26] to perform P-invariant analysis, 13 P-invariants are found. These invariants, and their constituent places including deactivated/inactive substrates are listed in

Table 3.2 below.

3.4.2 Interpretation of P-invariant

To make sense of the invariants, critical places in the invariants need to be identified first. Two such critical places are places that do not have any incoming edges, and those that do not have any outgoing edges, which can be aptly termed start and stop places, respectively. For example, in the first invariant, cFLIP is a start place and caspase_8$^{\wedge}$ is a stop place (refer to Figure 3.5). While there may be multiple stop places or start places per invariant, there must be at least one of each per invariant.

After these critical points have been identified, the rest of the places can be used to link the start place to the stop place using any combination of transitions and edges in between. In doing so, conservatory signalling pathways originating from the start place can be traced out to their final destination.

Identified P-invariants can thus be interpreted by considering them with respect to the start and stop places. Table 3.3 shows the list of P-invariants and their interpreted biological significance.

Invariants 6, 7, 8 and 9 are connected to each other because they all identify the same interactive pathway. The only difference lies in the starting places (procaspase_8, Fadd, Fas, and p53, respectively) that finally lead to the caspase cascade via the extrinsic pathway. In invariant 9, p53 is the starting place and its pro-apoptotic killing activity of anti-apoptotic signals originating from Bcl-2 and IAPs can be easily identified.

Invariants 10, 11, 12 and 13 are also connected because of the same four different starting points (as in invariants 6 to 9) mentioned above.

In fact, invariants 5, 9 and 13 are also connected and do not identify any unique interactive pathway. The only difference is the choice of the starting point for the "lower" segment of the intrinsic pathway. In invariant 5, it is Bax that starts the propagation of signal via cytochrome c, leading to the caspase cascade; in invariant 9, APAF-1 is implicated for the same reaction, and in invariant 13, it is procaspase_9.

Table 3.2: P-invariants and places involved

No.	Places in P-Invariant
1	cFLIP, caspase_8$^\wedge$
2	Bcl-2, Bcl-2$^\wedge$, Bax$^\wedge$/Bcl-2$^\wedge$
3	procaspase_3, caspase_3, caspase_3$^\wedge$
4	IAP, IAP$^\wedge$, procaspase_9$^\wedge$, caspase_3$^\wedge$
5	p53, NOXA PUMA, Bax, Bcl-2$^\wedge$, Bax$^\wedge$/Bcl-2$^\wedge$, cytochrome c,Apoptosome, caspase_9, caspase_3, caspase_3$^\wedge$, Fas, Fas_trimer, FADD_Fas, DISC, caspase_8, caspase_8$^\wedge$, caspase_8_i, Bid, tBid
6	APAF-1, Apoptosome, caspase_9, caspase_3, caspase_3$^\wedge$, procaspase_8, DISC, caspase_8, caspase_8$^\wedge$, caspase_8_i
7	APAF-1, Apoptosome, caspase_9, caspase_3, caspase_3$^\wedge$, Fadd, FADD_Fas, DISC, caspase_8, caspase_8$^\wedge$, caspase_8_i
8	APAF-1, Apoptosome, caspase_9, caspase_3, caspase_3$^\wedge$, FasL, Fas_trimer, FADD_Fas, DISC, caspase_8, caspase_8$^\wedge$, caspase_8_i
9	p53, NOXA PUMA, Bax, Bcl-2$^\wedge$, Bax$^\wedge$/Bcl-2$^\wedge$, SMAC, IAP$^\wedge$,APAF-1, Apoptosome, caspase_9, caspase_3, caspase_3$^\wedge$, Fas, Fas_trimer, FADD_Fas, DISC, caspase_8, caspase_8$^\wedge$,caspase_8_i, Bid, tBid
10	procaspase_8, DISC, caspase_8, caspase_8$^\wedge$, caspase_3, caspase_3$^\wedge$, procaspase_9, procaspase_9$^\wedge$, caspase_9
11	Fadd, FADD_Fas, DISC, caspase_8, caspase_8$^\wedge$, caspase_3, caspase_3$^\wedge$, procaspase_9, procaspase_9$^\wedge$, caspase_9
12	FasL, Fas_trimer, FADD_Fas, DISC, caspase_8, caspase_8$^\wedge$, caspase_3, caspase_3$^\wedge$, procaspase_9, procaspase_9$^\wedge$, caspase_9
13	p53, NOXA PUMA, Bax, Bcl-2$^\wedge$, Bax$^\wedge$/Bcl-2$^\wedge$, SMAC, IAP$^\wedge$, procaspase_9, procaspase_9$^\wedge$, caspase_9, caspase_3, caspase_3$^\wedge$, Fas, Fas_trimer, FADD_Fas, DISC, caspase_8, caspase_8$^\wedge$, caspase_8_i, Bid, tBid

Table 3.3: P-invariants interpretation

No.	P-Invariant Interpretation
1	cFLIP resulting in deactivation of caspase-8
2	Antiapoptotic Bcl-2 family used in inhibiting proapoptotic signals, or deactivated by NOXA and PUMA
3	procaspases being activated into caspase-3, -6, -7, and deactivated by IAP
4	IAPs deactivating procaspase-9 and caspase-3, -6, -7, or being deactivated by SMAC
5	p53 resulting in deactivation of Bcl-2, and activation of caspase -3, -6 and -7 via extrinsic and intrinsic pathways. Caspase-8 and tBid involvement into the intrinsic pathway is also shown.
6	same as 9
7	same as 9
8	same as 9
9	p53 resulting in deactivation of Bcl-2, deactivation of IAP, and activation of caspase-3, -6 and -7 via extrinsic pathway only. Caspase-8 and tBid activity into the intrinsic pathway is also shown. APAF-1 is shown as the stimulus for the intrinsic pathway.
10	same as 13
11	same as 13
12	same as 13
13	p53 resulting in deactivation of Bcl-2, deactivation of IAP, and activation of caspase-3, -6 and -7 via extrinsic pathway only. Caspase-8 and tBid activity into the intrinsic pathway is also shown. Procaspase-9 is shown as the stimulus for the intrinsic pathway.

Hence, invariants 5 to 13 collectively represent the workings of both the extrinsic and intrinsic pathway, each identifying a different substrate that is important as a starting requirement for the pathways. Such substrates include (as represented by their places) p53, death receptor and ligand, proteins with death domains, procaspase-8, APAF-1 and procaspase-9.

The presence of multiple invariants that have the same biological significance is primarily due to places that only serve a single function of being a substrate for the next reaction. Such places, like APAF-1, and Fadd, do not affect the signalling pathways in more meaningful "secondary" ways other than simply being present for the next step to occur. An extended Petri net that involves these places in other dynamic ways will eliminate this problem of "repeated invariants".

It is noted that caspase_3$^\wedge$ is the stop point in invariants 5 to 13. This, however, does not mean that both the intrinsic and extrinsic pathways always end up in deactivated caspases-3, -6 and -7 (and hence termination of apoptosis). Caspase_3$^\wedge$ is the stop point in this case because further pathways downstream of caspase_3, i.e. the execution pathway [90], are not included in the model.

In the model, caspase_3 is taken to be equivalent to the trigger of apoptosis itself, since it is assumed that high levels of caspases-3, -6 and -7 would always lead to the execution pathway. In this right, caspase_3$^\wedge$ should be viewed as a stop place while interpreting the P-invariants.

3.4.3 Model Validation using P-invariant

The full list of biological activities involved in the Petri net model, as identified through P-invariant analysis is thus given as follows:

- Main Extrinsic Pathway leading to apoptosis

- Main Intrinsic Pathway leading to apoptosis

- Deactivation of caspase-8 by cFLIP

- NOXA and PUMA's upregulation by p53, deactivating anti-apoptotic Bcl-2 family members

- Mutual inhibition between pro- and anti-apoptotic Bcl-2 family members

- IAPs deactivating procaspase-9 and caspases-3 and -7

- SMAC/DIABLO deactivating IAPs

- Upregulation of tBid by caspase-8 to facilitate intrinsic pathway

This list fully includes all biological processes that have described in the sources [96, 28, 92, 46, 90, 42, 61, 30, 76, 89, 86, 11]. It is thus evident that P-invariants can be employed to analyse and validate extended versions of this Petri net model and similar signalling transduction networks.

3.4.4 Discussion

P-invariant analysis has thus been established as an additional method of validating Petri net models of signalling transduction networks. It is recommended that additional validation techniques, such as T-invariant analysis, be used in conjunction with P-invariant analysis, described in this section, when validating models, so as to further increase model confidence and prevent over-reliance on a single technique.

Besides validation via P-invariant analysis, the model used in this section has also been validated using T-invariant analysis. The results of the T-invariant analysis are the same as those compiled in Section 3.4.3. Common mechanisms can also be found in Heiner's work.

Compared to P-invariant analysis, T-invariant analysis also requires a set of structural requirements. P-invariant analysis will provide some overlapping pathways as shown in the last section. In this aspect, the results by P-invariant is not so intuitive as the ones by T-invariant. However, the P-invariant results reveal the complementary evidence to T-invariant.

The confidence of protein-protein interactions indicated in this work is thus strong enough for it to be the subject for future developments and model extensions, so as to better study the intrinsic and extrinsic apoptotic pathways. Besides including more substrates as more discoveries are made by the scientific community, improvements such as a quantitative model can be considered.

3.5 Reachability Analysis Result

3.5.1 Problem Formulation

In this section, we will validate the Petri net model based on the reachability analysis. The necessary condition mentioned in Theorem 3.2.1 could be treated into the following two scenarios, which will give hints to do the model validation in systems biology.

First of all, the contrapositive of the necessary condition provides a sufficient condition for unreachability, i.e. if the state equations do not have non-negative integer solutions, M_d is not reachable from M_0. This condition can be used to test the unreachability of certain states. Furthermore, in the acyclic case, this condition is used to test both the reachability and unreachability of concerned states at the same time.

In the biological context, taking the genetic regulatory pathways for example, this condition could be utilized to test the expressions of certain genes. Initial marking could be generated according to the initial expression information. The incidence matrix could also be constructed when the relations among the genes are identified. Then, this condition may help to verify the biologists' hypotheses on the interested gene's expression. It can be imagined that we may often eliminate some impossible guesses through the state equations, which will save remarkable cost and energy for the experimenters.

Secondly, if the expression data are observed, which means, in Petri net formalism, the state equations should be definitely solvable. If it has non-negative integer solutions accordingly, the model of current Petri net is validated. It can be further analysed to get more insights and generate further hypotheses based on the current model. If there

does not exist a non-negative integer solution, the incidence matrix C is somehow wrong. We have to modify the structure of the Petri net, such as the connections and weights, reflecting the reactions and stoichiometric parameters.

To address this problem, we introduce the following two approaches. One is the Smith normal form based criterion, which is established to check the reachability of desired states. It is shown that if the condition violates the criterion, the state equations are unsolvable. Accordingly, the desired state cannot be reached. The other approach is to test the model structure by the solvability of state equations under different initial conditions. The model which agrees with known evidence will be retained. Integer programming is introduced to facilitate this problem.

3.5.2 Diophantine Equations

The discussion in the last section transforms the reachability problem to the existence of non-negative integer solutions of state equations (3.2). By inspection of the state equations, the entries of the incidence matrix are all integers. Additionally, the solutions are also required to be integers, which is correlated to the Diophantine Analysis.

In mathematics, a Diophantine equation is an indeterminate polynomial equation that allows the variables to be integers only. The solvability of Diophantine equations is proposed as the tenth of Hilbert problem. In general, Diophantine problems are unsolvable. But in the case of linear Diophantine equation, Bézout's identity is a perfect solution [44]. Bézout's identity is specific to one linear Diophantine equation with multiple unknowns. In the sequel, we will analyse a set of Diophantine equations using linear algebraic method first. Then, this problem will be reformulated as integer programming (IP), and the state equations are solved numerically.

3.5.3 Approach by Smith Normal Form Test

The Smith normal form applies for any matrix with entries in a principal ideal domain (PID). In particular, the integers are PID, so one can always calculate the Smith normal

form for an integer matrix [106].

Theorem 3.5.1 *[72] If $A = (a_{ij})$ is any integer matrix of order $m \times n$ with rank r, then there exists a unique integer matrix*

$$D_{m \times n} = \begin{bmatrix} d_1 & 0 & 0 & \cdots & 0 & & \cdots & 0 \\ 0 & d_2 & 0 & \cdots & 0 & & \cdots & 0 \\ 0 & 0 & \ddots & & & \vdots & & \vdots \\ \vdots & \vdots & & & d_r & 0 & \cdots & 0 \\ 0 & 0 & & & 0 & 0 & & \vdots \\ \vdots & \vdots & & & & & \ddots & \ddots \\ 0 & 0 & \cdots & & \cdots & & \cdots & 0 \end{bmatrix},$$

where $d_k > 0, k = 1, \cdots, r$, and d_k divides d_{k+1}. D is the Smith normal form of A, and $\{d_1, d_2, \cdots, d_r\}$ are called invariant factors. Furthermore, the matrices A and D are equivalent $(A \cong D)$, i.e. $A = PDQ$ for some unimodular square matrices P of order m and Q of order n.

The Smith normal form D can always be obtained through a series of elementary row and column operations on the original matrix A by left-multiplying matrix P and right-multiplying matrix Q correspondingly. Several software packages are also available for this computation, such as Fermat, GAP, MAGMA and so on.

The classical method using the Smith normal form to solve the linear Diophantine equations is described in [72]. In the following, one simple criterion to test the the existence of integer solution is presented. Before that, a lemma is given first.

Lemma 3.5.2 *There exist matrix $A_{m \times n} = (a_{ij})$ and vector $\boldsymbol{b} = [b_1, b_2, \cdots, b_m]^T$. Two matrices $A_0 = [A \ \boldsymbol{0}]_{m \times (n+1)}$ and $\overline{A} = [A \ \boldsymbol{b}]_{m \times (n+1)}$ are equivalent if and only if there exists an unimodular matrix C with order $n + 1$, such that $\overline{A} = A_0 C$.*

Proof The sufficiency is obvious. The necessity can be derived by the property that elementary row and column operations do not change the rank of matrix. And the

greatest common divisor of every k-th subdeterminant of two equivalent matrices is the same. ∎

Theorem 3.5.3 *Linear equations $AX = \boldsymbol{b}$, where $A, X, \boldsymbol{b}$ are integer matrices of size $m \times n$, $n \times 1$ and $m \times 1$, respectively, have integer solutions if and only if matrices A and $\overline{A} = [A \ \boldsymbol{b}]$ share the same set of invariant factors of the Smith normal form.*

Proof Necessity: Suppose $AX = \mathbf{b}$ has integer solutions $\mathbf{x} = [x_1, x_2, \cdots, x_n]^T$. Multiply the first n columns of $\overline{A}$ by $-x_1, -x_2, \cdots, -x_n$, respectively, and add the products to the $(n+1)$th column. Then, A_0 is obtained. Consequently, A and $\overline{A}$ share the same set of invariant factors.

Sufficiency: If A and $\overline{A}$ have the same invariant factors, so is for A_0 and $\overline{A}$. Therefore, $\overline{A}$ is equivalent to A_0. According to Lemma 3.5.2, there exists an unimodular matrix C with order $n + 1$, such that $\overline{A} = A_0 C$.

Suppose the last column of C is $C_{n+1} = [c_1, c_2, \cdots, c_n, c_{n+1}]^T$. The equations $AC_{n+1} = \mathbf{b}$ hold. Equivalently, the equations $AX = \mathbf{b}$ have integer solutions. ∎

By Theorem 3.5.3, it can be verified whether the given state equations of Petri net have integer solutions. This theorem can be efficiently applied by taking advantages of computational software. One limitation here is that the theorem only guarantees the existence of integer solutions. Non-negativity cannot be addressed by this theorem. More efforts are needed to completely solve this problem. However, it does provide an efficient way to do the first test. If this test cannot pass, the candidate will be definitely eliminated in the first round.

3.5.4 Approach by Integer Programming

Compared to the first approach, a more straightforward method is to solve the equations numerically. Because one main concern in practice is to confine the whole experimental procedure to be as simple and quick as possible, this concern can be formulated as an integer programming problem [85]. The total number of reactions or efforts can

be the objective function to be minimized, and the constrains are defined by the state equations. One example is shown as follows.

$$
\begin{aligned}
\text{minimize} \quad & \sum_{i=1}^{n} x_i \\
\text{subject to} \quad & C^T \mathbf{x} = \triangle M \\
& x \in Z, x \geq 0
\end{aligned}
$$

In addition, in genetic regulatory networks, the markings of places usually take binary values, i.e. 0 or 1. For instance, one gene is expressed high, which can be represented with "1" token in the corresponding place. Otherwise, it is represented with "0" token. On the other hand, the transitions usually represent these occurrences of reactions. Sometimes, to test the gene expression in terms of logical relations, some transitions are required to happen at most once, implying that the transforms are enabled or disabled, as the reasoning process does. The above considerations can be categorized as 0-1 integer programming or binary integer programming (BIP) where variables are required to be only 0 or 1. Several supporting software can solve BIP problems. Among them, LINDO is a powerful optimization tool for linear programming, integer programming, etc. [60] We may use the built-in numerical algorithm to solve the state equations. The solvability and non-negativity can be tested directly from the solutions.

3.5.5 Case Studies

In this part, two case studies will be presented to show applications of the aforementioned methods. In the first example, one biochemical reaction network is built in Petri net. Using Smith normal form test, one expected state is proven to be unreachable. The second example is derived from the lac operon genetic regulatory network in *E.coli*. The example will demonstrate the modification of model structure by binary integer programming.

Biochemical reactions

Suppose that there exist three molecules P1, P2 and P3. Three reactions can happen among them. The detailed relations are illustrated below. The enzyme molecules are omitted in the corresponding reactions.

Reaction 1: $P_2 \rightarrow P_1 + 4P_3$

Reaction 2: $2P_3 + P_2 \rightarrow 2P_1$

Reaction 3: $3P_1 \rightarrow P_2$
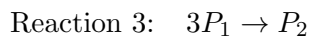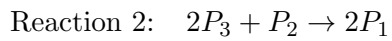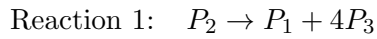
The Petri net model is shown in Figure 3.8 .
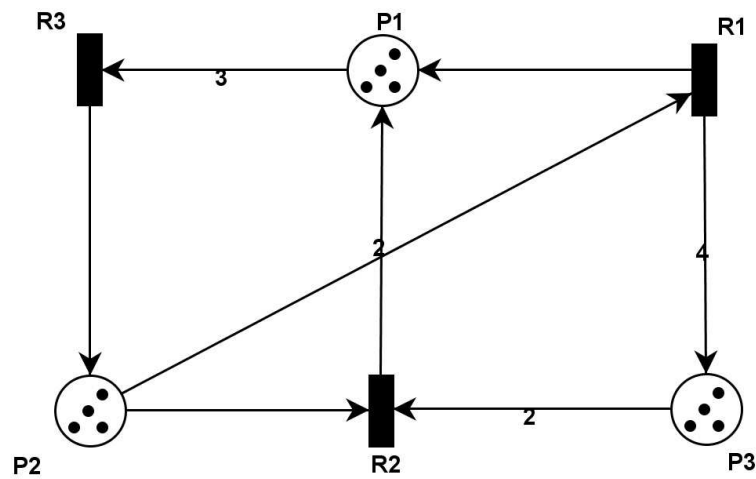
Figure 3.8: Petri net representation of case study 1. The numbers on the arcs denote the weights. Weights of 1 are omitted. Initial marking M_0 is $[4\ 4\ 4]^T$.

The model contains 3 places and 3 transitions. Therefore, it has a 3×3 incidence matrix.

$$
C_1 = \begin{bmatrix} 1 & -1 & 4 \\ 2 & -1 & -2 \\ -3 & 1 & 0 \end{bmatrix}.
$$

The state equations can be written as,

$$C_1^T x = \begin{bmatrix} 1 & 2 & -3 \\ -1 & -1 & 1 \\ 4 & -2 & 0 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \triangle M$$

The initial marking M_0 is $[4\ 4\ 4]^T$. The tokens can represent the numbers of molecules or scaled concentrations. One expected result M_{d1} is $[5\ 1\ 6]^T$.

$$\triangle M_1 = M_{d1} - M_0 = \begin{bmatrix} 5 \\ 1 \\ 6 \end{bmatrix} - \begin{bmatrix} 4 \\ 4 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 \\ -3 \\ 2 \end{bmatrix}$$

Next, we will perform the Smith test to check whether this expected state can be reached or not.

Smith normal form of C_1^T is

$$C_1^T \cong SNF1 = diag\{1, 1, 8\};$$

while, Smith normal form of $[C_1^T \ \triangle M_1]$ is

$$[C_1^T \ \triangle M_1] \cong SNF2 = [diag\{1, 1, 2\}, \mathbf{0}_{3\times1}].$$

The invariant factors are different. According to Theorem 3.5.3, these state equations cannot have integer solutions. It can be concluded that the desired state of three molecules cannot be realized. This conclusion can also be verified manually.

If the destination state is $M_{d2} = [6\ 2\ 4]^T$. Then, $\triangle M_2 = [2\ -2\ 0]^T$. Following the same procedure as above, the Smith normal form of new augmented matrix $[C_1^T \ \triangle M_2]$ is

$$[C_1^T \ \triangle M_2] \cong SNF3 = [diag\{1, 1, 8\}, \mathbf{0}_{3\times1}].$$

The same set of invariant factors guarantees the existence of integer solutions. And the solutions are $[x_1 \ x_2 \ x_3]^T = [1 \ 2 \ 1]^T$, which represent after firing R1 and R3 once, and firing R2 twice, the desired state can be obtained under the current initial marking. However, the solutions of state equations cannot determine the firing sequences. This drawback can be overcome by reachability graph complementarily provided that the token distribution allows so.

Lac Operon pathway

This section demonstrates how we can use the integer programming to solve the state equations, modify the undesired behaviour generated by the model and correct the model structure accordingly. Biological facts used for constructing the model are from the well-studied lac operon gene regulatory network happening in *E.coli* [2].

During the gene's transcription, the activity of a single promoter can be controlled by two different signals. The lac operon in *E.coli*, for example, is controlled by both the lac repressor and the activator protein CAP. Glucose and lactose levels control the initiation of the lac operon through their effects on the lac repressor protein and CAP. Addition of lactose increases the concentration of allolactose which binds to the repressor protein and removes it from the DNA. Addition of glucose decreases the concentration of cyclic AMP (cAMP), then the dissociation of cAMP with CAP removes the activator CAP from the DNA, which will lead to the shutdown of transcription. The transcripted gene of lacZ, lacY and lacA will also join this regulatory network. In this work, we only consider the initiation of the transcription under control of lactose and glucose.

It has been confirmed that, among four combinations of lactose and glucose, only the first condition illustrated in Table 3.4 will turn on the transcription.

Based on the description above, the relations among components are established. The first model is shown in Figure 3.9. Most places are the involved proteins, such as lactose, glucose, etc. However, place P_7 is used to represent the gene's status "transcription start". Transition T_1 indicates that the existence of lactose will lead to the gener-

Table 3.4: Transcription on/off by combinations of lactose and glucose

condition	lactose	glucose	transcription
1	+	-	yes
2	+	+	no
3	-	-	no
4	-	+	no

ation of allolactose. T_6 has the similar purpose. The transitions, $T_2, T_3, T_4, T_5, T_7, T_9$, which have only input arcs or output arcs are called sink transition or source transition, respectively. As the name implies, source transition can only produce tokens. For example, T_3, T_4, T_5 will produce the repressor protein, CAP and cAMP, respectively. On the other hand, sink transition will only consume tokens. For example, glucose will consume cAMP by T_7. T_2 represents that allolactose will make the repressor malfunction. T_9 is denoting the suppression by "repressor". Once T_9 fires, it will consume "complex" on the existence of "repressor". Thus, T_6 will be disabled, and transcription cannot start.
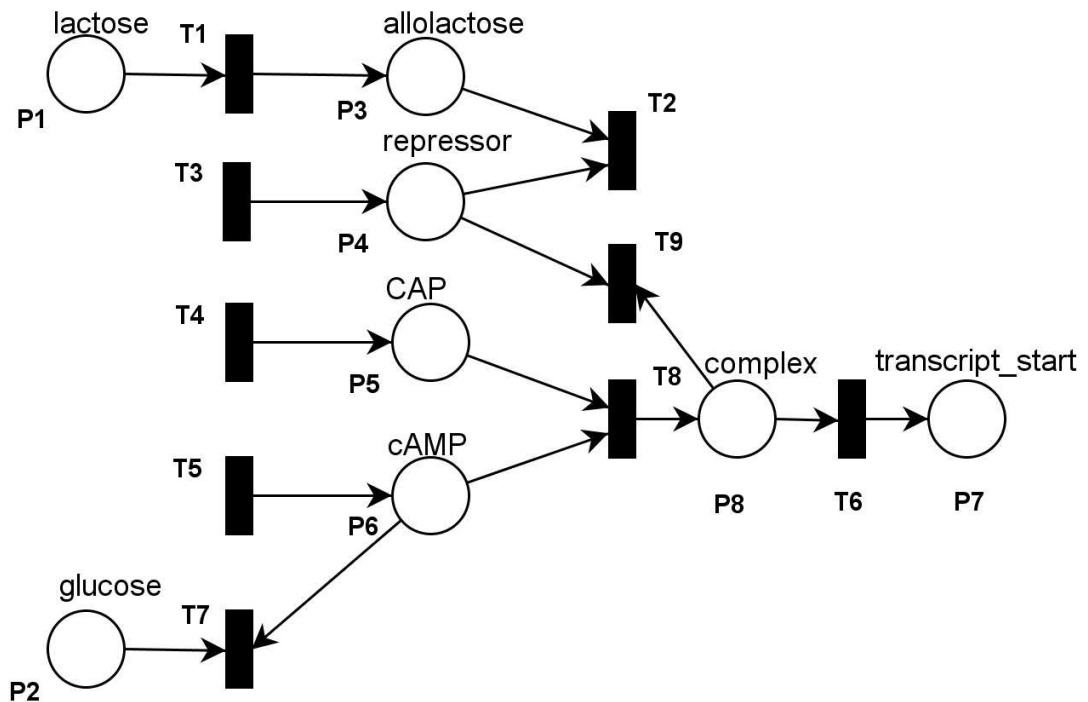
Figure 3.9: Petri net representation of Case study 2 Model 1

As explained earlier, this problem can be formulated as binary integer programming. The model should be tested by the four combinations in Table 3.4, which correspond to four initial markings $M_{10}(P_1, P_2, \cdots) = [1\ 0\ \mathbf{0}_6]^T$, $M_{20}(P_1, P_2, \cdots) = [1\ 1\ \mathbf{0}_6]^T$, $M_{30}(P_1, P_2, \cdots) = [0\ 0\ \mathbf{0}_6]^T$, $M_{40}(P_1, P_2, \cdots) = [0\ 1\ \mathbf{0}_6]^T$. The final marking is $M_F(P_1, P_2, \cdots, P_7, P_8) = [\mathbf{0}_6\ 1\ 0]^T$, in which the entry "1" at P_7 position denotes the initiation of transcription. Thus, there will be four state equations. If the state equations are expressed as $A^T \mathbf{x} = \mathbf{b}$, then there will be four candidates of $\mathbf{b}$. The incidence matrix of A is as follows.

$$
A_{9 \times 8} = \begin{bmatrix}
-1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -1 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\
0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & -1 & -1 & 0 & 1 \\
0 & 0 & 0 & -1 & 0 & 0 & 0 & -1
\end{bmatrix}
$$

Four candidates of b are

$$\mathbf{b}_1 = [\ \ -1\ \ 0\ \ \ 0\ \ 0\ \ 0\ \ 0\ \ 1\ \ 0\ \ ]^T,$$
$$\mathbf{b}_2 = [\ \ -1\ \ -1\ \ 0\ \ 0\ \ 0\ \ 0\ \ 1\ \ 0\ \ ]^T,$$
$$\mathbf{b}_3 = [\ \ \ 0\ \ \ 0\ \ \ 0\ \ 0\ \ 0\ \ 0\ \ 1\ \ 0\ \ ]^T,$$
$$\mathbf{b}_4 = [\ \ \ 0\ \ -1\ \ 0\ \ 0\ \ 0\ \ 0\ \ 1\ \ 0\ \ ]^T.$$

By Smith normal form test, all four conditions have same invariant factors set, implying that all four state equations have integer solutions. However, it is required that all the values of markings are binary value, i.e. "0" or "1". Therefore, Smith test cannot help further. Binary integer programming is used to solve this problem. The

formulation is listed below.

$$
\begin{aligned}
\text{minimize} \quad & \textstyle\sum_{i=1}^{9} x_i \\
\text{subject to} \quad & x_i = \{0, \ 1\} \\
-x_1 \quad & = \alpha \\
-x_7 \quad & = \beta \\
x_1 - x_2 \quad & = 0 \\
-x_2 + x_3 - x_9 \quad & = 0 \\
x_4 - x_8 \quad & = 0 \\
x_5 - x_7 - x_8 \quad & = 0 \\
x_6 \quad & = 1 \\
-x_6 + x_8 - x_9 \quad & = 0
\end{aligned}
$$

where $\{\alpha, \beta\}$ can be $\{-1, 0\}$, $\{-1, -1\}$, $\{0, 0\}$ or $\{0, -1\}$.

Using LINDO, conditions 2 and 4 cannot find the solutions subject to binary values, meaning that the initial markings under these two conditions cannot initiate the transcription, which accords with the evidence. Condition 1 can find the solution, which also coincides with the evidence that presence of lactose and absence of glucose initiate the transcription. One unexpected result is case 3. It is also solvable subject to binary value, and the solution is $\mathbf{x} = [\ 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1 \quad 0 \quad 1 \quad 0\ ]^T$.

Firing T_4, T_5, T_6, T_8 can also make the desired state reachable. This result contradicts the evidence, and provides the warning of incorrectness of the model structure. After check, the model cannot guarantee whenever "repressor" contains 1 token, it has to be consumed by transition T_9. So the model has to be modified accordingly.

In the new model shown in Figure 3.10, one intermediate place P_9, "rep_bar" is added to indicate the status that the repressor is consumed by allolactose. This place collaborates with "complex" to enable T_6 and finally initiates the transcription.
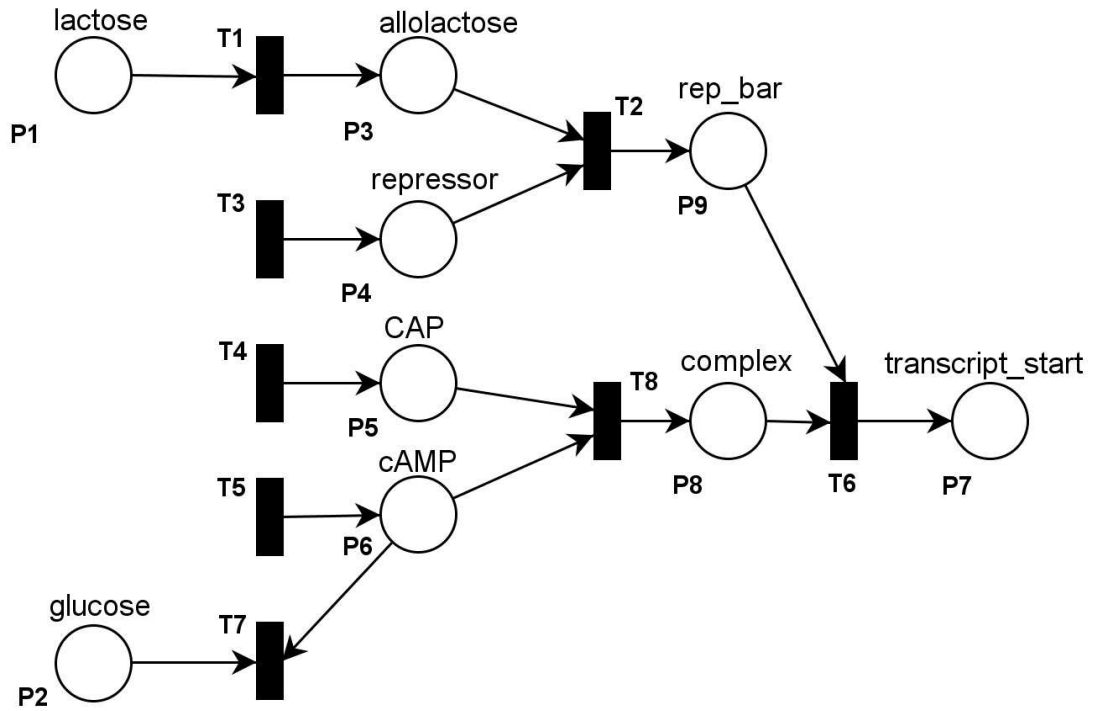
Figure 3.10: Petri net representation of Case study 2 Model 2

The incidence matrix becomes a 8×9 matrix with the following form.

$$
A_{8 \times 9} = \begin{bmatrix}
-1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 \\
0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -1 & -1 & 0 & 1 & 0
\end{bmatrix}
$$

The initial and final markings are similar with the previous model. The binary

integer programming problem becomes

$$
\begin{aligned}
\text{minimize} \quad & \sum_{i=1}^{8} x_i \\
\text{subject to} \quad & x_i = \{0, \ 1\} \\
-x_1 \quad & = \alpha \\
-x_7 \quad & = \beta \\
x_1 - x_2 \quad & = 0 \\
-x_2 + x_3 \quad & = 0 \\
x_4 - x_8 \quad & = 0 \\
x_5 - x_7 - x_8 \quad & = 0 \\
x_6 \quad & = 1 \\
-x_6 + x_8 \quad & = 0 \\
x_2 - x_6 \quad & = 0
\end{aligned}
$$

where $\{\alpha, \beta\}$ can be $\{-1, 0\}$, $\{-1, -1\}$, $\{0, 0\}$, or $\{0, -1\}$.

After performing the tests again, only condition 1 can be solved under current structure. This new model rules out other three impossible conditions. As concluded, this modification improves the model's reliability.

3.5.6 Discussion

In this section, through the analysis of solvability of state equations in terms of integer non-negative solutions, we may eliminate impossible biological states and, furthermore, check and improve the model structure. Two methods are proposed to solve Diophantine equations both in theory and numerically, followed by applications on two examples.

One limitation is that we need the complete marking information, which is for large systems often difficulty to specify. At this point, we will resort to model checking method in next chapter.

The Smith normal form test needs further improvements to take care of non-negative property. Moreover, as a necessary condition, solutions of state equations only provide

the information of firing times, rather than the firing orders and the existence of fire sequences. Although reachability graph does well at this aspect, it usually suffers from the dimension explosion. As for the integer programming method, the capability to perform this validation is smaller compared to model checking [25]. Many available tools, such as PEP, PROD, INA, etc., have the interface to perform large scaled model checking on Petri net model. So far, few studies have been done in the development of tools for Petri net validation using linear programming. However, the success of integer programming method will help to handle middle-scaled network efficiently. Therefore, users can choose the appropriate method to do the validations.

3.6 Conclusion

In this chapter, we proposed two methods to address model validation issue. Both were based on the state equations of Petri net.

In the first method part, it was demonstrated how apoptosis pathways modelled by Petri nets can be analysed and validated using P-invariants, similar to how T-invariants have been used to the same purpose in [38]. The structural requirements that the model must have for P-invariants were also described. The results from P-invariant, which agreed with the descriptions in the biological literature, proved the effectiveness of this method.

In second part, reachability analysis based model validation problem was discussed. Petri net was reused to model biological systems. An investigation on the existence of non-negative integer solutions of state equations introduced a new angle to rule out impossible biological states and further, check and improve the model structure. Two approaches based on the Smith normal form test and integer programming , were proposed to do the verifications and two subsequent case studies exemplified the successful applications in biological systems.

In next chapter, model checking based method is fully developed to address the target validation problem for microRNA. The method can be categorized into model

validation problem as well. However, it is investigated from a totally different angle through which the targets' validity is evaluated by multiple models which differ in the target involvement.

Chapter 4

MicroRNA Target Validation

The work in the previous two chapters lays the substantial foundations of our main problem. In this chapter, we evaluate the regulatory roles of microRNAs by studying the validation of repressive targets of microRNAs in the context of p53-dependent apoptosis pathway. Model checking based solution is developed and the experimentation is designed for the wet lab verification as well.

4.1 Introduction

4.1.1 MicroRNA

Being first discovered in 1993 [53], microRNA, a sort of tiny pieces of RNA, has attracted more and more scientists' interests since then, due to the mystery of regulatory effect in human cancer researches. MicroRNA (miRNA) refers to a large class of short length ($\sim$ 22 nucleotides) non-coding RNAs, which have negatively regulatory effects on the expression of other genes. MiRNAs act as a subunit of RNA-induced silencing complex (RISC), and specifically repress the target gene expression, through either direct cleavage and degradation of the target messenger RNA (mRNA) or blockage of the target mRNA translation process, by forming Watson-Crick base-pairings [111].

Mounting evidence shows that miRNAs are involved in many crucial biological pro-

cesses, such as development, differentiation, apoptosis and proliferation, and their aberrant expression can manifest various growth abnormalities, including cancer. Recent examination of miRNA expression profiles in tumour tissues has revealed that miRNA expression is nearly ubiquitously deregulated in various types of human cancers, and is found to have strong correlations with some known oncogenes and tumour suppressor genes. In addition, miRNA expression patterns are highly specific for cell-type and cellular differentiation status, which actually can be used to classify poorly characterized human tumours. Furthermore, more than 50% of miRNAs are located in cancer-associated genomic regions or in fragile sites. All these evidence suggests that miRNAs function in concert with classical tumour suppressors and oncogenes to regulate key pathways involved in cellular transformation and tumorigenesis. Therefore, a better understanding of miRNA regulatory effects in genetic pathways involved in cancer development should provide new and complementary insights into cancer pathogenesis, and should contribute to the diagnosis and treatment of this group of diseases [9].

4.1.2 Target validation problem

MicroRNA and the other typical type of small RNA molecule—small interfering RNA (siRNA)—usually participate into RNA interference (RNAi), which is a conserved biological mechanism for specific gene silencing [34]. This mechanism has been widely exploited in the fields, such as experimental biology, medicine, biotechnology, etc., among which the most potential relies on the provision of novel therapeutic approaches for caner treatment.

However, the target specificity of these two types of molecules is not often accurate enough to perfectly base-wise pair with the target mRNA due to RNA's structure. Compared to siRNA, miRNA suffers more serious troubles in accurate target pairing. For example, most of the animal miRNAs imperfectly complement the nucleotide pair; thus the screening of the genome finds much more targets than the true number. With the increase of evidence showing miRNAs' aberration expressions in tumour cells, miRNA

is considered to highly relate to carcinogenesis. The identification of miRNA repression target becomes the central job to investigate its role as tumour suppressor or oncogene. Several computational models are developed to do the target prediction. Despite the variety of prediction algorithms, the common issue of those predictions is that the tools return enormous number of targets so as to be considered that most of the results are false positive [56]. Taking hsa-mir-34a for example. Pictar [50] predicts 354 targets; TargetScan [56] gives 469 and the number from DIANA-microT V3.0 [63] is 575. Therefore, it is a critical task to help to accurately determine the targets out of the prediction pools.

4.1.3   Objective

However, it is hard to investigate the role of microRNA by itself; while it should be involved in a specific biological environment. Recent five papers [13, 21, 36, 78, 95] show the evidence connecting mir34a,b,c with p53 protein. P53 is considered as the most important nexus in human cancers, evidenced by the fact that mutations of p53 are founded in nearly 50% of human cancers. On the other hand, apoptosis, also called programmed cell death, effects highly the formation and death of tumor cells, in which many critical components are the targets of p53. Thus, it is naturally proposed to learn microRNA's role in p53-dependent apoptosis pathway.

We propose to use systems biology approach which is a holistic investigation of the dynamical behaviour of the biological pathways and utilize the properties to facilitate the screening process. Different from the one-by-one proof done by the biologists, we incorporate miRNA and its predicted targets into the context of certain pathways and investigate the dynamical profiles and causality relationships among the components in a batch mode. The basic idea goes as follows. We collect the real observed evidence from the wet lab as the evaluation criteria. If the involvement of one predicted target makes the pathways' behaviour violate the evidence, we will cancel its candidature provided that its absence leads to the agreement with the evidence, and vice versa. We

may confirm the target by the agreement with the evidence when it is present and the disagreement when it is absent.

To be more specific, we introduce the formal methods to accomplish this verification process. Among the several approaches to formal methods, in this work, we concentrate on the method of model checking due to the following two benefits [25]. First of all, it is fully automated. The checking process is executed without man's intervention. Secondly, if the checking result is false, it will return one counterexample to demonstrate the behaviour which falsifies the answer, which helps to locate the error source. We use the primitive idea of model checking and compare the satisfaction results from the models with and without considering miRNA inhibitions at the predicted places. Model checking technique has already been successfully applied in many researches of systems biology. For example, in the paper [10], model checking was introduced to validate the qualitative model about the nutritional stress regulation network in *E. coli*. In another work by [93], the authors used model checking to validate one model for *B. subtilis* sporulation. In both papers, model checking is mainly in charge of model validation, although a small portion of model predictions are suggested to be checked. In our work, the main focus is about the post-modelling stage to do further analyses and predictions on the models making full use of the power of model checking. To be further, we propose the experiment design suggestions to assist the miRNA target validation in the wet lab.

This chapter is organized as follows. In Section 4.2, we give a brief introduction of model checking technique, NuSMV model checker and the individual components, such as model and specification, to be used in this chapter. In Section 4.3, we develop our method in a problem-driven style and illustrate the whole procedure based on one artificial example. We then apply the method to miRNA targeting problem and elaborate the detailed checking results and possible experiment design schemes in Section 4.4. Finally, we end this chapter by the discussion of the accomplishments.

4.2 Model Checking

4.2.1 Introduction

Model checking, since firstly developed in the 1980s, has been widely used in the practical hardware and software design [25]. The purpose of model checking is to check the satisfaction of the desired property, technically called specification, by the considered model. In this approach, specifications are expressed in a propositional temporal logic, and circuit design and protocols are modelled as state-transition systems. An efficient search procedure is used to determine if the specification is true for the transition system. In other words, the transition system is checked to see whether it is a model of the specifications.

There are a number of applications of model checking in systems biology. The works differ in the modelling frameworks and studied topics [31]. The frameworks are of qualitative or semi-quantitative type, such as boolean networks [82, 32], Petri nets [93], state machine [3], etc. Some papers are used to demonstrate the functionality of model checkers [3, 19]. Among the literatures, most of the work focuses on modelling and analysis phases, where predictions are less made.

Compared to the old methods based on simulation and testing, model checking exhaustively explores and checks the satisfaction on each individual reachable state, avoiding the miss of any single behaviour. Several sophisticated techniques are also developed to assist the expansion of reachability tree to prevent state explosion, such ordered binary decision diagram (OBDD) [64], partial order reduction [25], unfolding technique [65], etc.

4.2.2 Transition System

The "Model" in "Model checking" is always related to reactive systems. Such systems may need to interact with the environment frequently. The state will represent an instantaneous description of the system that captures the values of the variables at a

particular instant of time. We also need to know the change of the state because of some actions or events. We may use a pair of states, state before and state after the action occurs, to determine the transition of the system. This kind of model is described by a Kripke Structure.

Let AP be a set of atomic propositions. A Kripke structure M over AP is a four tuple $M = (S, S_0, R, L)$ where

1. S is a finite set of states

2. $S_0 \subseteq S$ is the set of initial states

3. $R \subseteq S \times S$ is a transition relation that must be total, that is, for every state $s \in S$ there is a state $s' \in S$, such that $R(s, s')$

4. $L : S \to 2^{AP}$ is a function that labels each state with the set of atomic propositions true in that state

Thus, a path or a trajectory in the structure M initiated from a state s is an infinite sequence of states $\pi = s_0 s_1 s_2 \ldots$ such that $s_0 = s$ and $R(s_i, s_{i+1})$ holds for all $i \geq 0$.

Temporal logic has been proved to be useful for specifying concurrent systems, because they can describe the ordering of events in time without introducing time explicitly. Therefore, it is used to describe the desired specifications or properties of the system. It is often classified according to whether time is assumed to have a linear or branching structure, which correspond to linear temporal logic (LTL) and computational tree logic (CTL), respectively. They could be translated from human languages following the defined semantics.

4.2.3 Computational Tree Logic

Computational tree logic (CTL) is a branching-time logic, meaning that there are multiple options for the future state, which is non-determined. Therefore, there are many different paths from the stating state to reach the final state. Biological reactions inter-

leave to occur and there are non-deterministic choices for next state. Therefore, CTL is one ideal option to describe the evidence abstracted from the real observations.

The following part is mainly focused on how to build a syntactically correct CTL formulae and the interpretation of semantics.

Definition Let Φ be a set of atomic propositions. Then the following sentences are formulae of CTL:

1. For all $p \in \Phi$, p is a formula

2. True($\top$) and False($\bot$)

3. Let ϕ be a formula, then $\neg\phi$ is a formula

4. Let ϕ_1, ϕ_2 be formulae, then $(\phi_1 \wedge \phi_2), (\phi_1 \vee \phi_2), (\phi_1 \Rightarrow \phi_2), (\phi_1 \equiv \phi_2)$ are formulae

5. Let ϕ, ϕ_1, ϕ_2 be a formula, then $AX\phi$, $EX\phi$, $AF\phi$, $EF\phi$, $AG\phi$, $EG\phi$, $A[\phi_1 \mathrm{U} \phi_2]$, $E[\phi_1 \mathrm{U} \phi_2]$ are formulae.

The operators in the first four conditions are all traditional boolean logic operators, negation($\neg$), disjunction($\vee$), conjunction($\wedge$), implication($\Rightarrow$), equivalence($\equiv$).

The operators in the last condition are temporal operators, which are divided to the following two categories.

1. Path quantifier $A(\mathbf{A}\text{ll the paths}), E(\text{there } \mathbf{E}\text{xists at least one path})$

2. State operator $X(\text{ne}\mathbf{X}\text{t}), G(\mathbf{G}\text{lobally}), F(\mathbf{F}\text{inally}), U(\mathbf{U}\text{ntil})$

In CTL, the temporal operators must consist of any one from each category and one path quantifier must be followed by one state operator.

CTL formula is interpreted over transition systems in the structure of Kripke structure $K = \{S, S_0, R, L\}$, with $s \in S$, $\phi \in F$, where F is a set of well-defined formulae. Then the semantics of these operators are defined as below:

- $K, s \models \text{true}$ iff $K, s \nvDash \text{false}$

- $K, s \models p$ iff $p \in L(s_0)$

- $K, s \models \neg\phi$ iff $K, s \nvDash \phi$

- $K, s \models \phi_1 \wedge \phi_2$ iff $K, s \models \phi_1$ and $K, s \models \phi_2$

- $K, s \models \phi_1 \vee \phi_2$ iff $K, s \models \phi_1$ or $K, s \models \phi_2$

- $K, s \models \phi_1 \Rightarrow \phi_2$ iff $K, s \nvDash \phi_1$ and $K, s \models \phi_2$

- $K, s \models AX\phi$ iff $\forall\langle s \rightarrow s_1\rangle(K, s_1 \models \phi)$

- $K, s \models EX\phi$ iff $\exists\langle s \rightarrow s_1\rangle(K, s_1 \models \phi)$

- $K, s \models AG\phi$ iff $\forall\langle s_1 \rightarrow s_2 \rightarrow \cdots\rangle(s = s_1)\forall i(K, s_i \models \phi)$

- $K, s \models EG\phi$ iff $\exists\langle s_1 \rightarrow s_2 \rightarrow \cdots\rangle(s = s_1)\forall i(K, s_i \models \phi)$

- $K, s \models AF\phi$ iff $\forall\langle s_1 \rightarrow s_2 \rightarrow \cdots\rangle(s = s_1)\exists i(K, s_i \models \phi)$

- $K, s \models EF\phi$ iff $\exists\langle s_1 \rightarrow s_2 \rightarrow \cdots\rangle(s = s_1)\exists i(K, s_i \models \phi)$

- $K, s \models A[\phi_1 U \phi_2]$ iff $\forall\langle s_1 \rightarrow s_2 \rightarrow \cdots\rangle(s = s_1)\exists i((K, s_i \models \phi_2) \wedge (\forall(j < i)K, s_j \models \phi_1))$

- $K, s \models E[\phi_1 U \phi_2]$ iff $\exists\langle s_1 \rightarrow s_2 \rightarrow \cdots\rangle(s = s_1)\exists i((K, s_i \models \phi_2) \wedge (\forall(j < i)K, s_j \models \phi_1))$

where $K, s \models \phi$ denotes "satisfaction relation" between a model K, one of its states s, and a formula ϕ which holds if and only if the formula is true in that situation.

4.2.4   NuSMV

Many software tools are developed, which are providing fully automated approaches to perform the checking tasks. Among them, NuSMV is a famous tool to perform the checking task based on OBDD algorithm [24]. This tool provides a powerful checking ability and the built-in script platform offers a flexible solution to program the model

and specifications. Same as most model checkers, NuSMV also renders counterexample when the desired property is rejected by the model. In addition, the easy accessibility and configuration delight the using experience. Due to the above reasons, NuSMV is selected to facilitate the verification task.

Table 4.1 lists the comparison with two other model checkers, which are also investigated during the tool-hunting stage.

Table 4.1: Model checker comparison

Name	Formalism	LTL		CTL	
		Cap	C.E.	Cap	C.E.
Snoopy [80] Charlie [26]	Standard Petri net	yes	yes	yes	no
Process Analysis Toolkit (PAT) [84]	Process Algebra (Communicating Sequential Process)	yes	yes	no	no
NuSMV	Built-in modelling language	yes	yes	yes	yes

"Cap" stands for capability; "C.E" stands for counter example.

The models used in the verification process adopt a high-level abstracted discrete structure, usually in the form of automaton, Petri net, process algebra, etc. Petri net owns the natural affinity to describe the biochemical substances and reactions in an intuitive pictorial format [22]. In this work, we represent the biological network in extended Petri net, which includes the inhibition and read arc. Thus, activation, repression and catalysis reactions happened in biochemical network could be all considered under this framework. The places and the transitions in Petri net formalism stand for the biological components and the reactions respectively, both of which are written in the NuSMV build-in language. The detailed introduction of Petri net formalism could be referred to Section3.2. The programming rules of reactions are briefly appended into Appendix A.

In addition, CTL is employed to describes the behaviours of biological systems in this work. Therefore, CTL specifications are preferably derived to be checked by NuSMV.

The above descriptions could be summarised in the flowchart shown in Figure 4.1.
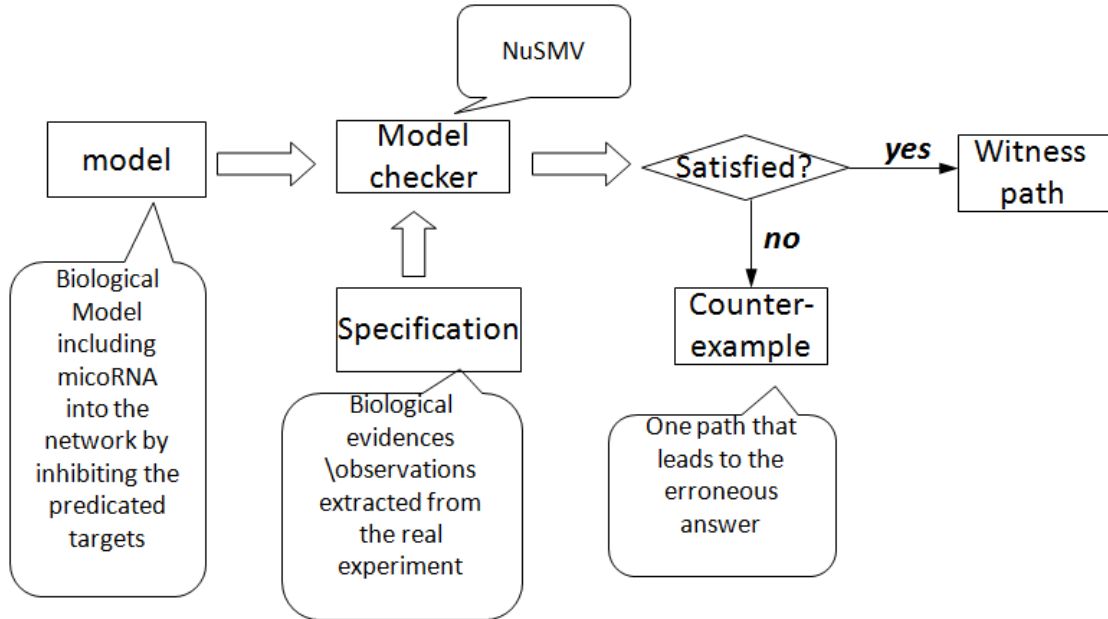
Figure 4.1: Model checking flowchart used in this work

4.3 Method Illustration

In this section, we will illustrate the method by using one artificial example. The illustration exemplifies the practical implementation and follows the problem-driven style. Prior to that, we introduce the example setup first.

4.3.1 Pilot Example

This example mimics a small-scale biological network including the typical biochemical reaction types. The structure is shown in Figure 4.2. In this example, it is assumed that the predicted targets are P2, P4, P10 and P13. P2 is the false target and the rest candidates are all true targets.

For the sake of convenience, three types of models are introduced herein. The same definitions apply in the subsequent sections.

- "phenotype model" is used to depict the original network structure without considering the miRNA's repression roles. The model shown in Figure 4.2 is the

"phenotype model" for this example.

- "miRNA model" initially involves the connections from miRNA to all the possible targets. As the check proceeds, this model updates step by step. If one target is confirmed to be false, it will be dropped in the subsequent rounds. If the target is confirmed to be true, it will be kept. Otherwise, the connections will be pending until the evaluation criteria can conclude.

- "true model" reflects the real connections. It is introduced for two purposes. On one hand, for illustration purpose, "true model" is used to generate the evaluation evidence. On the other hand, the final validation results should be consistent with "true model".
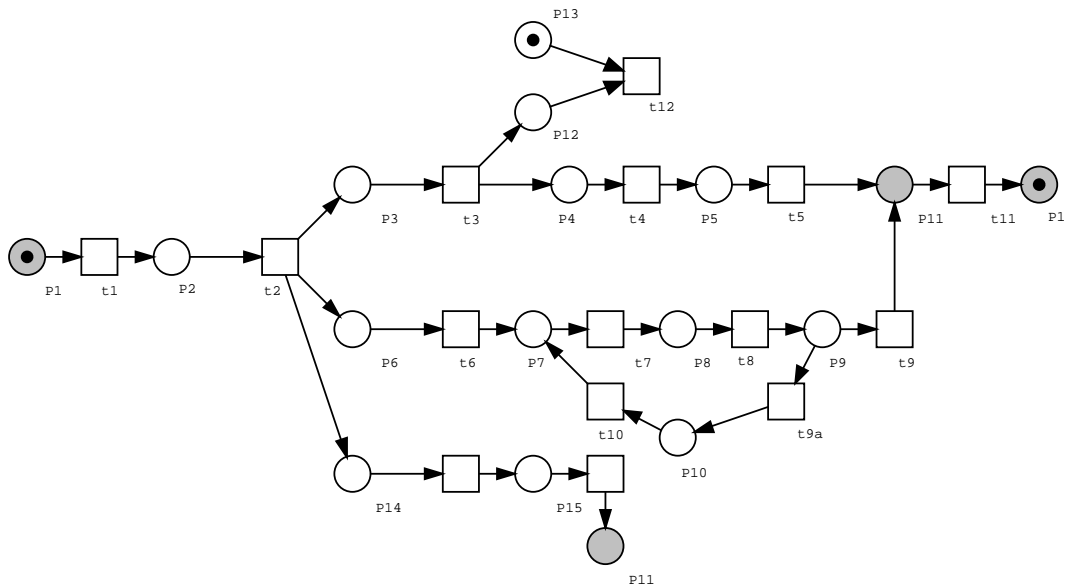
Figure 4.2: Prototype model of the pilot example. This model is built in Snoopy [80]. Circles represent places, and squares represent transitions. Grey places with the same name represent the identical component. The place marked by one dot is expressed and the corresponding state variable equals to 1. In this setup, it is assumed that P2 is the false target and P4, P10, P13 are true targets. MiRNA is introduced to disable the targets function by repression.

Table 4.2: Evaluation criterion for individual target

	Prototype model	miRNA model	Conclusion
Case 1	agree	disagree	false target
Case 2	disagree	agree	true target
Case 3	agree	agree	not decided
Case 4	disagree	disagree	not decided

"agree" and "disagree" are with respect to the results given by true model.

4.3.2   Target Validation

In this part, we will illustrate how to use model checking based approach to predict the targets validity and meanwhile how to effectively select the evidence to assist this checking process.

Evaluation Criteria

The evidence generated from "true model" is considered to be the fact. The CTL formulae converted from the evidence are evaluated by "prototype model" and "miRNA model". Four different cases arise. The comparisons and conclusions are summarized in Table 4.2. The idea is intuitive. Only the one which agrees with the answer from the true model will be considered to be correct. Case 1 and 2 can answer the correctness directly. As for Case 3, we have to change to another evidence because that evidence could not distinguish these two models. If Case 4 occurs, the improvement of prototype model must be made. Moreover, the evidence mentioned here is specifically designed for the individually investigated target, rather than general one for the whole set of targets. Because the interplay among targets may influence the effectiveness of the evaluations. The checking process is always stuck at certain critical positions in most cases. The selection criteria for evidence are discussed in Section 4.3.2.

For the sake of carefulness, we herein conclude the target is true or false with high possibility. The concern originates from the model's correctness and evidence's reliability. First of all, all the components and connection information to build the model is

from the biological discoveries. The information blast complexifies the modelling work. Moreover, the debates about certain proteins and genes are always existing in the biology society. It is inevitable to have some uncertainties in the model. Therefore, the model is only an abstraction of the reality. We pick the model which represents the interested aspects. On the other hand, we collect the experimental evidences from the reported data in the literature. Sometimes, the interpretation is not so strong as the data indicate. Some results may be against with the ones in other experiments. Furthermore, in most cases, the expected evidences are not always available. The evidence retrieval brings about many troubles in systems biology research. Regarding this point, we have a further discussion in Section 4.4. Due to these factors, based on currently available resources we can only present our results reticently.

Selection Criteria of Specifications
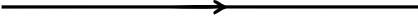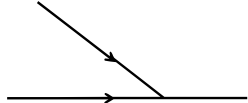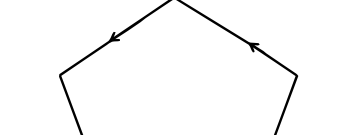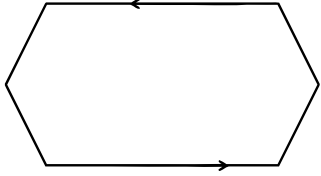
To make the checking process effective, we must deliberately select the specifications with respect to each target. Otherwise, the checking result could easily fall into Case 3 in Table 4.2.

We are considering the selection criterion from the following three aspects.

- component types

- component position

- query patterns

First of all, we must evaluate the component dynamical type, namely global and local types. Global components influence the whole dynamics seriously, such as P2, P11 in this example. Their positions usually lie on the main stream of the reaction chain. The dysfunction of the global components will influence the main outcome. On the other hand, local components influence the local dynamics and potentially cooperatively influence the global dynamics, such as P4, P7, P13, etc. P13 influences P12 only and P4 has the impact on its immediate downstream neighbour P5. If P4, P7 and P14

Table 4.3: Generic structure types

Type	Structure	Example
Line		P3-P4-P5
Collaborative		P12, P13
Feedback		P6-P7-P8-P9-P10-P7
Loop		P1-P2-P14-P15-P11-P1

are disabled simultaneously, the main stream will be blocked. Accordingly the global dynamics will be influenced.

Secondly, we must consider the component position information. We generalized into four position types—line type, collaborative type, feedback type and loop type. The illustrations and respective examples could be found in Table 4.3. All the other structures could be considered as the derivatives of the above four.

Thirdly, the query patterns summarize the translation rules from human descriptions of biological behaviours to CTL formulae. In the work done in [67, 20], the authors provided the detailed introductions and demonstrations. Here we re-summarize the patterns in Table 4.4 specially for our own purposes from [67].

Checking Procedure

As introduced before, in this example, four targets are predicted. They are P2, P4, P10 and P13. The specifications are designed as shown in Table 4.5. As can be seen, the specifications are specific for each candidate and meanwhile, the candidates are all of

Table 4.4: Query pattern

Query pattern	Human language description	CTL formula
Occurrence /Exclusion	It is possible for a state ϕ to occur	$EF\phi$
	It is necessary for a state ϕ to occur	$AF\phi$
	It is not possible for a state ϕ to occur	$!EF\phi$
Consequence	If a state ϕ occur, then it is possibly followed by a state ψ	$AG(\phi \rightarrow EF\psi)$
	If a state ϕ occur, then it is necessarily followed by a state ψ	$AG(\phi \rightarrow AF\psi)$
Sequence	A state ψ is reachable and is possibly preceded at some time by a state ϕ	$EF(\phi\&EF\psi)$
	A state ψ is reachable and is possibly preceded all the time by a state ψ	$EF(\phi U\psi)$
Invariance	A state ϕ can persist indefinitely	$EG\phi$
	A state ϕ must persist indefinitely	$AG\phi$
Repetitive	A state ϕ will appear infinitely often	$AGAF\phi$

local type. Therefore, the update order will not affect the final result. However, if the candidate is global, it will be prioritized to be the first.

Table 4.5: Checked formulae and results for pilot example

Target	CTL Formula	True model	Prototype model	miRNA model
P2	$AGAF(\text{P1} = 1)$	T	T	F
P4	$AG!(\text{P5} = 1)$	T	F	T
P13	$AG((\text{P12} = 1) \rightarrow AG(\text{P12} = 1))$	T	F	T
P10	$EF(\text{P8} = 1\&EF(\text{P7} = 1))$	T	T	T

"T" for true, "F" for false

The formula designed for P2 is $AGAF(\text{P1} = 1)$, which is interpreted as P1 will infinitely often appears and the loop behaviour can persist. After checking results from three models, "prototype model" which excludes the repression of miRNA on P2, generates the same answer with "true model". The inclusion of this repression in "miRNA model" violates the answer from "true model". Moreover, the counterexample zeros in on P2. Based on our evaluation criterion, P2 is confirmed as false target of miRNA and this repression connection should be dropped from "miRNA model" in the subsequent

steps.

The CTL formulae for P4 and P13 are designed as shown. The criterion tells that both are true targets. These selected specifications effectively distinguish between prototype model and miRNA model. However, if we choose the occurrence pattern formula $AF(\text{P}12 = 1)$ to check P13, it will become less effective. P13 will influence the future behaviour of P12 after P12 becomes active. However, $AF(\text{P}12 = 1)$ says P12 will eventually be active along all possible paths. It is not be able to handle the future behaviour after P12 becomes active.

P10 is a local and feedback type component. Via P10, P8 could occur before P7, against the main direction. If miRNA does not repress P10, this is true. Therefore, we preferably select the sequence pattern formula

$$EF(\text{P}8 = 1 \ \& \ EF(\text{P}7 = 1))$$

However, that formula is true under all three models, even in the case that the miRNA repression is considered. The witness path shows

$$\text{P1-t1-P2-t2-P6-t6-P7-t7-P8-t8-P9-t9-P11-t11}$$
$$\text{-P1-t1-P2-t2-P6-t6-P7.}$$

P8 could occur in the first cycle, which appears prior to P7 in the next cycle. Under such circumstance, our specification could not distinguish the results from "prototype model" and "miRNA model".

4.3.3 Experimental Design

As in the case of P10, in practice, we always encounter the similar problem of lacking evidence for the check use. We have to shift our focus to design the experiment to generate the useful evidence. This part addresses the issue that how we deliberately design the experiment to generate useful evidence when it is lacking. The design scheme

is proposed in terms of the experimental setup, initial conditions, observation points, etc. The design in this example is the simplest case. In reality, several factors should be involved, especially the technical feasibility. To avoid the occurrence of Case 4 in Table 4.2, some general guideline to is suggested in Table 4.6.

Table 4.6: General guideline for experiment design

Type	Query pattern	Observation point
Global	occurrence/exclusive	any places
Line/collaborative	consequence	downstream of parallel place
Loop/feedback	sequence	the place in loop/feedback loop
Loop	repetitive	the place in loop

Regarding the problem in this example, we design as follows. We zoom in to only consider the branching reactions starting from P6, ending at P9 shown in Figure 4.3. P2 and P11 could be knocked out to construct this environment. In the current design, the above formula could confirm that P10 is a true target.
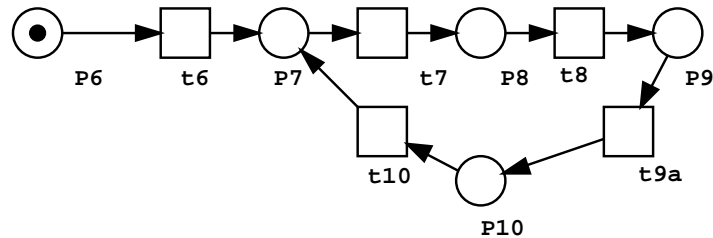
Figure 4.3: Branching reactions

Up to here, we have finished the identification for all the candidates. P2 is removed from the candidates list, and the other three are confirmed to be the true targets.

4.3.4   Model Modification

As mentioned before, the model is only an abstraction of the reality. Due to the current limitations of techniques and experimentation conditions, the model has to tolerate

uncertain and conflicting cases. Based on the available evidence, our methods could give the answer to the candidates. It could be the case that the final verifications by the biologists prove that the answer is wrong. Provided this new knowledge and the old information, we have to update the model to correct the connections and re-evaluate the roles of the components. The feedback from the biological society closes the loop for systems biology research.

To accomplish this model modification step, we still utilize model checking facilities. The idea is to use the counterexample to narrow down the searching range of error source and propose possible solutions. Take P13 for example. Our prediction confirms that P13 is the true target of miRNA. The real measurement tells that P13 is not the target. We have to delete the repression connection from miRNA to P13, which turns out to be the "prototype model". The old evidence $AG(\text{P12} = 1 \rightarrow AG(\text{P12} = 1)$ should be satisfied by "prototype model", which originally disapprove that evidence. The counterexample shows the path stops at the reaction t12, which means P13 will consume P12 to falsify the formula. The possible modification could be that there exists one missing component which has a higher priority to react with P13 than P12.

This problem always exists regardless of which formalism is chosen—continuous or discrete. With the advent of new knowledge, the model is changing all the time. The factors could be from the connections, parameters, components and even the principles. Therefore, the instructions to modify the model have to be problem-dependent. Our discussion here is providing one solution to identify the error source, and then propose the suggestion accordingly.

4.4 Mir-34 Target Validation

4.4.1 Mir-34

MiRNA was firstly discovered to regulate the larval development timing in *C. elegance* in 1993 [53]. Since then, more and more miRNAs have been identified and believed

that they may influence nearly all aspects of health and disease. Among them, aberrant expression of mir-34a in neroblastoma draw much attention from the biology society [103]. Especially, the correlation of mir-34a with the famous tumour suppressor p53 protein attracted intensive studies on this small RNA. Recent studies reveal that the expression of miR-34 family is robustly induced by DNA damage and oncogenic stress in a p53-dependent manner [37]. In addition, inactivation of miR-34a strongly attenuates p53-mediated apoptosis in cells exposed to genotoxic stress [78, 21]. Further studies confirm that miR-34a are directly transactivated by p53 [36]. Taken together, these findings strongly indicate that miR-34s can affect tumorigenesis in p53-dependent apoptosis [39]. However, it is unclear whether miR-34s, especially miR-34a, are *bona fide* tumour suppressors in their own right or just do fine-tuning of the gene expression program initiated by p53. A critical step to answer this question is to identify the targets of miR-34s in the apoptosis pathway. However, the identification of true targets represents a great challenge due to the current limitations of experimental techniques and bioinformatics tools.

In this section, we will apply the aforementioned model checking based approach to compensate the roles of biology and bioinformatics to facilitate the identification of mir-34a's true targets in apoptosis pathways.

4.4.2   Candidate Screening

As enumerated before, the listed computational tools render a great number of the target predictions in the whole genome. However, the modelling of a complete system is irrealizable and if possible, too complicated to do anything on it. Therefore, we choose a middle-scale network to realize the functionality of our proposed method and meanwhile, the chosen apoptosis pathway plays a significant role to regulate the cell death for tumorigenesis study.

To make the predication pools as complete as possible, we use the results from the database miRecords which fetch the enquiring results from 11 miRNA prediction

programs [110]. The union number reaches 1116. We overlap with the components in apoptosis pathways provided by KEGG pathway database [45] and finally narrow down to 15 targets listed in Table 4.7 into our investigation range. The complete list of predicted genes and the descriptions about the selected genes are summarized in Appendix B.

4.4.3   Modelling and Validation

As introduced in Section 3.1.1, classical apoptosis pathways consist of extrinsic and in-trinsic pathways. Extrinsic pathway is initiated by the ligation of the death receptors by external ligand stimuli whereas intrinsic pathway is mitochondria-dependent stimu-lated by non-receptor stimuli, such as DNA damage agents, oncogene stress, etc. Both processes can trigger the caspase cascade activation. Two typical initiator caspases are Caspase 8 and 9 in the extrinsic and intrinsic pathways respectively. p53's activation by the stimulating agents consequently transactivates some pro-apoptotic genes, BAX, PUMA, FAS, etc., which are playing the main roles to determine the cell-fate [90], and mir-34a as well. Anti-apoptotic proteins, such as FLIP, XIAP, BCL2, block the death procedure by antagonizing the pro-apoptotic proteins. Besides the classical path-ways, cell death is also regulated by survival factor withdrawal [55]. Therefore, survival pathways should be taken into account in the whole picture as well. In the apoptosis pathways we study presently, the survival pathways are mainly built around NF-κB and PI3K-AKT/PKB [91, 98]. The former one is another transcription factor, which can transactivate anti-apoptotic proteins XIAP, FLIP. The latter one is a kinase which phosphorylates and deactivates pro-apoptotic proteins, such as BAD, Caspase 9. The connection details could be found in the prototype model shown in Figure 4.4. The full list of place names and the representative genes are listed in Appendix B.

Before we make use of this model, we should confirm its correctness. As done in [10], we apply model checking technique to the prototype model to test the satisfaction of some desired specifications under the conditions without the existence of mir-34a.

Selected testing results are listed in Table 4.8. All the pro-apoptotic and anti-apoptotic properties could be fulfilled under specified conditions. With the confidence on our model, we proceed to the mir-34a's target validation task.

4.4.4 Checking Results

Different from the pilot example in Section 4.3, we extract all evaluation specifications from the experimental data. In most cases, mir-34a's level is too low to exert any biological functions [78]. Therefore, we have to consider two sources to get evidence. One is from p53-initiated apoptosis. Elevated level of p53 can raise mir-34a's repression ability. The other one is by considering the experiment with ectopic expression of mir-34a. To identify the role of mir-34a, the experiments usually introduce mir-34a from external sources.

We select the following examples to illustrate the results by our approaches. The summarized results could be found in Table 4.7.

Let us see the first candidate Caspase 8. In [12], it reported the occurrence of Fas-induced extrinsic pathway in the p53-mediated apoptosis. Fas and p53 were both activated, and accordingly mir-34a was active. Under this setup , Apoptosis occurred; namely Caspase 3 was necessarily expressed. This phenomenon could be expressed in CTL as follows.

$$AF(\text{ CASP3} = 1) \text{ true}$$

Caspase 8 is along the extrinsic pathway and has a direct impact on Caspase 3. Therefore, this property is ideally for the purpose to check Caspase 8. The checking results are listed in Table 4.7. The miRNA model which includes the repression of mir-34a on Caspase 8 falsifies that formula, whereas the prototype model which excludes the repression validates the description in the paper. Based on our criterion, Caspase 8 is concluded as false target. So far, no biological study is conducted on this verification. However, we may find the indirect data to prove our prediction from microarray dataset. Although the method by observing microarray data will miss the *bona fide* miR-34a targets that

are not affected at the level of mRNA abundance, successful identifications of miRNA targets by this method have been accomplished by many groups [51, 58]. The dataset GDS2755 from NCBI GEO [8] shows the gene expression data with overexpression of mir-34a reported in [21] . We capture Caspase 8' profile into Table 4.9. After statistical analysis, Caspase 8 is not included in the differentially expressed gene list provided in the supplementary material of [21]. The comparable levels of Caspase 8 indicate the irrelevance to mir-34a overexpression. Therefore, it increases the confidence to rule out the candidature of Caspase 8.

Next, let's see the application on BCL2 which is an experimentally confirmed target. Hence, our prediction should definitely agree with the existence of repression by mir-34a. As reported in [78], mir-34a is a necessary condition for apoptosis. When we set all the rest sufficient conditions satisfied, mir-34a becomes a necessary and sufficient condition to let p53-dependent intrinsic pathway occur. The above formula $AF($ CASP3 $= 1)$ still holds. Under this setting, we test the repression on BCL2. The prototype model falsifies the formula and the miRNA model validates the formula. Therefore, BCL2 is also confirmed to be the target of mir-34a by our approach.

Because of the intensive studies of classical p53-dependent apoptosis pathways, the evidence about extrinsic and intrinsic pathways is abundantly found in the literature. The direct success of applications is all from these categories. As seen in Figure 4.4, most of the rest candidates reside in survival pathways. The correlations with p53 or ectopic expressed mir-34a are rarely reported. Therefore, we have to consider the design scheme to assist the validation task.

4.4.5 Design Schemes

The effective selection of evidence plays an important role to distinguish two types of models and meanwhile the lack of evidence makes the trouble to conduct the verification. Therefore, the design task is mainly about the specification. Based on the designed specification, the experimental setup is thus prepared accordingly. For illustration pur-

pose, the four generic structures in Table 4.2 cover all the possible cases in the graphs. In reality, the situations are less complicated and share several common features of the structure, such as locality, line type, sequence pattern, etc.

First of all, in this design work, initial condition must be the presence of p53 activation or ectopic expression of mir-34a. The design guideline is as follows. If the candidate is of line type, the specification could be designed to concern the impact on its downstream neighbours in occurrence or sequence pattern. For example, to test on IKK, NF-κB could be selected as evaluation point. $AF($ NF-κ B $= 1)$ is selected. The observation in the experiment is on NFκB and to check whether its level will be expressed or depressed. Similar strategies could be also applied to IL1RAP, MYD88, PI3CD, etc. If the candidate is of collaborative type, the specification could be designed to concern the impact on its collaborator or its repressed target in consequence pattern. For example, XIAP deactivates Caspase 3 in a collaborative type. Caspase 3 could be selected as evaluation point. $AG(\text{CASP3} = 1 \rightarrow EF(\text{CASP3} = 0))$ is about the consistent expression level of Caspase 3. If XIAP is not repressed by mir-34a, Caspase 3 will not be consistently high level due to the blockage by XIAP.

Among the targets listed in Table 4.7, some candidates share the similar functions or positions so as to be clustered as one type of component, such as BCL2 with BCL2L1, PRKACB with PRKX. They cannot be distinguished in this macroscopic model in terms of biological functions. Moreover, Caspase 7 is an equivalent executioner caspase of Caspase 3. DFFA is the downstream of Caspase 3. Both are not involved in this model. Experiment design scheme should be subject to the biological methods.

4.5   Conclusion

In this chapter, we have developed a systematic approach to filter the true targets of miRNA from the enormous predictions made by the bioinformatics tools. We took full advantage of model checking technique. Rather than confining to the model validation issue, we further applied this technique to identify the target specificity by comparing the

satisfaction results of real evidence between prototype model and miRNA model. The shortlisted candidates could be provided to the biologists to conduct further verification experiments. Thus, the biologists may focus more attentions on the highly possible ones without wasting time on the targets with low possibilities. Meanwhile, this method could handle a moderate number of identifications at the same time. Therefore, it potentially provides a efficient and labour-saving facility for the miRNA study in the tumorigenesis.

When the application is impeded by the lack of real evidence, the generalized criteria for specification selection are helpful to design the experiment to generate the desired specifications. The drawback of these criteria is the difficulty for a non-expert to interpret the temporal language and the assistance from the modeller is compulsory as well. The proposed schemes in Table 4.7 are obtained under certain initial conditions which let specific pathway enabled. Therefore, the collaborations becomes necessary in the practical implementation.

Table 4.7: 15 candidates list

Candidate	CTL	Evid.	Prototype model	miRNA model	Conclusion /design pattern	Description
CASP8	$AF(\,$CASP3$=1)$	T	T	F	False target	extrinsic apoptosis occurs [12]
BCL2 /BCL2L1	$AF(\,$CASP3$=1)$	T	F	T	True target	intrinsic apoptosis occurs [78]
APAF1	$AF(\,$CASP3$=1)$	T	T	F	False target	intrinsic apoptosis occurs [90]
MYD88 /IRAK2	$EF(\,$TAK1$=1)$	n.a	T	F	occurrence pattern	TAK1 could be activated
XIAP	$AG(\,$CASP3$=1$ $\rightarrow EF($CASP3$=0))$	n.a	T	F	consequence pattern	CASP3's level will decrease after its activation
IL1RAP	$EF(\,$ILC$=1)$	n.a.	T	F	occurrence pattern	ILC could be formed
IKBKB	$EG(\,$NFκB$=0)$	n.a.	F	T	invariant pattern	NFκB could never be active or increase
PI3CD	$AF(\,$AktPKB$=1)$	n.a.	T	F	occurrence pattern	AktPKB could be phosphorylated
PRKACB /PRKX	$AF(\,$AktPKB$=1)$	n.a.	T	F	occurrence pattern	AktPKB could be phosphorylated
AKT2	$AG(\,$AktPKB$=1$ $\rightarrow AG($AktPKB$=1))$	n.a.	T	F	consequence pattern	phosphorylated AktPKB level will not influenced
CASP7					biological methods	CASP7 is equivalent to CASP3
DFFA					biological methods	DFFA is downstream of CASP3/7

"T" for true, "F" for false, "n.a." for not applicable because of lack of evidence, "Evid." for evidence, conditions on supportive components should be fulfilled accordingly.
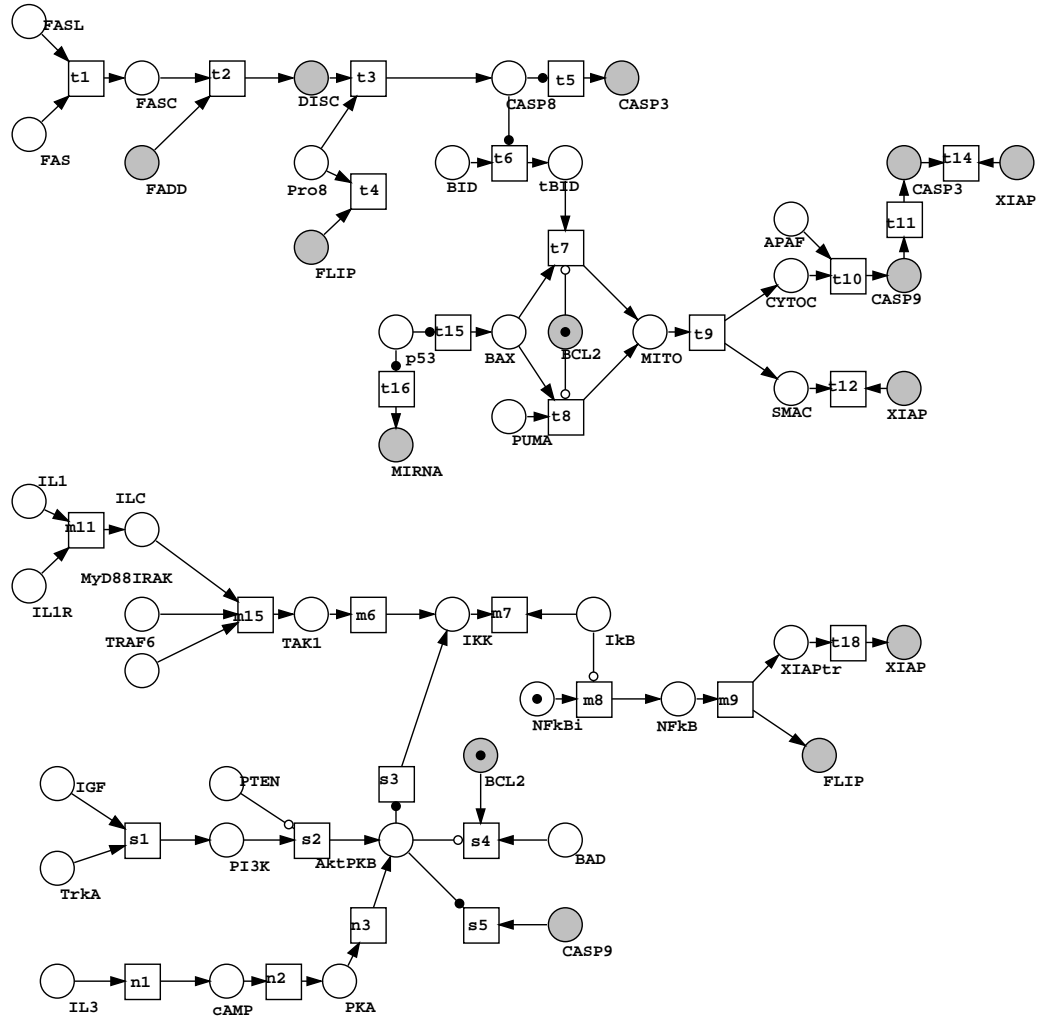
Figure 4.4: Prototype model of apoptosis pathways. Grey places with the same name represent the identical component. The pathways are intertwined by the grey components. The initial marking distributions may vary according to the investigated conditions. The marking for each place is 1 which means the activation or the expression level increase of that component, or 0 which means the deactivation or the expression level decrease of that component. Transitions represent typical biochemical reactions, such as activation, repression, phosphrylation, catalysis, combination, separation, etc. The detailed description about places and transitions could refer to Appendices A and B.

Table 4.8: Approved specifications by the prototype model

Condition	CTL	Description
FAS ligation	$AF($ CASP3 $= 1)$	Extrinsic apoptosis
	$AF($ CASP8 $= 1)$	pathway is enabled
p53 and PUMA active	$AF($ CASP3 $= 1)$	Intrinsic apoptosis
	$AF($ CASP9 $= 1)$	pathway is enabled
IL1 active and intrisic	$EG($ CASP3 $= 0)$	Apoptosis can be disabled
or extrinsic open	$EG($ CASP3 $= 1 \rightarrow$	by survival factor IL1
	$EF($ CASP3 $= 0))$	
IGF active and intrisic	$EG($ CASP3 $= 0)$	Apoptosis can be disabled
or extrinsic open		by growth factor IGF

Conditions on supportive components should be fulfilled accordingly.

Table 4.9: MicroArray dataset GDS2755

ID_ REF	Identi- fier	GSM 187633	GSM 187634	GSM 187631	GSM 187632
1553306_at	CASP8	8.7	6.1	44.9	39.6
207686_s_at	CASP8	174.3	143.2	115.6	107.7

GSM187633: HCT116 cells with empty vector, technical replicate 1
GSM187634: HCT116 cells with empty vector, technical replicate 2
GSM187631: HCT116 cells with enforced miR-34a expression, technical replicate 1
GSM187632: HCT116 cells with enforced miR-34a expression, technical replicate 2

Chapter 5

Conclusion

5.1 Contributions

Systems biology is an emerging area that attracts more and more attentions from various disciplines. It originates from the problems from biologists and finally feedback the solutions to the biology society. During the intermediate steps, the interactive collaborations happen frequently. Besides engineering approaches, the tools from computer science, bioinformatics, statistics are essential as well. Prior to systems biology, bioinformatics has been proven very helpful to the biologists. The emergence of tools in data mining, sequence analysis and protein structure modelling helps the biologists out of the daily grind. However, nowadays we would like to learn more about the global behaviours. The existing methods seem incompetent to accomplish this task. To break the bottleneck, systems biology is a potential candidate. To substantiate this idea, in this thesis, we proposed our solutions for microRNA target validation problem. During the exploration, we reach our final goal step by step.

In Chapter 2, p53's oscillatory behaviour was explained by one model in ordinary differential equations. The relations among the components obeyed the basic biochemical laws together with some reasonable approximations. The model was good enough to reproduce the phenomena observed in the experiment. Furthermore, two new predic-

tions were made from the mathematical model, which deserved a thorough investigation of this well-established protein. One prediction was that a high level of stimulation induced a dramatic change of the stability—from oscillation to steady state. The other one was the frequency drift while changing the stimulation levels. No literature and simulation work has reported these two phenomena so far. The verification experiment could provide further evidence to mature the model.

In Chapter 3, we put p53 into a bigger context—apoptosis pathways. Identifying a large number of parameters causes a serious trouble for modelling. Therefore, the formalism was changed to be of discrete type, which thus skipped the hurdle of the parameter identification for large-scale system. Under the discrete framework, the qualitative properties were emphasized, such as the existence of links, reaction types, the expression occurrences, whereas the reaction rates, accurate expression levels were neglected. Petri net was employed to describe the relations. To make a full use of the model, it has to run the validation test. To address this issue, two methods were proposed, i.e. invariant analysis and reachability analysis. Both methods were based on the examination of the solutions of state equations. The illustrative examples showed the convenience of these methods. The validation of p53-dependent apoptosis model paved the way for next-step exploration.

In Chapter 4, the targets of microRNA were identified in the context of apoptosis pathway. The bioinformatics tools provide a great number of candidates by sequence matching, which makes the task of manual verification impossible. Our target is to sift the highly possible targets from the predication pool. Model checking based approach was developed. The results were obtained by comparisons among the different models and the real evidence. The specifications were intentionally designed to effectively discriminate the models. Moreover, when the desired evidence was lacking, we proposed the experimental scheme by designing the specifications. In the case study, mir-34a was taken as one example to illustrate this approach. Some results were verified by the available findings. Some candidates were yet decided due to the lack of evidence and

instead, the design schemes were provided for assistance. This chapter exemplified the characteristics of interdisciplinary research which resorted to the various resources from computer science, bioinformatics, biotechnology, systems engineering.

5.2 Future Work

Based on the considered issues in this thesis, some further work could be done to extend the potential applications.

- It becomes more encouraging to draw attention from biologists to verify the predictions in our work. The hypotheses can be proven only by the support from practical realization. The wet lab experiments will finally complete the loop and provide the further knowledge to strengthen the models and methods. For example, the desired experiments could be conducted to generate the specifications for microRNA target validation. The p53 sampling guideline could facilitate the data generation to polish our model. After several rounds, the models and methods could become reliable to be fully utilized. In the meantime, the biologists will also benefit from this.

- Regarding the continuous modelling, one possible direction is to build the model to unify the conflicts in biological discoveries. In biological systems, there exist several proteins or genes playing contradictory roles. In certain environment, it performs conversely compared to other environment, even if its target gene or binding site of DNA keep consistent. One possible explanation could be thought as in different environments, different mechanisms of cell cycle or cell death are triggered to decide the cell fates. Bifurcation analysis gives the direct clue for this problem.

- Different mechanisms may generate the same phenomenon. For example, in p53-Mdm2 modelling, each group has their own model. Which one is the closest to the reality that is happening in cells? How could we differentiate these models?

Hence, it is interesting to extract the characteristics of each type of modelling, through which we may design the experiment to narrow down the differentiation conditions. Consequently, we may introduce the modelling structure to other biological systems which share the similar mechanisms.

- The strategy discussed for mir34a target validation in apoptosis pathways could be promoted to other perspectives in terms of pathways and miRNAs. For example, cell cycle pathway is of bigger size than apoptosis and more predicted candidates are involved. Many cellular malfunctions are also related to the components in cell cycle [17]. Other important miRNAs, such as mir-15/16, mir-127, etc., could also be investigated with high interest [101]. Regarding the technical part, the more quantitative model could be introduced to describe richer behaviours than the qualitative one.

- Moreover, the most promising thing will be the development of a fully automated checking program, which are fed by the information of connections and experimental evidence. The program can acquire the primary prediction results from the web server, and then return the false and true targets for the chosen miRNA. The approach developed may finally evolve into a formal method to comprehensively and rapidly identify targets genes for microRNAs. In addition, the successful accomplishment should pave the way to help the biologists to design new therapeutic strategies for cancer.

APPENDIX

Appendix A

Reaction Rules

All the transitions are defined in NuSMV script following the rules defined below. When microRNA is included, the reaction is modified according to the repressive position. The whole set of reactions shown in Figure 4.4 are listed in Table A.1. The descriptions of biological functions are provided as well.

```
reaction#1#2[r#][i#](reactant variable list, product variable list,
catalyst variable list, inhibitor variable list)
--#1: number of reactants
--#2: number of products
--r#: catalyst participation and number of catalyst, by default 1
--i#: inhibitor participation and number of inhibitor, by default 1
```

Table A.1: Reaction List

#	Reaction type	Biological function
t1	reaction21(FASL, FAS, FASC)	combination
t2	reaction21(FASC, FADD, DISC)	combination

#	Reaction type	Biological function
t3	reaction21(DISC, Pro8, CASP8)	activation
t4	reaction20(Pro8, FLIP)	Pro8 is consumed by FLIP
t5	reaction01r(CASP3, CASP8)	generating CASP3 catalysed by CASP8
t6	reaction11r(BID, tBID, CASP8)	generating tBID catalysed by CASP9
t7	reaction21i(tBID, BAX, MITO, BCL2)	stimulating Mitochondria, inhibited by BCL2
t8	reaction21i(BAX, PUMA, MITO, BCL2)	stimulating Mitochondria, inhibited by BCL2
t9	reaction12(MITO, CYTOC, SMAC)	Mitochondria release CytoC and SMAC
t10	reaction21(CYTOC, APAF, CASP9)	activation
t11	reaction11(CASP9, CASP3)	activating CASP3 by CASP9
t12	reaction20(XIAP, SMAC)	SMAC is consumed by XIAP
t14	reaction20(XIAP, CASP3)	CASP3 is consumed by XIAP
t15	reaction01r(BAX, p53)	p53 transcriptionally activate BAX
t16	reaction01r(MIRNA, p53)	p53 transcriptionally activate microRNA
t18	reaction11(XIAPtr, XIAP)	translation of XIAP
m6	reaction11(TAK1, IKK)	TAK1 phosphorylate IKK to active
m7	reaction20(IKK, IkB)	IKK consume IkB

#	Reaction type	Biological function
m8	reaction11i(NFkBi, NFkB, IkB)	generating NFkB, inhibited by IkB
m9	reaction12(NFkB, XIAPtr, FLIP)	NFkB transcriptionally activate XIAPtr and FLIP
m11	reaction21(IL1, IL1R, ILC)	combination
m15	reaction31(MyD88IRAK, TRAF6, ILC, TAK1)	combination
s1	reaction21(IGF, TrkA, PI3K)	activation of PI3K
s2	reaction11i(PI3K, AktPKB, PTEN)	PI3K phosphorylate AktPKB to active, inhibited by PTEN
s3	reaction01r(IKK, AktPKB)	IKK is phosphorylted by AktPKB
s4	reaction20i(BAD, BCL2, AktPKB)	AktPKB disable the consumption of BCL2 by BAD by phosphorylation
s5	reaction10r(CASP9, AktPKB)	AktPKB disable CASP9 by phosphorylation
n1	reaction11(IL3, cAMP)	cAMP is activated by IL3
n2	reaction11(cAMP,PKA)	PKA is activated by cAMP
n3	reaction11(PKA, AktPKB)	PKA phosphorylate AktPKB to active

Appendix B

Gene Names in Apoptosis Model

All the genes are from the genome of *Homo sapiens* (human). For certain genes, one representative of the homologs is shown. Regarding the places which correspond to the organelle and complex, the gene information is not applicable. All the gene information is retrieved from NCBI Entrez Gene [69].

Table B.1: Gene Name List

Place Name	Official Symbol	Gene Name	Gene ID
MIRNA	MIR34A	microRNA 34a	407040
FASL	FASLG	Fas ligand (TNF superfamily, member 6)	356
FAS	FAS	FAS(TNF receptor, member 6)	355
FASC	n.a	FASL-FAS Complex	n.a
FADD	FADD	Fas (TNFRSF6) -associated via death domain	8772
DISC	n.a.	Death-inducing signaling complex	n.a
Pro8	n.a	procaspase-8	n.a.
FLIP	CFLAR	CASP8 and FADD-like apoptosis regulator	8837

Place Name	Official Symbol	Gene Name	Gene ID
CASP8	CASP8	caspase 8, apoptosis-related cysteine peptidase	841
CASP3	CASP3	caspase 3, apoptosis-related cysteine peptidase	836
BID	BID	BH3 interacting domain death agonist	637
tBID	n.a	truncated BID	n.a
p53	TP53	tumour protein p53	7157
BAX	BAX	BCL2-associated X protein	581
PUMA	BBC3	BCL2 binding component 3	27113
BCL2	BCL2	B-cell CLL/lymphoma 2	596
MITO	n.a	mitochondria	n.a
CYTOC	CYCS	cytochrome c, somatic	54205
SMAC	DIABLO	diablo homolog (Drosophila)	56616
APAF	APAF1	apoptotic peptidase activating factor 1	317
CASP9	CASP9	caspase 9, apoptosis-related cysteine peptidase	842
CASP3	CASP3	caspase 3, apoptosis-related cysteine peptidase	836
XIAP	XIAP	X-linked inhibitor of apoptosis	331
IL1	IL1	interleukin 1	3551
IL1R	IL1R1	interleukin 1 receptor, type I	3554
ILC	n.a	IL1-IL1R complex	n.a

Table B.1: (Continued)

Place Name	Official Symbol	Gene Name	Gene ID
MyD88(IRAK)	MYD88	myeloid differentiation primary response gene (88)	4615
(MyD88)IRAK	IRAK1	interleukin-1 receptor-associated kinase 1	3654
TRAF6	TRAF6	TNF receptor-associated factor 6	7189
TAK1	MAP3K7	mitogen-activated protein kinase kinase kinase 7	6885
IKK	CHUK	conserved helix-loop-helix ubiquitous kinase	1147
IkB	NFKBIA	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha	4792
NFkBi	n.a	NFκB initiation	n.a
NFkB	NFKB1	nuclear factor of kappa light polypeptide gene enhancer in B-cells 1	4790
XIAPtr	n.a	XIAP transcript	n.a.
IGF	IGF1R	insulin-like growth factor 1 receptor	3480
TrkA	NTRK1	neurotrophic tyrosine kinase, receptor, type 1	4914
PI3K	PIK3CA	phosphoinositide-3-kinase, catalytic, alpha polypeptide	5290
PTEN	PTEN	phosphatase and tensin homolog	5728
AktPKB	AKT1	v-akt murine thymoma viral oncogene homolog 1	207

Table B.1: (Continued)

Place Name	Official Symbol	Gene Name	Gene ID
BAD	BAD	BCL2-associated agonist of cell death	572
IL3	IL3	interleukin 3 (colony-stimulating factor, multiple)	3562
cAMP	CAMP	cathelicidin antimicrobial peptide	820
PKA	PRKACB	protein kinase, cAMP-dependent, catalytic, beta	5567

Appendix C

Molecular Biological Background

As shown in the Figure 1.1, the starting point where we conduct our research is the demands from biology society. Moreover, our modelling and verification must be tested by the data from the biological experiments as well. Therefore, the understanding of the basic biological concepts and the underlying principles become essential. As long as we know their language, we can communicate well with the biologists and elucidate our results in a more meaningful and practical way.

The interpretation of the biological data plays the same important role to systems biology research. No matter where the data comes from, either from direct feed by collaborators or the reports in the literature, the collected data is directly translated for modelling and analysis. Therefore, it is another essential task to get familiar with the experimental methods and the data representations.

In this appendix, the elementary knowledge about biological systems is included in the first half and then the brief survey about some most common experimental tools is given. Those materials provide the background information for the work we undertake.

C.1 Elementary Molecular Biology

There exist various kinds of living things in the world. For example, animals, fungi, bacteria. There are also non-living matter, such as water, rock, etc. [2] How to characterize living things and distinguish them from the non-living matter?

Cell All living things are made of cells: small, membrane-enclosed units filled with a concentrated aqueous solution of chemicals and endowed with the extraordinary ability to create copies of themselves by growing and dividing in two. Typically, there are two kinds of cells, say eucaryotic cell and procaryotic cell. The critical difference is the ownership of nucleus or not. For `eucaryote`, living organism are composed of one or more cells with a distinct nucleus and cytoplasm, including all forms of life except viruses and bacteria. Contrarily, `procaryote` is a major category of living cells distinguished by the absence of a nucleus.
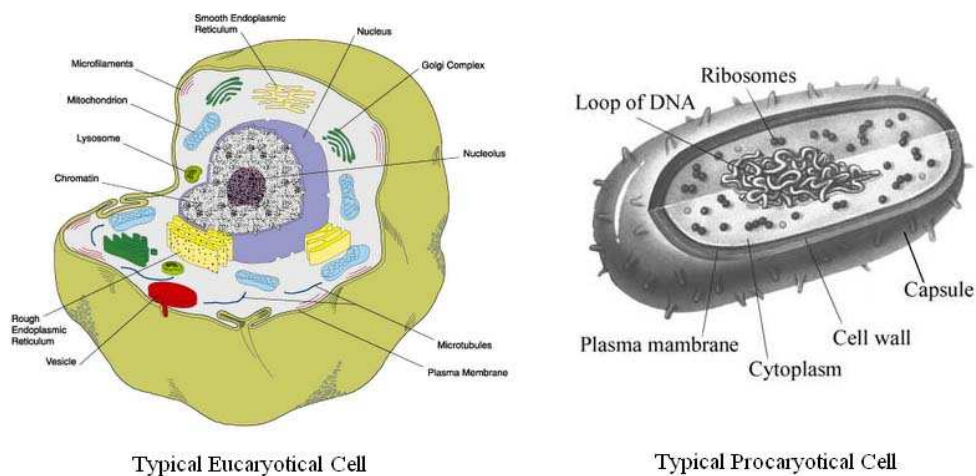

Figure C.1: Cell contents [2]

Besides nucleus, some other organelles are also enclosed in the cell, referring to Figure C.1. For example, mitochondria, chloroplast, ribosomes, etc. They all perform the functional roles in corresponding sub-environment. Nucleus is the information store of the cell. Mitochondria generates energy from food to power the cell. Ribosome helps

in the translation process of protein. Chloroplasts, which are found only in plants and algae, capture energy from sunlight. As for the structure difference between plants and animals, eucaryotes and procaryotes, it will not be mentioned much in this part. What concern us most is the activities happened in the nucleus of eucaryotic cells.

Nucleus The nucleus is usually the most prominent organelle in a eucaryotic cell. It is enclosed within two concentric membranes that form the nuclear envelope, and it contains molecules of DNA—extremely long polymers—that encode the genetic information of the organism. DNA also stores the genetic information in procaryotic cells; these cells lack a distinct nucleus not because they lack DNA, but because they do not keep it inside a nuclear envelope, segregated from the rest of the cell contents.

DNA The development of molecular biology is synchronized with the DNA's recognition and exploration. DNA is a store of genetic codes controlling and regulating the synthesis and functioning of proteins and other molecules. The term "DNA" is short for deoxyribonucleic acid. It is a macromolecule which is composed of large number of nucleotides, and is packaged into individual chromosomes. The most attractive feature is the double helix structure of DNA, which was determined by James Watson and Francis Crick in 1953.

> "Today, the fact that DNA is the genetic material is so fundamental to biological thought that is difficult to appreciate what an enormous intellectual gap this discovery filled." by Bruce Alberts, et al. [2]

Nucleotide has three parts: a nitrogen-containing base; a pentose(a sugar molecule containing 5 carbon atoms); one or more phosphate groups. According to the different bases of nucleotides, we will have two categories and totally five members, i.e. pyrimidines have uracil(U), cytosine(C), thymine(T); while purines have adenine(A) and guanine(G). The different combinations of bases represent different genetic information.

RNA(ribonucleic acid) is a similar molecule with DNA. The difference resides in the following three aspects. The sugar contained in RNA's nucleotide is ribose, not the deoxyed form. And what's more, RNA contains the bases A, C, G, U, while DNA contains A, C, G, T. Finally, RNA is in a single-stranded form, and DNA typically exists in the double-stranded form—two single strands of DNA—running anti-parallel and joined by base-base hydrogen bonds: A with T, G with C.

Protein   Proteins consist of one or more polypeptides. A polypeptide consist of a covalently linked long chain of amino acids. There are found totally 20 kinds of amino acids in nature. The order of amino acid constituting polypeptide is coded by genetic information stored in the DNA molecules. From DNA's base combinations to protein's amino acid sequences, thus enabling the cell to perform various functions, it is a complicated process, which is known as "Central Dogma of Molecular Biology".

Central Dogma of Molecular Biology   This dogma implies the process from DNA to RNA to protein. It contains mainly three major stages (shown in Figure C.2), which are all accomplished with the assistance of enzymes and co-factor proteins:

1. Replication of DNA

2. Transcription of RNA and processing to yield mRNA which migrate to the cytoplasm

3. Translation by ribosomes of the code carried by mRNA into proteins

Gene expression   Gene is a unit of heredity composed of DNA. It occupies a specific position on a chromosome. It concerns with a specific function, such as synthesis of a protein or an mRNA. Besides the coding sequences, regulatory gene sequences(known as operons) adjacent to the gene contains information about when and in which cell type the gene is to be expressed. Therefore, in process of expression, many enzymes
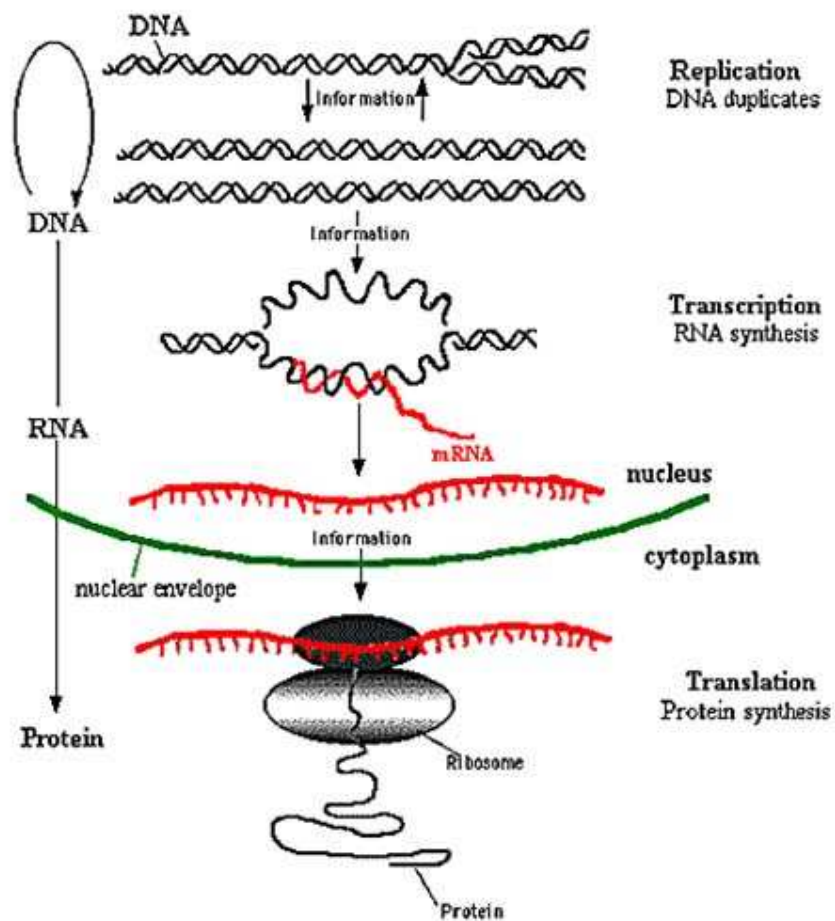
Figure C.2: Central Dogma of Molecular Biology [2]

and RNAs are involved to control the replication, transcription and translation, finally, determine the gene expression levels.

Pathways Biological systems consist of respiratory system, circulatory system, immune system, reproductive system, etc. The pathways underlying in these systems can be categorized into three-fold: metabolic pathways, signalling transduction pathways and gene regulatory networks. Gene regulatory network has been introduced before. It concerns about the positive or negative control and regulation of gene expression level by the components included. Metabolic pathways study the energy acquirement or loss and the transfer of energy among biochemical reactions. Signalling transduction pathways include the intra- and inter- cellular interactions among molecules, proteins, organelles and so on. The external stimuli are also necessary to be considered. All these networks must be modelled appropriately and then employed to explore more.

C.2   Experimental Methods

In this part, the one-sentence version of explanations could facilitate the understanding of the aims of frequently-encountered methods, and the principles behind those methods are omitted. Because the current measuring technology is not able to reflect the real accurate levels in the molecular biology laboratories. In most cases, the data at molecular scale is more qualitative. For example, in northern blot result, the levels are indicated by the grey density through which only high and low indices are obtained. PCR techniques are used to amplify the DNA samples; blotting methods are used to detect and measure the expressed level of DNA sequence, RNA and protein. DNA microarray is a high-throughput measurement of genome-scale RNA expression. Flow cytometry and luciferase are for a relative macroscopic purpose. The following information is collected from Wikipedia [104] key words search.

PCR In molecular biology, the polymerase chain reaction (PCR) is a technique to amplify a single or few copies of a piece of DNA across several orders of magnitude, generating thousands to millions of copies of a particular DNA sequence.

RT-PCR Reverse transcription polymerase chain reaction (RT-PCR) is a variant of polymerase chain reaction (PCR), a laboratory technique commonly used in molecular biology to generate many copies of a DNA sequence, a process termed "amplification". Reverse transcription PCR is not to be confused with real-time polymerase chain reaction (Q-PCR/qRT-PCR), which is also sometimes (incorrectly) abbreviated as RT-PCR.

Real-time PCR In molecular biology, real-time polymerase chain reaction, also called quantitative real time polymerase chain reaction (Q-PCR/qPCR) or kinetic polymerase chain reaction, is a laboratory technique based on the PCR, which is used to amplify and simultaneously quantify a targeted DNA molecule. It enables both detection and quantification (as absolute number of copies or relative amount when normalized to DNA input or additional normalizing genes) of one or more specific sequences in a DNA sample.

Southern Blot A southern blot is a method routinely used in molecular biology for detection of a specific DNA sequence in DNA samples.

Northern Blot The northern blot is a technique used in molecular biology research to study gene expression by detection of RNA (or isolated mRNA) in a sample.

Western Blot The western blot (alternatively, protein immunoblot) is an analytical technique used to detect specific proteins in a given sample of tissue homogenate or extract.

DNA Microarray DNA Microarray can be used to detect DNA (as in comparative genomic hybridization), or detect RNA (most commonly as cDNA after reverse tran-

scription) that may or may not be translated into proteins. The process of measuring gene expression via cDNA is called expression analysis or expression profiling.

Flow cytometry   Flow cytometry is a technique for counting and examining microscopic particles, such as cells and chromosomes, by suspending them in a stream of fluid and passing them by an electronic detection apparatus. It allows simultaneous multiparametric analysis of the physical and/or chemical characteristics of up to thousands of particles per second. Flow cytometry is routinely used in the diagnosis of health disorders, especially blood cancers, but has many other applications in both research and clinical practice.

Luciferase   Luciferase is a generic term for the class of oxidative enzymes used in bioluminescence and is distinct from a photoprotein. One famous example is the firefly luciferase (EC 1.13.12.7) from the firefly *Photinus pyralis.*

Bibliography

[1] B. D. Aguda, Y. Kim, M. G. Piper-Hunter, A. Friedman, and C. B. Marsh. Microrna regulation of a cancer network: consequences of the feedback loops involving mir-17-92, e2f, and myc. *Proceedings of the National Academy of Sciences*, 105(50):19678, 2008.

[2] B. Alberts, A. Johnson, A. Lewis, M. Raff, K. Roberts, and P. Walter. *The Molecular Biology of the Cell*. Garland Science, 4 edition, 2002.

[3] M. Antoniotti, A. Policriti, N. Ugel, and B. Mishra. Model building and model checking for biochemical processes. *Cell Biochemistry and Biophysics*, 38:271–286, 2003.

[4] E. Balsa-Canto, A.A. Alonso, and J.R. Banga. Computational procedures for optimal experimental design in biological systems. *IET Systems Biology*, 2(4):163–172, 2008.

[5] S. Banin, L. Moyal, S.Y. Shieh, Y. Taya, C.W. Anderson, L. Chessa, N.I. Smorodinsky, C. Prives, Y. Reiss, Y. Shiloh, et al. Enhanced phosphorylation of p53 by atm in response to dna damage. *Science*, 281(5383):1674–1677, 1998.

[6] A. Bar-Even, J. Paulsson, N. Maheshri, M. Carmi, E. O'Shea, Y. Pilpel, and N. Barkai. Noise in protein expression scales with natural protein abundance. *Nature genetics*, 38(6):636–643, 2006.

[7] R.L. Bar-Or, R. Maya, L.A. Segel, U. Alon, Levine A.J., and M. Oren. Generation of oscillations by the p53-mdm2 feedback loop: A theoretical and experimental study. *Proceedings of the National Academy of Sciences*, 97(21):11250, 2000.

[8] T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I.F. Kim, A. Soboleva, M. Tomashevsky, and K.A. Marshall. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic acids research*, 37(Database issue):D885, 2009.

[9] D.P. Bartel. Micrornas: Genomics, biogenesis, mechanism, and function. *Cell*, 116:281–297, 2004.

[10] G. Batt, D. Ropers, H. De Jong, J. Geiselmann, R. Mateescu, M. Page, and D. Schneider. Validation of qualitative models of genetic regulatory networks by model checking: Analysis of the nutritional stress response in Escherichia coli. *Bioinformatics*, 21(Suppl 1):i19, 2005.

[11] S. Benchimol. p53-dependent pathways of apoptosis. *Cell Death and Differentiation*, 8:1049–1051, 2001.

[12] M. Bennett, K. Macdonald, S.W. Chan, J.P. Luzio, R. Simari, and P. Weissberg. Cell surface trafficking of Fas: a rapid mechanism of p53-mediated apoptosis. *Science*, 282(5387):290, 1998.

[13] G.T. Bommer, I. Gerin, Y. Feng, A.J. Kaczorowski, R. Kuick, R.E. Love, Y. Zhai, T.J. Giordano, Z.S. Qin, and B.B. Moore. p53-mediated activation of mirna34 candidate tumor-suppressor genes. *Current Biology*, 17(15):1298–1307, 2007.

[14] S. Bottani and B. Grammaticos. Analysis of a minimal model for p53 oscillations. *Journal of theoretical biology*, 249(2):235–245, 2007.

[15] J.C. Bourdon, V.D. Laurenzi, G. Melino, and D. Lane. p53: 25 years of research and more questions to answer. *Cell Death and Differentiation*, 10(4):397–399, 2003.

[16] M. Calder, V. Vyshemirsky, D. Gilbert, and R. Orton. Analysis of signalling pathways using the prism model checker. *Proceedings of Computational Methods in Systems Biology*, 2005.

[17] M. Carleton, M.A. Cleary, and P.S. Linsley. MicroRNAs and cell cycle regulation. *Cell cycle*, 6(17):2127, 2007.

[18] P. Carole and G. Douglas. Explaining oscillations and variability in the p53-mdm2 system. *BMC Systems Biology*, 2, 2008.

[19] N. Chabrier and F. Fages. Symbolic model checking of biochemical networks. *Proceedings of Computational Methods in Systems Biology*, 2003.

[20] N. Chabrier-Rivier, M. Chiaverini, V. Danos, F. Fages, and V. Schachter. Modeling and querying biomolecular interaction networks. *Theoretical Computer Science*, 325(1):25–44, 2004.

[21] T.C. Chang, E.A. Wentzel, O.A. Kent, K. Ramachandran, M. Mullendore, K.H. Lee, G. Feldmann, M. Yamakuchi, M. Ferlito, and C.J. Lowenstein. Transactivation of mir-34a by p53 broadly influences gene expression and promotes apoptosis. *Molecular Cell*, 26(5):745–752, June 2007.

[22] C. Chaouiya. Petri net modelling of biological networks. *Briefings in Bioinformatics*, 8(4):210, 2007.

[23] A. Ciliberto, B. Novak, and J.J. Tyson. Steady states and oscillations in the p53/mdm2 network. *Cell Cycle*, 4(3):488–493, 2005.

[24] A. Cimatti, E. Clarke, E. Giunchiglia, F. Giunchiglia, M. Pistore, M. Roveri, R. Sebastiani, and A. Tacchella. Nusmv 2: An opensource tool for symbolic model checking. In *Computer Aided Verification*, pages 241–268. Springer, 2002.

[25] E.M. Clarke, J.O. Grumberg, and D.A. Peled. *Model Checking*. MIT Press, 1999.

[26] BTU Cottbus. Charlie—a tool for the analysis of place/transition nets. `http://www-dssz.informatik.tu-cottbus.de/software/charlie/charlie.html`.

[27] W.S. El-Deiry. Regulation of p53 downstream genes. *Seminars in Cancer Biology*, 8:345–357, 1998.

[28] S. Elmore. Apoptosis: a review of programmed cell death. *Toxicologic pathology*, 35(4):495, 2007.

[29] K. Engelborghs, T. Luzyanina, and G. Samaey. Dde-biftool v.2.00: a matlab package for bifurcation analysis of delay differential equations. Technical report, Department of Computer Science, K.U.Leuven, Leuven, Belgium, 2001.

[30] M. D. Esposti. The roles of bid. *Apoptosis*, 7:433–440, 2002.

[31] J. Fisher and T.A. Henzinger. Executable cell biology. *Nature Biotechnology*, 25(11):1239–1249, 2007.

[32] J. Fisher, N. Piterman, A. Hajnal, and T. Henzinger. Predictive modeling of signaling crosstalk during c. elegans vulval development. *PLoS Computational Biology*, 3:e92, 05 2007.

[33] N. Geva-Zatorsky, N. Rosenfeld, S. Itzkovitz, R. Milo, A. Sigal, E. Dekel, T. Yarnitzky, Y. Liron, P. Polak, G. Lahav, et al. Oscillations and variability in the p53 system. *Molecular Systems Biology*, 2(1), 2006.

[34] G.J. Hannon. RNA interference. *Nature*, 418(6894):244–251, 2002.

[35] S.L. Harris and A.J. Levine. The p53 pathway: positive and negative feedback loops. *Oncogene*, 24(17):2899–2908, 2005.

[36] L. He, X. He, et al. A microrna component of the p53 tumour suppressor network. *Nature*, 447(7148):1130–1134, June 2007.

[37] L. He, X. He, S.W. Lowe, and G.J. Hannon. micrornas join the p53 network - another piece in the tumour-suppression puzzle. *Nat Rev Cancer*, 7(11):819–822, November 2007.

[38] M. Heiner, I. Koch, and J. Will. Model validation of biological pathways using petri nets - demonstrated for apoptosis. *BioSystems*, 75:15–28, 2004.

[39] H. Hermeking. The mir-34 family in cancer and apoptosis. *Cell Death & Differentiation*, 17(2):193–199, 2010.

[40] H.R. Horvitz. Genetic control of programmed cell death in the nematode caenorhabditis elegans. *Cancer research*, 59:1701–1706, 1999.

[41] F. Hua, M.G. Cornejo, M.H. Cardone, C.L. Stokes, and D.A. Lauffenburger. Effects of bcl-2 levels on fas signaling-induced caspase-3 activation: Molecular genetic tests of computational model predictions. *The Journal of Immunology*, 175:985–995, 2005.

[42] Q. Huang, Q. L. Deveraux, S. Maeda, H. R. Stennicke, B. D. Hammock, and J. C. Reed. Cloning and characterization of an inhibitor of apoptosis protein (iap) from bombyx mori. *Biochem. Biophys. Acta.*, 1499:191–198, 2001.

[43] A. Ichikawa and K. Hiraishi. A class of petri nets that a necessary and sufficient condition for reachability is obtainable. *Trans. Society of Instrument and Control Engineers (SICE)*, 24(6), 1988.

[44] G. A. Jones and J. M. Jones. *Elementary Number Theory*. Berlin: Springer-Verlag, 1998.

[45] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research*, 38(Database issue):D355, 2010.

[46] F. C. Kischkel, S. Hellbardt, I. Behrmann, M. Germer, M. Pawlita, P. H. Krammer, and M. E. Peter. Cytotoxicity-dependent apo-1 (fas/cd95)-associated proteins form a death-inducing signaling complex (disc) with the receptor. *EMBO J.*, 14:5579–5588, 1995.

[47] H. Kitano. Computational systems biology. *Nature*, 420(6912):206–210, 2002.

[48] H. Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662, 2002.

[49] G.A. Korn and T.M. Korn. *Mathematical handbook for scientists and engineers*. McGraw-Hill, New York, 1968.

[50] A. Krek, D. Grun, M.N. Poy, R. Wolf, L. Rosenberg, E.J. Epstein, P. MacMenamin, I. da Piedade, K.C. Gunsalus, M. Stoffel, et al. Combinatorial microRNA target predictions. *Nature genetics*, 37(5):495–500, 2005.

[51] J. Krutzfeldt, N. Rajewsky, R. Braich, K.G. Rajeev, T. Tuschl, M. Manoharan, and M. Stoffel. Silencing of micrornas in vivo with 'antagomirs'. *Nature*, 438(7068):685–689, 2005.

[52] G. Lahav, N. Rosenfeld, A. Sigal, N. Geva-Zatorsky, A.J. Levine, M.B. Elowitz, and U. Alon. Dynamics of the p53-mdm2 feedback loop in individual cells. *Nature Genetics*, 36(2):147–150, 2004.

[53] R.C. Lee, R.L. Feinbaum, and V. Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, 1993.

[54] S. Legewie, N. Bluthgen, and H. Herzel. Mathematical modeling identifies inhibitors of apoptosis as mediators of positive feedback and bistability. *PLoS Computational Biology*, 2:1061–1073, 2006.

[55] A. Letai. Growth factor withdrawal and apoptosis: the middle game. *Molecular cell*, 21(6):728–730, 2006.

[56] B.P. Lewis, I. Shih, et al. Prediction of mammalian microRNA targets. *Cell*, 115(7):787–798, 2003.

[57] C. Li, M. Nagasaki, K. Ueno, and S. Miyano. Simulation-based model checking approach to cell fate specification during caenorhabditis elegans vulval development by hybrid functional petri net with extension. *BMC Systems Biology*, 3(1):42, 2009.

[58] L.P. Lim, N.C. Lau, P. Garrett-Engele, A. Grimson, J.M. Schelter, J. Castle, D.P. Bartel, P.S. Linsley, and J.M. Johnson. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433(7027):769–773, 2005.

[59] B. Lindner. Some unsolved problems relating to noise in biological systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2009:P01008, 2009.

[60] LINDO. `http://www.lindo.com/`.

[61] X. Luo, I. Budihardjo, H. Zou, C. Slaughter, and X. Wang. Bid, a bcl2 interacting protein, mediates cytochrome c release from mitochondria in response to activation of cell surface death receptors. *Cell*, 94:481–490, 1998.

[62] L. Ma, J. Wagner, J.J. Rice, W. Hu, A.J. Levine, and G.A. Stolovitzky. A plausible model for the digital response of p53 to dna damage. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40):14266, 2005.

[63] M. Maragkakis, P. Alexiou, et al. Accurate microrna target prediction correlates with protein repression levels. *BMC Bioinformatics*, 10(1):295, 2009.

[64] K.L. Mcmillan. *Symbolic Model Checking*. Kluwer Academic Publishers, 1993.

[65] K.L. Mcmillan. A technique of state space search based on unfolding. *Formal methods in system design*, 6:45–65, 1995.

[66] N.A.M. Monk. Oscillatory expression of hes1, p53, and nf-κb driven by transcriptional time delays. *Current Biology*, 13(16):1409–1413, 2003.

[67] P.T. Monteiro, D. Ropers, R. Mateescu, A.T. Freitas, and H. De Jong. Temporal logic patterns for querying dynamic models of cellular interaction networks. *Bioinformatics*, 24(16):i227, 2008.

[68] T. Murata. Petri nets: Properties, analysis and applications. *Proceedings of the IEEE*, 77(4):541–580, 1989.

[69] NCBI. Entrez gene. `www.ncbi.nlm.nih.gov/gene`.

[70] D.E. Nelson, A.E. Ihekwaba, et al. Oscillations in nf-κb signaling control the dynamics of gene expression. *Science*, 306(5696):704, 2004.

[71] J.R.S. Newman, S. Ghaemmaghami, J. Ihmels, D.K. Breslow, M. Noble, J.L. DeRisi, and J.S. Weissman. Single-cell proteomic analysis of s. cerevisiae reveals the architecture of biological noise. *Nature*, 441(7095):840–846, 2006.

[72] M. Newman. *Integral Matrices*. Academic Press, 1972.

[73] S. Nikolov, J. Vera, V. Kotev, O. Wolkenhauer, and V. Petrov. Dynamic properties of a delayed protein cross talk model. *BioSystems*, 91(1):51–68, 2008.

[74] A.V. Oppenheim, A.S. Willsky, and S. Hamid. *Signals and systems*. Prentice-Hall, 1997.

[75] J. Paulsson. Prime movers of noisy gene expression. *Nature Genetics*, 37(9):925–926, 2005.

[76] D. Perez and E. White. Tnf-alpha signals apoptosis through a bid-dependent conformational change in bax that is inhibited by e1b 19k. *Mol. Cell*, 6:53–63, 2000.

[77] S. Ramalingam, P. Honkanen, L. Young, T. Shimura, J. Austin, P.S. Steeg, and S. Nishizuka. Quantitative assessment of the p53-mdm2 feedback loop using protein lysate microarrays. *Cancer research*, 67(13):6247, 2007.

[78] N. Raver-Shapira, E. Marciano, E. Meiri, Y. Spector, N. Rosenfeld, N. Moskovits, N. Bentwich, and M. Oren. Transcriptional activation of mir-34a contributes to p53-mediated apoptosis. *Molecular Cell*, 26(5):731–743, June 2007.

[79] V.N. Reddy, M.N. Liebman, and M.L. Mavrovouniotis. Qualitative analysis of biochemical reaction systems. *Computers in Biology and Medicine*, 26:9–14, 1996.

[80] C. Rohr, W. Marwan, and M. Heiner. Snoopy: a unifying Petri net framework to investigate biomolecular networks. *Bioinformatics*, 26(7):974, 2010.

[81] FFT routine in Matlab. `http://www.mathworks.com/access/helpdesk/help/techdoc/ref/fft.html`.

[82] M. Schaub, T. Henzinger, and J. Fisher. Qualitative networks: a symbolic approach to analyze biological signaling networks. *BMC Systems Biology*, 1(1):4, 2007.

[83] H. Schmidt and E.W. Jacobsen. Linear systems approach to analysis of complex dynamic behaviours in biochemical networks. *IEE Systems Biology*, 1(1):149–158, 2004.

[84] NUS School of Computing. Pat— an enhanced simulator, model checker and refinement checker for concurrent and real-time systems. `http://www.comp.nus.edu.sg/~pat`.

[85] A. Schrijver. *Theory of Linear and Integer Programming*. John Wiley & Sons, 1998.

[86] M. Schuler and D. R. Green. Mechanisms of p53-dependent apoptosis. *Biochemical Society Transactions*, 29:684–688, 2001.

[87] P. Sethupathy, M. Megraw, and A.G. Hatzigeorgiou. A guide through present computational approaches for the identification of mammalian microrna targets. *Nature Methods*, 3:881–886, 2006.

[88] L.F. Shampine and S. Thompson. Solving delay differential equations with dde23. *Appl. Numer. Math*, 37(4):441–458, 2001.

[89] M.S. Sheikh and A.J. Fornace. Death and decoy receptors and p53-mediated apoptosis. *Leukemia*, 14:1509—1513, 2000.

[90] Y. Shen and E. White. p53-dependent apoptosis pathways. *Advances in cancer research*, 82:55–84, 2001.

[91] B. Skaug, X. Jiang, and Z.J. Chen. The role of ubiquitin in NF-κB regulatory pathways. *Biochemistry*, 78(1):769, 2009.

[92] E.A. Slee, C. Adrain, and S.J. Martin. Executioner caspase-3, -6, and -7 perform distinct, non-redundant roles during the demolition phase of apoptosis. *J. Biol. Chem.*, 276:7320–7326, 2001.

[93] L.J. Steggles, R. Banks, O. Shaw, and A. Wipat. Qualitatively modelling and analysing genetic regulatory networks: a Petri net approach. *Bioinformatics*, 23(3):336, 2007.

[94] S.H. Strogatz. *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering.* Addison-Wesley, 2004.

[95] V. Tarasov, P. Jung, B. Verdoodt, D. Lodygin, A. Epanchintsev, A. Menssen, G. Meister, and H. Hermeking. Differential regulation of micrornas by p53 revealed by massively parallel sequencing. *Cell Cycle*, 6(13):1586–1593, 2007.

[96] C. Thompson. Apoptosis in the pathogenesis and treatment of disease. *Science*, 267:1456–1462, 1995.

[97] J.J. Tyson, K.C. Chen, and B. Novak. Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Current Opinion in Cell Biology*, 15(2):221–231, 2003.

[98] A.M. Vincent and E.L. Feldman. Control of cell survival by IGF signaling pathways. *Growth Hormone and IGF Research*, 12(4):193–197, 2002.

[99] B. Vogelstein, D. Lane, A.J. Levine, et al. Surfing the p53 network. *Nature*, 408(6810):307–310, 2000.

[100] J. Wagner, L. Ma, J.J. Rice, W. Hu, A.J. Levine, and G.A. Stolovitzky. p53-mdm2 loop controlled by a balance of its feedback strength and effective dampening using atm and delayed feedback. *IEE Systems Biology*, 152:109–118, 2005.

[101] Y. Wang and C.G.L. Lee. MicroRNA and cancer–focus on apoptosis. *Journal of cellular and molecular medicine*, 13(1):12–23, 2009.

[102] R.L. Weinberg, D.B. Veprintsev, and A.R. Fersht. Cooperative binding of tetrameric p53 to dna. *Journal of molecular biology*, 341(5):1145–1159, 2004.

[103] C. Welch, Y. Chen, and R. L. Stallings. Microrna-34a functions as a potential tumor suppressor by inducing apoptosis in neuroblastoma cells. *Oncogene*, 26(34):5017–5022, 2007.

[104] Wikipedia. The free encyclopedia. `http://en.wikipedia.org`.

[105] Wikipedia. Law of mass action — wikipedia,the free encyclopedia. `http://en.wikipedia.org/wiki/Law_of_mass_action`.

[106] Wikipedia. Smith normal form — wikipedia, the free encyclopedia. `http://en.wikipedia.org/wiki/Smith_normal_form`.

[107] O. Wolkenhauer. Defining systems biology: An engineering perspective. *IET Systems Biology*, 1:204–206, 2007.

[108] O. Wolkenhauer, B.K. Ghosh, and K.H. Cho. Control and coordination in bio-chemical networks. *IEEE Control Systems Magazine*, 24(4):30–34, 2004.

[109] O. Wolkenhauer, H. Kitano, and K.H. Cho. Systems biology. *IEEE Control Systems Magazine*, 23:38–48, 2003.

[110] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao, and T. Li. miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Research*, 37:105–110, 2009.

[111] B. Zhang, X. Pan, G.P. Cobb, and T.A. Anderson. Micrornas as oncogenes and tumor suppressors. *Developmental Biology*, 302:1–12, 2007.

Publication List

Journal Publications

1. **Yang Y.**, Xiang C., Lin H., "Identification of RNA interference target by formal methods— Application on microRNA target validation", *submitted to IET Systems Biology, under review*

2. **Yang Y.**, Lee K.S., Xiang C., Lin H., "Biological mechanisms revealed by a mathematical model for p53-Mdm2 core regulation", *IET Systems Biology*, 2009 Jul, 3(4): 229-238

International Conference Publications:

1. **Yang Y.**, Xiang C.,Lin H., "MicroRNA Target Validation by Formal Methods", *Proceedings of the 9th IEEE International Conference on Control & Automation*, Santiago, Chile, Dec 19–21, 2011

2. **Yang Y.**, Lin H., "Reachability analysis based model validation in systems biology", *Proceedings of the 4th IEEE CIS & RAM*, Singapore, Jun 28–30, 2010

3. Low W.J.Ian, **Yang Y.**, Lin H., "Validation of Petri Net Apoptosis Models Using P-Invariant Analysis", *Proceedings of the 7th IEEE International Conference on Control & Automation*, Christchurch, New Zealand, Dec 9–11, 2009

4. **Yang Y.**, Lin H., "p53-Mdm2 Core Regulation Revealed by a Mathematical Model", *Proceedings of 2008 IEEE International Conference on Systems, Man and Cybernetics*, Singapore, Oct 13–15, 2008

5. **Yang Y.**, Lin H., Xiang C.,"Feedback control for siRNA-induced gene regulatory networks", *Proceedings of International Conference on Cellular & Molecular Bioengineering*, Singapore, Oct 10–12, 2007