# Mathematical modelling and inference of genetic regulation for cell fate determination in hematopoiesis


A thesis submitted in partial fulfillment of
the requirements for the degree of

**Doctor of Philosophy**

by

**Siyuan Wu**

Supervisor: A/Prof. Tianhai Tian
Associate Supervisor: Dr. Tiangang Cui

School of Mathematics
Monash University
Australia
February 2022

—This page intentionally left blank—

# Copyright notice

—This page intentionally left blank—

**Dedicated to my family**

**especially**

**my mother Rong He and father Yong Wu**

**and**

**my wife Wei Zhang**

—This page intentionally left blank—

# Abstract

This thesis focuses on the development of mathematical methods to explore and analyze the dynamical mechanisms of genetic regulation related to cell fate determination in hematopoiesis. Although substantial research studies have already been conducted, the detailed regulatory mechanisms are still uncertain. The mathematical modelling and inference of genetic regulatory networks is therefore of particular importance. The objective of this thesis is to design two mathematical frameworks. The first framework is developed to understand nonlinear dynamics of gene expression. The second framework is designed to achieve the multistability property of a system by embedding two bistable systems. This thesis consists of three parts.

Firstly, we focus only on the genetic regulatory networks of protein monomers. We use a Forward Search Algorithm and ordinary differential equations to analyze the genetic regulation of the network and the dynamical properties. By using genetic regulation in hematopoiesis as a testing system, we provide an effective framework for studying regulatory mechanisms.

Second, we extend our methods to analyze the regulatory roles of protein heterodimers and/or synergistic effects in the determination of cell fate during hematopoiesis. We propose a new algorithm, known as Extended Forward Search Algorithm, to infer the structure of networks with both linear terms (namely protein monomers) and nonlinear terms (namely protein heterodimers and/or synergistic effects). The Taylor expansion method is used to simplify the nonlinear ordinary differential equation for describing the dynamical properties of genetic regulatory networks. This proposed approach gives accurately simulated results based on the published data on hematopoiesis.

Finally, in the third part, we find that hematopoiesis can be treated as a system with multiple stable states. To better understand the problem of cell fate determination in hematopoiesis, we develop a novel framework to obtain a multistable system by embedding systems with bistable characteristics. To ensure the multistability of the embedding system, we demonstrate the conditions of existence for all possible equilibrium states between the bistable systems and the embedding system and the conditions of stability for each equilibrium state. Using the *GATA1-GATA2-PU.1* module as a testing system, our method with stochastic simulation successfully achieves the recent experimental results.

In summary, research results in this thesis demonstrate that the proposed modelling and inference methods are powerful tools to study genetic regulatory networks and other complex systems.



**Keywords:** mathematical modelling, stochastic modelling, differential equation, genetic regulatory networks, hematopiesis, cell fate determination, multistability, network inference

—This page intentionally left blank—

# Declaration

This thesis is an original work of my research and contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

**Name:** Siyuan Wu

**Date:** February 25, 2022

—This page intentionally left blank—

# List of publications

- **Wu S.**, Cui T. and Tian T. Mathematical Modelling of Genetic Network for Regulating the Fate Determination of Hematopoietic Stem Cells, Proceedings of 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2018), 2167-2173, IEEE Press.
- **Wu S.**, Cui T, Zhang X and Tian T. 2020. A non-linear reverse-engineering method for inferring genetic regulatory networks. PeerJ 8:e9065
- **Wu S.**, Zhou T. and Tian T. 2022. A robust method for designing multistable systems by embedding bistable subsystems. NPJ Systems Biology and Applications (Accepted)

—This page intentionally left blank—

*"Wir müssen wissen. Wir werden wissen."*

— David Hilbert

—This page intentionally left blank—

# Acknowledgements

Throughout the writing of this thesis I have received a great deal of support and assistance.

I would like to express my innermost gratitude to my supervisor, **Associate Professor Tianhai Tian** and **Dr. Tiangang Cui**, whose expertise was invaluable in formulating the research questions and methodology. Also for your constant guidance and patient support throughout the whole PhD life. Without your inspiring and encouragement, this thesis could not have been accomplished. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I would also like to acknowledge the members of my PhD Committee, **Associate Professor Jonathan Keith**, **Associate Professor Tim Garoni** and **Dr. Mark Flegg**. Your timely suggestions with kindness and valuable feedback have enabled me on the right track. I also thank my collaborators, **Professor Tianshou Zhou** from Sun Yet-sen University and **Professor Xinan Zhang** from Central China Normal University, for your efforts in putting this into our manuscripts.

I would also like to thank graduate research coordinator, **Associate Professor Heiko Dietrich**, and school postgraduate administrator, **Mr. John Chan**, for your assistance in many aspects during my PhD time at Monash University.

I have been very grateful for the friendship and support of my friends. Many thanks go to my brilliant colleagues at the School of mathematics, especially to my good friend, **Ziwen Zhong**. In addition, it is a pleasure to thank my wonderful friends from the WeChat group, 2021 Get Rich, for the wonderful times we shared, especially the countless board-game nights and dinners. I also want to thank my close friends **Peichun Yang** and **Ke Zhang** for always having faith in me and telling me "I know you can do it". I feel fortunate to have met you in Melbourne and made my time in Australia memorable. Moreover, I would like to express my special thanks to my best friend, **Shiyi Wang**. Thank you for being by my side in my life at the most difficult time, as well as many unforgettable chats and discussions on the topic of games, research, career and life. Your friendship means more to me than you can ever know.

Last but not least, my family deserves endless gratitude: my mother, **Rong He**, for teaching me how to fear and love life, and my father, **Yong Wu**, for teaching me how to be strong when faced with challenges. I will leave this last word to my wife, **Wei Zhang**, thank you for your everlasting love, including but not limited to the constant encouragement, spiritual companionship and fabulous food. Your twelve years of companionship was an essential foundation for me to be able to achieve my dream.

Big thank you to each and everyone!

—This page intentionally left blank—

# Contents

# List of Figures

# List of Tables

# 1 Introduction

Biological systems are highly ordered and complex, with the cell serving as the basic unit. The behaviour of cellular activities is then controlled by complex regulatory mechanisms. Researchers frequently choose to describe activity within a cell in terms of data (e.g., gene expression data), chemical reaction equations, or networks (e.g., gene regulatory networks/biochemical reaction networks). Regulatory networks provide a comprehensive examination of molecular mechanisms underlying biological activities [116]. With the advances of omics technologies, a large amount of gene expression data can be obtained by high-throughput technology with different experimental conditions, including time-series gene expression data [8, 115]. But how to harness such experimental data to uncover the nature of the mechanisms underlying the biological systems is still a challenge for researchers today, especially for the mechanism of cell fate determination. Recently, a large number of network inference algorithms have been designed to deduce the regulation of molecular components inside biological systems [83, 101, 115], and also a number of mathematical models have been proposed to describe the dynamical properties of regulatory network [28, 78, 83, 103, 146, 152, 158, 159]. Chapter 2 will introduce a needed biological background and also describes some classical computational models and methods for studying genetic regulatory networks. One of the key challenges in most of the inference methods is the large number of unknown parameters compared with the relatively small amount of data. Chapter 3 and Chapter 4 provide a general approach to address these issues in predicting the dynamical mechanism of genetic regulatory networks in hematopoiesis.

Another issue is that majority of current mathematical models only considered the function of each gene as monomer (namely using $x_i$) or homodimer (namely using $x_i^2$). The function of heterodimer (namely using $x_i x_j$) has not been considered. There is a lack of

investigations for the effect of possible protein heterodimers and/or synergistic effect in genetic regulations. The main reason is that the number of unknown parameters in the model will be the order of $n^3$ for a network with $n$ genes. It would be difficult to estimate such a large number of model parameters, and therefore a large amount of experimental data is needed to determine these parameters. More importantly, it is very challenging to infer the model parameters due to the complexity of parameter space and high computational cost. Given the large parameter number issue, a linear ODE model may have the least number of unknown parameters among the models for all possible relationships between genes and protein heterodimers and/or synergistic effects. However, as the linear model is limited to describing linear relationships, it is not appropriate to use a linear model to study systems with complex nonlinear dynamics. Chapter 4 addresses these issues by proposing a inference method to the reconstruction of genetic regulatory networks in hematopoiesis with genes and the protein heterodimers and/or synergistic effects.

Multistability is the characteristic of a system that exhibits two or more mutually exclusive stable states. This phenomenon has been observed in many different disciplines of science, including genetic regulatory networks [37, 75, 106, 117], cell signalling pathways [4, 48, 132], metabolic networks [24], ecosystems [9, 79], neuroscience [58], laser systems [44, 68], and quantum systems [67]. When external and/or internal conditions change, the system may switch from one stable state to another either randomly by perturbations or in a desired way according to control strategies. In recent years mathematical models with multistability have been developed for theoretical analysis and computer simulations, which shed light on the mechanisms that generate multistability and control the transition between stable states [7, 38, 63, 108].

Mathematical modelling is a powerful tool to explore the regulatory mechanisms of multistability for controlling the transitions between different cell types [3, 29, 103]. A number of mathematical models have been proposed to describe the underlying mechanisms of multistability inside biological systems [13, 21, 52, 95, 119, 136]. Although these attempts have realized multistability by using different assumptions, it is still a challenge to develop mathematical models for realizing tristability by using both realistic regulatory mechanisms and experimental data. On the other hand, substantial research studies have been conducted to develop mathematical models for realizing bistability properties [37, 39, 71, 84, 107, 118, 135]. Thus, the question is whether we can develop mathematical models with tristability or higher-order of multistability by using the bistable models. Chapter 5 addresses this problem by proposing a novel and robust method to develop multistable models by embedding bistable models, and it uses this methodology to analyze the problem of cell fate determination in hematopoiesis.

## 1.1 Thesis outline

Motivated by these issues, the principal aim of this thesis is to use mathematical and computational methods to analyze the underlying mechanisms of genetic regulatory networks involved in the cell fate determination in hematopoiesis. Here, a brief description of each chapter contained in this thesis is now presented.

- Chapter 2: This chapter contains various biological and mathematical backgrounds needed for this thesis.
- Chapter 3: This chapter contains a general approach that is used to predict the dynamical mechanism of genetic regulatory networks in hematopoiesis by combining both the top-down approach (for reducing the complexity of the network structure) and bottom-up approach (for deriving the detailed dynamic properties) [153].
- Chapter 4: This chapter presents a general approach with a new inference algorithm, Extended Forward Search Algorithm, and a new mathematical model to infer the genetic regulatory networks of genes and the protein heterodimers and/or synergistic effects [154].
- Chapter 5: This chapter provides a new framework to achieve multistability from bistable systems. The methodology is also used to discuss the cell fate determination in hematopiesis [155].
- Chapter 6: This chapter discusses all studies in this thesis, and also remarks some interesting topics for future research.

# 2

# Background

The objective of this chapter is to provide the necessary biological and mathematical background knowledge needed for the follow-up chapters.

## 2.1 Introduction to genetic regulation

In 1953, with the introduction of the double helix structure of DNA by James D. Watson and Francis Crick, the nature of genes was further recognized as DNA segments that contain genetic information [150]. It has been shown that there are one or two DNA molecules per chromosome, each with multiple genes, each containing large numbers of deoxyribonucleotides. Therefore, deoxyribonucleotides are the material basis of DNA. There are four bases in deoxyribonucleotides (i.e., adenine, guanine, cytosine, and thymine) that determine biodiversity. Different genes contain different genetic information due to the different base sequences. The genetic information stored in DNA represents the genotype. The most fundamental level at which the genotype results in an individual's phenotype is gene expression. In genetics, gene expression is the process by which information from a gene is used to synthesize a functional gene product that allows the gene to produce end products such as protein or noncoding RNA and, as a result, alter the phenotype [80]. Although every cell in an organism contains all the genetic information, the gene expression varies for different cell types. The specificity of gene expression supports the basic construction and performance of life. Gene expression is regulated by genetic regulatory networks at several stages in protein synthesis, including DNA transcription, RNA processing, RNA translation

and post-translational modification of a protein [28]. Genetic regulatory networks are collections of molecular regulators that interact with each other and with other substances in the cell to control the timing, location and levels of gene expression of mRNAs and proteins within the cell, which in turn determines the cell's structure and function [148]. Consequently, uncovering the mechanisms of gene expression and regulation is essential to the understanding of basic intracellular processes and is also critical for understanding the onset and development of various diseases and their treatment processes [53, 72, 87, 89, 121]. Although the process of gene expression is complex, its underlying principles can be described by the central dogma of molecular biology.

The central dogma of molecular biology is one of significant milestones in biology. It describes the transition flow of genetic information within a biological system. The original dogma is first stated by Francis Crick in 1958 [26]. The dogma states that the genetic information contained in a protein-coding gene is expressed through transcription and translation. RNA polymerases use one of two DNA strands as a template to form a complementary RNA strand that carries the genetic information contained in the DNA into the ribosome in the form of mRNA, where it controls the synthesis of proteins [80]. That is, the flow of genetic information within cells follows that outline DNA $\rightarrow$ RNA $\rightarrow$ Protein. In 1970, Howard Temin and David Baltimore discovered a reverse transcriptase that catalyzes the synthesis of DNA using RNA as a template from two RNA tumour viruses: Rous sarcoma virus and murine leukemia virus [6, 130]. The discovery of this reverse transcriptase revealed that genetic information can flow not only from DNA to RNA, but also from RNA to DNA, which further developed and refined the central dogma of molecular biology. The updated central dogma with the path RNA $\rightarrow$ DNA was restated by Francis Crick in the same year [25]. We therefore understand that genetic information flow is a two-way transfer between DNA and RNA, but the information is transmitted one-way from nucleic acids to proteins as shown in Figure 2.1.

By studying gene regulatory systems in depth, researchers have uncovered a number of complex molecular regulatory mechanisms that are grounded in the central dogma. For example, in eukaryotic cells, a sequence of DNA is transcribed into a pre-mRNA by RNA polymerases with the aid of transcription factors (TFs). The pre-mRNA must be processed in to mRNA by the following steps

1. Adding a 5' cap to the beginning of the pre-mRNA.
2. Adding a 3' poly-A tail to the end of the pre-mRNA.
3. Removing the introns (non-coding sequences) from the pre-mRNA.
4. Joining the exons (protein-coding sequences) together.

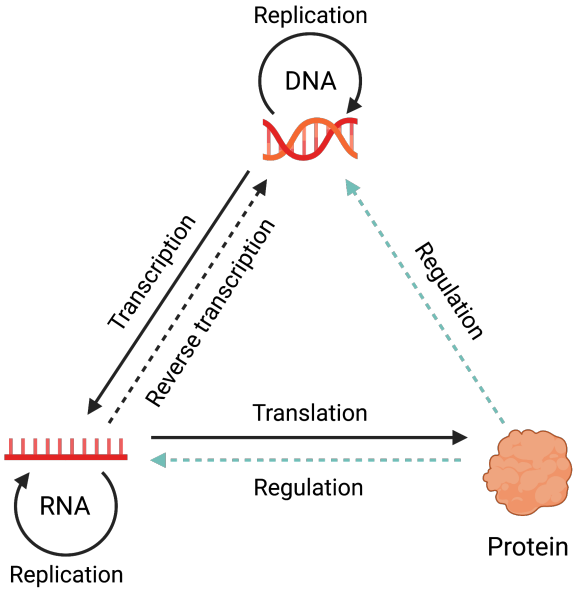


Figure 2.1: **Description of the Central dogma of molecular biology:** Black solid arrows represent general information transfer; Black dashed arrow represents the special information transfer from RNA to DNA with specific reverse transcriptase; Cyan dashed arrows indicate that the protein will bind to a specified gene site to regulate the expression of DNA and RNA, rather than the information flow stated in the central dogma. Created with BioRender.com.

Then, the mRNA is exported to the cytoplasm from the nucleus for translation. When the translation process is activated, ribosomes will move along the mRNA from 5' end to 3' end. At the same time, amino acids are transported to the ribosomes by transfer RNAs and paired with mRNA codons to synthesize the corresponding polypeptide. After being modified, the polypeptide folds into its characteristic and functional 3D structure to become a mature protein. However, the detailed mechanisms of these chemical reactions are beyond the scope of this thesis, which we will not discuss further.

## 2.2 Introduction to hematopoiesis

Hematopoiesis is a process for blood cell production, which is a highly complex process that controls the proliferation, differentiation and maturation of hematopoietic stem cells (HSCs) [97]. In adult mammals, all blood cell types arise from HSCs that reside mainly in the bone marrow [12, 113]. HSCs have the feature of self-renewal and multipotency as well as the ability to differentiate into multipotent progenitors (MPPs). Then, MPPs will differentiate into two main lineages of blood cells, namely the myeloid lineage which starts

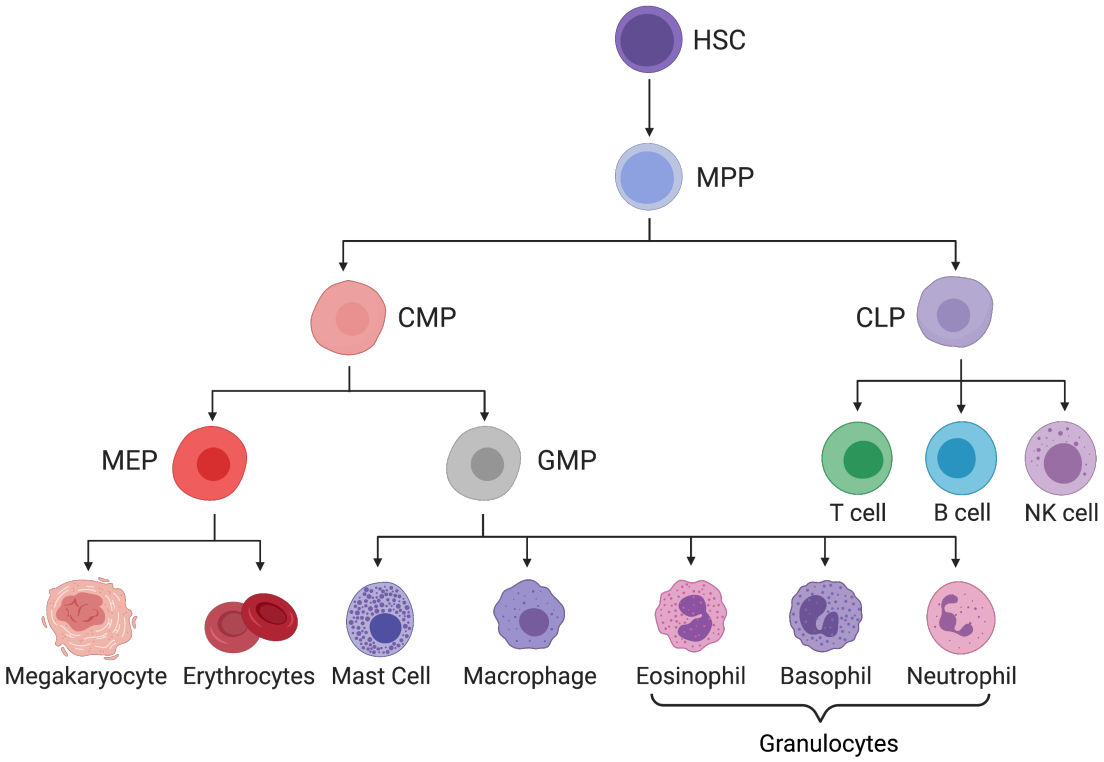


Figure 2.2: **Diagram of developmental process of hematopoietic stem cell.** Created with BioRender.com.

at common myeloid progenitors (CMPs) and the lymphoid lineage which starts at common lymphoid progenitors (CLPs). In addition, the myeloid lineage has two distinct progenitors, namely megakaryocyte-erythroid progenitors (MEPs) and granulocyte-macrophage progenitors (GMPs). MEPs can differentiate into megakaryocytes and erythrocytes, and GMPs can give rise to mast cells, macrophages and granulocytes. Lymphoid lineage cells include T lymphocytes (T-cells), B lymphocytes (B-cells) and natural killer cells (NK-cells) [104]. Figure 2.2 gives the brief developmental process of HSC to different blood cells.

Blood is one of the most regenerative tissues in the organism, and new blood cells are constantly replenished into the organism's metabolic cycle [113]. Due to this ability of blood cells, the hematopoietic system has been successfully used in regenerative medicine for more than three decades. At present, many malignant blood disorders, such as leukemia, are treated primarily through HSCs transplantation. Transplanting healthy HSCs can help patients quickly restore the health of their hematopoietic system by taking advantage of its highly regenerative power. To maintain proper function of the hematopoietic system, an organism must generate the appropriate amount of specific blood cells at the appropriate

time and location. Hence, HSCs must continually determine their cell fates correctly, such as when to initiate proliferation, self-renew, differentiate, which lineage to develop into, and when to undergo apoptosis process. Whereas each cell has the potential to select for different fates, the output of distinct mature cell types is balanced and regulated at the population level. If the normal mechanisms of cell fate determination of HSCs are disrupted so that the distribution of different cell types produced is skewed [47]. This can lead to a variety of blood disorders, such as leukemia. Thus, it is imperative to unravel the regulatory mechanisms for cell fate determination of HSCs.

At the molecular level, a number of TFs act as key regulators to control the cell fate determination of HSCs and operate within organized regulatory programs that can be modelled as regulatory networks [2, 16, 47]. For example, *Mef2c* is abundantly expressed in HSCs and CLPs. Whereas the expression declines when CMPs differentiate into GMPs and MEPs [32]. Studies reported that *Mef2c* is a direct target gene of of *Scl/Tal1* and regulates the megakaryopoiesis and B-cell homeostasis [43]. In addition, *Scl/Tal1* is essential for hematopoiesis, maturation of both megakaryopoiesis and erythropoiesis [32]. The Runx family is another collection of key transcription factors. *Runx1* is primarily involved in the differentiation and self-renewal of HSCs, while *Runx2* and *Runx3* are important in its maintenance [15]. The GATA family is also play a significant role in the development and differentiation of HSCs. Especially, the *GATA1/2/3* are required for HSC proliferation and differentiation into erythrocytes and megakaryocytes [143]. *Gata3* is abundantly expressed in CMPs, and it is a master regulator of T cell differentiation while inhibiting their differentiation into B-cells [138].

In this thesis, we mainly focus on the fate determination of HSCs in the myeloid lineage for the choice between MEPs lineage and GMPs lineage. The genetic complex *GATA1-GATA2-PU.1* is a very important module for the cell fate determination of HSCs between erythrocytes or granulocytes differentiation [41, 76, 77]. *GATA2* is mainly expressed in the HSCs and MPPs and regulates the hematopoiesis. The initial expression levels at HSCs are high for gene *GATA2* but low for genes *GATA1* and *PU.1*, since gene *GATA2* controls the production and proliferation of HSCs by an autoregulation [77]. The erythroid commitment of progenitors decreases the expression of *GATA2*, while expression of *GATA1* increases with erythroid differentiation. Experimental studies suggested that *GATA1* directly represses *GATA2* transcription and *GATA2* and *GATA1* sequentially bind the same *cis*-elements (as shown in Figure 2.3). This transition process is referred to as the GATA-switching [56, 124]. Therefore, the GATA-switching is an essential driver of hematopoiesis [14]. Moreover, the *GATA1-PU.1* complex forms a double negative feedback module, in which each gene inhibits the expression of the other [41]. HSCs are more likely to choose MEPs lineage with high expression levels of *GATA1*, or conversely to choose GMPs lineage with high expression

levels of *PU.1*. Recently it has been elucidated that the fate determination of HSCs between MEPs and GMPs lineage was defined not only by the ratio of *GATA1* and *PU.1* [51], but also by a third party during the regulation. For example, *FOG-1* is a significant third party to regulate the *GATA1-PU.1* module [17, 85]. Erythropoietin receptor (EpoR) signalling also acts the essential role in regulating the *GATA1-PU.1* Module [166]. Although the regulatory mechanisms of *GATA1-GATA2-PU.1* module in hematopoiesis are relatively well studied, the connection of this triad complex with other significant genes as well as the role of these genes in hematopoiesis are mostly unclear [76].

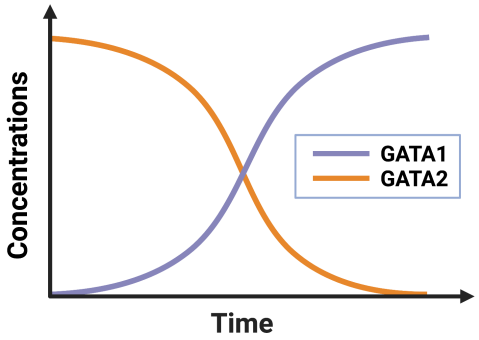


Figure 2.3: **Illustrative diagram of GATA-Switching.** Expression levels of *GATA1* and *GATA2* during GATA-switching process. Created with BioRender.com.

As mentioned above, the cell fate determination of HSCs is governed by genetic regulatory networks. Although the regulatory mechanisms have been studied over a century, there are still many challenging questions regarding cell fate determination in hematopoiesis [105]. One of the reasons is the experimental validation of functional relationships between regulator and target genes does not readily scale to a system-wide approach [47]. Therefore, mathematical modelling and inference methods have become widely used to study the mechanism of genetic regulations. This use of mathematical models to investigate regulatory networks will provide a novel perspective on gene regulation mechanisms. This is important for researchers to comprehensively analyze the mechanisms of cell fate determination in the hematopoietic system and to better understand the pathogenesis of malignant blood disorder, prevention methods, and the development of new treatments.

## 2.3 Mathematical models for genetic regulation

Mathematical modelling is an important method for studying the detailed regulatory mechanisms. Mathematical models can describe the functioning of biological systems more objectively and accurately than the data itself. Starting in the 20th century, a large number of mathematical models are used to quantitatively analyze biological systems [28, 50, 103, 120]. These proposed mathematical formalisms are developed by known genetic regulatory mechanisms[28]. In this section, I will give a brief introduction to the most common quantitative approaches to describe the regulatory mechanism for determining stem cell fates.

### 2.3.1 Ordinary differential equation models

Ordinary differential equations (ODEs) are most used to analyze genetic regulatory mechanism. ODEs describe the production rate of an RNA or a protein as a function of the presence of other reactants in the transcriptional process with the mathematical form [28] as follows

$$\frac{dX_i}{dt} = f_i(\boldsymbol{X}), \tag{2.1}$$

where $\boldsymbol{X}$ is a vector of concentrations of reactants with non-negative real valued components. Next, I will introduce some classical ODE models in analysis of regulatory mechanisms.

#### 2.3.1.1 Linear model

Linear modelling is the most direct method to analyze how the production rate of a protein or RNA is regulated by different reactants [28].

$$\frac{dX_i}{dt} = \sum_{j=1}^{n} a_{ij} X_j \tag{2.2}$$

where $a_{ij}$ is the regulation parameter from the reactant $X_j$ to $X_i$. If $a_{ij}$ is positive (negative), which means the concentration of reactant $X_j$ will up-regulate (down-regulate) the production rate of $X_i$.

#### 2.3.1.2 Michaelis-Menten formalism

In 1913, Leonor Michaelis and Maud Menten proposed the mathematical model to describe the kinetics of an enzymatic reaction mechanism [91]. This formalism is based on the two important rules.

**Theorem 2.3.1** (Conservation of mass)**.** For any system closed to all transfers of matter and energy, the mass of the system must remain constant over time, as the system's mass cannot change, so quantity can neither be added nor be removed. Therefore, the quantity of mass is conserved over time.

**Theorem 2.3.2** (Law of mass action)**.** The rate of the chemical reaction is directly proportional to the product of the activities or concentrations of the reactants.

Let us first consider reaction equation for transcriptional activation,

$$\mathrm{TF} + \mathrm{DNA} \underset{\mathrm{k}_b}{\overset{\mathrm{k}_f}{\rightleftharpoons}} \mathrm{TF{-}DNA} \xrightarrow{\mathrm{k_{cat}}} \mathrm{P} + \mathrm{TF{-}DNA}, \tag{2.3}$$

which describes the process of a TF binds to a DNA to form a complex TF-DNA. For simplicity, we neglect the mRNA production process. Then this complex activates the production of protein (P). Moreover, the TF binding is a reversible process, and we assume the binding process is fast enough. In this equation $k_f$ is a forward reaction rate, $k_b$ is a backward reaction rate and $k_{cat}$ is a catalytic rate. Before we move forward, I need to point out that we will use the notation [.] represents the concentration of a reactant. Based on the Theorem 2.3.1, we have the following relationship about the concentration of DNA.

$$[\mathrm{DNA}]_{total} = [\mathrm{DNA}]_{free} + [\mathrm{DNA}]_{combined}, \tag{2.4}$$

where $[\mathrm{DNA}]_{free}$ and $[\mathrm{DNA}]_{combined}$ are actually the concentration of unbinding DNA and the concentration of the complex TF-DNA, respectively. That is,

$$[\mathrm{DNA}]_{total} = [\mathrm{DNA}] + [\text{TF-DNA}]. \tag{2.5}$$

More specifically, we have

$$\begin{aligned} &\text{the concentration of free DNA} \\ =\ &\text{the concentration obtained from the backward reaction of TF-DNA} \\ -\ &\text{the concentration consumed from the forward reaction to TF-DNA.} \end{aligned}$$

Then, based on the Theorem 2.3.2, we have the following differential equation to describe the rate of change of the concentration of enzyme with respect to the time $t$

$$\frac{d[\mathrm{DNA}]}{dt} = k_b[\text{TF-DNA}] - k_f[\mathrm{TF}][\mathrm{DNA}]. \tag{2.6}$$

The similar consideration applied to other reactants, we also have the following three equations. Since concentration of the complex TF-DNA before and after the transcription

remains same, there is no consumption for TF-DNA in the second (right) reaction.

$$\frac{d[\text{TF}]}{dt} = k_b[\text{TF-DNA}] - k_f[\text{TF}][\text{DNA}], \tag{2.7}$$

$$\frac{d[\text{TF-DNA}]}{dt} = k_f[\text{TF}][\text{DNA}] - k_b[\text{TF-DNA}], \tag{2.8}$$

$$\frac{d[\text{P}]}{dt} = k_{cat}[\text{TF-DNA}]. \tag{2.9}$$

Assume the reaction between forward and backward reaches the chemical equilibrium, which means that the both reactants and products are present in the rate of change of the concentration which have no further tendency to change, that is,

$$k_f[\text{TF}][\text{DNA}] = k_b[\text{TF-DNA}]. \tag{2.10}$$

Then substitutes (2.5) into (2.10), we have

$$k_f[\text{TF}]([\text{DNA}_{total}] - [\text{TF-DNA}]) = k_b[\text{TF-DNA}] \tag{2.11}$$

Rearrange the equation, we have

$$[\text{TF-DNA}] = \frac{k_f[\text{DNA}]_{total}[\text{TF}]}{k_b + k_f[\text{TF}]}, \tag{2.12}$$

$$= \frac{[\text{DNA}]_{total}[S]}{\frac{k_b}{k_f} + [\text{TF}]}, \tag{2.13}$$

$$= \frac{[\text{DNA}]_{total}[\text{TF}]}{K_d + [\text{TF}]}, \tag{2.14}$$

where $K_d = k_b/k_f$ is the dissociation rate of the complex TF-DNA. Finally, we substitute (2.14) into (2.9), we will have the rate of product P is produced by the enzyme reaction. The expression is given by

$$\frac{d[\text{P}]}{dt} = k_{cat}\frac{[\text{DNA}]_{total}[\text{TF}]}{K_d + [\text{TF}]} \tag{2.15}$$

$$= V_{max}\frac{[\text{TF}]}{K_d + [\text{TF}]}, \tag{2.16}$$

where $V_{max} = k_{cat}[\text{DNA}]_{total}$ is the maximum reaction rate of product P. This is a increasing function of TF concentration.

Then, we consider the reaction for transcriptional repression, which means that if a TF binds to a DNA, blocking the production of protein P. In this case, we need to consider the

equation separately.

$$\mathrm{TF} + \mathrm{DNA} \underset{\mathrm{k}_b}{\overset{\mathrm{k}_f}{\rightleftharpoons}} \mathrm{TF{-}DNA}, \tag{2.17}$$

$$\mathrm{DNA} \xrightarrow{\mathrm{k_{cat}}} \mathrm{P} + \mathrm{DNA}. \tag{2.18}$$

The same idea as above, based on the Theorem 2.3.1 and Theorem 2.3.2, we know that

$$[\mathrm{DNA}]_{total} = [\mathrm{DNA}] + [\mathrm{TF\text{-}DNA}], \tag{2.19}$$

$$\frac{d[\mathrm{P}]}{dt} = k_{cat}[\mathrm{DNA}], \tag{2.20}$$

$$k_f[\mathrm{TF}][\mathrm{DNA}] = k_b[\mathrm{TF\text{-}DNA}]. \tag{2.21}$$

Then, we substitute (2.19) into (2.21) and rearrange the equation, we have

$$[\mathrm{TF\text{-}DNA}] = \frac{k_f[\mathrm{DNA}]_{total}[\mathrm{TF}]}{k_b + k_f[\mathrm{TF}]}. \tag{2.22}$$

Next, we substitute (2.21) into (2.22),

$$\frac{k_f}{k_b}[\mathrm{TF}][\mathrm{DNA}] = \frac{k_f[\mathrm{DNA}]_{total}[\mathrm{TF}]}{k_b + k_f[\mathrm{TF}]} \tag{2.23}$$

$$\Longrightarrow [\mathrm{DNA}] = \frac{\frac{k_b}{k_f}[\mathrm{DNA}]_{total}}{\frac{k_b}{k_f} + [\mathrm{TF}]} = \frac{K_d[\mathrm{DNA}]_{total}}{K_d + [\mathrm{TF}]} \tag{2.24}$$

Finally, we substitute (2.24) into (2.20). The rate equation for the production of protein is given by

$$\frac{d[\mathrm{P}]}{dt} = k_{cat}\frac{K_d[\mathrm{DNA}]_{total}}{K_d + [\mathrm{TF}]} \tag{2.25}$$

$$= V_{max}\frac{K_d}{K_d + [\mathrm{TF}]} \tag{2.26}$$

Thus, we have discussed the production rate of a protein for transcriptional activation and repression, respectively. However, a protein has a half-life that allows it to degrade. Based on these results, we can have the following key points:

1. If the production a protein is up-regulated by a TF, The concentration of a protein is given by

$$\frac{d[\mathrm{P}]}{dt} = V_{max}\frac{[\mathrm{TF}]}{K_d + [\mathrm{TF}]} - k^*[\mathrm{P}], \tag{2.27}$$

where $k^*$ is the degradation rate of protein P.

2. If the production a protein is down-regulated by a TF, The concentration of a protein is given by

$$\frac{d[\mathrm{P}]}{dt} = V_{max}\frac{[K_d]}{K_d + [\mathrm{TF}]} - k^*[\mathrm{P}]. \tag{2.28}$$

3. We can also add the logic operator into the Michaelis-Menten formalism with multiple different TFs [103].

   - If the production a protein is down-regulated by a transcriptional factor A AND up-regulated by a transcriptional factor B, we can develop the following model

$$\frac{d[\mathrm{P}]}{dt} = \frac{V_A K_A}{K_A + [\mathrm{TF}_A]} \cdot \frac{V_B[\mathrm{TF}_B]}{K_B + [\mathrm{TF}_B]} - k^*[\mathrm{P}]. \tag{2.29}$$

   - If the production a protein is down-regulated by a transcriptional factor A OR up-regulated by a transcriptional factor B, we can develop the following model

$$\frac{d[\mathrm{P}]}{dt} = \frac{V_A K_A}{K_A + [\mathrm{TF}_A]} + \frac{V_B[\mathrm{TF}_B]}{K_B + [\mathrm{TF}_B]} - k^*[\mathrm{P}]. \tag{2.30}$$

#### 2.3.1.3 Hill equation

In 1910, Archibald Hill proposed a classical nonlinear ODE model to describe the sigmoidal oxygen binding curve of haemoglobin [50]. Since then, Hill equation have been applied to explore the mechanisms in a wide range of genetic regulatory networks and biological systems. For example, the genetic toggle switching was achieved by the models with Hill equations [42]. In addition, Hill equation was employed to formalize the mechanisms of cell fate determination [52, 69, 156] and Hill equation with high cooperativity was used to realize the tristability [52]. Recently, Hill equation was also used to discover a regulatory network of 52 genes with the uniform activation and repression strengths [73].

To better understand the concept of Hill equation, we need to know the following definition.

**Definition 2.3.1** (Cooperative binding)**.** The binding of a ligand to a molecule is often enhanced if there are already other ligands present on the same molecule.

For example, let us consider the chemical reaction in gene transcription for activation.

$$\mathrm{DNA} + n\,\mathrm{TF} \underset{\mathrm{k}_b}{\overset{\mathrm{k}_f}{\rightleftharpoons}} \mathrm{TF}_n{-}\mathrm{DNA} \xrightarrow{\mathrm{k_{cat}}} \mathrm{P} + \mathrm{TF}_n{-}\mathrm{DNA}, \tag{2.31}$$

which describes the process of $n$ same TFs bind simultaneously to a DNA. The Cooperative binding is still a reversible process. In this equation, $n$ is the number of TFs in the chemical reaction. The same idea is applied when consider the transcriptional repression with cooperative binding. The derivation of Hill functions is highly similar with the Michaelis-Menten formalism. The only difference is we replace [TF] with $[\mathrm{TF}]^n$ during the algebraic operation. We therefore can summarise the following key points for Hill functions:

1. If the production a protein is up-regulated by TF, The concentration of a protein is given by

$$\frac{d[\mathrm{P}]}{dt} = V_{max}\frac{[\mathrm{TF}]^n}{K_d + [\mathrm{TF}]^n} - k^*[\mathrm{P}], \tag{2.32}$$

$$= V_{max}\frac{[\mathrm{TF}]^n}{K^n + [\mathrm{TF}]^n} - k^*[\mathrm{P}], \tag{2.33}$$

   where $n$ is called the Hill coefficient, which measures degree of cooperativity. $K_d$ can be rewritten as $K_d = K^n$, we say that $K$ is the dissociation constant with the Hill coefficient $n$.

2. If the production a protein is down-regulated by a TF, The concentration of a protein is given by

$$\frac{d[\mathrm{P}]}{dt} = V_{max}\frac{K^n}{K^n + [\mathrm{TF}]^n} - k^*[\mathrm{P}], \tag{2.34}$$

3. We can also add the logic operator into Hill equation with multiple different TFs [103].

   - If the production a protein is down-regulated by a transcriptional factor A AND up-regulated by a transcriptional factor B, we can develop the following model

$$\frac{d[\mathrm{P}]}{dt} = \frac{V_A K_A^{n_A}}{K_A^{n_A} + [\mathrm{TF}_A]^{n_A}} \cdot \frac{V_B[\mathrm{TF}_B]^{n_B}}{K_B^{n_B} + [\mathrm{TF}_B]^{n_B}} - k^*[\mathrm{P}]. \tag{2.35}$$

   - If the production a protein is down-regulated by a transcriptional factor A OR up-regulated by a transcriptional factor B, we can develop the following model

$$\frac{d[\mathrm{P}]}{dt} = \frac{V_A K_A^{n_A}}{K_A^{n_A} + [\mathrm{TF}_A]^{n_A}} + \frac{V_B[\mathrm{TF}_B]^{n_B}}{K_B^{n_B} + [\mathrm{TF}_B]^{n_B}} - k^*[\mathrm{P}]. \tag{2.36}$$

It is clear to see that the Michaelis-Menten formalism is actually a special case of Hill function with the Hill coefficient $n = 1$. That is, Michaelis-Menten formalism does not consider any cooperativity in chemical reactions.

#### 2.3.1.4 Shea-Ackers formalism

Another widely used approach is the Shea-Ackers formalism for studying the mechanism of genetic regulatory networks [120]. It models gene transcriptions from the thermodynamic view. For example, Tian used the Shea-Ackers formalism to realize the mechanisms of GATA-switching and designed an effective algorithm to realize tristability [136]. The structure of this formalism is similar to the Michaelis-Menten formalism. To understand this formalism, we need the following biological terminologies.

**Definition 2.3.2** (Promoter)**.** In genetics, a promoter is a sequence of DNA to which RNA polymerase bind that initiate transcription of a single RNA from the DNA downstream of it.

**Definition 2.3.3** (RNA polymerase (RNAP))**.** RNAP is an enzyme that binds to a promoter for basal transcription.

**Definition 2.3.4** (Basal transcription)**.** Basal transcription is defined as the level of transcript produced by RNAP in the absence of regulation by TF.

Consider both TF and RNAP, there are four possible binding situations: (1) Nothing bound, (2) TF bound, (3) RNAP bound and (4) TF-RNAP bound. The first two situations mean that transcription is "off" and the last two situations mean that transcription is "on". The Shea-Acker formalism describes the transcription rate by the ratio of the statistical weight of all "on" states ($Z_{\text{on}}$) to the total statistical weight of all "off" and "on" states ($Z_{\text{total}} = Z_{\text{off}} + Z_{\text{on}}$). Let us consider the following example, suppose a production rate of protein is regulated by a TF and RNAP. The four scenarios are presented in Table 2.1. In the columns of "TF binding site" and "RNAP binding site", 0 means unbound and 1 means bound. In the column of "Status", 0 means transcription is inactive and 1 means transcription is active. In addition, Case 1 is the reference state, which represents there is nothing bound to the DNA, the binding constant for the reference state is always 1. Case 2 is the case of basal transcription since only RNAP bound to the DNA. Moreover, the rate is calculated based on the Theorem 2.3.1. We will use the (partial) sum of these rates to obtain the statistical weight for both "off" states and "on" states.

In this case, we can have the following ODE model to describe the production rate of protein P

$$\frac{d[\text{P}]}{dt} = \frac{Z_{\text{on}}}{Z_{\text{off}} + Z_{\text{on}}} = \frac{\beta_1[\text{RNAP}] + \beta_3[\text{TF}][\text{RNAP}]}{1 + \beta_1[\text{RNAP}] + \beta_2[\text{TF}] + \beta_3[\text{TF}][\text{RNAP}]}. \tag{2.37}$$

| | TF binding site | RNAP binding site | Status | Binding constant | Rate |
|---|---|---|---|---|---|
| Case 1 | 0 | 0 | 0 | 1 | 1 |
| Case 2 | 0 | 1 | 1 | $\beta_1$ | $\beta_1$[RNAP] |
| Case 3 | 1 | 0 | 0 | $\beta_2$ | $\beta_2$[TF] |
| Case 4 | 1 | 1 | 1 | $\beta_3$ | $\beta_3$[TF][RNAP] |

Table 2.1: **Example of Shea-Ackers formalism**

### 2.3.2 Stochastic differential equation models

There are many different sources of variation that exist in gene expression data. In particular, the chemical reactions in gene transcription are intrinsically stochastic, leading to variation in the production of proteins [35]. However, deterministic models cannot describe the noise in gene transcription. Stochastic model was therefore proposed to investigate the function of noise in regulating cell fate determination. Numerical results also suggested that fluctuations of protein numbers may lead stem cells to different developmental pathways. Developing stochastic models to understand noise in gene expression data therefore has significant implications for understanding regulatory mechanisms. The stochastic differential equation (SDE) model is constructed from an ODE model with an extra noise term. Suppose the general ODE to describe the dynamics of gene transcription is given by

$$\frac{d[\mathrm{P}]}{dt} = f(\mathbf{x}) - k^*[\mathrm{P}], \tag{2.38}$$

where $\mathbf{x}$ is a vector of TFs and $f(\mathbf{x})$ represents the regulation of these TFs. The noise term in SDE is based on the Wiener process.

**Definition 2.3.5** (Wiener Process)**.** Wiener process $\{W(t)\}$ is a stochastic process with the following properties:

1. (Independence of increments) $W(t) - W(s)$, for $t > s$, is independent of the past $W(u)$, $0 \leq u \leq s$.

2. (Normal increments) $W(t) - W(s)$ has normal distribution with mean 0 and variance $t - s$.

3. (Continuity of paths) $W(t)$, $t \geq 0$ are continuous function of $t$.

Based on these basis, we can have the following SDEs with a noise strength parameter $\mu$ [133]

- SDE model with additive noise

$$d[\mathrm{P}] = (f(\mathbf{x}) - k^*[\mathrm{P}])dt + \mu dW(t). \tag{2.39}$$

This model only consider the stochastic properties explicitly in gene expression data. Stochasticity may occur during the process of protein degradation. Thus, the following two models are proposed.

- SDE model with multiplicative noise in protein degradation

$$d[\mathrm{P}] = (f(\mathbf{x}) - k^*[\mathrm{P}])dt + \mu k^*[P]dW(t). \tag{2.40}$$

- SDE model with Langevin noise in protein degradation

$$d[\mathrm{P}] = (f(\mathbf{x}) - k^*[\mathrm{P}])dt + \mu\sqrt{k^*[P]}dW(t). \tag{2.41}$$

These three SDE models above fail to take stochasticity into account in the transcriptional process. We therefore also have two SDE models that consider stochasticity in gene transcription to address this issue as follows.

- SDE model with multiplicative noise in the process of gene transcription

$$d[\mathrm{P}] = (f(\mathbf{x}) - k^*[\mathrm{P}])dt + \mu b f(\mathbf{x})dW(t). \tag{2.42}$$

- SDE model with Langevin noise in the process of gene transcription

$$d[\mathrm{P}] = (f(\mathbf{x}) - k^*[\mathrm{P}])dt + \mu\sqrt{b f(\mathbf{x})}dW(t). \tag{2.43}$$

Moreover, we can also combine any one of degradation's and transcriptional noise, respectively, together to form a SDE model, such as

- SDE model with Langevin noise in the process of gene transcription and protein degradation

$$d[\mathrm{P}] = (f(\mathbf{x}) - k^*[\mathrm{P}])dt + \mu_1\sqrt{b f(\mathbf{x})}dW_1(t) + \mu_2\sqrt{k^*[P]}dW_2(t), \tag{2.44}$$

where $W_1(t)$ and $W_2(t)$ are two independent Wiener processes.

## 2.4 Inference methods for genetic regulation

Genome-wide gene expression and kinase activity measurements have become feasible due to significant advancements in high-throughput technologies such as microarray gene expression data, DNA/RNA sequencing data, and single-cell sequencing data. Given the abundance of experimental data available, researchers hope to gain new insights and inspiration for unknown regulatory mechanisms from experimental observations. Therefore, finding an efficient inference method based on experimental data becomes a major challenge in computational biology and bioinformatics. There are many inference methods that have been proposed for inferring genetic regulatory networks. For those interested in this field, it is recommended to read these papers for more details [20, 83, 98, 110, 123, 129]. In this section, I only introduce two types of methods for inferring genetic regulatory networks that will be used in this thesis.

### 2.4.1 Static model

An inference method, which identifies the structure of regulatory networks from experimental data, is termed as a static model. Probabilistic graphical model is a useful tool as a static model to predict the regulatory network between different components in this system. A gene network is represented by a graph $G$ with a set of nodes (genes) $K$ and a set of edges $E$. The nodes of the graph are modelled as random variables and the edges represent the interaction between them [99]. There are two popular graphical model used for regulatory network inference.

One type of probabilistic graphical models is the Gaussian graphical model (GGM), which provides a simple and effective method to characterize the regulatory relationship between genes. The GGM is based on the calculation of the conditional dependencies among genes using the gene expression data. If a $n$-dimensional continuous random vector $\boldsymbol{X} \in \mathbb{R}^n$ ($n$ genes) has the probability density function

$$f(\boldsymbol{X}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\right\}, \tag{2.45}$$

we say that random vector $\boldsymbol{X}$ follows a multivariate Gaussian distribution $\boldsymbol{\mathcal{N}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with the mean parameter $\boldsymbol{\mu} \in \mathbb{R}^n$ and the positive definite covariance matrix $\boldsymbol{\Sigma}$ [140].

Given an undirected graph $G$ with a set of nodes (genes) $K$ and a set of edges $E$. The partial correlation between genes $i$ and $j$ is estimated to search for the best independence graph [147, 149]. The edge connecting two genes in the graph is neglected if they are conditionally independent given all other genes [64]. That is, a $n$-dimensional continuous random vector $\boldsymbol{X}$ ($n$ genes) with a multivariate normal distribution said to satisfy the

Gaussian graphical model with graph $G$, if $\boldsymbol{X}$ follows a multivariate Gaussian distribution $\boldsymbol{\mathcal{N}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$(\boldsymbol{\Sigma}^{-1})_{i,j} = 0 \text{ for all} (i,j) \notin E. \tag{2.46}$$

Therefore, the graph $G$ is described by the sparse matrix of $\boldsymbol{\Sigma}^{-1}$. Based on the theory, the GGM with Forward Search Algorithm (FSA) was proposed to infer the network structure [147]. Here, I give a brief description of the FSA, which is used in Chapter 3, as follows,

**Forward Search Algorithm (FSA)**

1. Let $X = (X_1, \ldots, X_N)$ be a vector with $N$ elements, $n$ be the number of genes. An initial empty graph $G$ is built by the identity matrix with $n$-dimensions.

2. An iterative maximum likelihood estimates algorithm [30] is used to compute the covariance matrix , $Cov(G)$, of the initial graph $G$ until the maximum deviance difference is almost unchanged. In our study, we set this threshold value as 0.0001.

3. An edge $E_i$ is added into the initial graph and then we compute a new covariance matrix, denoted as $Cov(E_i)$, by the iterative maximum likelihood estimates as step 2. The significance of the added edge $E_i$ is tested by the deviance difference, which is approximately Chi-square distribution with one degree of freedom. A p-value of the Chi-square test is used as the model selection criteria.

4. After all edges have been tested, all p-values are ranked by descending order. If the smallest p-value of an added edge $E_i$ is smaller than a predefined cutoff p-value (e.g. cutoff p-value $= 0.05$), the edge is added to the initial graph $G$, and then go back to step 2. Repeatedly computing from step 2 to step 4 until the ranked p-value of added edge is larger than the predefined cutoff p-value.

5. Based on the recording of added edges from step 4, the algorithm shows the sparse matrix of graphical model and the partial correlation coefficient matrix.

6. Finally, nodes represent genes; edges describe the regulatory relationship between a pair of nodes. Moreover, the sparse matrix elucidates the regulatory relationship existing between genes and the partial correlation coefficient matrix tells us regarding the positive or negative regulation between two components.

### 2.4.2 Dynamic model

An inference method, which examines the mechanisms through dynamical properties of interactions between network components, is termed as a dynamic model. For a network

with $N$ genes, the following dynamic model comprising all network components is used to capture the dynamical behavior of genetic regulation,

$$\frac{dX_i}{dt} = \frac{\sum_{j=1}^{N} a_{ij} X_j^{n_j}}{1 + \sum_{j=1}^{N} b_{ij} X_j^{n_j}} - k_i X_i \text{ for } i = 1, \cdots, N, \tag{2.47}$$

where $X_i$ denotes expression levels of the $i$-th gene and coefficient $a_{ij}$ denotes regulations from the $j$-th gene to the $i$-th gene. The regulation may be positive ($a_{ij} > 0$) or negative ($a_{ij} = 0$) if the corresponding coefficient ($b_{ij} > 0$). In addition, when $b_{ij} > 0$, it is assumed that the $i$-th gene can autoregulate itself if $a_{ii} > 0$. Moreover, if $a_{ij} = b_{ij} = 0$, there will be no regulation from the $j$-th gene to the $i$-th gene. Coefficients $n_j$ and $k_i$ are the Hill coefficient of gene $j$ and degradation rate of gene $i$, respectively.

As the dynamic model has many unknown parameters, it is generally defined that this inference method is a problem for estimating unknown parameters from experimental observations. Based on the estimated results, we can then infer the regulatory types and strengths of the unknown regulatory mechanisms.

## 2.5 Numerical methods for parameter estimation

In previous sections, I have introduced several models, both mathematical models that study dynamical properties based on known regulatory mechanisms and dynamic models that make predictions about unknown mechanisms based on experimental data. The parameter estimation of these models is an essential part to study the genetic regulations in detail. In this section, I will outline several existing numerical methods for parameter estimation.

### 2.5.1 Genetic algorithm

Genetic algorithms are an effective stochastic search method for finding unknown parameters of a mathematical model when the search space is associated with a complex error landscape. Given the gene expression data and developed mathematical model, the algorithm begins by generating a population of initial parameter values. Each initial value is called an individual and the whole population is called one generation. Then it calculates the fitness value for each individual of current generation. Highly fit individuals will have higher probability of being selected to take part in the next stage than the less fit individuals [22]. Based on the fitness values, the algorithm next creates new values for each individual by modifying the selected individuals and thus forms a population of the next generation. This process is repeated until a pre-defined number of generations have been calculated. In

Chapter 3 and Chapter 4, we employed the following functions in MATLAB to implement the Genetic algorithm for estimating the unknown dynamic model parameters, namely function *crtbp* to generate initially binary populations, function *reins* to effect fitness-based reinsertion, function *select* to give a convenient interface to the selection routines, function *recombine* to conduct crossover operators, and function *mut* to conduct binary and integer mutations. The detailed information of these functions and their alternatives can be found in the relevant reference [22].

### 2.5.2 Approximate Bayesian computation

Since the likelihood function reflects the probability of observed data fitting a particular statistical model, it quantifies the evidence supporting certain parameter values and models selection in all simulation-based statistical inferences. It is generally possible to find an analytic solution for the likelihood function for simple models. However, analytical solutions for likelihood functions are difficult or impossible to obtain for more complex models and may be prohibitively expensive to compute. Approximate Bayesian computation (ABC) is a computational methodology based on Bayesian statistics which can be used to estimate posterior distributions of model parameters. This method is highly useful for simulation-based models since it does not need to determine the likelihood function. The algorithm begins with a set of candidate model parameters, $\widehat{\theta}$, which is generated by Monte Carlo sampling techniques with a user-defined prior probability distribution, $\pi(\theta)$. Based on the mathematical model, the algorithm then use these candidate parameters to simulate the given gene expression data. Then, a distance function, $\rho(\boldsymbol{X}, \boldsymbol{X^*})$, such as absolute deviations and root mean squared errors is used to measure the similarity between the experimental data ($\boldsymbol{X}$) and the simulated data ($\boldsymbol{X^*}$). Finally, if the distance obtained is less than the pre-defined cutoff value ($\varepsilon_0$), we will keep this set of candidate model parameters. Otherwise, we will discard them, resample the parameters and repeat the steps until $\rho(\boldsymbol{X}, \boldsymbol{X^*}) < \varepsilon_0$. In Chapter 5, the ABC rejection-sampling algorithm is used to estimate the mathematical model parameters [10, 139]. The pseudo-code of this algorithm is shown as follows.

### 2.5.3 Robustness analysis

Robustness is described in biological systems as a system's capacity to operate appropriately in the midst of both internal and external uncertainty [27, 147]. If a model with the estimated parameters is not robust, a perturbation to the parameters might lead to substantial variations of the model output. Therefore, robustness has been more popular in recent years as a critical criterion for determining the ideal network structure or model parameters from estimated results [5, 86]. The robustness property of a mathematical model

**Algorithm 1:** ABC rejection-sampling algorithm

**Data:** Gene expression data
**Result:** Estimated model parameters
Given gene expression data ($\boldsymbol{X}$)
Set the prior distribution $\pi(\theta)$
Set the cutoff value $\varepsilon_0$
Define the distance function $\rho(\boldsymbol{X}, \boldsymbol{X^*})$
**for** $1 \leq i \leq N$ **do**
  **while** $\rho(\boldsymbol{X}, \boldsymbol{X^*}) > \varepsilon_0$ **do**
    Sample the candidate parameters $\widehat{\theta}$ from the prior distribution $\pi(\theta)$
    Solve mathematical models and generate simulated data $\boldsymbol{X^*}$ with $\widehat{\theta}$
    Calculate the distance $\rho(\boldsymbol{X}, \boldsymbol{X^*})$
  **end**
  Store $\theta_i \leftarrow \widehat{\theta}$
**end**

with respect to a set of perturbations $P$ is defined as the average of an evaluation function $D^S_{a,P}$ of the system over all perturbations $p \in P$. Here we propose that evaluation function is $\Sigma_{i,t} x_{it}(p)$, where $x_{it}(p)$ is the expression level of gene $i$ at time point $t$ with a specific perturbation $p \in P$. Thus, the average gene expression level over all perturbed model parameters is given by

$$R^M_{a,p} = \sum_{i,t} \int_{p \in P} \mathbb{P}(p) x_{it}(p) dp, \tag{2.48}$$

where $\mathbb{P}(p)$ is the probability of the perturbation $p$. In addition, the impact of perturbations on the system performance is evaluated by

$$R^N_{a,p} = \sum_{i,t} \int_{p \in P} \mathbb{P}(p)(x_{it} - x_{it}(p))^2 dp. \tag{2.49}$$

This average value should be close to the simulated gene expression levels $x_{it}$ obtained from the unperturbed rate constants. In perturbation test, based on the inferred parameter $k_i$ that is assumed to be the unperturbed one, the perturbed parameter is generated by

$$\overline{k_i} = k_i \times (1 + \mu \times \varepsilon), \tag{2.50}$$

where $\varepsilon$ is a sample generated from either the normal distribution or the uniform distribution. In addition, $\mu$ is the control parameter to determine the perturbation strength.

For a set of estimated parameters, we can firstly obtained $N$ sets of perturbed model parameters by using (2.50) and then use these perturbed parameter sets to obtain $N$ corre-

sponding simulations. We used $x_{it}^{(k)}(p)$ and $x_{it}^{(k)}$ to denote the simulation value of variable $x_i$ $(1 \leq i \leq m)$ at time point $t$ $(1 \leq t \leq M)$ obtained by the $k$-th perturbed and unperturbed model parameters, respectively. Then, we defined

$$E^{(k)} = \sqrt{\sum_{i=1}^{m}\sum_{t=1}^{M}(x_{it}^{(k)}(p) - x_{it}^{(k)})^2} \tag{2.51}$$

as the measure for the robustness property of the model with the $k$-th perturbed parameter set. Afterwards, we defined the robust average for the given parameter set as

$$RA = \frac{1}{N}\sum_{k=1}^{N} E^{(k)}, \tag{2.52}$$

and robust standard deviation as

$$RSTD = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(E^{(k)} - RA)^2} \tag{2.53}$$

over $N$ perturbation tests. Smaller values of RA and RSTD mean that the model with the given parameter set is more robust. In Chapter 3 and Chapter 4, we use robustness property (2.52) and (2.53) of the model to select the optimal model parameter sets from estimated candidates.

## 2.6 Numerical methods for simulation of mathematical models

We always try to describe complex and nonlinear systems, such as genetic regulatory systems, with some complicated mathematical models. It is either time-consuming or impossible to find analytical solutions for most nonlinear deterministic or stochastic models. In this case, we can use appropriate numerical methods to find solutions to these mathematical models within acceptable approximation error ranges to further simulate and analyze the mechanism of the described system. In this section, I will introduce some classical methods for simulating ordinary differential equations and stochastic differential equations, respectively.

### 2.6.1 Numerical simulation of ODE models

This subsection, based on textbooks written by Chapra [19] and Quarteroni et al. [111], is devoted to solving ODEs of the form

$$\frac{dy}{dt} = f(t, y). \tag{2.54}$$

Recalling the first principle of differentiation, the rate of change (slope) for a function is defined as

$$\frac{dy}{dt} = \lim_{\Delta t \to 0} \frac{\Delta y}{\Delta t} = \lim_{\Delta t \to 0} \frac{y(t_{i+1}) - y(t_i)}{t_{i+1} - t_i}, \tag{2.55}$$

where $\Delta y = y(t_{i+1}) - y(t_i)$ and $\Delta t = t_{i+1} - t_i$ are differences in values of $y$ and time $t$ computed over finite intervals, $y(t_i)$ is the value of $y$ at time $t_i$, and $y(t_{i+1})$ is the value of $y$ at one-step later time $t_{i+1}$. If $\Delta t$ is small enough, the slope can be approximated by differences over finite intervals. That is,

$$\frac{dy}{dt} \approx \frac{\Delta y}{\Delta t} = \frac{y(t_{i+1}) - y(t_i)}{t_{i+1} - t_i}. \tag{2.56}$$

Thus, the equation (2.56) can be rewritten more concisely as

$$y(t_{i+1}) = y(t_i) + \frac{dy}{dt}\Delta t, \tag{2.57}$$

$$= y(t_i) + \Phi \Delta t, \tag{2.58}$$

where the slope $\frac{dy}{dt} = \Phi$ is called an increment function. From this general form, we can easily determine the future value $y(t_{i+1})$ if we know the previous value $y(t_i)$, the time step $\Delta t$ and the increment function $\Phi$. These approaches, which determine a future value by previous value at a single time point, are called the Runge-Kutta methods. The Euler's method is the simplest Runge-Kutta method, which Leonhard Euler described in his book and published in 1768 [36].

#### 2.6.1.1 Euler's method

Euler's method is a first order numerical method for solving ODEs with a given initial value. Let the increment function $\Phi$ is provided by the ODE evaluated at $t_i$ and $y(t_i)$. That is,

$$\Phi = f(t_i, y(t_i)). \tag{2.59}$$

Then, we will have the following formula of Euler's method,

$$y(t_{i+1}) = y(t_i) + f(t_i, y(t_i))\Delta t. \tag{2.60}$$

Euler's method uses the slope and old value at the beginning of the time interval $t_i$ to find the new value $y(t_{i+1})$ at the end of the time interval $t_{i+1}$. The pseudocode of Euler's method is given below.

**Algorithm 2:** Euler's method

Given the increment function $\Phi = \frac{dy}{dt} = f(t, y)$
Given the initial value $y(t_0)$, finite time span $t = [t_0, t_T]$ and time step $dt$
Calculate the number of iterations $n = \frac{t_T - t_0}{dt}$
Set a result vector $\boldsymbol{Y}$ with $n$ components and let $\boldsymbol{Y}(1) = y(t_0)$
Set $t_{old} = t_0$, $steps = 1$ and $y_{old} = \boldsymbol{Y}(1) = y(t_0)$
**for** $2 \leq i \leq n$ **do**
    $y_{new} = y_{old} + dt * f(t_{old}, y_{old})$
    $steps = steps + 1$
    $t_{old} = t_{old} + dt$
    Store $\boldsymbol{Y}(steps) \leftarrow y_{new}$
    $y_{old} = y_{new}$
**end**
Display $\boldsymbol{Y}$ as the trajectory of the solution over the time span $t = [t_0, t_T]$

There are two types of errors for the numerical solution of Euler's method. One is the round-off error resulting from the limited significant figures that can be retained by a computer. The second is a global truncation error (GTE) that consists of a local truncation error (LTE) and a propagated truncation error. The local truncation error is the error produced by a single step approximation for the value of $y$ that is the difference between the numerical solution $y(t_{i+1})$ and the exact solution at time $t_{i+1}$. Furthermore, the propagated truncation error is caused by the approximation during the preceding steps. Thus, the global truncation error is the sum of these errors, which is the cumulative effect of the local truncation errors determined at each step over the whole time span. To find the local truncation error, we need to find the exact solution at time $t_{i+1}$. First, we consider the Taylor expansion of the function $y$ around time $t_i$

$$y(t_{i+1}) = y(t_i) + \Delta t y'(t_i) + (\Delta t)^2 y''(t_i) + \mathcal{O}((\Delta t)^3). \tag{2.61}$$

Moreover, the numerical solution of $y(t_{i+1})$ by Euler's method is $y(t_{i+1}) = y(t_i) + \Delta t y'(t_i)$.

Thus,

$$\begin{aligned} \text{LTE} &= \text{Exact solution} - \text{Numerical solution}, && (2.62)\\ &= y(t_{i+1}) - y(t_i) - \Delta t y'(t_i), && (2.63)\\ &= (\Delta t)^2 y''(t_i) + \mathcal{O}((\Delta t)^3), && (2.64)\\ &= \mathcal{O}((\Delta t)^2). && (2.65) \end{aligned}$$

It is clear to show that the LTE is proportional to $(\Delta t)^2$. In addition, the number of iteration over the time span $t = [t_0, t_T]$ is determined by $\frac{t_T - t_0}{\Delta t}$, which is proportional to $\frac{1}{\Delta t}$. Therefore, the GTE is proportional to $\Delta t$, that is, GTE $= \mathcal{O}(\Delta t)$. Moreover, this conclusion tell us that we can reduce the LTE/GTE of Euler's method by decreasing the time step size, $\Delta t$, and if the solution of the differential equation (increment function) is linear, the Euler's method will give an exact solution for $y(t_{i+1})$.

Noted that Euler's method provides an exact solution if the increment function is a first-order polynomial function (linear function), this is why Euler's method is also referred to as a first-order method. In general, an $n$th-order method will provide an exact solution if its increment function is an $n$th-order polynomial function, and the LTE and GTE of an $n$th-order method are $\mathcal{O}((\Delta t)^{n+1})$ and $\mathcal{O}((\Delta t)^n)$, respectively. In this thesis, we use the Euler's method to solve all differential equations models.

#### 2.6.1.2 Runge-Kutta methods

A key rationale behind the application of Euler's method is that the derivative at the time point, $y_{t_i}$, is assumed to be unchanged over the entire one-step time interval, $[y_{t_i}, y_{t_{i+1}}]$. Thus, in addition to decreasing the $\Delta t$, we can also improve Euler's method by applying other methods to estimate the derivative over the time interval. These methods belong to a family of iterative methods called Runge-Kutta methods.

The general form of Runge-Kutta methods is the following equation

$$y(t_{i+1}) = y(t_i) + \Phi \Delta t \tag{2.66}$$

with the increment function

$$\Phi = a_1 k_1 + a_2 k_2 + \cdots + a_n k_n, \tag{2.67}$$

where the coefficients $a$ are constants and the coefficients $k$ are

$$k_1 = f(t_i, y(t_i)), \tag{2.68}$$
$$k_2 = f(t_i + p_2\Delta t, y(t_i) + q_{2,1}k_1\Delta t), \tag{2.69}$$
$$k_3 = f(t_i + p_3\Delta t, y(t_i) + (q_{3,1}k_1 + +q_{3,2}k_2)\Delta t), \tag{2.70}$$
$$\vdots$$
$$k_n = f(t_i + p_n\Delta t, y(t_i) + (q_{n,1}k_1 + q_{n,2}k_2 + \cdots + q_{n,n-1}k_{n-1})\Delta t), \tag{2.71}$$

where the coefficients $p$ and $q$ are constants. All coefficients $a$, $p$ and $q$ are determined by setting Eq. (2.66) with $n$th-order equal to the corresponding $n$th-order Taylor expansion. Coefficient $k_1$ is obtained from the ODE evaluated at $t_i$ and $y(t_i)$, then the remaining coefficients $k$ are determined by recurrence relationships. I will then introduce the various types of Runge-Kutta methods that are most commonly used by researchers. The detailed derivations of these methods are beyond the scope of this thesis, which we will not discuss here.

1. **First-order Runge-Kutta methods:** When $n = 1$ in the increment function, we can derive the following equation from the Eq. (2.66),

$$y(t_{i+1}) = y(t_i) + a_1 f(t_i, y(t_i))\Delta t \tag{2.72}$$

   If $a_1$ is assumed to be 1, we can get the formula for Euler's method. The LTE and GTE of the first-order Runge-Kutta methods are $\mathcal{O}((\Delta t)^2)$ and $\mathcal{O}(\Delta t)$, respectively. As we discussed in the last subsection, the first-order method will provide an exact solution if the increment function is linear

2. **Second-order Runge-Kutta methods:** When $n = 2$ in the increment function, we can derive the following equation from the Eq. (2.66),

$$y(t_{i+1}) = y(t_i) + (a_1k_1 + a_2k_2)\Delta t, \tag{2.73}$$

   where

$$k_1 = f(t_i, y(t_i)), \tag{2.74}$$
$$k_2 = f(t_i + p_2\Delta t, y(t_i) + q_{2,1}k_1\Delta t). \tag{2.75}$$

   By setting Eq. (2.73) equal to the corresponding second-order Taylor expansion, we can have the following methods with different combinations of coefficients.

- When $a_1 = a_2 = \frac{1}{2}$ and $p_2 = q_{2,1} = 1$, we have

$$y(t_{i+1}) = y(t_i) + (\frac{1}{2}k_1 + \frac{1}{2}k_2)\Delta t, \tag{2.76}$$

where

$$k_1 = f(t_i, y(t_i)), \tag{2.77}$$
$$k_2 = f(t_i + \Delta t, y(t_i) + k_1\Delta t). \tag{2.78}$$

This method is known as Heun's method. Since $k_1$ and $k_2$ are the derivative evaluated at the beginning and end of the time interval $[y_{t_i}, y_{t_{i+1}}]$, respectively, the key rationale behind the application of Heun's method is that the increment function is assumed to be the average of start and end gradients across the whole time interval. The pseudocode of Heun's method is given below.

**Algorithm 3:** Heun's method

```
Given the function f(t, y)
Given the initial value y(t_0), finite time span t = [t_0, t_T] and time step dt
Calculate the number of iterations n = (t_T − t_0)/dt
Set a result vector Y with n components and let Y(1) = y(t_0)
Set t_old = t_0, steps = 1 and y_old = Y(1) = y(t_0)
for 2 ≤ i ≤ n do
    k_1 = f(t_old, y_old)
    k_2 = f(t_old + dt, y_old + k_1 * dt)
    y_new = y_old + (dt/2) * (k_1 + k_2)
    steps = steps + 1
    t_old = t_old + dt
    Store Y(steps) ← y_new
    y_old = y_new
end
Display Y as the trajectory of the solution over the time span t = [t_0, t_T]
```

- When $a_1 = 0$, $a_2 = 1$ and $p_2 = q_{2,1} = \frac{1}{2}$, we have

$$y(t_{i+1}) = y(t_i) + k_2\Delta t, \tag{2.79}$$

where

$$k_1 = f(t_i, y(t_i)), \tag{2.80}$$
$$k_2 = f(t_i + \frac{\Delta t}{2}, y(t_i) + k_1\frac{\Delta t}{2}). \tag{2.81}$$

This method is known as the Midpoint method. Since $k_2$ is the derivative evaluated at the midpoint of the time interval $[y_{t_i}, y_{t_{i+1}}]$, the key rationale behind the application of the Midpoint method is that the increment function is assumed to be the midpoint gradient across the whole time interval.

The LTE and GTE of the second-order Runge-Kutta methods are $\mathcal{O}((\Delta t)^3)$ and $\mathcal{O}((\Delta t)^2)$, respectively. Therefore, the second-order method will provide an exact solution if the increment function is quadratic. The pseudocode of the Midpoint method is given below.

**Algorithm 4:** The Midpoint method

Given the function $f(t, y)$
Given the initial value $y(t_0)$, finite time span $t = [t_0, t_T]$ and time step $dt$
Calculate the number of iterations $n = \frac{t_T - t_0}{dt}$
Set a result vector $\boldsymbol{Y}$ with $n$ components and let $\boldsymbol{Y}(1) = y(t_0)$
Set $t_{old} = t_0$, $steps = 1$ and $y_{old} = \boldsymbol{Y}(1) = y(t_0)$
**for** $2 \leq i \leq n$ **do**
    $k_1 = f(t_{old}, y_{old})$
    $k_2 = f(t_{old} + \frac{dt}{2}, y_{old} + k_1 * \frac{dt}{2})$
    $y_{new} = y_{old} + dt * k_2$
    $steps = steps + 1$
    $t_{old} = t_{old} + dt$
    Store $\boldsymbol{Y}(steps) \leftarrow y_{new}$
    $y_{old} = y_{new}$
**end**
Display $\boldsymbol{Y}$ as the trajectory of the solution over the time span $t = [t_0, t_T]$

3. **Fourth-order Runge-Kutta methods:** When $n = 4$ in the increment function, we can derive the following equation from the Eq. (2.66),

$$y(t_{i+1}) = y(t_i) + (a_1 k_1 + a_2 k_2 + a_3 k_3 + a_4 k_4)\Delta t, \tag{2.82}$$

where

$$k_1 = f(t_i, y(t_i)), \tag{2.83}$$
$$k_2 = f(t_i + p_2 \Delta t, y(t_i) + q_{2,1} k_1 \Delta t), \tag{2.84}$$
$$k_3 = f(t_i + p_3 \Delta t, y(t_i) + (q_{3,1} k_1 + q_{3,2} k_2)\Delta t), \tag{2.85}$$
$$k_4 = f(t_i + p_4 \Delta t, y(t_i) + (q_{4,1} k_1 + q_{4,2} k_2 + q_{4,3} k_3)\Delta t). \tag{2.86}$$

The most commonly used method in the Runge-Kutta family is the classical fourth-order Runge-Kutta method with $a_1 = a_4 = \frac{1}{6}$, $a_2 = a_3 = \frac{1}{3}$, $p_2 = p_3 = q_{2,1} = q_{3,2} = \frac{1}{2}$,

$q_{3,1} = q_{4,1} = q_{4,2} = 0$ and $p_4 = q_{4,3} = 1$. That is,

$$y(t_{i+1}) = y(t_i) + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)\Delta t, \tag{2.87}$$

where

$$k_1 = f(t_i, y(t_i)), \tag{2.88}$$

$$k_2 = f(t_i + \frac{1}{2}\Delta t, y(t_i) + \frac{1}{2}k_1\Delta t), \tag{2.89}$$

$$k_3 = f(t_i + \frac{1}{2}\Delta t, y(t_i) + \frac{1}{2}k_2\Delta t), \tag{2.90}$$

$$k_4 = f(t_i + \Delta t, y(t_i) + k_3\Delta t). \tag{2.91}$$

The key rationale behind the application of the classical fourth-order Runge-Kutta method is similar to the Heun's method in that the increment function is assumed to be an improved average gradient by multiple estimates across the whole time interval. In addition, the LTE and GTE of the fourth-order Runge-Kutta methods are $\mathcal{O}((\Delta t)^5)$ and $\mathcal{O}((\Delta t)^4)$, respectively. Therefore, the fourth-order method will provide an exact solution if the increment function is a 4th-order polynomial. The pseudocode of the classical fourth-order Runge-Kutta method is given below.

**Algorithm 5:** The classical fourth-order Runge-Kutta method

Given the function $f(t, y)$
Given the initial value $y(t_0)$, finite time span $t = [t_0, t_T]$ and time step $dt$
Calculate the number of iterations $n = \frac{t_T - t_0}{dt}$
Set a result vector $\boldsymbol{Y}$ with $n$ components and let $\boldsymbol{Y}(1) = y(t_0)$
Set $t_{old} = t_0$, $steps = 1$ and $y_{old} = \boldsymbol{Y}(1) = y(t_0)$
**for** $2 \leq i \leq n$ **do**
    $k_1 = f(t_{old}, y_{old})$
    $k_2 = f(t_{old} + \frac{dt}{2}, y_{old} + k_1 * \frac{dt}{2})$
    $k_3 = f(t_{old} + \frac{dt}{2}, y_{old} + k_2 * \frac{dt}{2})$
    $k_4 = f(t_{old} + dt, y_{old} + k_3 * dt)$
    $y_{new} = y_{old} + \frac{dt}{6} * (k_1 + 2k_2 + 2k_3 + k_4)$
    $steps = steps + 1$
    $t_{old} = t_{old} + dt$
    Store $\boldsymbol{Y}(steps) \leftarrow y_{new}$
    $y_{old} = y_{new}$
**end**
Display $\boldsymbol{Y}$ as the trajectory of the solution over the time span $t = [t_0, t_T]$

Depending on the choice of order $n$, we can also derive other Runge-Kutta methods. The different methods are similar in concept in terms of the solution of the coefficients and

follow the pattern we discussed before for both LTE and GTE. Higher-order Runge-Kutta methods are certainly more accurate, but the number of steps required for computation is also greater, thus reducing the computational efficiency. Therefore, when using these methods to solve ODEs, we should balance accuracy and computational efficiency to choose a more appropriate method, rather than pursuing more complex and advanced methods.

### 2.6.2 Numerical simulation of SDE models

This subsection is devoted to solving the Itô SDE driven by a one-dimensional Wiener process of the form

$$dX(t) = \mu(X(t), t)dt + \sigma(X(t), t)dW(t) \tag{2.92}$$

for $0 \leq t \leq T$ with initial condition $X(0) = X_0 \leq \infty$. In this SDE, functions $\mu(X(t), t)$ and $\sigma(X(t), t)$ are called the drift and the diffusion coefficient, respectively; $X(t)$ is an arbitrary process and $W(t)$ is a Wiener process. If for all $t >\geq 0$ we have $\int_0^t |\mu(X(s), s)|ds < \infty$ and $\int_0^t \sigma^2(X(s), s)ds < \infty$, then the solution of the Itô SDE is given by

$$X(t) = X_0 + \int_0^t \mu(X(s), s)ds + \int_0^t \sigma(X(s), s)dW(s). \tag{2.93}$$

To find the approximate solution of the Itô SDE, it is useful to discretize the SDE by dividing the time interval $[0, T]$ into $n$ equal subintervals of width $\Delta t = \frac{T}{n} > 0$, such as

$$0 = t_0 < t_1 < t_2 < t_3 < \cdots < t_{n-1} < t_n = T. \tag{2.94}$$

Then, the following two methods are introduced to solve this SDE, namely Euler-Maruyama method and Milstein method. I will not give a comprehensive introduction here on the theoretical foundations behind these methods. A rigorous description and explanation of these underlying theories can be found in Kloeden and Platen's textbook [61] and Higham's paper [49]. The methods presented in this subsection are also based on these references.

#### 2.6.2.1 Euler-Maruyama method

The simplest numerical method for solving the Itô SDE is the Euler-Maruyama (EM) method. EM method is also referred to as the explicit Euler method, which is an extension of the Euler method for ODEs. Since the time interval is discretized, then the SDE can be solved by evaluating the $X(t)$ at each time point recurrently if the initial condition is

given. That is,

$$X(t_{j+1}) = X_{t_j} + \int_{t_j}^{t_{j+1}} \mu(X(s), s)ds + \int_{t_j}^{t_{j+1}} \sigma(X(s), s)dW(s). \tag{2.95}$$

The EM method approximates the two integrals as follows:

$$\int_{t_j}^{t_{j+1}} \mu(X(s), s)ds \approx \mu(X(t_j), t_j)(t_{j+1} - t_j) = \mu(X(t_j), t_j)\Delta t, \tag{2.96}$$

$$\int_{t_j}^{t_{j+1}} \sigma(X(s), s)dW(s) \approx \sigma(X(t_j), t_j)(W(t_{j+1}) - W(t_j)) = \sigma(X(t_j), t_j)\Delta W(t_j) \tag{2.97}$$

Then, for $0 \leq j \leq n-1$, the EM method takes the form of

$$X(t_{j+1}) = X_{t_j} + \mu(X(t_j), t_j)\Delta t + \sigma(X(t_j), t_j)\Delta W(t_j). \tag{2.98}$$

Noted that if $\sigma(X(t_j).t_j) = 0$ and $X_{t_j}$ is a constant, the EM method reduces to Euler's method for solving ODEs. Since the increment of the Wiener process, $DeltaW(t_j) = (W(t_{j+1}) - W(t_j))$, is normally distributed with mean 0 and variance $\Delta t$, the independent standard Gaussian pseudorandom number generator *randn* in the MATLAB can be used to generate the increment of Wiener process $\Delta W(t_j)$ at each time point by setting $\Delta W(t_j) = \sqrt{\Delta t} * randn$. The pseudocode of the EM method is given below.

**Algorithm 6:** The Euler-Maruyama method

```
Set an arbitrary random seed value R for repeated experiments randn('seed', R)
Given the function μ(x,t) and σ(x,t)
Given the initial value x_0, finite time span t = [0,T] and time step dt
Calculate the number of iterations n = T/dt
Set a result vector X with n components and let X(1) = x_0
Set t_old = 0, steps = 1 and x_old = X(1) = x_0
for 2 ≤ i ≤ n do
    x_new = x_old + μ(x_old, t_old) * dt + σ(x_old, t_old) * sqrt(dt) * randn
    steps = steps + 1
    t_old = t_old + dt
    Store X(steps) ← x_new
    x_old = x_new
end
Display X as the trajectory of the solution over the time span t = [0,T]
```

There are also two other methods based on the similar idea, the only difference being that these two methods use different time points to evaluate functions $\mu(X(t_j), t_j)$ and $\sigma(X(t_j), t_j)$

1. **The semi-implicit Euler method:**

$$X(t_{j+1}) = X_{t_j} + \mu(X(t_{j+1}), t_{j+1})\Delta t + \sigma(X(t_j), t_j)\Delta W(t_j). \tag{2.99}$$

2. **The implicit Euler method:**

$$X(t_{j+1}) = X_{t_j} + \mu(X(t_{j+1}), t_{j+1})\Delta t + \sigma(X(t_{j+1}), t_{j+1})\Delta W(t_j). \tag{2.100}$$

In the EM and semi-implicit Euler methods, the left-endpoint value at $t_j$ is used to evaluate the integral $\int_{t_j}^{t_{j+1}} \sigma(X(s), s)dW(s)$ based on the definition of the Itô integral. Therefore, the numerical solutions of these two methods converge to exact solution of the Itô SDE. In contrast, in the implicit Euler method, the right-endpoint value at $t_{j+1}$ is used to evaluate the integral $\int_{t_j}^{t_{j+1}} \sigma(X(s), s)dW(s)$ based on the definition of backward stochastic integrals. Therefore, the numerical solution of this methods converge to exact solution of the right-endpoint SDE rather than the Itô SDE [134]. Even though EM is the simplest method for solving SDE, this method is only strongly convergent with order 0.5 if the drift and diffusion functions $\mu(X(t_j), t_j)$ and $\sigma(X(t_j), t_j)$ satisfy appropriate conditions (see detail in [61]). It implies that if we want to decrease the approximation error by 10 times, we need to make the time step size 100 times smaller than before. As a result, the computation time will be 100 times greater than before.

#### 2.6.2.2 Milstein method

Milstein method is a numerical scheme for solving SDE, which is strongly convergent with order 1 if the drift and diffusion functions $\mu(X(t_j), t_j)$ and $\sigma(X(t_j), t_j)$ satisfy some appropriate conditions. That means if we want to decrease the approximation error by 10 times, we only need to make the time step size 10 times smaller than before. The Milstein method is obtained by truncating the stochastic Taylor series in the form of

$$\begin{aligned} X(t_{j+1}) =& X_{t_j} + \mu(X(t_j), t_j)\Delta t + \sigma(X(t_j), t_j)\Delta W(t_j) \\ &+ \frac{1}{2}\sigma(X(t_j), t_j)\sigma'(X(t_j), t_j)((\Delta W(t_j))^2 - \Delta t). \end{aligned} \tag{2.101}$$

The pseudocode of the Milstein method is implemented in a similar way as the EM method. See the detail below.

**Algorithm 7:** Milstein method

Set an arbitrary random seed value $R$ for repeated experiments randn('seed', $R$)
Given the function $\mu(x,t)$ and $\sigma(x,t)$
Find the derivative of $\sigma(x,t)$, $\sigma'(x,t)$
Given the initial value $x_0$, finite time span $t=[0,T]$ and time step $dt$
Calculate the number of iterations $n=\frac{T}{dt}$
Set a result vector $\boldsymbol{X}$ with $n$ components and let $\boldsymbol{X}(1)=x_0$
Set $t_{old}=0$, $steps=1$ and $x_{old}=\boldsymbol{X}(1)=x_0$
**for** $2\leq i\leq n$ **do**
    $dw=\sqrt{dt}*randn$
    $x_{new}=$
    $x_{old}+\mu(x_{old},t_{old})*dt+\sigma(x_{old},t_{old})*dw+\frac{1}{2}*\sigma(x_{old},t_{old})*\sigma'(x_{old},t_{old})*((dw)^2-dt)$
    $steps=steps+1$
    $t_{old}=t_{old}+dt$
    Store $\boldsymbol{X}(steps)\leftarrow x_{new}$
    $x_{old}=x_{new}$
**end**
Display $\boldsymbol{X}$ as the trajectory of the solution over the time span $t=[0,T]$

## 2.7 Review of mathematical modelling in hematopoiesis

As we discussed in Section 2.2, the genetic module *GATA1-GATA2-PU.1* is significant in the process of lineage specification of HSC. Therefore, several mathematical models have been proposed to investigate the interactions within this module. This section presents a brief review of existing models for studying genetic regulation in hematopoiesis.

The first mathematical model to study the regulatory mechanism of genes *GATA1* and *PU.1* is proposed in the form of the Shea–Ackers formalism [114]. The assumption of the model is based on the experimental observations as outlined below:

1. Both *GATA1* and *PU.1* positively regulate self-transcription and activate transcription of each other in the form of GATA1 and PU.1 homodimers.

2. The formation of *GATA1/PU.1* heterodimers inhibits both *GATA1* and *PU.1* expression.

Moreover, some regulatory effects are ignored in the model studies for simplicity, such as post-transcriptional regulation, time lags and interactive regulation with other TFs. We assume that $x$ and $y$ are the molecular concentrations of *GATA1* and *PU.1*, respectively,

and the model is proposed in a dimensionless form as

$$\frac{dx}{dt} = \frac{sx^2 + uk_u y^2}{1 + x^2 + k_u y^2 + k_r xy} - x, \tag{2.102}$$

$$\frac{dy}{dt} = \frac{sy^2 + uk_u x^2}{1 + k_u x^2 + y^2 + k_r xy} - y. \tag{2.103}$$

Detailed descriptions of this model can be found in the referenced paper [114]. This model sheds new light on the mechanisms underlying HSCs differentiation. In case the strength of each regulation type varies (i.e., the parameters vary), the system may have a different number of stable solutions. The model also indicates that if a gene is over-expressed suddenly at a stable stable, the entire system can be shifted from one stable state to another. As a first mathematical model of *GATA1-PU.1* module,

Hill equation has also been used to study the cell fate determination of HSCs. Based on the double-negative feedback loop of *GATA1* and *PU.1* with positive autoregulations, Huang et al. proposed the following model

$$\frac{dx}{dt} = a_1 \frac{x^n}{\theta_1^n + x^n} + b_1 \frac{\theta_2^n}{\theta_2^n + y^n} - k_1 x, \tag{2.104}$$

$$\frac{dy}{dt} = a_2 \frac{y^n}{\theta_3^n + y^n} + b_2 \frac{\theta_4^n}{\theta_4^n + x^n} - k_2 y, \tag{2.105}$$

where $x$ and $y$ represent expression levels of *GATA1* and *PU.1*, respectively; $a_{1,2}$, $b_{1,2}$ and $\theta_{1,2,3,4}$ are non-negative parameters. In addition, $k_1$ and $k_2$ are degradation rates of *GATA1* and *PU.1*, respectively. This model suggested that there are two stable states and one unstable state for some parameter values. The unstable state can be found at the boundary between the two basins of attraction of two stable states. The unstable state can be seen as a progenitor cell capable of choosing MEP lineage or GMP lineage. This binary state choice of the system also provided me with great inspiration for my later work in Chapter 5. While we can easily design a binary choice model, is it possible to construct a higher-order multistable model by embedding multiple bistable systems so that we can more accurately describe a larger system? In Chapter 5, I will introduce a novel framework for constructing multistable systems.

Tian and Smith-Miles proposed a mathematical model for *GATA1-GATA2-PU.1* module based on the Shea–Ackers formalism [136]. This is the first stochastic model to study the mechanism of GATA-switching and the function of noise in cell fate determination of HSCs. The model is based on the regulatory mechanism shown in Figure 2.4. The detailed assumptions can be found in the referenced paper [136]. Let $x$, $y$ and $z$ be the molecular concentrations of *GATA1*, *GATA2* and *PU.1*, respectively. To describe the mechanisms of GATA-switching, Tian introduced an additional rate constant $k_2^*$ over a time interval

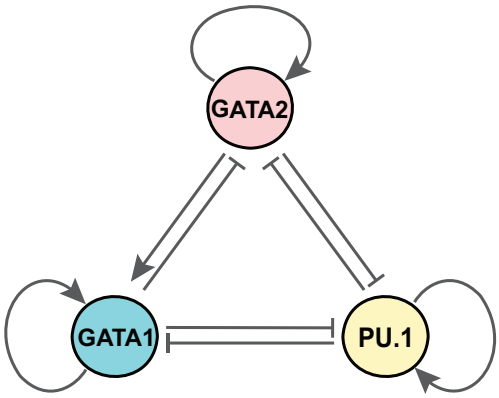


Figure 2.4: **The network structure of GATA1-GATA2-PU.1.** '→' and '⊣' denote the activating and inhibiting regulations, respectively.

$[t_1, t_2]$ for the displacement rate of *GATA2* proteins during the process of GATA-switching. Since the displacement of *GATA2* protein increasing, the concentration of *GATA1* proteins around the binding site will increase proportionally to $k_2^*$. Hence, the rate $\mu k_2^* y$ for the increase of *GATA1* during GATA-switching, where $\mu$is a control parameter to adjust the availability of *GATA1* proteins around chromatin sites. Then, the mathematical model is proposed as follows

$$\frac{dx}{dt} = \frac{a_1 x + a_2 y}{a_3 + a_4 x + a_5 y + a_6 z + a_7 xy} - k_1 x + \mu k_2^* y, \tag{2.106}$$

$$\frac{dy}{dt} = \frac{b_1 y}{b_2 + b_3 x + b_4 y + b_5 z + b_6 yz} - k_2 y - k_2^* y, \tag{2.107}$$

$$\frac{dz}{dt} = \frac{c_1 z}{c_2 + c_3 x + c_4 y + c_5 z + c_6 xz + c_6 yz} - k_3 z. \tag{2.108}$$

The simulation result shows that the model successfully realize the tristability if the model parameters satisfy some necessary conditions. However, the deterministic model cannot describe the heterogeneity in cell fate determination of HSCs. Therefore, Tian also introduced the following stochastic model by using Poisson $\tau$-leap method [135].

$$x(t+\tau) = x(t) + \mathbb{P}\left[\frac{a_1 x + a_2 y}{a_3 + a_4 x + a_5 y + a_6 z + a_7 xy}\tau\right] - \mathbb{P}[k_1 x\tau] + \mathbb{P}[\mu k_2^* y\tau], \tag{2.109}$$

$$y(t+\tau) = y(t) + \mathbb{P}\left[\frac{b_1 y}{b_2 + b_3 x + b_4 y + b_5 z + b_6 yz}\tau\right] - \mathbb{P}[k_2 y\tau] - \mathbb{P}[k_2^* y\tau], \tag{2.110}$$

$$z(t+\tau) = z(t) + \mathbb{P}\left[\frac{c_1 z}{c_2 + c_3 x + c_4 y + c_5 z + c_6 xz + c_6 yz}\tau\right] - \mathbb{P}[k_3 z\tau]. \tag{2.111}$$

In this stochastic model, the molecular concentrations of three genes are treated as numbers of molecular copies in the differentiation process. The model successfully revealed a range of

cell proportions resulting in a variety of differentiation pathways. This is also the first time a discrete stochastic model has been used to simulate the cell fate determination of HSCs with a multimodal distribution, and it also gives testable predictions about the mechanisms behind the realization of distinct differentiation lineages.

In summary, mathematical modelling is a powerful tool to accurately describe the dynamics of hematopoiesis and to explore the regulatory mechanisms for controlling the transitions between different cell types. Beyond the models described above, many more have been used to study the regulatory mechanism of hematopoiesis [33, 92, 96]. In addition, bifurcation theory is also an efficient method to explore the mechanisms of *GATA1-PU.1* module [13]. Moreover, the underlying mechanisms of how the stem/progenitor cells leave the stable steady states and commit to a specific lineage were also revealed with the assistance of mathematical models [95]. At the single-cell level, mathematical modelling and inference methods also helped to reconstruct the genetic regulatory network in hematopoiesis from the experimental data [47, 102]. Moreover, mathematical models have also been used to study the dynamical properties of blood diseases such as periodic hematological disorders and leukemia. [82, 127].

# 3

# Forward search algorithm for inferring genetic regulatory networks

The objective of this chapter is to study a framework for inferring the detailed dynamical mechanism of genetic regulatory networks. Inference of genetic networks is an important task to explore and predict the regulatory mechanism inside the cell. Although a number of algorithms have been designed to reverse-engineer regulatory networks effectively, it is still a challenge to introduce nonlinearity into dynamic models effectively. Recently, Wang and his team have proposed the probabilistic graphical models for microarray data analysis [145]. Based on this method, in this chapter, we introduce a novel framework for inferring genetic networks with nonlinearity to address these issues [153]. A new dynamic model using ordinary differential equations with exponential function is introduced to understand the nonlinearity. Using hematopoietic stem cell fate determination as a test problem, this work successfully constructs two networks for erythroid and granulocyte differentiation respectively, each of which involves 11 genes. Numerical results suggest that our new framework is able to provide accurate realizations of the system states. This work provides new ideas to infer regulatory networks effectively and explore novel regulatory mechanisms. In this chapter, we first introduce experimental data and selected candidate genes, then reconstruct the genetic regulatory networks in fate determination of HSCs by Forward Search Algorithm as the top-down approach, and we also introduce the nonlinear dynamic model in genetic regulation as the bottom-top approach. To reduce the number of unknown parameters, we combine both top-down and bottom-up approaches together. Finally we present numerical results.

## 3.1 Experimental data

### 3.1.1 Database background

In this work, we used the sub-series GSE49987 as the experimental data from the published microarray dataset GSE49991 [88]. This dataset contains the expression profiles collected by experiments using the cell line FDCPmix. This dataset was generated with the probe name version of Agilent Whole Mouse Genome Microarray $4 \times 44K$ [88]. It provides microarray gene expression profiles of hematopoietic stem cells (HSCs) differentiating into erythrocytes and neutrophils. To convert all microarray probe IDs to gene names, we pre-processed this dataset based on the Ensembl BioMart and GO Enrichment Analysis [131]. From a previous study, the regulatory network of 18 core genes during the hematopoiesis has been curated [93]. Moreover, the same research team studied the regulatory interaction of 26 core genes during the hematopoiesis [94]. The total number of distinct genes in these two studies is 30. Thus, in our work we considered 30 genes whose names are listed in Table 3.1. There are three repeated experiments for each developmental process, each of which contains the expression levels of 30 genes from HSCs to differentiated cells at 30 time points spanning over one week. The observation time points are those starting from the HSCs/progenitors stage (1 point), then every 2 hours over the first day (12 points), every 3 hours over the second day (8 points), every 4 hours over the third day (6 points), every 24 hours until the fifth day (2 points), and the seventh day (1 point). In this study, we used the average data of these three repeated tests as the experimental data for each time point.

### 3.1.2 Selection of candidate genes

Based on our research experience [147], it is challenging to study a dynamic network with 30 genes. Thus, we conducted an extensive literature review for selecting a smaller number of important genes based on their relationship with the three genes *GATA1, GATA2* and *PU.1*. These candidate genes should be essential for the cell-fate choice in hematopoiesis, or they significantly interact with these three genes. For example, gene *Scl/Tal1* interacts with *GATA1, Eto2/Cbfa2t3* and *Ldb1* [45], and is a regulator in the differentiation of hematopoietic stem cells (HSCs) [109, 112, 122, 164]. In addition, *Eto2/Cbfa2t3* regulates the differentiation of HSCs by repressing the expression of target gene *Scl/Tal1* [45]. Moreover, *Ldb1* is a significant TF for the differentiation of erythroid lineage [125]. According to the ChIPSeq analysis, *Ldb1* is necessary for HSCs to control their maintenance since it binds to the majority of enhancer elements in hematopoiesis [74].

We also included a number of genes with potential regulatory relationship with the three genes *GATA1, GATA2* and *PU.1*. For example, it was indicated that there might be

| Reference Paper | Number of Genes | Name of Genes |
|---|---|---|
| Moignard et al., 2013, Figure 1 | 18 | GATA1, GATA2, PU.1, Gfi1, Gfi1b, Hhex, Ldf1, Lmo2, Lyl1, Meis1, Mitf, Nfe2, Runx1, Tal1, Etv6, Erg, Cbfa2t3 |
| Moignard et al., 2015, Figure 3 | 26 | Mesi1, Mitf, Etv2, Fli1, Tal1, GATA1, Hoxb4, Lyl1, Notch1, Sox7, PU.1, Ets1, Erg, Nfe2, Cbfa2t3, Lmo2, Myb, Hoxb2, Sox17, Gfi1, Gfi1b, Hhex, Tbx3, Tbx20, FoxH1, Ikaros |
| This chapter | 30 | GATA1, GATA2, PU.1, Gfi1, Gfi1b, Hhex, Ldb1 Lmo2, Lyl1, Meis1, Mitf, Nfe2, Runx1,Tal1, Etv6, Erg, Cbfa2t3, Etv2, Fli1, Hoxb4, Notch1, Sox7, Ets1, Hoxb2, Sox17, Tbx3, Tbx20, FoxH1, Ikaros, Myb |

Table 3.1: **Information of the 30 candidate genes for differentiation of hematopoietic stem cells.** The 30 genes in "This chapter" are the combination of the genes in two published studies.

unclear regulations between *GATA2* and *Gfi1* [93]. *Gfi1* is an important TF in the regulation of HSCs differentiation [66, 141]. *Gfi1* is required for the differentiation of common lymphoid progenitors (CLPs) and common myeloid progenitors (CMPs) from HSCs and exists in the majority of HSCs, CLPs and CMPs. Similar to gene *Gfi1*, gene *Runx1* is also expressed in most HSCs and progenitor cells as well. Then, *Gfi1* and/or *Runx1* are expressed continually in most cells which differentiate into the granulocyte lineage [100]. *Lmo2* is a master regulator of hematopoiesis [54]. However, its specific role in regulation is still unclear. Experimental studies suggested that the knockdown of *Lmo2* does not affect the expression of *GATA1* and *Scl/Tal1* [54]. However, the overexpression of *Lmo2* gene also inhibited erythroid differentiation [144]. In addition, gene *Ets1* is a suppressor in the erythrocyte differentiation. It is downregulated in erythrocyte differentiation by binding to and activating the GATA2 promoter [81]. The last candidate gene is *Notch1* that inhibits the differentiation of granulocyte lineage by maintaining the expression of gene *GATA2*. It also enhances the HSCs differentiate to CLPs [65, 128]. Therefore, in this study we considered the regulatory networks with the following 11 genes: *GATA1, GATA2, PU.1/Sfpi1, Runx1, Eto2/Cbfa2t3, Ets1, Notch1, Scl/Tal1, Ldb1, Gfi1* and *Lmo2*. The detailed information of the references for these 11 genes is also given in Table 3.2.

| Name of Genes | Reference |
|---|---|
| GATA1 | Friedman, 2007; Liew et al., 2006; Ling et al., 2004 |
| GATA2 | Friedman, 2007; Liew et al., 2006; Ling et al., 2004 |
| PU.1 | Friedman, 2007; Liew et al., 2006; Ling et al., 2004 |
| Runx1 | North et al., 2004 |
| Cbfa2t3 | Goardon et al., 2006 |
| Ets1 | Lulli et al., 2006 |
| Notch1 | Kumano et al., 2001; Stier et al., 2002 |
| Tal1 | Goardon et al., 2006; Shivdasani et al., 1995; Zhang et al., 2005; Porcher et al., 1996; Real et al., 2012 |
| Ldb1 | Soler et al., 2010; Li et al., 2011 |
| Gfi1 | North et al., 2004; van der Meer et al., 2010; Lancrin et al., 2012 |
| Lmo2 | Inouea et al., 2013; Visvader et al., 1997 |

Table 3.2: **Literature information for the selected 11 genes in this study.** These genes are selected from Table 3.1 based on their relationship with the three genes *GATA1*, *GATA2* and *PU.1*.

## 3.2 Methods

### 3.2.1 Top-down approach: static model

In this study, we use the Gaussian graphical model with a forward search algorithm (FSA) [146] as the static model for inferring the structure of regulatory networks. FSA is used to predict gene-gene interactions based on the time series data from microarray dataset. However, since the number of hetero-dimers is much larger than the number of genes, we cannot derive the probabilistic graphic model using the algorithm mentioned above directly. We will provide an extended algorithm to construct regulatory networks with a large number of putative hetero-dimers in the next chapter.

### 3.2.2 Bottom-up approach: dynamic model

In this work, we introduce a dynamic model that considers nonlinearity in genetic regulations as the bottom-up approach for studying the detailed regulatory mechanism. For a gene network with $n$ genes, the expression level of the $i$-th gene is denoted at $x_i(t)$ at time $t$. The following general model describes the dynamics of the network as follows:

$$\dot{\boldsymbol{x}} = \frac{d\boldsymbol{x}}{dt} = F(t, \boldsymbol{x}), \tag{3.1}$$

where $\boldsymbol{x} = (x_1, \ldots, x_n)$ is the expression level of $n$ genes which consist of eleven genes we discussed above. A number of dynamic models have been proposed to describe the regulatory relationship, such as the linear model

$$\frac{dx_i}{dt} = \sum_j a_{ij} x_j \tag{3.2}$$

Although linear model is more efficient than nonlinear model, nonlinear model capture more information than linear model. In our study, it assume that there are nonlinear regulatory relationships existing in the proposed network. Thus, we propose the following dynamic model, given by

$$\frac{dx_i}{dt} = F_i(t, \boldsymbol{x}) = \frac{b_i}{c_i + d_i \exp\{\sum_{j=1, j\neq i}^{n} \alpha_{ij} x_j\}} - k_i x_i, \tag{3.3}$$

where $x_i$ represents the expression level of a single gene, and $k_i$ is the degradation rate of $x_i$. In addition, $b_i$, $c_i$ and $d_i$ are arbitrary constants in this model. To avoid the confusion between degradation and self-regulation, we assume that $\alpha_{ii} = 0$ for each gene and do not discuss the self-regulation.

In this model, the value of the model coefficients represents the putative regulations. If the value of the $\alpha_{ij}$ is positive (negative or zero), it means that gene $x_j$ may activate (inhibit or has no relation to) the expression of gene $x_i$. The network we studied in this work contains 11 genes, thus, the derived system (3.3) has 11 ordinary differential equations in total.

### 3.2.3 Parameter inference

In this study, we used a MATLAB toolbox [22] to infer the unknown ODE model parameters based on the inferred network structures. We use 500 generations and 100 individuals per generation for each estimate of model parameters in the simple genetic algorithm. There are three different types of parameters $(\alpha_{ij}, b_i, c_i, d_i, k_i)$ in our dynamic model. We assume that the initial rate constants follow the uniform distribution $[0, W_{max}]$. To ensure our model works properly, based on the numerical test, we determine the value of $W_{max}$ for parameters $(\alpha_{ij}, b_i, c_i, d_i, k_i)$ are $(1, 1, 1, 1, 1)$, respectively by numerical tests. For each parameter, we first select an initial value of $W_{max}$ to infer model parameters. If certain estimates are very close to $W_{max}$, the value of $W_{max}$ will be increased. Otherwise, the value of $W_{max}$ is decreased if the estimated values are substantially smaller than $W_{max}$. The final estimate of rate constants are different by using different setup of random seed in the algorithm (i.e., the initial estimate of rate constants are changed). For each dynamic model, we infer 200 sets of model parameters and select the top ten sets with minimal errors for further

analysis.

The error of an estimation was measured by the L2-norm between the simulated expression level and original microarray date. The total error is calculated by

$$E = \sum_{i=1}^{N} \sum_{j=1}^{M} (x_i(t_j) - x_{ij}^*)^2 \tag{3.4}$$

where $x_i(t_j)$ and $x_{ij}^*$ are the experimentally and simulated measured gene expression levels at time point $t_j (j = 1, 2, \ldots, M)$, respectively.

### 3.2.4 Robustness analysis

As described in Chapter 2, we next used the robustness property of the model to select the inferred model parameter sets from the Genetic algorithm. It is assumed that the perturbation sample $\varepsilon$ in (2.50) is generated from a standard Gaussian distribution $N(0, 1)$. We have tested various values of $\mu$, to ensure the perturbation has enough impact on the simulation process, we use $\mu = 0.1$ in this robustness analysis.

For each of the top ten sets of parameters determined in the previous subsection, we firstly obtained $N = 5000$ sets of perturbed model parameters by using (2.50) and then used these parameter sets to obtain 5000 corresponding simulations. Then, based on the robustness property (2.52) and (2.53), we determine the optimal model parameter sets from estimated candidates.

## 3.3 Results

### 3.3.1 Inference of regulatory network

To improve the precision of dynamic model. We first use the published MATLAB package - FSA [145] to predict the regulatory network in the fate determination of HSCs. The cutoff p-value controls the number of edges in this network. Since there are eleven genes in our regulatory network. We setup the p-value so that there are no more than 25 edges in this study. When the graph is non-directional, it means that on average each gene receives about 5 regulations from other genes, which is a reasonable value for the sparse genetic regulation. In order to have 25 edges between these eleven genes, we setup the p-value as 0.1 for HSCs choosing MEL and p-value as 0.05 for HSCs choose granulocyte and macrophage lineage (GML).

Using FSA algorithm, an undirected graph between single gene is obtained which is a sparse structure of the network for single genes. This network consist of 25 undirected

edges between different single genes. The predicted regulatory network for two different lineages are presented in Figure 3.1 and Figure 3.2.

### 3.3.2 Inference of dynamic model

After the success implementation of the top-down approach, we obtain a predicted regulatory network of 11 genes. In the next step, we will derive the detailed dynamics of this network. The first step is to estimate the unknown parameters in our dynamic model. When a full connection network in (3.3) is considered, there are $N^2 + 4N = 165$ unknown parameter in our model. After the sparse approximation by FSA the number of unknown parameter is reduced to $50 + 44 = 94$ (50 directed edges, 11 self-degradation terms and 33 arbitrary constants). Based on the partial correlation coefficient obtained by FSA and normalized microarray data, we apply the genetic algorithm to estimate these 94 unknown parameters. The genetic algorithm is implemented by using different random seeds for samples, which leads to different initial samples and then different estimates of parameters. We obtain ten sets of estimated model parameter with minimal simulation errors for the erythroid and neutrophil networks, respectively. Then we analyze the robustness property for these ten sets of estimated parameters. Numerical results suggest that the optimal estimation of unknown parameters for erythroid differentiation is the estimation set with estimation error 0.6003, moreover, the robust mean and standard deviation is 272.4482 and 99.7386, respectively. In addition, the optimal estimation of unknown parameters for neutrophil differentiation is the estimation set with estimation error 0.7328, robust mean 266.9088 and robust standard deviation 120.1733.

Figure 3.3 and Figure 3.4 show simulation results with the optimal estimation set for the expression levels of four genes - *GATA1*, *GATA2*, *PU.1* and *Ets1*, for two different cell fate choices. We clearly see that *GATA1* activity steady climb in the microarray data and our simulation during the erythroid differentiation, however, it keeps fluctuation in the microarray data and we estimate the expression level of *GATA1* almost unchanged during neutrophil differentiation. For *GATA2*, during both differentiated processes, the expression level instantly increase and then keep steady, and slightly increase for the expression level of *GATA2* during neutrophil differentiation. Our simulation results indicates that the expression level of *GATA2* gradually increase for both cell fate choices. The expression level of *PU.1* displays the completely opposite trend for two differentiated processes. Similarly, the expression level of *Ets1* rises with fluctuation during erythroid differentiation and decline with fluctuation during neutrophil differentiation. To summary, the simulation result almost fits the trend of expression level of genes.

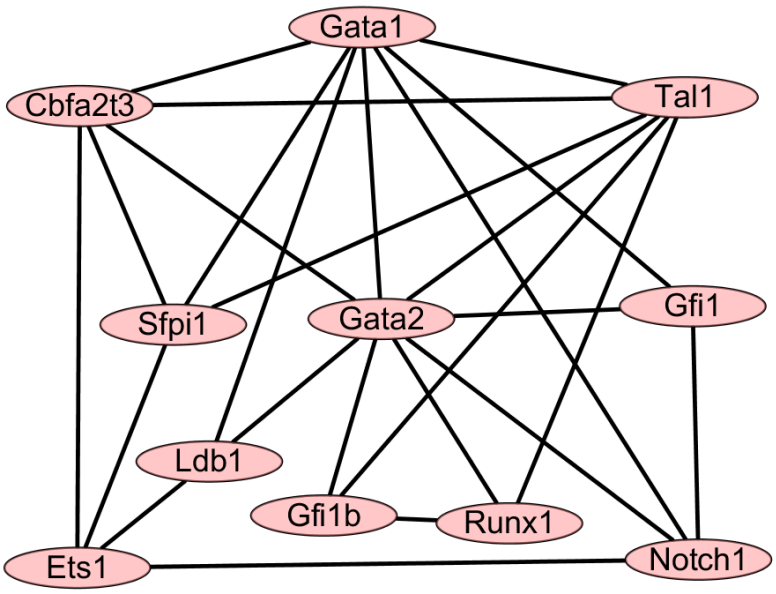


Figure 3.1: **The genetic regulatory networks of eleven genes predicted by FSA are related to fate determination of HSCs.** Regulatory network for HSCs choose megakaryocyte-erythroid lineage. The network is visualized by Cytoscape software.

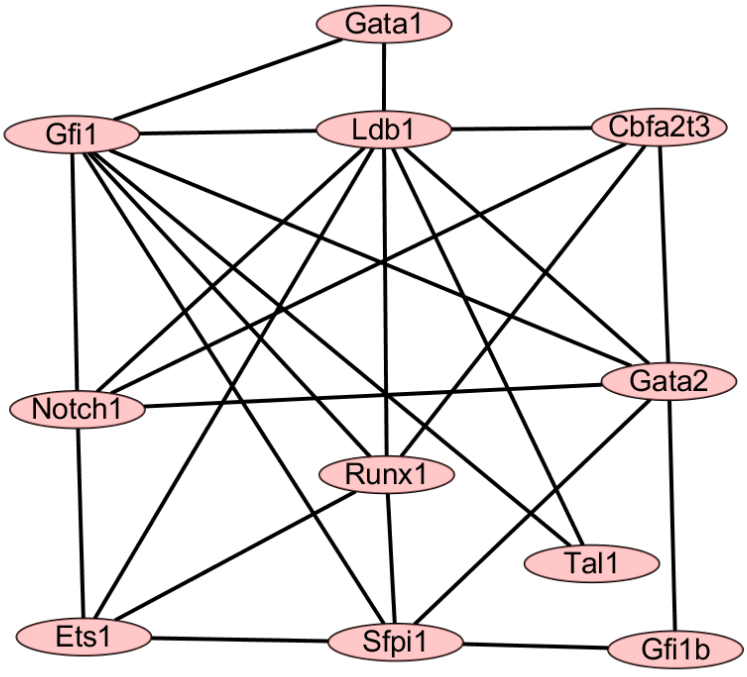


Figure 3.2: **The genetic regulatory networks of eleven genes predicted by FSA are related to fate determination of HSCs.** Regulatory network for HSCs choose granulocyte-macrophage lineage. The network is visualized by Cytoscape software.

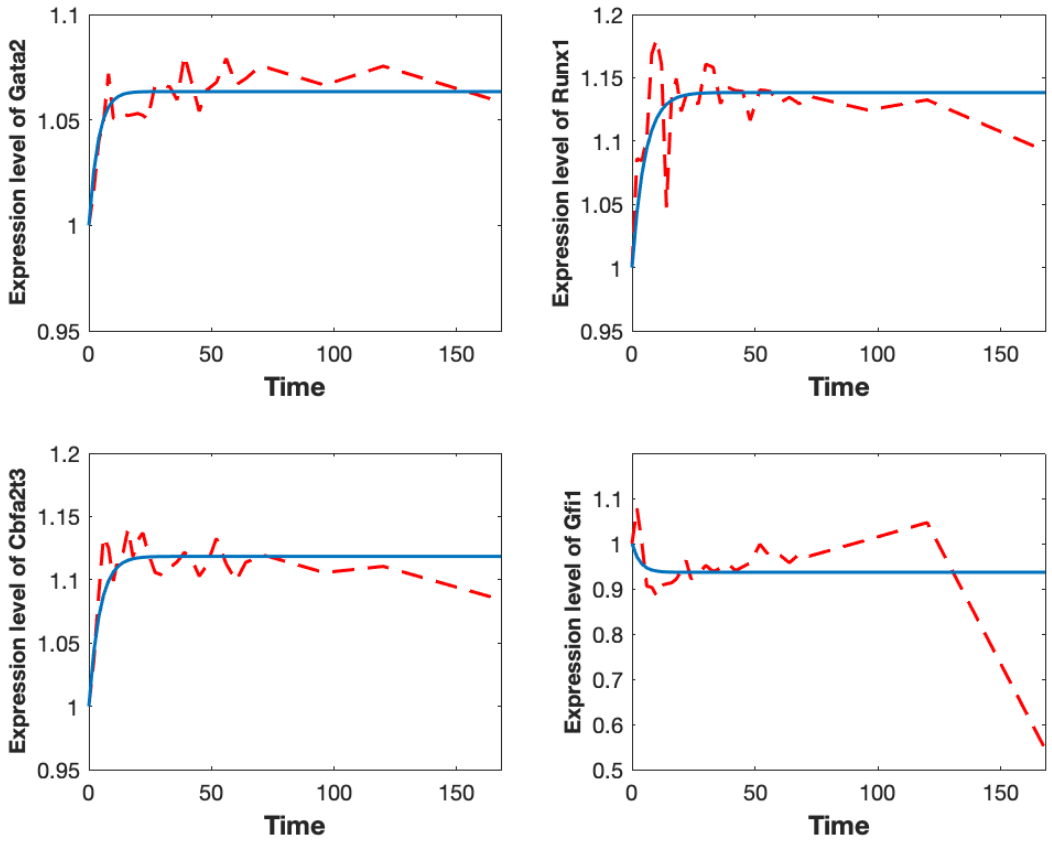


Figure 3.3: **Simulation result of the regulatory network with eleven genes for erythroid differentiation.** Red dash line: microarray data; Blue solid line: simulation of the regulatory network

### 3.3.3 Reduction of network model - edge deletion

The networks predicted by the FSA algorithm are undirected among eleven core genes. There are 25 undirected edges in each of our two predicted network with 11 genes (Figure 3.1 and Figure 3.2). Next we simplify the network structure by reducing certain regulations from the network. In the test of regulation deletion, we evaluated 25 mutual regulatory interaction in erythroid and neutrophil differentiation, respectively, to test the potential insignificant regulation that should be removed from our predicted network. We test the first edge based on 25 edges, then based on the simulations of the 25 systems, we delete one edge if this deletion generates the smallest estimation error among the 25 systems. Then we further test the system by deleting one of the edges based on the remaining 24 edges, and then delete one edge with the same standard as mentioned above. The test is repeated until there is substantial increase in the estimation error and/or decrease in robustness property.

Table 3.3 suggests that, (*GATA1, Gfi1*) edge is significant in this regulatory network,

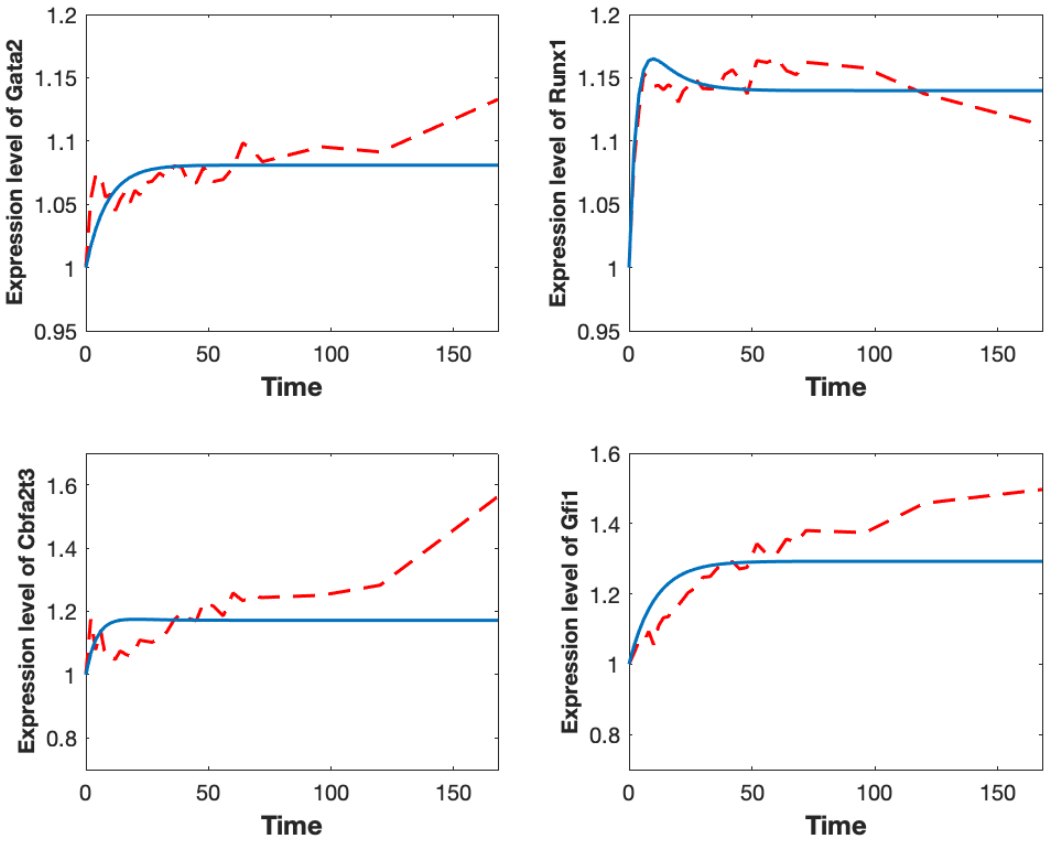


Figure 3.4: **Simulation result of the regulatory network with eleven genes for neutrophil differentiation.** Red dash line: microarray data; Blue solid line: simulation of the regulatory network

since if we delete this edge the simulation error and robustness property become worse than optimal estimation set. However, when we delete (*Notch1, Gfi1*) edge, the estimation error is lower than the error of optimal estimation set, and the robustness property of the systems with deletion is better than that of the optimal model. For edge (*GATA2, Gfi1*), although the estimation error is lower than the optimal estimation error, the robustness property is slightly worse. In Table 3.4, it indicates that deleting both (*Notch1, Gfi1*) and (*GATA2, Gfi1*) edges cause poor robustness. This suggests that the variations in the parameters related to these edges in the optimal model have much influences on the simulation error. Thus, for erythroid differentiation, single (*Notch1, Gfi1*) or (*GATA2, Gfi1*) edges is removable. However, deletion of both edges is not recommended for the regulatory network of erythrocytic differentiation. An interesting observation is that these two deleted edges are related to gene *Gfi1*. Similarly, Table 3.5 indicates that *Gfi1* is a core regulator in neutrophil differentiation because of removing the edges connected with *Gfi1*

have a worse robustness property than before. These numerical results suggest that edge deletion is not recommended for this network.

## 3.4 Summary

In this study we proposed a new dynamic model for inferring genetic regulatory network between eleven genes in the fate determination of HSCs. To improve the accuracy and efficiency of dynamic models, we first applied FSA to infer the network topology. Based on different cutoff p-value of FSA, we derived the regulatory network of eleven genes for MEL (p-value = 0.09) and GML (p-value = 0.05). Combining the partial correlation coefficient, sparse matrix and microarray data, we estimated and tested our dynamic model with different random seed. Subsequently, we simulate the tested model with optimal estimation error and robustness property. Finally, we tested the possibility of removing regulation edge from our predicted regulation network, according to simulation error and robustness property. Numerical results indicated that our proposed method and model is able to provide accurate prediction for inferring the regulatory network among genes. The proposed new methods can be applied to other complex networks.

Table 3.3: **Edge deletion test for erythroid differentiation.** OES represents the network without any deletion. (RA: robustness property in the mean, RSTD: robustness property in standard deviation).

| **Edge pair** | **Estimation Error** | **RA** | **RSTD** |
|---|---|---|---|
| OES | 0.6003 | 272.4482 | 99.7386 |
| GATA1 ↔ Gfi1 | 0.6221 | 256.9349 | 111.5465 |
| Notch1 ↔ Gfi1 | 0.5820 | 270.4757 | 95.2828 |
| GATA2 ↔ Gfi1 | 0.5603 | 279.2123 | 97.9913 |

Table 3.4: **Edge deletion test for erythroid differentiation.** OES represents the network without any deletion. (RA: robustness property in the mean, RSTD: robustness property in standard deviation).

| **Edge pair** | **Estimation Error** | **RA** | **RSTD** |
|---|---|---|---|
| OES | 0.6003 | 272.4482 | 99.7386 |
| Notch1 ↔ Gfi1 | 0.5820 | 270.4757 | 95.2828 |
| Notch1 ↔ Gfi1<br>GATA2 ↔ Gfi1 | 0.5504 | 295.4578 | 101.2557 |

Table 3.5: **Edge deletion test for neutrophil differentiation.** OES represents the network without any deletion. (RA: robustness property in the mean, RSTD: robustness property in standard deviation).

| **Edge pair** | **Estimation Error** | **RBNM** | **RSTD** |
|---|---|---|---|
| OES | 0.7328 | 266.9088 | 120.1733 |
| GATA2 ↔ Gfi1 | 0.7725 | 284.0381 | 100.9540 |
| Pu.1 ↔ Gfi1 | 0.8339 | 295.4406 | 105.4037 |
| GATA1 ↔ Gfi1 | 0.8487 | 295.7750 | 105.0168 |
| Tal1 ↔ Gfi1 | 0.8026 | 281.7749 | 110.3516 |
| Runx1 ↔ Gfi1 | 0.8798 | 295.8294 | 107.7127 |
| Ldb1 ↔ Gfi1 | 0.8391 | 298.1686 | 110.4269 |
| Notch1 ↔ Gfi1 | 0.7205 | 282.2148 | 103.5167 |

# 4

# Extended forward search algorithm for inferring genetic regulatory networks

The objective of this chapter is to extend the method developed in Chapter 3 to study the protein heterodimers and/or synergistic effects involved in genetic regulatory networks. As introduced in Chapter 2, hematopoiesis is a highly complex developmental process that produces various types of blood cells [16, 97]. Although substantial progress has been made for understanding hematopoiesis [18, 21, 52, 103, 137], the detailed regulatory mechanisms of protein monomers, heterodimers and/or synergistic effect for the fate determination of HSCs are still unravelled. In this chapter, we introduce a novel approach to infer the detailed regulatory mechanisms. This work is designed to develop a framework that is able to realize nonlinear gene expression dynamics accurately. In particular, we intended to investigate the effect of possible protein heterodimers and/or synergistic effects in genetic regulation. This approach includes the Extended Forward Search Algorithm to infer network structure (top-down approach) and a nonlinear dynamic model to infer dynamical property (bottom-up approach). Based on the published experimental data, we study two regulatory networks of 11 genes for regulating the erythrocyte differentiation pathway and the neutrophil differentiation pathway. The proposed algorithm is first applied to predict the network topologies among 11 genes and 55 nonlinear terms which may be for heterodimers and/or synergistic effects. Then, the unknown model parameters are estimated by fitting simulations to the expression data of two different differentiation pathways. In addition, the edge deletion test is conducted to remove possible insignificant regulations from the inferred networks. Furthermore, the robustness property of the dynamic model is employed as an additional criterion to choose better network reconstruction results. Our

simulation results successfully realized experimental data for two different differentiation pathways, which suggests that the proposed approach is an effective method to infer the topological structure and dynamic property of genetic regulations. In this chapter, based on the selected candidate genes in Chapter 3, I develop a dynamic model and a inference algorithm, Extended Forward Search Algorithm, to reconstruct the network structure, and then present numerical results.

## 4.1 Methods

### 4.1.1 Top-down approach: static model

To reduce the number of unknown parameters in dynamic model, we used the Gaussian graphical models as the static model to infer the topological structure of gene regulatory networks. In this work it is assumed that a system includes genes $\{G_1, \ldots, G_m\}$ with expression levels $x_{ij}$ for gene $G_i$ at time point $j$. Compared with the existing methods that study networks with genes only, this work will study gene networks that include not only genes in the form of monomers $\{G_1, \ldots, G_m\}$, which are represented by the linear terms in the model, but also protein heterodimers and/or synergistic effect $\{G_k\text{-}G_l\}$ $(k, l = 1, \ldots, m)$, which are represented by the non-linear terms (NLTs) in the model. There are two reasons for using the NLTs $\{G_k\text{-}G_l\}$. Firstly, we can use the product of two variables to represent the synergistic effect of these two genes. Secondly, if the NLT represents the protein heterodimer, we assumed that the binding and disassociation reactions for the heterodimer $\{G_k\text{-}G_l\}$ reach an equilibrium state quickly. Thus the level of the heterodimer $\{G_k\text{-}G_l\}$ can be written as $C_{kl} \times G_k \times G_l$, where $C_{kl}$ is the equilibrium constant. We can consider this constant $C_{kl}$ as a coefficient in our dynamic model. In both cases, we only need to consider the product of the expression levels of these two genes, namely $y_{klj} = x_{kj}x_{lj}$, as the level of NLT $\{G_k\text{-}G_l\}$ at time $t_j$ for our algorithm computation. Since the number of possible regulations from NLTs to genes is much larger than that of possible regulations among genes (i.e., 726 vs 110), the regulations from NLTs to genes will dominate the whole genetic regulatory system with high probability. However, the regulations among genes should be the core mechanisms rather than the regulations from NLTs to genes. To avoid the dominance of NLTs regulations, we assume that the number of regulations from NLTs to genes does not exceed that between genes.

According to the GGM [145, 147], we proposed a new algorithm, named Extended Forward Search Algorithm (EFSA), to infer the topological structure of regulatory networks that includes both genes and NLTs. Let $\boldsymbol{X} = (x_1, x_2 \ldots, x_N)$ be a vector that consists of $m$ genes and $n$ NLTs $(N = m + n)$. The following three matrices are constructed, namely a $m \times m$ covariance matrix $\boldsymbol{A}$ of $m$ genes, a $m \times n$ covariance matrix $\boldsymbol{B}$ to measure the

covariance between $m$ genes and $n$ NLTs, and a $n \times n$ covariance matrix $\boldsymbol{C}$ of $n$ NLTs. The $N$-dimensional matrix $\boldsymbol{M}$ is defined by

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B'} & \boldsymbol{C} \end{bmatrix}, \tag{4.1}$$

where $\boldsymbol{B'}$ is the transpose of $\boldsymbol{B}$. An initial empty graph $\boldsymbol{G}$ is built by the $N$-dimensional identity matrix. This initial graph $\boldsymbol{G}$ consists of four matrices $\boldsymbol{G_1}$, $\boldsymbol{G_2}$, $\boldsymbol{G_3}$ and $\boldsymbol{G_4}$ which have the same dimensions as $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{B'}$ and $\boldsymbol{C}$, respectively, namely

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{G_1} & \boldsymbol{G_2} \\ \boldsymbol{G_3} & \boldsymbol{G_4} \end{bmatrix}, \tag{4.2}$$

where $\boldsymbol{G_1}$ and $\boldsymbol{G_4}$ are identity matrix with dimensions $m$ and $n$, respectively, and $\boldsymbol{G_2}$ and $\boldsymbol{G_3}$ are $m \times n$ and $n \times m$ zero matrices, respectively.

The proposed algorithm is given below [154].

**Extended Forward Search Algorithm (EFSA)**

1. Let $\boldsymbol{X} = (x_1, x_2 \ldots, x_N)$ be a vector with $N$ elements, and $N$ be the number of components consist of $m$ genes and $n$ NLTs. An initial empty graph $\boldsymbol{G}$ is built by the $N$-dimensional identity matrix, which is defined by (4.2).

2. Substitute all covariance values from the diagonal positions of sub-matrix $\boldsymbol{A}$ into the corresponding positions of sub-matrix $\boldsymbol{G_1}$, and then based on the updated $\boldsymbol{G_1}$, use the Iterative Maximum Likelihood Estimates Algorithm (IMLEA) to compute the new covariance matrix [30].

3. Add an undirected edge $E_{ij}^1$ $((i,j) \in [1,m]^2)$ into $\boldsymbol{G_1}$, namely add the symmetrical covariance value between the $i^{th}$ gene and $j^{th}$ gene from the positions $\boldsymbol{A}(i,j)$ and $\boldsymbol{A}(j,i)$ into the positions $\boldsymbol{G_1}(i,j)$ and $\boldsymbol{G_1}(j,i)$, respectively. Then compute a new covariance matrix by the IMLEA. Based on the deviance difference between the new covariance matrix and that before addition, test the significance of the added edge $E_{ij}^1$ by using the Chi-square distribution with one degree of freedom. The p-value of the Chi-square test is used in the next step as the edge selection criterion. Record the p-value of this tested edge and remove it from $\boldsymbol{G_1}$.

4. Add a new undirected edge into $\boldsymbol{G_1}$. Then, repeat the computation in Step 3. After all possible undirected edges have been tested, sort all tested edges in ascending order by their p-values. If the smallest p-value is lower than the predefined cut-off value, add the edge with the smallest p-value into the sub-graph $\boldsymbol{G_1}$ permanently.

5. Go back to step 3, add the second edge in the updated sub-graph $\boldsymbol{G_1}$. Repeat the computation in steps 3 and 4 until the smallest p-value of an added edge is larger than the cutoff p-value.

6. Based on the last updated undirected graph $\boldsymbol{G_1}$, the graph orientation rules are applied to transform the undirected graph into a directed acyclic graph (DAG) [90]. The inferred DAG with $m_1$ directed edges, denoted as $\boldsymbol{A_s}$, represents the predicted regulatory network among $m$ genes.

7. Test the possible edges between $m$ genes and $n$ NLTs. Based on the latest matrix $\boldsymbol{G}$, add an undirected edge $E_{ij}^2$ between the $i^{th}$ gene and the $j^{th}$ NLT. That is, add the symmetrical covariance value between the $i^{th}$ gene and $j^{th}$ NLT from the positions $\boldsymbol{B}(i,j)$ and $\boldsymbol{B'}(j,i)$ into the positions $\boldsymbol{G_2}(i,j)$ and $\boldsymbol{G_3}(j,i)$, respectively. Then, compute a new covariance matrix by the IMLEA [30]. Based on the deviance difference between the new covariance matrix and that before addition, test the significance of the added edge $E_{ij}^2$ by using the Chi-square distribution with one degree of freedom. The p-value of the Chi-square test is used as the edge selection criterion. Record the p-value of this tested edge $E_{ij}^2$ and remove it from $\boldsymbol{G}$.

8. Repeat the computation in steps 7 for the regulation between genes and NLTs. The last updated sub-graph $\boldsymbol{G_3}$ with $n_1$ edges, denoted as $\boldsymbol{B'_s}$, is the predicted directed regulatory network from $n$ NLTs to $m$ genes. Since we only consider regulations among genes and those from NLTs to genes, the result matrix is given as follows,

$$\boldsymbol{G_s} = \begin{bmatrix} \boldsymbol{A_s} \\ \boldsymbol{B'_s} \end{bmatrix}. \tag{4.3}$$

   The output network includes $m_1$ directed edges among $m$ gene and $n_1$ directed edges from $n$ NLTs to $m$ genes.

Note that we have initially applied the GGM in our previous work to the whole matrix $\boldsymbol{M}$ directly [147]. However, since the number of NLTs is much larger than that of genes, numerical results showed that the majority of selected edges connect NLTs, but few edges are selected to connect genes. This result is not appropriate because the regulations between genes should be the primary mechanisms of the network. Then we conducted another test, in which we did not consider the regulations between NLTs by changing matrix $\boldsymbol{C}$ into an identity matrix $\boldsymbol{I}_m$. Matrix $\boldsymbol{M}$ now is

$$\boldsymbol{M_1} = \begin{bmatrix} \boldsymbol{A_s} & \boldsymbol{B} \\ \boldsymbol{B'} & \boldsymbol{I_m} \end{bmatrix}. \tag{4.4}$$

However, when we applied the GGM to $\boldsymbol{M_1}$ directly, the singular problem arose during the computation of IMLEA. To satisfy our intention and make the algorithm stable, we proposed EFSA which is executed in two steps. The first step selects regulations between genes and the second step finds regulations from NLTs to genes. The EFSA can be used to predict the gene-gene interactions and the effect from NLTs to genes based on the time-course experimental data.

### 4.1.2 Bottom-up approach: dynamic model

For a regulatory network with $m$ genes, the expression levels of the $i$-th gene at time $t$ is denoted as $x_i(t)$. We used the following ordinary differential equation (ODE) model to describe the dynamics of the network [28],

$$\frac{d\boldsymbol{x}}{dt} = F(t, \boldsymbol{x}), \tag{4.5}$$

where $\boldsymbol{x} = (x_1, \ldots, x_m)$ is a vector representing the expression levels of $m$ genes. A number of dynamic formalisms have been proposed to describe the dynamical interactions between different genes in the network, such as the models with linear functions [28],

$$F_i(t, \boldsymbol{x}) = \sum_{j=1, j\neq i}^{n} a_{ij}x_j - k_i x_i \tag{4.6}$$

or the models with non-linear functions [103],

$$F_i(t, \boldsymbol{x}) = \frac{\sum_{j=1}^{n} a_{ij}x_j}{1 + \sum_{j=1}^{n} b_{ij}x_j} - k_i x_i. \tag{4.7}$$

The advantage of the model (4.5) with the linear functions (3.2) is that it has a much smaller number of unknown parameters than the non-linear functions (4.7). However, the non-linear model is able to describe the non-linear dynamics more precisely. Therefore, we proposed a method that combines the feature of additive terms in the linear model and the advantages of non-linear model. We applied the second truncated Taylor series approach to approximate the non-linear function (4.7). Here the Taylor series is a dynamic formula to approximate a function by using a polynomial function [126]. Thus, we proposed an ODE model (4.5) with the following functions [154]

$$F_i(t, \boldsymbol{x}) = \sum_{j=1, j\neq i}^{m} \alpha_{ij}x_j + \sum_{1\leq j<k\leq n} \beta_{ijk}x_j x_k - k_i x_i, \tag{4.8}$$

where $k_i$ is the degradation rate of $x_i$. This proposed model (4.5) with the non-linear

function (4.8) is based on the following assumptions:

1. The regulations from different genes to a particular gene are additive. Similarly, the regulations from non-linear terms (NLTs) to a particular gene are also additive.

2. The regulations from gene $j$ to gene $i$ is represented by $\alpha_{ij}x_j$, where $\alpha_{ij}$ is the coefficient of regulation strength.

3. The regulation of NLT $x_jx_k$ to gene $i$ is represented by $\beta_{ijk}x_jx_k$, where $\beta_{ijk}$ consists of the regulation strength and equilibrium constant $C_{ij}$, as we discussed in the subsection Top-down Approach.

4. The auto-regulation is not considered, namely $\alpha_{ii} = 0$, to avoid confusion between auto-regulation term $\alpha_{ii}x_i$ and degradation term $k_ix_i$. Note that the issue of auto-regulation may be addressed using a model with non-linear function (4.7). In addition, we just consider the effect of NLTs $x_jx_k$ for $j \neq k$ since the expression levels of $x_j$ may be highly correlated to that of $x_j^2$. Therefore, we assume that $\beta_{ijj} = 0$.

5. If the value of $\alpha_{ij}$ is positive (negative or zero), it means that gene $x_j$ activates (represses or has no regulation to) the expression of gene $x_i$. Similar assumption is applied to the value of $\beta_{ijk}$.

We emphasize that the proposed method in this work is substantially different from our previous work [147]. The first difference is that the proposed non-linear model (4.8) is different from the non-linear model in [147]. This new model not only can study the regulations from genes to genes, as we considered in our previously proposed model [147], but also can investigate the effects of heterodimers and/or synergistic effect in genetic regulation. This new model also leads to the second difference compared with our previous top-down approach, namely the proposed Extended Forward Search Algorithm (EFSA) not only includes the probabilistic graphical model in our previous work [147] but also can predict the possible regulations from NLTs to genes. In addition, in this work, we will infer a medium-sized network first by using EFSA and then reduce the network size by removing regulations from the network in the Results section, rather than inferring a core network first and then adding regulations to the core network in our previous approach [147].

When considering the full connected graph among $m$ genes and $n$ non-linear terms (NLTs), we have an ordinary differential equation (ODE) system with $m$ differential equations. The total number of all unknown coefficients is $m(m+n)$. After applying the Extended Forward Search Algorithm (EFSA), we have an inferred regulatory network which contains only $m_1$ edges among genes and $n_1$ edges from NLTs to genes. Thus, the numbers

of coefficients $\alpha_{ij}$ and $\beta_{ijk}$ are reduced from $m(m-1)$ to $m_1$ and from $mn$ to $n_1$, respectively. It is easier to estimate the parameters for the inferred network than for the fully connected network.

In this work, we used a MATLAB toolbox of Genetic Algorithm to estimate the parameters in the proposed dynamic model [22]. To ensure the accuracy of estimates, we set the number of generations as 1000 and the number of individuals for each generation as 300. For the parameter vector $(\alpha_{ij}, \beta_{ijk}, k_i)$, we used the uniform distribution over the interval $[W_{min}, W_{max}]$ to generate the initial estimates. Here $W_{min}$ and $W_{max}$ are the minimal value and maximal value, respectively, for choosing the samples of the parameters. The values of $W_{min}$ and $W_{max}$ are adjusted by computation. For example, if the majority of estimated parameters all are close to $W_{min}$, then we will further decrease the value of $W_{min}$. However, if the majority of estimated values are well above $W_{min}$, then we need to increase the value of $W_{min}$ accordingly. The similar consideration is applied to $W_{max}$. In this study, for the erythroid lineage pathway, numerical results suggest that the values of $W_{min}$ and $W_{max}$ for $(\alpha_{ij}, \beta_{ijk}, k_i)$ are $(-3, -3, 0)$ and $(3, 3, 1)$, respectively. In addition, for the neutrophil lineage pathway, numerical results suggest that the values of $W_{min}$ and $W_{max}$ for $(\alpha_{ij}, \beta_{ijk}, k_i)$ are $(-2.5, -2.5, 0)$ and $(2.5, 2.5, 1)$, respectively. We run the algorithm using an initial random number to generate an initial set of model parameters, which leads to a set of estimated parameters. For each model, we used 200 different initial random numbers, which lead to 200 different sets of estimated model parameters. Denote $x_i(t_j)$ and $x_i^*(t_j)$ as the observation data and numerical simulations at time point $t_j$ for $j = \{1, 2, \ldots, M\}$, respectively. The simulation error is calculated by

$$E = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{M}(x_i(t_j) - x_i^*(t_j))^2}. \tag{4.9}$$

We selected the top ten sets with the minimal estimated errors out of 200 estimates for further analysis and comparison.

### 4.1.3 Robustness analysis

As described in Chapter 2, we next used the robustness property of the model to select the inferred model parameter sets from the Genetic algorithm. It is assumed that the perturbation sample $\varepsilon$ in (2.50) is generated from a standard Gaussian distribution $N(0, 1)$. The value of parameter $\mu$ determines the variations of simulations. Numerical results suggest that when the value of $\mu$ is small, perturbation has small effect on the system dynamics, and it is difficult to distinguish the robustness properties of the model with different parameter sets. However, if the value of $\mu$ is large, perturbation will make the

model output substantially different, and it will be difficult to measure the robustness property. To make the variations of simulations appropriately for robustness analysis, $\mu = 0.4$ was employed in this study.

For each of the top ten sets of parameters determined in the previous subsection, we firstly obtained $N = 5000$ sets of perturbed model parameters by using (2.50) and then used these parameter sets to obtain 5000 corresponding simulations. Then, based on the robustness property (2.52) and (2.53), we determine the optimal model parameter sets from estimated candidates.

## 4.2 Results

### 4.2.1 Inference of regulatory network

To reduce the complexity of regulatory networks, we first used the Extended Forward Search Algorithm (EFSA) to predict the topological structure of genetic networks. The algorithm controls the number of edges by adjusting a pre-defined cut-off value. This value is equivalent to the significant value in statistics. If the threshold is too low, we may miss some significant regulations. However, if the threshold is relatively high, it is quite possible to select insignificant regulations. This work considers the networks including 11 genes and 55 non-linear terms (NLTs). For the sub-network of 11 genes only, i.e. matrix $\boldsymbol{A_s}$ in (4.3), to ensure the statistically significant, we set a specific threshold as 0.1 for both the erythroid regulatory network and neutrophil regulatory network. The selection of this threshold value (i.e. 0.1) is based on the balance between neither selecting much insignificant regulations nor choosing a small number of candidate regulations. Then we had 46 and 40 directed edges for the erythroid regulatory network and neutrophil network, respectively.

For the regulations from NLTs to genes, i.e. matrix $\boldsymbol{B'_s}$ in (4.3), the size of matrix $\boldsymbol{B'_s}$ is much larger than that of $\boldsymbol{A_s}$. To avoid the dominance of the regulations from NLTs to genes, we also set the cut-off value as 0.1 for the two networks, or take the first 46 and 40 directed edges from NLTs to genes for the erythrocyte and neutrophil differentiation, respectively, if more edges are selected when using the cut-off value 0.1. The reason we still applied threshold 0.1 here is that the number of selected edges that satisfy this value is much larger than the required number (i.e. 46 for the erythroid regulatory network and 40 for the neutrophil regulatory network). Since the edges are selected and ranked by their significance, we can simply select the top 46 edges and 40 edges for the erythroid and neutrophil pathway, respectively, without conducting any further numerical tests.

Figure 4.1 and Figure 4.2 present the inferred regulatory networks for the erythroid and neutrophil networks, respectively. Note that there are 11 and 17 isolated NLTs in the erythroid and neutrophil networks, respectively, since no significant edges have been selected

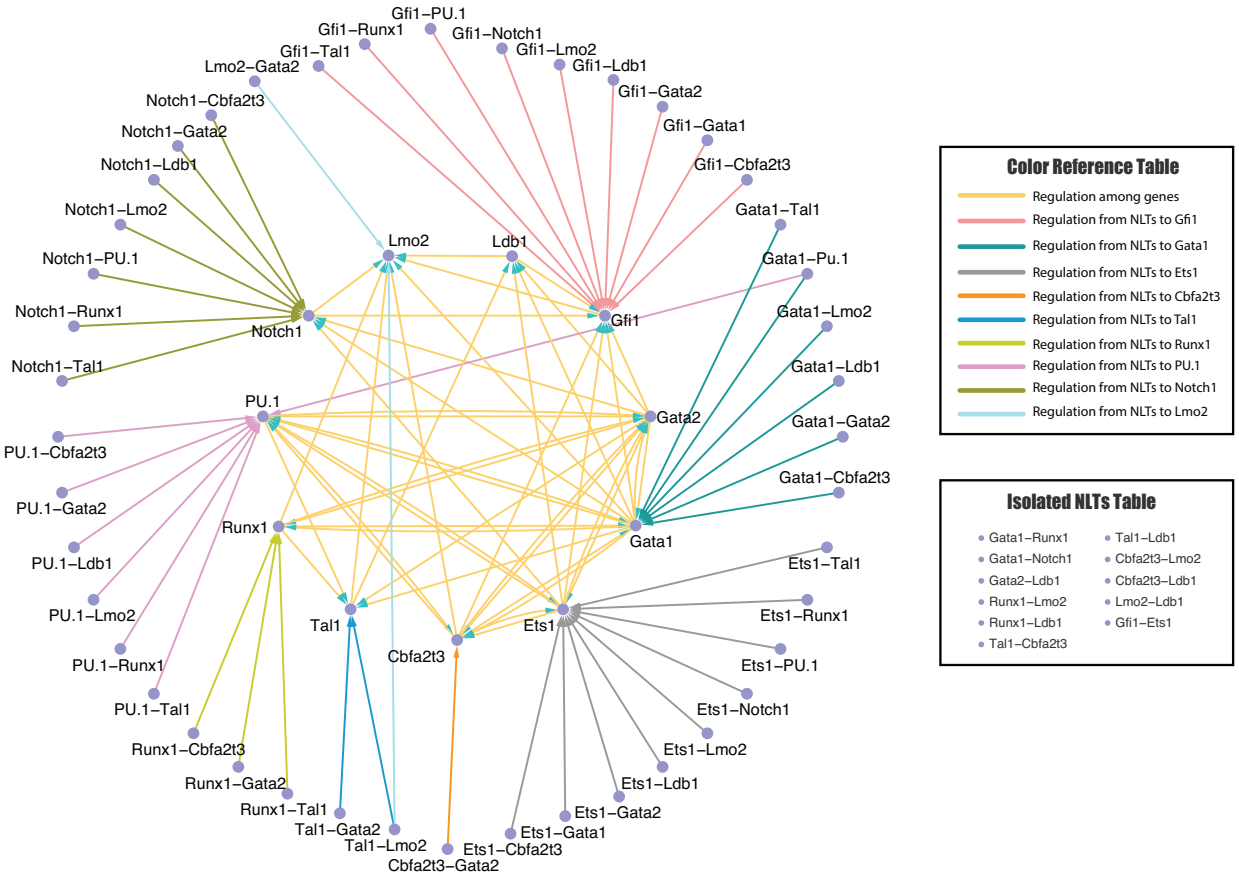


Figure 4.1: **Inferred regulatory network for the differentiation of erythrocyte by EFSA.** The genetic regulatory network predicted by EFSA with 11 genes and 44 NLTs (11 isolated terms excluded), which is related to the fate determination of erythrocyte pathway: Regulatory network for hematopoietic stem cells differentiate to megakaryocyte-erythroid progenitors. The network is visualized by Cytoscape software.

from these NLTs by our algorithm. All these isolated NLTs are listed in the "Isolated NLTs Table". Moreover, all arrows in these figures only represent the direction of regulations, rather than the types of regulations (i.e. positive or negative regulation). We will study the detailed regulatory mechanisms in the next subsection. We found that the targeted gene of the protein heterodimer is a component of that heterodimer in all situations. The possible explanation of this observation is that the expression levels of a heterodimer are the product of the expression levels of the two corresponding genes (namely $x_i x_j$ for genes $i$ and $j$ with expression levels $x_i$ and $x_j$, respectively). Thus, the expression data of the NLTs $\{x_i x_j\}$ may be highly correlated to those of the component genes, namely $\{x_i\}, \{x_j\}$.

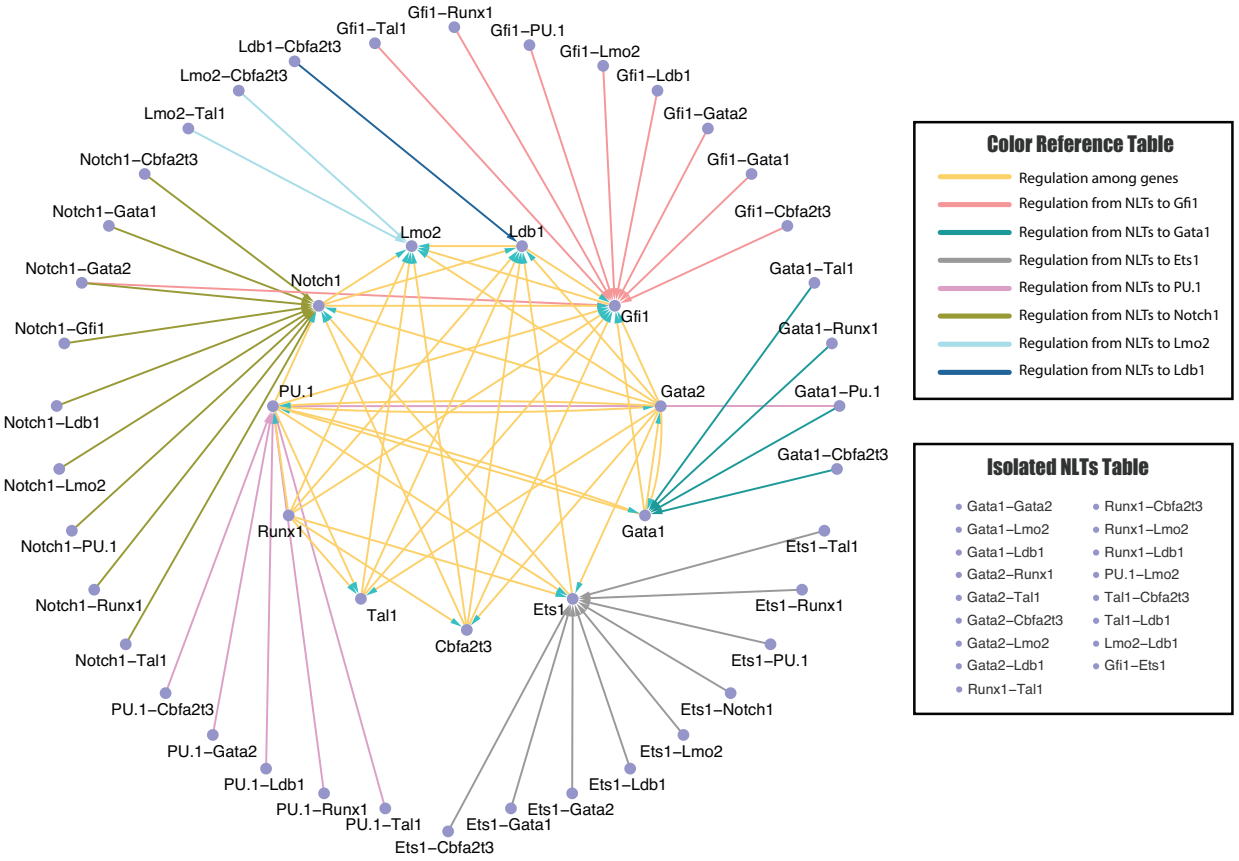


Figure 4.2: **Inferred regulatory network for the differentiation of neutrophil by EFSA.** The genetic regulatory networks predicted by EFSA with 11 genes and 38 NLTs (17 isolated terms excluded), which is related to the fate determination of neutrophil pathway: Regulatory network for hematopoietic stem cells differentiate to granulocyte-macrophage progenitors. The network is visualized by Cytoscape software.

### 4.2.2 Inference of dynamic model

After the success of constructing regulatory networks in the previous sub-section, we next study the detailed dynamics of genetic networks in fate determination of hematopoietic stem cells (HSCs) by using our proposed dynamic model. The major step is to infer the values of unknown parameters in the model (4.8). If we consider the fully connected model, there should be $11 \times (11 + 55) = 726$ parameters. However, after the application of EFSA, the number of unknown parameters is reduced to 103 (including 46 directed edges between genes, 46 directed edges from non-linear terms (NLTs) to genes and 11 self-degradation rate constants) for the differentiation of erythrocytes and 91 (including 40 directed edges between genes, 40 directed edges from NLTs to genes and 11 self-degradation rate constants) for the differentiation of neutrophils. We next applied the Genetic-Algorithm to estimate these unknown parameters for two networks. We used 200 different random numbers to obtain

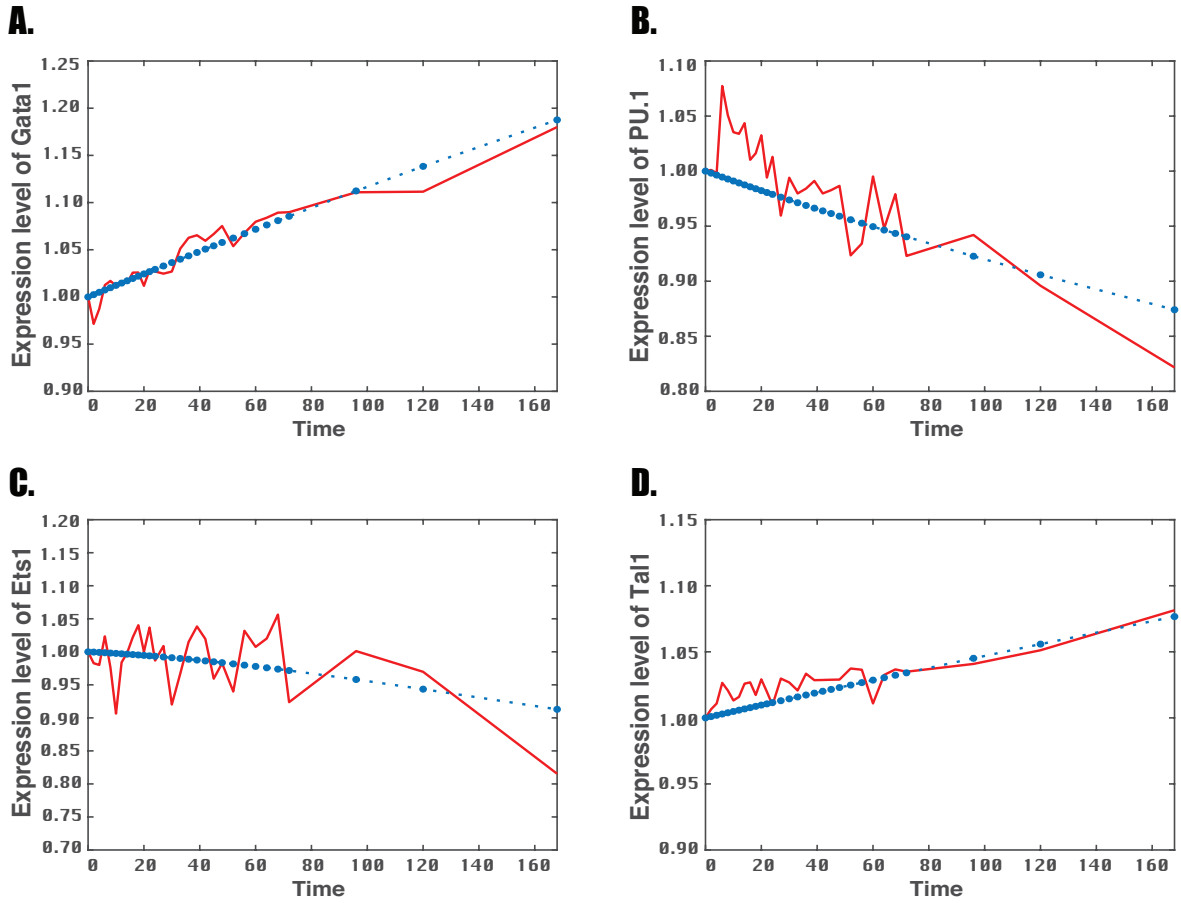


Figure 4.3: **Simulation results and experimental data of the regulatory network for erythrocyte differentiation.** Red solid line: experimental microarray data; Blue star dash line: simulation of the regulatory network.

different initial values of rate constants $(\alpha_{ij}, \beta_{ijk}, k_i)$ over the defined range $[W_{min}, W_{max}]$, which was discussed in the Methods section. This leads to 200 different sets of estimated parameters. Then, we chose the top ten sets of estimated results for each differentiated lineage with the smallest estimation errors for further robustness analysis. According to the definition of estimation error (4.9), the optimal inferred network for the erythrocyte differentiation in our tests has estimation error 0.9902. In addition, based on the robustness defined in Chpater 3, the robust average (2.52) and robust standard deviation (2.53) are 0.3977 and 0.1066, respectively. For the neutrophil differentiation, the optimal inferred network has estimation error 0.8726, robust average 0.3983 and robust standard deviation 0.1275.

Figure 4.3 and Figure 4.4 present the simulation results based on the optimal estimated parameters for the expression levels of four genes, namely genes *GATA1, PU.1, Ets1* and *Tal1*, for the differentiation of erythrocyte and neutrophil, respectively. The expression levels of *GATA1* increase continuously in both simulated and experimental data during the

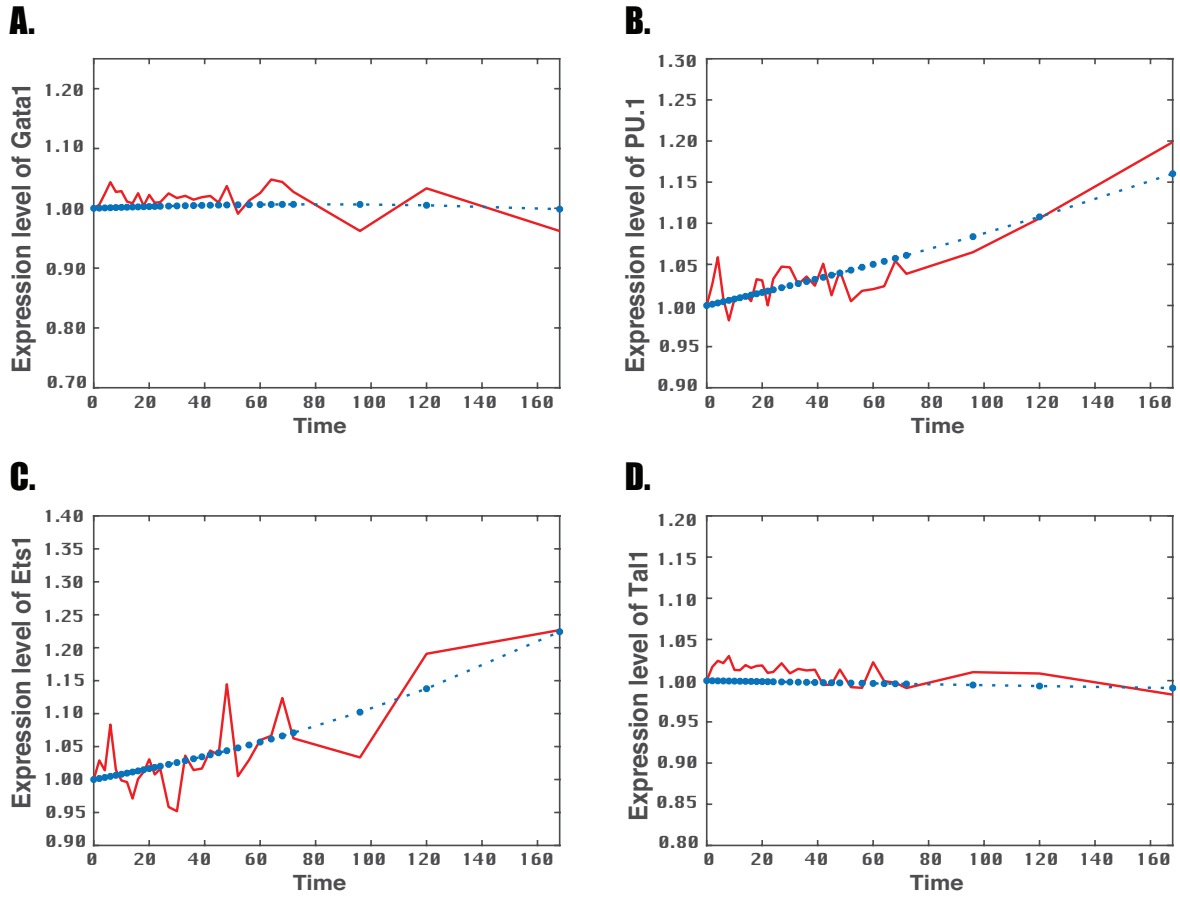


Figure 4.4: **Simulation results and experimental data of the regulatory network for neutrophil differentiation.** Red solid line: experimental microarray data; Blue star dash line: simulation of the regulatory network.

erythrocyte differentiation. However, during the neutrophil differentiation, experimental data of *GATA1* keep fluctuations and then turn to slightly decreasing at the end of differentiation, which is matched by our simulation. For *PU.1* , both microarray and simulated data decline in the differentiation of erythrocyte but climb during the differentiation of neutrophil. Similarly, the expression levels of *Ets1* in microarray data increase during erythrocyte differentiation but decrease during neutrophil differentiation. Simulation results also fit the trends for both differentiation pathways. The experimental data of *Tal1* increase with fluctuations during the first 60 hours of erythrocyte differentiation, but then rises rapidly after the first 60 hours. Our simulated results are consistent with the expression levels of *Tal1* with the same trend in expression levels. Thus, our simulation results fit the trend of expression levels of these genes very well during two developmental processes.

### 4.2.3 Reduction of network model - edge deletion

We have obtained two regulatory networks with 92 directed edges and 80 directed edges for erythroid and neutrophil differentiation, respectively. Next we tested the possibility to delete the potential insignificant edges from our predicted regulatory networks. In the first step, we tested the deletion of regulations from non-linear terms (NLTs) to genes. We removed one edge in each test to form a temporary system model, and then examined the simulation error and robustness property of the new model. Afterwards, we removed one specific edge permanently if the corresponding new system has the minimal change in simulation error and robustness property, and then formed an updated model. This test is repeated until both the simulation error and robustness property of the updated model are much worse than the original network without any removal. In the second step, we evaluated the regulatory interactions between 11 genes using the same method in the first step.

For the erythrocyte differentiation, Table 4.1 suggests that after removing 3 regulations from NLTs to genes, the estimation error (4.9) is improved (shown in DEL1). Then, we tested the regulation reduction from gene to gene. The final result suggests that, after we deleted ($Ldb1 \rightarrow Lmo2$), ($Notch1 \rightarrow Lmo2$), ($Cbfa2t3 \rightarrow Lmo2$) and ($Runx1 \rightarrow Lmo2$) edges, the estimation error (4.9) is slightly increased. However, the robustness property is better than that of the DEL1 model since the robust average (2.52) is decreased. Thus, for the erythroid differentiation, numerical tests recommended to remove total seven edges from our predicted regulatory network. We stopped the deletion test after obtaining the DEL5 model. If we proceed further deletion, both simulation error and robustness property of the temporary network are much worse than the original network without removal.

Table 4.2 shows that, for the neutrophil differentiation, there are no insignificant regulations from NLTs to genes, because the removal of any edge from NLTs to genes will increase the simulation error (4.9) substantially and/or decrease the robustness property by increasing the values of robust average (2.52) and robust standard deviation (2.53). For the regulations between genes, we have removed the following four regulations, namely ($GATA2 \rightarrow Ldb1$), ($Runx1 \rightarrow Cbfa2t3$), ($Ldb1 \rightarrow Lmo2$) and ($Tal1 \rightarrow Lmo2$), and formed an updated system. This table shows that the simulation error and robustness property of the updated system are close to those of the original system without any removal of edges. Thus, for the neutrophil differentiation, numerical tests recommended to remove only four edges from our predicted regulatory network. Coincidentally, we stopped the deletion test after obtaining the DEL5 model because of the same reason for the erythrocyte differentiation.

Figure 4.5 and Figure 4.6 present the inferred regulatory networks after edge deletion test for erythroid and neutrophil differentiation, respectively. Initially, we have 92 directed

edges for the erythrocyte pathway and 80 directed edges for the neutrophil pathway. After the edges deletion, seven and four directed edges have been taken away from the erythrocyte network and neutrophil network, respectively, since the removal of these edges has not much negative influence on simulation error (4.9), robust average (2.52) and robust standard deviation (2.53). Thus, there are 85 and 76 directed edges left for the erythrocyte and neutrophil pathways, respectively.

## 4.3 Summary

This work was designed to develop a dynamic framework that was able to realize nonlinear gene expression dynamics accurately. In particular, we intended to investigate the effect of possible protein heterodimers and/or synergistic effect in genetic regulation. In this study, we designed the Extended Forward Search Algorithm (EFSA) to predict the topology of regulatory networks connecting genes and heterodimers. We also proposed a new dynamic model for inferring dynamic mechanisms of regulatory networks. Using the EFSA, we derived two regulatory networks of 11 genes for erythrocyte and neutrophil differentiation pathways. According to the predicted networks and experimental data, we estimated parameters in our proposed dynamic model based on the criteria of simulation error and robustness property. By removing regulations with less importance based on simulation error and robustness property, we developed two gene networks that regulate erythrocyte and neutrophil differentiation pathways. Numerical results suggested that our proposed method is capable of reconstructing genetic regulatory networks effectively and accurately.

Table 4.1: **Edge deletion test for erythrocyte differentiation.** RR: Removed regulation; SE: Simulation error, defined by (4.9); RA: Robust average, defined by (2.52); RSTD: Robust standard deviation, defined by (2.53).

| **Model** | **RR** | **SE** | **RA** | **RSTD** |
|---|---|---|---|---|
| OES | N/A | 0.9902 | 0.3977 | 0.1066 |
| DEL1 | GATA2-Notch1 → Notch1<br>Tal1-Gfi1 → Gfi1<br>Cbfa2t3-Gfi1 → Gfi1 | 0.9826 | 0.4594 | 0.1259 |
| DEL2 | Ldb1 → Lmo2 | 0.9955 | 0.3938 | 0.1124 |
| DEL3 | Notch1 → Lmo2 | 0.9861 | 0.4506 | 0.1263 |
| DEL4 | Cbfa2t3 → Lmo2 | 1.0451 | 0.3820 | 0.0962 |
| DEL5 | Runx1 → Lmo2 | 1.0298 | 0.3471 | 0.0904 |

**Description of different models:** OES: The original model without any deletion; DEL1: Model based on OES by removing regulations from NLTs to genes; DEL2: Model based on DEL1 by removing a regulation among genes; DEL3: Model based on DEL2 by removing a regulation among genes; DEL4: Model based on DEL3 by removing a regulation among genes; DEL5: Model based on DEL4 by removing a regulation among genes.

Table 4.2: **Edge deletion test for neutrophil differentiation.** RR: Removed regulation; SE: Simulation error, defined by (4.9); RA: Robust average, defined by (2.52); RSTD: Robust standard deviation, defined by (2.53).

| **Model** | **RR** | **SE** | **RA** | **RSTD** |
|---|---|---|---|---|
| OES | N/A | 0.8726 | 0.3983 | 0.1275 |
| DEL1 | No Suggestion | N/A | N/A | N/A |
| DEL2 | GATA2 $\rightarrow$ Ldb1 | 0.8726 | 0.3943 | 0.1273 |
| DEL3 | Runx1 $\rightarrow$ Cbfa2t3 | 0.8726 | 0.3928 | 0.1265 |
| DEL4 | Ldb1 $\rightarrow$ Lmo2 | 0.8748 | 0.4183 | 0.1333 |
| DEL5 | Tal1 $\rightarrow$ Lmo2 | 0.8809 | 0.3925 | 0.1237 |

**Description of different models:** OES: The original model without any deletion; DEL1: Model based on OES by removing regulations from NLTs to genes; DEL2: Model based on DEL1 by removing a regulation among genes; DEL3: Model based on DEL2 by removing a regulation among genes; DEL4: Model based on DEL3 by removing a regulation among genes; DEL5: Model based on DEL4 by removing a regulation among genes.

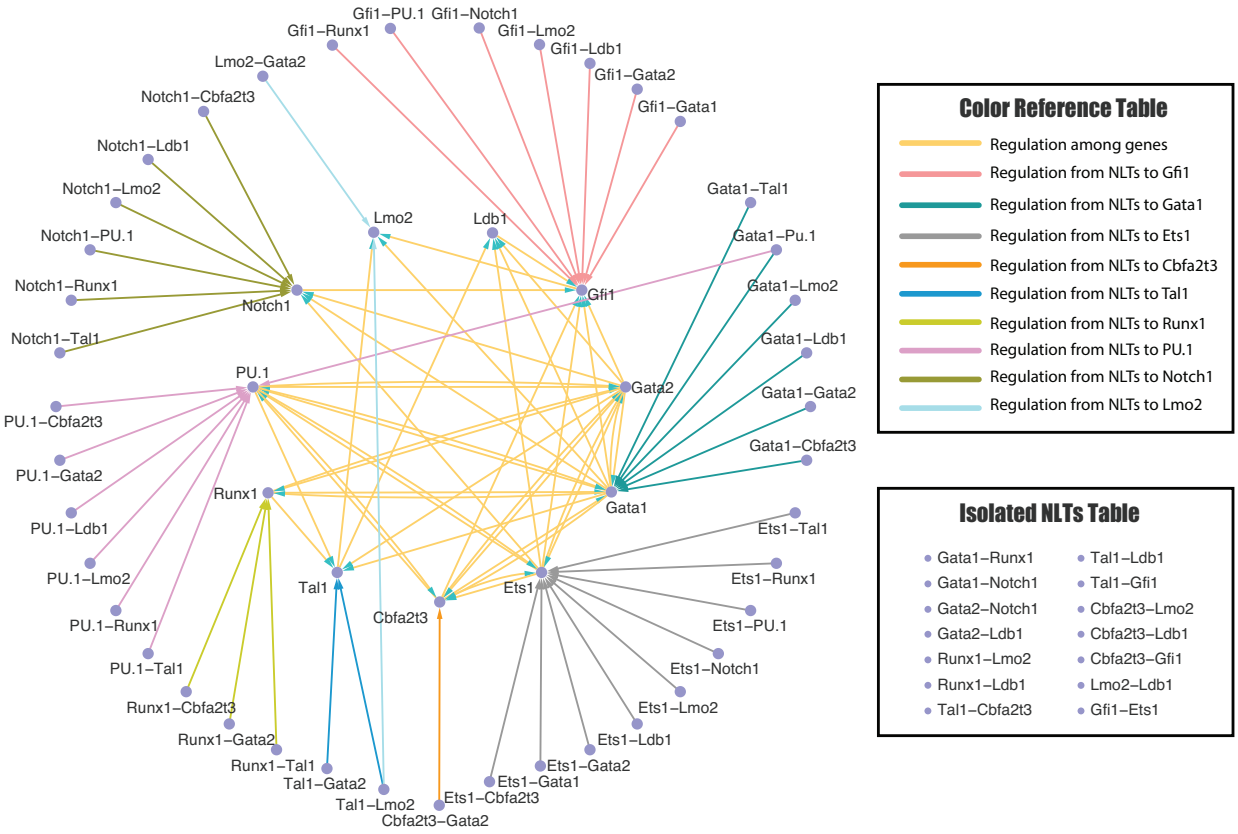


Figure 4.5: **Predicted genetic regulatory network of erythrocyte pathway after edges deletion.** The genetic regulatory network predicted by the Extended Forward Search Algorithm with 11 genes and 41 non-linear terms (NLTs) (14 isolated NLTs excluded) after edges deletion test, which is related to the fate determination of erythrocyte pathway: Regulatory network for hematopoietic stem cells differentiate to megakaryocyte-erythroid progenitors. The network is visualized by Cytoscape software.

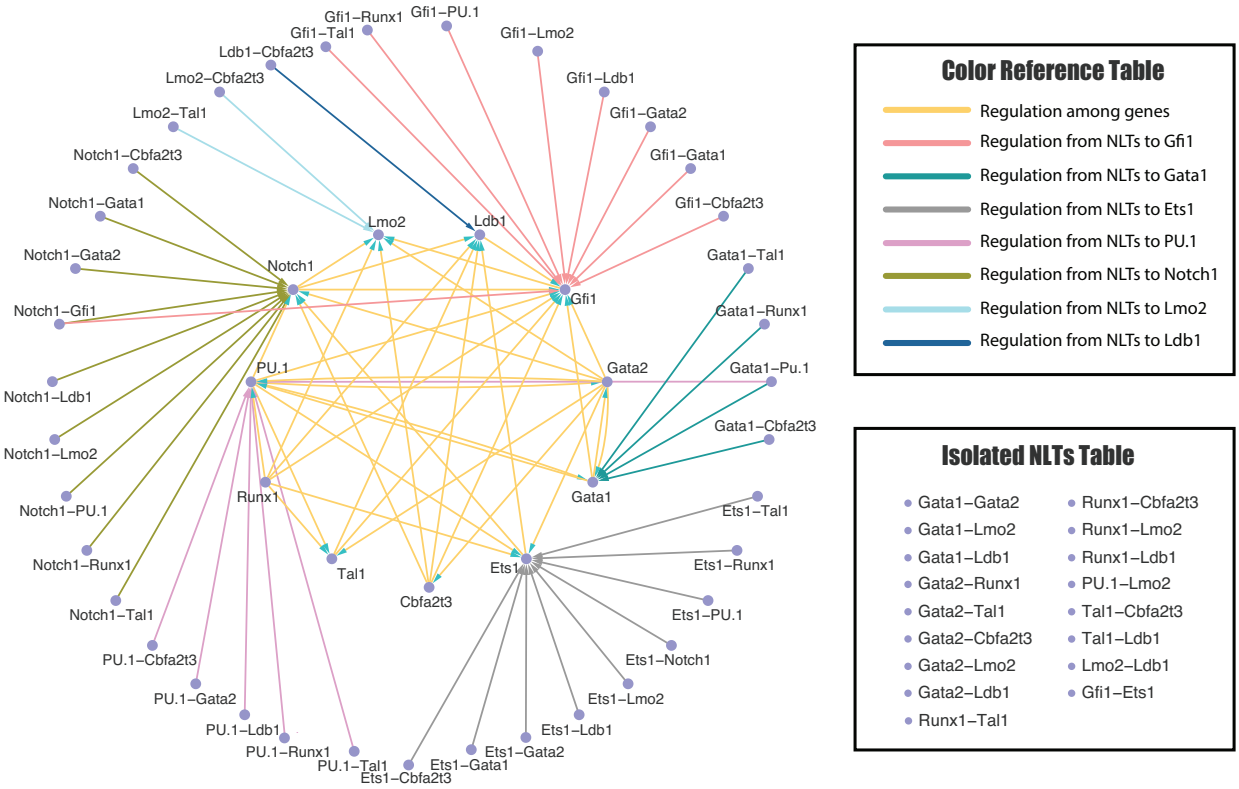


Figure 4.6: **Predicted genetic regulatory network of neutrophil pathway after edges deletion.** The genetic regulatory networks predicted by the Extended Forward Search Algorithm with 11 genes and 38 non-linear terms (NLTs) (17 isolated NLTs excluded) after edges deletion test, which is related to the fate determination of neutrophil pathway: Regulatory network for hematopoietic stem cells differentiate to granulocyte-macrophage progenitors. The network is visualized by Cytoscape software.

# 5

# A robust method for designing multistable systems by embedding bistable subsystems

The objective of this chapter is to introduce a novel and robust method to develop multistable mathematical models by embedding bistable models together. This study uses the *GATA1-GATA2-PU.1* module in hematopoiesis as the test system, we first develop a tristable model based on two bistable models without any high cooperative coefficients, and then modify the tristable model based on the experimentally determined mechanisms. The modified model successfully realize four stable steady states and accurately reflects a recent experimental observation showing four transcriptional states. In addition, we develop a stochastic model, and stochastic simulations successfully realize the experimental observations in single cells. These results suggest that the proposed method is a general approach to develop mathematical models for realizing multistability and heterogeneity in complex systems. In this chapter, I first outline principle for obtaining a multistable system by embedding bistable systems, then introduce the model development for both bistable systems and the embedding multistable system. Finally, I apply the methodology to the problem of cell fate determination in hematopoiesis.

## 5.1 Principle of embeddedness

### 5.1.1 Embedding method for designing multistable models

We propose a framework to model regulatory networks with multiple stable steady states based on the embedding of sub-systems with less stable steady states [155]. It is assumed

that we need to study a regulatory network that consists of two regulatory modules. The first module has genes $X_i$, and it is modelled by following equations

$$\frac{dX_i}{dt} = \mathcal{F}_i(X_1, X_2, \cdots, X_n, X_{n+1}, \cdots, X_{n+N}, \boldsymbol{\Theta}_1, t) \tag{5.1}$$

for $i = 1, 2, \cdots, n+N$, where $\boldsymbol{\Theta}_1$ includes model parameters of $\mathcal{F}_i$. The second module has the following model

$$\frac{dY_j}{dt} = \mathcal{G}_j(Y_1, Y_2, \cdots, Y_m, \boldsymbol{\Theta}_2, t) \tag{5.2}$$

for $j = 1, 2, \cdots, m$, where $\boldsymbol{\Theta}_2$ includes model parameters of $\mathcal{G}_j$. In these two models, $\mathcal{F}(\boldsymbol{X}, \boldsymbol{\Theta}_1, t)$ and $\mathcal{G}(\boldsymbol{Y}, \boldsymbol{\Theta}_2, t)$ are non-linear vector fields. To develop mathematical models with more stable steady states, we propose an embedding method by assuming that $X_{n+k}$ $(k = 1, ..., N)$ are functions of variables $Y_1, Y_2, \cdots, Y_m$, given by

$$X_{n+k} = \mathcal{H}_k(Y_1, Y_2, \cdots, Y_m). \tag{5.3}$$

In this way, we obtain an embedding system

$$\frac{d\boldsymbol{W}}{dt} = \boldsymbol{F}(\boldsymbol{W}, \boldsymbol{\Theta}^*, t), \tag{5.4}$$

where $\boldsymbol{W} = (X_1, X_2, \cdots, X_n, Y_1, Y_2, \cdots, Y_m)$ represents all genes in the system, $\boldsymbol{F}$ denotes embedding systems from two modules with gene $X_i$ and $Y_i$ with function $\mathcal{H}_k$. In addition, $\boldsymbol{\Theta}^* = \boldsymbol{\Theta}_1 \cup \boldsymbol{\Theta}_2$ is the model parameters space. This embedding system (5.4), consists of two components:

$$\begin{aligned} \frac{dX_i}{dt} &= \mathcal{F}_i(X_1, X_2, \cdots, X_n, \mathcal{H}_k(Y_1, Y_2, \cdots, Y_m), \boldsymbol{\Theta}^*, t), \\ \frac{dY_j}{dt} &= \mathcal{R}_j(X_1, X_2, \cdots, X_n, Y_1, Y_2, \cdots, Y_m, \boldsymbol{\Theta}^*, t) \end{aligned} \tag{5.5}$$

for $i = 1, 2, \cdots, n$, $k = 1, ..., N$ and $j = 1, 2, \cdots, m$. Since each $X_i$ is regulated by the $X_{n+k}$ $(k = 1, ..., N)$, and $X_{n+k}$ are functions of $Y_1, Y_2, \cdots, Y_m$, the expressions of each gene $Y_j$ is also regulated by $X_i$ $(i = 1, ..., n)$. The non-linear vector field $\mathcal{G}(\boldsymbol{Y}, \boldsymbol{\Theta}_2, t)$ in (5.2) will then be transformed to a new non-linear vector field $\mathcal{R}(\boldsymbol{W}, \boldsymbol{\Theta}^*, t)$, which includes both genes $X_i$ and $Y_i$ from two sub-systems with their corresponding regulations. Note that this is a general idea to develop mathematical models with more stable steady states. Depending the specific formalism and properties of sub-systems, the embedding system may have different results regarding multiple stable steady states with different conditions. In this study, we only focus on the systems with Shea-Ackers formalism [1].

### 5.1.2 Effectiveness of embedding method

The motivation of this work is to develop a mathematical model to realize the tristable property of the HSC genetic regulatory network in Figure 5.1A based on experimental observations. Figure 5.1B and Figure 5.1E illustrate the embedding method to couple two bistable modules in a network together. Variable $U$ in the first $Z$-$U$ module is an auxiliary node, which is assumed to be $U = \mu X + \delta Y$, where $\mu$ and $\delta$ are two positive parameters. When the system stays in the state with a high expression level of $Z$ and a low level of $U$, the expression levels of $X$ and $Y$ are low. However, when the system has a low expression level of $Z$ and a high level of $U$, the system triggers the second module $X$-$Y$ to choose either a high level of $X$ and a low level of $Y$ or a low level of $X$ and a high level of $Y$. In this way we realize the system with three stable states in which one of the three variables (namely $Z$, $X$ or $Y$) is at the high expression state but the other two are at low expression states.

To demonstrate the effectiveness of the proposed embedding method, we use the toggle switch network as the test system [62]. This network consists of two genes that form a double negative feedback loop and is modelled by the following equations with parameter space $\boldsymbol{\Theta}_1 = \{a = 0.2, b = 4, c = 3\}$, given by

$$\begin{aligned} \frac{dz}{dt} &= \mathcal{F}_1(z, u, \boldsymbol{\Theta}_1, t) = 0.2 + \frac{4}{1+u^3} - z, \\ \frac{du}{dt} &= \mathcal{F}_2(z, u, \boldsymbol{\Theta}_1, t) = 0.2 + \frac{4}{1+z^3} - u. \end{aligned} \tag{5.6}$$

It is assumed that the first $Z$-$U$ module follows model (5.6) and the second $X$-$Y$ module satisfies the same model with same parameter space $\boldsymbol{\Theta}_1$, but different variables $x$ and $y$, given by

$$\begin{aligned} \frac{dx}{dt} &= \mathcal{G}_1(x, y, \boldsymbol{\Theta}_1, t) = 0.2 + \frac{4}{1+y^3} - x, \\ \frac{dy}{dt} &= \mathcal{G}_2(x, y, \boldsymbol{\Theta}_1, t) = 0.2 + \frac{4}{1+x^3} - y. \end{aligned} \tag{5.7}$$

Now we embed these two sub-systems together using $u = \mathcal{H}(x, y) = x + y$. Since gene $z$ is negatively regulated by gene $u$ in the sub-system (5.6), and $u$ is a function of genes $x$ and $y$, the expressions of genes $x$ and $y$ are also negatively regulated by gene $z$ in the new embedding model. Then the non-linear vector fields $\mathcal{G}_{1,2}(x, y, \boldsymbol{\Theta}_1, t)$ are transformed to new non-linear vector fields $\mathcal{R}_{1,2}(x, y, z, \boldsymbol{\Theta}_1, t)$, respectively, which include both genes $x$, $y$ and $z$ from two sub-systems with negative regulations from gene $z$ to genes $x$ and $y$. Therefore, the new model with three variables is given by

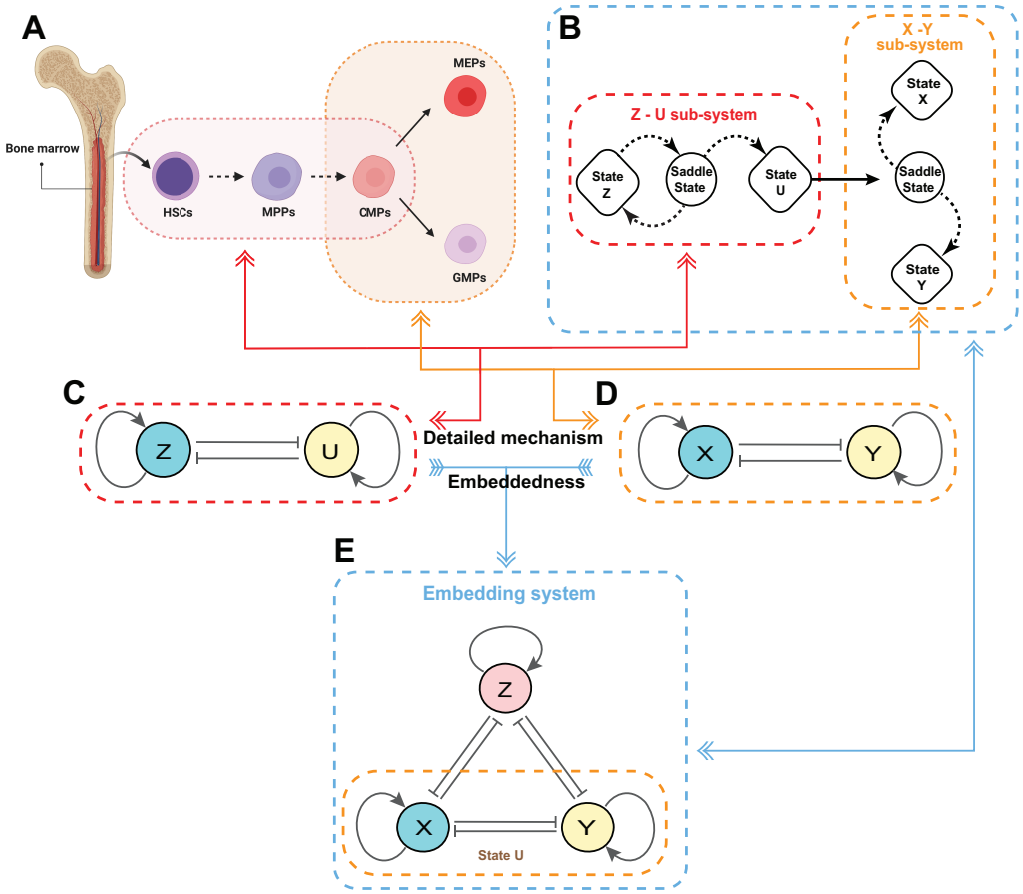


Figure 5.1: **Methodology for developing multistable models by embedding two sub-systems with bistability together.**
(A) Brief flowchart of hematopoietic hierarchy that is created with BioRender.com. HSCs, hematopoietic stem cells; MPPs, multipotent progenitors; MEPs, megakaryocyte-erythroid progenitors; GMPs, granulocyte-macrophage progenitors.
(B) The principle of embeddedness: $Z$-$U$ module is the first bistable sub-system. Once this module crosses the saddle point from state $Z$ to state $U$, it enters the $X$-$Y$ sub-system that has two stable steady states $X$ and $Y$, reaching either state $X$ and state $Y$ via the imaginary state $U$.
(C,D) The structure of two double-negative feedback loops with positive autoregulations, which is the mechanisms for bistable sub-systems in HSCs.
(E) The structure of regulatory network after embeddedness. The $X$-$Y$ sub-system is embedded into the state $U$. ('$\rightarrow$' and '$\dashv$' denote the activating and inhibiting regulations, respectively.)

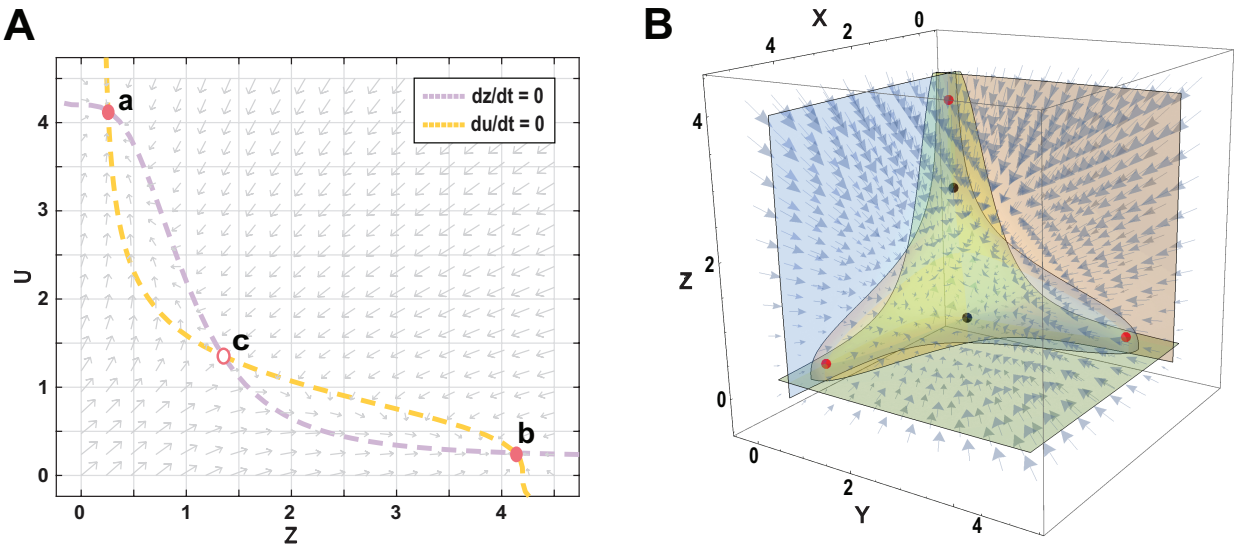


Figure 5.2: **Realization of tristability by embedding two bistable sub-systems.**
(A) The phase plane of the toggle switch sub-system (5.6) with bistability (a and b: stable steady states, c: saddle state).
(B) The 3D phase portrait of the embedded system (5.8) with tristability (Three red points: stable steady states; two black points: saddle states)

$$\begin{aligned}
\frac{dx}{dt} &= \mathcal{R}_1(x, y, z, \boldsymbol{\Theta}_1, t) = 0.2 + \frac{4}{(1+y^3)(1+z^3)} - x, \\
\frac{dy}{dt} &= \mathcal{R}_w(x, y, z, \boldsymbol{\Theta}_1, t) = 0.2 + \frac{4}{(1+x^3)(1+z^3)} - y, \\
\frac{dz}{dt} &= \mathcal{F}_1(z, u = x + y, \boldsymbol{\Theta}_1, t) = 0.2 + \frac{4}{1+(x+y)^3} - z.
\end{aligned} \tag{5.8}$$

Figure 5.2A shows the phase plane of the toggle switch sub-system (5.6), with bistability properties, and Figure 5.2B provides the 3D phase portrait of the embedded model (5.8), with three stable steady states. The embedded model successfully realized the tristability, which validates our embedding method for developing mathematical models with multistability.

## 5.2 Model development for embedding method

### 5.2.1 Model development with bistability properties

We first develop a model for the network in Figure 5.1C and Figure 5.1D with bistability properties. Suppose that two sub-systems, namely the $Z$-$U$ system and $X$-$Y$ sub-

system, have the same structure of a double-negative feedback loop and positive autoregulations. For the $Z$-$U$ system, based on the formalism (5.1) with $\boldsymbol{X} = \{z, u\}$ and $\boldsymbol{\Theta}_1 = \{a_1, b_1, b_2, c_1, d_1, d_2, k_1, k_2\}$, we propose the following model to describe the dynamics, given by

$$\begin{aligned} \frac{dz}{dt} &= \mathcal{F}_1(z, u, \boldsymbol{\Theta}_1, t) = \frac{a_1 z}{1 + b_1 z} \frac{1}{1 + b_2 u} - k_1 z, \\ \frac{du}{dt} &= \mathcal{F}_2(z, u, \boldsymbol{\Theta}_1, t) = \frac{c_1 u}{1 + d_1 u} \frac{1}{1 + d_2 z} - k_2 u. \end{aligned} \tag{5.9}$$

Similarly, based on the formalism (5.2) with $\boldsymbol{Y} = \{x, y\}$ and $\boldsymbol{\Theta}_2 = \{\alpha_1, \beta_1, \beta_2, \gamma_1, \sigma_1, \sigma_2, k_3, k_4\}$, the dynamics of the $X - Y$ subsystem is modelled by

$$\begin{aligned} \frac{dx}{dt} &= \mathcal{G}_1(x, y, \boldsymbol{\Theta}_2, t) = \frac{\alpha_1 x}{1 + \beta_1 x} \frac{1}{1 + \beta_2 y} - k_3 x, \\ \frac{dy}{dt} &= \mathcal{G}_2(x, y, \boldsymbol{\Theta}_2, t) = \frac{\gamma_1 y}{1 + \sigma_1 y} \frac{1}{1 + \sigma_2 x} - k_4 y, \end{aligned} \tag{5.10}$$

where $x$ and $y$ are expression levels of genes $X$ and $Y$, respectively; $\alpha_1$ and $\gamma_1$ represent expression rates; $\beta_1, \beta_2, \sigma_1$ and $\sigma_2$ represent association rates of corresponding proteins on the binding-sites; and $k_3$ and $k_4$ are self-degradation rates. The model of the $Z - U$ subsystem has the same structure but may have different values of model parameters. To obtain the bistability, we establish following theorems for our proposed models for these two sub-systems. Since they have the same structure, we only give the theorems for the $X$-$Y$ sub-system.

**Theorem 5.2.1.** There are at most five sets of non-negative equilibria for the model of the $X$-$Y$ system.

1. There are three equilibria: $(0, 0)$, $(x_e, 0)$ and $(0, y_e)$, where $x_e = \frac{\alpha_1 - k_3}{k_3 \beta_1}$ and $y_e = \frac{\gamma_1 - k_4}{k_4 \sigma_1}$, if $\alpha_1 > k_3$ and $\gamma_1 > k_4$.

2. There are two other equilibria: $(x_1^*, y_1^*)$ and $(x_2^*, y_2^*)$. If $-\frac{\mathcal{B}}{\mathcal{A}} > 0$, $\frac{\mathcal{C}}{\mathcal{A}} > 0$ and $\mathcal{B}^2 - 4\mathcal{A}\mathcal{C} \geq 0$, then $x_1^*$ and $x_2^*$ are positive real solutions of the following equation,

$$\mathcal{A} m^2 + \mathcal{B} m + \mathcal{C} = 0, \tag{5.11}$$

   where $m = \beta_1 x, \mathcal{A} = A_1 B_1 - B_1, \mathcal{B} = A_1 - B_1 - 1 + A_1 B_1 - A_1 B_2 + A_2 B_1$, $\mathcal{C} = A_1 + A_2 - 1 - A_1 B_2$, $A_1 = \frac{\beta_2}{\sigma_1}, A_2 = \frac{\alpha_1}{k_3}, B_1 = \frac{\sigma_2}{\beta_1}$ and $B_2 = \frac{\gamma_1}{k_4}$.

3. To have positive values of $y_1^*$ and $y_2^*$, the following conditions should be satisfied,

$$x_{1,2}^* < \frac{A_2 - 1}{\beta_1} \text{ or } x_{1,2}^* < \frac{B_2 - 1}{\sigma_2}. \tag{5.12}$$

**Proof.** Suppose the the equilibrium state exists, then we have

$$\frac{\alpha_1 x}{1+\beta_1 x}\frac{1}{1+\beta_2 y} - k_3 x = 0, \tag{5.13}$$

$$\frac{\gamma_1 y}{1+\sigma_1 y}\frac{1}{1+\sigma_2 x} - k_4 y = 0. \tag{5.14}$$

Then, we consider the following three cases.

1. 
   - The trivial solution: $(0, 0)$.
   - When the equilibrium state is $(x_e, 0)$, where $x_e \neq 0$. According to (5.13), we have

     $$x_e = \frac{\alpha_1 - k_3}{k_3 \beta_1}. \tag{5.15}$$

     Since $\alpha_1, k_3$ and $\beta_1$ are positive, we have the positive equilibrium solution if $\alpha_1 > k_3$.
   - When the equilibrium state is $(0, y_e)$, where $y_e \neq 0$. According to (5.14), we have

     $$y_e = \frac{\gamma_1 - k_4}{k_4 \sigma_1}. \tag{5.16}$$

     Since $\gamma_1, k_4$ and $\sigma_1$ are positive, we have the positive equilibrium solution if $\gamma_1 > k_4$.

2. When the equilibria are $(x_1^*, y_1^*)$ and $(x_2^*, y_2^*)$, where all values here are not zero, according to (5.13) and (5.14), we have

   $$\frac{\alpha_1}{(1+\beta_1 x)(1+\beta_2 y)} = k_3, \tag{5.17}$$

   $$\frac{\gamma_1}{(1+\sigma_1 y)(1+\sigma_2 x)} = k_4. \tag{5.18}$$

   Let $m = \beta_1 x$ and $n = \sigma_1 y$, we have

   $$(1+m)(1+\frac{\beta_2}{\sigma_1} n) = \frac{\alpha_1}{k_3}, \tag{5.19}$$

   $$(1+n)(1+\frac{\sigma_2}{\beta_1} m) = \frac{\gamma_1}{k_4}. \tag{5.20}$$

Let $A_1 = \frac{\beta_2}{\sigma_1}, A_2 = \frac{\alpha_1}{k_3}, B_1 = \frac{\sigma_2}{\beta_1}$ and $B_2 = \frac{\gamma_1}{k_4}$. Finally, we can get

$$(1+m)(1+A_1 n) = A_2, \tag{5.21}$$

$$(1+n)(1+B_1 m) = B_2. \tag{5.22}$$

From (5.21) and (5.22), we have

$$n = \frac{A_2 - 1 - m}{A_1 + A_1 m} = \frac{B_2 - 1 - B_1 m}{1 + B_1 m}. \tag{5.23}$$

That is,

$$(A_1B_1 - B_1)m^2 + (A_1 - B_1 - 1 + A_1B_1 - A_1B_2 + A_2B_1)m + (A_1 + A_2 - 1 - A_1B_2) = 0.$$

Let $\mathcal{A} = A_1B_1 - B_1$, $\mathcal{B} = A_1 - B_1 - 1 + A_1B_1 - A_1B_2 + A_2B_1$ and $\mathcal{C} = A_1 + A_2 - 1 - A_1B_2$. Then, we have the following quadratic function

$$\mathcal{A}m^2 + \mathcal{B}m + \mathcal{C} = 0. \tag{5.24}$$

(a) If $\Delta = \mathcal{B}^2 - 4\mathcal{A}\mathcal{C} = 0$, there is only one solution, namely $m = \frac{-\mathcal{B}}{2\mathcal{A}}$. Thus, the solution of $m = \frac{-\mathcal{B}}{2\mathcal{A}}$ is positive if $-\frac{\mathcal{B}}{\mathcal{A}} > 0$. Then we have

$$x_1^* = x_2^* = \frac{-\mathcal{B}}{2\beta_1\mathcal{A}} \text{ and } y_1^* = y_2^* = \frac{A_2 - 1 - \beta_1 x_{1,2}^*}{\beta_2(1 + \beta_1 x_{1,2}^*)} = \frac{B_2 - 1 - \sigma_2 x_{1,2}^*}{\sigma_1(1 + \sigma_2 x_{1,2}^*)}. \tag{5.25}$$

(b) If $\Delta > 0$, there are two distinct real solutions. If the following conditions are satisfied, we will have two distinct positive real solutions of $m$:

i. $m_1 + m_2 = -\frac{\mathcal{B}}{\mathcal{A}} > 0$,

ii. $m_1 m_2 = \frac{\mathcal{C}}{\mathcal{A}} > 0$,

iii. $\mathcal{B}^2 - 4\mathcal{A}\mathcal{C} > 0$

In this case, the solution of (5.24) is

$$m = \frac{-\mathcal{B} \pm \sqrt{\mathcal{B}^2 - 4\mathcal{A}\mathcal{C}}}{2\mathcal{A}} \tag{5.26}$$

the solution of $n$ satisfies (5.23). Substitute $x^* = \frac{m}{\beta_1}$ and $y^* = \frac{n}{\sigma_1}$ into the

solution, we have the solution of $(x_1^*, y_1^*)$ and $(x_2^*, y_2^*)$

$$x_{1,2}^* = \frac{-\mathcal{B} \pm \sqrt{\mathcal{B}^2 - 4\mathcal{A}\mathcal{C}}}{2\beta_1 \mathcal{A}} \tag{5.27}$$

$$y_{1,2}^* = \frac{A_2 - 1 - \beta_1 x_{1,2}^*}{\beta_2(1 + \beta_1 x_{1,2}^*)} = \frac{B_2 - 1 - \sigma_2 x_{1,2}^*}{\sigma_1(1 + \sigma_2 x_{1,2}^*)}. \tag{5.28}$$

3. From the proof of both part 2(a) and 2(b), if $\Delta \geq 0$, we have the expression of $y_{1,2}^*$, as follows

$$y_{1,2}^* = \frac{A_2 - 1 - \beta_1 x_{1,2}^*}{\beta_2(1 + \beta_1 x_{1,2}^*)} = \frac{B_2 - 1 - \sigma_2 x_{1,2}^*}{\sigma_1(1 + \sigma_2 x_{1,2}^*)}. \tag{5.29}$$

It is clear that, if $x_{1,2}^* < \frac{A_2 - 1}{\beta_1}$ or $x_{1,2}^* < \frac{B_2 - 1}{\sigma_2}$, the corresponding value of $y_{1,2}^*$ is positive as well.

□

Moreover, to study the bistability, it is necessary to establish the conditions of stability/instability for each equilibrium state. We first give the following conditions for each equilibrium state that locates on an axis

**Theorem 5.2.2.** The model of the $X$-$Y$ system has three equilibria: $(0, 0)$, $(x_e, 0)$ and $(0, y_e)$.

1. The equilibrium state $(0, 0)$ is unstable if $\alpha_1 > k_3$ and $\gamma_1 > k_4$.
2. The equilibrium state $(x_e, 0)$ is stable if $\frac{\gamma_1}{1+\sigma_2 x_e} < k_4$.
3. The equilibrium state $(0, y_e)$ is stable if $\frac{\alpha_1}{1+\beta_2 y_e} < k_3$.

**Proof.** The Jacobian matrix $\boldsymbol{J}_{(x,y)} = [\boldsymbol{J}_{ij}]_{2\times 2}$ of the $X$-$Y$ system is defined by

$$\boldsymbol{J}_{11} = \frac{\partial \dot{x}}{\partial x}|_{(x,y)=(x_0,y_0)} = \frac{\alpha_1}{(1 + \beta_1 x_0)^2} \frac{1}{1 + \beta_2 y_0} - k_3, \tag{5.30}$$

$$\boldsymbol{J}_{12} = \frac{\partial \dot{x}}{\partial y}|_{(x,y)=(x_0,y_0)} = \frac{\alpha_1 x_0}{1 + \beta_1 x_0} \frac{-\beta_2}{(1 + \beta_2 y_0)^2}, \tag{5.31}$$

$$\boldsymbol{J}_{21} = \frac{\partial \dot{y}}{\partial x}|_{(x,y)=(x_0,y_0)} = \frac{\gamma_1 y_0}{1 + \sigma_1 y_0} \frac{-\sigma_2}{(1 + \sigma_2 x_0)^2}, \tag{5.32}$$

$$\boldsymbol{J}_{22} = \frac{\partial \dot{y}}{\partial y}|_{(x,y)=(x_0,y_0)} = \frac{\gamma_1}{(1 + \sigma_1 y_0)^2} \frac{1}{1 + \sigma_2 x_0} - k_4. \tag{5.33}$$

(1). The Jacobian matrix at the equilibrium state $(0,0)$ is

$$\boldsymbol{J}_{(0,0)} = \begin{bmatrix} \alpha_1 - k_3 & 0 \\ \\ 0 & \gamma_1 - k_4 \end{bmatrix}. \tag{5.34}$$

The eigenvalues of the Jacobian matrix are $\lambda_1 = \alpha_1 - k_3$ and $\lambda_2 = \gamma_1 - k_4$. Obviously, the equilibrium state $(0,0)$ is unstable if any one of the following conditions are satisfied

$$\alpha_1 > k_3, \quad \gamma_1 > k_4. \tag{5.35}$$

Notice that the above conditions are also the existence conditions for equilibria $(x_e, 0)$ and $(0, y_e)$ which has been proved in Theorem 5.2.1. In this case, we prove that when $(0,0)$ is an unstable state, there exist two positive equilibria $(x_e, 0)$ and $(0, y_e)$.

(2). The Jacobian matrix at the equilibrium state $(x_e, 0) = (\frac{\alpha_1 - k_3}{k_3 \beta_1}, 0)$ is

$$\boldsymbol{J}_{(x_e,0)} = \begin{bmatrix} \frac{\alpha_1}{(1+\beta_1 x_e)^2} - k_3 & \frac{-\alpha_1 \beta_2 x_e}{1+\beta_1 x_e} \\ \\ 0 & \frac{\gamma_1}{1+\sigma_2 x_e} - k_4 \end{bmatrix}. \tag{5.36}$$

When $x_e \neq 0$, the eigenvalues of the Jacobian matrix are

$$\lambda_1 = \frac{\alpha_1}{(1+\beta_1 x_e)^2} - k_3 = \frac{\alpha_1}{(1+\beta_1 x_e)^2} - \frac{\alpha_1}{1+\beta_1 x_e} = \frac{-\alpha_1 \beta_1 x_e}{(1+\beta_1 x_e)^2}, \tag{5.37}$$

$$\lambda_2 = \frac{\gamma_1}{1+\sigma_2 x_e} - k_4 \tag{5.38}$$

It is clear that $\lambda_1 < 0$. Thus, this equilibrium state is stable if

$$\frac{\gamma_1}{1+\sigma_2 x_e} < k_4. \tag{5.39}$$

(3). The Jacobian matrix at the equilibrium state $(0, y_e) = (0, \frac{\gamma_1 - k_4}{k_4 \sigma_1})$ is

$$\boldsymbol{J}_{(0,y_e)} = \begin{bmatrix} \frac{\alpha_1}{1+\beta_2 y_e} - k_3 & 0 \\ \\ \frac{-\gamma_1 \sigma_2 y_e}{1+\sigma_1 y_e} & \frac{\gamma_1}{(1+\sigma_1 y_e)^2} - k_4 \end{bmatrix}. \tag{5.40}$$

When $y_e \neq 0$, the eigenvalues of the Jacobian matrix are

$$\lambda_1 = \frac{\alpha_1}{1+\beta_2 y_e} - k_3, \tag{5.41}$$

$$\lambda_2 = \frac{\gamma_1}{(1+\sigma_1 y_e)^2} - k_4 = \frac{\gamma_1}{(1+\sigma_1 y_e)^2} - \frac{\gamma_1}{1+\sigma_1 y_e} = \frac{-\gamma_1\sigma_1 y_e}{(1+\sigma_1 y_e)^2} \tag{5.42}$$

It is clear that $\lambda_2 < 0$. Thus, this equilibrium state is stable if

$$\frac{\alpha_1}{1+\beta_2 y_e} < k_3. \tag{5.43}$$

□

In addition, we give the following stable conditions for each equilibrium state that locates within the 2-dimensional positive real space.

**Theorem 5.2.3.** The positive equilibria $(x_1^*, y_1^*)$ and $(x_2^*, y_2^*)$ are stable if the following condition is satisfied.

$$\beta_1\sigma_1\eta_y\xi_x - \beta_2\sigma_2\theta_x\rho_y > 0, \tag{5.44}$$

where $\theta_x = 1+\beta_1 x$, $\eta_y = 1+\beta_2 y$, $\rho_y = 1+\sigma_1 y$ and $\xi_x = 1+\sigma_2 x$.

**Proof.** The Jacobian matrix at the equilibrium state $(x, y)$ is

$$\boldsymbol{J}_{(x,y)} = \begin{bmatrix} \frac{\alpha_1}{(1+\beta_1 x)^2(1+\beta_2 y)} - k_3 & \frac{-\alpha_1\beta_2 x}{(1+\beta_1 x)(1+\beta_2 y)^2} \\ \\ \frac{-\gamma_1\sigma_2 y}{(1+\sigma_1 y)(1+\sigma_2 x)^2} & \frac{\gamma_1}{(1+\sigma_1 y)^2(1+\sigma_2 x)} - k_4 \end{bmatrix} \tag{5.45}$$

When $x$ and $y$ are not zero, we substitute (5.17) and (5.18) into the Jacobian matrix (5.45). Then, we have

$$\boldsymbol{J}_{(x,y)} = \begin{bmatrix} \frac{\alpha_1(1-\theta_x)}{\theta_x^2\eta_y} & \frac{-\alpha_1\beta_2 x}{\theta_x\eta_y^2} \\ \\ \frac{-\gamma_1\sigma_2 y}{\rho_y\xi_x^2} & \frac{\gamma_1(1-\rho_y)}{\rho_y^2\xi_x} \end{bmatrix}. \tag{5.46}$$

The eigenvalues of the Jacobian matrix (5.46) are

$$\lambda = \frac{-\Phi \pm \sqrt{\Phi^2 - \tau}}{2\theta_x^2\eta_y^2\rho_y^2\xi_x^2}, \tag{5.47}$$

where

$$\Phi = \eta_y \xi_x (\rho_y^2 \xi_x \alpha_1 \beta_1 x + \theta_x^2 \eta_y \gamma_1 \sigma_1 y), \tag{5.48}$$

$$\tau = 4\theta_x^2 \eta_y^2 \rho_y^2 \xi_x^2 \alpha_1 \gamma_1 xy(\beta_1 \sigma_1 \eta_y \xi_x - \beta_2 \sigma_2 \theta_x \rho_y). \tag{5.49}$$

For both $(x_1^*, y_1^*)$ and $(x_2^*, y_2^*)$ to be stable, it requires that both eigenvalues are negative or have negative real parts. Thus, the stability conditions of $(x_1^*, y_1^*)$ and $(x_2^*, y_2^*)$ are $\Phi > 0$ and $\tau > 0$. Note that, $\Phi > 0$ is always true, since all values in condition (5.48) are positive. However, $\tau > 0$ if and only if $\beta_1 \sigma_1 \eta_{y^*} \xi_{x^*} - \beta_2 \sigma_2 \theta_{x^*} \rho_{y^*} > 0$. Thus, we have proved that if $\beta_1 \sigma_1 \eta_{y^*} \xi_{x^*} - \beta_2 \sigma_2 \theta_{x^*} \rho_{y^*} > 0$, the positive equilibria $(x_1^*, y_1^*)$ and $(x_2^*, y_2^*)$ are stable.

□

In summary, Theorem 5.2.1 gives the existence conditions of the equilibria for our proposed two-node systems. Theorem 5.2.2 and Theorem 5.2.3 provide necessary conditions for the stability properties of these equilibria. According to these theorems, we can easily check whether the two-node systems have bistability based on the generated samples of model parameters. The proofs of these theorems are given in *Supplementary Information*.

### 5.2.2 Perturbation analysis of bistable models

We have proved that systems (5.9) and (5.10) have bistable steady states under the conditions in Theorem 5.2.2 or Theorem 5.2.3. Next we use the random search method to find the model parameters with which the system has bistable steady states. We first generate a sample for each model parameter from the uniform distribution over the interval $[0, A]$ and then test whether the system with the sampled parameters satisfies the conditions in Theorem 5.2.2 or Theorem 5.2.3. If the conditions are satisfied, we solve nonlinear equations of the system to find the steady states. We test different values of $A$ and find that the system has bistable steady states when $A = 10$. To find more types of bistable states, we test 10000 sets of parameters from the uniform distribution over the interval $[0, 10]$. Table 5.1 gives three types of bistable steady states, namely Case 1: $(x_e, 0)$ and $(0, y_e)$; Case 2: $(x_e, 0)$ and $(x_1^*, y_1^*)$; and Case 3: $(0, y_e)$ and $(x_2^*, y_2^*)$. All stable states in case 1 are located on the coordinate axis. We add a perturbation to each estimated coefficient $c$ as $c^* = [\varepsilon \times (P - 0.5) + 1] \times c$, where $P$ is a uniformly distributed random variable over the interval $[0, 1]$, and $\varepsilon$ is the strength of perturbation. Table 5.2 shows that the two other cases of bistability can be obtained by the perturbed coefficients from Case 1.

t

| | $\alpha_1$ | $\beta_1$ | $\beta_2$ | $\gamma_1$ | $\sigma_1$ | $\sigma_2$ | $k_3$ | $k_4$ | Equilibrium State | Characteristic | Stability |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Type 1** | 4.0252 | 1.1393 | 8.5862 | 7.5202 | 2.1524 | 4.6084 | 0.5924 | 0.6515 | (0.0000, 4.8982) | Nodal Sink | **Stable** |
| | | | | | | | | | (1.6517, 0.1581) | Saddle Point | **Unstable** |
| | | | | | | | | | (5.0862, 0.0000) | Nodal Sink | **Stable** |
| **Type 2** | 8.0486 | 2.0932 | 0.3926 | 8.4293 | 1.1974 | 4.5864 | 1.2308 | 0.3546 | (0.0000, 19.0173) | Nodal Sink | **Stable** |
| | | | | | | | | | (0.2556, 8.3039) | Saddle Point | **Unstable** |
| | | | | | | | | | (1.1683, 2.2871) | Nodal Sink | **Stable** |
| | | | | | | | | | (2.6463, 0.0000) | Saddle Point | **Unstable** |
| **Type 3** | 8.6817 | 3.6357 | 5.7823 | 2.7531 | 1.3007 | 0.5855 | 1.1452 | 1.4741 | (0.0000, 0.6671) | Saddle Poing | **Unstable** |
| | | | | | | | | | (0.3034, 0.4505) | Nodal Sink | **Stable** |
| | | | | | | | | | (1.2241, 0.0676) | Saddle Point | **Unstable** |
| | | | | | | | | | (1.8101, 0.0000) | Nodal Sink | **Stable** |

Table 5.1: **Three types of the bistable model whose stable steady states locate at different positions.** Type 1: two stable states are in the axis; Type 2 and Type 3, one of the stable states is in an axis but the other is out of the axis.

| | $\alpha_1$ | $\beta_1$ | $\beta_2$ | $\gamma_1$ | $\sigma_1$ | $\sigma_2$ | $k_3$ | $k_4$ | Equilibrium State | Characteristic | Stability |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Type 1** | 4.0252 | 1.1393 | 8.5862 | 7.5202 | 2.1524 | 4.6084 | 0.5924 | 0.6515 | (0.0000, 4.8982) | Nodal Sink | **Stable** |
| | | | | | | | | | (1.6517, 0.1581) | Saddle Point | **Unstable** |
| | | | | | | | | | (5.0862, 0.0000) | Nodal Sink | **Stable** |
| **Perturbed case 1** | 4.2582 | 0.3682 | 1.7541 | 11.8512 | 3.8062 | 1.4729 | 0.6912 | 0.4352 | (0.0000, 6.8918) | Nodal Sink | **Stable** |
| | | | | | | | | | (2.5527, 1.2403) | Saddle Point | **Unstable** |
| | | | | | | | | | (9.2824, 0.2249) | Nodal Sink | **Stable** |
| | | | | | | | | | (14.0157, 0.0000) | Saddle Point | **Unstable** |
| **Perturbed case 2** | 4.9263 | 1.6689 | 9.6312 | 0.8750 | 3.1029 | 0.8549 | 0.5510 | 0.2678 | (0.0000, 0.7307) | Saddle Point | **Unstable** |
| | | | | | | | | | (0.2915, 0.5206) | Nodal Sink | **Stable** |
| | | | | | | | | | (1.0317, 0.2372) | Saddle Point | **Unstable** |
| | | | | | | | | | (4.7580, 0.0000) | Nodal Sink | **Stable** |

Table 5.2: **Perturbation analysis with strength $\varepsilon = 1.8$.** Type 1 is the Type 1 case in Table 5.1. Perturbed cases 1 and 2 are obtained from Type 1 by perturbing the model parameters. In these two cases, one the stable state is in an axis but the other is out of the axis.

### 5.2.3 Model development for tristability properties

Figure 5.1E shows structure of the network of three genes is formed by embedding the $X$-$Y$ system into the $Z$-$U$ system. For simplicity, let $u = \mathcal{H}(x, y) = x + y$. Since gene $z$ is negatively regulated by the gene $u$ in the sub-system (5.9), and $u$ is a function of genes $x$ and $y$, the expressions of genes $x$ and $y$ are also negatively regulated by the gene $z$ in the new embedding model. The non-linear vector fields $\mathcal{G}_{1,2}(x, y, \boldsymbol{\Theta}_1, t)$ are then transformed into new non-linear vector fields $\mathcal{R}_{1,2}(x, y, z, \boldsymbol{\Theta}^*, t)$, respectively, which include genes $x$, $y$ and $z$ from two sub-systems with the negative regulations from gene $z$ to genes $x$ and $y$. Using the embedding method (5.5) and sub-system models (5.9, 5.10), we obtain the following model to describe the embedded $X$-$Y$-$Z$ system,

$$\begin{aligned}
\frac{dx}{dt} &= \mathcal{R}_1(x, y, z, \boldsymbol{\Theta}^*, t) = \frac{\alpha_1 x}{1+\beta_1 x}\frac{1}{1+\beta_2 y}\frac{1}{1+d_2 z} - k_3 x, \\
\frac{dy}{dt} &= \mathcal{R}_2(x, y, z, \boldsymbol{\Theta}^*, t) = \frac{\gamma_1 y}{1+\sigma_1 y}\frac{1}{1+\sigma_2 x}\frac{1}{1+d_2 z} - k_4 y, \\
\frac{dz}{dt} &= \mathcal{F}_1(z, u = x + y, \boldsymbol{\Theta}^*, t) = \frac{a_1 z}{1+b_1 z}\frac{1}{1+b_2(x+y)} - k_1 z.
\end{aligned} \tag{5.50}$$

To verify the tristability of model (5.50), we give the following conditions for the existence of equilibria and necessary conditions for the stability properties of these equilibria.

**Theorem 5.2.4.** 1. If $(x_e, 0)$ and $(0, y_e)$ are the equilibria of $X$-$Y$ sub-system and $(z_e, 0)$ is a equilibrium state of $Z$-$U$ sub-system, where $x_e = \frac{\alpha_1 - k_3}{k_3\beta_1}$, $y_e = \frac{\gamma_1 - k_4}{k_4\sigma_1}$ and $z_e = \frac{a_1 - k_1}{k_1 b_1}$, then $(x_e, 0, 0)$, $(0, y_e, 0)$ and $(0, 0, z_e)$ are three equilibria of the embedding $X$-$Y$-$Z$ system.

2. If $(x_1^*, y_1^*)$ and $(x_2^*, y_2^*)$ are two positive equilibria of $X$-$Y$ system as stated in Theorem 5.2.1, then $(x_1^*, y_1^*, 0)$ and $(x_2^*, y_2^*, 0)$ are still two equilibria of the embedding $X$-$Y$-$Z$ system.

**Proof.** From Theorem 5.2.1, we have proved that the $Z$-$U$ system has the equilibrium state $(z_e, 0)$, where $z_e = \frac{a_1 - k_1}{k_1 b_1}$ if $a_1 > k_1$. Moreover, the $X$-$Y$ system has the equilibria $(x_e, 0)$ and $(0, y_e)$, where $x_e = \frac{\alpha_1 - k_3}{k_3\beta_1}$ and $y_e = \frac{\gamma_1 - k_4}{k_4\sigma_1}$, if $\alpha_1 > k_3$ and $\gamma_1 > k_4$. Let us consider the $X$-$Y$-$Z$ system. Suppose the equilibrium state exists. Then we have

$$\frac{\alpha_1 x}{1+\beta_1 x}\frac{1}{1+\beta_2 y}\frac{1}{1+d_2 z} - k_3 x = 0, \tag{5.51}$$

$$\frac{\gamma_1 y}{1+\sigma_1 y}\frac{1}{1+\sigma_2 x}\frac{1}{1+d_2 z} - k_4 y = 0, \tag{5.52}$$

$$\frac{a_1 z}{1+b_1 z}\frac{1}{1+b_2(x+y)} - k_1 z = 0. \tag{5.53}$$

1. (a) When $y = z = 0$, according to (5.51), we have

$$x_e = \frac{\alpha_1 - k_3}{k_3\beta_1}. \tag{5.54}$$

Since $\alpha_1, k_3$ and $\beta_1$ are positive, we have the positive equilibrium solution if $\alpha_1 > k_3$.

(b) When $x = z = 0$, according to (5.52), we have

$$y_e = \frac{\gamma_1 - k_4}{k_4\sigma_1}. \tag{5.55}$$

Since $\gamma_1, k_4$ and $\sigma_1$ are positive, we have the positive equilibrium solution if $\gamma_1 > k_4$.

(c) When $x = y = 0$, according to (5.53), we have

$$z_e = \frac{a_1 - k_1}{k_1 b_1}. \tag{5.56}$$

Since $a_1, k_1$ and $b_1$ are positive, we have the positive equilibrium solution if $a_1 > k_1$.

2. When $z = 0$, the system will reduced to

$$\frac{\alpha_1 x}{1 + \beta_1 x}\frac{1}{1 + \beta_2 y} - k_3 x = 0, \tag{5.57}$$

$$\frac{\gamma_1 y}{1 + \sigma_1 y}\frac{1}{1 + \sigma_2 x} - k_4 y = 0, \tag{5.58}$$

where (5.57) and (5.58) are the same as equations (5.17) and (5.18) in the proof of Theorem 5.2.1.

It is clear to see that, for all cases, the conditions for the existence of these equilibria in the $X$-$Y$-$Z$ system are the same as those in the two bistable sub-systems $Z$-$U$ and $X$-$Y$.

□

This theorem shows that the existence conditions of equilibria in the embedded system are the same as those of the two-node sub-systems. Thus, the information of two-node sub-systems can be directly applied to the embedded system. For each equilibrium state located on the axis, we give the following conditions of stability.

**Theorem 5.2.5.** If $(x_e, 0)$ and $(0, y_e)$ are both stable states of $X$-$Y$ system and $(z_e, 0)$ is a stable state of $Z$-$U$ system.

1. The equilibrium state $(x_e, 0, 0)$ is stable if $\frac{a_1}{1+b_2 x_e} < k_1$.
2. The equilibrium state $(0, y_e, 0)$ is stable if $\frac{a_1}{1+b_2 y_e} < k_1$.
3. The equilibrium state $(0, 0, z_e)$ is stable if $\frac{\alpha_1}{1+d_2 z_e} < k_3$ and $\frac{\gamma_1}{1+d_2 z_e} < k_4$.

**Proof.** The Jacobian matrix $\boldsymbol{J}_{(x,y,z)} = [\boldsymbol{J}_{ij}]_{3\times 3}$ of the $X$-$Y$-$Z$ system is defined by

$$\boldsymbol{J}_{11} = \frac{\partial \dot{x}}{\partial x}|_{(x,y,z)=(x_0,y_0,z_0)} = \frac{\alpha_1}{(1+\beta_1 x_0)^2}\frac{1}{1+\beta_2 y_0}\frac{1}{1+d_2 z_0} - k_3, \tag{5.59}$$

$$\boldsymbol{J}_{12} = \frac{\partial \dot{x}}{\partial y}|_{(x,y,z)=(x_0,y_0,z_0)} = \frac{\alpha_1 x_0}{1+\beta_1 x_0}\frac{-\beta_2}{(1+\beta_2 y_0)^2}\frac{1}{1+d_2 z_0}, \tag{5.60}$$

$$\boldsymbol{J}_{13} = \frac{\partial \dot{x}}{\partial z}|_{(x,y,z)=(x_0,y_0,z_0)} = \frac{\alpha_1 x_0}{1+\beta_1 x_0}\frac{1}{1+\beta_2 y_0}\frac{-d_2}{(1+d_2 z_0)^2}, \tag{5.61}$$

$$\boldsymbol{J}_{21} = \frac{\partial \dot{y}}{\partial x}|_{(x,y,z)=(x_0,y_0,z_0)} = \frac{\gamma_1 y_0}{1+\sigma_1 y_0}\frac{-\sigma_2}{(1+\sigma_2 x_0)^2}\frac{1}{1+d_2 z_0}, \tag{5.62}$$

$$\boldsymbol{J}_{22} = \frac{\partial \dot{y}}{\partial y}|_{(x,y,z)=(x_0,y_0,z_0)} = \frac{\gamma_1}{(1+\sigma_1 y_0)^2}\frac{1}{1+\sigma_2 x_0}\frac{1}{1+d_2 z_0} - k_4, \tag{5.63}$$

$$\boldsymbol{J}_{23} = \frac{\partial \dot{y}}{\partial z}|_{(x,y,z)=(x_0,y_0,z_0)} = \frac{\gamma_1 y_0}{1+\sigma_1 y_0}\frac{1}{1+\sigma_2 x_0}\frac{-d_2}{(1+d_2 z_0)^2}, \tag{5.64}$$

$$\boldsymbol{J}_{31} = \frac{\partial \dot{z}}{\partial x}|_{(x,y,z)=(x_0,y_0,z_0)} = \frac{a_1 z_0}{1+b_1 z_0}\frac{-b_2}{(1+b_2(x_0+y_0))^2}, \tag{5.65}$$

$$\boldsymbol{J}_{32} = \frac{\partial \dot{z}}{\partial y}|_{(x,y,z)=(x_0,y_0,z_0)} = \frac{a_1 z_0}{1+b_1 z_0}\frac{-b_2}{(1+b_2(x_0+y_0))^2}, \tag{5.66}$$

$$\boldsymbol{J}_{33} = \frac{\partial \dot{z}}{\partial z}|_{(x,y,z)=(x_0,y_0,z_0)} = \frac{a_1}{(1+b_1 z_0)^2}\frac{1}{1+b_2(x_0+y_0)} - k_1. \tag{5.67}$$

1. The Jacobian matrix at the equilibrium state $(x_e, 0, 0) = (\frac{\alpha_1 - k_3}{k_3\beta_1}, 0, 0)$ is

$$\boldsymbol{J}_{(x_e,0,0)} = \begin{bmatrix} \frac{\alpha_1}{(1+\beta_1 x_e)^2} - k_3 & \frac{-\alpha_1\beta_2 x_e}{1+\beta_1 x_e} & \frac{-\alpha_1 d_2 x_e}{1+\beta_1 x_e} \\ 0 & \frac{\gamma_1}{1+\sigma_2 x_e} - k_4 & 0 \\ 0 & 0 & \frac{a_1}{1+b_2 x_e} - k_1 \end{bmatrix} \tag{5.68}$$

When $x_e \neq 0$, the eigenvalues of the Jacobian matrix are

$$\lambda_1 = \frac{\alpha_1}{(1+\beta_1 x_e)^2} - k_3 = \frac{\alpha_1}{(1+\beta_1 x_e)^2} - \frac{\alpha_1}{1+\beta_1 x_e} = \frac{-\alpha_1\beta_1 x_e}{(1+\beta_1 x_e)^2}, \tag{5.69}$$

$$\lambda_2 = \frac{\gamma_1}{1+\sigma_2 x_e} - k_4, \tag{5.70}$$

$$\lambda_3 = \frac{a_1}{1+b_2 x_e} - k_1. \tag{5.71}$$

It is clear that $\lambda_1 < 0$ and $\lambda_2 < 0$ since $(x_e, 0) = (\frac{\alpha_1 - k_3}{k_3\beta_1}, 0)$ is a stable state of the $X$-$Y$ system. Thus, this equilibrium state is stable if

$$\frac{a_1}{1 + b_2 x_e} < k_1. \tag{5.72}$$

2. The Jacobian matrix at the equilibrium state $(0, y_e, 0) = (0, \frac{\gamma_1 - k_4}{k_4\sigma_1}, 0)$ is

$$\boldsymbol{J}_{(0,y_e,0)} = \begin{bmatrix} \frac{\alpha_1}{1+\beta_2 y_e} - k_3 & 0 & 0 \\ \frac{-\sigma_2\gamma_1 y_e}{1+\sigma_1 y_e} & \frac{\gamma_1}{(1+\sigma_1 y_e)^2} - k_4 & \frac{-d_2\gamma_1 y_e}{1+\sigma_1 y_e} \\ 0 & 0 & \frac{a_1}{1+b_2 y_e} - k_1 \end{bmatrix} \tag{5.73}$$

When $y_e \neq 0$, the eigenvalues of the Jacobian matrix are

$$\lambda_1 = \frac{\alpha_1}{1 + \beta_2 y_e} - k_3, \tag{5.74}$$

$$\lambda_2 = \frac{\gamma_1}{(1 + \sigma_1 y_e)^2} - k_4 = \frac{\gamma_1}{(1 + \sigma_1 y_e)^2} - \frac{\gamma_1}{1 + \sigma_1 y_e} = \frac{-\gamma_1\sigma_1 y_e}{(1 + \sigma_1 y_e)^2}, \tag{5.75}$$

$$\lambda_3 = \frac{a_1}{1 + b_2 y_e} - k_1. \tag{5.76}$$

It is clear that $\lambda_2 < 0$ and $\lambda_1 < 0$ since $(0, y_e) = (0, \frac{\gamma_1 - k_4}{k_4\sigma_1})$ is a stable state of the $X$-$Y$ system. Thus, this equilibrium state is stable if

$$\frac{a_1}{1 + b_2 y_e} < k_1. \tag{5.77}$$

3. The Jacobian matrix at the equilibrium state $(0, 0, z_e) = (0, 0, \frac{a_1 - k_1}{k_1 b_1})$ is

$$\boldsymbol{J}_{(0,0,z_e)} = \begin{bmatrix} \frac{\alpha_1}{1+d_2 z_e} - k_3 & 0 & 0 \\ 0 & \frac{\gamma_1}{1+d_2 z_e} - k_4 & 0 \\ \frac{-a_1 b_2 z_e}{1+b_1 z_e} & \frac{-a_1 b_2 z_e}{1+b_1 z_e} & \frac{a_1}{(1+b_1 z_e)^2} - k_1 \end{bmatrix} \tag{5.78}$$

When $z_e \neq 0$, the eigenvalues of the Jacobian matrix are

$$\lambda_1 = \frac{\alpha_1}{1+d_2 z_e} - k_3, \tag{5.79}$$

$$\lambda_2 = \frac{\gamma_1}{1+d_2 z_e} - k_4, \tag{5.80}$$

$$\lambda_3 = \frac{a_1}{(1+b_1 z_e)^2} - k_1 = \frac{a_1}{(1+b_1 z_e)^2} - \frac{a_1}{1+b_1 z_e} = \frac{-a_1 b_1 z_e}{(1+b_1 z_e)^2}. \tag{5.81}$$

$$\tag{5.82}$$

It is clear that $\lambda_3 < 0$. Thus, this equilibrium state is stable if

$$\frac{\alpha_1}{1+d_2 z_e} < k_3 \text{ and } \frac{\gamma_1}{1+d_2 z_e} < k_4. \tag{5.83}$$

□

In addition, we give the following stable conditions for each equilibrium state that locates within the 3-dimensional positive real space.

**Theorem 5.2.6.** Suppose $(x^*, y^*)$ is a stable state of $X$-$Y$ system, then the equilibrium state $(x^*, y^*, 0)$ is also a stable state of the $X$-$Y$-$Z$ system if

$$\frac{a_1}{1+b_2(x^*+y^*)} < k_1. \tag{5.84}$$

**Proof.** The Jacobian matrix of $X$-$Y$-$Z$ system $\boldsymbol{J}_{(x,y,z)} = [\boldsymbol{J}_{ij}]_{3\times 3}$ is defined in the proof of Theorem 5.2.5. Assume that $x^*$ and $y^*$ are both non-zero, then the Jacobian matrix at $(x^*, y^*, 0)$ is

$$\boldsymbol{J}_{(x^*,y^*,0)} = \begin{bmatrix} \frac{\alpha_1}{(1+\beta_1 x^*)^2(1+\beta_2 y^*)} - k_3 & -\frac{\alpha_1\beta_2 x^*}{(1+\beta_1 x^*)(1+\beta_2 y^*)^2} & -\frac{\alpha_1 d_2 x^*}{(1+\beta_1 x^*)(1+\beta_2 y^*)} \\ -\frac{\gamma_1\sigma_2 y^*}{(1+\sigma_1 y^*)(1+\sigma_2 x^*)^2} & \frac{\gamma_1}{(1+\sigma_1 y^*)^2(1+\sigma_2 x^*)} - k_4 & -\frac{\gamma_1 d_2 y^*}{(1+\sigma_1 y^*)(1+\sigma_2 x^*)} \\ 0 & 0 & \frac{a_1}{1+b_2(x^*+y^*)} - k_1 \end{bmatrix} \tag{5.85}$$

Substitute (5.17) and (5.18) into the Jacobian matrix (5.85). Since $\theta_x = 1+\beta_1 x$, $\eta_y =$

$1+\beta_2 y$, $\rho_y = 1+\sigma_1 y$ and $\xi_x = 1+\sigma_2 x$. Then, we have

$$\boldsymbol{J}_{(x^*,y^*,0)} = \begin{bmatrix} \frac{\alpha_1(1-\theta_x)}{\theta_x^2\eta_y} & -\frac{\alpha_1\beta_2 x^*}{\theta_x\eta_y^2} & -\frac{\alpha_1 d_2 x^*}{\theta_x\eta_y} \\ -\frac{\gamma_1\sigma_2 y^*}{\rho_y\xi_x^2} & \frac{\gamma_1(1-\rho_y)}{\rho_y^2\xi_x} & -\frac{\gamma_1 d_2 y^*}{\rho_y\xi_x} \\ 0 & 0 & \frac{a_1}{1+b_2(x^*+y^*)} - k_1 \end{bmatrix}. \tag{5.86}$$

The eigenvalues of the Jacobian matrix (5.86) are

$$\lambda_1 = \frac{a_1}{1+b_2(x^*+y^*)} - k_1 \tag{5.87}$$

$$\lambda_{2,3} = \frac{-\Phi \pm \sqrt{\Phi^2 - \tau}}{2\theta_x^2\eta_y^2\rho_y^2\xi_x^2}. \tag{5.88}$$

where $\Phi$ and $\tau$ are defined by (5.48) and (5.49) in the proof of Theorem 5.2.3, as follows:

$$\Phi = \eta_y\xi_x(\rho_y^2\xi_x\alpha_1\beta_1 x + \theta_x^2\eta_y\gamma_1\sigma_1 y), \tag{5.89}$$

$$\tau = 4\theta_x^2\eta_y^2\rho_y^2\xi_x^2\alpha_1\gamma_1 xy(\beta_1\sigma_1\eta_y\xi_x - \beta_2\sigma_2\theta_x\rho_y). \tag{5.90}$$

Since $(x^*, y^*)$ is a stable state of the $X$-$Y$ system, based on Theorem 5.2.3, it is clear that $\tau$ is positive. Moreover, $\Phi$ is always positive. The two eigenvalues $\lambda_2$ and $\lambda_3$ are negative or have negative real part. Thus, the equilibrium state $(x^*, y^*, 0)$ is a stable state of the $X$-$Y$-$Z$ system if

$$\frac{a_1}{1+b_2(x^*+y^*)} < k_1. \tag{5.91}$$

□

Theorem 5.2.5 and Theorem 5.2.6 describe the necessary conditions for the stability properties of the equilibria in the embedding $X$-$Y$-$Z$ system. By applying these theorems, we can further constrain the estimated parameters obtained from two-node systems so that the embedding system can achieve tristability.

## 5.3 Application in hematopoiesis

### 5.3.1 Bistable models for *GATA1-PU.1* and GATA-switching modules

For the two double-negative feedback loops with positive autoregulation in Figure 5.1C and Figure 5.1D, we develop two mathematical models for the $Z$-$U$ module (5.9) and $X$-$Y$ module (5.10). These two models have the same structure but with different model parameters. Theorem 5.2.1 shows that there are five possible non-negative equilibrium states in these models. Theorem 5.2.2 indicates that two steady states located on the axis are stable under the given conditions. In addition, Theorem 5.2.3 gives the conditions under which the two possible steady states out of the axis are stable.

We further search for stable steady states of the model with randomly sampled parameters. Table 5.1 gives three types of bistable steady states. However, we have not found any parameter samples to realize tristability. To test robustness properties, we conduct perturbation tests by examining the bistable property of the model with slightly changed model parameters [59, 60]. Computation results show that, for the model with two stable steady states located on the axis, we can find a perturbed bistable model that has one stable steady state is located on an axis but another is located out of the axis (see Table 5.2). These results suggest that the developed model has very good robustness properties in terms of parameter variations.

We next use the approximate Bayesian computation (ABC) rejection algorithm [10, 139] to estimate model parameters based on the experimental data for erythroiesis and granulopoiesis [151]. The data used here is obtained by using single-molecule RNA fluorescent in situ hybridization (smFISH) on mouse stem cells derived from hematopoietic tissue to measure the transcription dynamics of genes *GATA1*, *GATA2* and PU.1 [151]. We first estimate parameters in the $X$-$Y$ module that describes regulations between genes *GATA1* and *PU.1* (5.10). It is assumed that the prior distribution of each parameter is a uniform distribution over the interval $[0, 100]$. The distance between experimental data and simulations are measured by

$$\rho(\boldsymbol{X}, \boldsymbol{X^*}) = \sum_{i=1}^{m} [|x_i - x_i^*| + |y_i - y_i^*|],$$

where $(x_i, y_i)$ and $(x_i^*, y_i^*)$ are the observed data and simulated data of the model at time point $t_i$ for genes $(X, Y)$, respectively. Table 5.3 gives the estimated parameters of this module. Figure 5.3A shows that the phase plane of the *GATA1-PU.1* sub-system based on estimated parameters, which shows that this system is bistable.

Regarding the $Z$-$U$ module (5.9) that describes the regulation of GATA-switching, to be consistent with the module structure, we first assume that *GATA1* and *GATA2* form a double negative feedback module with autoregulations, and will modify this assumption later based on the experimentally observed mechanisms. Here the data of the auxiliary variable $U$ is the sum of *GATA1* and *PU.1*. Table 5.4 gives the estimated parameters of the $Z$-$U$ module.

| | | | |
|---|---|---|---|
| $\alpha_1 = 15.665$ | $\beta_1 = 0.4263$ | $\beta_2 = 0.9047$ | $k_3 = 0.1587$ |
| $\gamma_1 = 89.4$ | $\sigma_1 = 1.0724$ | $\sigma_2 = 0.4535$ | $k_4 = 0.752$ |

Table 5.3: **Estimated model parameter values for module $X$-$Y$.**

| | | | |
|---|---|---|---|
| $a_1 = 16.5$ | $b_1 = 0.6024$ | $b_2 = 1.1$ | $k_1 = 0.6090$ |
| $c_1 = 4.2934$ | $d_1 = 0.3340$ | $d_2 = 3.6$ | $k_2 = 0.2143$ |

Table 5.4: **Estimated model parameter values for module $Z$-$U$.**

An experimental study has identified *GATA2* at chromatin sites in early-stage erythroblasts [14], when expression levels of *GATA1* increase as erythropoiesis progresses, *GATA1* displaces *GATA2* from chromatin sites. To describe the mechanisms of GATA-switching, we introduce an additional rate constant $k^*$ over a time interval $[t_1, t_2]$ for the displacement rate of *GATA2* proteins during the process of GATA-switching, given by

$$k^* = \begin{cases} k_0^* & t \in [t_1, t_2], \\ 0 & \text{otherwise.} \end{cases} \tag{5.92}$$

Since the displacement of *GATA2* protein increasing, the concentration of *GATA1* proteins around the binding site will increase proportionally to $k^*$. Hence, we use rate $\psi k^* z$ for the increase of *GATA1* during GATA-switching, where $\psi$ is a control parameter to adjust the availability of *GATA1* proteins around chromatin sites. Then the GATA-switching module is modelled by

$$\begin{aligned} \frac{dz}{dt} &= \frac{a_1 z}{1 + b_1 z} \frac{1}{1 + b_2 u} - k_1 z - k^* z, \\ \frac{du}{dt} &= \frac{c_1 u}{1 + d_1 u} \frac{1}{1 + d_2 z} - k_2 u + \psi k^* z, \end{aligned} \tag{5.93}$$

where $z$ and $u$ are expression levels of *GATA2* and *GATA1*, respectively. Note that the bistability property of this module is realized by model (5.93) using $k^* = 0$. Figure 5.3B gives two simulations for an unsuccessful switching and a successful switching. It is assumed

that the GATA-switching occurs over the interval $[t_1, t_2] = [500, 3500]$. Simulations show that an adequate displacement of *GATA2* is the key to achieve GATA-switching using a relatively large value of $k_0^* \leq 1$.

### 5.3.2 Tristable model of the *GATA1-GATA2-PU.1* network

After successfully realizing the bistability in double-negative feedback loops with positive autoregulation, we next incorporate the *GATA1-PU.1* regulatory module into the GATA-switching module to realize the tristability of HSC differentiation. We use expression levels of *GATA1* in the GATA-switching module to represent the total levels of *GATA1* plus *PU.1*, and embed these two modules together (5.50) (see Theorems 5.2.4 to 5.2.6). The model parameters have the same values as the corresponding parameters in the *Z-U* module or the *X-Y* module. Figure 5.4 gives the 3D phase portrait of the embedded system, which shows that the embedding model faithfully realizes the three stable steady states in the two sub-modules, which also suggests that the proposed embedding method is a robust approach to develop high order multistable models based on bistable models.

As mentioned in the previous subsection, the GATA-switching module is not a perfect double-negative feedback loop. In fact, experimental studies suggest that *GATA2* moderately simulates the expression of gene *GATA1* [46] (shown in Figure 5.5). Thus we make a modification to model (5.50) by adding the term $d^*z$ in the first equation to represent a weak positive regulation from *GATA2* to *GATA1*. In addition, to avoid zero basal gene expression levels, we add a constant to each equation of the proposed model (5.50). The modified model is given by,

$$\begin{aligned}
\frac{dx}{dt} &= \frac{\alpha_0 + \alpha_1 x}{1 + \beta_1 x} \frac{1}{1 + \beta_2 y} \frac{1 + d^* z}{1 + d_2 z} - k_3 x + \psi k^* z, \\
\frac{dy}{dt} &= \frac{\gamma_0 + \gamma_1 y}{1 + \sigma_1 y} \frac{1}{1 + \sigma_2 x} \frac{1}{1 + d_2 z} - k_4 y, \\
\frac{dz}{dt} &= \frac{a_0 + a_1 z}{1 + b_1 z} \frac{1}{1 + b_2 (x + y)} - k_1 z - k^* z,
\end{aligned} \tag{5.94}$$

where $x$, $y$, $z$ represent expression levels of genes *GATA1*, *GATA2* and *PU.1*, respectively. The values of $\alpha_0, \gamma_0, a_0$ and $d^*$ are carefully selected so that the model simulation still matches experimental data and the model has at least three stable steady states (see Table 5.5). Figure 5.3C gives the 3D phase portrait of system (5.94) with $k^* = 0$. Using the estimated parameters (see Table 5.3 to Table 5.5), the modified system (5.94) actually achieves quad-stability. In three stable states, one of the three genes has high expression levels but the other two have low expression levels. The fourth stable state has low expression levels $(2.3364, 0.7417, 8.6664)$ of the three genes. In fact, these are exact four transcriptional

| $\alpha_0 = 0.045$ | $\gamma_0 = 0.1$ | $a_0 = 1$ | $d^* = 0.01$ |
|---|---|---|---|

Table 5.5: **Estimated additional model parameter values for modified model.**

states that have been observed in experimental studies, namely a $PU.1^{high}Gata1/2^{low}$ state (P1H); a $Gata1^{high}GATA2/PU.1^{low}$ state (G1H); a $Gata2^{high}GATA1/PU.1^{low}$ state (G2H); and a state with low expression of all three genes (LES CMP) [151]. Compared with existing modelling studies, our embedding model (5.94) for the first time realizes the state with low expression levels of all three genes. Note that the embedding model is based on the assumption of GATA-switching, namely the exchange of *GATA1* for *GATA2* at key chromatin sites, which controls the expression of genes *GATA1* and *GATA2*. However, a low level of *GATA2* at the chromatin site does not mean the total level of *GATA2* in cells is also low. This may be the reason for the difference between the simulated state $Gata1^{high}GATA2/PU.1^{low}$ state (G1H) (namely only *GATA1* has high expression) and the experimentally observed state $Gata1/2^{high}PU.1^{low}$ state (G1/2H) (namely both *GATA1* and *GATA2* have high expression levels) [151].

### 5.3.3 Stochastic model for realizing heterogeneity

Although the modified embedding model has successfully realized the quad-stability properties, this deterministic model cannot describe the heterogeneity in the cell fate commitment. Thus, the next question is whether we can use a stochastic model to realize experimental data showing different gene expression levels in single cells [151]. To answer this question, we propose a stochastic differential equations model in Itô form to describe the functions of noise during the cell lineage specification [155], given by

$$\begin{aligned}
dX(t) &= \left[\frac{\alpha_0+\alpha_1 X(t)}{1+\beta_1 X(t)}\frac{1}{1+\beta_2 Y(t)}\frac{1+d^*Z(t)}{1+d_2 Z(t)} - k_3 X(t) + \psi k^* Z(t)\right]dt + [\omega_1(k_3 X(t) + \psi k^* Z(t))]dW_t^1,\\
dY(t) &= \left[\frac{\gamma_0+\gamma_1 Y(t)}{1+\sigma_1 Y(t)}\frac{1}{1+\sigma_2 X(t)}\frac{1}{1+d_2 Z(t)} - k_4 Y(t)\right]dt + [\omega_2 k_4 Y(t)]dW_t^2,\\
dZ(t) &= \left[\frac{a_0+a_1 Z(t)}{1+b_1 Z(t)}\frac{1}{1+b_2(X(t)+Y(t))} - k_1 Z(t) - k^* Z(t)\right]dt + [\omega_3(k_1+k^*)Z(t)]dW_t^3,
\end{aligned} \tag{5.95}$$

where $W_t^1$, $W_t^2$ and $W_t^3$ are three independent Wiener processes whose increment is a Gaussian random variable $\Delta W_t = W(t+\Delta t) - W(t) \sim N(0, \Delta t)$, and $\omega_1, \omega_2$ and $\omega_3$ represent noise strengths. The reason for selecting Itô form is to maintain the mean of the stochastic system (5.95) as the corresponding deterministic system (5.94). To test the influence

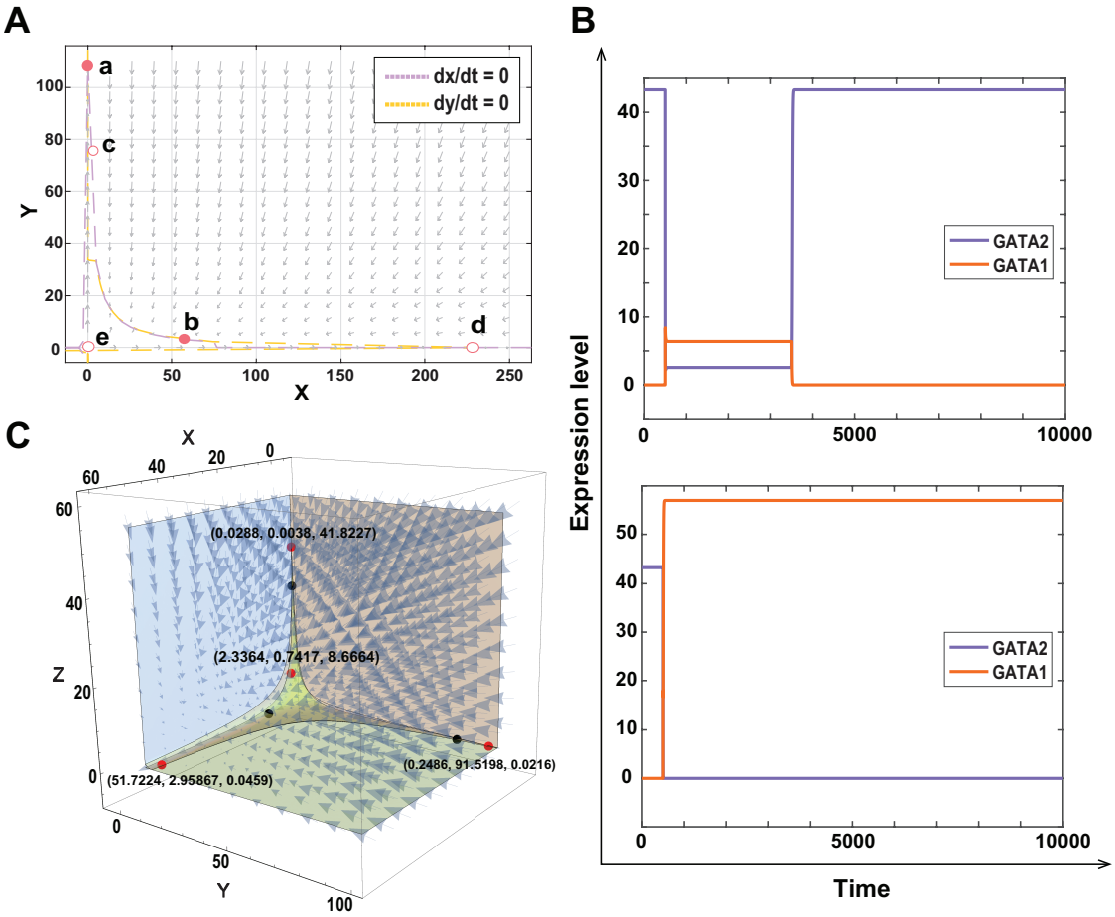


Figure 5.3: **Realization of tristability by embedding two bistable sub-systems in hematopoiesis.**
(A) Phase plane of the *GATA1-PU.1* module showing the bistable property of the proposed model, where a and b are stable steady states; c, d and e are saddle states.
(B) Simulations of GATA-switching of model (5.93). Upper panel: An unsuccessful switching with a small value of $k_0^*$ due to the displacement of *GATA2* not being enough for cells to leave the HSCs state ($Z$ state); Lower panel: A successful switching with sufficient displacement of *GATA2* by using a large value of $k_0^*$. Cells leave the HSCs state and enter the U state.
(C) The 3D phase portrait of the modified embedding model (5.94) with $k^* = 0$. Four red points are stable steady states, while the three black points are saddle states.

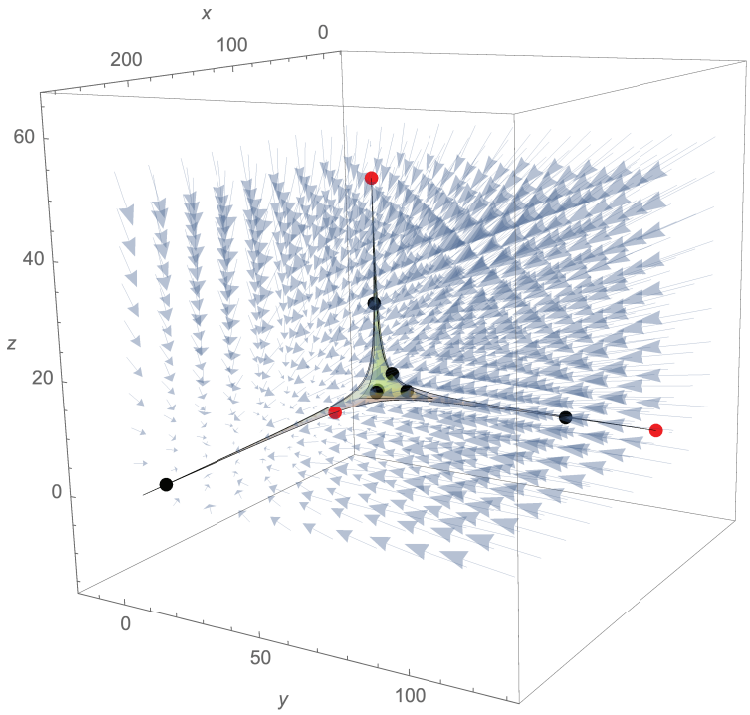


Figure 5.4: **The 3D phase portrait of the embedded system.** Based on the experimental data, the proposed model successfully realize the tristability properties, with the same parameter values presented in the Table 5.3 and Table 5.4. Red points: stable steady states; Black points: saddle states.

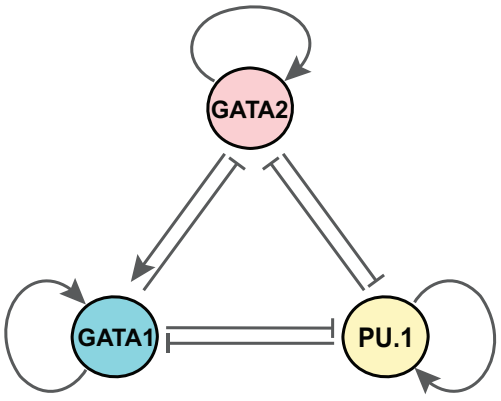


Figure 5.5: **The network structure of GATA1-GATA2-PU.1.** '→' and '⊣' denote the activating and inhibiting regulations, respectively.

of GATA-switching on determining the transitions between different states, we introduce noise to coefficient $k^*$ and consequently to the three degradation processes in the model. We use the semi-implicit Euler method to simulate the proposed model [134]. Figure 5.6 provides four stochastic simulations for four different types of cell fate commitments with model parameters $k_0^* = 0.52$, $\psi = 0.0005$, $\omega_1 = 0.04$, and $\omega_2 = \omega_3 = 0.08$. Figure 5.6A and Figure 5.6B show two simulations of unsuccessful GATA switching when the displacement of *GATA2* is not large enough. However, the sufficient displacement of *GATA2* can trigger successful GATA switching, which leads to either the GMP state with high expression levels of PU.1 in Figure 5.6C or the MEP state with high expression levels of *GATA1* in Figure 5.6D.

To examine the heterogeneity of hematopoiesis with different displacement rates $k_0^*$ and $\psi$ together, we generate 20000 stochastic simulations for each set of $k_0^*$ and $\psi$ values over the range of $[0.04, 1]$ and $[0, 0.001]$, respectively. The ranges of $k_0^*$ and $\psi$ are determined by numerical testing. If all stochastic simulations move to a single stable state for the given $k_0^*$ and $\psi$ values, we change the lower bound and/or upper bound of the value range in order that simulations may move to different stable states for the given $k_0^*$ and $\psi$ values. To show the boundary of parameter space, we also keep certain sets of parameter values with which simulations move to one specific stable state. Figure 5.7A gives proportions of simulations that have successful switching in 20000 simulations. When the value of $k_0^*$ is between 0.1 and 0.2, the displacement speed of *GATA2* is low, which gives limited relief of negative regulation to *PU.1* but *GATA1* increases gradually due to GATA-switching and weak positive regulation from *GATA2* to *GATA1*. Thus nearly all cells choose the MEP state with high expression levels of *GATA1*. However, if the value of $k_0^*$ is larger, the negative regulation from *GATA2* to *PU.1* is eliminated quickly, thus the competition between *GATA1* and *PU.1* will lead cells to different lineages. When the value of $k_0^*$ is relatively large but the value of $\psi$ is relatively small, the increase of *GATA1* is slow due to the smaller value of $\psi$ in GATA-switching. However, the negative regulation from *GATA2* to *PU.1* declines rapidly due to the larger value of $k_0^*$. Thus, Figure 5.7B shows that the combination of larger $k_0^*$ and smaller $\psi$ values allows more cells to move to the GMP lineage with high expression level of *PU.1*. If there is no winner in the competition between *GATA1* and *PU.1*, the cell then goes to the state with low expression levels of three genes (namely LE3G). Figure 5.7C shows that, when the value of $k_0^*$ is larger than 0.2, there are four types of simulations as shown in Figure 5.6 for a set of $k_0^*$ and $\psi$ values. We use a MATLAB package [11] to give the violin plot for the expression distributions of three genes in three different cellular states. The violin plot is a combination of a box plot and a kernel density plot that illustrates data peaks. The violin plots in Figure 5.7D match the experimental observations shown in Fig.1e in [151].

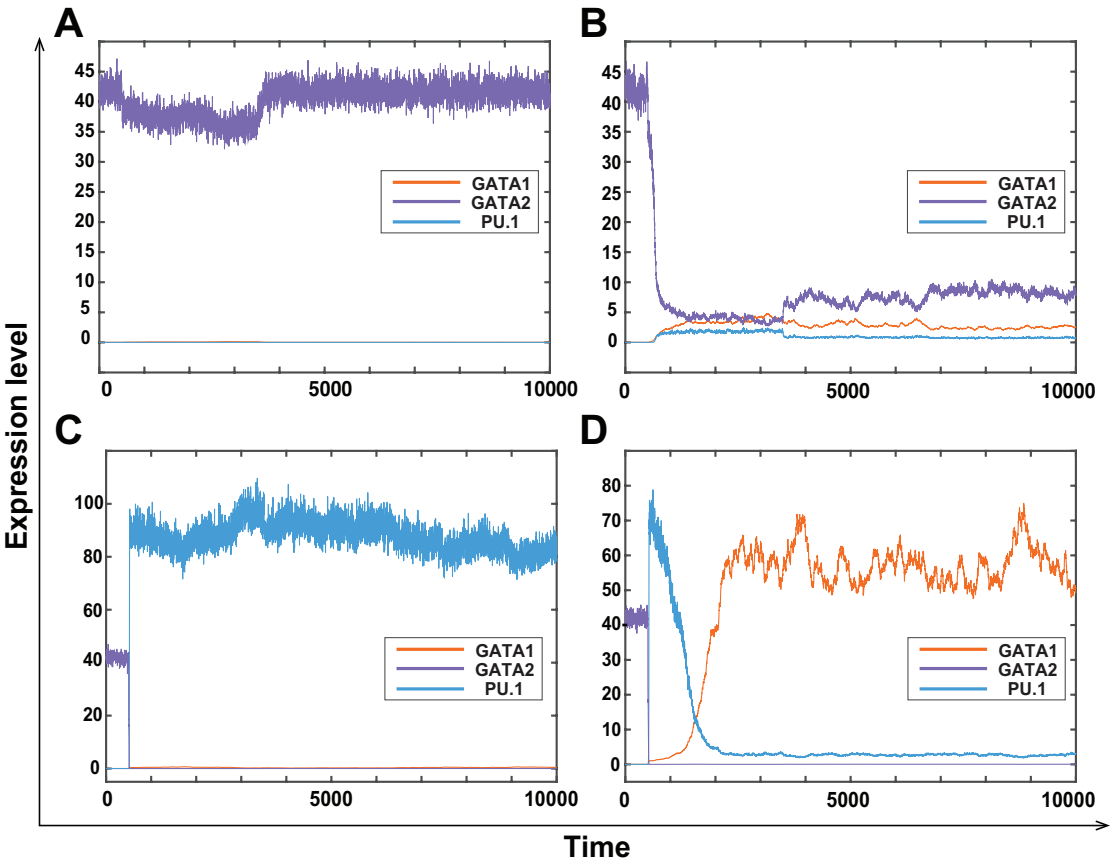


Figure 5.6: **Stochastic simulations showing four stable states that correspond to the experimentally observed four different states.**
(A) Simulation of unsuccessful GATA switching that makes the cell stay at the HSC state, which is the G2H state.
(B) Simulation of unsuccessful GATA switching but the cell enters the state with low expression of all three genes, which is the LES CMP state.
(C) Simulation of successful switching that leads to the GMP state with high expression levels of *PU.1*, which is the P1H state.
(D) Simulation of successful switching that leads to the MEP state with high expression levels of *GATA1*, which is the G1H state.

Regarding the size of basins of attraction, we first calculate the distances between the stable states and saddle points in Figure 5.3C, which are given in Table 5.6. The minimal distance between the G1H state and three saddle points is much larger than the minimal distances of the other three stable states to the saddle points, which suggests that the size of basin of attraction for the G1H state is larger than those of the other three stable states. In addition, we observe the variability of stable states in 20000 stochastic simulations. Table 5.7 shows that the variations of GATA1 in the G1H state are much larger than those of the other two genes when having high expression levels.

We also study the relative frequency of LE3G state. Figure 5.8 shows that, for a fixed

value of parameter $\psi$, the frequency increases as the value of $k_0^*$ increases. In addition, for a fixed value of $k_0^*$, the frequency decreases as the value of $\psi$ increases. The variation of parameter $\psi$ is much more important than that of parameter $k_0^*$. For the simulations showing in Figure 5.7D, the frequency is 0.1080 with $k_0^* = 0.52$ and $\psi = 0.0005$. Figure 5.7D and Figure 5.8 suggest that more cells stay at the LE3G or P1H (GMP) state if *GATA2* leaves the chromatin site fast (i.e. a large $k_0^*$ value) and the expression of *GATA1* is slow (i.e. a small $\psi$ value). However, if the expression of *GATA1* is fast (i.e. a large $\psi$ value), more cells will transit to G1H (MEP) state and the frequency of the LE3G state is low, which is consistent with the results in a recent study [31].

## 5.4 Summary

Waddington's epigenetic landscape is a famous metaphor for how gene regulation drives cell development. Marbles symbolise cells that roll downhill over a landscape of bifurcating valleys. Each new valley represents a possible cell fate, while the ridges between the valleys keep the cell fate after it has been determined [40]. Inspired by Waddington's epigenetic landscape model, we assume that a multistable system makes a series of binary decisions for the selection of multiple evolutionary pathways. Compared with modelling studies for multistable networks, it is relatively easy to develop models with bistability and there is a rich literature for studying the bistable networks [37, 39, 71, 84, 107, 118, 135]. Thus, our proposed embedding method is a novel and effective approach to develop multistable models based on well-studied models with bistable properties. In addition, using the cell fate commitment in hematopoiesis as the test problem, we have successfully realized tristability in the *GATA-PU.1* module by embedding two bistable modules together. More importantly, by modifying the model using the experimentally determined regulatory mechanisms, the developed model, that have no high co-operativity coefficients, successfully realizes four stable states that have been observed recently in a recent experimental study [151].

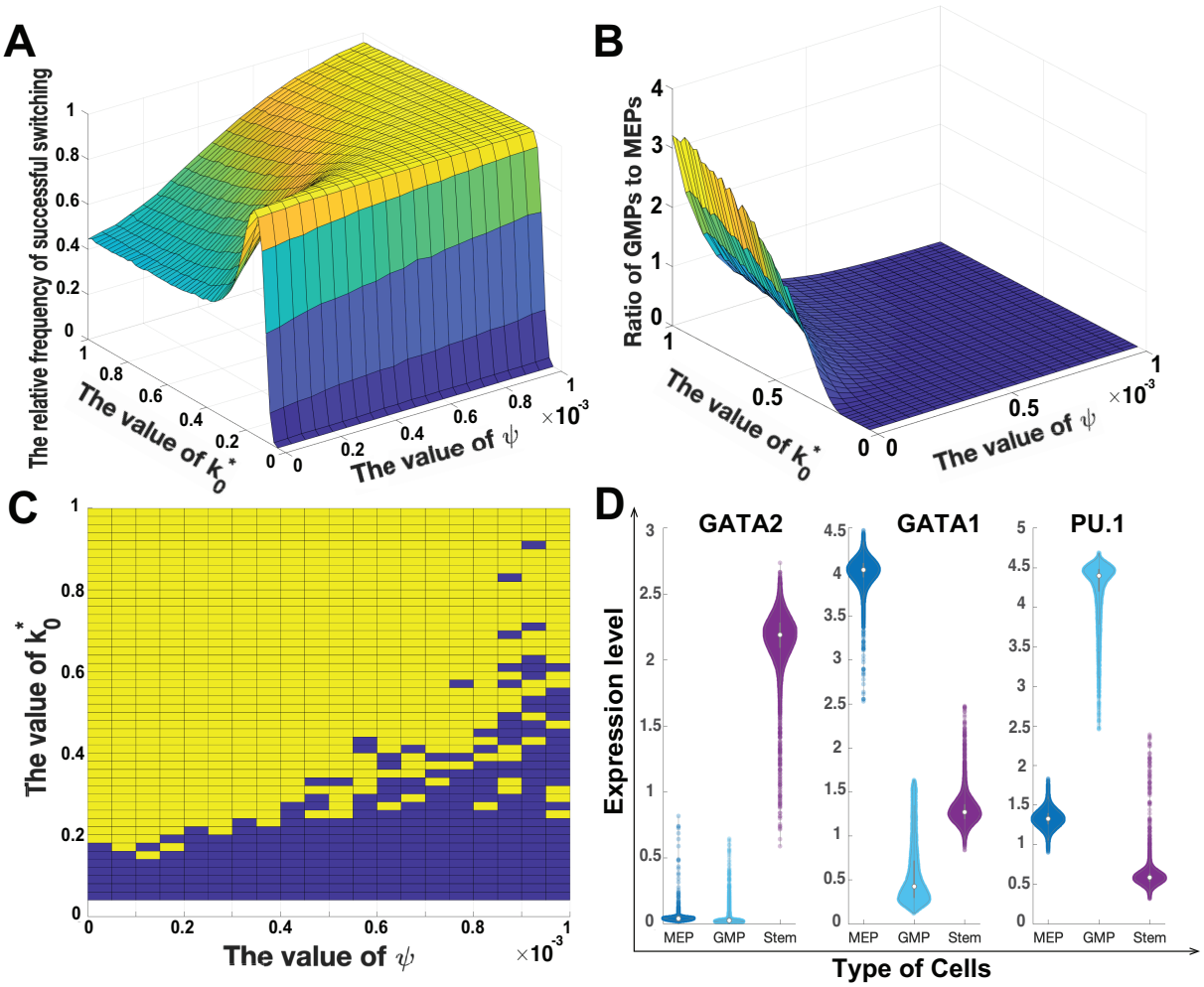


Figure 5.7: **Distributions of different cell types derived from stochastic simulations.**
(A) Frequencies of cells having successful switching for each set of parameters $(k_0^*, \psi)$.
(B) Ratios of GMP cells to MEP cells when the cells have successful switching in (A) for each set of parameters $(k_0^*, \psi)$.
(C) Parameter sets of $(k_0^*, \psi)$ that generate stochastic simulations with four steady states as shown in Figure 5.6 (yellow part) or with two or three states (blue part).
(D) Violin plots of the natural log normalised (expression level per cell + 1) distributions for three genes in different cell states derived from stochastic simulations with parameters $k_0^* = 0.52$ and $\psi = 0.0005$.

| | | Unstable States | | |
|---|---|---|---|---|
| Stable States | | **(0.3170, 0.0100, 31.4818)** | **(0.6179, 78.1130, 0.0265)** | **(13.3231, 2.7597, 0.9070)** |
| **G1H** | **(51.7224, 2.9587, 0.0459)** | 60.3277 | 90.8837 | 38.4095 |
| **P1H** | **(0.2486, 91.5198, 0.0216)** | 96.7667 | 13.4119 | 89.7222 |
| **G2H** | **(0.0288, 0.0038, 41.8227)** | 10.3449 | 88.5907 | 43.1095 |
| **LE3G** | **(2.3364, 0.7417, 8.6664)** | 22.9163 | 77.8712 | 13.6010 |

Table 5.6: **Distances between four stable states and three saddle points shown in the phase portrait of Figure 5.3C - Related to Results.**

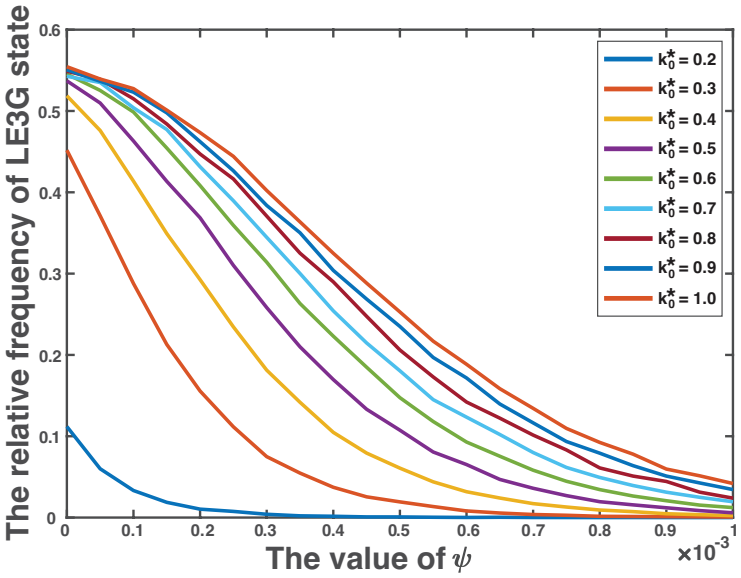


Figure 5.8: **The relative frequency of LE3G state with different values of $k_0^*$ - Related to Results.**

| | | G1H (MEP) State | P1H (GMP) State | G2H (HSC) State | LE3G State |
|---|---|---|---|---|---|
| **GATA1** | **Deterministic Solution** | **51.7224** | **0.2486** | **0.0288** | **2.3364** |
| | **Min** | 34.8137 | 0.1344 | 0.0243 | 1.5030 |
| | **Max** | 77.0405 | 1.0789 | 0.0374 | 4.2895 |
| **PU.1** | **Deterministic Solution** | **2.9587** | **91.5298** | **0.0038** | **0.7414** |
| | **Min** | 1.6972 | 70.6865 | 0.0026 | 0.4675 |
| | **Max** | 4.3167 | 105.4327 | 0.0057 | 1.2441 |
| **GATA2** | **Deterministic Solution** | **0.0459** | **0.0216** | **41.8227** | **8.6664** |
| | **Min** | 0.0265 | 0.0180 | 36.5418 | 4.8823 |
| | **Max** | 0.0868 | 0.0351 | 47.7615 | 12.7441 |

Table 5.7: **The expression variations in stochastic simulations around the four stable states of the corresponding deterministic model - Related to Results.** The deterministic solutions (*GATA1, PU.1, GATA2*) for G1H, P1H, G2H and LE3G states are (51.7224, 2.9587, 0.0459), (0.2486, 91.5298, 0.0216), (0.0288, 0.0038, 41.8227) and (2.3364, 0.7414, 8.6664), respectively (also shown in Figure 5.3C). The minimal/maximal expression levels of each gene are obtained from 20000 stochastic simulations for each state.

# 6 Conclusions and open questions

Understanding the dynamical mechanism of genetic regulatory networks in cell fate determination during hematopoiesis is crucial for biologists to control the cell differential pathways. This is essential for preventing and treating many diseases caused majorly by genetic factors such as leukemia. This thesis has contributed to providing novel and effective methods for describing and studying the dynamical properties of genetic regulation and accurately realizing the experimental results. This doctoral work aims at developing mathematical and computational methods to better understand the detailed mechanism of cell fate determination based on the experimental data [47, 88, 92, 151]. This chapter consists of two parts. The first part reviews and summaries the contribution made by each study. The second part identifies the limitations of each study and remarks some interesting question for future research.

## 6.1 Conclusion

To infer the underlying regulatory network, Chapter 3 and Chapter 4 provide a novel method by combining both top-down approaches (i.e. probabilistic graphical model) and bottom-up approaches (i.e. mathematical model). We first applied the Forward Search Algorithm (FSA) in Chapter 3 and Extended Forward Search Algorithm (EFSA) in Chapter 4 to simplify the network topology and reduced the number of unknown parameters in the mathematical model. Then the Genetic-Algorithm was used to estimate the unknown parameters. The combination of these two approaches reduced the errors in simulation and also improved the robustness property of the mathematical model. We then reduced the network complexity by removing edges from the network, rather than studying the core

network and then adding the edges to the network in Tian's previous study [147]. The reason for changing the method from "adding edge" to "removing edge" in these two works is mainly due to the high computational cost in the "adding edge" tests since the number of candidate edges in the "removing edge test" is much smaller than that in the "adding edge test". Thus, in these two works, we used the FSA and EFSA, respectively, to obtain more candidate edges and then used the dynamic model to remove unimportant edges. If the number of potential regulations derived from the probabilistic graphical model is relatively large, the removal of one single regulation from the potential network may not have any changes in simulation error. Numerical results suggested that a couple of regulations should be removed simultaneously in order to achieve changes in simulation error, especially for the study in Chapter 4.

The inferred regulatory networks from our proposed methods are partially supported by experimental observations. For example, the regulation of *GATA1-GATA2-PU.1* complex in our inferred networks agrees with the experimental results [88]. The *GATA1-PU.1* heterodimer plays an important role in regulating the hematopoiesis [161], which is also included in our inferred model. In addition, the *Ldb1-Lmo2* dimer is activated with significantly expression profiles during the erythroid differentiation process [157], which is consistent with our prediction. However, not all of the predictions can be confirmed by the existing experimental observations, especially for the regulation from protein heterodimers and/or synergistic effects to genes. The first explanation is that the non-linear terms in our mathematical model are introduced by mathematical operation (i.e. the Taylor series). Some of these non-linear terms may be needed for realizing the nonlinear dynamics accurately, but not supported by biological mechanisms. Note that another inference method, called semi-supervised method, can include the validated regulations first and then infer the invalidated regulations [83]. Secondly, our inferred regulatory network may predict some potential possible regulations between genes and from non-linear terms to genes, which may be confirmed by future experimental studies. Thus, the first contribution of these two chapters is the inferred regulations in this work may provide testable prediction for further experimental studies to explore the detailed mechanism of hematopoiesis. Another contribution is that, in Chapter 4, we not only use EFSA to predict network structures, but also use the $2^{nd}$ order truncated Taylor expansion as a dynamic model to study detailed regulatory mechanisms. The proposed EFSA provides a method for inferring network structure with both genes, protein heterodimers and/or synergistic effects. In addition, the truncated Taylor expansion gives our dynamic model the ability to describe a nonlinear system while having relatively few unknown parameters, as a linear model does. Therefore, it provides a new idea to modelling the genetic regulatory networks. The proposed method can be applied to model other regulatory networks and biological systems as well.

In Chapter 5, we first propose an embeddedness principle, then used the toggle switch as a testing system to test the effectiveness of our methodology. We then selected the core driver of cell fate commitment, *GATA1-GATA2-PU1*, as the second testing system to study cell fate determination in hematopoiesis. Despite the assumption of a binary choice in each sub-module, the developed model is able to realize a rich variety of dynamics. Our research suggests that, depending on the properties of bistable systems, the embedding model of two bistable modules may have more than three stable steady states. In addition, using the embedding method in Figure 5.1, the state $U$ is not a meta-stable state but actually disappears from the system. Simulations show that, when the system leaves the high *GATA2* expression state due to GATA-switching, genes *GATA1* and *PU.1* begin to increase their expression levels. Each stochastic simulation will reach one of the steady states with either high *GATA1* levels or high *PU.1* levels or return to the stem cell state. These simulations are consistent with the CLOUD-HSPC model in which differentiation is a process of uncommitted cells in transitory states that gradually acquire uni-lineage priming [47, 70, 142]. In addition, stochastic simulations demonstrate that noise plays a key role in determining different cell differentiation pathways. Therefore, our proposed model successfully contributes a novel approach to develop mathematical models for realising multistability and heterogeneity in complex systems.

## 6.2 Limitations of study and open questions

As discussed in Chapter 3 and Chapter 4, the study raised a number of important issues in the study of genetic regulations. One problem is that our nonlinear model cannot fit all the expression data very well if there is noise in the data. The noise in expression data may increase the simulation error of our proposed model. It is a challenging issue in mathematical modelling, if the noise ratio in expression data is large. Large variations in the data may lead to incorrect inference results. In that case, stochastic modelling therefore may be a more appropriate approach to describe the noise in gene expression data [23, 34, 133]. However, a high dimensional system of stochastic models is difficult to find the analytic solutions. Thus, the development and convergence analysis of numerical schemes for such a huge stochastic system with the network topological information is essential. This is also an open question for future research. In addition, the Gaussian graphical model is based on a covariance matrix, which only measures the linear relationship between different components. Currently, other approaches, such as mutual information and conditional mutual information, have been used to measures both linear and nonlinear correlation between the gene expression data [162, 163, 165]. The question is how to develop the graphical model with nonlinear information? Moreover, this research determines

the regulatory mechanisms based on numerical simulation and robustness property. More information from experimental studies will be important to improve the accuracy of the model and make more reasonable predictions. We may also use other key criteria to select mathematical models, such as Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), and Bayesian factor [55]. The another problem is how to speed up the edge deletion or addition in the selection of Gaussian Graphical model? Since we have the slow network model selection in either adding and deleting edges in Gaussian graphical mode. Finally, there is a high computational cost in using a genetic algorithm to process model parameter estimation. Exploring other efficient inference methods is necessary if we want to apply our method to a large regulatory network with more genes and their protein heterodimers and/or synergistic effects. All these issues will be the interesting topics for future research.

In Chapter 5, even though the proposed model has successfully achieved multistability, some equilibrium states have gene expression levels at zero, which is different from experimental observations. The first issue is therefore the development of mathematical models that are able to understand experimental data accurately. In addition, this work uses differential equation models to determine stable states and then employs corresponding stochastic models to realize the functions of noise in determining the fluctuations of gene expression. However, this continuous stochastic process is different from the discrete nature of the gene expression process. The challenge is how to determine conditions for realizing the multistable properties in stochastic models with discrete bursting processes. When considering the gene expression process, existing gene expression bursting models will use a fixed constant to represent the time step at which bursting occurs and the amount of transcription per burst. A more realistic scenario would be that depending on the propensity of each chemical reaction, the time step at which each burst occurs and the transcription size after it occurs should not be a constant, but a function related to the propensity function. We now need to put aside the perception that the time interval of each bursting occurrence and the amount of transcription are constants, and instead model the bursting process innovatively in a functional form. However, the question is how to determine the distribution of time steps and bursting size? Moreover, in this study the stable states are achieved by a model without high cooperativity (i.e. Hill coefficient $n = 1$). Recently, the dynamics of toggle triad with self-activations have attracted much attention [31, 160]. Mathematical models with high cooperativity have been developed to achieve pentastable, namely a hybrid $X/Y$ state with high $X$, high $Y$ and low $Z$. We tried to realise pentastability by using our proposed model with high cooperativity ($n = 2$ or 3), but numerical tests were not successful. Thus, high cooperativity in self-activation may be essential to realise pentastable. Lastly, as we introduced in Chapter 2, HSCs have

the capacity to differentiate into all blood cells, and different cell types can be considered as different equilibrium states within a system. Therefore, this is an ideal test system to develop mathematical models with multistable dynamics. We achieved tristability or even quadra-stability by embedding two binary systems, which produces results comparable to those obtained in biological experiments [151]. However, the hematopoietic system is complex, and modelling the entire system directly is difficult. If we can identify some bistable modules within the hematopoietic system and find connections between them. In that case, our approach could theoretically construct a higher-order multistable system by embedding more bistable systems, thereby allowing us to describe the mechanism of cell fate determination in hematopoiesis. However, the important question is how to embed more modules with more TFs to develop mathematical models with more stable states. All these questions will be interesting topics for further research. A recent study designed a powerful approach to study the stochastic transitions by using the energy landscape approach [57]. This is also an interesting and important topic for future study.

# Bibliography


[1] Ackers, G. K., Johnson, A. D., and Shea, M. A. (1982). Quantitative model for gene regulation by lambda phage repressor. *Proc. Natl. Acad. Sci. U.S.A.*, 79(4):1129–1133.

[2] Aggarwal, R., Lu, J., Pompili, V. J., and Das, H. (2012). Hematopoietic stem cells: transcriptional regulation, ex vivo expansion and clinical application. *Curr. Mol. Med.*, 12(1):34–49.

[3] Ali Al-Radhawi, M., Del Vecchio, D., and Sontag, E. D. (2019). Multi-modality in gene regulatory networks with slow promoter kinetics. *PLoS Comput. Biol.*, 15(2):1–27.

[4] Angeli, D., Ferrell, J. E., and Sontag, E. D. (2004). Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems. *Proc. Natl. Acad. Sci. U.S.A.*, 101(7):1822–1827.

[5] Apri, M., Molenaar, J., Gee, M. d., and Voorn, G. v. (2010). Efficient Estimation of the Robustness Region of Biological Models with Oscillatory Behavior. *PLoS ONE*, 5(4):e9865.

[6] Baltimore, D. (1970). Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of RNA Tumour Viruses. *Nature*, 226(5252):1209–1211.

[7] Banaji, M. and Pantea, C. (2018). The inheritance of nondegenerate multistationarity in chemical reaction networks. *SIAM J. Appl. Math.*, 78(2):1105–1130.

[8] Bar-Joseph, Z., Gitter, A., and Simon, I. (2012). Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.*, 13(8):552–564.

[9] Bastiaansen, R., Jaïbi, O., Deblauwe, V., Eppinga, M. B., Siteur, K., Siero, E., Mermoz, S., Bouvet, A., Doelman, A., and Rietkerk, M. (2018). Multistability of model and real dryland ecosystems through spatial self-organization. *Proc. Natl. Acad. Sci. U.S.A.*, 115(44):11256–11261.

[10] Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

[11] Bechtold, B. (2015). Violin Plots for Matlab.

[12] Birbrair, A. and Frenette, P. S. (2016). Niche heterogeneity in the bone marrow. *Ann. N. Y. Acad.*, 1370(1):82–96.

[13] Bokes, P., King, J. R., and Loose, M. (2009). A bistable genetic switch which does not require high co-operativity at the promoter: a two-timescale model for the PU.1-GATA-1 interaction. *Math. Med. Biol.*, 26(2):117–32.

[14] Bresnick, E. H., Lee, H.-Y., Fujiwara, T., Johnson, K. D., and Keles, S. (2010). Gata switches as developmental drivers. *J. Biol. Chem.*, 285(41):31087–31093.

[15] Bruijn, M. d. and Dzierzak, E. (2017). Runx transcription factors in the development and function of the definitive hematopoietic system. *Blood*, 129(15):2061–2069.

[16] Cedar, H. and Bergman, Y. (2011). Epigenetics of haematopoietic cell development. *Nat. Rev. Immunol.*, 11(7):478–88.

[17] Chang, A. N., Cantor, A. B., Fujiwara, Y., Lodish, M. B., Droho, S., Crispino, J. D., and Orkin, S. H. (2002). GATA-factor dependence of the multitype zinc-finger protein FOG-1 for its essential role in megakaryopoiesis. *Proc. Natl. Acad. Sci. U.S.A.*, 99(14):9237–9242.

[18] Chang, H. H., Oh, P. Y., Ingber, D. E., and Huang, S. (2006). Multistable and multistep dynamics in neutrophil differentiation. *BMC Cell Biol.*, 7(1):11.

[19] Chapra, S. C. (2012). *Applied Numerical Methods with MATLAB for Engineers and Scientists*. McGraw-Hill.

[20] Chen, S. and Mar, J. C. (2018). Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinform.*, 19(1):232.

[21] Chickarmane, V., Enver, T., and Peterson, C. (2009). Computational modeling of the hematopoietic erythroid-myeloid switch reveals insights into cooperativity, priming, and irreversibility. *PLoS Comput. Biol.*, 5(1):e1000268.

[22] Chipperfield, A. J., Fleming, P. J., and Fonseca, C. M. (1994). Genetic algorithm tools for control systems engineering. In *Proceedings of Adaptive Computing in Engineering Design and Control*, volume 128, page 133.

[23] Chowdhury, A. R., Chetty, M., and Evans, R. (2015). Stochastic S-system modeling of gene regulatory network. *Cogn. Neurodyn.*, 9(5):535–547.

[24] Craciun, G., Tang, Y., and Feinberg, M. (2006). Understanding bistability in complex enzyme-driven reaction networks. *Proc. Natl. Acad. Sci. U.S.A.*, 103(23):8697–8702.

[25] Crick, F. (1970). Central Dogma of Molecular Biology. *Nature*, 227(5258):561–563.

[26] Crick, F. H. (1958). On protein synthesis. In *The Symposia of the Society for Experimental Biology 12*, pages 138–163.

[27] Csete, M. E. and Doyle, J. C. (2002). Reverse engineering of biological complexity. *Science*, 295(5560):1664–1669.

[28] de Jong, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.*, 9(1):67–103.

[29] Del Sol, A. and Jung, S. (2020). The importance of computational modeling in stem cell research. *Trends Biotechnol.*, 39(2):126–136.

[30] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Series. B Stat. Methodol.*, 39(1):1–38.

[31] Duddu, A. S., Sahoo, S., Hati, S., Jhunjhunwala, S., and Jolly, M. K. (2020a). Multistability in cellular differentiation enabled by a network of three mutually repressing master regulators. *J. R. Soc. Interface*, 17(170):20200631.

[32] Duddu, S., Chakrabarti, R., Ghosh, A., and Shukla, P. C. (2020b). Hematopoietic Stem Cell Transcription Factors in Cardiovascular Pathology. *Front. Genet.*, 11:588602.

[33] Duff, C., Smith-Miles, K., Lopes, L., and Tian, T. (2012). Mathematical modelling of stem cell differentiation: the PU.1–GATA-1 interaction. *J. Math. Biol.*, 64(3):449–468.

[34] El Samad, H., Khammash, M., Petzold, L., and Gillespie, D. (2005). Stochastic modelling of gene regulatory networks. *Int. J. Robust Nonlinear Control*, 15(15):691–711.

[35] Eling, N., Morgan, M. D., and Marioni, J. C. (2019). Challenges in measuring and understanding biological noise. *Nat. Rev. Genet.*, 20(9):536–548.

[36] Euler, L. (1768). *Institutiones calculi integralis*. Lipsiae Et Berolini.

[37] Fang, X., Liu, Q., Bohrer, C., Hensel, Z., Han, W., Wang, J., and Xiao, J. (2018). Cell fate potentials and switching kinetics uncovered in a classic bistable genetic switch. *Nat. Commun.*, 9(1):2787.

[38] Feliu, E., Rendall, A. D., and Wiuf, C. (2020). A proof of unlimited multistability for phosphorylation cycles. *Nonlinearity*, 33(11):5629–5658.

[39] Feng, J., Kessler, D. A., Ben-Jacob, E., and Levine, H. (2014). Growth feedback as a basis for persister bistability. *Proc. Natl. Acad. Sci. U.S.A.*, 111(1):544–549.

[40] Ferrell, J. E. (2012). Bistability, bifurcations, and waddington's epigenetic landscape. *Curr. Biol.*, 22(11):R458–R466.

[41] Friedman, A. D. (2007). Transcriptional control of granulocyte and monocyte development. *Oncogene*, 26(47):6816–6828.

[42] Gardner, T. S., Cantor, C. R., and Collins, J. J. (2000). Construction of a genetic toggle switch in Escherichia coli. *Nature*, 403(6767):339–342.

[43] Gekas, C., Rhodes, K. E., Gereige, L. M., Helgadottir, H., Ferrari, R., Kurdistani, S. K., Montecino-Rodriguez, E., Bassel-Duby, R., Olson, E., Krivtsov, A. V., Armstrong, S., Orkin, S. H., Pellegrini, M., and Mikkola, H. K. A. (2009). Mef2C is a lineage-restricted target of Scl/Tal1 and regulates megakaryopoiesis and B-cell homeostasis. *Blood*, 113(15):3461–3471.

[44] Gelens, L., Beri, S., Sande, G. V. d., Mezosi, G., Sorel, M., Danckaert, J., and Verschaffelt, G. (2009). Exploring multistability in semiconductor ring lasers: theory and experiment. *Phys. Rev. Lett.*, 102(19):193904.

[45] Goardon, N., Lambert, J. A., Rodriguez, P., Nissaire, P., Herblot, S., Thibault, P., Dumenil, D., Strouboulis, J., Romeo, P.-H., and Hoang, T. (2006). Eto2 coordinates cellular proliferation and differentiation during erythropoiesis. *EMBO J.*, 25(2):357–366.

[46] Grass, J. A., Boyer, M. E., Pal, S., Wu, J., Weiss, M. J., and Bresnick, E. H. (2003). GATA-1-dependent transcriptional repression of GATA-2 via disruption of positive autoregulation and domain-wide chromatin remodeling. *Proc. Natl. Acad. Sci. U.S.A.*, 100(15):8811–8816.

[47] Hamey, F. K., Nestorowa, S., Kinston, S. J., Kent, D. G., Wilson, N. K., and Göttgens, B. (2017). Reconstructing blood stem cell regulatory network models from single-cell molecular profiles. *Proc. Natl. Acad. Sci. U.S.A.*, 114(23):5822–5829.

[48] Harrington, H. A., Feliu, E., Wiuf, C., and Stumpf, M. P. (2013). Cellular compartments cause multistability and allow cells to process more information. *Biophys. J.*, 104(8):1824–1831.

[49] Higham, D. J. (2001). An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Review*, 43(3):525–546.

[50] Hill, A. V. (1910). The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *J. Physiol.*, 40(Suppl):iv–vii.

[51] Hoppe, P. S., Schwarzfischer, M., Loeffler, D., Kokkaliaris, K. D., Hilsenbeck, O., Moritz, N., Endele, M., Filipczyk, A., Gambardella, A., Ahmed, N., Etzrodt, M., Coutu, D. L., Rieger, M. A., Marr, C., Strasser, M. K., Schauberger, B., Burtscher, I., Ermakova, O., Bürger, A., Lickert, H., Nerlov, C., Theis, F. J., and Schroeder, T. (2016). Early myeloid lineage choice is not initiated by random PU.1 to GATA1 protein ratios. *Nature*, 535(7611):299–302.

[52] Huang, S., Guo, Y.-P., May, G., and Enver, T. (2007). Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Dev. Biol.*, 305(2):695–713.

[53] Ingusci, S., Verlengia, G., Soukupova, M., Zucchini, S., and Simonato, M. (2019). Gene Therapy Tools for Brain Diseases. *Front. Pharmacol.*, 10:724.

[54] Inoue, A., Fujiwara, T., Okitsu, Y., Katsuoka, Y., Fukuhara, N., Onishi, Y., Ishizawa, K., and Harigae, H. (2013). Elucidation of the role of lmo2 in human erythroid cells. *Exp. Hematol.*, 41(12):1062–1076.e1.

[55] Kadane, J. B. and Lazar, N. A. (2004). Methods and criteria for model selection. *J. Am. Stat. Assoc.*, 99(465):279–290.

[56] Kaneko, H., Shimizu, R., and Yamamoto, M. (2010). GATA factor switching during erythroid differentiation. *Curr. Opin. Hematol.*, 17(3):163–168.

[57] Kang, X. and Li, C. (2021). A Dimension Reduction Approach for Energy Landscape: Identifying Intermediate States in Metabolism-EMT Network. *Adv. Sci.*, 8(10):2003133.

[58] Kelso, J. A. S. (2012). Multistability and metastability: Understanding dynamic coordination in the brain. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 367(1591):906–918.

[59] Kitano, H. (2004). Biological robustness. *Nat. Rev. Genet.*, 5(11):826–37.

[60] Kitano, H. (2007). Towards a theory of biological robustness. *Mol. Syst. Biol.*, 3(1):137.

[61] Kleden, P. E. and Platen, E. (1999). *Numerical Solution of Stochastic Differential Equations*. Spinger.

[62] Kobayashi, H., Kærn, M., Araki, M., Chung, K., Gardner, T. S., Cantor, C. R., and Collins, J. J. (2004). Programmable cells: Interfacing natural and engineered gene networks. *Proc. Natl. Acad. Sci. U.S.A.*, 101(22):8414–8419.

[63] Kothamachu, V. B., Feliu, E., Cardelli, L., and Soyer, O. S. (2015). Unlimited multistability and Boolean logic in microbial signalling. *J. R. Soc. Interface*, 12(108).

[64] Krämer, N., Schäfer, J., and Boulesteix, A.-L. (2009). Regularized estimation of large-scale gene association networks using graphical Gaussian models. *BMC Bioinform.*, 10(1):384.

[65] Kumano, K., Chiba, S., Shimizu, K., Yamagata, T., Hosoya, N., Saito, T., Takahashi, T., Hamada, Y., and Hirai, H. (2001). Notch1 inhibits differentiation of hematopoietic cells by sustaining gata-2 expression. *Blood*, 98(12):3283–3289.

[66] Lancrin, C., Mazan, M., Stefanska, M., Patel, R., Lichtinger, M., Costa, G., Vargel, O., Wilson, N. K., Möröy, T., Bonifer, C., Göttgens, B., Kouskoff, V., and Lacaud, G. (2012). GFI1 and GFI1B control the loss of endothelial identity of hemogenic endothelium during hematopoietic commitment. *Blood*, 120(2):314–322.

[67] Landa, H., Schiró, M., and Misguich, G. (2020). Multistability of driven-dissipative quantum spins. *Phys. Rev. Lett.*, 124(4):043601.

[68] Larger, L., Penkovsky, B., and Maistrenko, Y. (2015). Laser chimeras as a paradigm for multistable patterns in complex systems. *Nat. Commun.*, 6(1):7752.

[69] Laslo, P., Spooner, C. J., Warmflash, A., Lancki, D. W., Lee, H.-J., Sciammas, R., Gantner, B. N., Dinner, A. R., and Singh, H. (2006). Multilineage transcriptional priming and determination of alternate hematopoietic cell fates. *Cell*, 126(4):755–766.

[70] Laurenti, E. and Göttgens, B. (2018). From haematopoietic stem cells to complex differentiation landscapes. *Nature*, 553(7689):418–426.

[71] Lebar, T., Bezeljak, U., Golob, A., Jerala, M., Kadunc, L., Pirš, B., Stražar, M., Vučko, D., Zupančič, U., Benčina, M., Forstnerič, V., Gaber, R., Lonzarić, J., Majerle, A., Oblak, A., Smole, A., and Jerala, R. (2014). A bistable genetic switch based on designable DNA-binding domains. *Nat. Commun.*, 5(1):5007.

[72] Lee, T. I. and Young, R. A. (2013). Transcriptional Regulation and Its Misregulation in Disease. *Cell*, 152(6):1237–1251.

[73] Li, C. and Wang, J. (2013). Quantifying cell fate decisions for differentiation and reprogramming of a human stem cell network: landscape and biological paths. *PLoS Comput. Biol.*, 9(8):e1003165.

[74] Li, L., Jothi, R., Cui, K., Lee, J. Y., Cohen, T., Gorivodsky, M., Tzchori, I., Zhao, Y., Hayes, S. M., Bresnick, E. H., Zhao, K., Westphal, H., and Love, P. E. (2011). Nuclear adaptor Ldb1 regulates a transcriptional program essential for the maintenance of hematopoietic stem cells. *Nat. Immunol.*, 12(2):129–136.

[75] Li, Q., Wennborg, A., Aurell, E., Dekel, E., Zou, J.-Z., Xu, Y., Huang, S., and Ernberg, I. (2016). Dynamics inside the cancer cell attractor reveal cell heterogeneity, limits of stability, and escape. *Proc. Natl. Acad. Sci. U.S.A.*, 113(10):2672–2677.

[76] Liew, C. W., Rand, K. D., Simpson, R. J. Y., Yung, W. W., Mansfield, R. E., Crossley, M., Proetorius-Ibba, M., Nerlov, C., Poulsen, F. M., and Mackay, J. P. (2006). Molecular analysis of the interaction between the hematopoietic master transcription factors gata-1 and pu.1. *J. Biol. Chem.*, 281(38):28296–28306.

[77] Ling, K.-W., Ottersbach, K., van Hamburg, J. P., Oziemlak, A., Tsai, F.-Y., Orkin, S. H., Ploemacher, R., Hendriks, R. W., and Dzierzak, E. (2004). Gata-2 plays two functionally distinct roles during the ontogeny of hematopoietic stem cells. *J. Exp. Med.*, 200(7):871–882.

[78] Liu, P. and Wang, F. (2008). Inference of biochemical network models in S-system using multi-objective optimization approach. *Bioinformatics*, 24(8):1085–1092.

[79] Liu, Q., Herman, P. M. J., Mooij, W. M., Huisman, J., Scheffer, M., Olff, H., and Van de Koppel, J. (2014). Pattern formation at multiple spatial scales drives the resilience of mussel bed ecosystems. *Nat. Commun.*, 5(1):5234.

[80] Lodish, H., Berk, A., Kaiser, C. A., Krieger, M., Bretscher, A., Ploegh, H., Martin, K. C., Yaffe, M. B., and Amon, A. (2021). *Molecular Cell Biology*. Macmillan Learning, New York, NY.

[81] Lulli, V., Romania, P., Morsilli, O., Gabbianelli, M., Pagliuca, A., Mazzeo, S., Testa, U., Peschle, C., and Marziali, G. (2006). Overexpression of ets-1 in human hematopoietic progenitor cells blocks erythroid and promotes megakaryocytic differentiation. *Cell Death Differ.*, 13(7):1064–74.

[82] Mackey, M. C. (2020). Periodic hematological disorders: Quintessential examples of dynamical diseases. *Chaos*, 30(6):063123.

[83] Maetschke, S. R., Madhamshettiwar, P. B., Davis, M. J., and Ragan, M. A. (2013). Supervised, semi-supervised and unsupervised inference of gene regulatory networks. *Brief Bioinform.*, 15(2):195–211.

[84] Maity, I., Wagner, N., Mukherjee, R., Dev, D., Peacock-Lopez, E., Cohen-Luria, R., and Ashkenasy, G. (2019). A chemically fueled non-enzymatic bistable network. *Nat. Commun.*, 10(1):4636.

[85] Mancini, E., Sanjuan-Pla, A., Luciani, L., Moore, S., Grover, A., Zay, A., Rasmussen, K. D., Luc, S., Bilbao, D., O'Carroll, D., Jacobsen, S. E., and Nerlov, C. (2012). FOG-1 and GATA-1 act sequentially to specify definitive megakaryocytic and erythroid progenitors. *EMBO J.*, 31(2):351–365.

[86] Masel, J. and Siegal, M. L. (2009). Robustness: mechanisms and consequences. *Trends Genet.*, 25(9):395–403.

[87] Matharu, N. and Ahituv, N. (2020). Modulating gene regulation to treat genetic disorders. *Nat. Rev. Drug Discov.*, 19(11):757–775.

[88] May, G., Soneji, S., Tipping, A. J., Teles, J., McGowan, S. J., Wu, M., Guo, Y., Fugazza, C., Brown, J., Karlsson, G., Pina, C., Olariu, V., Taylor, S., Tenen, D. G., Peterson, C., and Enver, T. (2013). Dynamic analysis of gene expression and genome-wide transcription factor binding during lineage specification of multipotent progenitors. *Cell Stem Cell*, 13(6):754–768.

[89] McKnight, S. and Schibler, U. (1995). Differentiation and gene regulation. *Curr. Opin. Genet. Dev.*, 5(5):549–551.

[90] Meek, C. (1995). Causal inference and causal explanation with background knowledge. *Uncertainty in Artificial Intelligence*, 11:403–410.

[91] Michaelis, L., Menten, M. L., Johnson, K. A., and Goody, R. S. (2011). The original michaelis constant: translation of the 1913 michaelis-menten paper. *Biochemistry*, 50(39):8264–8269.

[92] Missal, K., Cross, M. A., and Drasdo, D. (2006). Gene network inference from incomplete expression data: transcriptional control of hematopoietic commitment. *Bioinformatics*, 22(6):731–738.

[93] Moignard, V., Macaulay, I. C., Swiers, G., Buettner, F., Schütte, J., Calero-Nieto, F. J., Kinston, S., Joshi, A., Hannah, R., Theis, F. J., Jacobsen, S. E., Bruijn, M. F. d., and Göttgens, B. (2013). Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nat. Cell Biol.*, 15(4):363–372.

[94] Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A. J., Tanaka, Y., Wilkinson, A. C., Buettner, F., Macaulay, I. C., Jawaid, W., Diamanti, E., Nishikawa, S.-I., Piterman, N., Kouskoff, V., Theis, F. J., Fisher, J., and Göttgens, B. (2015). Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol.*, 33(3):269–276.

[95] Mojtahedi, M., Skupin, A., Zhou, J., Castaño, I. G., Leong-Quong, R. Y. Y., Chang, H., Trachana, K., Giuliani, A., and Huang, S. (2016). Cell fate decision as high-dimensional critical state transition. *PLoS Biol.*, 14(12):1–28.

[96] Narula, J., Williams, C., Tiwari, A., Marks-Bluth, J., Pimanda, J. E., and Igoshin, O. A. (2013). Mathematical model of a gene regulatory network reconciles effects of genetic perturbations on hematopoietic stem cell emergence. *Dev. Biol.*, 379(2):258–269.

[97] Ng, A. P. and Alexander, W. S. (2017). Haematopoietic stem cells: past, present and future. *Cell Death Discov.*, 3(1):17002.

[98] Nguyen, H., Tran, D., Tran, B., Pehlivan, B., and Nguyen, T. (2021). A comprehensive survey of regulatory network inference methods using single-cell RNA sequencing data. *Brief. Bioinformatics*, 22(3):bbaa190.

[99] Noor, A., Serpedin, E., Nounou, M., Nounou, H., Mohamed, N., and Chouchane, L. (2013). An overview of the statistical methods used for inferring gene regulatory networks and protein-protein interaction networks. *Adv. Bioinform.*, 2013:953814.

[100] North, T. E., Stacy, T., Matheny, C. J., Speck, N. A., and Bruijn, M. F. d. (2004). Runx1 is expressed in adult mouse hematopoietic stem cells and differentiating myeloid and lymphoid cells, but not in maturing erythroid cells. *Stem Cells*, 22(2):158–168.

[101] Novère, N. L. (2015). Quantitative and logic modelling of molecular and gene networks. *Nat. Rev. Genet.*, 16(3):146–158.

[102] Ocone, A., Haghverdi, L., Mueller, N. S., and Theis, F. J. (2015). Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics*, 31(12):i89–i96.

[103] Olariu, V. and Peterson, C. (2019). Kinetic models of hematopoietic differentiation. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, 11(1):e1424.

[104] Orkin, S. H. and Zon, L. I. (2008). Hematopoiesis: an evolving paradigm for stem cell biology. *Cell*, 132(4):631–644.

[105] Ottersbach, K., Smith, A., Wood, A., and Göttgens, B. (2010). Ontogeny of haematopoiesis: recent advances and open questions. *Br. J. Haematol.*, 148(3):343–55.

[106] Ozbudak, E. M., Thattai, M., Lim, H. N., Shraiman, B. I., and Oudenaarden, A. v. (2004). Multistability in the lactose utilization network of Escherichia coli. *Nature*, 427(6976):737–740.

[107] Perez-Carrasco, R., Barnes, C. P., Schaerli, Y., Isalan, M., Briscoe, J., and Page, K. M. (2018). Combining a toggle switch and a repressilator within the AC-DC circuit generates distinct dynamical behaviors. *Cell Syst.*, 6(4):521–530.e3.

[108] Pisarchik, A. N. and Feudel, U. (2014). Control of multistability. *Phys. Rep.*, 540(4):167–218.

[109] Porcher, C., Swat, W., Rockwell, K., Fujiwara, Y., Alt, F. W., and Orkin, S. H. (1996). The T cell leukemia oncoprotein SCL/tal-1 is essential for development of all hematopoietic lineages. *Cell*, 86(1):47–57.

[110] Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., and Murali, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods*, 17(2):147–154.

[111] Quarteroni, A., Sacco, R., and Saleri, F. (2006). *Numerical Mathematics*. Springer.

[112] Real, P. J., Ligero, G., Ayllon, V., Ramos-Mejia, V., Bueno, C., Gutierrez-Aranda, I., Navarro-Montero, O., Lako, M., and Menendez, P. (2012). SCL/TAL1 regulates hematopoietic specification from human embryonic stem cells. *Mol. Ther.*, 20(7):1443–1453.

[113] Rieger, M. A. and Schroeder, T. (2012). Hematopoiesis. *Cold Spring Harb. Perspect. Biol.*, 4(12):a008250.

[114] Roeder, I. and Glauche, I. (2006). Towards an understanding of lineage specification in hematopoietic stem cells: A mathematical model for the interaction of transcription factors GATA-1 and PU.1. *J. Theor. Biol.*, 241(4):852–865.

[115] Rung, J. and Brazma, A. (2013). Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.*, 14(2):89–99.

[116] Saint-Antoine, M. M. and Singh, A. (2020). Network inference in systems biology: recent developments, challenges, and applications. *Curr. Opin. Biotechnol.*, 63:89–98.

[117] Santos-Moreno, J., Tasiudi, E., Stelling, J., and Schaerli, Y. (2020). Multistable and dynamic CRISPRi-based synthetic circuits. *Nat. Commun.*, 11(1):2746.

[118] Semenov, S. N., Kraft, L. J., Ainla, A., Zhao, M., Baghbanzadeh, M., Campbell, V. E., Kang, K., Fox, J. M., and Whitesides, G. M. (2016). Autocatalytic, bistable, oscillatory networks of biologically relevant organic reactions. *Nature*, 537(7622):656–660.

[119] Semrau, S., Goldmann, J. E., Soumillon, M., Mikkelsen, T. S., Jaenisch, R., and Oudenaarden, A. v. (2017). Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. *Nat. Commun.*, 8(1):1096.

[120] Shea, M. A. and Ackers, G. K. (1985). The or control system of bacteriophage lambda. a physical-chemical model for gene regulation. *J. Mol. Biol.*, 181(2):211–30.

[121] Shivdasani, R. A. (2006). MicroRNAs: regulators of gene expression and cell differentiation. *Blood*, 108(12):3646–3653.

[122] Shivdasani, R. A., Mayer, E. L., and Orkin, S. H. (1995). Absence of blood formation in mice lacking the T-cell leukaemia oncoprotein Tal1/SCL. *Nature*, 373(6513):432–434.

[123] Skinnider, M. A., Squair, J. W., and Foster, L. J. (2019). Evaluating measures of association for single-cell transcriptomics. *Nat. Methods*, 16(5):381–386.

[124] Snow, J. W., Trowbridge, J. J., Johnson, K. D., Fujiwara, T., Emambokus, N. E., Grass, J. A., Orkin, S. H., and Bresnick, E. H. (2011). Context-dependent function of "GATA switch" sites in vivo. *Blood*, 117(18):4769–4772.

[125] Soler, E., Andrieu-Soler, C., de Boer, E., Bryne, J. C., Thongjuea, S., Stadhouders, R., Palstra, R.-J., Stevens, M., Kockx, C., van Ijcken, W., Hou, J., Steinhoff, C., Rijkers, E., Lenhard, B., and Grosveld, F. (2010). The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. *Genes Dev.*, 24(3):277–289.

[126] Stewart, J. (2018). *Calculus*, chapter Infinite Sequences and Series. Cengage Learning.

[127] Stiehl, T. and Marciniak-Czochra, A. (2012). Mathematical Modeling of Leukemogenesis and Cancer Stem Cell Dynamics. *Math. Model. Nat. Phenom.*, 7(1):166–202.

[128] Stier, S., Cheng, T., Dombkowski, D., Carlesso, N., and Scadden, D. T. (2002). Notch1 activation increases hematopoietic stem cell self-renewal in vivo and favors lymphoid over myeloid lineage outcome. *Blood*, 99(7):2369–2378.

[129] Stumpf, M. P. (2021). Inferring better gene regulation networks from single-cell data. *Curr. Opin. Syst. Biol.*, 27:100342.

[130] Temin, H. M. and Mizutani, S. (1970). Viral RNA-dependent DNA polymerase: RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature*, 226(5252):1211–1213.

[131] The Gene Ontology Consortium (2017). Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res.*, 45(D1):D331 – D338.

[132] Thomson, M. and Gunawardena, J. (2009). Unlimited multistability in multisite phosphorylation systems. *Nature*, 460(7252):274–277.

[133] Tian, T. (2010). Stochastic models for inferring genetic regulation from microarray gene expression data. *Biosystems*, 99(3):192–200.

[134] Tian, T. and Burrage, K. (2001). Implicit taylor methods for stiff stochastic differential equations. *Appl. Numer. Math.*, 38(1):167–185.

[135] Tian, T. and Burrage, K. (2006). Stochastic models for regulatory networks of the genetic toggle switch. *Proc. Natl. Acad. Sci. U.S.A.*, 103(22):8372–8377.

[136] Tian, T. and Smith-Miles, K. (2014a). Mathematical modeling of GATA-switching for regulating the differentiation of hematopoietic stem cell. *BMC Syst. Biol.*, 8(Suppl 1):S8.

[137] Tian, T. and Smith-Miles, K. (2014b). Mathematical modeling of gata-switching for regulating the differentiation of hematopoietic stem cell. *BMC Syst. Biol.*, 8 Suppl 1:S8.

[138] Tindemans, I., Serafini, N., Di Santo, J. P., and Hendriks, R. W. (2014). GATA-3 Function in Innate and Adaptive Immunity. *Immunity*, 41(2).

[139] Turner, B. M. and Van Zandt, T. (2012). A tutorial on approximate bayesian computation. *J. Math. Psychol.*, 56(2):69–85.

[140] Uhler, C. (2017). Gaussian Graphical Models: An Algebraic and Geometric Perspective. *arXiv*.

[141] van der Meer, L. T., Jansen, J. H., and van der Reijden, B. A. (2010). Gfi1 and Gfi1b: key regulators of hematopoiesis. *Leukemia*, 24(11):1834–1843.

[142] Velten, L., Haas, S. F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B. P., Hirche, C., Lutz, C., Buss, E. C., Nowak, D., Boch, T., Hofmann, W.-K., Ho, A. D., Huber, W., Trumpp, A., Essers, M. A. G., and Steinmetz, L. M. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.*, 19(4):271–281.

[143] Viger, R. S., Guittot, S. M., Anttonen, M., Wilson, D. B., and Heikinheimo, M. (2008). Role of the GATA Family of Transcription Factors in Endocrine Development, Function, and Disease. *Mol. Endocrinol.*, 22(4):781–798.

[144] Visvader, J. E., Mao, X., Fujiwara, Y., Hahm, K., and Orkin, S. H. (1997). The lim-domain binding protein ldb1 and its partner lmo2 act as negative regulators of erythroid differentiation. *Proc. Natl. Acad. Sci. U.S.A.*, 94(25):13707–13712.

[145] Wang, J., Myklebost, O., and Hovig, E. (2003). MGraph: graphical models for microarray data analysis. *Bioinformatics*, 19(17):2210–2211.

[146] Wang, J. and Tian, T. (2010). Quantitative model for inferring dynamic regulation of the tumour suppressor gene P53. *BMC Bioinform.*, 11(1):36.

[147] Wang, J., Wu, Q., Hu, X. T., and Tian, T. (2016). An integrated approach to infer dynamic protein-gene interactions, a case study of the human P53 protein. *Methods*, 110:3–13.

[148] Wang, Y. (2013). *Gene Regulatory Networks*, pages 801–805. Springer New York, New York, NY.

[149] Wang, Y. X. R., Li, L., Li, J. J., and Huang, H. (2021). Network modeling in biology: statistical methods for gene and brain networks. *Statist. Sci.*, 36(1):89–108.

[150] Watson, J. D. and Crick, F. H. C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738.

[151] Wheat, J. C., Sella, Y., Willcockson, M., Skoultchi, A. I., Bergman, A., Singer, R. H., and Steidl, U. (2020). Single-molecule imaging of transcription dynamics in somatic stem cells. *Nature*, 583(7816):431–436.

[152] Woods, M. L., Leon, M., Perez-Carrasco, R., and Barnes, C. P. (2016). A statistical approach reveals designs for the most robust stochastic gene oscillators. *ACS Synth. Biol.*, 5(6):459–470.

[153] Wu, S., Cui, T., and Tian, T. (2018). Mathematical modelling of genetic network for regulating the fate determination of hematopoietic stem cells. In *Proceedings of 2018 IEEE International Conference on Bioinformatics and Biomedicine*, pages 2167–2173.

[154] Wu, S., Cui, T., Zhang, X., and Tian, T. (2020). A non-linear reverse-engineering methodfor inferring genetic regulatory networks. *PeerJ*, 8:e9065.

[155] Wu, S., Zhou, T., and Tian, T. (2022). non-linear reverse-engineering methodfor inferring genetic regulatory networks. *NPJ Syst. Biol. Appl.*

[156] Xiong, W. and Ferrell, J. E. (2003). A positive-feedback-based bistable 'memory module' that governs a cell fate decision. *Nature*, 426(6965):460–465.

[157] Xu, Z., Huang, S., Chang, L.-S., Agulnick, A., and Brandt, S. (2003). Identification of a Tal1 target gene reveals a positive role for the LIM domain-binding protein Ldb1 in erythroid gene expression and differentiation. *Mol. Cell. Biol.*, 23(21):7585–7599.

[158] Yang, B. and Bao, W. (2019). RNDEtree: Regulatory Network With Differential Equation Based on Flexible Neural Tree With Novel Criterion Function. *IEEE Access*, 7:58255–58263.

[159] Yang, B., Bao, W., Huang, D., and Yuehui, C. (2018). Inference of large-scale time-delayed gene regulatory network with parallel MapReduce Cloud platform. *Sci. Rep.*, 8(1):17787.

[160] Yang, L., Sun, W., and Turcotte, M. (2021). Coexistence of Hopf-born rotation and heteroclinic cycling in a time-delayed three-gene auto-regulated and mutually-repressed core genetic regulation network. *J. Theor. Biol.*, 527:110813.

[161] Zhang, P., Zhang, X., Iwama, A., Yu, C., Smith, K. A., Mueller, B. U., Narravula, S., Torbett, B. E., Orkin, S. H., and Tenen, D. G. (2000). PU.1 inhibits GATA-1 function and erythroid differentiation by blocking GATA-1 DNA binding. *Blood*, 96(8):2641–2648.

[162] Zhang, X., Zhao, J., Hao, J.-K., Zhao, X.-M., and Chen, L. (2015). Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. *Nucleic Acids Res.*, 43(5):e31.

[163] Zhang, X., Zhao, X., He, K., Lu, L., Cao, Y., Liu, J., Hao, J., Liu, Z., and Chen, L. (2012). Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics*, 28(1):98–104.

[164] Zhang, Y., Payne, K. J., Zhu, Y., Price, M. A., Parrish, Y. K., Zielinska, E., Barsky, L. W., and Crooks, G. M. (2005). SCL expression at critical points in human hematopoietic lineage commitment. *Stem Cells*, 23(6):852–860.

[165] Zhao, J., Zhou, Y., Zhang, X., and Chen, L. (2016). Part mutual information for quantifying direct associations in networks. *Proc. Natl. Acad. Sci. U.S.A.*, 113(18):5130–5135.

[166] Zhao, W., Kitidis, C., Fleming, M. D., Lodish, H. F., and Ghaffari, S. (2006). Erythropoietin stimulates phosphorylation and activation of GATA-1 via the PI3-kinase/AKT signaling pathway. *Blood*, 107(3):907–915.