Mathematical modeling and simulation of stem cell differentiation and reprogramming by Waddington's epigenetic landscape

Guo, Jing

2018

Guo, J. (2018). Mathematical modeling and simulation of stem cell differentiation and reprogramming by Waddington's epigenetic landscape. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/88079>

<https://doi.org/10.32657/10220/45611>

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

**MATHEMATICAL MODELING AND SIMULATION
OF STEM CELL DIFFERENTIATION AND
REPROGRAMMING BY WADDINGTON'S
EPIGENETIC LANDSCAPE**

GUO JING

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

2018

**MATHEMATICAL MODELING AND SIMULATION
OF STEM CELL DIFFERENTIATION AND
REPROGRAMMING BY WADDINGTON'S
EPIGENETIC LANDSCAPE**

GUO JING

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

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

2018

Declaration

I hereby declare I have written this doctoral thesis independently and quoted all the sources of information used in accordance with methodological instructions on ethical principles for writing an academic thesis. Moreover, I state that this thesis has neither been submitted nor accepted for any other degree.

.....

Jing Guo, 2018

Abstract

Recently, the field of computational stem cell biology, a new subdiscipline of developmental biology, has received a remarkable level of attention from the researchers. It leverages computational algorithms and tools to study the mechanisms of regulating the behaviors of embryonic stem cells (ESCs). In particular, the invention of induced pluripotent stem cells (iPSCs) in 2006 raised a question of intense focus: “What are the underlying molecular mechanisms by which iPSCs regain pluripotent potential?” Dynamical systems modeling has revealed the complexity of the regulation of cell fate determination and thus has proven to be a useful method for answering this question.

Towards exploring the molecular basis of stem cell fate regulation, the transcriptional regulatory network of core pluripotency transcription factors (TFs) was first modeled to uncover critical network structures that contribute to the cell fate transitions. Subsequently, we integrated the epigenetic network and signaling pathways that support the maintenance of stem cells and the determination of lineage specifications with the TF-based genetic layer of regulation, thereby constructing a hybrid network model of multilayered regulations. However, most of the computational models are mainly limited to exploring the system dynamics by theoretical modeling. More recently, a trend has emerged in computational stem cell biology to incorporate genome-wide sequencing data, especially single-cell profiling, with system-level modeling of stem cells.

In this thesis, we concentrate on these two computational methodologies, i.e., the knowledge-based modeling and the data-driven modeling, to facilitate the study of stem cell fate transitions using a well-known metaphor in development biology, namely Waddington’s epigenetic landscape. It describes the outcome of the percolation of genetic lineages by linking the cell phenotypes with the genotypes, through the coordination of gene regulatory networks. Thus, the trajectories of cell state transition on the landscape surface reflect the progressive dynamics of stem cell fate determination.

We developed a software tool named “NetLand” to facilitate the knowledge-based modeling and simulation study of cell state transitions driven by kinetics of gene regulatory network in Waddington’s epigenetic landscape. It serves as a useful tool for understanding the dynamics of stem cells. For example, it can be used to study the barriers to the conversion of iPSCs. Although there have been many a plenty of protocols for the generation of iPSCs, the low conversion rate is still a bottleneck. To apply iPSCs in

disease modeling and regenerative medicine, it is crucial to identify the key barriers that prevent the transition of cell types from a somatic state to an ESC-like state. NetLand identifies epigenetic blocks which have been shown to have an impact on the conversion rate of iPSCs, in the reprogramming process after analyzing a reconstructed hybrid network.

The leverage of the next generation sequencing techniques using data-driven approaches dramatically benefits the study of molecular regulations in stem cells. Here, we proposed a data-driven method, named HopLand, to model the process of stem cell fate determination by using the continuous Hopfield Network (CHN) to map cells in a Waddington's epigenetic landscape. The combinatorial regulatory interactions among genes constitute a kinetic model which can be calibrated by learning from single-cell gene expression data. HopLand is applied to estimate the pseudotimes of individual cells from the single-cell data, demonstrating its potential to generate fundamental insights into cell fate regulation.

The mathematical modeling methods proposed in this thesis, i.e., knowledge-based and data-driven approaches, yield insights into stem cell fate transitions. Subsequently, they have enhanced our capability to engineer cells, for example, increasing the conversion rate from somatic cell types to iPSCs. Many open challenges are still impeding cell fate engineering from broad clinical applications. We believe that the research described in this thesis will provide instrumental quantitative techniques to facilitate the understanding of cell dynamics and contribute to human health care.

Keywords: Waddington's epigenetic landscape, iPSC, stem cell reprogramming, differentiation, single-cell sequencing, neural network, pseudotime recovery.

Acknowledgements

I would like to express my most sincere gratitude to my supervisors Prof. Zheng Jie and Dr. Vivek Tanavde for their patient guidance and immense support during my research. This work would not be possible without their invaluable advice.

I would also like to show my gratitude to Prof. Lin Feng and Prof. Kwoh Chee Keong for their helpful advice and comments. My thanks also go to my colleagues and friends for valuable discussion and critical comments.

Especially, I would like to express my profound thanks to Dr. Wang Ying, Ms. Shamima Banu Binte Sm Rashid, Ms. Carolin Wall, and Dr. Swagato Barman Roy, for their comforting, understanding and encouragement, even during the difficult times of this research.

At last, I also greatly thank my parents and brother for their unconditional support.

Author's Publications

1. Jing Guo, Jie Zheng. HopLand: Single-cell pseudotime recovery using continuous Hopfield network based modeling of Waddington's epigenetic landscape. *Bioinformatics* 33, no. 14 (2017): i102-i109.
2. Jing Guo, Feng Lin, Xiaomeng Zhang, Vivek Tanavde, Jie Zheng. NetLand: quantitative modeling and visualization of Waddington's epigenetic landscape using probabilistic potential. *Bioinformatics* 33, no. 10 (2017): 1583-1585.
3. Jing Guo, Hui Liu, Jie Zheng. SynLethDB: synthetic lethality database toward discovery of selective and sensitive anticancer drug targets. *Nucleic acids research* 44, no. D1 (2015): D1011-D1017.
4. Jing Guo, Hao Chen, Peng Yang, Yew Ti Lee, Min Wu, Teresa M. Przytycka, Chee Keong Kwoh, Jie Zheng. LDSplitDB: A database for studies of meiotic recombination hotspots in MHC using human genomic data. *BMC Medical Genomics* (Accepted).
5. Jing Guo, Ritika Jain, Peng Yang, Rui Fan, Chee-Keong Kwoh, Jie Zheng. Reliable and fast estimation of recombination rates by convergence diagnosis and parallel Markov Chain Monte Carlo. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 11, no. 1 (2014): 63-72.
6. Jing Guo, Jie Zheng. Mathematical modeling for stem cell reprogramming by lineage specifiers and pluripotency factors. Asia-Oceania Top University League on Engineering (AOTULE) 2014 student conference.
7. Hao Chen, Peng Yang, Jing Guo, Chee Keong Kwoh, Teresa M. Przytycka, Jie Zheng. ARG-walker: inference of individual specific strengths of meiotic recombination hotspots by population genomics analysis. *BMC Genomics* 16, no. 12 (2015): S1.
8. Haifen Chen, Jing Guo, Shital K Mishra, Paul Robson, Mahesan Niranjan, Jie Zheng. Single-cell transcriptional analysis to uncover regulatory circuits driving cell fate decisions in early mouse development. *Bioinformatics* 31, no. 7 (2014): 1060-1066.

9. Peng Yang, Min Wu, Jing Guo, Chee Keong Kwoh, Teresa M. Przytycka, Jie Zheng. LDsplit: screening for cis-regulatory motifs stimulating meiotic recombination hotspots by analysis of DNA sequence polymorphisms. *BMC Bioinformatics* 15, no. 1 (2014): 48.

List of Tables

2.1	The list of essential signalling pathways governing ESC differentiation and reprogramming.	13
2.2	Examples of mathematical models for stem cell reprogramming or differentiation.	19
3.1	Steady states identified from synchronous updating simulation.	36
4.1	The default values of parameters in Eq. 4.1 and 4.2.	51
5.1	Numbers of embryos and cells in “GUO2010”.	79
5.2	Numbers of embryos and cells in “YAN2013”.	81
5.3	Numbers of embryos and cells in “DENG2014”.	82
5.4	Numbers of cells in “LPS”.	82
5.5	Numbers of cells in “HSMM”.	82
5.6	Numbers of cells in “ES_MEF”.	83
5.7	Top 10 key interactions identified from the weight matrix ranked by the absolute value of the weight in CHN.	89
S1	The detailed references for verifying interactions in HopLand’s result using GeneGo (Part 1)	107
S2	The detailed references for verifying interactions in HopLand’s result using GeneGo (Part 2)	108
S3	The detailed references for verifying interactions in HopLand’s result using GeneGo (Part 3)	109
S4	Logical rules in the Boolean model.	110
S5	ODE equations of the dynamical model of Figure 3.1.	111
S6	The inferred parameters of the network model.	113

List of Figures

1.1	The illustration of Waddington's epigenetic landscape. The marbles represent cells. Changes of cell fates are represented by transitions in the landscape under the control of different regulatory factors, e.g. epigenetic modifications, transcriptional regulations and signaling pathways.	3
2.1	The generation and differentiation of embryonic stem cells (ESCs).	9
2.2	The generation and derivation of iPSCs. Figure taken from http://learn.genetics.utah.edu/content/stemcells/quickref/ . .	11
2.3	An example network with two nodes. Each node is auto-regulated and mutually inhibited by the other node.	22
3.1	The 10-node transcriptional network with epigenetic modifiers in mouse embryonic stem cells. Gata6 and Sox1 are lineage specifiers. Klf4, Myc, Oct4, Nanog and Sox2 are pluripotency factors. Edges with flat heads indicate inhibition, and edges with arrowheads represent activation.	35
3.2	The probabilistic landscape of the constructed gene network.	40
3.3	The projection of a multi-dimensional data space to a 2-dimensional latent space using GPDM. The colour represents the level of quasi-potential U	42
3.4	Visualization of Waddington's epigenetic landscape. (a) Model without epigenetic modifiers. (b) Model with epigenetic modifiers. The dark areas indicate cell states with low potentials, and the light regions have high potential values.	44
4.1	A case study of a multi-attractor system by NetLand. Screenshot of NetLand on a 4-node network for mouse stem cell reprogramming. The 4-gene regulatory network (a) for mouse stem cell reprogramming [18] has three attractors shown in (d). Based on the mathematical model, time-series data of reprogramming experiment are simulated (b), and the network dynamics is visualized in Waddington's epigenetic landscape (c).	47

4.2	An overview of the functional modules in NetLand. There are mainly five functional modules, namely the input module, the mathematical modeling module, the modification module, the simulation module and the landscape module. It takes networks as input, and generates dynamic models for timeseries simulation and landscape construction.	48
4.3	The two visualization methods of the landscape in NetLand (a) by selecting two marker genes and (b) by mapping to a 2D latent space.	57
4.4	The three components in the output of NetLand. (a) The constructed landscape mapping to two marker genes (left) and mapping to a 2D latent space (right). (b) Control buttons. (c) The table of the grid data in the landscape colored according to the value of quasi-potentials.	58
4.5	(a) A two-gene network with mutual inhibitions and self-loops. The probabilistic landscapes of the two-gene network when (b) mapping to two marker genes and (c) a 2D latent space. (d) The three attractors.	60
4.6	The topology structure of the seesaw model with 4 genes and 16 interactions.	61
4.7	The probabilistic landscapes of the seesaw model when (a) mapping to two nodes, MEs and ECTs and (b) the 2D latent space using GPDM. (c) The table of attractors.	62
4.8	(a) The probabilistic landscapes of the modified network when mapping to MEs and ECTs. (b) The table of attractors. . .	63
4.9	Time series concentrations in terms of the concentration levels of $Signal_A$ indicating the final steady state values. (a) $Signal_A = 1$, the trophectoderm is the only possible state of the system. (b-c) $Signal_A = 10$, the bistable system gives either the embryonic or differentiated stem cell (endoderm) lineage. (d) $Signal_A = 25$, the endoderm lineage is the only possible state.	65
4.10	The topology structure of the cancer model.	66
4.11	The probabilistic landscapes of the network when (a) mapping to two marker genes (P53 and Zeb) and (b) a 2D latent space. (c) Identified attractors.	67
5.1	(a) The topology and (b) The dynamic feature of Hopfield network.	70
5.2	Accuracies of pseudotime recovery on 5 synthetic datasets using different pseudotime recovery methods. Pearson correlation coefficient between the predicted and observed times is used to evaluate the result.	84

5.3	The contour plots of the constructed Waddington's epigenetic landscape using the five synthetic datasets (a-e). The contour lines represent heights in the landscape. The dark areas indicate low energy, and the light regions have high energy. The cells from early stages to late stages are colored from dark red to white.	85
5.4	(a) The Waddington's epigenetic landscape recovered using HopLand. (b) The contour plot of the constructed Waddington's epigenetic landscape. The dots are colored according to the developmental stages of the represented cells.	86
5.5	Mapping gene expression values to Waddington's epigenetic landscape using (a) Fgf4, (b) Gata4, (c) Cdx2 and (d) Sox2. The value decreases from dark red to white.	87
5.6	The minimum spanning tree constructed from Waddington's epigenetic landscape. The dots are colored according to the developmental stages of the cells in the dataset of "GUO2010".	88
5.7	Accuracies of pseudotime recovery on the qPCR dataset "GUO2010" using different pseudotime recovery methods. Pearson correlation coefficient between the predicted and observed times is used to evaluate the result.	88
5.8	The core network found by searching GeneGo MetaCore for interactions overlapping with HopLand's top 20 interactions and genes. The red and green lines represent inhibition and activation respectively. The volcano icon indicates transcription factors. The orange bullet represents enzyme. The blue curve represents a binding protein. The green anchor represents receptor ligand. The purple X represents channel. . .	90
5.9	The contour plot of the landscape constructed from the dataset of "DENG2014". The cells are connected in a minimum spanning tree.	91
5.10	(a) The constructed landscape of the dataset "YAN2013". (b) The contour plot of the landscape with labeled lineages. (c) Unsupervised hierarchical clustering of 30 samples from the late blastocyst (including 3 lineages: EPI, PE, TE) and heatmaps of 8 selected differentially expressed genes [146]. .	92
S1	The modified reactions after removing self-loops.	101
S2	The probabilistic landscapes of the modified two-gene network when (a) mapping to two marker genes and (b) a 2D latent space. (c) The table of attractors.	103
S3	The table of attractor(s).	103
S4	(a) The topology structure of the model with 6 genes and 23 interactions. The probabilistic landscapes of the network when (b) mapping to two marker genes (Nanog and Gata6) and (c) a 2D latent space. (d) The two attractors.	104

S5	The weight matrix contains $N \times N$ interactions of CHN learned from the mouse embryonic early development dataset. N is the number of genes.	105
S6	Accuracies of pseudotime recovery on the qPCR dataset “DENG2014” using different pseudotime recovery methods. Pearson correlation coefficient between the predicted and observed times is used to evaluate the result.	105
S7	The contour plot of the landscape constructed from the dataset (a) “ES_MEF”, (b) “LPS” and (c) “HSSM”.	106

Contents

Abstract	iv
Acknowledgements	vi
Author's Publications	vii
List of Tables	ix
List of Figures	x
1 Introduction	1
1.1 Background	1
1.2 Motivation and objectives	4
1.3 Organization	6
2 Literature review	8
2.1 Biological Background	9
2.1.1 Stem cell differentiation	9
2.1.2 Stem cell reprogramming	10
2.1.3 Regulation of cell fate allocation	11
2.2 Waddington's epigenetic landscape	15
2.2.1 Theory	15
2.2.2 Implementation	16
2.3 Dynamical systems modeling	18
2.3.1 Boolean model	19
2.3.2 Continuous model	20
2.3.3 Current progress	23
2.4 Data-driven approach	25
2.4.1 Single-cell technology	25
2.4.2 Hopfield network	26
2.4.3 Current progress	29
2.5 Summary	30
3 Knowledge-based mathematical modelling	32
3.1 Construction of the dynamic model	32
3.1.1 Gene regulatory network	32
3.1.2 Dynamic formulation	34

3.1.3	Parameter estimation	37
3.2	Construction of the landscape	38
3.2.1	The probabilistic landscape	38
3.2.2	Visualization of multi-dimensional landscape	39
3.3	Epigenetic barriers in cell fate alterations	43
3.4	Summary	44
4	NetLand: quantitative modeling and visualization of Waddington's epigenetic landscape using probabilistic potential	46
4.1	Overview of NetLand	47
4.2	Dynamic model construction and simulation	49
4.2.1	Constructing a dynamic model	49
4.2.2	Running a simulation	52
4.3	Waddington's epigenetic landscape construction	55
4.3.1	Constructing a probabilistic landscape	55
4.3.2	Visualizing a probabilistic landscape	56
4.3.3	Result rendering	56
4.4	Case studies	57
4.4.1	A classic two-gene network regulating cell fate	58
4.4.2	Studies on stem cell state transitions	60
4.4.3	Cancer development	64
4.5	Summary	67
5	Data-driven approach in modelling stem cell fate determination	69
5.1	Workflow of HopLand	70
5.2	Data-driven kinetic modelling	71
5.2.1	Network formulation	72
5.2.2	Parameter estimation	73
5.3	Construction of the landscape	74
5.3.1	Energy function	76
5.3.2	Visualizing the landscape	76
5.4	Pseudotime recovery	77
5.4.1	Single-cell sequencing datasets	78
5.4.2	Pseudotimes inferred from single-cell sequencing data	83
5.5	Summary	92
6	Conclusion	95
6.1	Summary of the thesis	96
6.2	Future work	97
	References	127

Chapter 1

Introduction

By their nature, embryonic stem cells (ESCs) can irreversibly develop into all types of somatic cells. However, in 2006 the reprogramming of somatic cells by the replenishment of four transcription factors (TFs) generated ESC-like induced pluripotent stem cells (iPSCs) [1]. The remarkable success in cell engineering raised a question of intense focus: “What are the underlying molecular mechanisms by which iPSCs regain pluripotent potential?”

Towards exploring the molecular basis, *in silico* system modeling of Waddington’s epigenetic landscape leverages computational algorithms and tools to study the mechanisms of stem cell fate conversions. These methods have proven to be useful in answering the above question by revealing the complexity of the regulation of cell fate determination.

This chapter first introduces briefly the background of stem cell fate determination and the state-of-the-art modeling methods. Then the concept and application of Waddington’s epigenetic landscape in the field of computational stem cell biology are illustrated along with the motivations and objectives of the research. Finally, it outlines the structure of the thesis.

1.1 Background

Cell fate determination is the result of the coordination of multilayer regulations. Stem cell differentiation and reprogramming are the two represen-

tative examples in which multiple regulators work at the cellular and tissue levels to create the final cell types.

During the differentiation process, an ESC becomes specialized with a specific function, such as a muscle cell, a liver cell, or a neuron. These cells constitute the multicellular organism with various tissues and cell types after many times of differentiation. The characteristics of cells, such as ESCs or adult stem cells, that can differentiate into all (or a collection of) cell types is known as pluripotency. Thus differentiation is also a process that a cell loses its pluripotency along with dramatic changes in cell morphology and functionality. Modifications in transcriptional patterns lead to these changes involving switching on lineage-specific genes and off genes not needed in a particular tissue through epigenetic modification and cellular signaling.

This cell development process was generally considered to be irreversible, until the invention of iPSC reprogramming experiment in 2006 [1]. In this research, fully differentiated cells are successfully converted to iPSCs which possess similar functionality and morphology as ESCs, by the replenishment of the four TFs, i.e., Oct4, Sox2, Klf4, and cMyc. The regained pluripotency feature of iPSCs endows this technology with great potential in cell therapy and disease modeling [2]–[4]. Thus it is considered as a significant advance in the manipulation of cell fates. However, iPSCs still retain a unique gene expression signature and epigenetic profiles although they outwardly appear to be indistinguishable from ESCs [5]. The underlying molecular mechanisms by which iPSCs regain pluripotent potential are still unclear.

To understand how pluripotent cells develop into a final cell type, and how the cell fate is determined, *in silico* systems modeling of Waddington's epigenetic landscape has proven to be useful. The quantified epigenetic landscape provides a performing stage on which the cell fate determination and cell type maturation are choreographed according to the inherent multilayered regulations.

Based on the essential idea of Waddington's epigenetic landscape, changes of cell fates are represented by state transitions overcoming energy barriers

between different attractors (Figure 1.1). A cellular system tends to converge to low-energy stable states represented by attractors in the state space of the biological system [6]. These attractors emerge from the dynamics of the biological system. For example, basins in the Waddington’s epigenetic landscape of the cell development process represent cell types, such as ESCs or somatic cells.

Figure 1.1: The illustration of Waddington’s epigenetic landscape. The marbles represent cells. Changes of cell fates are represented by transitions in the landscape under the control of different regulatory factors, e.g. epigenetic modifications, transcriptional regulations and signaling pathways.

However, the original concept of Waddington’s epigenetic landscape is a theoretical metaphor that lacks a concrete mechanistic foundation. It still requires the quantification using mathematical methods to simulate complex biological systems. Knowledge-based dynamical modeling of kinetic mechanisms and data-driven statistical modeling are the two commonly used computational methods for modeling Waddington’s epigenetic landscape.

The knowledge-based computational models integrate the transcriptional, signaling and epigenetic factors at a system level according to the prior knowledge of regulatory mechanisms. Differential equations (DEs) in a continuous system with continuous expression levels of genes and Boolean network (BN) in a discrete system with binary switches of genes are the two main types of formulations [7]–[10]. These equations ultimately determine the interactions among the genes in a network, and they have been shown

to capture some characteristics and dynamical trends of gene expression.

However, intricate knowledge about regulatory systems and molecular kinetics that is required for BN- and DE-based models may not be available in many cases. In addition, the computational complexity of *in silico* simulation of such a system often increases exponentially with the increase of network size. Data-driven models can help address those challenges. With the ongoing development of high-throughput genomic sequencing technologies, not only bulk data but also single-cell expression profiles are available these days, which provide valuable resources and possibilities for the inference analysis of large gene regulatory networks. But it needs efficient methods for the data analysis and modeling.

A data-driven method for modeling Waddington’s epigenetic landscape using Hopfield network (a type of artificial neural network) was proposed recently [11]–[13]. The energy of the Hopfield network is defined as the potential of the landscape. It offers an attractive framework for quantitative modeling of cell differentiation and reprogramming from experimental data. Compared with BN and DE models, it significantly reduced the computational time for networks of larger sizes. Moreover, neither network structure nor knowledge of kinetic parameters is required.

1.2 Motivation and objectives

Embryonic stem cells and iPSCs share the common properties of self-renewal and the pluripotency which means they can be differentiated into various types of cell lineages. In addition, iPSCs overcome the ethical issues of ESCs in clinics to make it possible to be applied in disease modeling, regenerative medicine, and drug discovery [14]. The establishment of more authentic methods for iPSC generation and differentiation requires a thorough understanding of the underlying mechanism.

Current studies have revealed unprecedented complexity in stem cell differentiation and reprogramming involving various molecules in different

molecular mechanisms, e.g., TFs in transcriptional regulation, miRNAs in RAN interference. However, our current understanding is only a piece of the entire story.

Motivated by the demand for the study of stem cell fate determination, we concentrate on the two computational methodologies in this thesis: the knowledge-based modeling and the data-driven analysis, to facilitate the study of stem cell fate transitions using a popular visual metaphor for lineage specification, called Waddington’s epigenetic landscape:

1. It is estimated that epigenetic barriers in the reprogramming process are the main reason for the low conversion rate in reprogramming experiments [15], [16]. To verify the result, we proposed a knowledge-based mathematical model based on a hybrid network which integrates essential regulatory factors in ESCs, i.e., external signals, epigenetic modifiers and core transcriptional circuits. The *in silico* mathematical model successfully predicted a low cell-conversion rate in ordinary reprogramming experiments and identified the barriers in the simulation of reprogramming.
2. There is a lack of user-friendly software tools to calculate and visualize the Waddington’s epigenetic landscape, especially to visualize the 3D landscape for a regulatory network with more than two genes. To address the limitations of the practical applications of current modeling methods, we have developed a software tool called “NetLand” which is intended for modeling, simulation and visualization of gene regulatory networks (GRNs) and their corresponding quasi-potential landscapes. It promotes the broad applications of the quantitative models of Waddington’s epigenetic landscape by researchers of diverse backgrounds.
3. Taking advantage of the advanced single-cell technologies, a data-driven model is proposed to infer Waddington’s epigenetic landscape from single-cell gene expression data using Hopfield network. It re-

veals the transcriptome dynamics from real data, such as the recovery of the pseudotimes which is a latent dimension that measures the cells' progress through the transition. Therefore, it can help us understand the transition of cell states represented by gene expression profiles.

The mathematical modeling methods proposed in this thesis, i.e., knowledge-based and data-driven approaches, yield deep insights into stem cell fate determination. Subsequently, they can enhance our capability to engineer cells, for example, the reprogramming of somatic cell types to iPSCs. Many open challenges impede cell fate engineering from broad clinical applications. We believe that our methods and tools will provide a quantitative basis to facilitate the understanding of cell dynamics and contribute to human health care.

1.3 Organization

The thesis is organized into chapters to cover different aspects of the research on stem cell fate determination using mathematical modeling of Waddington's epigenetic landscape.

Chapter 2 introduces the biological background of stem cell fate regulation. The original theory and existing methods for the reconstruction of Waddington's epigenetic landscape are reviewed and summarized. A survey of the state-of-the-art modeling methods covers two types of techniques, i.e. knowledge-based modeling and data-driven modeling. It also introduces the fundamentals of Hopfield network and the protocols of different single-cell genomic sequencing technologies used for data-driven analysis.

Chapter 3 provides a detailed description of the formulation, implementation, and simulation of a hybrid network model constructed for the study of stem cell differentiation and reprogramming, as well as the construction and visualization of the corresponding landscape. In particular, the barrier role of epigenetic modifiers in reprogramming experiments is recapitulated in this model.

Chapter 4 introduces the software tool of NetLand for quantitative modeling and visualization of Waddington’s epigenetic landscape using probabilistic potential. We provide a list of case studies to illustrate the successful application of NetLand in the area of stem cell differentiation and reprogramming and cancer development.

Chapter 5 introduces a data-driven method of modeling stem cell fate conversion using the Hopfield network. The kinetic modeling including parameter estimation is introduced in this chapter, as well as the definition of energy function and the visualization of the landscape. It was applied to the estimation of pseudotimes from different types of single-cell sequencing data and compared with other existing methods.

Finally, Chapter 6 concludes the thesis by summarizing the methods and results introduced in the previous chapters, and maps out possible areas for future research.

Chapter 2

Literature review

Stem cell differentiation is the process where an ESC acquires cellular traits becoming functional specialized cells. Somatic cells differentiated from ESCs were considered as fully programmed that cannot be reprogrammed until the invention of induced pluripotent stem cells by reprogramming experiments. Even though reprogramming experiments require more critical conditions to reverse the process of cellular differentiation [1], [17], [18], they are considered promising in the application of regenerative therapy and disease modeling [19]–[22]. However, the achievement of Yamanaka’s reprogramming experiment and the following studies of stem cell reprogramming would just be the tip of the iceberg in a long-standing problem: how could human manipulate the fate of a cell eventually?

In this chapter, we will introduce the biological background of stem cell fate determination, including stem cell differentiation and reprogramming. The concept and the current development of Waddington’s epigenetic landscape are introduced as well. We also review and summarize the state-of-the-art computational methods in the studies of cell fate conversions.

2.1 Biological Background

2.1.1 Stem cell differentiation

ESCs are derived from the inner cell mass (ICM) of the blastocyst. They can proliferate indefinitely and retain pluripotency before differentiating into different types of cells [23]. The process from the fertilized embryo to a somatic tissue or organ is called stem cell differentiation [24]. During this process, different cell lineages are formed with specific combinations of media and factors. Figure 2.1 shows the origin of ESCs from the blastocyst and the commitment to different lineages.

Figure 2.1: The generation and differentiation of embryonic stem cells (ESCs).

Thus, stem cell differentiation is a natural process that produces the tissues and organs of living organisms. It generates different types of cells with various functions and structures. On the contrary, stem cell reprogramming converts the differentiated cells into iPSCs, which are ESC-like cells.

2.1.2 Stem cell reprogramming

The attempts of reprogramming adult cells into ESC-like cells have lasted for decades. Back to 1960s, scientists have successfully used adult cells to generate the same organisms using nuclear transfer and cell fusion. The remarkable experiment of Yamanaka in 2006 [1] heralded a new era of stem cell reprogramming. The development of iPSC reprogramming technology challenges the traditional idea of irreversible cell development by putting forward the possibility of the conversion from a completely programmed cell back to a state similar to ESCs.

In Yamanaka's experiment, the generation of iPSCs provides a TF transduction method to convert adult cells to pluripotent cells compared to nuclear transfer and cell fusion [22], [25], [26] (Figure 2.2). iPSCs show remarkable similarities to ESCs in both cell morphology and functionality, such as proliferative properties [27]. A variety of cells including fat, skin, and fibroblasts could be the sources to generate iPSCs when treated with different cocktails, e.g., the combinations of TFs [18], [22], [26], [28], small chemical compounds, growth factors stimulus and epigenetic modifiers [29]. Although many cocktails have been identified to facilitate the reprogramming, the underlying mechanisms for the generation of iPSCs are still unclear.

The reprogrammed cells offer the possibility of an unlimited source of material for the replacement of damaged tissues and disease study since they can be differentiated into a variety of cell types due to their infinite differentiation potential. Moreover, the production of iPSCs from a patient's cells provides a patient-specific source of pluripotent stem cells to overcome immune attacks, the most formidable barrier to transplantation. And the source somatic cells are relatively easy to obtain. The first attempt to use iPSCs in clinical treatment was performed in 2014 when a woman with eye disease was transplanted with retinal cells derived from iPSCs. However, genetic alterations arise from the reprogramming experiment leading to the uncertainty about the safety of using iPSCs in patients [22]. Therefore, the evaluation and prediction of iPSCs' lineage bias become critical with the

Figure 2.2: The generation and derivation of iPSCs. Figure taken from <http://learn.genetics.utah.edu/content/stemcells/quickref/>.

rapid development of stem cell engineering technologies.

Despite uncertainties and issues, the reprogramming technology is still progressing rapidly. Scientists try to search for alternative and more efficient methods to generate iPSCs and are getting to grips with the molecular basis of pluripotency and lineage plasticity, although the underlying mechanisms remain controversial.

2.1.3 Regulation of cell fate allocation

Although the underlying mechanisms of stem cell fate conversion are still unclear, some attempts to uncover them have been made. It is hypothesized that GRN plays a dominant role in the cell fate determination [30]. Besides signaling pathways work as triggers controlling the expression status of a set of genes [31]–[33]. Epigenetic modifiers embellish chromatin states leading to the change of transcriptional rates of a proportion of target genes [29], [34], [35]. The intricate interactions among various mechanisms form a multilayered system for cell fates determination.

Transcriptional regulation

Experiments have suggested that the system-level stable cell state arises from genetic interactions [36]. Cellular decisions controlling development, self-renewal, and differentiation are thought to be regulated by the expression of specific TFs. So does reprogramming. In Yamanaka's reprogramming experiment, four reprogramming factors, i.e. Oct4, Sox2, Klf4 and cMyc, were replenished to convert mouse fibroblast cells into iPSCs, which demonstrated that the over expression of specific TFs is sufficient to reprogram somatic cells back to pluripotent cells. Therefore, the transcriptional regulations within the GRN contribute to the driving force behind the change of cell identity.

In the early stage of development, TFs promote the construction of all three vertebrate germ layers, i.e., ectoderm, mesoderm, and endoderm (Figure 2.1). Different lineages are characterized by their marker genes with specific regulators. The activation of one group of TFs will inhibit their mutual regulators.

The reprogramming process is a multi-step process characterized by transcriptome and proteome resetting [37]. Gene expression patterns are changed dramatically from differentiated states to stem cell states. At the initial stage, differentiation-associated and lineage marker genes are down-regulated combined with the continuous activation of genes responsible for proliferation, metabolism and cytoskeleton organization [38]. The core pluripotency network is assembled and maintained when genes responsible for the embryonic development and stem cell maintenance are activated [39], [40]. Then a stable pluripotent state is established.

Signaling pathways

The cellular programs of ESCs, including proliferation, potency, and cell fate determination, are mediated by signalling pathways through modulating the expression of TFs or chromatin structures in a concerted manner. The signaling pathways have emerged as important control devices of cell fate, in-

Table 2.1: The list of essential signalling pathways governing ESC differentiation and reprogramming.

Signalling Pathway	Function	References
cytokine leukaemia inhibitory factor (LIF)	maintain pluripotency	[32], [41], [46], [48]–[50]
bone morphogenetic protein 4 (BMP4)	Induce differentiation of human ES cells to trophoblast; promoting MET; cooperates with LIF in the maintenance of pluripotency in mouse ESCs;	[24], [33], [41], [46], [48]–[50]
Transforming growth factor- β (TGF- β)	Inhibit reprogramming; Embryonic stem cells can be maintained; Inhibition of TGF- β	[42], [46], [48], [49], [51], [52]
phosphatidylinositol-3-kinase (PI3K)	Improves Mouse Fibroblast Reprogramming enhance reprogramming	[41], [48]
Fibroblast growth factor (FGF)	promotes the early stages of reprogramming; connected with the exit from pluripotency and epigenetic priming for differentiation; Embryonic stem cells can be maintained	[45], [46], [48], [49], [53]
Wnt	enhances the maturation phase of mouse somatic cell reprogramming; enhances somatic-cell reprogramming	[41], [42], [44], [46], [48], [54], [55]
Notch	cell-to-cell contact via surface proteins; operates in lateral inhibition roles for a single cell within a group of equivalent cells, and as a control of lineage decisions and inductive	[42], [49], [54]
Hippo	signaling among non-equivalent cells cell-to-cell contact via surface proteins	[49]

cluding Leukemia inhibitory factor (LIF), Wingless-type (Wnt), Notch and Bone Morphogenetic Protein 4 (BMP4) [41]–[43]. Some of them are essential for the maintenance of embryonic stem cell pluripotency [41], [44], whilst some of them trigger the differentiation process [45], [46]. At different developmental stages, one signaling pathway may simultaneously promote antagonistic lineages [47]. Table 2.1 shows the predominant signaling pathways involved in ESC differentiation and iPSC reprogramming.

The extrinsic BMP and LIF signaling pathways collaboratively maintain mouse pluripotency [56]. The Wnt pathway enhances cell reprogramming. TGF- β is the predominant signaling pathway involved in ESC pluripotency and self-renewal. Hippo pathway transduces signals through the cell-cell contact via surface proteins. FGF is associated with stem cell maintenance and promotes the early stages of reprogramming. Besides their functions in ESC proliferation, maintenance and differentiation, the LIF/STAT3, BMP, PI3K, FGF2, Wnt, TGF- β and MAPK pathways have been found to also govern the reprogramming process [48].

The derivations of the three primary germ layers, i.e., endoderm, mesoderm, and ectoderm, are controlled by signaling pathways. BMP is one of the primary signaling pathways that promote differentiation. It inhibits the

expression of Nanog and activates the expression of lineage-specific genes.

The regulation of signaling functions is through critical TFs, e.g., Oct4, Sox2, and Nanog, to control the state of ESCs. These TFs which often serve as markers, activate the expression of specific genes and suppress other genes involved in different lineages.

The signaling network is a complex system which may contain multiple regulatory circuits with variant downstream targets. The effect of regulations could reach a steady state through the balance of the counteraction between parallel circuits in the signaling pathway [32]. However, the intricate molecular mechanisms of the signaling pathways that regulate cell fate decisions have remained elusive.

Epigenetic modification

Besides transcriptional regulation and signaling pathways, the change of the chromatin states, e.g., covalent chromatin modification, histone modifications, DNA methylation and demethylation, is not only a marker of cell states but also a regulatory factor on developmental genes in ESCs [57]. The activities of TFs rely on the states of the chromatin to permit the TFs to bind to specific gene promoters.

During the reprogramming process, chromatin and its regulators work as essential controllers, especially in the early epigenetic priming events which may be critical for pluripotency induction later. The change of chromatin states and transcriptional profiles occurs in a stepwise manner: the epigenetic signature of the somatic cell is erased first, and then it is reshaped to an ESC-like epigenome including DNA demethylation of promoter regions of pluripotency genes and genome-wide resetting of histone modifications [37]. By contrast, distinct patterns of chromatin reorganization are discovered in the differentiation process. The changes in gene expression patterns during the differentiation process rely on dynamic epigenetic regulation.

Stem cell reprogramming and differentiation are complex biological processes, the global dynamics of which cannot be explained by a single pathway

or a GRN. Computational biology makes use of mathematical methods to model and simulate complex biological systems. Thus it is widely accepted as a powerful tool in stem cell research [29].

2.2 Waddington's epigenetic landscape

2.2.1 Theory

Waddington's epigenetic landscape [31] was originally introduced as a theoretical model in 1950s to explain the generation of distinct cell types [36]. Back to that time, the fate of a cell was considered irreversible, i.e. once a cell is fully developed, its identity will keep the same.

In this model, the differentiation process is depicted as cells rolling down a hillside. On the summit of the hill is a progenitor cell located, and in response to triggers, it would roll into different valleys corresponding to various cell types. Ridges separate the valleys thereby preventing the mixture of different cell types. The final positions of cells would be the attractors, located at the end of each valley. The cells will be trapped there even with external disturbances.

One defect of Waddington's epigenetic landscape is the lack of details about biological mechanisms. A question we try to answer is what is the best principled way to quantify the Waddington's epigenetic landscape according to the real biological systems.

In Waddington's depiction, the differentiation process goes along with the loss of potential energy which corresponds to the altitude in the landscape. Compared to the entry point, the valleys possess relatively lower potential representing more stable states. According to statistical physics, open systems in nature tend to change to lower energy and lower relative entropy which reflects the amount of overall disorder, randomness or uncertainty [58]. The barriers between valleys as energy shields restrain the cross-talks and force the cells to move along their canals. To some extent, two valleys with lower barriers between them indicate that the two cell types

have more similar phenotypes.

Thus in the theory of landscape a cell can be abstracted as a physical system. A pluripotent stem cell would have a higher potential energy and relative entropy since it possesses the potential to be committed to diverse cell fates. By contrast, a fully differentiated cell would have relatively lower entropy since it only needs to maintain its specific state. Physical theories and mathematical formulations could be adapted to quantify the model to provide more information [36], [59].

Since a cell type can be simplified as expression levels of a set of genes, a single point on the Waddington's epigenetic landscape corresponds to a specific cell state represented by a set of gene expression levels, and the attractor state defines a stable cell type. Then, the change of cell fate can be visualized as the movement of a point on the landscape. In that case, physical forces that determine the position of a cell in this high-dimensional space are analogous to regulations functioning in a complex biological system that is in control of gene expression.

The Waddington's epigenetic landscape, as a theoretical model, can be extended not only be used for understanding the transition from a progenitor cell toward differentiated stable states [60] but also for depicting the reprogramming from somatic cells to pluripotent states. Based on the essential idea of Waddington's epigenetic landscape, changes of cell fates are represented by transitions overcoming energy barriers between different attractors. Thus the quantified epigenetic landscape is the stage on which the play of cell fate decisions and cell type maturation is choreographed according to some physical laws [36].

2.2.2 Implementation

By following ideas from theoretical physics, the intrinsic kinetics of genes is encoded by physical laws that determine the developmental potentials of cells. As such, the theoretical concept becomes a quantitative model that is capable of describing the global dynamics of a system.

Langevin dynamics which describes the Brownian motion has been applied to studying the stochasticity in gene transcriptional regulation [61]–[63]. In genetic circuits, the change rates of molecular species can be formulated as Langevin equation (Eq. 2.1).

$$dX(t) = -\Delta U dt + \sqrt{D} dW_t \quad (2.1)$$

where $X(t)$ denotes the number of molecules of each gene in the network at time t . D is the noise term associated with each reaction. U is the particle interaction potential. W_t denotes a Wiener process (Standard Brownian motion). The gradient force F is equal to $-\Delta U$. The drifting part is composed of the noise term.

The Langevin equation can be reformulated as a Fokker–Planck (FP) equation which governs the time evolution of the probability density function of each $X(t)$ under the influence of driving forces and random fluxes (Eq. 2.2) [64]. FP equation is a partial differential equation (PDE).

$$\frac{\partial P(X, t)}{\partial t} = -\sum_{i=1}^N \frac{\partial [F(x_i) P(x_i, t)]}{\partial x_i} + \sum_{i=1}^N \sum_{j=1}^N \frac{\partial^2 [D_{ij}(X) P(X, t)]}{\partial x_i \partial x_j}, \quad (2.2)$$

where $P(X, t)$ is the probability density function of a time-dependent variable $X(t)$. $X(t)$ is assumed to be a first-order Markov process. Thus $P(x, t)$ is the probability of certain expression state x at time t , which possesses the quasi-potential of

$$U = -\ln(P(x, t)). \quad (2.3)$$

The quasi-potential landscape is an implementation of Waddington’s epigenetic landscape. It is based on the assumption that the noise in gene expression follows Gaussian distribution and the individual probability of each gene is independent. Moreover, the quasi-potential landscape is based on the notion that the probability of gene expression state determines the

stability of cell [65], [66].

The numerical result of U can be solved by self-consistent mean field approximation method demonstrated in [65]. It assumes that the probabilities of gene concentrations are independent and follow a Gaussian distribution. Thus $P(x, t) = \prod_{i=1}^n P(x_i, t)$, where n is the number of genes, and $P(x_i, t) \sim N(\mu_i, \sigma_i)$ where μ_i and σ_i are the mean and variance of each gene.

Therefore, there are $2n$ differential equations to be solved including n equations for the means and n equations for the variances. In a multi-stable system, there are two or more attractors with different sizes of basins. The steady state probability is defined as $P = \sum_{i=1}^m w_i P(x_i)$, where $\sum_{i=1}^m w_i = 1$, and m is the number of attractors.

2.3 Dynamical systems modeling

Dynamical systems modelling leverages computational algorithms and tools to study the mechanisms of stem cell fate regulation. It involves the use of mathematical modeling methods and computer simulations of cellular subsystems, such as transcriptional regulation, signal transduction pathways and epigenome that contribute to the change of cell types, to both analyze and visualize the complex interactions among these cellular processes.

The dynamics of reactions can be modeled in different ways, e.g., Boolean logic and differential equations. Discrete models based on Boolean rules are relatively independent of detailed information of system thermodynamics. It applies discretization to the real values which might compromise the accuracy [67]. By contrast, continuous modeling approaches using real-valued parameters can generate more accurate data over a continuous time scale [68]. There are some theoretical models constructed to depict the biological phenomena of transcriptional regulations, signaling regulations and epigenetic modifications (Table 2.2).

Table 2.2: Examples of mathematical models for stem cell reprogramming or differentiation.

Description	Method	Publications	Related work
A model for a two-gene network	ODE (Mension equations) Landscape	[7]	Sui Huang's quasi-potential landscape
A combination of fuzzy theory and petri network	Fuzzy petri network	[8]	Boolean network
Two isolated models for Wnt and Notch respectively and a combined model	ODE (Mension equations)	[9]	General method for signaling modeling [69], [70]
A model for Notch and BMP4 with core GRN	ODE (non-contact model Narula, 2010) Consider enhancer, promoter	[71]	
A model for three-gene network	ODE (Hill equation, considering protein complex binding)	[7], [10]	
A model for epigenetic regulations during reprogramming	Epigenetic regulatory rules with assigned probabilities ODE (Mension equations)	[72]	Probabilistic Boolean network
A model for 22-gene (out of 52) network during reprogramming or differentiation	Landscape (Langevin dynamics)	[73]	Sui Huang's probabilistic landscape
A model for 10-gene network during reprogramming or differentiation	ODE (Wilson-Cowan function) Epigenetic state network	[74]	
A model for lineage determination with external signals	ODE (Mension equations)	[75]	
A model for lineage commitment	ODE and landscape	[76]	Study the entropy evolution

2.3.1 Boolean model

The Boolean network model, which was first proposed by Kauffman [77], discretizes the real value into discrete levels, such as two levels of 1 (active state) and 0 (inactive state). It is instrumental for understanding how the network structure and logical rules determine the flow of information [78]. For example, the study of network stability can be achieved by modeling regulations with logic functions using only the network's structure [79]. It has also been used to model the dynamics of cell signaling networks involving pathway crosstalk and produced insights about the operation of signaling pathways [80].

The primary advantage of the discrete model is its simplicity as the values comprise limited discrete levels. The regulatory interactions between entries are formulated by logic relations. Thus the computational complexity of small networks is reduced compared to the calculation of DEs. Despite the simplicity, Boolean network models can provide important insights into network dynamics, such as the existence of steady states or network robustness, without need of detailed information of kinetic parameters [78].

One of the main limitations of Boolean logic models is the loss of information due to the biochemically unrealistic discretization of values. The simplified formulation may lead to limited details of real dynamics in biological models. Another problem is that for large-scale networks, the computational

cost of steady states simulation still grows exponentially.

2.3.2 Continuous model

The continuous modeling approach represents the concentration change of a component over time [81]. Differential equations (ODE or SDE) are widely used to describe the kinetics of a gene regulatory system. The algebraic equations depict the instantaneous change of concentrations over time in a GRN and construct a thermodynamic model. This model can provide detailed information about global kinetics and individual interactions [82].

A profound advantage and also disadvantage is that the quality of the model relies on the accuracy of kinetic parameters [83]. Due to measurement difficulties and lack of knowledge on the functions of many proteins and their interactions, some of the parameters of continuous models are usually based on estimations or inference [68]. Therefore this method has been applied to only a few well-studied fine-scale systems.

Despite these drawbacks, continuous modeling is still a practical method to study the kinetics of a biological system. The Hill and Michaelis-Menten functions are examples of such analytical models of small systems [81]. Here we give some recent cases for modeling the dynamics of regulatory networks using ODEs.

Michaelis-Menten kinetics

Michaelis-Menten kinetics was initially proposed to describe the synthetic rate of a product. It is assumed that the concentration of enzyme is much lower than the substrate concentration when the enzyme is not allosteric in a quasi-steady state [81]. In a general case, E denotes an enzyme binding to its substrate S and P is the product (Eq. 2.4). E catalyzes the reaction in the form of a temporary complex with S , denoted as ES . The total

concentration is not produced or consumed.

The concentration change rates are listed in the following equations:

$$\frac{dP}{dt} = k_{cat}[ES] \quad (2.5)$$

$$\frac{d[ES]}{dt} = k_f ES - k_r[ES] - k_{cat}[ES] \quad (2.6)$$

$$\frac{dE}{dt} = k_r[ES] + k_{cat}[ES] - k_f ES \quad (2.7)$$

$$\frac{dS}{dt} = k_r[ES] - k_f ES \quad (2.8)$$

where $[x]$ denotes the concentration of x . In the steady state, the synthetic rate of $[ES]$ would be a constant, i.e. $d[ES]/dt = 0$. The total amount of E is also a constant as $E_T = [ES] + E$. Then the following results can be derived:

$$[ES] = \frac{SE_r}{S + \frac{k_r + k_{cat}}{k_f}} \quad (2.9)$$

$$\frac{dP}{dt} = \frac{k_{cat}SE_r}{S + \frac{k_r + k_{cat}}{k_f}} \quad (2.10)$$

$K_m = \frac{k_r + k_{cat}}{k_f}$ is called the Michaelis constant. When $k_r, k_f \gg k_{cat}$, we can assume that $K_m = k_r/k_f$.

The Michaelis-Menten equation represents a special case of the Hill equation (explained next), where the Hill coefficient is set to one. If the protein under study has more than one substrate binding sites, and if there is cooperativity with respect to substrate binding to the protein, the reaction rate as a function of the substrate concentration is no longer hyperbolic, because a hyperbolic function suggests non-cooperativity but may assume a sigmoidal shape.

Hill equation

The Hill equation is a sigmoidal function that describes the enhancement of a ligand binding to its target, e.g., a promoter region. In cases when the receptor is a dimer or tetramer rather than a monomer, the affinity of the receptor to the latter bound ligand is significantly increased [81]. The Hill coefficient provides a way to quantify this effect. For a two-gene network, the reaction rates of the entities can be formulated as Hill equations:

Figure 2.3: An example network with two nodes. Each node is auto-regulated and mutually inhibited by the other node.

$$\frac{dx}{dt} = v_{xx} \frac{x^{n_{x1}}}{K_{xd1}^{n_{x1}} + x^{n_{x1}}} + v_{yx} \frac{1}{K_{yd1}^{n_{y1}} + y^{n_{y1}}} - d_x * x \quad (2.11)$$

$$\frac{dy}{dt} = v_{xy} \frac{x^{n_{x2}}}{K_{xd2}^{n_{x2}} + x^{n_{x2}}} + v_{yy} \frac{y^{n_{y2}}}{K_{yd2}^{n_{y2}} + y^{n_{y2}}} - d_y * y \quad (2.12)$$

where K_{*d} is the dissociation constant, v_* is the maximal constant expression level, n_* is the Hill coefficient describing the cooperative binding which governs the slope, and d_* is the constant degradation rate.

Stochastic models

In regulatory networks, the law of mass action can be used to accurately calculate the change in concentrations when the number of involved molecules of each species is large. By contrast, when the number of molecules is small, significant stochastic effects may occur. The dynamics will manifest dif-

ferent behaviors, even starting from the same initial conditions. For example, in Hanna’s reprogramming experiment on early Pre-B cells, a continuous stochastic process was revealed with continuous supplementation of pluripotency factors [16].

Stochastic differential equations (SDEs) are a type of differential equations with at least one inherent random component. Its solution is a stochastic process described in terms of probability distributions. Thus SDE is suitable for modeling the stochastic nature of regulation by accounting for the fluctuations that occur at the molecular level [61].

A quantitative model is proposed in [65] based on a 52-gene regulatory network with 123 edges in human stem cells. The network includes both pluripotency factors and lineage markers. To account for the effect of stochastic noise, a system of SDEs with linear diffusion coefficients was employed (Eq. 2.13). $F(x)$ is the driving force composed of transcriptional regulations on gene x . ξ is defined with Gaussian distribution and denotes the noise term generated from either the external environmental fluctuations or intrinsic fluctuations.

$$\frac{dx}{dt} = F(x) + \xi \quad (2.13)$$

2.3.3 Current progress

In recent years, there have been extensive studies on dynamical systems modeling for understanding stem cell fate determination. Table 2.2 lists some existing representative models for stem cell differentiation or reprogramming using Boolean rules or differential equations. A few works have adopted the concept of Waddington’s epigenetic landscape to visualize the system in a 3D plot.

Most of the recent works explained cell fate decisions by purely modeling GRNs. One of the representative work is from [7]. In this model, a two-gene network with mutual inhibitions and self-activations is modeled in ODEs to simulate the lineage commitment during differentiation. Another

continuous model of a GRN with 52 nodes was proposed for the study of stem cell differentiation and reprogramming [65]. Not only continuous models, but rule-based discrete models are also widely used. [8] presents a fuzzy reasoning model based on the Fuzzy Petri Net. The model considers the regulatory triplets using predicting changes in expression level of the target gene based on input expression levels.

Besides focusing on core GRNs, signaling pathways and epigenetic mechanisms are encoded into the models to reveal more details of molecular interactions. For example, Flöttmann [72] made use of the probabilistic Boolean network to integrate the known regulatory factors, e.g., gene expression, chromatin modifications, and DNA methylation, to study stem cell differentiation and the production of iPSCs. [75] constructed an ODE model for understanding stem cell maintenance and lineage determination with external signals. It incorporates two external signals on Gata6 and Nanog respectively to induce reprogramming, i.e., signal Sg which represses Gata6, and signal Sn which induces Nanog. Also, the level of an external signal A on Oct4 determines the fate of cells.

Due to the computational complexity and limitation of biochemical information, the dynamical modeling method is only applicable to small core networks with detailed biological information. Also, the simulation and prediction results generated from the constructed model need verification. Moreover, existing computational models tend to focus on only one or two regulatory mechanisms, and as such may have oversimplified the molecular mechanisms of stem cell fate determination. A dynamical model that integrates TF-DNA interactions, epigenetic dynamics and signaling regulation is required for the aims of increasing our understanding of the regulatory mechanisms of stem cell fate decision and finding new ways to improve the safety, reproducibility, and efficiency of stem cell reprogramming and guided differentiation.

In addition, an ever-increasing amount of data from new high-throughput techniques in stem cell research makes the data-driven methods possible.

These methods can accelerate new discoveries in the field as reviewed in the next section.

2.4 Data-driven approach

Bulk sequencing data, which measure the average expression level for each gene across a large population of cells, have been widely used for transcriptomics, quantifying expression signatures from ensembles. However, it is insufficient for studying heterogeneous systems, e.g., early development studies, complex tissues (for example brain) and tumour cells. An ever-increasing amount of single-cell data from new high-throughput techniques in stem cell research makes the data-driven methods popular in the field which are reviewed in this section.

2.4.1 Single-cell technology

The field of single-cell genomics is advancing rapidly and generating more high-resolution single-cell sequencing data, which enable new biological discoveries [84]. We provide an overview of the state of the art in the field.

qPCR

Quantitative polymerase chain reaction (qPCR) is one of the high-throughput RNA counting techniques for measuring the genome-wide gene expression of previously known genes of single cells [43]. It uses two-sided primers of the known genes for the PCR amplification, which however limits the number of measured genes immensely but makes dropout events less likely. The gene expression is the count of detected DNA molecules. The targeted nature of qPCR makes this method less prone to dropout events in lowly expressed genes, but there is an expression detection cap at the high end, making it difficult to differentiate between highly expressed genes.

scRNA-seq

Single-cell RNA sequencing (scRNA-seq) has rapidly gained popularity for profiling transcriptomes of hundreds of thousands of single cells since 2009 [85]. This technology has led to the discovery of novel cell types and insights into the development of complex tissues, e.g., cell type identification, heterogeneity in cellular responses to drugs, stochasticity of gene expression, the inference of gene regulatory networks, etc.

Currently, there are several different protocols in use, e.g. SMART-seq2 [86], CELL-seq [87] and Drop-seq [88] *etc.* However, due to minute amounts of starting material, samples undergo extensive amplification, which might increase technical variability. Dropout events are another main source of discrepancy between the libraries. Through proper normalization and corrections, it is possible to alleviate some of these issues.

Mass cytometry

Mass cytometry is a mass spectrometry technique based on inductively coupled plasma mass spectrometry and time of flight mass spectrometry used for the determination of the properties of cells [89].

Mass cytometry can investigate millions of single cells from an individual sample with over 40 parameters [89]–[91]. The parameters are not restricted to one level of cellular products. Rather, posttranslational modifications, proteolysis products, etc. can all be quantified from a single experiment [92].

Therefore it is applied to analyze complex cellular systems by observing the diversity of cellular phenotypes and behaviors in a single sample [93].

2.4.2 Hopfield network

The artificial neural networks (ANNs) that build on multiple similar information processing blocks imitate the architecture and function of a human brain. They are capable of modeling and processing linear or nonlinear re-

relationships between inputs and outputs in parallel. Hopfield network (HN) is a type of recurrent neural network in which all neurons are both inputs and outputs. It offers an attractive framework for quantitative modeling of cell states from empirical data.

HN was proposed by John J. Hopfield in 1982 based on spin-glass models [94]. It consists of a set of N interconnected neurons which update their activation values asynchronously or synchronously. A matrix $W = \{W_{ij}, i, j = 1, 2, \dots, N\}$ represents the interactions between these nodes. The changes in the activation values of the output neurons are significant such that the dynamical properties of the network are important.

A primary application of the HN is to model the recall of associative memories which are a form of long-term memory. When the network is cued with a noisy or incomplete test memory, it will render the incorrect or missing data by iterating to a stable state which is in some sense near the cued pattern. The behavior of the system can be described with an energy function determined by the weights of the connections between the neurons. In this case, the weight matrix has to be set such that the states of the system corresponding to the memories that are to be stored in the network are stable. These states can be seen as dips in an energy space. Every time the network is updated the energy level decreases (or is unchanged). The activation values of the units with low energy indicate stable states in which these activations do not change anymore.

Neuronal networks such as HN have diverse applications. For example, it has been used for making stock market predictions [95], time series prediction [96], super-resolution mapping of land cover features [97] and recognition of blurred images [98].

Discrete Hopfield network

The original HN proposed in 1982 is a two-state model, i.e. each neuron has two states (e.g. $+1, -1$). The activation of each node is assigned either $+1$ or -1 . During clamping, the activation state y_i of the i th unit can

be updated either synchronously or asynchronously in discrete time steps. The strength of the synapse from node j to node i is defined in the weight matrix w_{ij} . For synchronous updates, the state y_i of each node is updated simultaneously according to the rule:

$$y_i = \text{sgn}\left(\sum_{j=0}^N w_{ij}y_j\right) \quad (2.14)$$

where the function sgn is defined as

$$\text{sgn}(z) = \begin{cases} +1 & \text{if } z \geq 0 \\ -1 & \text{if } z < 0 \end{cases} \quad (2.15)$$

In asynchronous updating, one node that is chosen randomly or according to some rule is updated at each time point. For example, in the update rule of Glauber dynamics, one gene is selected at random and probabilistically updated according to its local field at each time step [99].

Continuous Hopfield network

Considering the state of neurons in a real biological system is continuous, Hopfield proposed the continuous Hopfield network (CHN) in 1984 [100]. It also consists of a set of N interconnected neurons which update their activation values synchronously or asynchronously. Compared with the original two-state HN proposed by Hopfield himself in 1982 [94], CHN uses continuous variables and predicts continuous responses.

In the framework of CHN, the change rate of neuron i is modeled by

$$\frac{dy_i}{dt} = C_i \sum_{j=1}^N W_{ij}U_j - \delta_i y_i + I_i, \quad (2.16)$$

$$U_j = g_j(y_j), \quad (2.17)$$

where W_{ij} is an entry of the weight matrix of CHN representing the interconnection weight coefficient from neuron j to neuron i , and C_i is an amplifier

on the synaptic connections. The external input I_i represents a combination of propagation delays. The activation function $g_j(y_j)$ represents the input-output relationship of a nonlinear amplifier with negligible response time. The activation function is required to be a monotonically increasing function to make the system stable [101], e.g. a sigmoid activation function (Eq. 2.18).

$$g_j(y_j) = 1/(1 + e^{-y_j}). \quad (2.18)$$

The behavior of a CHN system can be described with an energy function (Eq. 2.19) which is a Lyapunov function:

$$E = -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N W_{ij} U_i U_j + \sum_{i=1}^N I_i U_i + \sum_{i=1}^N \delta_i \int_0^{U_i} g_i^{-1}(u) du. \quad (2.19)$$

2.4.3 Current progress

The rapid increase of single-cell transcriptomic data provides new opportunities for exploring landscape models using data-driven approaches. HN offers an attractive framework for quantitative modeling of stem cell fate determination from single-cell data.

The discrete HN has been used to study the biological process of cell development [11]–[13]. Maetschke and Ragan [11] constructed a discrete HN with attractors denoting cancer subtypes. The energy of the network was defined as potentials of the landscape. For the visualization of the landscape, the first two components of PCA were selected as the X- and Y-coordinates. A similar approach was developed to model the stem cell fate transition processes [13]. It successfully explained partially reprogrammed cells and identified potential TFs that can be used to induce reprogramming. Taherian Fard used the same model to study the transition process in stem cell development [12].

However, the landscape constructed by discrete HN cannot fully capture the dynamics of a real biological system in which the values of elements

(e.g., the expression values of genes) are continuous. In particular, the heterogeneity characteristic of single-cell sequencing data requires high resolution of models. Thus we adopt the framework of CHN to construct the Waddington’s epigenetic landscape as a valuable tool to guide future studies and engineering of stem cell fates.

2.5 Summary

Experimental manipulations of cell fates provide the unprecedented opportunities for scientists to contribute to human health care. For example, the induction of pluripotent cells generates patient-specific disease models and transplant resources.

However, the application of cell manipulation faces challenging problems. First, the transitions between two cell states are only partially understood due to the lack of knowledge of the underlying mechanisms. For example, the discovery of the cocktails for stem cell reprogramming relies on experimental selections without fully understanding the mechanisms [1]. Secondly, the current studies and applications are hindered by the low transition efficiency and the low quality of products. Some researchers are trying to discover the bottlenecks [16], and new cocktails are identified that could significantly increase the conversion rate. Meanwhile, for further usage of the engineered cells, rigorous evaluation of the quality is crucial.

To address these issues, computational modeling methods have been applied to discover molecular regulations in stem cells. In a mathematical model of gene regulation, the update of a gene is determined by its interactions in a network (e.g., GRNs, signaling pathways). The connections between genes have been assigned specific regulatory functions according to the knowledge of mechanisms or inferred from experimental data. The system dynamics are visualized in the Waddington’s epigenetic landscape. The changes of cell states which are determined by the dynamic regulations are along the trajectories on the surface of the landscape. Both the knowledge-

based methods and the data-driven approaches yield inspiring insights into stem cell fate transitions. Subsequently, they can enhance our capability to engineer cells, for example, the conversion from somatic cell types to iPSCs.

In the next chapter, a knowledge-based mathematical model is proposed based on a hybrid network which integrates crucial regulatory factors in ESCs, i.e., external signals, epigenetic modifiers and core transcriptional network. A Waddington's epigenetic landscape is constructed to study the cell fate transitions in differentiation and reprogramming. It captures the roles of epigenetic factors as barriers to cell reprogramming.

Chapter 3

Knowledge-based mathematical modelling

The transitions of cell fates, like embryonic stem cell differentiation, as a result of the time evolution of cellular kinetics, are determined by the collective dynamics of regulatory networks. However, the underlying mechanisms need further elucidation. For example, it remains unclear how the ectopic expression of a few defined TFs can induce dynamic changes in gene expression profiles that establish a pluripotent state during stem cell reprogramming.

In this chapter, a model of core gene regulatory network is proposed to explore the dynamic characteristics of the stem cell reprogramming and differentiation processes, and especially the critical role of epigenetic modification in the reprogramming will be analyzed.

3.1 Construction of the dynamic model

3.1.1 Gene regulatory network

A core network with four nodes, i.e., Oct4, Sox2, MEs, and ECTs, has been proposed in [18] to analyze the cell state transitions in reprogramming experiments using lineage specifiers as substitutes for Oct4 and Sox2. In this network, MEs represents a group of marker genes of the mesoderm

lineage, while ECTs represents ectodermal genes. The 4-node transcriptional network is extended to a hybrid network including three layers of regulations, i.e., signaling, genetic and epigenetic regulations, which are considered to determine the transitions of cell fates. The three categories of regulators are briefly reviewed below.

The first category is the TFs, such as pluripotency factors, and lineage specifiers. The pluripotency-associated genes are the primary regulators that control the maintenance and self-renewal of ESCs. They often form regulatory complexes to regulate a group of genes thereby weaving a regulatory network. For example, Sox2 binds to promoters cooperatively with Oct4 at non-palindromic sequences to activate the transcription of crucial pluripotency factors. It was also found that these ESC markers can promote lineage commitment, e.g., Oct4 promotes the mesoderm lineage (ME) and primitive endoderm differentiation, and at the same time Oct4 suppresses the ectoderm lineage (ECT) [102]. On the contrary, the lineage specifiers are essential markers of specific lineages. In general, the lineage commitment is the result of the mutual inhibitions between pluripotency factors and lineage markers.

The second category is the epigenetic modifiers, which are enzymes and considered to be critical in cell fate determination [103]. Six types of epigenetic modifications including histone modification, DNA methylation *etc.*, are considered in the network model. The chromatin states have a significant impact on the interactions between TFs and DNA sequences. For example, the binding of epigenetic activators to MLL complex can alter the epigenetic states of pluripotency genes (such as Oct4, Sox2, and Nanog) from suppressed states (e.g. DNA methylation and H3K27me3) to activated states (e.g. histone acetylation and H3K4me3), which leads to the up-regulation of these genes [104]. Meanwhile, the abundances and activities of these epigenetic modifiers are also regulated by TFs, e.g., Oct4 [105].

The third category is the cell signaling pathways that coordinate the activities of individual cells. For instance, Notch signaling pathway functions

at all stages of development to regulate cell proliferation, survival, and differentiation [106]. BMPs play an important role in the regulation of stem cell pluripotency. It blocks ESC neural differentiation and maintains ESC pluripotency by forming a balance between growth factor (LIF and FGF) signaling pathways [107]. ESCs remain undifferentiated in the presence of the cytokine LIF. Specification and differentiation of germ layer derivatives in mouse embryos require signaling induction [108]. To study the influence of signal transduction on cell fate determination, we integrated three cell signals that govern Gata6, Sox1, and Oct4 respectively. Then we simulated the cellular dynamics using the model with different strength of signals to investigate the influence of signaling pathways on cell fate changes.

The final constructed circuit contains six epigenetic modifiers, two lineage specifiers (Sox1, Gata6), five ESC-related genes (Klf4, Myc, Oct4, Nanog, and Sox2), and three common genes (Gcnf, Pax6, Cdx2) in Figure 3.1. Five additional nodes representing the amounts of the regulatory DNA sequences are included as well. Klf4, Myc, Oct4, and Sox2 are also the Yamanaka factors used in the iPSC reprogramming experiment. They control various downstream target genes that are essential in ESCs. Lineage markers including Gata6 and Sox1 are selected to represent lineage commitment. Gata6, which is derived from lateral mesoderm, plays an important role in the regulation of cellular differentiation and organogenesis during vertebrate development [109]. Sox1 is one of the earliest transcription factors to be expressed in the ectodermal cells. They are the marker genes of mesoderm and ectoderm lineages respectively. The interactions between them have been verified by experimental observations, such as ChIP-chip or ChIP-Seq experiments. There are 75 edges in this network, where 34 edges are activating and 41 are inhibitory.

3.1.2 Dynamic formulation

The time evolution of the concentrations of TFs includes two timescales: a fast timescale if we assume rapid kinetics of protein-DNA binding, and a

Figure 3.1: The 10-node transcriptional network with epigenetic modifiers in mouse embryonic stem cells. Gata6 and Sox1 are lineage specifiers. Klf4, Myc, Oct4, Nanog and Sox2 are pluripotency factors. Edges with flat heads indicate inhibition, and edges with arrowheads represent activation.

slow timescale for the processes of transcription and protein degradation. We assume that the binding-unbinding reactions are so fast that we can use their equilibrium concentrations to compute the gene occupancy. A discrete model and a continuous model are proposed below.

Boolean model

A Boolean model was constructed based on the network. The construction and simulation were performed using the software “GINSim” [110]. Table S4 lists the transition functions. Three external signals, Sig0, Sig1 and Sig2, work on Gata6, Sox1 and Oct4 respectively driving the generation of lineages.

The attractor system contains four stable states in Table 3.1 when the external cues are muted, including stem cell state (S1), trophectoderm lineage (S2), ectoderm lineage (S3) and mesoderm lineage (S4). After adding signals to the network, e.g., Sig2 on Oct4, Sig0 on Gata6, ESC can maintain

its status or differentiate into different lineages.

Table 3.1: Steady states identified from synchronous updating simulation.

#	Oct4	Sox2	Nanog	Gcnf	Cdx2	Pax6	Sox1	Gata6	Myc	Klf4
S1	1	1	1	0	0	0	0	0	*	*
S2	0	0	0	0	1	*	0	0	0	0
S3	0	0	0	0	0	*	1	0	0	0
S4	0	0	0	0	0	*	0	1	0	0
S5	0	0	0	0	0	*	0	0	0	0

Note: * indicates both 0 and 1

The state transition graph shows the differences between the network with and without epigenetic modifications in both differentiation and reprogramming experiments. Only a proportion of ESCs in the start stage is differentiated into somatic cells under the control of epigenetic modifiers, while the conversion rate is 100% in the network without epigenetic modifications. The same result is observed in the in silico simulation of Yamanaka's reprogramming experiment by raising the basic rates of synthesis of Oct4, Sox2, Klf4, and Myc, and we also repeat the result in the seesaw reprogramming experiment using Gata6, Sox1 as substitutes of Oct4 and Sox2 on this discrete model with epigenetic modifiers.

Continuous model

In the continuous model, the change of the concentration of each gene's product was encoded in the form of ODEs. The expression level of a gene can be regulated by the exogenous expression, transcription factor binding to its DNA sequences, temporal epigenetic state, and self-degradation (Eq. 3.1).

$$\begin{aligned}
 \frac{d[Gene_i]}{dt} = & Basal_{Gene_i} + E_{Gene_i} \times S_{Gene_i} \times Sig_{Gene_i} \\
 & \times \varphi(Gene_1, Gene_2 \dots Gene_N) - \\
 & K^-[Gene_i] + D \cdot \theta(t),
 \end{aligned} \tag{3.1}$$

where $[Gene_i]$ denotes the concentration of $Gene_i$. $Basal_{Gene_i}$ is the exogenous expression of $Gene_i$. E_{Gene_i} is the maximum transcriptional rate determined by the epigenetic state of $Gene_i$. S_{Gene_i} is the state of $Gene_i$'s DNA sequence. K^- is the degradation rate, $\varphi(\cdot)$ represents the cis-regulation of $Gene_i$ by other nodes, $\theta(t)$ is the noise term, and D controls the strength of noise. External signals are transmitted into the cells through phosphorylation, de-phosphorylation and protein binding. They are simplified as a parameter Sig_{Gene_i} on the target genes.

For each target gene, $\varphi(\cdot)$ consists of two parts, i.e., an active component and an inhibitive component, denoting the regulations of activators and inhibitors by Hill equations (Eq. 3.2).

$$\varphi(\cdot) = \sum_{i=1}^{m_a} \frac{Gene_i^{n_a}}{Gene_i^{n_a} + k_a^{n_a}} + \sum_{i=1}^{m_b} \frac{k_i^{n_i}}{Gene_i^{n_i} + k_i^{n_i}} \quad (3.2)$$

In the above equation, m_a and m_b denote the numbers of activators and inhibitors respectively. k_a and k_b are activation constant and inhibition constant. n_a and n_i are Hill coefficients.

Two modules each comprising three mutually exclusive elements constitute the epigenetic network in Figure 3.1. The epigenetic modifiers control the accessibility of target genes' cis-regulatory elements and subsequently regulate the expression of genes. The same formula is applied to encode the change of the concentration.

3.1.3 Parameter estimation

The simulated annealing (SA) method was used to infer the parameters (Algorithm 1). Since the model is intended to simulate the stem cell reprogramming, we assume that there are three specific steady states representing the stem cell state ($Attractor_3$), ectoderm lineage ($Attractor_1$) and mesoderm lineage ($Attractor_2$) respectively. The attractors are represented by expression values of marker genes. The cost function in Eq. 3.3 is defined to capture the three groups of attractors representing ESC, mesoderm lineage

and ectoderm lineage with their marker genes highly expressed.

In our model, we define *Attractor*₃ as distinct cell states with highly expressed Oct4, Sox2 and down-regulated Sox1 and Gata6. Oct4, Sox2 are inhibited in both *Attractor*₁ and *Attractor*₂. In *Attractor*₁, Sox1 is down-regulated but Gata6 is activated. *Attractor*₂ is another lineage with Sox1 up-regulated and Gata6 inhibited. Considering the computational tractability and our previous experience, we calculated $N = 50$ trajectories with random initial states in each update of parameters, and then checked how many of the final states have converged to the predefined three types of attractor states. Ideally, all time-series will be stable and converge to the three attractors which constitute three parts of the cost function.

$$\begin{aligned} cost = & |N/3 - \#of Attractor_1| + |N/3 - \#of Attractor_2| \\ & + |N/3 - \#of Attractor_3| + (N - \#of Attractor_1 - \#of Attractor_2 - \\ & \#of Attractor_3) + |3 - \#of attractors| \end{aligned} \quad (3.3)$$

The full list of the ODEs and inferred parameters of the above continuous model can be found in Appendix (Eq. S6.2, Table S6).

3.2 Construction of the landscape

3.2.1 The probabilistic landscape

After the knowledge-based dynamical model is constructed, the next step of modeling is to quantify and visualize the Waddington's epigenetic landscape for stem cell fate transitions. The quasi-potential function of the probabilistic landscape is based on the probability density function of steady states [66]. The self-consistent mean field approximation method [65] described in Section 2.2, Chapter 2 can be used to find the numerical solution of FPE.

Under the assumption that gene concentration is independent and follows

Algorithm 1 The pseudocode of the Simulated Annealing algorithm for parameter estimation

INPUT: $iterations_{max}, temperature_{max}$

OUTPUT: $Param = [a, b, ka, ki, na, ni, a1, b1, ki1, ka1, na1, ni1]$

Initialization: $Param_{current}$ is set to random values

Iteration:

```

1: for  $i = 1$  to  $iterations_{max}$  do
2:    $Param_{new} = createNewSolution(Param_{current});$ 
3:    $temperature_{current} = calculateTemperature(i, temperature_{max})$ 
4:   if  $cost(Param_{new}) \leq cost(Param_{current})$  then
5:      $Param_{current} = Param_{new}$ 
6:   else
7:     if  $exp[(cost(Param_{current}) - cost(Param_{new})) / temperature_{current}] >$ 
 $random()$  then
8:        $Param_{current} = Param_{new}$ 
9:     end if
10:  end if
11: end for
12: return  $Param_{current}$ 

```

Gaussian distribution, two differential equations for mean μ_i and standard deviation σ_i of each gene, were constructed according to [65]. Based on these equations, multiple trajectories were generated to get the steady solutions. The initial states were chosen as quasi-random points produced from the Sobol sequence which is a base-2 digital sequence that fills space in a highly uniform manner. $P(x)$ is calculated as the product of $P(x_i)$ for gene x_i .

3.2.2 Visualization of multi-dimensional landscape

Waddington's epigenetic landscape is visualized as a 3D plot with the z-axis representing quasi-potential, while the XY-plane represents the expression space. Usually, the other two coordinates are the expression values of two marker genes selected from the gene regulatory network. Figure 3.2 shows the landscape of the network in Figure 3.1 with expression values of Sox1 and Gata6 as x-axis and y-axis. Three attractors can be found representing stem cell state (low expression of both lineage markers), ectoderm (up-regulation of Sox1) and mesoderm (up-regulation of Gata6) respectively.

However, in most of the cases, the expression space of two genes are not

Figure 3.2: The probabilistic landscape of the constructed gene network.

enough to cover the whole system dynamics when more than two genes are involved. Thus a suitable dimensionality reduction method is required to project a high-dimensional space to a 2D space.

Dimensionality reduction

Linear projections have traditionally been preferred for the ease of computing. Principal component analysis (PCA) is one of the most widely used methods. It uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. Kernel principal component analysis (kernel PCA) is an extension of PCA using a kernel to map the originally non-linear space to a linear mapping. Locally linear embedding (LLE) seeks an embedded space for which locally linear relationships are preserved. Isomap computes an approximation to the geodesic distance through constructing a neighborhood graph. Semidefinite embedding maximises the variance in the latent space with a constraint that local distances should be preserved. These methods focus on preserving local distances but not long distances. This might cause the overlap of mapping in the resulting embedding. When there exists noise in the system, it might lead to failure

to capture the underlying manifold structure [111].

In conclusion, the projection of high-dimension data to a latent space requires the solutions of two problems. One is how to reduce the dimensionality while preserving the distances and similarities between data points. The other is how to map data from the latent space back to the original space.

Gaussian process dynamical model (GPDM) is a probabilistic approach to modeling high-dimensional time-series data in a low-dimensional latent space with a dynamical model [112]. It has been used for nonlinear time-series analysis, with applications in learning models of human pose and motion from high-dimensional motion capture data [113], [114]. In the GPDM framework, the original data consisting of N points, $Y = [y_1 \dots y_N]$, with D dimensions are recovered from the latent space with the corresponding two-dimensional data $X = [x_1 \dots x_N]$. The relationship between the latent space data x and the high-dimensional data y with Gaussian noise would be

$$y_i = Wx_i + \sigma_i, \quad (3.4)$$

where $W \in R^{D \times 2}$ which has a prior as a spherical Gaussian distribution with zero mean and $\sigma_i \in R^{D \times 1}$ sampled from a spherical Gaussian distribution with 0 mean and $\beta^{-1}I$ covariance. Thus the likelihood of all data points is

$$p(Y|X, \beta) = \prod_{d=1}^D p(y_{.d}|X, \beta) = \prod_{d=1}^D N(y_{.d}|0, K) \quad (3.5)$$

where the kernel matrix is $K = XX^T + \beta^{-1}I$. The mapping between the latent space and observational space is achieved through maximizing the marginal likelihood in Eq. 3.5. Complex kernels could be applied, e.g. linear + RBF [113], to guarantee the smoothness and nonlinearity of the output function. Normally the data are independent in the analysis, while the time-series data have dynamic connections between consecutive points which are not considered in Eq. 3.5. GPDM can model time-series data using a dynamic model initialized by PCA and trained using scaled conjugate

gradient (SCG).

In addition, to make the points that are close in the original data space also close in the latent space, a back constraint is added to force the latent points to be a smooth function of the observational data points [111]. The latent points are composed by a function of input points.

Visualization of the landscape

A complex kernel of RBF, with linear and white noise, was selected to learn the latent coordinates. The back constraints made use of a multi-layer perceptron with ten hidden units. To generate a smooth surface in the latent space, new points were interpolated in the latent space and mapped back to the original space to obtain their quasi-potentials using the projection model learned from the simulated data.

The constructed landscape is shown in Figure 3.3. It shows a similar distribution of attractors as in Figure 3.2, i.e. one stem cell state, and two differentiated cell states.

Figure 3.3: The projection of a multi-dimensional data space to a 2-dimensional latent space using GPDM. The colour represents the level of quasi-potential U .

3.3 Epigenetic barriers in cell fate alterations

We performed the *in silico* reprogramming experiment using the deterministic models, one with epigenetic modifiers and the other without epigenetic modifiers, with the initial state selected as a differentiated state when Gata6 has highly expressed and pluripotency factors, as well as other lineage markers, are down-regulated. The difficulties of cell state transitions were evaluated by the energy barriers between attractors in the landscapes with and without epigenetic modification.

Gillespie stochastic algorithm, an algorithm for simulating the kinetics of a set of coupled chemical reactions, is employed to solve the stochastic model. In Gillespie's approach, the chemical reaction kinetics are modeled as a Markov process in which reactions occur at specific instants of time defining intervals that are Poisson-distributed, with a mean reaction time interval that is recomputed after each chemical reaction occurs. For each chemical reaction interval, a specific chemical reaction occurs, randomly selected from the set of all possible reactions with a weight given by the individual reaction rates.

After the replenishment of the four pluripotency factors (Oct4, Sox2, Klf4, and cMyc), the conversion rate from somatic cells to iPSCs is calculated from the two models. The landscape structures with three attractors compared to one attractor indicate the barrier role of epigenetic during the reprogramming events (Figure 3.4). In the model without epigenetic factors, only one attractor representing the stem cell state was revealed after the reprogramming treatment (Figure 3.4 (a)) indicating a 100% conversion rate. On the contrary, the epigenetic modification made all three types of attractors preserved (Figure 3.4 (b)) reaching a rate of near 0.6% in the full model with epigenetic modifications. The different topologies of the landscapes verified the effects of epigenetic barrier to iPSC reprogramming which have been previously validated in Jacob Hanna's reprogramming experiments [15].

In Jacob's experiment, Mbd3 was found to be a major epigenetic barrier

Figure 3.4: Visualization of Waddington's epigenetic landscape. (a) Model without epigenetic modifiers. (b) Model with epigenetic modifiers. The dark areas indicate cell states with low potentials, and the light regions have high potential values.

to the iPSC reprogramming using the Yamanaka's reprogramming factors that the repression of *Mbd3* expression can improve the conversion rate to nearly 100%. We simulated the *Mbd3* depletion in our proposed model and predicted the reprogramming rates. The simulation result is 100% in our model which is consistent with the experimental observations.

3.4 Summary

The mathematical model constructed in this chapter provides an example of knowledge-based modeling to study the mechanisms of stem cell reprogramming and differentiation. It integrates multiple regulatory factors and can be used to generate testable hypotheses for future research.

The probabilistic landscape of the mathematical model of the constructed gene network is composed of three stable states theoretically representing the stem cells, mesendodermal cells and ectodermal cells according to the gene expression levels of lineage marker genes. We successfully simulated the reprogramming process under experimental conditions, e.g., the overexpression of *Oct4*, *Sox2*, *Klf4*, and *cMyc*, and the predicted low conversion rate is consistent with published wet-lab experiments. By comparing the landscape structures of models with and without epigenetic modifiers, it recapitulates

the role of epigenetic modification as barriers to state transition during the process of reprogramming from somatic cells to iPSCs.

With quantitative modeling of Waddington's epigenetic landscape, the systems dynamics of transcriptional regulation driving cellular phenotypic changes can be more easily analyzed and predicted. Therefore, the modeling and visualization of Waddington's landscape can facilitate the studies of cell fate conversion and help design innovations in cell fate engineering and regenerative medicine.

Chapter 4

NetLand: quantitative modeling and visualization of Waddington's epigenetic landscape using probabilistic potential

Waddington's epigenetic landscape is a powerful metaphor for cellular dynamics driven by regulatory networks. Its quantitative modeling and visualization, however, remain a challenge, especially when there are more than two genes in the network. A software tool for Waddington's epigenetic landscape has not been available in the literature.

In this chapter, a software tool named NetLand for quantitative modeling and visualization of Waddington's epigenetic landscape using probabilistic potential has been introduced. With an interactive and graphical user interface, NetLand can facilitate the knowledge discovery and experimental design in the study of cell fate regulation (e.g., stem cell differentiation and reprogramming) (Figure 4.1).

Figure 4.1: A case study of a multi-attractor system by NetLand. Screenshot of NetLand on a 4-node network for mouse stem cell reprogramming. The 4-gene regulatory network (a) for mouse stem cell reprogramming [18] has three attractors shown in (d). Based on the mathematical model, time-series data of reprogramming experiment are simulated (b), and the network dynamics is visualized in Waddington’s epigenetic landscape (c).

4.1 Overview of NetLand

Practical applications of current modeling methods need address several limitations. First, there is a lack of user-friendly software tools to calculate and visualize the landscape. Secondly, it is still difficult to visualize the 3-dimensional landscape for a regulatory network with more than two genes. Although a 52-gene regulatory network has been recently used to plot a potential landscape [65], their method either selects only two marker genes for the coordinates of the state space or requires the system to have just two attractors.

For the theoretical models of GRN and landscape to be widely used, we need user-friendly software to automate the steps of model construction, dynamic simulation, and visualization of the Waddington’s epigenetic landscape, etc., so that users without experience in computational modeling can also benefit from these models.

To address these issues, we have developed a software tool called “NetLand” which is intended for modeling, simulation, and visualization of GRNs and their corresponding quasi-potential landscapes. It provides a user-

friendly interface to facilitate model construction, dynamical simulation and landscape plotting without restriction on the number of genes in a GRN (Figure 4.2). Hopefully, NetLand can help make the quantitative models of Waddington’s epigenetic landscape widely used by researchers of diverse backgrounds.

There are mainly five functional modules, i.e., the input module, the mathematical modeling module, the modification module, the simulation module and the landscape module. Users can import models of GRNs from a file in formats such as TSV, SBML, etc., and manually edit the network structure. Then, NetLand will automatically encode differential equations for modeling the kinetics of transcriptional regulations. Users can also modify model equations and parameters, e.g. adding external inputs as signals to the equations. As such, a computational model is constructed and can be used to simulate the dynamics of the input networks and generate the Waddington’s epigenetic landscape.

Figure 4.2: An overview of the functional modules in NetLand. There are mainly five functional modules, namely the input module, the mathematical modeling module, the modification module, the simulation module and the landscape module. It takes networks as input, and generates dynamic models for timeseries simulation and landscape construction.

To display the 3D landscape for a network of more than two genes, NetLand allows users to either project the landscape to a state space of two chosen marker genes or project to a latent space using a dimensionality re-

duction method called the Gaussian process dynamical model [112], [113]. Based on a nonlinear latent variable model, GPDM can take into account the dynamics of time-series gene expression data and is capable of projection in both directions between the low-dimensional latent space and the high-dimensional observation space.

Therefore, NetLand can provide a global picture of cellular dynamics for a user-specified GRN. Although we designed the NetLand software initially for modeling stem cell fate transitions, it can also be used to study other cellular phenotypes, such as cancer cell death, cellular aging, etc.

The software NetLand can be downloaded from [link:NetLand](#).

4.2 Dynamic model construction and simulation

In this section, we introduce the automatic modeling process in NetLand, as well as the three types of simulators integrated for model simulations. NetLand also provides interfaces for the modification of a network structure and elements of a dynamic model. A brief introduction to the model operations is also included.

4.2.1 Constructing a dynamic model

Loading a model

NetLand is capable of loading network structure files (in formats of TSV, GML, or DOT) or computational models in the systems biology markup language (SBML), e.g. models extracted from the BioModels database [115].

A network is often represented as a graph having a set of vertices and edges. The nodes correspond typically to genes, mRNAs, proteins, complexes, *etc.*, while the edges represent interactions, regulation, co-expression, co-location, *etc.*. Network structure files, such as TSV, are formats for storing the topology of a network. The dynamic model is automatically gener-

ated using the formulation rules introduced in the next section.

Computational models of biological processes are stored in representation formats, such as SBML based on XML. It has been applied for the representation of many different classes of biological phenomena, including metabolic networks, cell signaling pathways, regulatory networks, *etc.*. NetLand can process SBML files that describe transcriptional regulations.

Formulation

A computational model of a biological network typically contains three key components, i.e., molecular species (nodes), biochemical reactions (edges), and parameters. The molecular species are the products of genes, functioning as reactants or modifiers in a reaction. Biochemical reactions describe the regulatory interactions between species. Parameters control the kinetics of the reactions, such as the production and degradation rates.

In NetLand, the transcriptional regulation is formulated in ODEs or SDEs. The algebraic equations depict the instantaneous changes of molecular concentrations over time in the form of Hill equations (Eq. 4.1).

$$\frac{d[Gene1]}{dt} = K^+ \varphi(Gene1, Gene2 \dots GeneN) - K^- [Gene1] + D \cdot \theta(t), \quad (4.1)$$

where $[Gene1]$ denotes the concentration of the products of a gene named "Gene1". K^+ is the maximum transcriptional rate of "Gene1", and K^- is the degradation rate. $\varphi(\cdot)$ represents the cis-regulation of *Gene1* by other genes, $\theta(t)$ is the noise term, and D controls the strength of noise. For each target gene, $\varphi(\cdot)$ consists of two parts, i.e., an active component and an inhibitive component, denoting the regulations of activators and inhibitors by Hill equations (Eq. 4.2).

$$\begin{aligned}
\varphi(Gene1, Gene2 \dots GeneN) = & BasalExpression_{Gene1} \\
& + \sum_{i=1}^{m_a} (BasalExpression_{m_{ai}} + a_i * \frac{Genei^{n_a}}{Genei^{n_a} + k_a^{n_a}}) \\
& + \sum_{i=1}^{m_b} (BasalExpression_{m_{bi}} + b_i * \frac{k_b^{n_b}}{Genei^{n_b} + k_b^{n_b}})
\end{aligned} \tag{4.2}$$

In the above equation, m_a and m_i denote the numbers of activators and inhibitors respectively. Two coefficients, a_i and b_i , represent the strength of activators and inhibitors to their target. k_a and k_b are activation constant and inhibition constant. n_a and n_b are Hill coefficients. $BasalExpression_{Gene1}$ is the basic expression value of *Gene1* besides the regulation from other genes in the network, while the two parameters, i.e., $BasalExpression_{m_{ai}}$ and $BasalExpression_{m_{bi}}$, denote the basic values of the regulation of a specific regulator. The default values of the above parameters are defined in Table 4.1.

Table 4.1: The default values of parameters in Eq. 4.1 and 4.2.

Parameter	Default value
K^+	1
K^-	1
D	0.05
$BasalExpression_{Genei}$	0
$BasalExpression_{m_{ai}}$	0
$BasalExpression_{m_{bi}}$	0
a_i	1
b_i	1
n_a	4
n_b	4
k_a	0.5
k_b	0.5

This formulation may not apply to some other types of biochemical reaction networks, e.g., signaling systems, metabolic networks. For the analysis of those networks or models other than GRNs, an existing model in SBML format (e.g., extracted from BioModels database or saved from previous

model editing tools such as CellDesigner [116]) is preferred, and then the model can be edited in NetLand which will be elaborated in the following section.

Model operation

NetLand provides flexible control to load and modify a network model. In this section, we describe how to use NetLand to modify the network model, e.g., deleting nodes, adding new nodes or edges, editing parameters and reactions. Using the user-friendly interactive interfaces, users can easily modify the network structure and the kinetic model.

By modifying the structure of a network, users can simulate biological experiments. For example, a node in a gene regulatory network can be blocked (i.e., inhibited or suppressed) to reproduce the gene knockout experiment. In addition to blocking genes, users can also add genes and edges to build new nodes and regulatory interactions. Once a new gene is inserted, users will be asked to verify a set of parameters including the initial value, degradation rate, *etc.* The kinetics of the new regulatory relations (associated with the edges) will be encoded with Hill equations (Eq. 4.2) and automatically added to the target genes' original formulas. Besides the network structure, the elements of the dynamic model including the initial values of the variables for genes' activities (e.g., expression levels), the reactions of genes and parameters, can be directly modified to encode the interactions in a more flexible way.

4.2.2 Running a simulation

Three simulators are implemented in NetLand based on numerical algorithms of differential equations. In this section, we will introduce how to use NetLand to perform dynamical simulation using different simulators.

Different simulators are integrated to solve different types of models. The Runge-Kutta method using finite difference methods is implemented for the approximate solution of a deterministic model with a set of ODEs. The

stochastic model described as Langevin equations (SDEs) [61] is numerically solved using the Euler-Maruyama method [117]. The above two methods are suitable for models with high molecular concentrations. When the copy numbers of molecules are small, leading to significant stochastic effects, users can choose Gillespie algorithm [118].

ODE solver

A deterministic simulator, i.e., the 5(4) Dormand-Prince (RKDP) integrator [119], implemented in the package of `org.apache.commons.math3.ode` which provides classes to solve non-stiff ordinary differential equations problems is used to solve ODEs. The method belongs to the Runge-Kutta family of ODE solvers that it calculates fourth- and fifth-order accurate solutions. The same algorithm has been used in Matlab command `ODE45`.

In NetLand, the minimum step size is set 0.01. The maximal step size is set 1,000. The allowed absolute and relative scalar errors are both 0.000001.

SDE solver

SDEs are used to model many situations including population dynamics, protein kinetics, turbulence, finance, and engineering [64]. Knowing the solution of the SDE in question leads to an interesting analysis of the trajectories. Most SDE are unsolvable analytically, and other methods must be used to analyze properties of the stochastic process. From the SDE, a partial differential equation can be derived to give information on the probability transition function of the stochastic process. Knowing the transition function provides information on the equilibrium distribution (if one exists), and convergence to the equilibrium distribution.

In the stochastic model, the changes of species' concentrations are described by Langevin equations which are SDEs in Itô scheme (Eq. 4.3).

$$\frac{dx_t}{dt} = f(x_t) + g(x_t)\frac{dW_t}{dt} \quad (4.3)$$

where W_t is the Wiener process, as used in the stochastic integral, is characterized by its first two moments, $E[W_t - W_s] = 0$ for all $t, s \geq 0$ and $E[(W_t - W_s)^2] = (t - s)$. For the assumption that $g(x_t)$ is constant means that there is no feedback in the noise to the system. To simplify the calculation, we assume $g(x_t)$ is constant.

They are numerically solved by the Euler-Maruyama method implemented in the package “libSDE” [120]. In Itô calculus, the Euler-Maruyama method is a method for the approximate numerical solution of an SDE. It is a straightforward generalization of the Euler method for ordinary differential equations to stochastic differential equations.

Stochastic simulation algorithm (SSA)

The Gillespie stochastic algorithm is used to simulate the stochastic processes which are an algorithm for modeling the kinetics of a set of coupled chemical reactions, taking into account stochastic effects from low copy numbers of the chemical species.

In Gillespie’s approach, the chemical reaction kinetics are modeled as a Markov process in which reactions occur at specific instants of time defining intervals that are Poisson-distributed, with a mean reaction time interval that is recomputed after each chemical reaction occurs. For each chemical reaction interval, a specific chemical reaction occurs, randomly selected from the set of all possible reactions with a weight given by the individual reaction rates. NetLand integrates the Java implementation of the Gillespie algorithm in Dizzy [121].

The steps to run the Gillespie algorithm is as follows:

- Step 1** Set the number of molecules and reactions constants in the system.
- Step 2** Generate random numbers to determine the next reaction to occur as well as the time interval. The probability of a given reaction to be chosen is proportional to the number of substrate molecules.
- Step 3** Update the time and the number of molecules according to Step 2.

Step 4 Go back to Step 2. Stop the simulation when the number of reactants is zero or the simulation time has been exceeded.

4.3 Waddington’s epigenetic landscape construction

How to construct a probabilistic landscape from the dynamic model is discussed in this section. More importantly, the visualization of a landscape generated from high-dimensional gene expression data is crucial in its application. The implementation of a dimensionality reduction method is introduced.

4.3.1 Constructing a probabilistic landscape

NetLand provides a probabilistic implementation of Waddington’s epigenetic landscape based on the assumption that the probability of a gene expression state determines the cellular stability [65], [66]. The quasi-potential (altitude) of the landscape is defined as

$$U = -\log(p), \quad (4.4)$$

$$\frac{\partial p(X, t)}{\partial t} = \sum_{i=1}^N -\frac{\partial(f(X)p(X, t))}{\partial X_i} - \sum_{i=1}^N \sum_{j=1}^N \frac{\partial(g(X)p(X, t))}{\partial X_i X_j}. \quad (4.5)$$

Eq. 4.5 is the Fokker–Planck equation which is a partial differential equation that describes the time evolution of the probability density function of the velocity of a particle under the influence of drag forces and random forces, as in Brownian motion. It is used here to describe the probability of a gene expression state X to calculate its quasi-potential U .

NetLand adopted the self-consistent mean field approximation (SCMF) method in [65], [66] to solve the Fokker-Planck equation. The probability of individual genes is assumed independent and follows a Gaussian distribution,

and the probability of a cell state is the products of individual ones. For the multi-dimensional system, there may be several probability distributions localized at every attractor, but with different variations. Therefore, the total probability is the weighted sum of all these probability distributions. The weighting factors are the size of the basin, representing the relative size of the different basin of attraction. Once we have the total probability, we can construct the potential landscape.

4.3.2 Visualizing a probabilistic landscape

To visualize the landscape of a network with more than two genes, NetLand provides two options. The first option is to select two marker genes whose expression levels will be set as the x-axis and y-axis in the landscape (Figure 4.3(a)). However, the two marker genes may not represent the global dynamics in many cases.

To avoid the loss of information, NetLand offers the second option. A probabilistic dimensionality reduction method, GPDM is employed to project high dimensional gene expression time-series data to a low dimensional latent space with a dynamical model [112], [113], [122]. GPDM is a latent variable model for nonlinear time series analysis. It has been applied to learning models of human pose and motion from high-dimensional motion capture data. It comprises a low dimensional latent space with associated dynamics, as well as a map from the latent space to an observation space. In this case, the x-axis and y-axis are two constructed components (Figure 4.3(b)). For both options, the z-axis is the quasi-potential U . This method uses all genes to provide a global view of the network dynamics and thereby making the landscape plot more comprehensive.

4.3.3 Result rendering

NetLand provides an interactive interface for the visualization and analysis of Waddington's epigenetic landscape. The output of NetLand contains

(a)

(b)

Figure 4.3: The two visualization methods of the landscape in NetLand (a) by selecting two marker genes and (b) by mapping to a 2D latent space.

three sections, i.e., the plotted landscape, four function buttons and a data grid. Figure 4.4 shows an example of the constructed probabilistic landscape. The quasi-potential value and coordinates of a point in the landscape can be extracted by double click the cells in the data grid. Users can also plot the contour map or draw a mesh of points in the landscape using the control buttons. The attractors can be retrieved by the “Show attractors” function.

4.4 Case studies

In this section, we provide a list of case studies to illustrate the application of NetLand in the area of stem cell differentiation and reprogramming and cancer development. These case studies show how our software has successfully analyzed and visualized the systems, and inspire researchers to find ideas in

(a)

(b)

(c)

Figure 4.4: The three components in the output of NetLand. (a) The constructed landscape mapping to two marker genes (left) and mapping to a 2D latent space (right). (b) Control buttons. (c) The table of the grid data in the landscape colored according to the value of quasi-potentials.

the study of network kinetics using Waddington's epigenetic landscape.

4.4.1 A classic two-gene network regulating cell fate

A cell as a dynamical system is intuitively governed by genetic interaction networks that regulate cell fate. A simple circuit of two genes a and b that inhibit each other, forming a double-negative feedback loop structure works as a toggle switch with two stable steady states: one state with high b and low a expression, and the other state with high a and low b expression [7]. Such "bistable" switches formed by mutual antagonism of a pair of critical regulatory genes underlie many binary cell fate choices. When adding positive auto-regulations to the network, a new attractor appears with high

expressions of both nodes. The circuit can be described as a two-variable dynamical system, with the rate of change in expression of each of the two genes given as a function of their expression levels:

$$\frac{da}{dt} = \max_a * (\text{BasalExpression}_a + \quad (4.6)$$

$$\begin{aligned} & (\text{BasalExpression}_{a_0} + I_{a_0} * \frac{(a/K_{a_0})^{N_{a_0}}}{1 + (a/K_{a_0})^{N_{a_0}}}) + \\ & (\text{BasalExpression}_{a_1} + I_{a_1} * (\frac{1}{1 + (b/K_{a_1})^{N_{a_1}}})) - \text{deg}_a * a \\ & \frac{db}{dt} = \max_b * (\text{BasalExpression}_b + \quad (4.7) \end{aligned}$$

$$\begin{aligned} & (\text{BasalExpression}_{b_0} + I_{b_0} * \frac{(b/K_{b_0})^{N_{b_0}}}{1 + (b/K_{b_0})^{N_{b_0}}}) + \\ & (\text{BasalExpression}_{b_1} + I_{b_1} * (\frac{1}{1 + (a/K_{b_1})^{N_{b_1}}})) - \text{deg}_b * b \end{aligned}$$

Here we used NetLand to construct the landscape of this model (Figure 4.5(a)). The probabilistic landscapes mapping to two marker genes and a 2D latent space are shown in Figure 4.5(b-c). The system has three attractors listed in Figure 4.5(d). Then we removed the self-loops from the network by editing the reactions (Figure S1). After altering the structure of the deterministic network, instead of three attractors, only two attractors were identified (Figure S2).

NetLand creates a quantitative map of Waddington's epigenetic landscape for the underlying cell fate choice. It provides mechanistic insights into the "forces" that direct cellular differentiation in the context of physiological development, as well as during artificially induced cell lineage reprogramming. Rigorous quantification of the gene regulatory circuits that govern cell lineage choice and subsequent mapping of the epigenetic landscape can potentially help identify optimal routes of cell fate reprogramming.

Figure 4.5: (a) A two-gene network with mutual inhibitions and self-loops. The probabilistic landscapes of the two-gene network when (b) mapping to two marker genes and (c) a 2D latent space. (d) The three attractors.

4.4.2 Studies on stem cell state transitions

The seesaw model

To study the stem cell differentiation and reprogramming process, a computational model, named ‘seesaw model’ with four nodes (Figure 4.6) was constructed in [18]. The pluripotency module consists of self-activations of Oct4 and Sox2. Mutual antagonism between the ME and ECT lineages are

the components of the differentiation module. The seesaw model focuses on reprogramming using lineage specifiers as substitutions of Yamanaka’s pluripotency factors. In agreement with the two-gene model introduced above, mutually exclusive lineage specifiers facilitate reprogramming to reach the pluripotency state when the forces are balanced otherwise drive the cell to a differentiated fate.

Figure 4.6: The topology structure of the seesaw model with 4 genes and 16 interactions.

To understand the implications of the reprogramming landscape, NetLand was applied to simulate and analyze the model. The regulatory interactions between genes were extracted into a TSV file. After loading it to NetLand, the differential equations including the parameters were modified accordingly. Figure 4.7 shows the Waddington’s epigenetic landscape constructed by NetLand. Three attractors are identified representing the stem cell state and two lineages respectively.

Then a parameter sensitivity analysis was performed to test the system dynamics. We set the parameter KM to 0, which is the concentration of Klf4 and cMyc, then the cells cannot maintain their pluripotency and start to differentiate. NetLand detected three attractors in this system including two states representing ME (#1) and ECT (#3) states respectively and a rare state (#2) (Figure 4.8(b)).

Figure 4.7: The probabilistic landscapes of the seesaw model when (a) mapping to two nodes, MEs and ECTs and (b) the 2D latent space using GPDM. (c) The table of attractors.

To simulate the stem cell reprogramming experiment with two strategies, i.e., using Yamanaka’s factors and using lineage specifiers, we set the parameter KM to 0.4. Balanced equilibrium of the lineage-specifying forces facilitates the induction of pluripotency. At the dynamically stable point, the pluripotency module could be activated, wherein the ME and ECT lineage fates are blocked by Sox2 and Oct4, respectively, and the pluripotency state is maintained.

Cocktail 1: Using Yamanaka’s pluripotency factors

With the replenishment of Yamanaka’s pluripotency factors (Oct4, Sox2, Klf4, cMyc), the somatic cells will be reprogrammed into iPSCs (similar

Figure 4.8: (a) The probabilistic landscapes of the modified network when mapping to MEs and ECTs. (b) The table of attractors.

to stem cells) [1]. We tried to repeat the reprogramming experiment using NetLand following the steps below:

Step1. Set the initial state as ME state.

Step2. Increase the base production rate of Sox2 and Oct4.

Step3. Simulate ten trajectories using the Gillespie algorithm.

Step4. Calculate attractors from the simulated data. It shows that all ME cells turned into stem cell state with Oct4 and Sox2 highly expressed (Figure S3).

Cocktail 2: Using lineage specifiers

The somatic cells can be reprogrammed into iPSCs (similar to stem cells) using lineage specifiers [18]. In this reprogramming experiment, the lineage specifiers, MEs, and ECTs worked as the substitutions of Oct4 and Sox2. Then we followed the steps stated in the above experiment. We got qualitatively similar result that all ME cells are reprogrammed into stem cells in which Oct4 and Sox2 are highly expressed.

A dynamic model from BioModels

A dynamical model of lineage determination based upon a minimal circuit which contains the Oct4/Sox2/Nanog core network as well as other vital

genes are discussed in [75] (BIOMD0000000209.xml from BioModels). A self-organized core network of embryonic stem cells, consisting of Oct4, Sox2, and Nanog, maintains pluripotency and self-renewal functioning as a bistable switch. The mutual antagonistic interaction pairs Cdx2-Oct4 and Gata6-Nanog give rise to the endoderm lineage.

Assembling the entire network shown in Figure S4(a), we obtain equations for the gene concentrations. We probe this system about system dynamics using NetLand. The probabilistic landscapes mapping to two marker genes and a 2D latent space are shown in Figure S4(b)-(c). The system has two attractors listed in Figure S4(d) representing endoderm and stem cell fates. In particular, we studied the transitions of gene expression profiles as functions of the Oct4 concentration under the induction of an external factor $Signal_A$.

For low values of $Signal_A$, the external signal activating Oct4, the stem cell is switched off, and Cdx2 reach high levels (Figure 4.9(a)). For $Signal_A$ within an intermediate range, the core triplet genes in stem cells, i.e., Oct4, Sox2, and Nanog, switch on. This region is bistable (Figure 4.9(b)-(c)) that either Gata6 and Cdx2 are at higher levels, or the system reaches the stem cell state. For yet higher values of $Signal_A$, Oct4 is highly activated to induce Gata6 and simultaneously suppress Cdx2. At the same time, Nanog is shut down by the over-expression of Gata6, in turn, weakening the positive feedback to Oct4 and Sox2, which is therefore unable to maintain Sox2. Ultimately only Oct4, Gata6 and Genf are on (Figure 4.9(d)).

Therefore, the dynamics of the network consisting of the master TFs in stem cells and lineages suggests three stable attractors, corresponding to the three lineages. The external factors determine which state the system will stabilize.

4.4.3 Cancer development

Cancer presents a serious threat to human health regulated by the underlying gene networks. The understanding of the emergence of cancer states as

Figure 4.9: Time series concentrations in terms of the concentration levels of $Signal_A$ indicating the final steady state values. (a) $Signal_A = 1$, the trophectoderm is the only possible state of the system. (b-c) $Signal_A = 10$, the bistable system gives either the embryonic or differentiated stem cell (endoderm) lineage. (d) $Signal_A = 25$, the endoderm lineage is the only possible state.

well as the transformation remains challenging in current cancer biology. A simplified gene regulatory network of cancer and development including 6 genes and 16 edges is constructed in [123] (Figure 4.10).

Figure 4.10: The topology structure of the cancer model.

The core network contains two classes of major marker genes: two cancer marker genes (MDM2 and P53) and two stem cell markers (ZEB and OCT4). The interactions within the network determine the stable states of the system. From the topology of the network, there are some mutual inhibitions between pairs of genes with diverse functions, such as miR-200 and Zeb, miR-145 and Zeb, and miR-145 and Oct4. P53, the primary tumor repressor, represses Zeb and Oct4 through two miRNAs, miR-200 and miR-145, whereas the activation of Zeb promotes the epithelial-mesenchymal transition (EMT), which gives rise to the formation of cancer stem cells (CSCs). In addition, its expression is also under the control of oncogene Mdm2.

We used NetLand to develop a potential global landscape to quantify cancer and associated processes. Stem cell, CSC, cancer, and normal cell types emerge as basins of attraction on the associated landscape mapping to P53 and Zeb (Figure 4.11(a)). However, when the landscape is mapped to a 2D latent space, a complete depiction of the system dynamics in high-dimensional spaces is visualized (Figure 4.11(b)). Additional attractors are identified representing different types of intermediate cells (Figure 4.11(c)). This landscape provides a quantitative way to understand the underlying mechanisms of CSC formation and interplay between cancer and development.

Figure 4.11: The probabilistic landscapes of the network when (a) mapping to two marker genes (P53 and Zeb) and (b) a 2D latent space. (c) Identified attractors.

4.5 Summary

The famous drawing of Waddington’s epigenetic landscape depicts how a cell progresses from an undifferentiated state to one of some discrete, distinct, differentiated cell fates during development. Inspired by dynamical systems theory, it has been applied for a comprehensive understanding of the determination of cellular phenotypes (e.g., cell fate decision in differentiation and reprogramming). Each dimension of the landscape corresponds to a physical quantity, a ‘representative’ output of the regulatory system, for example, the concentration of a critical TF. But a complete depiction of a cell’s state with high-dimensional spaces is difficult to visualize. To address these issues, we present NetLand, an open-source software tool for modeling and simulating the kinetic dynamics of a biological system, and visualizing the corresponding Waddington’s epigenetic landscape in three dimensions without restriction on the number of genes.

We have applied NetLand on three case studies illustrating the great potential of NetLand in the study of cell fate transitions. With quantita-

tive modeling of Waddington’s epigenetic landscape, the systems dynamics of transcriptional regulation driving cellular phenotypic changes are easily analyzed and predicted. Therefore, by automating the modeling and visualization of Waddington’s landscape, NetLand can facilitate the studies of cell fate conversion and help design innovations in cell fate engineering and regenerative medicine.

However, the knowledge-based mathematical modeling method is often time-consuming and only applicable to a finely crafted network. In the next chapter, a data-driven model is proposed to infer a Waddington’s epigenetic landscape from single-cell sequencing data using Hopfield network. It interprets the transcriptome dynamics from empirical data, such as the recovery of the pseudo-times. Therefore, it can help us understand the transition of gene expression profiles.

Chapter 5

Data-driven approach in modelling stem cell fate determination

This chapter introduces a data-driven method, named HopLand, to model stem cell fate determination using continuous Hopfield network (CHN) to map cells in a Waddington’s epigenetic landscape. The combinatorial regulatory interactions of genes construct a kinetic model learned from single-cell sequencing data. Through the analysis of such a model, essential genes and regulatory interactions are identified which drive the transition of cell states. Therefore, HopLand has the potential to generate fundamental insights into cell fate regulation.

HopLand is applied for the estimation of pseudotimes of individual cells in different types of single-cell transcriptome data. Pseudotime is a latent dimension which measures the cells’ progress through the transition. Pseudotime is related to but not necessarily the same as laboratory capture time. The recovery of pseudotime increases the temporal resolution of single-cell transcriptional data to help understand the transition of gene expression profiles. It achieved high accuracies of pseudotime prediction compared to existing methods.

5.1 Workflow of HopLand

Motivated by the famous Waddington's epigenetic landscape metaphor in developmental biology, different proposals are made to realize this metaphor in a rationalized way. Data-driven approaches using single-cell sequencing are at the cutting edge of single-cell biology which can reveal the subtle changes that make each one unique. The virtual individual cells modeled based on the single-cell gene expression data are to be placed on the surface regions of the landscape corresponding to their developmental stages.

To plot such a landscape, we constructed a kinetic model from the sequencing data, using neural networks to simulate the transcriptional regulation. CHN is a type of recurrent neural network characterized by memory recall. Its topology structure is similar to a transcriptional network, and its dynamic feature is corresponding to the attractor system as Waddington's epigenetic landscape (Figure 5.1). For each target gene, the model associates its change rate with the adaptation of the neurons.

Figure 5.1: (a) The topology and (b) The dynamic feature of Hopfield network.

Based on the above framework, the HopLand algorithm is designed as follows:

Step 1. Normalize the gene expression data using z-score and select

differentially expressed genes (e.g., by filtering out genes with low variances).

Step 2. Construct the kinetic model by neural network inference from data.

Step 3. Construct the Waddington’s epigenetic landscape based on the kinetic model.

Algorithm 2 illustrates the steps of HopLand which are further elaborated in the following subsections.

Algorithm 2 HopLand algorithm

INPUT: Single-cell gene expression data $D \in R^{S \times N}$ where S is the number of cells and N is the number of genes, and temporal information *cellStages* (which is not compulsory)

OUTPUT: Kinetic model of Waddington’s epigenetic landscape *landModel*

```

1: if cellStages is available then
2:    $\theta = \text{PARAMETEROPTIMIZATION}(D, \text{cellStates})$ 
3: else
4:    $\theta = \text{INITIALIZEPARAM}(D)$ 
5: end if
6: landModel =  $\text{LANDSCAPECONSTRUCT}(D, \theta)$ 
7: return landModel

8: procedure  $\text{PARAMETEROPTIMIZATION}(D, \text{cellStates}, \text{realTraj}, w,$ 
   gmmModels)
9:   Algorithm 3
10: end procedure

11: procedure  $\text{LANDSCAPECONSTRUCT}(D, \theta)$ 
12:   Algorithm 4
13: end procedure

```

5.2 Data-driven kinetic modelling

This section outlines the modeling process in HopLand. Section 5.2.1 introduces the formulation of the change rate for each gene in a CHN and how gene-gene interactions can be encoded in ODEs. The gradient descent algorithm used in the estimation of parameters of the model is introduced

in Section 5.2.2.

5.2.1 Network formulation

The discrete HN has been used to study biological systems with each neuron representing a gene [11]–[13]. However, gene expression values tend to have continuous input-output relations which cannot be fully characterized by the simplified discrete state of neurons in the two-state HN. CHN uses continuous variables and predicts continuous responses. Thus the framework of CHN is adopted to model the system dynamics with each neuron corresponding to an individual gene whose adaptation indicates the change of gene expression value.

In the framework of CHN, the gene expression of a cell is characterized by the output of the neurons $\mathbf{V} = \{V_i, i = 1, 2, \dots, N\}$, where N is the number of genes. The inputs of each neuron come from two sources, the background noise and inputs from other neurons. The time evolution of the system is represented by ODEs. The change rate of neuron i is modeled by

$$\frac{dV_i}{dt} = C_i \sum_{j=1}^N W_{ij} U_j - \delta_i V_i + I_i, \quad (5.1)$$

$$U_j = g_j(V_j), \quad (5.2)$$

where W_{ij} is an entry of the weight matrix of CHN representing the interconnection weight coefficient from neuron j to neuron i , and C_i is an amplifier on the synaptic connections. The external input I_i represents a combination of propagation delays, regulations by other genes not in our model, and noise in transcriptional regulation. δ_i denotes the degradation rate of gene i . The activation function $g_j(V_j)$ represents the input-output relationship of a nonlinear amplifier with negligible response time. The activation function is required to be a monotonically increasing function to make the system stable [101]. In our model, a sigmoid activation function is used (Eq. 5.3) which has been used in [124], [125] to describe the regulatory function of a

gene:

$$g_j(V_j) = 1/(1 + e^{-(V_j - \mu_j)/\sigma_j}). \quad (5.3)$$

In Eq. 5.3, μ_j and σ_j are the mean and standard derivation of the expression levels of the j -th gene in all cells.

5.2.2 Parameter estimation

There are several parameters in the ODE model of kinetics in Eq. 5.1. To infer these parameters $\theta = \{\delta_i, I_i, C_i, W_{ij}, i, j = 1, 2, \dots, N\}$ from the sequencing data, an optimization method (Algorithm 3) which fits the simulated and observed single-cell data is proposed based on the premise that a realistic model should be able to generate simulated data consistent with the real data.

The consistency between experimental data and simulated data is measured in two aspects. First, the gene expression values $D_i = \{D_{it}, t = t_1, t_2, \dots, t_T\}$ where T is the number of time points (or cell stages) in the single-cell data, should follow a similar distribution. Normally it is believed to follow the Gaussian mixture distribution with the mean values of components as the representative gene expression values in different lineages [15], [126] which are calculated by *Function'fitMixtureGaussian'* in *AlgorithmS1*. The second aspect of consistency lies in the change of a single gene, e.g. down-regulated or up-regulated, along the time evolution of cell states. Thus the objective functions are defined as follows:

$$OBJ_1 = \sum_{i=1}^N (DF_{data}^i - DF_{simulate}^i(\theta)), \quad (5.4)$$

$$OBJ_2 = \frac{1}{mN} \sum_{t=t_1}^{t_m} \sum_{i=1}^N \sigma_{it} (D_{it} - S_{it}(\theta))^2, \quad (5.5)$$

where DF_{data}^i and $DF_{simulate}^i$ are the density functions for the observed and simulated expression levels of the i th gene. σ_{it} is the standard derivation of the expression values of gene i at the time point (or cell developmental

stage) t . $S_i = \{S_{it_1}, S_{it_2}, \dots, S_{it_m}\}$ is an average of the simulated trajectories with m time points by simulating the CHN of Eq. 5.1 using the generated initial states.

The gradient descent learning algorithm [127] is used to optimize the parameters in the CHN. The update of a parameter value at k -th iteration is defined as

$$\Delta\theta^{(k)} = -\eta\left(\frac{\partial OBJ_2}{\partial\theta^{(k)}}\right), \quad (5.6)$$

$$\theta^{(k+1)} = \theta^{(k)} + \Delta\theta^{(k)}, \quad (5.7)$$

where η is the learning rate between 0 and 1 which controls the rate of parameter adjustment. We also iteratively adapt the learning rate according to the Bold Driver technique [128]. The weight matrix is initialized as the Pearson correlation coefficients between samples. To simulate the dynamic trajectories, we used the Euler's method (the first order Runge-Kutta) to solve the ODEs with the initial states generated near the given starting points (Refer to *Fucntion'generateRandomInitialStates'* in *AlgorithmS1*). In each iteration of the gradient descent learning, we calculated the value of Eq. 5.4 using the current parameters. In the end, the optimized parameters are selected with the minimum sum of Eq. 5.4 and Eq. 5.5.

5.3 Construction of the landscape

This section introduces the definition and visualization of the Waddington's epigenetic landscape constructed in HopLand. The landscape is created based on the model constructed in the above section. Compared to the quasi-potential landscape in Section 3.2.1, it uses a similar Gaussian process based dimensionality reduction method in NetLand as an approach to the visualization of high dimensional data. The altitude of the landscape is defined according to the network energy of HN.

Algorithm 3 Parameter optimization

INPUT: Single-cell gene expression data D , temporal information $cellStages$ **OUTPUT:** Optimized parameters $\theta = \{\delta_i, I_i, C_i, W_{ij}, i, j = 1, 2, \dots, N\}$ **Initialization:** Set $\delta_i = 1$, $I_i = 0$, $C_i = 1$, $W_{ij} = corr(D)$, $\eta = 0.3$, $maxIts = 2000$.Set $startPoints$ as the earliest samples in $cellStages$;Random initial states $randomInitials = \text{GENERATERANDOMINITIALSTATES}(startPoints)$;Coefficient matrix $\sigma = \text{GENERATETRAJECTORY}(D, cellStates)$ Gaussian mixture models $gmmModels = \text{FITMIXTUREGAUSSIAN}(D)$ **Iteration:**

```

1: for  $k = 1, 2, \dots, maxIts$  do
2:    $\Delta\theta^{(k)} = -\eta(\frac{\partial OBJ_2(\theta^{(k)})}{\partial \theta^{(k)}})$ ;
3:    $\theta^{(k+1)} = \theta^{(k)} + \Delta\theta^{(k)}$ ;
4:   if  $OBJ_2(\theta^{(k)}) > OBJ_2(\theta^{(k+1)})$  then
5:      $\eta = \eta * 1.2$ ;
6:   else
7:      $\eta = \eta * 0.5$ ;
8:   end if
9: end for
10:  $k^* = \underset{k}{\operatorname{argmin}} OBJ_1(\theta^{(k)}) + OBJ_2(\theta^{(k)})$ ;
11: return  $\theta = \theta^{(k^*)}$ .
```

5.3.1 Energy function

Under certain conditions, the activation values of the units in a CHN undergo a relaxation process such that the network will converge to a stable state in which these activation values will not change anymore. These conditions include that the weight matrix $\mathbf{W} = \{W_{ij}, i, j = 1, 2 \dots N\}$ has to be symmetric, and the activation functions must be continuous, bounded, and strictly monotonically increasing, such as a sigmoid function (Eq. 5.3). The behavior of a CHN system can be described with an energy function (Eq. 5.8) which is a Lyapunov function:

$$E = -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N W_{ij} U_i U_j + \sum_{i=1}^N I_i U_i + \sum_{i=1}^N \delta_i \int_0^{U_i} g_i^{-1}(u) du. \quad (5.8)$$

Based on the essential idea of Waddington's landscape, the differentiation processes follow the paths of losing potential energy determined by the topography of the landscape. Compared to the entry point, the valleys possess relatively lower potentials representing more stable states. Here we use the network energy in Eq. 5.8 to quantify the altitude of the landscape, similar to previous works [11]–[13].

5.3.2 Visualizing the landscape

To visualize the cellular dynamics in a landscape whose shape is determined by the energy function in Eq. 5.8, we project the high-dimensional data to a two-dimensional latent space as the XY-plane in the landscape. The non-linear dimensionality reduction method, named Gaussian process latent variable model (GP-LVM), is used to generate the mapping between the original space and the latent space [112], [113]. GP-LVM is a probabilistic approach to modeling high-dimensional time-series data in a low dimensional latent space with a dynamical model. It has been used for computer vision, biological data analysis, etc. We have used this method to process gene

expression time-series data in the visualization of landscape [129].

Using GP-LVM to reduce the dimensions of the observed data to two dimensions, we create a mapping from the original space to the latent space. Then we generate a 2D grid in the reduced space covering all the cells to plot a continuous surface. Since the energy is calculated in the high-dimensional space, we project the grid points back to their original space. The landscape is plotted on the grid data. The pseudocode of the landscape construction method is shown in Algorithm 4.

Algorithm 4 Landscape construction

INPUT: Single-cell gene expression data D with S samples and N genes, parameter vector θ

OUTPUT: A landscape model $landModel$

- 1: Generate mapping $X = GPLVM(D)$, where latent variables are encoded in matrix $X \in R^{S \times 2}$;
 - 2: Define a 2D grid $Grid = [\min(X^{1\cdot}) - \varepsilon, \max(X^{1\cdot}) + \varepsilon]' \times [\min(X^{2\cdot}) - \varepsilon, \max(X^{2\cdot}) + \varepsilon]$, where $X^{1\cdot}$ and $X^{2\cdot}$ are the first and second components of samples, and ε is a small positive constant which determines the size of margins around the observed data in the latent space;
 - 3: Perform inverse dimensionality reduction $Y = GPLVM^{-1}(Grid)$, where $Y \in R^{S^{Grid} \times N}$, S^{Grid} is the number of points in $Grid$;
 - 4: Calculate the *energy* according to Eq. 5.8;
 - 5: $landModel = \{X, Grid, energy\}$;
 - 6: **return** $landModel$.
-

5.4 Pseudotime recovery

The heterogeneities among cells from single-cell datasets show the gene expression dynamics during the gradual transition of their transcriptomes. The estimation of the pseudotimes of the single cells is to construct a pseudo-temporal path to order cells. Then, the pseudotime can be estimated by calculating the geodesic distance between every two cells in the landscape.

In Waddington's epigenetic landscape, a single cell with specific gene expression pattern is simplified as a point, hence the time evolution of cell states is defined as the state-transition movement on the landscape which is determined by the topography of the landscape surface. Thus the geodesic

distances between two cells can be calculated from the coordinates of the cells on the landscape. The geodesic distances to the starting point are considered proportional to the pseudotimes, setting the pseudotime of the starting point to zero. The pseudotime of the i th cell is defined as the geodesic distance to the starting point.

We used the fast marching algorithm [130], which is a numerical method for solving boundary value problems of the Eikonal equation, to perform geodesic extraction on a triangulated mesh generated from the landscape surface.

By extracting the shortest path between every two points on the surface, a minimum spanning tree (MST) is constructed to connect all cells with the given starting point as the root. An MST is a subset of the edges of a connected, edge-weighted undirected graph that connects all the vertices, without any cycles and with the minimum possible total edge weight. That is, it is a spanning tree whose sum of edge weights is as small as possible. More generally, Due to the sparsity of single-cell sequencing data, the connections between consecutive states may lack sample data to link them. In such an undirected graph, a minimum spanning forest exists which is a union of the minimum spanning trees for its connected components. Kruskal's algorithm is commonly used to find the forest in a graph.

The pseudocode of the pseudotime estimation is shown in Algorithm 5. Considering the significant differences among the three coordinates in the landscape, a.k.a. the network energy and x,y- two coordinates, we normalized the three values to the same scale in the range $[0,1]$. Then a triangulated mesh was created to perform fast marching on the landscape surface to generate a distance matrix. The MST was built based on this distance matrix and the given start point.

5.4.1 Single-cell sequencing datasets

In this section, the testing datasets using different sequencing protocols are introduced. Totally 11 testing datasets are included, a qPCR dataset

Algorithm 5 Pseudotime recovery**INPUT:** *landModel*, *StartPoint***OUTPUT:** pseudotimes of cells *PT*

- 1: *manifold* = *Normalize*(*landModel*)
- 2: [*Vetices*, *Faces*] = *delaunayTriangulation*(*manifold*);
- 3: *distMatrix* = *fast_marching*(*Vetices*, *Faces*);
- 4: *T* = *minimumSpanningTree*(*distMatrix*, *StartPoint*);
- 5: *PT* = *calculateDistance*(*T*, *StartPoint*);
- 6: **return** *PT*.

“GUO2010” [131], 5 synthetic datasets and 5 scRNA-seq datasets, i.e. “DENG2014” using Smart-seq2 [132], “YAN2013” [133] using scRNA-seq method demonstrated in [85], “ES_MEF” using STRT [134], “LPS” [135] and “HSMM” [136]

qPCR

As introduced in Section 2.4.1, quantitative polymerase chain reaction (qPCR) is a directed way of measuring the expression of previously known genes in down to singular cells. It combines PCR amplification and detection into a single step.

We extracted the dataset “GUO2010” from a mouse embryonic single-cell qPCR experiment [131]. It contains 438 cells, captured the differentiation process from zygote to 64-cell state. In this experiment, 48 genes are sequenced for each cell (Table 5.1).

Table 5.1: Numbers of embryos and cells in “GUO2010”.

Stage	No. of single cells
zygote	9
2-cell	19
4-cell	23
8-cell	44
16-cell	75
32-cell	109
64-cell	159
Total	438

Bifurcation events happened at the 32-cell stage and the 64-cell stage.

Two lineages, i.e., inner cell mass (ICM) and trophectoderm (TE), are generated in the 32-cell stage when the blastocyst starts to grow from the morula. ICM consists of cells from the inside of the morula, while the others located on the outside form TE. ICM cells remain totipotent that they can differentiate into any of the cells in the embryo. In differentiation, they will further differentiate into the epiblast (EPI) and the extraembryonic primitive endoderm (PE) according to the concentration of fibroblast growth factor-4 (FGF4). Thus, cells are differentiated into three different cell states in the 64-cell stage: TE, EPI, and PE.

Synthetic data

Besides of the experimental qPCR dataset, five differentiation patterns of cells were simulated using the code from [137] based on the qPCR dataset. The simulation was done by simulating a branching process in two dimensions with increasing probability of splitting in a cell stage along the timeline which ensures a random pattern of splitting. Gene expression profiles are simulated by sampling from independent Gaussian processes with differing priors for 48 different genes. The priors for the mapping include a linear mapping and both long-term and short-term non-linear trends to simulate the complexity of biological processes. This makes the simulations more complex than any of the models we applied to recover the data.

scRNA-Seq

Genome-wide single-cell mRNA sequencing (scRNA-seq) profiles with high temporal and spatial resolution offer possibilities to characterize cellular states and understand regulatory mechanisms among heterogeneous cell populations. A variety of methods has been recently developed making it possible to reveal new biological and medical insights.

For the dataset published by Yan et al. [133], the transcriptional profiles of 124 individual cells covering 22,687 maternally expressed genes from human preimplantation embryos and human embryonic stem cells (hESCs) at

seven crucial stages (Table 5.2) were screened using RNA-Seq analysis. The sequenced embryos were picked at different passages of pre-implantation: metaphase II oocyte, zygote, 2-cell, 4-cell, 8-cell, morula and late blastocyst at the hatching stage. Approximately 85% genes with low variances were filtered out, and we normalized the rest data using zscore.

Table 5.2: Numbers of embryos and cells in “YAN2013”.

Stage	No. of single cells
Oocyte	3
Zygote	3
2-cell	6
4-cell	12
8-cell	20
Morula	16
Late blastocyst	30
hESC	34
Total	124

The dataset “DENG2014” [132] covers the same time span in early mouse embryo development as the qPCR dataset analyzed “GUO2010”, a.k.a. from zygote to the late blastocyst stage. In addition, two adult cell types are included in the dataset, fibroblasts and liver cells. The RNA-seq experiments detected a total of 22,958 genes in 317 single cells (Table 5.3), but many genes were expressed at a low level and subject to considerable technical variation. Thus, a subset of 1,000 genes are extracted that were likely to be discriminative and expressed (> 1 reads per kilobase of transcript per million reads mapped) in at least 30% of the cells [138]. This filtered RNA-seq gene signature is further normalized.

The “LPS” dataset consists of single-cell RNA-seq samples collected after stimulating bone-marrow-derived dendritic cells by lipopolysaccharide (LPS) [135]. A total of 306 cells collected at 1, 2, 4 and 6h after the stimulation were used for our analysis. We selected a subset of the data with two classes, i.e. unstimulated cells and treated cells at 6h, in order to test the clustering function of HopLand (Table 5.4).

The dataset “HSMM” consists of single-cell RNA-seq samples from dif-

Table 5.3: Numbers of embryos and cells in “DENG2014”.

Stage	No. of single cells
zygote	4
early 2-cell	8
middle 2-cell	20
late 2-cell	10
4-cell	14
8-cell	47
16-cell	58
early blastocyst	43
middle blastocyst	60
late blastocyst	30
liver	13
fibroblast	10
Total	317

Table 5.4: Numbers of cells in “LPS”.

Class	No. of single cells
unstimulated	57
treated	74
Total	131

differentiating human skeletal muscle myoblasts (HSMM) under high-mitogen conditions (GM) [136] (Table 5.5). Around 49 and 77 cells are captured at each of the four-time points after the switch to DM using the Fluidigm C1 microfluidic system. Complete gene expression profiles of 271 cells are collected at 0, 24, 48 and 72 h after switching human myoblasts to low serum. This dataset is also used to test the clustering function in HopLand.

Table 5.5: Numbers of cells in “HSMM”.

Class	No. of single cells
T1	69
T2	74
T3	79
T4	49
Total	271

The transcriptomes of 92 single cells of two distinct types, ESC and

mouse embryonic fibroblasts (MEF), in the dataset “ES_MEF” are generated using single-cell tagged reverse transcription (STRT) [134] (Table 5.6).

Table 5.6: Numbers of cells in “ES_MEF”.

Class	No. of single cells
ESC	48
MEF	44
Total	92

5.4.2 Pseudotimes inferred from single-cell sequencing data

We evaluate the performance of HopLand by comparing it with 6 state-of-the-art methods, i.e. Monocle [136], Topslam [137], Wanderlust [139], Wishbone [140], SCUBA [138] and Diffusion map [141].

Synthetic data

We tested HopLand on five synthetic datasets generated by simulating the early development of mouse embryos. Each dataset contains a randomly generated differentiation pattern by angled linear splits in two dimensions [137]. We extracted the pseudotimes of the cells using the HopLand algorithm and compared it with other methods (Figure 5.2).

The singular cells in each synthetic dataset are mapped in the contour plot of the landscape surface according to the extracted pseudotimes (Figure 5.3). The movement directions of the cells following the shape of the landscape could reflect the irreversible transitions of cell states during the differentiation in the embryonic development.

Most of the data contain one diverging event splitting cells into two lineages, except for Synthetic data 3 which contains two bifurcations resulting in four lineages. The cells at early developmental stages (dark red dots) are located in the regions of the landscape with high energy, while the derived lineages (white or light red dots) rest in valleys.

Figure 5.2: Accuracies of pseudotime recovery on 5 synthetic datasets using different pseudotime recovery methods. Pearson correlation coefficient between the predicted and observed times is used to evaluate the result.

Single-cell data on early embryonic development

The single-cell dataset “GUO2010” of mouse preimplantation development contains the expression profiles of 438 cells with 48 genes per cell covering the developmental stage from the 1-cell to 64-cell stages [131]. Two distinct cell lineages, i.e., trophectoderm (TE) and inner cell mass (ICM), emerge from the 16-cell to 32-cell stages. During the transitions from the 32-cell to 64-cell stages, another two cell lineages, i.e., primitive endoderm (PE) and epiblast (EPI), are split from ICM. We applied the HopLand algorithm to this dataset and recovered the pseudotimes of the cells.

Analyzing the expression profiles of individual genes, we found that the expression of marker genes follows a mixture of Gaussian distributions. The mean values in different components indicate differential expression in separate lineages. We inferred the moments of the Gaussian distributions from the data. The cells from the 1-cell stage were used to generate the training data as the initial states for the simulation of Hopfield network to calibrate the model. Then we learned a dynamical model to capture the kinetics of the cell differentiation process of early mouse embryonic development.

After projecting the high-dimensional data to a 2D latent space using

Figure 5.3: The contour plots of the constructed Waddington's epigenetic landscape using the five synthetic datasets (a-e). The contour lines represent heights in the landscape. The dark areas indicate low energy, and the light regions have high energy. The cells from early stages to late stages are colored from dark red to white.

GP-LVM, we calculated the energy values according to Eq. 5.8 which are used for the z-axis of the landscape (Fig. 5.4a). In the contour plot of the landscape (Fig. 5.4b), two bifurcations are shown corresponding to the cell fate decisions made at the 16-cell to 32-cell stages (cyan dots to light blue

Figure 5.4: (a) The Waddington's epigenetic landscape recovered using HopLand. (b) The contour plot of the constructed Waddington's epigenetic landscape. The dots are colored according to the developmental stages of the represented cells.

dots) and the 32-cell to 64-cell stages (light blue dots to dark blue dots). Then, we mapped the expression values of marker genes into the landscape to trace the differentiation process (Fig. 5.5). The expression profiles of the cells in different branches are separated using the marker genes for different cell lineages (e.g., ICM, TE, PE, EPI). The result shows that cells belonging to the same stage are located together in the landscape, and they follow the developmental orders. In the landscape, the change of the network energy shows a decreasing trend along the differentiation process which confirms our premise that cell differentiation is a process with decreasing energy.

Using the topography of Waddington's epigenetic landscape as a correction, we extracted the pseudotime information. The cells were connected in a minimum spanning tree (or a forest of multiple trees) which contains the estimated pseudotimes revealing the transitions of cell states (Fig. 5.6). We identified a small group of cells (in dark red) isolated from others. It contains cells from the 1-cell stage due to the sparsity of data during the early mouse embryonic development.

We also compared HopLand to other methods. The accuracies of pseudotime recovery measured by the correlation coefficient between the predicted and observed pseudotimes are shown in Figure 5.7.

Figure 5.5: Mapping gene expression values to Waddington’s epigenetic landscape using (a) *Fgf4*, (b) *Gata4*, (c) *Cdx2* and (d) *Sox2*. The value decreases from dark red to white.

The weight matrix in the CHN inferred from the dataset of “GUO2010” (Figure S5) contains information about the interactions between genes which can help us find the key regulatory relations and reveal the dynamics of gene expression during cell differentiation. The top 10 significant interactions in the learned weight matrix of the mouse pre-implantation data are listed in Table 5.7. Seven of the top 10 gene pairs are confirmed. Some of them have direct interactions, e.g., transcriptional regulation. Although we have not yet found evidence for the rest of interactions, some genes from these interactions, e.g., *DPPA1*, *HAND1*, are known to be involved in mouse embryonic development. For example, *HAND1* is a transcription factor expressed in extra-embryonic mesoderm and trophoblast. *DPPA1* is developmental

Figure 5.6: The minimum spanning tree constructed from Waddington’s epigenetic landscape. The dots are colored according to the developmental stages of the cells in the dataset of “GUO2010”.

Figure 5.7: Accuracies of pseudotime recovery on the qPCR dataset “GUO2010” using different pseudotime recovery methods. Pearson correlation coefficient between the predicted and observed times is used to evaluate the result.

pluripotency associated.

To further validate the results, we used GeneGo MetaCore, an integrated “knowledge-based” platform for pathway analysis of OMICs data and gene lists, to verify the interactions. 14 out of the top 20 (70%) gene pairs in HopLand’s results can be checked by searching GeneGo MetaCore (Figure 5.8), including some well-known critical interactions involved in embryonic development (Table S1,S2,S3), e.g., the interactions between ESRRB and FGF4.

From the weight matrix, we also ranked genes by the sum of weights of incident edges and identified a few essential regulators, e.g. FGF4, OCT4,

Table 5.7: Top 10 key interactions identified from the weight matrix ranked by the absolute value of the weight in CHN.

Rank	Gene 1	Gene 2	References (PMID)
1	GATA4	LCP1	18555785,22083510,16153702,14990861
2	GATA4	GATA4	15987774
3	ATP12A	DPPA1	-
4	ESRRB	ESRRB	16767105,19136965
5	AQP3	DPPA1	-
6	AQP3	LCP1	18700969,19884255
7	HNF4A	LCP1	21852396,15159395
8	GRHL1	HAND1	-
9	ESRRB	FGF4	26206133
10	KLF4	KLF4	18264089,18358816,19030024,18555785

GATA4 and ESRRB, which have been experimentally tested to be essential for early embryonic development [142], [131], [143], [144], [145]. These key factors play important roles in the regulation of embryonic development, cell proliferation, and cell differentiation.

We also compared the HopLand algorithm with other methods on the monoallelic mouse and human pre-implantation embryo RNA-Seq data, “DENG2014” and “YAN2013”.

The single-cell RNA-seq dataset “DENG2014” comprises transcriptome profiles of 317 cells from zygote to blastocyst and two mature cell types including 11 stages. The constructed landscape is shown in Figure 5.9. Splitting occurs in both 8-cell and 16-cell stages. The two developed cell types, fibroblast and adult liver, are separated from the early embryonic developmental lineages. The blastocyst cells (colored in green, cyan and light blue) are clustered together with lower energy than cells of the early lineages. Early blastocyst cells (colored in green) and middle blastocyst cells (colored in cyan) are mixed but separated from the late-stage cells (colored in light blue) indicating closer developmental relations. The mature cells (colored in dark blue) located in a valley have low energy values.

The result of comparing the accuracies of different methods in estimating pseudotimes are shown in Figure S6. HopLand is not as good as most other

Figure 5.8: The core network found by searching GeneGo MetaCore for interactions overlapping with HopLand’s top 20 interactions and genes. The red and green lines represent inhibition and activation respectively. The volcano icon indicates transcription factors. The orange bullet represents enzyme. The blue curve represents a binding protein. The green anchor represents receptor ligand. The purple X represents channel.

methods for this dataset, partly due to the shortage of time information from the early blastocyst stage to the late blastocyst stage. In addition, due to the lack of specific temporal details between the late blastocyst samples and mature cells, HopLand cannot accurately recover the trajectories from the pre-implantation development to the adult cells. Nevertheless, HopLand can successfully reconstruct the progress from zygote to blastocyst (Figure 5.9) achieving a correlation coefficient of 0.91 with real data.

The landscape of “YAN2013” is shown in Figure 5.10(a). From oocyte to morula, it follows a single lineage. We identified three lineages in the late stage of the blastocyst, i.e., TE, PE and EPI (Figure 5.10(b)-(c)). The development path from a late blastocyst (EPI) to ESC is established. However, a separate group of cells showing a profile of the TE lineage is identified in the late blastocyst stage which is consistent with the clustering result in [146]. It may result from the first bifurcation during the 16-cell stage.

Figure 5.9: The contour plot of the landscape constructed from the dataset of “DENG2014”. The cells are connected in a minimum spanning tree.

Single-cell data without temporal information

For datasets without continuous temporal data, the algorithm HopLand mainly serves as a clustering method omitting the learning process. The dataset “ES_MEF”, “LPS” and “HSMM” contain distinct groups of cells rather than continuous time series expression data. HopLand was applied to analyze the samples.

Another learning scheme, Hebbian learning, is adopted for the construction of Hopfield network. The Hebbian learning rule is commonly used to calculate changes in connection strengths in a neural network, in which a change in the strength of a connection is a function of the pre- and postsynaptic neural activities [147].

Two clusters are identified in the dataset “ES_MEF”, “LPS” and four clusters in “HSMM” (Figure S7). The two types of cells in “ES_MEF” were well separated, and an incline of network energy was revealed from ESCs to MEF cells. However, the rug landscape surfaces of the other two outputs may be due to the insufficient analysis, e.g., the problematic selection of differentially expressed genes and inappropriate filtering of the sequencing data.

Figure 5.10: (a) The constructed landscape of the dataset “YAN2013”. (b) The contour plot of the landscape with labeled lineages. (c) Unsupervised hierarchical clustering of 30 samples from the late blastocyst (including 3 lineages: EPI, PE, TE) and heatmaps of 8 selected differentially expressed genes [146].

5.5 Summary

In this chapter, we proposed a novel method, named HopLand, to recover the pseudotime from single-cell data using CHN based modeling of Waddington’s epigenetic landscape. The order of cells is determined by the geodesic distances in the landscape. Waddington’s epigenetic landscape constructed from the neural network model serves as a stage on which the progression of cell fate decision is simulated. In addition, our method models the dynamics of gene regulation using the framework of CHN which generates simulation results consistent with the observed data. The constructed model has al-

lowed us to make novel experimentally testable hypotheses about transcriptional mechanisms that control the cell fate conversions.

Applied to the prediction of pseudotimes in real single-cell gene expression data from different types of biological experiments and compared against other methods, HopLand outperformed most of the other methods in most cases. Our method could also be used to identify key regulators and interactions which is helpful for the understanding of underlying mechanisms.

The simulation and analysis results have shown that HopLand has some advantages, whereas the other methods fail in certain circumstances. First, our approach does not rely on any prior knowledge of essential marker genes. Secondly, the non-linear dimensionality reduction method used in HopLand generates a non-linear mapping between the landscape and the phenotype space respecting the non-linear structures of biological systems. Thirdly, we construct the landscape based on biological interactions between genes that allows simulating real biological processes. Fourthly, using the geodesic distances in the landscape to refine the estimation of distances between cells is biologically plausible.

However, in the application of HopLand as a pseudotime estimation algorithm using dynamical systems modeling, it still needs to address several issues. First, the mathematical modeling approach makes use of temporal information, e.g., the physical time points within the single-cell data, which is not required by some other methods. Nevertheless, our method tries to recover the underlying regulatory mechanisms from the data using the extracted information. The physical time points provide the consecutive updates during the process which is useful for the modeling, but for a dataset without temporal details, HopLand can skip the training process and directly predict the pseudotimes based on a landscape constructed using initial settings of parameters. Secondly, HopLand is computationally costly compared with other methods. The parameter learning process is time-consuming partially due to the repeated numerical solution of ODEs.

Moreover, our method was implemented in MATLAB which is not suitable for intensive computation. In the future, we will implement our approach in C/C++ to speed it up. Thirdly, our algorithm was proposed under the premise that the single-cell transcriptional data cover the critical stages along a biological process. If that is not the case, however, the predicted model might give misleading results. Fourthly, HopLand makes use of GP-LVM to extract a 2D latent space from a high-dimensional space which may suffer from the high technical noise in the single-cell data. Although the recovery of pseudotime relies on not only the reduced components but also a third value, i.e., network energy, which can alleviate the influence of the noise, we still recommend users to preprocess their data using some noise reduction techniques, e.g., PAGODA [148], scLVM [149].

The result of HopLand on the qPCR dataset is better than those on the scRNA-seq data. It is probably because the protocols of qPCR make data less prone to the dropout effect [150]. Among the 5 RNA-seq datasets, HopLand has unstable performances, which may be partly caused by the different scRNA-seq protocols used in generating the data [151]. In the future, We will try to analyze HopLand in different sequencing datasets and make it satisfy specific needs of different types of sequencing technologies.

Chapter 6

Conclusion

The past few decades witnessed much of the development and advances of mathematical modeling of Waddington's epigenetic landscape. It has been applied in computational cell biology leading to the design of powerful computational algorithms and tools to reveal the complexity in the regulation of cell fate determination.

Through the modeling of integrated genomic regulatory elements that contribute to the cell fate transitions, a quasi-potential landscape is constructed based on physical theories to explore the system dynamics. However, most of the computational models based on theoretical mechanisms are limited by the lack of biological knowledge and constraint to fine-crafted networks. The rapid advancement of the next generation sequencing techniques provides the opportunity to discover molecular regulations through data analysis and system-level modeling of stem cell functions.

Although different methods have been proposed to realize the famous Waddington's epigenetic landscape metaphor in a principled way, the visualization of the 3D model remains to be solved when the gene regulatory network underlying the landscape contains more than two genes.

This chapter first summarizes the research work presented in this thesis for mathematical modeling of Waddington's epigenetic landscape applied in the study of stem cell fate determination. Then, we will introduce possible directions for future research.

6.1 Summary of the thesis

In this thesis, two computational methodologies, i.e., the knowledge-based modeling method and the data-driven modeling approach, to realize the metaphor of Waddington’s epigenetic landscape are formulated and implemented to facilitate the study of stem cell fate transitions.

The knowledge-based modeling and simulation method is packaged into the software tool “NetLand” for more accessible user experience which serves as a useful tool for understanding the dynamics of stem cells. It was successfully applied to analyzing the bottlenecks in the reprogramming process and identifying the epigenetic barriers which have been shown to cause the low success rate of iPSC reprogramming. Detailed analysis of a few case studies further confirmed the great potential of this method in the study of cell fate transitions, as well as the design of innovations in cell fate engineering and regenerative medicine.

However, the knowledge-based modeling and simulation method relies on intricate knowledge of regulatory subsystems and molecular kinetics. In addition, the complexity of the *in silico* simulation of such a system increases exponentially corresponding to the increasing network size. To address these issues, a data-driven approach named “HopLand” was introduced to model and visualize the Waddington’s epigenetic landscape from single-cell gene expression data using machine learning techniques.

The analysis and construction of large-scale attractor networks with high resolutions have greatly benefited from leveraging the next generation sequencing techniques. This data-driven method of HopLand achieved satisfactory results in estimating the pseudotimes of individual cells in the sequencing data, demonstrating its potential to provide fundamental insights into cell fate regulation.

The mathematical modeling methods proposed in this thesis, i.e., knowledge-based and data-driven approaches, provide rationalized means for quantitatively modeling and visualizing the Waddington’s epigenetic landscape,

thereby enabling the utilization of the constructed landscape to explain and predict the dynamical behaviors of cell fate decisions.

6.2 Future work

Although a lot of effort has been put into the research work presented in this thesis, the quantitative modeling methods for Waddington’s epigenetic landscape and its application in cell fate conversion studies still have much room for further improvement. This section briefly introduces a few possible directions for future work on this topic.

Automatic construction of biological networks

In the current design, the knowledge-based methods implemented in NetLand rely on the input of a constructed gene regulatory network. However, the construction process requires extensive efforts and technical proficiency to search and integrate the information from diverse resources, it is highly desirable to extend the NetLand software with functionalities to support automatic and customizable biological network construction.

The interactions (edges) representing regulatory relationships among the genes (nodes) are thought to have either a physical interaction through their gene products, e.g., proteins, or to be that one of the genes alters or affects the activity of other genes of interest. High-throughput methodologies, such as RNA-seq, or Chip-seq, have been widely used to identify these two general categories of interactions: physical protein-protein interactions and genetic interactions.

Almost all of the interactions that have been discovered through wet-lab experiments are collected in public databases for free sharing. There are also computational approaches that are used to predict protein-protein interactions. These methods utilize genomic data to establish structural or evolutionary links among protein pairs or to predict novel interactions by analyzing known interactions. Utilizing these publicly available resources,

we can build an interaction network. A tentative list of steps that future users of NetLand may follow is given here:

1. Identify the essential genes involved in the biological process. Users can upload the set of target nodes or extract them from public datasets.
2. Obtain a list of interactions from publicly available interaction databases and literature. Querying task is automatically provoked in the saved resources.
3. Remove redundant items and curate inconsistent records.
4. Determine the model parameters as described in Section 3.1.3.

In many cases, when the network topology is identified, the required parameters can be obtained with some effort, e.g. through qualitative simulation or statistical inference from empirical data. However, the inference of parameters in biological networks remains a big challenge.

Another challenge for the automatic construction of biological networks is the establishment of context-specific networks with not only customized parameters but also topological rewiring. There is still a long way before we can construct a master model to explain different biological processes under different conditions with simple calibration. As a result, a repository of context-specific networks and regulatory factors would benefit the construction of the genetic circuits.

Extended formulation coverage of NetLand

Currently, the kinetics of the regulatory relations (associated with the edges) are formulated with Hill equations (Eq. 3.1). These equations have been widely used for modeling multiple biological networks (Table 2.2). However, many different approaches have been developed to model and simulate gene regulatory networks [152]. In addition, while it is expected to apply this software to some other types of biochemical reaction networks, e.g., metabolic

networks, the formulation module needs to be adapted according to specific kinetic laws for biochemical reactions.

Although a growing number of modeling and analysis techniques have been employed to study different types of biological systems [153], [154], a deeper understanding of the kinetics and molecular interactions is still needed to produce better software tools and more realistic simulations of complex biological processes. After integrating the available computational modeling methods and tools, the execution of a computer program can provide an *in silico* numerical evaluation of hypotheses, thereby accelerating knowledge discovery in molecular and cell biology.

Efficient optimization algorithm in HopLand

HopLand is computationally costly partially due to the parameter learning process. The optimization algorithm used in HopLand is gradient descent (GD) in which a set of parameters are updated iteratively to minimize an error function. Thus, if the number of training samples is big, then using GD may take too long because, in every iteration, it runs through the complete training set. To achieve scalability in the future, using stochastic gradient descent (SGD) would be a better choice, because it uses only a subset of training sample to do the update for a parameter in a particular iteration. Moreover, the optimization method in HopLand can be parallelized on a multicore HPC (High-performance Computing) Cluster to achieve significant acceleration compared to the single threaded CPU implementation [155].

Furthermore, HopLand was implemented in MATLAB which is not suitable for intensive computation. To speed up HopLand in C/C++ would also be a practically useful future work.

Data-specific optimization in HopLand

The application of the data-driven approach of HopLand in the pseudotime recovery has been validated in Chapter 5 with satisfactory results. However,

HopLand achieved unstable performances on different types of sequencing datasets, e.g., the results of qPCR datasets are better than those of scRNA-seq, which may be partly caused by the different protocols used in generating the data [151]. Thus customizing HopLand for different sequencing datasets and making it satisfy specific needs of different types of sequencing technologies could also be a future direction of this work.

Another important challenge of optimization in HopLand is data pre-processing. The analysis of single-cell sequencing data is challenging due to the high variability in gene expression between individual cells. Moreover, the dimensionality reduction method GP-LVM used in HopLand may suffer from the high technical noise in the single-cell data. Therefore, it is of great importance to choose the proper method to pre-process the data.

Appendix

$$\text{max_a} * (\text{BasalExpression_a} + (\text{BasalExpression_a_1} + \text{I_a_1} * (1 / (1 + (\text{b} / \text{K_a_1})^{\text{N_a_1}})))) - \text{deg_a} * \text{a}$$

$$\text{max_b} * (\text{BasalExpression_b} + (\text{BasalExpression_b_1} + \text{I_b_1} * (1 / (1 + (\text{a} / \text{K_b_1})^{\text{N_b_1}})))) - \text{deg_b} * \text{b}$$

Figure S1: The modified reactions after removing self-loops.

Algorithm 6 Functions

```

1: function GENERATERANDOMINITIALSTATES(startPoints)
2:   number of simulated initial states numSeq = 1000,
3:   Fluctuation intensity  $\alpha = 0.01$ ;
4:   stdValue = std(startPoints);
5:    $\gamma = 0.001$ ;
6:   for i = 1 to numSeq do
7:     randomly select one sample sampleSeq from startPoints;
8:     randomly select a list of genes sampleGene;
9:     randomNoise =  $\gamma * \text{rand}(1, N)$ ;
10:    perturbation = sampleGene *  $\alpha * \text{stdValue}$ ;
11:    randomInits(i) = sampleSeq + randomNoise + perturbation;
12:  end for
13:  return randomInits
14: end function

15: function FITMIXTUREGAUSSIAN(D)
16:   for all gene in D do
17:     gmmModels(gene) = gmdistribution.fit(D(gene));
18:   end for
19:   return gmmModels
20: end function

21: function GENERATETRAJECTORY(D, cellStates)
22:   for all gene in D do
23:     for all state in cellStates do
24:        $\sigma = 1./\text{std}(D(\text{gene}, \text{state}))$ ;
25:     end for
26:   end for
27:   return  $\sigma$ 
28: end function

29: function INITIALIZEPARAM(D)
30:   Let  $\theta = \{\delta_i, I_i, C_i, W_{ij}, i, j = 1, 2, \dots, N\}$ , N is the number of genes;
31:   Set  $\delta_i = 1$ ,  $I_i = 0$ ,  $C_i = 1$ ,  $W_{ij} = \text{corr}(\mathbf{D})$ ;
32:   return  $\theta$ 
33: end function

```

Figure S2: The probabilistic landscapes of the modified two-gene network when (a) mapping to two marker genes and (b) a 2D latent space. (c) The table of attractors.

Gene Names	#1
percentage	1.0 (10)
Oct4	1.38
Sox2	1.4
ECTs	0.51
MEs	0.52

Figure S3: The table of attractor(s).

Figure S4: (a) The topology structure of the model with 6 genes and 23 interactions. The probabilistic landscapes of the network when (b) mapping to two marker genes (Nanog and Gata6) and (c) a 2D latent space. (d) The two attractors.

Figure S5: The weight matrix contains $N \times N$ interactions of CHN learned from the mouse embryonic early development dataset. N is the number of genes.

Figure S6: Accuracies of pseudotime recovery on the qPCR dataset “DENG2014” using different pseudotime recovery methods. Pearson correlation coefficient between the predicted and observed times is used to evaluate the result.

(a)

(b)

(c)

Figure S7: The contour plot of the landscape constructed from the dataset (a) "ES_MEF", (b) "LPS" and (c) "HSSM".

Table S1: The detailed references for verifying interactions in HopLand's result using GeneGo (Part 1)

Description	PMID
Fgf and Esrrb integrate epigenetic and transcriptional networks that regulate self-renewal of trophoblast stem cells	26206133
KLF2 binds to gene ERR2 promoter and activates ERR2 expression.	18264089;20959618
HNF4-alpha can bind to gene Fibronectin promoter.	19761587;21384409;25786776
NANOG directly activates Oct-3/4 transcription.	16153702;16518401;16790525
ID2 promoter has a putative GATA-3 binding site.	15608294
SOX2 binds to Fibronectin promoter.	27411517;18358816;19884255
Oct-3/4 decreases Hand1 expression	16325584;20014010;17183653
Probably, HAND1 is one of the Nanog genomic targets in murine ES cells. NANOG inhibits expression of HAND1.	19567677;21062744;18700969
HAND1 promoter has a putative GATA-3 binding site.	15608294
SOX2 decreases HAND1 mRNA expression.	17506876;16767105;17515932
Oct-3/4 probably regulates transcription of C14orf92.	18555785;22083510
SAHH was identified as Oct-3/4 target gene by ChIP-on-Chip assay.	16751344;19884255;23376973
Esrrb positively regulates expression of the Oct4 target gene Nanog by binding to its promoter.	18662995;18957414;19740934
ERR2 binds to Oct-3/4 and modulates its transcriptional activity.	18662995;18957414;19740934
Esrrb positively regulates the expression of reprogramming genes Klf4 and Sox2.	23019124;19740934
ERR2 activates transcription of KLF5 .	23169531
In mouse ES cells gene ERR2 was identified as the downstream target of Oct3/4 by chromatin-immunoprecipitation assays.	20362541;23508100;24905168
KLF2 binds to gene ERR2 promoter and activates ERR2 expression.	18264089;20959618
KLF5 binds to gene ERR2 promoter and activates ERR2 expression.	18264089;20959618
NANOG physically interacts with ERR2.	19884255;18957414;18454139
SOX2 can bind to gene ESRRB promoter.	19740934;24120664;22334693

Table S2: The detailed references for verifying interactions in HopLand's result using GeneGo (Part 2)

Description	PMID
Using chromatin immunoprecipitation was found that Med12 and Nanog co-occupied Nanog target promoter BMP4 in embryonic stem cells.	19036726;19884255;19567677
Oct-3/4 downregulates BMP4 expression	16518401;23041284;17506876
SOX2 decreases BMP4 mRNA expression.	17506876;18238855
BMP-4 gene was identified as a direct downstream target for GATA-4 and GATA-6 which can activate BMP-4 transcription.	12606287;19491195;26395571
NANOG can bind to gene GATA-4 promoter and suppress GATA-4 expression.	16456133;16391521;15983365
Oct-3/4-directly downregulates GATA-4 expression.	14990861;16153702;17567999
qRT-PCR revealed that the levels of mRNAs encoding FOXA1, FOXA2, FOXA3, HNF1B, HNF1A, GATA4, GATA6 and HHEX, all of which have crucial roles during early stages of hepatic development, were reduced in HNF4A-depleted cells.	20054869;21852396
KLF2 regulates transcription of GATA-4 .	23457456
Interaction of the Id 1-3 Proteins with Gata4 and Nkx2.5 was demonstrated. The Id proteins inhibited the synergistic transactivation of reporter constructs by Gata4 and Nkx2.5 in consequence of inhibiting the sequence-specific binding of Gata4 and Nkx2.5 to DNA and the binding of Gata4 to Nkx2.5.	16556596
GATA-4 can bind to gene ID2 promoter and activates ID2 expression.	17559079;21330551
GATA-4 inhibits transcription of NANOG .	23132827
GATA-4 binds to gene FGF4 promoter and activates FGF4 expression.	15942670;15183731
BMP-4 gene was identified as a direct downstream target for GATA-4 and GATA-6 which can activate BMP-4 transcription.	12606287;19491195;26395571
GATA-4 physically interacts with HNF4-alpha and increases its activity.	15159395;25053715;18426912
KLF5 5' upstream region has a putative NANOG-binding site.	20875146;19544440;16153702
KLF5 5' upstream region has a putative Oct-3/4-binding site.	17925037;17183653;16153702
ERR2 activates transcription of KLF5 .	23169531

Table S3: The detailed references for verifying interactions in HopLand's result using GeneGo (Part 3)

Description	PMID
KLF5 binds to gene NANOG promoter and activates NANOG expression.	24770696;18264089;20875146
KLF5 activates transcription of HNF4-alpha	25053715;24770696
.	
KLF5 binds to gene ERR2 promoter and activates ERR2 expression.	18264089;20959618
putative homologs were found interacting (Actb and Lcp1)	24758171
KLF2 binds to gene ERR2 promoter and activates ERR2 expression.	18264089;20959618
KLF4 binding to gene KLF2 promoter was identified on a global scale in mouse Embryonic Stem cells and validated by quantitative ChIP-PCR.	18358816;19796622;23667633
Esrrb positively regulates the expression of reprogramming genes Klf4 and Sox2.	23019124;19740934;19136965

Table S4: Logical rules in the Boolean model.

Product	Equations
Oct4	$\text{geneOct4} \& (\text{Sig2} \text{OS} \text{OSN} \text{Myc} \text{Klf4}) \& !\text{OC} \& !\text{Gcnf} \& !\text{Sox1} \& !\text{Gata6}$
Sox2	$\text{geneSox2} \& (\text{OS} \text{OSN} \text{Myc} \text{Klf4}) \& !\text{Sox1} \& !\text{Gcnf} \& !\text{Gata6}$
Nanog	$\text{geneNanog} \& (\text{OS} \text{OSN} \text{Myc} \text{Klf4}) \& !\text{OG} \& !\text{Oct4} \& !\text{Gcnf} \& !\text{Cdx2} \& !\text{Sox1} \& !\text{Gata6}$
Cdx2	$\text{Cdx2} \& !(\text{OC}) \& !\text{Nanog} \& !\text{Sox1} \& !\text{Gata6}$
Gcnf	$\text{Cdx2} \text{Gata6} \text{---} \text{Gcnf}$
Pax6	$(\text{Sox2} \text{Pax6}) \& !\text{Oct4} \& !\text{Nanog}$
Sox1	$\text{geneSox1} \& (\text{Sig1} \text{Sox2}) \& !\text{Nanog} \& !\text{Oct4} \& !\text{Cdx2} \& !\text{Gata6}$
Gata6	$\text{geneGata6} \& (\text{Sig0} \text{Oct4}) \& !\text{Nanog} \& !\text{Sox2} \& !\text{Cdx2} \& !\text{Sox1}$
Myc	$\text{Klf4} \& !\text{Sox1} \& !\text{Gcnf} \& !\text{Cdx2}$
Klf4	$(\text{Myc} \text{Klf4}) \& !\text{Gcnf} \& !\text{Sox1} \& !\text{Cdx2} \& !\text{Gata6}$
OG	$\text{Oct4} \& \text{Gata6}$
OC	$\text{Oct4} \& \text{Cdx2}$
OS	$\text{Oct4} \& \text{Sox2}$
OSN	$\text{Oct4} \& \text{Sox2} \& \text{Nanog}$
Prc2	$(\text{Oct4} \text{Sox2} \text{Klf4} \text{Myc}) \& !\text{Mbd3} \& !\text{Nurd}$
Mbd3	$\text{Gata6} \& !\text{Prc2} \& !\text{Nurd}$
Nurd	$\text{Sox1} \& !\text{Prc2} \& !\text{Mbd3}$
EA1	$(\text{Oct4} \text{Sox2} \text{Klf4} \text{Myc}) \& !\text{EA2} \& !\text{EA3}$
EA2	$\text{Gata6} \& !\text{EA1} \& !\text{EA3}$
EA3	$\text{Sox1} \& !\text{EA1} \& !\text{EA2}$
Gene	$\text{EA1} \& !\text{Mbd3} \& !\text{Nurd}$
Oct4	
Gene	$\text{EA1} \& !\text{Mbd3} \& !\text{Nurd}$
Sox2	
Gene	$\text{EA1} \& !\text{Mbd3} \& !\text{Nurd}$
Nanog	
Gene	$\text{EA2} \& !\text{Prc2} \& !\text{Nurd}$
Gata6	
Gene	$\text{EA3} \& !\text{Prc2} \& !\text{Mbd3}$
Sox1	

Table S5: ODE equations of the dynamical model of Figure 3.1.

$$\begin{aligned}
\frac{dOct4}{dt} &= geneOct4 * (kbase + a * (\frac{(Oct4 * Sox2)^{n1}}{k1^{n1} + (Oct4 * Sox2)^{n1}} + \\
&\quad \frac{Klf4^{n1}}{k1^{n1} + Klf4^{n1}} + \frac{(Oct4 * Sox2 * Nanog)^{n1}}{k1^{n1} + (Oct4 * Sox2 * Nanog)^{n1}} + \\
&\quad \frac{Myc^{n1}}{k1^{n1} + Myc^{n1}}) * \frac{k2^{n2}}{k2^{n2} + Gata6^{n2}} \\
&\quad * \frac{k2^{n2}}{k2^{n2} + Sox1^{n2}} + (\frac{k2^{n2}}{k2^{n2} + (Oct4 * Cdx2)^{n2}} + \\
&\quad \frac{k2^{n2}}{k2^{n2} + Gcnf^{n2}}) * b) - gamma * Oct4 \\
\frac{dSox2}{dt} &= geneSox2 * (kbase + a * (\frac{(Oct4 * Sox2)^{n1}}{k1^{n1} + (Oct4 * Sox2)^{n1}} + \\
&\quad \frac{(Oct4 * Sox2 * Nanog)^{n1}}{k1^{n1} + (Oct4 * Sox2 * Nanog)^{n1}} + \frac{Myc^{n1}}{k1^{n1} + Myc^{n1}} + \\
&\quad \frac{Klf4^{n1}}{k1^{n1} + Klf4^{n1}}) * \frac{k2^{n2}}{k2^{n2} + Sox1^{n2}} * \frac{k2^{n2}}{k2^{n2} + Gata6^{n2}} + \\
&\quad \frac{k2^{n3}}{k2^{n3} + Oct4^{n3}} * b) - gamma * Sox2 \\
\frac{dNanog}{dt} &= geneNanog * (kbase + a * (\frac{(Oct4 * Sox2)^{n1}}{k1^{n1} + (Oct4 * Sox2)^{n1}} + \\
&\quad \frac{(Oct4 * Sox2 * Nanog)^{n1}}{k1^{n1} + (Oct4 * Sox2 * Nanog)^{n1}} + \frac{Myc^{n1}}{k1^{n1} + Myc^{n1}} + \\
&\quad \frac{Klf4^{n1}}{k1^{n1} + Klf4^{n1}}) * \frac{k2^{n2}}{k2^{n2} + (Oct4 * Gata6)^{n2}} * \frac{k2^{n2}}{k2^{n2} + Sox1^{n2}} + \\
&\quad \frac{k2^{n2}}{k2^{n2} + Oct4^{n2}} * b) - gamma * Nanog \\
\frac{dCdx2}{dt} &= kbase + a_{10} * \frac{Cdx2^{n1}}{k1^{n1} + Cdx2^{n1}} * \frac{k2^{n3}}{k2^{n3} + Nanog^{n3}} * \\
&\quad \frac{k2^{n3}}{k2^{n3} + (Oct4 * Cdx2)^{n3}} * b - gamma * Cdx2 \\
\frac{dGcnf}{dt} &= kbase + aa_{10} * (\frac{Cdx2^{n1}}{k1^{n1} + Cdx2^{n1}} + \frac{Gata6^{n1}}{k1^{n1} + Gata6^{n1}} + \\
&\quad \frac{Gcnf^{n1}}{k1^{n1} + Gcnf^{n1}}) - gamma * Gcnf \\
\frac{dPax6}{dt} &= kbase + aa_{10} * (\frac{Sox2^{n1}}{k1^{n1} + Sox2^{n1}} + \frac{Pax6^{n1}}{k1^{n1} + Pax6^{n1}}) * \\
&\quad \frac{k2^{n3}}{k2^{n3} + Oct4^{n3}} * \frac{k2^{n3}}{k2^{n3} + Nanog^{n3}} * b - gamma * Pax6
\end{aligned}$$

$$\begin{aligned}
\frac{dSox1}{dt} &= geneSox1 * (kbase + aa_{10} * \frac{Sox2^{n1}}{k1^{n1} + Sox2^{n1}} * \frac{k2^{n3}}{k2^{n3} + Gata6^{n3}} * \\
&\quad \frac{k2^{n3}}{k2^{n3} + Oct4^{n3}} + \frac{k2^{n3}}{k2^{n3} + Nanog^{n3}} * b - gamma * Sox1 \\
\frac{dGata6}{dt} &= geneGata6 * (kbase + aa_{10} * \frac{Oct4^{n1}}{(k1^{n1} + Oct4^{n1}} * \frac{k2^{n3}}{k2^{n3} + Sox1^{n3}} * \\
&\quad \frac{k2^{n3}}{k2^{n3} + Sox2^{n3}} + \frac{k2^{n3}}{k2^{n3} + Nanog^{n3}} * b - gamma * Gata6 \\
\frac{dMyc}{dt} &= kbase + a * \frac{Klf4^{n1}}{k1^{n1} + Klf4^{n1}} * \frac{k2^{n3}}{k2^{n3} + Sox1^{n3}} * \frac{k2^{n3}}{k2^{n3} + Gata6^{n3}} * \\
&\quad \frac{k2^{n3}}{k2^{n3} + Cdx2^{n3}} + \frac{k2^{n3}}{k2^{n3} + Gcnf^{n3}} * b - gamma * Myc \\
\frac{dKlf4}{dt} &= kbase + a * \frac{Klf4^{n1}}{k1^{n1} + Klf4^{n1}} * \frac{k2^{n3}}{k2^{n3} + Sox1^{n3}} * \frac{k2^{n3}}{k2^{n3} + Cdx2^{n3}} * \\
&\quad \frac{k2^{n3}}{k2^{n3} + Gata6^{n3}} + \frac{k2^{n3}}{k2^{n3} + Gcnf^{n3}} * b - gamma * Klf4 \\
\frac{dPrc2}{dt} &= kbase + a_1 * \frac{Oct4^{n4}}{k3^{n4} + Oct4^{n4}} + a_1 * \frac{Sox2^{n4}}{k3^{n4} + Sox2^{n4}} + \\
&\quad a_1 * \frac{Myc^{n4}}{k4^{n4} + Myc^{n4}} + a_1 * \frac{Klf4^{n4}}{k4^{n4} + Klf4^{n4}} + \\
&\quad a_2 * \frac{k4^{n4}}{k4^{n4} + Mbd3^{n4}} + a_2 * \frac{k4^{n4}}{k4^{n4} + Nurd^{n4}} - gamma * Prc2 \\
\frac{dMbd3}{dt} &= kbase + a_3 * \frac{Gata6^{n4}}{k3^{n4} + Gata6^{n4}} + a_4 * \frac{k4^{n4}}{k4^{n4} + Prc2^{n4}} + \\
&\quad a_4 * \frac{k4^{n4}}{k4^{n4} + Nurd^{n4}} - gamma * Mbd3 \\
\frac{dNurd}{dt} &= kbase + a_3 * \frac{Sox1^{n4}}{k3^{n4} + Sox1^{n4}} + a_4 * \frac{k4^{n4}}{k4^{n4} + Prc2^{n4}} + \\
&\quad a_4 * \frac{k4^{n4}}{k4^{n4} + Mbd3^{n4}} - gamma * Nurd \\
\frac{dEA1}{dt} &= a_5 * \frac{Oct4^{n4}}{k3^{n4} + Oct4^{n4}} + a_5 * \frac{Sox2^{n4}}{k3^{n4} + Sox2^{n4}} + a_5 * \frac{Myc^{n4}}{k4^{n4} + Myc^{n4}} + \\
&\quad a_5 * \frac{Klf4^{n4}}{k4^{n4} + Klf4^{n4}} + a_6 * \frac{k4^{n4}}{k4^{n4} + EA2^{n4}} + \\
&\quad a_6 * \frac{k4^{n4}}{k4^{n4} + EA3^{n4}} - gamma * EA1
\end{aligned}$$

$$\begin{aligned}
\frac{dEA2}{dt} &= a_7 * \frac{Gata6^{n4}}{k3^{n4} + Gata6^{n4}} + a_7 * \frac{k4^{n4}}{k4^{n4} + EA1^{n4}} + \\
&\quad a_7 * \frac{k4^{n4}}{k4^{n4} + EA3^{n4}} - gamma * EA2 \\
\frac{dEA3}{dt} &= a_7 * \frac{Sox1^{n4}}{k3^{n4} + Sox1^{n4}} + a_7 * \frac{k4^{n4}}{k4^{n4} + EA1^{n4}} + \\
&\quad a_7 * \frac{k4^{n4}}{k4^{n4} + EA2^{n4}} - gamma * EA3 \\
geneOct4 &= a_8 * \frac{k3^{n4}}{k3^{n4} + Mbd3^{n4}} + \frac{k3^{n4}}{k3^{n4} + Nurd^{n4}} + \frac{EA1^{n4}}{k4^{n4} + EA1^{n4}} \\
geneSox2 &= a_8 * \frac{k3^{n4}}{k3^{n4} + Mbd3^{n4}} + \frac{k3^{n4}}{k3^{n4} + Nurd^{n4}} + \frac{EA1^{n4}}{k4^{n4} + EA1^{n4}} \\
geneNanog &= a_8 * \frac{k3^{n4}}{k3^{n4} + Mbd3^{n4}} + \frac{k3^{n4}}{k3^{n4} + Nurd^{n4}} + \frac{EA1^{n4}}{k4^{n4} + EA1^{n4}} \\
geneGata6 &= a_9 * \frac{k3^{n4}}{k3^{n4} + Prc2^{n4}} + \frac{k3^{n4}}{k3^{n4} + Nurd^{n4}} + \frac{EA2^{n4}}{k4^{n4} + EA2^{n4}} \\
geneSox1 &= a_9 * \frac{k3^{n4}}{k3^{n4} + Prc2^{n4}} + \frac{k3^{n4}}{k3^{n4} + Mbd3^{n4}} + \frac{EA3^{n4}}{k4^{n4} + EA3^{n4}}
\end{aligned}$$

Table S6: The inferred parameters of the network model.

Parameter	Defination	Value
kbase	exogenous expression	0
gamma	degradation rate	0.1
a	maximum transcrip- tional rates	0.1
b	maximum transcrip- tional rates	0.01
a_{1-5}	maximum transcrip- tional rates	0.4802,0.7065,0.6033,0.2405,0.8269
a_{6-10}	maximum transcrip- tional rates	0.0386,0.7158,0.3503,0.6858,0.7
n1	Hill coefficient	1
n2	Hill coefficient	5
n3	Hill coefficient	7
n4	Hill coefficient	2
k1	dissociation constant	0.1
k2	dissociation constant	0.4
k3	dissociation constant	0.0252
k4	dissociation constant	0.4086

Bibliography

- [1] K. Takahashi and S. Yamanaka, “Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors”, *Cell*, 2006, ISSN: 00928674. DOI: 10.1016/j.cell.2006.07.024.
- [2] J. Sandoe and K. Eggan, “Opportunities and challenges of pluripotent stem cell neurodegenerative disease models”, *Nature neuroscience*, vol. 16, no. 7, pp. 780–789, 2013.
- [3] D. A. Robinton and G. Q. Daley, “The promise of induced pluripotent stem cells in research and therapy”, *Nature*, vol. 481, no. 7381, p. 295, 2012.
- [4] S. M. Wu and K. Hochedlinger, “Harnessing the potential of induced pluripotent stem cells for regenerative medicine”, *Nature cell biology*, vol. 13, no. 5, pp. 497–505, 2011.
- [5] M. H. Chin, M. J. Mason, W. Xie, S. Volinia, M. Singer, C. Peterson, G. Ambartsumyan, O. Aimiwu, L. Richter, J. Zhang, *et al.*, “Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures”, *Cell stem cell*, vol. 5, no. 1, pp. 111–123, 2009.
- [6] R. A. Meyers, *Encyclopedia of complexity and systems science*. Springer, 2009.
- [7] S. Bhattacharya, Q. Zhang, and M. E. Andersen, “A deterministic map of waddington’s epigenetic landscape for cell fate specification”, *BMC systems biology*, vol. 5, no. 1, p. 85, 2011.
- [8] R. I. Hamed, S. Ahson, and R Parveen, “A new approach for modelling gene regulatory networks using fuzzy petri nets”, *Journal of Integrative Bioinformatics (JIB)*, vol. 7, no. 1, pp. 1–16, 2010.
- [9] H.-y. Wang, Y.-x. Huang, Y.-f. Qi, Y. Zhang, Y.-l. Bao, L.-g. Sun, L.-h. Zheng, Y.-w. Zhang, Z.-q. Ma, and Y.-x. Li, “Mathematical models for the notch and wnt signaling pathways and the crosstalk between them during somitogenesis”, *Theoretical Biology and Medical Modelling*, vol. 10, no. 1, p. 27, 2013.
- [10] T. Tian and K. Smith-Miles, “Mathematical modeling of gata-switching for regulating the differentiation of hematopoietic stem cell”, *BMC systems biology*, vol. 8, no. 1, S8, 2014.

- [11] S. R. Maetschke and M. A. Ragan, “Characterizing cancer subtypes as attractors of Hopfield networks”, *Bioinformatics*, vol. 30, no. 9, pp. 1273–1279, May 2014, ISSN: 1367-4803.
- [12] A. Taherian Fard, S. Srihari, J. C. Mar, and M. A. Ragan, “Not just a colourful metaphor: Modelling the landscape of cellular development using hopfield networks”, *Npj Systems Biology and Applications*, vol. 2, p. 16 001, 2016.
- [13] A. H. Lang, H. Li, J. J. Collins, and P. Mehta, “Epigenetic landscapes explain partially reprogrammed cells and identify key reprogramming genes”, *PLoS computational biology*, vol. 10, no. 8, e1003734, 2014.
- [14] V. K. Singh, M. Kalsan, N. Kumar, A. Saini, and R. Chandra, “Induced pluripotent stem cells: Applications in regenerative medicine, disease modeling, and drug discovery”, *Frontiers in cell and developmental biology*, vol. 3, p. 2, 2015.
- [15] Y. Rais, A. Zviran, S. Geula, O. Gafni, E. Chomsky, S. Viukov, A. A. Mansour, I. Caspi, V. Krupalnik, M. Zerbib, *et al.*, “Deterministic direct reprogramming of somatic cells to pluripotency”, *Nature*, vol. 502, no. 7469, pp. 65–70, 2013.
- [16] J. Hanna, K. Saha, B. Pando, J. Van Zon, C. J. Lengner, M. P. Creighton, A. van Oudenaarden, and R. Jaenisch, “Direct cell reprogramming is a stochastic process amenable to acceleration”, *Nature*, vol. 462, no. 7273, p. 595, 2009.
- [17] T. Graf and T. Enver, “Forcing cells to change lineages”, *Nature*, vol. 462, no. 7273, p. 587, 2009.
- [18] J. Shu, C. Wu, Y. Wu, Z. Li, S. Shao, W. Zhao, X. Tang, H. Yang, L. Shen, X. Zuo, W. Yang, Y. Shi, X. Chi, H. Zhang, G. Gao, Y. Shu, K. Yuan, W. He, C. Tang, Y. Zhao, and H. Deng, “XInduction of pluripotency in mouse somatic cells with lineage specifiers”, *Cell*, 2013, ISSN: 00928674. DOI: 10.1016/j.cell.2013.05.001.
- [19] S. Diecke, S. M. Jung, J. Lee, and J. H. Ju, “Recent technological updates and clinical applications of induced pluripotent stem cells.”, *The Korean journal of internal medicine*, vol. 29, no. 5, pp. 547–557, 2014.
- [20] C. Eguizabal, N. Montserrat, A. Veiga, and J. C. I. Belmonte, “Dedifferentiation, transdifferentiation, and reprogramming: Future directions in regenerative medicine”, in *Seminars in reproductive medicine*, Thieme Medical Publishers, vol. 31, 2013, pp. 082–094.
- [21] H. Okano, M. Nakamura, K. Yoshida, Y. Okada, O. Tsuji, S. Nori, E. Ikeda, S. Yamanaka, and K. Miura, “Steps toward safe cell therapy using induced pluripotent stem cells”, *Circulation research*, vol. 112, no. 3, pp. 523–533, 2013.

- [22] K. Takahashi and S. Yamanaka, “Induced pluripotent stem cells in medicine and biology”, *Development*, vol. 140, no. 12, pp. 2457–2461, 2013.
- [23] J. A. Thomson, J. Itskovitz-Eldor, S. S. Shapiro, M. A. Waknitz, J. J. Swiergiel, V. S. Marshall, and J. M. Jones, “Embryonic stem cell lines derived from human blastocysts”, *science*, vol. 282, no. 5391, pp. 1145–1147, 1998.
- [24] X. Ren-He, X. Chen, D. S. Li, R. Li, G. C. Addicks, C. Glennon, T. P. Zwaka, and J. A. Thomson, “Bmp4 initiates human embryonic stem cell differentiation to trophoblast”, *Nature biotechnology*, vol. 20, no. 12, p. 1261, 2002.
- [25] Y.-S. Chan, J. Göke, J.-H. Ng, X. Lu, K. A. U. Gonzales, C.-P. Tan, W.-Q. Tng, Z.-Z. Hong, Y.-S. Lim, and H.-H. Ng, “Induction of a human pluripotent state with distinct regulatory circuitry that resembles preimplantation epiblast”, *Cell stem cell*, vol. 13, no. 6, pp. 663–675, 2013.
- [26] N. Montserrat, E. Nivet, I. Sancho-Martinez, T. Hishida, S. Kumar, L. Miquel, C. Cortina, Y. Hishida, Y. Xia, C. R. Esteban, and J. C. Izpisua Belmonte, “Reprogramming of human fibroblasts to pluripotency with lineage specifiers”, *Cell Stem Cell*, 2013, ISSN: 19345909. DOI: 10.1016/j.stem.2013.06.019.
- [27] N. Maherali, R. Sridharan, W. Xie, J. Utikal, S. Eminli, K. Arnold, M. Stadtfeld, R. Yachechko, J. Tchieu, R. Jaenisch, *et al.*, “Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution”, *Cell stem cell*, vol. 1, no. 1, pp. 55–70, 2007.
- [28] J.-C. D. Heng, B. Feng, J. Han, J. Jiang, P. Kraus, J.-H. Ng, Y. L. Orlov, M. Huss, L. Yang, T. Lufkin, *et al.*, “The nuclear receptor nr5a2 can replace oct4 in the reprogramming of murine somatic cells to pluripotent cells”, *Cell stem cell*, vol. 6, no. 2, pp. 167–174, 2010.
- [29] E. Apostolou and K. Hochedlinger, “Chromatin dynamics during cellular reprogramming”, *Nature*, vol. 502, no. 7472, pp. 462–471, 2013.
- [30] S. Huang and S. A. Kauffman, “Complexgrn complex gene complex grn regulatory networks—from structure to biological observables: Cell fate determination gene regulation, cell fate determination”, in *Encyclopedia of complexity and systems science*, Springer, 2009, pp. 1180–1213.
- [31] C. H. Waddington *et al.*, “The strategy of the genes. a discussion of some aspects of theoretical biology. with an appendix by h. kacser.”, *The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser.*, 1957.
- [32] H. Niwa, K. Ogawa, D. Shimosato, and K. Adachi, “A parallel circuit of lif signalling pathways maintains pluripotency of mouse es cells”, *Nature*, vol. 460, no. 7251, p. 118, 2009.

- [33] P. Samavarchi-Tehrani, A. Golipour, L. David, H.-k. Sung, T. A. Beyer, A. Datti, K. Woltjen, A. Nagy, and J. L. Wrana, “Functional genomics reveals a bmp-driven mesenchymal-to-epithelial transition in the initiation of somatic cell reprogramming”, *Cell stem cell*, vol. 7, no. 1, pp. 64–77, 2010.
- [34] J. M. Polo, E. Anderssen, R. M. Walsh, B. A. Schwarz, C. M. Nefzger, S. M. Lim, M. Borkent, E. Apostolou, S. Alaei, J. Cloutier, *et al.*, “A molecular roadmap of reprogramming somatic cells into ips cells”, *Cell*, vol. 151, no. 7, pp. 1617–1632, 2012.
- [35] B. Papp and K. Plath, “Epigenetics of reprogramming to induced pluripotency”, *Cell*, vol. 152, no. 6, pp. 1324–1343, 2013.
- [36] S. Huang, I. Ernberg, and S. Kauffman, “Cancer attractors: A systems view of tumors from a gene network dynamics and developmental perspective”, in *Seminars in cell & developmental biology*, Elsevier, vol. 20, 2009, pp. 869–876.
- [37] Y. Buganim, D. A. Faddah, and R. Jaenisch, “Mechanisms and models of somatic cell reprogramming”, *Nature reviews. Genetics*, vol. 14, no. 6, p. 427, 2013.
- [38] A. M. Singh, T. Hamazaki, K. E. Hankowski, and N. Terada, “A heterogeneous expression pattern for nanog in embryonic stem cells”, *Stem cells*, vol. 25, no. 10, pp. 2534–2542, 2007.
- [39] E. E. Romero-Lanman, S. Pavlovic, B. Amlani, Y. Chin, and R. Benezra, “Id1 maintains embryonic stem cell self-renewal by up-regulation of nanog and repression of brachyury expression”, *Stem cells and development*, vol. 21, no. 3, pp. 384–393, 2011.
- [40] Z. Wei, Y. Yang, P. Zhang, R. Andrianakos, K. Hasegawa, J. Lyu, X. Chen, G. Bai, C. Liu, M. Pera, *et al.*, “Klf4 interacts directly with oct4 and sox2 to promote reprogramming”, *Stem cells*, vol. 27, no. 12, pp. 2969–2978, 2009.
- [41] K. Okita and S. Yamanaka, “Intracellular signaling pathways regulating pluripotency of embryonic stem cells”, *Current stem cell research & therapy*, vol. 1, no. 1, pp. 103–111, 2006.
- [42] O. Dreesen and A. H. Brivanlou, “Signaling pathways in cancer and embryonic stem cells”, *Stem cell reviews*, vol. 3, no. 1, pp. 7–17, 2007.
- [43] J. Trott and A. M. Arias, “Single cell lineage analysis of mouse embryonic stem cells at the exit from pluripotency”, *Biology open*, BIO20135934, 2013.
- [44] S. Y. Sokol, “Maintaining embryonic stem cell pluripotency with wnt signaling”, *Development*, vol. 138, no. 20, pp. 4341–4350, 2011.
- [45] F. Lanner and J. Rossant, “The role of fgf/erk signaling in pluripotent cells”, *Development*, vol. 137, no. 20, pp. 3351–3360, 2010.

- [46] L. Sui, L. Bouwens, and J. K. Mfopou, “Signaling pathways during maintenance and definitive endoderm differentiation of embryonic stem cells”, *International Journal of Developmental Biology*, vol. 57, no. 1, pp. 1–12, 2013.
- [47] L. Vallier, T. Touboul, Z. Chng, M. Brimpari, N. Hannan, E. Millan, L. E. Smithers, M. Trotter, P. Rugg-Gunn, A. Weber, *et al.*, “Early cell fate decisions of human embryonic stem cells and mouse epiblast stem cells are controlled by the same signalling pathways”, *PloS one*, vol. 4, no. 6, e6082, 2009.
- [48] K. Hawkins, S. Joy, and T. McKay, “Cell signalling pathways underlying induced pluripotent stem cell reprogramming”, *World journal of stem cells*, vol. 6, no. 5, p. 620, 2014.
- [49] V. Valdespino, P. M. Valdespino, and V. V. Junior, “Genomic-epigenomic signaling pathways changes in cellular differentiation process”, *American Journal of Biomedical Research*, vol. 1, no. 2, pp. 35–42, 2013.
- [50] X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V. B. Vega, E. Wong, Y. L. Orlov, W. Zhang, J. Jiang, *et al.*, “Integration of external signaling pathways with the core transcriptional network in embryonic stem cells”, *Cell*, vol. 133, no. 6, pp. 1106–1117, 2008.
- [51] K. Woltjen and W. L. Stanford, “Inhibition of tgf- β signaling improves mouse fibroblast reprogramming”, *Cell stem cell*, vol. 5, no. 5, pp. 457–458, 2009.
- [52] L. Mishra, R. Derynck, and B. Mishra, “Transforming growth factor- β signaling in stem cells and cancer”, *Science*, vol. 310, no. 5745, pp. 68–71, 2005.
- [53] G. Ficiz, T. A. Hore, F. Santos, H. J. Lee, W. Dean, J. Arand, F. Krueger, D. Oxley, Y.-L. Paul, J. Walter, *et al.*, “Fgf signaling inhibition in escs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency”, *Cell stem cell*, vol. 13, no. 3, pp. 351–359, 2013.
- [54] U. Blank, G. Karlsson, and S. Karlsson, “Signaling pathways governing stem-cell fate”, *Blood*, vol. 111, no. 2, pp. 492–503, 2008.
- [55] P. Hayward, T. Kalmar, and A. M. Arias, “Wnt/notch signalling and information processing during development”, *Development*, vol. 135, no. 3, pp. 411–424, 2008.
- [56] Z. Li, T. Fei, J. Zhang, G. Zhu, L. Wang, D. Lu, X. Chi, Y. Teng, N. Hou, X. Yang, *et al.*, “Bmp4 signaling acts via dual-specificity phosphatase 9 to control erk activity in mouse embryonic stem cells”, *Cell stem cell*, vol. 10, no. 2, pp. 171–182, 2012.
- [57] S. Xiao, D. Xie, X. Cao, P. Yu, X. Xing, C.-C. Chen, M. Musselman, M. Xie, F. D. West, H. A. Lewin, *et al.*, “Comparative epigenomic annotation of regulatory dna”, *Cell*, vol. 149, no. 6, pp. 1381–1392, 2012.

- [58] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.
- [59] S. Huang, “The molecular and mathematical basis of waddington’s epigenetic landscape: A framework for post-darwinian biology?”, *Bioessays*, vol. 34, no. 2, pp. 149–157, 2012.
- [60] R. Morris, I. Sancho-Martinez, T. O. Sharpee, and J. C. I. Belmonte, “Mathematical approaches to modeling development and reprogramming”, *Proceedings of the National Academy of Sciences*, vol. 111, no. 14, pp. 5076–5082, 2014.
- [61] M. L. Simpson, C. D. Cox, and G. S. Sayler, “Frequency domain chemical langevin analysis of stochasticity in gene transcriptional regulation”, *Journal of theoretical biology*, vol. 229, no. 3, pp. 383–394, 2004.
- [62] D. T. Gillespie, “The chemical langevin equation”, *The Journal of Chemical Physics*, vol. 113, no. 1, pp. 297–306, 2000.
- [63] R. Khanin and D. J. Higham, “Chemical master equation and langevin regimes for a gene transcription model”, *Theoretical Computer Science*, vol. 408, no. 1, pp. 31–40, 2008.
- [64] H. Risken, “Fokker-planck equation”, in *The Fokker-Planck Equation*, Springer, 1996, pp. 63–95.
- [65] C. Li and J. Wang, “Quantifying cell fate decisions for differentiation and reprogramming of a human stem cell network: Landscape and biological paths”, *PLoS computational biology*, vol. 9, no. 8, e1003165, 2013.
- [66] J. X. Zhou, M. Aliyu, E. Aurell, and S. Huang, “Quasi-potential landscape in complex multi-stable systems”, *Journal of the Royal Society Interface*, vol. 9, no. 77, pp. 3539–3553, 2012.
- [67] A. Saadatpour and R. Albert, “Boolean modeling of biological regulatory networks: A methodology tutorial”, *Methods*, vol. 62, no. 1, pp. 3–12, 2013.
- [68] G. Karlebach and R. Shamir, “Modelling and analysis of gene regulatory networks”, *Nature reviews. Molecular cell biology*, vol. 9, no. 10, p. 770, 2008.
- [69] N. J. Eungdamrong and R. Iyengar, “Modeling cell signaling networks”, *Biology of the Cell*, vol. 96, no. 5, pp. 355–362, 2004.
- [70] S. R. Neves and R. Iyengar, “Modeling of signaling networks”, *Bioessays*, vol. 24, no. 12, pp. 1110–1117, 2002.
- [71] J. Narula, C. Williams, A. Tiwari, J. Marks-Bluth, J. E. Pimanda, and O. A. Igoshin, “Mathematical model of a gene regulatory network reconciles effects of genetic perturbations on hematopoietic stem cell emergence”, *Developmental biology*, vol. 379, no. 2, pp. 258–269, 2013.

- [72] M. Flöttmann, T. Scharp, and E. Klipp, “A stochastic model of epigenetic dynamics in somatic cell reprogramming”, *Frontiers in physiology*, vol. 3, 2012.
- [73] X. Lu, P. R. Kensche, M. A. Huynen, and R. A. Notebaart, “Genome evolution predicts genetic interactions in protein complexes and reveals cancer drug targets”, *Nature communications*, vol. 4, 2013.
- [74] P. Wang, C. Song, H. Zhang, Z. Wu, X.-J. Tian, and J. Xing, “Epigenetic state network approach for describing cell phenotypic transitions”, *Interface focus*, vol. 4, no. 3, p. 20130068, 2014.
- [75] V. Chickarmane and C. Peterson, “A computational model for understanding stem cell, trophoderm and endoderm lineage determination”, *PLoS one*, vol. 3, no. 10, e3478, 2008.
- [76] J. F. Rabajante and A. L. Babierra, “Branching and oscillations in the epigenetic landscape of cell-fate determination”, *Progress in biophysics and molecular biology*, vol. 117, no. 2, pp. 240–249, 2015.
- [77] S. Kauffman, C. Peterson, B. Samuelsson, and C. Troein, “Random boolean network models and the yeast transcriptional network”, *Proceedings of the National Academy of Sciences*, vol. 100, no. 25, pp. 14796–14799, 2003.
- [78] A. MacNamara, C. Terfve, D. Henriques, B. P. Bernabé, and J. Saez-Rodriguez, “State-time spectrum of signal transduction logic models”, *Physical biology*, vol. 9, no. 4, p. 045003, 2012.
- [79] A. Fauré and D. Thieffry, “Logical modelling of cell cycle control in eukaryotes: A comparative study”, *Molecular BioSystems*, vol. 5, no. 12, pp. 1569–1581, 2009.
- [80] B. B. Aldridge, J. Saez-Rodriguez, J. L. Muhlich, P. K. Sorger, and D. A. Lauffenburger, “Fuzzy logic analysis of kinase pathway crosstalk in tnfr/egf/insulin-induced signaling”, *PLoS computational biology*, vol. 5, no. 4, e1000340, 2009.
- [81] Z. Zi, “Mathematical modeling and kinetic analysis of cellular signaling pathways”, PhD thesis, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät I, 2008.
- [82] E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach, *Systems biology in practice: concepts, implementation and application*. John Wiley & Sons, 2008.
- [83] K. Raza and R. Parveen, “Evolutionary algorithms in genetic regulatory networks model”, *arXiv preprint arXiv:1205.1986*, 2012.
- [84] C. Gawad, W. Koh, and S. R. Quake, “Single-cell genome sequencing: Current state of the science”, *Nature reviews. Genetics*, vol. 17, no. 3, p. 175, 2016.

- [85] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, *et al.*, “Mrna-seq whole-transcriptome analysis of a single cell”, *Nature methods*, vol. 6, no. 5, pp. 377–382, 2009.
- [86] S. Picelli, Å. K. Björklund, O. R. Faridani, S. Sagasser, G. Winberg, and R. Sandberg, “Smart-seq2 for sensitive full-length transcriptome profiling in single cells”, *Nature methods*, vol. 10, no. 11, p. 1096, 2013.
- [87] T. Hashimshony, F. Wagner, N. Sher, and I. Yanai, “Cel-seq: Single-cell rna-seq by multiplexed linear amplification”, *Cell reports*, vol. 2, no. 3, pp. 666–673, 2012.
- [88] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, *et al.*, “Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets”, *Cell*, vol. 161, no. 5, pp. 1202–1214, 2015.
- [89] D. R. Bandura, V. I. Baranov, O. I. Ornatsky, A. Antonov, R. Kinach, X. Lou, S. Pavlov, S. Vorobiev, J. E. Dick, and S. D. Tanner, “Mass cytometry: Technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry”, *Anal. Chem*, vol. 81, no. 16, pp. 6813–6822, 2009.
- [90] S. C. Bendall, E. F. Simonds, P. Qiu, D. A. El-ad, P. O. Krutzik, R. Finck, R. V. Bruggner, R. Melamed, A. Trejo, O. I. Ornatsky, *et al.*, “Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum”, *Science*, vol. 332, no. 6030, pp. 687–696, 2011.
- [91] O. Ornatsky, D. Bandura, V. Baranov, M. Nitz, M. A. Winnik, and S. Tanner, “Highly multiparametric analysis by mass cytometry”, *Journal of immunological methods*, vol. 361, no. 1, pp. 1–20, 2010.
- [92] S. C. Bendall, G. P. Nolan, M. Roederer, and P. K. Chattopadhyay, “A deep profiler’s guide to cytometry”, *Trends in immunology*, vol. 33, no. 7, pp. 323–332, 2012.
- [93] M. H. Spitzer and G. P. Nolan, “Mass cytometry: Single cells, many features”, *Cell*, vol. 165, no. 4, pp. 780–791, 2016.
- [94] J. J. Hopfield, “Neural networks and physical systems with emergent collective computational abilities (associative memory/parallel processing/categorization/content-addressable memory/fail-soft devices)”, *Biophysics*, vol. 79, pp. 2554–2558, 1982.
- [95] K. A. Smith and J. N. Gupta, “Neural networks in business: Techniques and applications for the operations researcher”, *Computers & Operations Research*, vol. 27, no. 11, pp. 1023–1044, 2000.
- [96] J. T. Connor, R. D. Martin, and L. E. Atlas, “Recurrent neural networks and robust time series prediction”, *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 240–254, 1994.

- [97] A. J. Tatem, H. G. Lewis, P. M. Atkinson, and M. S. Nixon, “Super-resolution land cover pattern prediction using a Hopfield neural network”, *Remote Sensing of Environment*, vol. 79, no. 1, pp. 1–14, 2002, ISSN: 00344257. DOI: 10.1016/S0034-4257(01)00229-2.
- [98] D. Stauffer, A. Aharony, L. Da Fontoura Costa, and J. Adler, “Efficient Hopfield pattern recognition on a scale-free neural network”, *European Physical Journal B*, vol. 32, no. 3, pp. 395–399, 2003, ISSN: 14346028. DOI: 10.1140/epjb/e2003-00114-7. arXiv: 0212601 [cond-mat].
- [99] D. A. Levin, M. J. Luczak, and Y. Peres, “Glauber dynamics for the mean-field ising model: Cut-off, critical power law, and metastability”, *Probability Theory and Related Fields*, vol. 146, no. 1, pp. 223–265, 2010.
- [100] J. J. Hopfield, “Biophysics Neurons with graded response have collective computational properties like those of two-state neurons”, vol. 81, pp. 3088–3092, 1984.
- [101] H. Zhang, Z. Wang, and D. Liu, “A Comprehensive Review of Stability Analysis of Continuous-Time Recurrent Neural Networks”, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 7, pp. 1229–1262, Jul. 2014, ISSN: 2162-237X. DOI: 10.1109/TNNLS.2014.2317880.
- [102] J. Fujikura, E. Yamato, S. Yonemura, K. Hosoda, S. Masui, K. Nakao, J.-i. Miyazaki, and H. Niwa, “Differentiation of embryonic stem cells is induced by gata factors”, *Genes & development*, vol. 16, no. 7, pp. 784–789, 2002.
- [103] C. A. Gifford, M. J. Ziller, H. Gu, C. Trapnell, J. Donaghey, A. Tsankov, A. K. Shalek, D. R. Kelley, A. A. Shishkin, R. Issner, *et al.*, “Transcriptional and epigenetic dynamics during specification of human embryonic stem cells”, *Cell*, vol. 153, no. 5, pp. 1149–1163, 2013.
- [104] T. Kouzarides, “Chromatin modifications and their function”, *Cell*, vol. 128, no. 4, pp. 693–705, 2007.
- [105] H Lopez-Bertoni, B Lal, A Li, M Caplan, H Guerrero-Cazares, C. Eberhart, A Quinones-Hinojosa, M Glas, B Scheffler, J Laterra, *et al.*, “Dnmt-dependent suppression of microrna regulates the induction of gbm tumor-propagating phenotype by oct4 and sox2”, *Oncogene*, vol. 34, no. 30, pp. 3994–4004, 2015.
- [106] C. Richard, C. Drevon, P.-Y. Canto, G. Villain, K. Bollérot, A. Lempereur, M.-A. Teillet, C. Vincent, C. R. Castillo, M. Torres, *et al.*, “Endothelio-mesenchymal interaction controls runx1 expression and modulates the notch pathway to initiate aortic hematopoiesis”, *Developmental cell*, vol. 24, no. 6, pp. 600–611, 2013.

- [107] J. Zhang and L. Li, “Bmp signaling and stem cell regulation”, *Developmental biology*, vol. 284, no. 1, pp. 1–11, 2005.
- [108] D. A. Loebel, C. M. Watson, R. A. De Young, and P. P. Tam, “Lineage choice and differentiation in mouse embryos and embryonic stem cells”, *Developmental biology*, vol. 264, no. 1, pp. 1–14, 2003.
- [109] M. Koutsourakis, A. Langeveld, R. Patient, R. Beddington, and F. Grosveld, “The transcription factor gata6 is essential for early extraembryonic development”, *Development*, vol. 126, no. 4, pp. 723–732, 1999.
- [110] A. G. Gonzalez, A. Naldi, L. Sanchez, D. Thieffry, and C. Chaouiya, “Ginsim: A software suite for the qualitative modelling, simulation and analysis of regulatory networks”, *Biosystems*, vol. 84, no. 2, pp. 91–100, 2006.
- [111] N. D. Lawrence and J. Quiñonero-Candela, “Local distance preservation in the gp-lvm through back constraints”, in *Proceedings of the 23rd international conference on Machine learning*, ACM, 2006, pp. 513–520.
- [112] N. Lawrence, “Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models”, *Journal of Machine Learning Research*, vol. 6, pp. 1783–1816, 2005.
- [113] J. M. Wang, D. J. Fleet, and A. Hertzmann, “Gaussian process dynamical models for human motion”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008, ISSN: 01628828. DOI: 10.1109/TPAMI.2007.1167.
- [114] A. Yao, J. Gall, L. V. Gool, and R. Urtasun, “Learning probabilistic non-linear latent variable models for tracking complex activities”, in *Advances in Neural Information Processing Systems*, 2011, pp. 1359–1367.
- [115] V. Chelliah, N. Juty, C. Laibe, H. Hermjakob, and N. L. Novère, “Biomodels database: A public repository for sharing models of biological processes”, *Encyclopedia of Computational Neuroscience*, pp. 389–393, 2015.
- [116] A. Funahashi, Y. Matsuoka, A. Jouraku, M. Morohashi, N. Kikuchi, and H. Kitano, “Celldesigner 3.5: A versatile modeling tool for biochemical networks”, *Proceedings of the IEEE*, vol. 96, no. 8, pp. 1254–1265, 2008.
- [117] D. J. Higham, “An algorithmic introduction to numerical simulation of stochastic differential equations”, *SIAM review*, vol. 43, no. 3, pp. 525–546, 2001.
- [118] D. T. Gillespie, “A general method for numerically simulating the stochastic time evolution of coupled chemical reactions”, *Journal of computational physics*, vol. 22, no. 4, pp. 403–434, 1976.

- [119] J. R. Dormand and P. J. Prince, “A family of embedded runge-kutta formulae”, *Journal of computational and applied mathematics*, vol. 6, no. 1, pp. 19–26, 1980.
- [120] T. Schaffter, “Libsde user manual”, 2012.
- [121] S. Ramsey, D. Orrell, and H. Bolouri, “Dizzy: Stochastic simulation of large-scale genetic regulatory networks (supplementary material)”, *Journal of bioinformatics and computational biology*, vol. 3, no. 02, pp. 437–454, 2005.
- [122] N. Gopalan and I. T. Darmstadt, “Gaussian process latent variable models for dimensionality reduction and time series modeling”,
- [123] C. Li and J. Wang, “Quantifying the landscape for development and cancer from a core cancer stem cell circuit”, *Cancer research*, vol. 75, no. 13, pp. 2607–2618, 2015.
- [124] A. Ay and D. N. Arnosti, “Mathematical modeling of gene expression: A guide for the perplexed biologist”, *Critical reviews in biochemistry and molecular biology*, vol. 46, no. 2, pp. 137–151, 2011.
- [125] K.-C. Chen, T.-Y. Wang, H.-H. Tseng, C.-Y. F. Huang, and C.-Y. Kao, “A stochastic differential equation model for quantifying transcriptional regulatory network in *Saccharomyces cerevisiae*”, *Bioinformatics*, vol. 21, no. 12, pp. 2883–2890, Jun. 2005, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti415.
- [126] T. Kalmar, C. Lim, P. Hayward, S. Muñoz-Descalzo, J. Nichols, J. Garcia-Ojalvo, and A. M. Arias, “Regulated fluctuations in Nanog expression mediate cell fate decisions in embryonic stem cells”, *PLoS Biology*, 2009, ISSN: 15449173. DOI: 10.1371/journal.pbio.1000149.
- [127] P. Baldi, “Gradient Descent Learning Algorithm Overview: A General Dynamical Systems Perspective”, *IEEE Transactions on Neural Networks*, vol. 6, no. 1, pp. 182–195, 1995, ISSN: 19410093. DOI: 10.1109/72.363438.
- [128] S. Ruder, “An overview of gradient descent optimization algorithms”, *Web Page*, pp. 1–12, 2016. arXiv: 1609.04747.
- [129] J. Guo, F. Lin, X. Zhang, V. Tanavde, and J. Zheng, “Netland: Quantitative modeling and visualization of waddington’s epigenetic landscape using probabilistic potential”, *Bioinformatics*, vol. 33, no. 10, pp. 1583–1585, 2017.
- [130] J. a. Sethian, “Fast Marching Methods”, *SIAM Review*, vol. 41, no. 2, pp. 199–235, 1999, ISSN: 0036-1445. DOI: 10.1137/S0036144598347059.
- [131] G. Guo, M. Huss, G. Q. Tong, C. Wang, L. Li Sun, N. D. Clarke, and P. Robson, “Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyst”, *Developmental Cell*, 2010, ISSN: 15345807. DOI: 10.1016/j.devcel.2010.02.012.

- [132] Q. Deng, D. Ramsköld, B. Reinius, R. Sandberg, D. Ramsköld, B. Reinius, and R. Sandberg, “Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells”, *Science*, vol. 343, no. 6167, pp. 193–196, 2014, ISSN: 0036-8075. DOI: 10.1126/science.1245316.
- [133] L. Yan, M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, *et al.*, “Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells”, *Nature structural & molecular biology*, vol. 20, no. 9, pp. 1131–1139, 2013.
- [134] S. Islam, U. Kjällquist, A. Moliner, P. Zajac, J.-B. Fan, P. Lönnerberg, and S. Linnarsson, “Characterization of the single-cell transcriptional landscape by highly multiplex rna-seq”, *Genome research*, vol. 21, no. 7, pp. 1160–1167, 2011.
- [135] I. Amit, M. Garber, N. Chevrier, A. P. Leite, Y. Donner, T. Eisenhaure, M. Guttman, J. K. Grenier, W. Li, O. Zuk, *et al.*, “Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses”, *Science*, vol. 326, no. 5950, pp. 257–263, 2009.
- [136] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn, “The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells”, *Nature biotechnology*, vol. 32, no. 4, pp. 381–386, 2014.
- [137] M. Zwiessle and N. D. Lawrence, “Topslam: Waddington Landscape Recovery for Single Cell Experiments”, 2016. DOI: 10.1101/057778.
- [138] E. Marco, R. L. Karp, G. Guo, P. Robson, A. H. Hart, L. Trippa, and G.-C. Yuan, “Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape”, *Proceedings of the National Academy of Sciences*, vol. 111, no. 52, E5643–E5650, 2014.
- [139] S. C. Bendall, K. L. Davis, E.-a. David Amir, M. D. Tadmor, E. F. Simonds, T. J. Chen, D. K. Shenfeld, G. P. Nolan, and D. Pe, “Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development”, *Cell*, vol. 157, no. 3, pp. 714–725, 2014.
- [140] M. Setty, M. D. Tadmor, S. Reich-Zeliger, O. Angel, T. M. Salame, P. Kathail, K. Choi, S. Bendall, N. Friedman, and D. Pe’er, “Wishbone identifies bifurcating developmental trajectories from single-cell data”, *Nature biotechnology*, vol. 34, no. 6, pp. 637–645, 2016.
- [141] L. Haghverdi, F. Büttner, and F. J. Theis, “Diffusion maps for high-dimensional single-cell analysis of differentiation data”, *Bioinformatics*, vol. 31, no. 18, pp. 2989–2998, 2015.

- [142] B. Sozen, A. Can, and N. Demir, “Cell fate regulation during preimplantation development: A view of adhesion-linked molecular interactions”, *Developmental Biology*, vol. 395, no. 1, pp. 73–83, 2014, ISSN: 1095564X. DOI: 10.1016/j.ydbio.2014.08.028. arXiv: arXiv:1011.1669v3.
- [143] G. Martello, T. Sugimoto, E. Diamanti, A. Joshi, R. Hannah, S. Ohtsuka, B. Göttgens, H. Niwa, and A. Smith, “Esrrb is a pivotal target of the gsk3/tcf3 axis regulating embryonic stem cell self-renewal”, *Cell stem cell*, vol. 11, no. 4, pp. 491–504, 2012.
- [144] Y. Li, J. McClintick, L. Zhong, H. J. Edenberg, M. C. Yoder, and R. J. Chan, “Murine embryonic stem cell differentiation is promoted by soxs-3 and inhibited by the zinc finger transcription factor klf4”, *Blood*, vol. 105, no. 2, pp. 635–637, 2005.
- [145] I. Kehat, D. Kenyagin-Karsenti, M. Snir, H. Segev, M. Amit, A. Gepstein, E. Livne, O. Binah, J. Itskovitz-Eldor, and L. Gepstein, “Human embryonic stem cells can differentiate into myocytes with structural and functional properties of cardiomyocytes”, *The Journal of clinical investigation*, vol. 108, no. 3, pp. 407–414, 2001.
- [146] P. Blakeley, N. M. Fogarty, I. Del Valle, S. E. Wamaitha, T. X. Hu, K. Elder, P. Snell, L. Christie, P. Robson, and K. K. Niakan, “Defining the three cell lineages of the human blastocyst by single-cell rna-seq”, *Development*, vol. 142, no. 18, pp. 3151–3165, 2015.
- [147] C. Keysers and D. I. Perrett, “Demystifying social cognition: A hebbian perspective”, *Trends in cognitive sciences*, vol. 8, no. 11, pp. 501–507, 2004.
- [148] J. Fan, N. Salathia, R. Liu, G. E. Kaeser, Y. C. Yung, J. L. Herman, F. Kaper, J.-B. Fan, K. Zhang, J. Chun, *et al.*, “Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis”, *Nature methods*, vol. 13, no. 3, pp. 241–244, 2016.
- [149] F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle, “Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells”, *Nature biotechnology*, vol. 33, no. 2, pp. 155–160, 2015.
- [150] T. Kalisky and S. R. Quake, “Single-cell genomics”, *Nature methods*, vol. 8, no. 4, pp. 311–314, 2011.
- [151] C. Ziegenhain, B. Vieth, S. Parekh, B. Reinius, M. Smets, H. Leonhardt, I. Hellmann, and W. Enard, “Comparative analysis of single-cell rna sequencing methods”, *bioRxiv*, p. 035758, 2016.
- [152] T. Schlitt and A. Brazma, “Current approaches to gene regulatory network modelling”, *BMC bioinformatics*, vol. 8, no. 6, S9, 2007.

- [153] E. Bartocci and P. Lió, “Computational modeling, formal analysis, and tools for systems biology”, *PLoS computational biology*, vol. 12, no. 1, e1004591, 2016.
- [154] K. Sriyudthsak, F. Shiraishi, and M. Y. Hirai, “Mathematical modeling and dynamic simulation of metabolic reaction systems using metabolome time series data”, *Frontiers in molecular biosciences*, vol. 3, 2016.
- [155] M. Zinkevich, M. Weimer, L. Li, and A. J. Smola, “Parallelized stochastic gradient descent”, in *Advances in neural information processing systems*, 2010, pp. 2595–2603.