TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Xiaofeng Dai

Data Fusion Methods and an Application on Exploration of Gene Regulatory Mechanisms

Julkaisu 866 • Publication 866

Tampere 2010

Tampereen teknillinen yliopisto. Julkaisu 866
Tampere University of Technology. Publication 866

Xiaofeng Dai

Data Fusion Methods and an Application on Exploration of Gene Regulatory Mechanisms

Thesis for the degree of Doctor of Philosophy to be presented with due permission for public examination and criticism in Tietotalo Building, Auditorium TB104, at Tampere University of Technology, on the 12th of January 2010, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2010

ISBN 978-952-15-2299-4 (printed)
ISBN 978-952-15-2312-0 (PDF)
ISSN 1459-2045

Abstract

Understanding the regulatory mechanisms of gene regulatory networks (GRN) is an important topic in the field of Systems Biology. It has been widely accepted that holistic approaches are needed to explore biological systems given, for example, the noisy dynamics of gene expression and the complex interactions between genes and between gene expression products and other cellular components. As new advanced high throughput technologies emerge, i.e., as more information sources become available, thorough investigation of this problem is becoming feasible to be addressed from multiple perspectives.

The main objective of this thesis is to provide solutions to problems related to gene regulatory mechanisms with data fusion methods, aiming at a more precise understanding of a GRN's structure and its dynamics. This thesis can be divided into two parts: the presentation of the new data fusion methods here proposed to explore GRNs' topologies and, subsequently, the application of one method to investigate the dynamics of such networks.

In the 'Methods' chapter, two methods are proposed: one for transcription factor binding sites (TFBS) prediction and the other for gene clustering. The results from TFBS prediction can be used as an input for the gene clustering algorithm. Particularly, a new data fusion method is developed and novel information sources are explored to improve TFBS prediction accuracy in comparison with previous methods. Three finite joint mixture models are developed to cluster genes from multiple data sources: the beta-Gaussian mixture model (BGMM), the stratified beta-Gaussian mixture model (sBGMM) and the Gaussian-Bernoulli mixture model (GBMM). These methods are shown to significantly improve the accuracy of TFBS predictions and clustering results.

In the 'Application' chapter, one of the developed methods is applied to detect noisy attractors in delayed stochastic models of GRNs. The detection of noisy attractors is carried out for a model of a genetic toggle switch (TS) and for a model of an excitable genetic circuit of *Bacillus subtilis* responsible for phenotypic changes, by fusing multiple data sources extracted from the dynamics of the corresponding GRN. The results suggest that resorting to a single data source alone is, in general, insufficient to reveal the underlying structure of the GRN or to capture the changes in the dynamics of a GRN

modeled according to the delayed stochastic framework.

In summary, this thesis focuses on developing and applying data fusion methods to explore the topology and dynamics of a GRN, including TFBS prediction, gene clustering and noisy attractor detection. The developed algorithms and strategies are applicable to investigate real biological phenomena, and the findings can be used to guide future wet- or dry-lab experiments.

Preface

The work presented in this thesis was carried out at the Computational System Biology Group in the Department of Signal Processing of the Tampere University of Technology during 08/2007 to 08/2009.

I would like to address my deepest gratitude to Prof. Olli Yli-Harja for offering me this wonderful opportunity to touch the cutting-edge research programmes across multiple disciplines, and providing me with two strong supervisors to help me fulfilling my research. Further, his cordial advices on my academic career deserve my devout thankfulness. As my first supervisor, Prof. Harri Lähdesmäki introduced me to the world of Computational System Biology and gave me lots of supports and guidances during the past years, for which I would like to express my warmest acknowledgement to him wholeheartedly. Also, I owe a tremendous debt of thanks to assistant Prof. Andre S. Ribeiro, my current supervisor, for his numerous help, precious suggestions, and endless care, both in my academic research and personal growth. In addition, I'm extremely grateful to my two reviewers, Dr. Sampsa Hautaniemi and Dr. Andreas Beyer, for their pertinent and constructive comments given to my thesis.

Meanwhile, I would like to express my special thanks to the coauthors, Timo Erkkilä, M.Sc., and Dr. Shannon Healy, for their efforts and contributions to our work. Also, sincere appreciations are recorded to Dr. Kirsi Rautajoki for her help on improving my thesis, and to Dr. Reija Autio for providing me with the thesis template as well as some other supports.

I would like to acknowledge cordially the Tampere Graduate School in Information Science and Engineering (TISE) for its financial support of this research, with special gratitude goes to Dr. Pertti Koivisto, the coordinator of TISE, for his considerable help and advices besides the duty.

Also, I wish to thank honestly the department sectaries Virve Larmila and Kirsi Järnström, and the coordinators Elina Orava and Ulla Siltaloppi, for their kind help on all the miscellaneous problems I'v encountered during my studies here.

In addition, I will never forget the delightful moments I spent with my friends, which make my life full of joys and happinesses. I will always remember the comforts paid by my companions when I was frustrated, which help me go through many down moments. To them, I have nothing but grateful-

ness and would like to devote my sincere gratitude with all my heart.

Last but not least, I would like to tender my eternal thanks to my families and relatives, no matter near or far, live or dead, for all their ceaseless supports and blessings. Special acknowledgement goes to my dear husband, Bin Hong, without whose love, care, understanding and encouragement, I would never be so determined and devoted, and the work would not have been completed and progressed so fast.

For all the people that have helped or supported me during my doctoral studies, I dedicate this thesis to them and may they be blessed.

Tampere, December 2009
Xiaofeng Dai

Abbreviations

AP-MS	affinity purification followed by mass spectrometry
AIC	Akaike information criterion
AIC3	modified Akaike information criterion
AUC	area under the curve
BIC	Bayesian information criterion
CDF	cumulative density function
cDNA	complementary DNA
ChIP	chromatin immunoprecipitation
DBD	DNA binding domain
DDE	Delayed differential equation
EM	expectation maximization
FCM	fuzzy C-means
FPR	false positive rate
GO	gene ontology
GRN	gene regulatory network
GTF	general transcription factor
HMM	hidden Markov model
ICL-BIC	integrated classification likelihood - Bayesian information criterion
MCMC	Markov chain Monte Carlo
MLE	maximum likelihood estimate
mRNA	messenger RNA
PAM	partitioning around medoids
PDF	probability density function
PPI	protein-protein interaction
PSFM	position specific frequency matrix
PTM	post-translational modification
PSWM	position specific weight matrix
RBP	RNA binding protein
RNAi	RNA interference
ROC	receiver operating characteristic
RRM	RNA recognition motif

SDE	stochastic differential equation
SIDD	stress induced duplex destabilization
siRNA	short interference RNA
SOM	self organizing map
SSA	stochastic simulation algorithm
SVM	support vector machine
TF	transcription factor
TFBS	transcription factor binding site
TLR	Toll-like receptor
TRED	transcriptional regulatory element database
TS	toggle switch
UTR	untranslated region
VOMM	variable order Markov Model
Y2H	yeast two-hybrid

Contents

List of Publications

This thesis is a compound based on the following seven publications. In the text, they are referred to as [Publication **I**], [Publication **II**], and so on.

List of Publications

I X.F. Dai, O. Yli-Harja and H. Lähdesmäki, "Incorporating DNA duplex stability and nucleosome positioning information into genome-level data fusion for transcription factor target prediction", *BMC System Biology*, 2009, submitted.

II X.F. Dai, T. Erkkilä, O. Yli-Harja and H. Lähdesmäki, "A joint finite mixture model for clustering genes from independent Gaussian and beta distributed data", *BMC Bioinformatics*, 2009, vol. 10, pp. 165.

III X.F. Dai, H. Lähdesmäki and O. Yli-Harja, "A stratified beta-Gaussian finite mixture model for clustering genes with multiple data sources", *International Journal On Advances in Life Sciences*, 2009, vol. 1, no. 1, pp. 14–25.

IV X.F. Dai and H. Lähdesmäki, "A unified probabilistic framework for clustering genes from gene expression and protein-protein interaction data", in *Proceedings of the Sixth International Workshop on Computational System Biology*, Aarhus, Denmark, 10–12 June 2009, pp. 31–34.

V A.S. Ribeiro, X.F. Dai and O. Yli-Harja, "Variability of the distribution of differentiation pathway choices regulated by a multipotent delayed stochastic switch", *Journal of Theoretical Biology*, 2009, vol. 260, no. 1, pp. 66–76.

VI X.F. Dai, O. Yli-Harja and A.S. Ribeiro, "Determining noisy attractors of delayed stochastic Gene Regulatory Networks from multiple data sources", *Bioinformatics*, 2009, vol. 25, no. 18, pp. 2362–2368.

VII X.F. Dai, S. Healy, O. Yli-Harja and A.S. Ribeiro, "Tuning cell differentiation patterns and single cell dynamics by regulating proteins' functionalities in a Toggle Switch", *Journal of Theoretical Biology*, 2009, vol. 261, no. 3, pp. 441–448.

The work included in this thesis are joint efforts with co-authors, where the author's contributions are described below.

In [Publication **I**], the author and H. Lähdesmäki co-designed the study. The author developed and implemented the methods, and analyzed the results. The author and H. Lähdesmäki co-wrote the manuscript.

In [Publication **II**], the author and H. Lähdesmäki co-designed the study and co-developed the methods. The author implemented the algorithms and did the performance tests. T. Erkkilä and H. Lähdesmäki derived the EM algorithms. The author wrote the manuscript with coauthors' help.

In [Publication **III**], the author designed the study, developed and implemented the method. Under the supervision of H. Lähdesmäki, the author derived the algorithm and wrote the manuscript.

In [Publication **IV**], the author and H. Lähdesmäki co-designed the study. The author developed the method, implemented the algorithms, and did the performance tests. Under the supervision of H. Lähdesmäki, the author derived the algorithm and wrote the manuscript.

In [Publication **V**], the author did the simulations, provided the results, and was involved in manuscript drafting.

In [Publication **VI**], the author co-designed this study with A.S. Ribeiro. The author derived the algorithm, did the simulations, involved in building the stochastic model of MeKS module, and analyzed the results. With much help and supervision of A.S. Ribeiro, the author drafted the paper.

In [Publication **VII**], the author designed the study, did the simulations. With the help of A.S. Ribeiro, the author analyzed the results. The author and S. Healy co-drafted the paper under A.S. Ribeiro's guidance.

Relevant work that are not included in this thesis

- X.F. Dai, H. Lähdesmäki and O. Yli-Harja, "BGMM: a Beta-Gaussian mixture model for clustering genes with multiple data sources", in *Proceedings of the Fifth international workshop on computational system biology*, Leipzig, Germany, 11–13 June 2008, pp. 25–28.
- X.F. Dai, H. Lähdesmäki and O. Yli-Harja, "sBGMM: a stratified Beta-Gaussian mixture model for clustering genes with multiple data sources", in *International Conference on Biocomputation, Bioinformatics, and Biomedical Technologies*, Bucharest, Romania, 29 June–5 July 2008, pp. 94–99.

Chapter 1

Introduction

A gene, typically composed of the regulatory and information coding DNA sequence regions [1], is not an independent inheritance unit in the genome. Instead, genes are organized in a network, called the 'gene regulatory network' (GRN) [2], to regulate one another's expression. In other words, it is the process of converting genotypes encoded by genes into phenotypes exhibited by their products such as various types of RNAs, proteins and protein complexes. Further, besides genes and their products, other molecules such as some metabolites may also contribute to gene expression and be involved in a GRN [3]. These multiple components collaborate in concordance to orchestrate numerous cellular events, such as transcription, translation, post-transcriptional or post-translational modifications, and signal transduction cascades. Thus, a GRN can be viewed as an intertangled regulatory circuit governing processes such as gene expression, signal transduction and metabolism [3]. Each step of any cellular event in a GRN is stochastic [4], which may cause non-neglectable consequences. For example, for genes which can only express one of their two copies such as olfactory receptor genes and antigen-specific receptors, the stochastic choice of the allele to be expressed may result in cells' phenotypic difference [5]. Also, despite the possible mutations, the genomes of certain types of cells may undertake random remodeling during the development such as the stochastic genetic recombination that forms different immunoglobulin molecules to fight with diverse antigens [6], which adds even more noises to the genome and the GRN.

Understanding the gene regulatory mechanisms of a GRN is one of the long-term goals of Systems Biology, to which enormous efforts have been devoted [7–10]. However, given the complex regulatory relationships among the multiple components of a GRN, and the stochasticity of the cellular events that contribute to the phenotypic variations of the cells, studying the gene regulatory mechanisms of GRNs with single data sources may not be sufficient to fully capture the characteristics of a GRN and reveal its true

regulatory nature. Thus, the author targets on understanding the topology and the dynamics of a GRN via integrating information from multiple data sources, i.e., exploring such problems from a higher dimensional space with multiple coordinates.

To do so, this thesis focuses on both the development of novel data fusion methods and their applications. In particular, the data fusion methods of two interconnected problems are developed, which are transcription factor binding site (TFBS) prediction, i.e., predicting the binding sites of a TF on a DNA sequence, and gene clustering, i.e., grouping genes with similar features together. TFBS prediction explores the gene regulatory mechanisms at the sequence and physical protein-DNA binding levels, which offers us detailed information on how TFs bind to the genes and the links between the TFs and their targets. Gene clustering, on the other hand, investigates the regulatory relationships among genes at the gene level, which provides us a global view on how genes interact with each other and work in a concert to regulate gene expression. TFBS prediction and gene clustering, although from different perspectives and at different levels, both reveal the regulatory relationships among genes and contribute to the understanding of a GRN's topology. Further, the output of TFBS prediction, which contains the probabilities of the genes being bound by a set of TFs, can be used as the input of gene clustering to study the genes' relationships regarding their given potential regulators. This is particularly important here since obtaining protein-DNA binding data from experimental techniques are limited in measuring TFBSs of all TFs, largely due to the difficulties in finding specific antibodies for TFs which are needed in chromatin immunoprecipitation (ChIP) related experiments. Thus, these two problems interwind with each other, and work together in a complementary fashion on the exploration of a GRN's topology. Besides the application of each data integration method in what they are originally developed for, the data fusion framework for gene clustering is also applied to study the dynamics of GRNs at a single cell and cell population levels. To be specific, the background and motivation of each of the three studied problems are described, separately, below.

Transcriptional processes are largely controlled by TFs that bind to gene regulatory elements in a sequence specific manner [11; 12]. Thus, correctly predicting the binding sites of a TF to its target genes can provide us the detailed information of how genes regulate one another at the sequence level. While novel experimental techniques for measuring protein-DNA binding specificities keep emerging [13–16], computational predictions are proven to be a good facilitation in unveiling TFBSs genome-wide [17–19]. However, relying on the sequence specificities alone, the fundamental basis of current computational methods, is insufficient to accurately predict TFBSs due to the high level of noises within the genome [19]. Lähdesmäki et al. developed an algorithm called ProbTF [19], which can predict TFBSs via integrat-

ing multiple data sources. In particular, they have explored evolutionary conservations, regulatory potentials and nucleosome positioning predictions from [20] using their algorithm, where no performance improvement is reported with the nucleosome positioning data they employed. This negative result may be associated with the integration method used and the quality of the data under study. Thus, it is necessary to see how much further the performance can be improved if a new data fusion principle is developed, the data of better quality is employed, and novel information sources are explored.

Functionally related genes may be regulated or regulate the other genes' expression in a similar fashion [21], and thus can be viewed as a block when studying the topology of a GRN. Therefore, gene clustering can facilitate as the first step towards understanding the regulatory relations among genes within a GRN. Among many genomic data, gene expression data has been widely used for this purpose, with the assumption that genes that share similar expression patterns have similar cellular functions and are likely to be involved in the same process [21]. This assumption has been challenged by many evidences, e.g., genes participating in different processes may share similar profiles, and patterns of functionally related genes may not be well correlated [22; 23]. This, however, can be compensated by, e.g., observing physical interactions such as protein-protein interactions [24; 25]. Thus, in order to gain a holistical view of genes' functional relationships, it is necessary to develop methods that can cluster genes from multiple data sources.

Another important goal of this thesis is to study the dynamics of a GRN. The number of a GRN's possible states is immense, far more than that of cell types. For example, even assuming that genes are either on or off, given a human has $30000 - 35000$ genes, the human genome may encode 2^{30000} to 2^{35000} states [26]; while, only around 411 distinct cell types exist in an adult human body [27]. Thus, cell types are most likely to be constrained patterns of genes' activities, i.e., the attractors of GRNs' dynamics [2]. However, due to high level of genome noise, real cells, strictly speaking, do not have attractors [28]. Thus, the concept of noisy attractor is proposed [29] to study the dynamics of a GRN. While techniques for finding noisy attractors are well-established in noisy Boolean networks [29], it is not a simple task under a delayed stochastic framework. In [29], noisy attractors were detected by binarizing (using K-means [30] clustering algorithm) protein time series of a delayed stochastic GRN, which may not capture the full richness of the GRN's dynamics due to the information loss caused by binarization. Also, given the regulatory role of some cellular components at other levels, e.g., miRNA can cause sooner degradation of mRNA in eukaryotes [31], observing a single data source alone may be insufficient to capture the behavior of a GRN. Thereby, jointly utilizing multiple data sources is critical in noisy

attractor detection, which is a novel and apropos problem to apply the data fusion method that is originally developed for gene clustering.

Taken together, with the goal of understanding the gene regulatory mechanisms of a GRN, the author investigates GRNs' topologies and dynamics by developing and applying data fusion methods. Specifically, the objectives of this thesis are summarized below.

- Developing efficient multiple data fusion method for TFBS prediction, and exploring novel information sources to improve the prediction accuracy.
- Developing suitable data fusion clustering framework to group genes from different data sources.
- Applying the developed data fusion clustering framework to detect noisy attractors of delayed stochastic GRNs and explore such networks' dynamics.

The three problems studied in this thesis are presented in two chapters, i.e., 'Methods' and 'Application'. TFBS prediction and gene clustering, which focus on the method development, are introduced in the 'Methods' chapter, and noisy attractor detection, which is a novel application of one of the data fusion methods developed, is put in the 'Application' chapter. Specifically, this thesis is organized as following.

- Chapter 1: introduces the motivation, objectives and outline of this thesis.
- Chapter 2: introduces the basis of the central topics covered by this thesis, i.e., the biological background of gene regulation and GRN.
- Chapter 3: presents the key concepts, data sources, algorithms and models that are encountered, explored, used and developed when developing the methods. In particular, ProbTF, the TFBS prediction algorithm used in [Publication **I**] and the data fusion method developed are introduced. Also, the data fusion framework and all the joint finite mixture models built (including models presented in [Publication **II**] to [Publication **IV**] and [Publication **VI**]) are summarized in a systematic way. This chapter concentrates on the methods employed and developed in TFBS prediction and gene clustering. The exploration of novel information sources ([Publication **I**]) in TFBS prediction, and the simulation test ([Publication **II**] to [Publication **IV**]) and real case application ([Publication **II**] and [Publication **III**]) of each clustering model are summarized in Chapter 5, whose details can be found in each publication.

- Chapter 4: presents the key concepts, data sources, algorithm and GRNs that are used or investigated in an application of the gene clustering framework introduced in Chapter 3, i.e., detecting noisy attractors of delayed stochastic GRNs. Specifically, the modeling strategies and delayed stochastic simulation algorithms which are used to build models and generate data are introduced. Further, the GRNs explored for noisy attractor detection are described. This chapter focuses on introducing the background of this application. The results and conclusions of each publication are summarized in Chapter 5, with details available in [Publication **V**] to [Publication **VII**].
- Chapter 5: summarizes the main results of the listed publications, draws conclusions and proposes the future directions.

Chapter 2

Biological Background

This chapter gives a brief overview of the background of gene regulation and gene regulatory networks (GRN), which are the focused problems of this thesis.

2.1 Central Dogma

The backbone of molecular biology is the central dogma (as shown in Fig. 2.1), which was first proposed by Francis Crick in 1958 [32], and restated in 1970 [33].

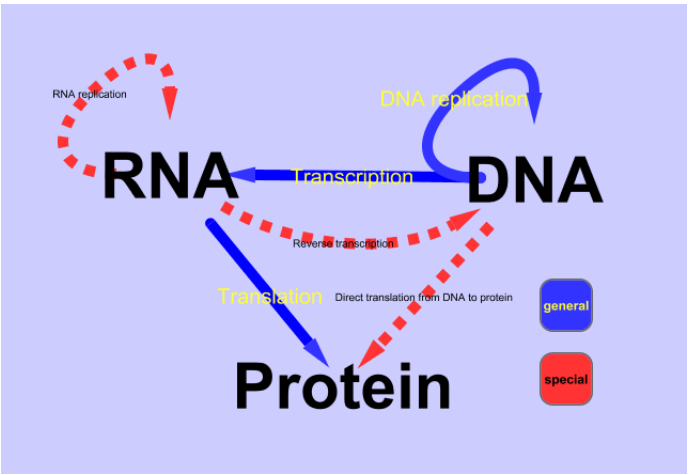


Figure 2.1: The central dogma of molecular biology. Figure is drawn using Cytoscape [34] based on [33].

There are three key information transfer stages according to the central dogma [33]:

- Replication: a double stranded DNA replicates itself, perpetuating the genetic information.
- Transcription: the genetic information is transferred from one DNA strand to a complementary RNA strand, called messenger RNA (mRNA).
- Translation: the genetic information is read by the ribosome as triplet codons, and transferred from mRNA to protein. In prokaryotic cells where there is no nucleolus, translation occurs simultaneously with transcription. In eukaryotic cells, mRNA must be transported into the cytoplasm to find the ribosome for translation to occur.

While, generally, information is transferred from DNA via mRNA to protein, some exceptions also exist, including reverse transcription (transferring information from RNA to DNA), RNA replication (RNA copying itself), and direct translation from DNA to protein [33]. Specifically, reverse transcription is reported to occur in retroviruses [35]. RNA replication and the direct translation from DNA to protein, which are known by hypothesis at the time the central dogma was enunciated [33], are found to exist in RNA viruses, such as Ebola virus [36], and are experimentally verified *in vitro*, e.g., using the extract from *E. coli* that contains ribosomes [37; 38], respectively.

Note that RNA includes many other types besides mRNA, such as ribosome RNA (rRNA) and transfer RNA (tRNA) [11; 12], but, in this thesis, it only refers to mRNA if no special claim is made.

2.2 Gene regulation

Definition 1 (Gene)**.** *Gene is a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions [1].*

Gene expression is the process that converts genotypes encoded by genes into phenotypes exhibited by gene products, where a gene product often refers to a protein and in some cases, such as for non-protein coding genes, can be an RNA (any type of RNA) [11]. Any step involved in gene expression may be regulated, and the regulation process can be stratified into at least four layers, i.e., transcriptional regulation, post-transcriptional regulation, translational regulation and post-translational regulation [11].

2.2.1 Transcriptional regulation

Transcriptional regulation refers to the process that regulates gene expression levels by altering the time for a transcription to occur and the amount of RNA that is produced [39]. This is the main regulatory mechanism in

prokaryotes, where promoter, operator and the protein encoding genes, organized as an operon, work in a concert to regulate themselves [40]. Transcriptional regulation is much more complex in eukaryotic cells, which typically involves one more level of regulation, i.e., chromosome packaging [41] via for example post-translationally modifying histones and regulating molecules involved in chromatin organization such as Polycomb and Trithorax proteins [42], two cis-acting elements other than the operator, i.e., enhancer and silencer, which may locate at varying points along the chromosome [40], and more number of trans-acting factors [40]. Further, the promoter of prokaryotes is bounded by the RNA polymerase (the enzyme for transcription) and initiates transcription; however, in eukaryotes, the transcription start site is separated from the promoter, and the promoter is recognized by transcription factors (TF) [40].

Among many trans-acting regulatory factors (such as TFs and coactivators), TFs are of the most interest, both in prokaryotes and eukaryotes, due to their universal existence and important regulatory roles [11; 12]. A TF is defined as a protein that controls the transcription of genetic information from DNA to RNA via binding to specific part(s) of DNA sequence(s) [43; 44]. One distinguishable feature of TFs compared with other trans-acting factors is that they contain one or more DNA binding domains (DBDs), which can attach them to specific DNA sequences such as the promoter [45; 46]. Correspondingly, the bounded DNA sequences are called transcription factor binding sites (TFBS) [47]. TFs perform their functions by promoting (as an activator) or blocking (as a repressor) the recruitment of RNA polymerase to specific genes, either alone or together with other proteins in the form of a protein complex [48–50]. In eukaryotes, another class of TFs, general TFs (GTFs), also exit, which do not activate or repress gene transcription but are necessary for the transcription to occur [50].

2.2.2 Post-transcriptional regulation

Post-transcriptional regulation is the process that controls gene expression by manipulating the RNA transcripts after RNA synthesis has begun [12; 51]. While it has long been accepted to exert important regulatory roles in eukaryotes [41], it is also found to exist in prokaryotes, basically by affecting mRNAs' stabilities [52]. Generally, regulation at this layer refers to the mechanisms occurring in eukaryotes, such as transcription attenuation, alternative splicing, RNA editing, nuclear transport and degradation [11; 12], thus the following text focuses on eukaryotes only. It is reported that differences at mRNA level only contribute to 20% to 40% proteins' concentration differences, indicating the importance of gene regulation after transcription [53; 54]. Further, studies on transcription, translation and protein turnover in yeast also suggest the significant role of

post-transcriptional regulation in controlling protein levels [7].

RNA binding protein (RBP) is the controlling factor that regulates the stability and distribution of different transcripts via controlling the steps and rates of various events involved in post-transcriptional regulation [51]. Similar with TFs, a RBP contains RNA recognition motif (RRM) that binds to a specific sequence or secondary structure, typically at the 5' and 3' untranslated region (UTR), of a transcript [51; 55]. Small RNAs can also post-transcriptionally regulate gene expression in many eukaryotes. The most well studied example would be RNA interference (RNAi), where siRNAs (short interfering RNA) induce the degradation of mRNAs [56].

2.2.3 Translational regulation

Translational regulation refers to the control of translation efficiency and is featured by the differential usage of mRNAs [11; 12]. This level of control exists in both prokaryotes and eukaryotes [57]. In prokaryotes, the known mechanisms include, e.g., controlling the initiation rate and programmed frame-shifting, and are shown to be important in many special cases [58], e.g., the differential choice of the translational initiation codon in the RNA of foot-and-mouth disease virus results in two different proteins with identical carboxy termini [59]. In eukaryotes, translational regulation can occur via, e.g., altering translation initiation rate and schemes [60], alternating translation elongation [60; 61] and modulating the length of poly(A) tail [60; 62], which is critical in controlling a variety of physiological processes in eukaryotic cells, such as cell differentiation, proliferation and self-protection [63].

Similar with transcription initiation, trans-acting factors are also important in translational regulation, e.g., translational repressors can stop translation via binding to the ribosome binding site in prokaryotes [58], and mRNA-specific initiation factors need to recognize and interact with the 5' and/or 3' UTR of a particular mRNA before the start of its translation in eukaryotes [60].

2.2.4 Post-translational regulation

Post-translational regulation refers to any process that affects the amount or activities of proteins after translation in eukaryotic cells [11; 12]. Regulation at this level is realized via either reversible events, i.e., post-translational modification (PTM), or irreversible events, such as proteolysis [11; 12].

PTM is the most common way for post-translational regulation, during which a protein undergoes specific chemical modifications [64]. Generally, PTMs can be viewed as three alternatives, which are attaching or removing other biochemical functional groups (such as phosphorylation, acylation [65], formylation [66], and glycation [67]), changing the chemical property of an

amino acid (e.g., citrullination [68], deamidation [69; 70], and eliminylation [71]), and making structural or length changes (such as forming disulfide bridges [69] and proteolytic cleavage [72]).

2.3 Gene regulatory network

Definition 2 (Gene Regulatory Network)**.** *A gene regulatory network is a set of highly interconnected processes that govern the rate at which different genes in a cell are expressed in time, space, and amplitude [3].*

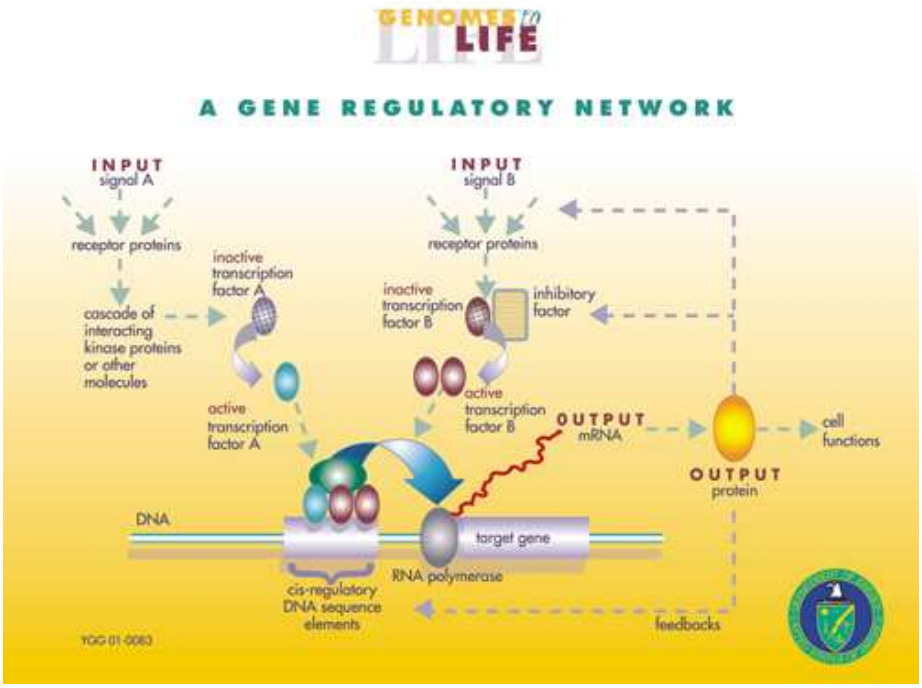
A typical scheme of a GRN is shown in Fig. 2.2 (a) [73], where TFs, responsive to the signal cascade caused by the external inputs, are the main players. Through the activated or inactivated responses of TFs, gene expression is up- or down- regulated, with the output signals affecting cell functions. A network can be modeled as static or dynamic, and its complexity and content may vary with time and space [3]. The whole control process is depicted in Fig. 2.2 (b) [73]. The consequences of a GRN can be viewed as primary outputs, i.e., RNAs and proteins, and terminal outputs, i.e., changes in the cell's phenotype and function, both of which in return act as the network's inputs (besides external signals) through the feedback circuitry. Hautaniemi et al. proposed a decision tree analysis approach to study the relationship between cell functional responses and extracellular signals, and found a joint role of multiple inputs in controlling cellular outputs by studying cell migration process [8].

Models of GRNs describe various aspects of the complex relationships among genes, their products and other cellular components, which can be considered over a wide range of systems, e.g., gene interaction networks, protein interaction networks, and signal transduction networks [3]. A simplified scheme of intracellular regulation circuits is illustrated by a bipartite graph in Fig. 2.3 [3]. It is shown that gene regulation largely intertwines with signal transduction (notice the significant overlap between Box I and Box II), and the process not only involves genes and their products but also requires metabolites.

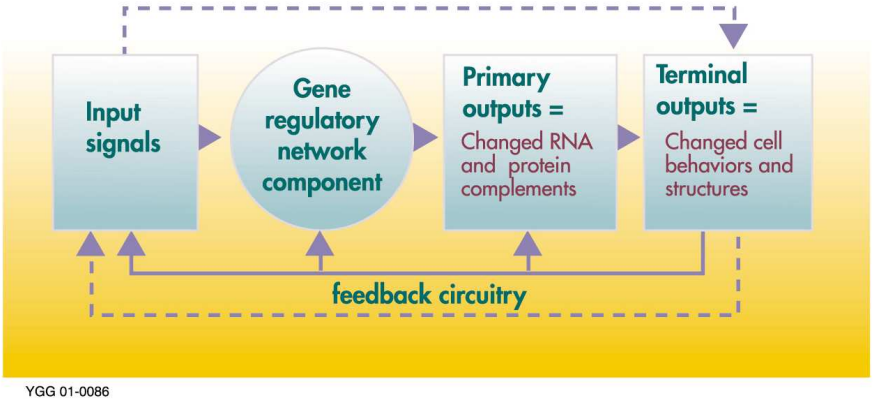
The collective information of a GRN is often extracted and represented as the network structure [74]. In such a structure, genes or gene products (e.g., proteins, RNAs, and protein complexes) are represented as nodes, and molecular interactions (i.e., one gene affects the other via its products) are symbolized as edges [74]. In directed graphs, an arrow is used to indicate the causal relationship between two nodes, and the shape of an arrow head is used to represent the regulatory effect, i.e., inductive or inhibitory, if such information is available [74]. Finally, the dependencies within a GRN are shown as a series of edges, with cycles illustrating feedback loops [74]. In practice, such a structure is often inferred from biological literature or

experimental evidence by certain modeling methods, whose results can be used, e.g., to make predictions or suggest new exploratory approaches.

Many modeling approaches have been used to model GRNs, such as Boolean networks [75], ordinary differential equations [76], Bayesian networks [77] and stochastic models [78].



(a)



(b)

Figure 2.2: (a) The structure and (b) the control process of a gene regulatory network. Dashed lines in (b) shows signaling responses which do not involve gene expression regulation but act directly on proteins or protein machine assemblies. Figures are retrieved from [73] with permission.

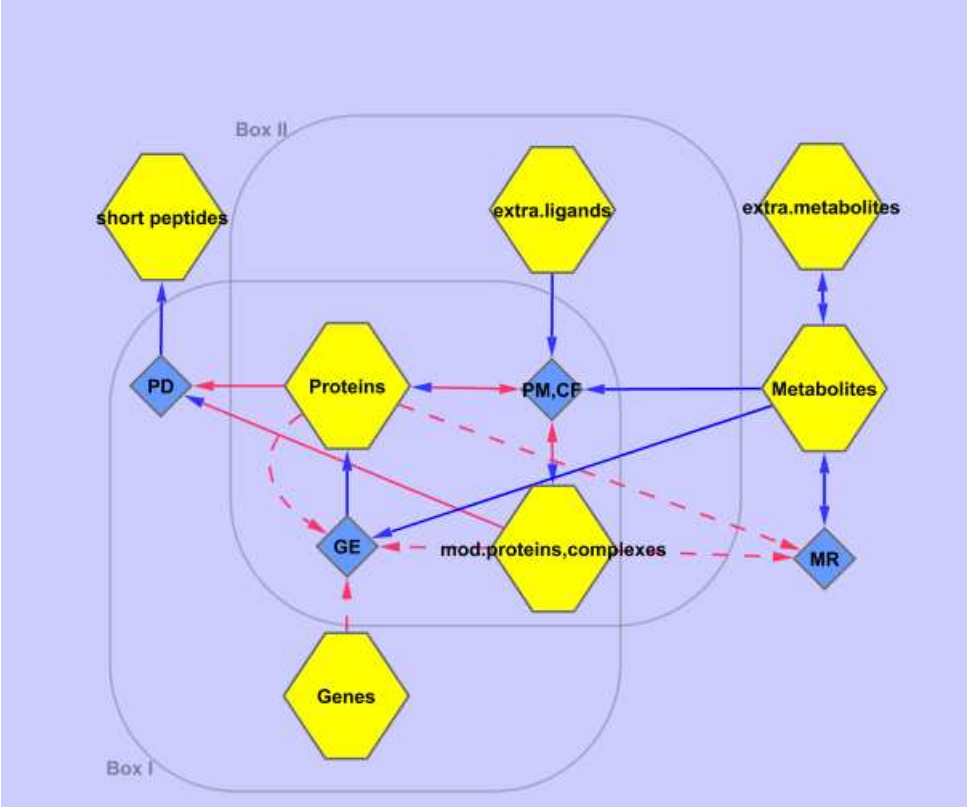


Figure 2.3: Simplified scheme of intracellular regulation circuits with gene expression (Box I), signal transduction (Box II), and metabolic processes (shown outside of the boxes on the righthand side). Yellow hexagons represent molecular entities, and blue diamonds stand for regulatory events. Molecular entities are genes ('Genes'), proteins ('Proteins'), modified proteins ('mod.proteins'), protein complexes ('complexes'), peptides ('short peptides'), extracellular metabolites ('extra.metabolites'), metabolites ('Metabolites'), and extracellular ligands ('extra.ligands'). Regulatory events are gene expression ('GE'), protein modification and complex formation ('PM, CF'), protein degradation ('PD'), and metabolic reactions ('MR'). Blue solid arrows represent the mass flow, red dashed arrows show the catalytic action of molecular entities on the corresponding regulatory event, and red solid arrows stand for both the mass flow and the catalytic event. Note that catalysts are themselves not consumed during the catalytic processes. This graph is drawn using Cytoscape [34] based on [3].

Chapter 3

Methods

A gene regulatory network's (GRN) structure reveals the regulatory relationships among genes, which is an important aspect in understanding a GRN's regulatory mechanisms.

This chapter focuses on exploring the topology of a GRN and, specifically, introduces the key concepts, data sources, algorithms, and models that are used or developed in [Publication **I**] to [Publication **IV**]. Studies discussed in this chapter involve two interconnected problems, transcription factor binding site (TFBS) prediction ([Publication **I**]) and gene clustering ([Publication **II**] to [Publication **IV**]), which investigate GRNs' topologies at sequence and gene levels, respectively, and the result of the first problem is used as one input of the second one.

3.1 TFBS prediction

Transcription factors (TF) recognize and bind to the promoters of their target genes, exerting roles such as activation or repression. Analyzing the binding sites of the TFs at their target genes reveals the links between TFs and their targets and offers us a detailed map of a GRN's topology at the sequence and protein-DNA physical binding level. TFBS analysis comprises of TFBS discovery and TFBS prediction, and this thesis puts its emphasis on TFBS prediction.

[Publication **I**] improves TFBS prediction accuracy by developing a new data fusion strategy and exploring two novel information sources. Besides the data fusion principle proposed, this section also introduces the key concepts, data sources, and the algorithm that the work is built on. The content for novel data source exploration and the prediction results after implementing the new data fusion method and using novel information sources are summarized in Chapter 5.

3.1.1 Key concepts

In [Publication **I**], the following key concepts are encountered.

Definition 3 (Transcription Factor Binding Site Discovery)**.** *Transcription factor binding site discovery, also called motif discovery, is a computational approach of transcription factor binding site analysis, which searches for novel binding motifs from a collection of short sequences that are assumed to contain a common regulatory motif [79].*

Definition 4 (Transcription Factor Binding Site Prediction)**.** *Transcription factor binding site prediction is a computational approach of transcription factor binding site analysis, which makes use of given transcription factors' DNA-binding specificities to predict putative transcription factor binding sites. Transcription factors' DNA-binding specificities can be either the output of a transcription factor discovery algorithm or experimentally measured [19].*

Definition 5 (Transcription Factor Binding Preference)**.** *Transcription factor binding preference is the preference of a transcription factor towards binding to single- or double-stranded DNA [Publication **I**].*

3.1.2 Data sources

TFBS prediction methods rely on specific sequence patterns which, although highly specific, may have both poor sensitivity and high false positive rate (FPR) when the patterns are degenerate [18; 19]. One way for improving TFBS prediction accuracy is to guide the algorithm via incorporating additional information [19]. In this thesis, additional data sources explored include evolutionary conservation, regulatory potential, nucleosome positioning and DNA duplex stability.

Evolutionary conservation data

Evolutionary conservation data stores information of conserved sequences across species [80]. It has been widely applied to find functional sequence motifs [81–84], with the rational that essential genes evolve more slowly than nonessential ones. Thereby, orthologous sequences that are significantly more similar than what is expected are likely to be functionally critical if they evolve under neutral evolution [85]. Many tools have been developed for multiple sequence alignment [86–88], making conservation study practically more feasible. Sequences that are predicted to be functional can either encode gene products or exert regulatory roles such as TFBSs [11]. Thus, besides stimulating new hypotheses and driving experimentation on gene function discovery [81; 82; 89; 90], conservation data also facilitates TFBS

analysis [19; 83; 84]. Although only $\sim$ 50% human regulatory sites are reported to be conserved in mouse [91; 92], evolutionary conservations are proven to be informative in discovering or predicting TFBSs [19; 83; 84].

Numerous computational algorithms are developed to compute conservation scores, via pair-wise [93] or multiple sequences [94–97] alignment. The evolutionary conservation data used in this thesis is obtained from phastCons [80], which predicts the conserved elements using the Viterbi algorithm and computes the conservation scores by the forward/backward algorithm, based on a two-state phylogenetic hidden Markov model.

Regulatory potential data

It is reported that many functional genomic elements are not conserved, and lots of constrained regions do not overlap with known functional elements [80; 98]. Thus, further improvement of TFBS prediction accuracy calls for information other than interspecies sequence conservation. Regulatory potentials, defined as the data that discriminate regulatory regions from neutral sites [99], can be used to assess whether a conserved sequence is functional or not.

ESPERR [99] is used to provide the regulatory potentials analyzed in this thesis. ESPERR first retrieves information from multiple genome alignments via appropriate dimension reduction and alphabet selection. Then it applies two variable-order Markov models (VOMM), trained from known regulatory and neutral sites, respectively, to estimate the likelihoods of a site being regulatory and neutral. Finally, the regulatory potential scores are computed as the log-odds based on the two VOMMs. Given the ability of measuring variable-length position dependencies and the usage of multiple sequence alignment, ESPERR is believed to be able to capture evolutionary patterns that span multiple sequences [99].

Nucleosome positioning data

Eukaryotic genomic DNA exists in a highly compact form, namely chromatin [100]. The chromatin is composed of nucleosomes, which are 147 base pairs (*bps*) DNA tightly wrapped around a histone protein octamer and linked by $10 - 50$ *bps* long unwrapped short DNA sequences (namely 'linker DNAs') [100]. Facilitated by specific dinucleotides, DNA sharply bends at every DNA helical repeat ($\sim$ 10 *bps*) when DNA's major groove faces towards the octamer, and $\sim$ 5 *bps* away to the opposite direction when the major groove faces reversely [20; 101]. It is reported that polymerase and complexes, e.g., used for regulatory, repair and recombination, are occluded from accessing wrapped DNAs buried in nucleosomes [20]. Thus, nucleosome locations may play important regulatory roles in gene expression, whose in-

trinsic genomic organization is hypothesized to guide the recognition process of TFs to their binding sites, i.e., TFs bind more easily to sites that are free of nucleosomes [20].

Genome-wide nucleosome positions have been experimentally identified with high resolution in yeast [102–104], *Caenorhabditis elegans* [105], *Drosophila* [106], and human [107; 108]. Also, four computational methods to compute the nucleosome occupancy probabilities for each sequence have been developed [20; 109–111]. The algorithms of [20] and [110] recognize the nucleosome sequences' patterns by counting the dinucleotide frequencies, against which matches are scanned across the genomic sequences. The method presented in [111] searches sequence patterns using k-mer enumeration (k from 1 to 6) from a training data set, and applies a support vector machine (SVM) to distinguish the nucleosome forming sequences from the background. While the methods of [20], [110] and [111] use direct information from nucleosome, the algorithm of [109] focuses on long-range sequence information. It uses wavelet transformation to extract periodic features of genomic sequences, among which those that are associated with nucleosome positioning are selected with a statistical model. It is reported that the methods presented in [109] and [111] perform similar, and are superior to the other two algorithms [109]. Thus, in order to see whether more accurate nucleosome positioning data could improve TFBS prediction, the data computed from [20] and [109] are compared in this thesis.

DNA duplex stability data

DNA is confined into the form of either a circular molecule or closed loops within chromosomes *in vivo* [11; 112; 113]. Constraints in both forms are precisely equivalent, with the loops formed by periodic attachments of the chromatin fiber to the nuclear matrix [11; 112; 113]. The number of times one strand winds around the other, namely the linking number, may be changed by transient strand breakage and religation, resulting in a linking difference which imposes DNA superhelicity on the domain [113]. It is reported that DNA superhelicity, a force driving the formation of locally unpaired regions at specific genomic sites (such as regulatory regions [114]) [115], is closely regulated by enzymatic and other processes *in vivo* [113]. The destabilization energy of DNA double helices induced by DNA superhelicity, namely stress induced duplex destabilization (SIDD), is shown to be involved in transcriptional regulation [113]. Many molecular binding sites, including TFBSs, are susceptible to SIDD. For example, the $ilvp_G$ promoter of *Escherichia coli* is activated by an IHF (integration host factor)-mediated translocation of destabilization from the binding site to the -10 downstream region of the promoter [116]. Also, evidences show that regulatory proteins require locally denatured DNA for binding [117]. Further, SIDD sites are reported to oc-

cur at chromosomal attachment regions [118], which are known to augment transcription and separate independent regulatory domains [113].

While measuring DNA duplex stability *in vivo* is not currently possible, the computational method, WebSIDD [113], is developed to address this problem. It calculates the transition probability and destabilization energy of a given sequence based on a statistical mechanical SIDD analysis procedure [113]. Data computed from WebSIDD, although not directly measured from experiments, are considered quantitatively accurate, since all the thermodynamic parameter values used in WebSIDD are taken from experimental measurements [113]. In this thesis, WebSIDD is used to compute the destabilization energies for TFBS prediction.

3.1.3 TFBS prediction algorithm

In this thesis, multiple data sources are integrated to improve the prediction accuracy of ProbTF [19], which is a TFBS prediction algorithm under probabilistic framework. The basis of most TFBS prediction algorithms (including ProbTF), i.e., the probability models for binding sites and background sequences, and the ProbTF algorithm are described below.

Probability models for TFBSs and background sequences

In the probabilistic methods, the motif, represented as a position probability matrix, is assumed to be buried in the noisy background [119]. The most widely used probabilistic models for binding sites and background sequences are the position specific frequency matrix (PSFM) model [17; 120] and the Markovian model [119], respectively, based on which ProbTF is developed [19].

- Markovian background model [119]: The d^{th} order Markovian model means that, the probability of finding a nucleotide s_i ($s_i \in \{A,\ C,\ G,\ T\}$) at position i ($i \in \{1, \ldots, N\}$) depends on the d previous nucleotides in the sequence. Assuming that the d previous nucleotides before the start of the actual sequence S is accessible, the probability of the sequence of length N being generated by this background model ϕ_d is given by Equation 3.1.

$$P(S|\phi_d) = P(s_1, \ldots, s_d) \prod_{i=1}^{N} P(s_i|s_{i-1}, \ldots, s_{i-d}) \tag{3.1}$$

- PSFM model [17]: The motif of length l is represented by a position probability matrix θ as shown in Equation 3.2, where entry $\theta(s_i, i)$ is the probability of finding nucleotide s_i ($s_i \in \{A,\ C,\ G,\ T\}$) at position

i ($i \in \{1, \ldots, l\}$) in the motif.

$$\theta = \begin{bmatrix} \theta(A,1) & \theta(A,2) & \ldots & \theta(A,l) \\ \theta(C,1) & \theta(C,2) & \ldots & \theta(C,l) \\ \theta(G,1) & \theta(G,2) & \ldots & \theta(G,l) \\ \theta(T,1) & \theta(T,2) & \ldots & \theta(T,l) \end{bmatrix} \tag{3.2}$$

Probabilistic framework for TFBS prediction

ProbTF, a TFBS prediction algorithm, is used as the platform for data fusion method testing and novel information source exploration. This subsubsection is dedicated to introduce ProbTF's basic principles, where [19] is the key reference material.

In ProbTF, the non-binding site (i.e. background) sequence locations are modeled by the d^{th} order Markovian background model ϕ_d, and TFBSs are modeled with the standard PSFM model which is a product of independent multinomial distributions. Let Q denote the number of (unknown) binding sites and A be the (hidden) start positions of non-overlapping binding sites in sequence S, i.e., if $Q = c$ then $A = \{a_1, \ldots, a_c\}$. Assume a TF is characterized by M PSFMs, $\Theta = (\theta^{(1)}, \ldots, \theta^{(M)})$, and define $\pi \in \{1, \ldots, M\}^c$ as the configuration of motif models from Θ in A, i.e., π_i specifies the motif model $\theta^{(\pi_i)}$, which starts from location a_i and is of length l_{π_i}.

ProbTF The probability that a TF binds to a promoter sequence S that is of length N, $P(\Theta \rightarrow S|S, \Theta, \phi_d)$, is defined as the probability that at least one of the motif models in Θ has a binding site in S, which is computed by

$$\begin{aligned} P(\Theta \rightarrow S|S, \Theta, \phi_d) &= P(Q > 0|S, \Theta, \phi_d) && (3.3) \\ &= \sum_{c=1}^{\lfloor \frac{N}{l_{min}} \rfloor} P(Q = c|S, \Theta, \phi_d) \\ &= 1 - P(Q = 0|S, \Theta, \phi_d). && (3.4) \end{aligned}$$

$P(Q = c|S, \Theta, \phi_d)$ is the probability that a sequence S has c binding sites, which can be obtained with the Bayes' rule

$$P(Q = c|S, \Theta, \phi_d) = \frac{P(S|Q = c, \Theta, \phi_d) P(Q = c|\Theta, \phi_d)}{P(S|\Theta, \phi_d)}. \tag{3.5}$$

Solving Equation 3.5 depends on the normalization factor $P(S|\Theta, \phi_d)$, the prior of the number of motif instances $P(Q = c|\Theta, \phi_d)$, and the probability $P(S|Q = c, \Theta, \phi_d)$. Computations of each of these components (in Equation 3.5) are shown, separately, below.

First, $P(S|\Theta, \phi_d) = \sum_{c=0}^{\lfloor \frac{N}{l_{min}} \rfloor} P(S|Q = c, \Theta, \phi_d) P(Q = c|\Theta, \phi_d)$, where $\lfloor \frac{N}{l_{min}} \rfloor$ is the maximum number of non-overlapping motifs in an N-length sequence.

Second, $P(Q = c|\Theta, \phi_d)$, which is assumed to be independent of Θ and ϕ_d, has an exponential form, as represented by

$$P(Q = c) \sim \left[\frac{1}{2}, \frac{1}{C}, \frac{\kappa}{C}, \frac{\kappa^2}{C}, \ldots, \frac{\kappa^{\lfloor \frac{N}{l_{min}} - 1 \rfloor}}{C}\right], \tag{3.6}$$

where $C = 2\sum_{i=0}^{\lfloor \frac{N}{l_{min}} \rfloor - 1} \kappa^i$. This formula shows that, for a fixed value of Q, the prior over binding site positions A and configurations π is uniform and inversely proportional to the number of different binding site positions and configurations.

Finally, the probability $P(S|Q = c, \Theta, \phi_d)$ is obtained by summing over all possible positions and configurations, as shown in Equation 3.7.

$$P(S|Q = c, \Theta, \phi_d) = \sum_{\pi \in \{1,\ldots,M\}^c} \sum_{A:|A|=c} P(S|A, \pi, Q = c, \Theta, \phi_d) P(A, \pi|Q = c, \Theta, \phi_d) \tag{3.7}$$

The following text shows how $P(S|Q = c, \Theta, \phi_d)$ is obtained based on Equation 3.7.

$P(S|A, \pi, Q = c, \Theta, \phi_d)$ is the probability of sequence S, given non-overlapping motif positions, and the motif and background models. It is computed by Equation 3.8, where $|A| = Q = c$, and $W_{a_j}^{\pi_j}$ is shown in Equation 3.9. Recall that the notation of θ is defined in Equation 3.2.

$$\begin{aligned} P(S|A, \pi, Q = c, \Theta, \phi_d) &= \prod_{i=1}^{N} \phi_d(s_i) \prod_{j=1}^{|A|} \prod_{k=0}^{l_{\pi_j}-1} \frac{\theta^{(\pi_j)}(s_{a_j+k}, k+1)}{\phi_d(s_{a_j+k})} \\ &= P(S|\phi_d) \prod_{j=1}^{|A|} W_{a_j}^{\pi_j}, \end{aligned} \tag{3.8}$$

$$W_{a_j}^{\pi_j} = \begin{cases} \prod_{k=0}^{l_{\pi_j}-1} \frac{\theta^{(\pi_j)}(s_{a_j+k}, k+1)}{\phi_d(s_{a_j+k})} & \text{if } 1 \le a_j \le N - l_{\pi_j} + 1 \\ 0 & \text{otherwise.} \end{cases} \tag{3.9}$$

Equation 3.7 becomes Equation 3.10 after plugging in Equation 3.8, where $S_{a_1+l_{\pi_1}}$ is a subsequence of S covering the locations from $a_1 + l_{\pi_1}$ to N. Equation 3.10 is a recursive formula apart from $P(A, \pi|Q = c, \Theta, \phi_d)$, where $P(A, \pi|Q = c, \Theta, \phi_d)$ is a constant prior for a fixed Q and can be computed numerically in a similar recursive fashion (derivation details can be found in [19]).

$$\begin{aligned} P(S|Q = c, \Theta, \phi_d) &= \sum_{\pi_1 \in \{1,\ldots,M\}} \sum_{a_1=1}^{N - cl_{min} + 1} W_{a_1}^{\pi_1} P(S_{a_1+l_{\pi_1}}|Q = c - 1, \Theta, \phi_d) \\ &\quad \times P(A, \pi|Q = c, \Theta, \phi_d) \end{aligned} \tag{3.10}$$

In this thesis, 0^{th} order Markovian model ($d = 0$) is used, and κ in Equation 3.6 is set to 0.5, according to [19].

ProbTF with additional data sources ProbTF allows integrating multiple data sources in TFBS prediction. Assume that the data sources are in the form of $D = (P_1, \ldots, P_N)$ where P_i is the probability that the i^{th} base pair location is a binding site. D can be derived from a single or multiple data source(s).

Similarly, the probability of a TF binding to a promoter sequence S with additional knowledge of data source D is computed by Equation 3.11, where $P(Q = c|S, D, \Theta, \phi_d)$ can be obtained from the Bayes' rule as shown in Equation 3.12.

$$\begin{aligned} P(\Theta \rightarrow S|S, D, \Theta, \phi_d) &= P(Q > 0|S, D, \Theta, \phi_d) \\ &= \sum_{c=1}^{\lfloor \frac{N}{l_{min}} \rfloor} P(Q = c|S, D, \Theta, \phi_d) \\ &= 1 - P(Q = 0|S, D, \Theta, \phi_d) \end{aligned} \tag{3.11}$$

$$P(Q = c|S, D, \Theta, \phi_d) = \frac{P(S, D|Q = c, \Theta, \phi_d)P(Q = c|\Theta, \phi_d)}{P(S, D|\Theta, \phi_d)} \tag{3.12}$$

The normalization factor $P(S, D|\Theta, \phi_d)$ is calculated in a similar way as the case where no additional data source is used.

The prior $P(Q = c|\Theta, \phi_d)$ is defined using Formula 3.6.

$P(S, D|A, \pi, \Theta, \phi_d)$ is needed to obtain $P(S, D|Q = c, \Theta, \phi_d)$, which can be factorized by Equation 3.13. Note the assumption used here is that S and D are conditionally independent and the probability of D does not depend on the background and PSFM models.

$$P(S, D|A, \pi, \Theta, \phi_d) = P(S|A, \pi, \Theta, \phi_d)P(D|A, \pi). \tag{3.13}$$

In Equation 3.13, $P(S|A, \pi, \Theta, \phi_d)$ is obtained by Equation 3.8, and $P(D|A, \pi)$ can be further factorized as Equation 3.14, where $P(D|\phi_d) = \prod_{i=1}^{N}(1 - P_i)$, $D_{a_j}^{(\pi_j)} = \prod_{k=0}^{l_{\pi_j}-1} \frac{P_{a_j+k}}{1-P_{a_j+k}}$, $I = \{1, \ldots, N\}$ denotes the base pair indices of a promoter, and $I_{A,\pi} = \{a_1, \ldots, a_1 + l_{\pi_1} - 1, a_2, \ldots, a_2 + l_{\pi_2} - 1, \ldots, a_M, \ldots, a_M + l_{\pi_M} - 1\}$.

$$\begin{aligned} P(D|A, \pi) &= \prod_{i \in I \setminus I_{A,\pi}} (1 - P_i) \prod_{i \in I_{A,\pi}} P_i \\ &= \prod_{i \in I}(1 - P_i) \prod_{i \in I_{A,\pi}} \frac{P_i}{1 - P_i} \\ &= \prod_{i=1}^{N}(1 - P_i) \prod_{j=1}^{|A|} \prod_{k=0}^{l_{\pi_j}-1} \frac{P_{a_j+k}}{1 - P_{a_j+k}} \\ &= P(D|\phi_d) \prod_{j=1}^{|A|} D_{a_j}^{(\pi_j)} \end{aligned} \tag{3.14}$$

Thus, the joint probability $P(S, D|A, \pi, \Theta, \phi_d)$ can be written compactly as

$$P(S, D|A, \pi, \Theta, \phi_d) = P(S|\phi_d)P(D|\phi_d)\prod_{j=1}^{|A|}(W_{a_j}^{(\pi_j)} \cdot D_{a_j}^{(\pi_j)}). \quad (3.15)$$

Finally, a similar efficient recursive formula as Equation 3.10 for $P(S, D|Q = c, \Theta, \phi_d)$ is obtained.

3.1.4 Data fusion methods for TFBS prediction

A general approach to fuse multiple data sources is shown in Formula 3.16 [19], assuming that there are n data sources, the m^{th} data source is denoted as $D^{(m)}$ ($D^{(m)} = (\mathcal{P}_1^{(m)}, \ldots, \mathcal{P}_N^{(m)})$, $1 \leq m \leq n$, N is the length of the promoter sequence), and the probability of the position i contributing to a binding site is $\mathcal{P}_i^{(m)}$.

$$P_i \propto \prod_{m=1}^{n}(\mathcal{P}_i^{(m)})^{L_m} \quad (3.16)$$

Notice from this formula that a weight, L_m, is assigned to the prior probability provided by each data source, and the final prior for a particular position is computed as the multiplication of these weighted probabilities. This method (as shown by Equation 3.16), although is proven capable of improving the TFBS prediction accuracy when data such as evolutionary conservation is used, does not perform well when other information, e.g., nucleosome positioning, is employed [19]. Thus, [Publication **I**] studies how TFBS prediction could be further improved by exploring a new way to integrate additional data (whose principle is illustrated in Fig. 3.1), and utilizing novel information sources (see [Publication **I**]).

In particular, the proposed data fusion method first filters out the positions whose probabilities of being TFBSs are below a certain threshold ($T^{(m)}$, which is data specific) for each data source using Equation 3.17. Then, the probability of a position i being a TFBS is computed by fusing its thresholded probabilities from multiple data sources ($\widetilde{\mathcal{P}}_i = (\widetilde{\mathcal{P}}_i^{(1)}, \ldots, \widetilde{\mathcal{P}}_i^{(n)})$, $1 \leq i \leq N$) with Equation 3.18. In this equation, j_i is the number of data sources whose thresholded probabilities at location i are above zero as defined in Equation 3.19, and the weight L_{j_i} ($1 \leq m \leq n$) increases with the value of its subindex. In short, the new data fusion method assigns higher probabilities to positions that are indicated to be the binding sites by more

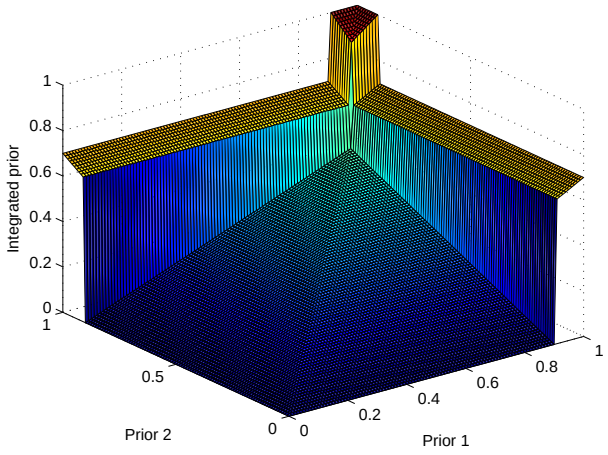


Figure 3.1: An illustration of the new prior integration method. A case of two additional information sources is illustrated. x and y axes each corresponds to one data source. z-axis shows the integrated prior. $L_0 = 0.5$, $L_1 = 0.7$, $L_2 = 1$ and $T^{(1)} = T^{(2)} = 0.9$ percentile of the distribution of each data source. This figure is retrieved from [Publication **I**] with permission.

evidence sources.

$$\widetilde{\mathcal{P}}_i^{(m)} = \begin{cases} \mathcal{P}_i^{(m)} & \text{if } \mathcal{P}_i^{(m)} \geq T^{(m)} \\ 0 & \text{otherwise,} \end{cases} \tag{3.17}$$

$$\widetilde{P}_i = \begin{cases} \max(\widetilde{\mathcal{P}}_i) \times L_{j_i} & \text{if } j_i \geq 1 \\ \min(\mathcal{P}_i) \times L_0 & \text{otherwise} \end{cases} \tag{3.18}$$

$$j_i = \left| \left\{ \widetilde{\mathcal{P}}_i^{(m)} \mid \widetilde{\mathcal{P}}_i^{(m)} > 0, 1 \leq m \leq n \right\} \right| \tag{3.19}$$

In practice, each raw data source for the i^{th} position, $\mathcal{P}_{i,raw}^{(m)}$, is scaled by a multiplicative factor f_1 and an additive factor f_2 before being used (as shown in Equation 3.20), and the integrated prior $\widetilde{P}_i$ for location i is scaled by f_3 before TFBS prediction (as shown by Equation 3.21). These scaling parameters, i.e., f_i ($i \in \{1, 2, 3\}$), are data specific and chosen by a grid search method via optimizing the receiver operating characteristic (ROC) curves of the TFBS prediction results.

$$\mathcal{P}_i^{(m)} = f_1 \times \mathcal{P}_{i,raw}^{(m)} + f_2 \tag{3.20}$$

$$P_i = 2 \times f_3 \times \widetilde{P}_i + 0.5 - f_3 \tag{3.21}$$

The results show that the new data fusion principle outweighs the one illustrated in Equation 3.16, and is characterized by lower FPR. This is because a non-binding site may still satisfy one or several criteria of being

a binding site, and the previous method, assigning all the positions with the same weight for each data source, is highly subject to the adverse effects caused by such positions. The proposed method, on the other hand, distinguishes positions by assigning them different weights according to the number of data sources that indicate them to be binding sites. Thus, the proposed method can significantly reduce the risk of choosing a position that is not a binding site in TFBS prediction. To do so, thresholding is needed to binarize each data source, based on which each position is labeled as being a binding site or not; the number of positive labels (being a binding site) for each position is then counted to assign a corresponding weight to each site's statistic.

3.2 Gene clustering

Closely related genes are likely to collaborate in orchestrating a particular pathway and can be viewed as a block when constructing and analyzing a GRN. Thus, the precision of gene clustering is of great importance in further data analysis. One way to improve the clustering accuracy is to observe genes' relationships from multiple perspectives, motivated by which this thesis presents a model based method to jointly utilize multiple data sources in gene clustering.

In [Publication **II**] to [Publication **IV**], different joint finite mixture models are developed for gene clustering, whose performances are tested by simulations ([Publication **II**] to [Publication **IV**]), and real data ([Publication **II**] and [Publication **III**]). [Publication **VI**] extends this clustering framework and builds a new model to solve problems in another research domain, i.e., noisy attractor detection.

This section presents the key concept, the data sources and some background information of [Publication **II**] to [Publication **IV**], and summarizes all the developed models (including [Publication **II**] to [Publication **IV**] and [Publication **VI**]) in a holistic way. The simulation tests and biological applications of [Publication **II**] to [Publication **IV**] are not discussed here but summarized in Chapter 4, whose details can be found in each corresponding publication.

The background of the problem studied in [Publication **VI**] is excluded in this chapter but explained explicitly in Chapter 5.

3.2.1 Key concept

The key concept encountered in [Publication **II**] to [Publication **IV**] is gene clustering, which is defined below.

Definition 6 (Gene Clustering)**.** *Gene clustering is the process of grouping*

or segmenting a collection of genes into subsets or clusters, such that genes within each cluster have more common features or are more closely related to each other than genes assigned to different clusters [121].

3.2.2 Data sources

In this thesis, data fusion methods are developed to cluster genes from multiple data sources, with the aim of more accurately identifying functionally related genes or genes that are co-regulated. Information sources involved include gene expression data, protein-DNA binding data, and protein-protein interaction (PPI) data.

Gene expression data

Gene expression data is obtained from DNA microarrays or DNA chips which measure the messenger RNA (mRNA) levels in particular cells or tissues in a high throughput manner [122]. There are two widely used DNA microarray platforms, i.e., spotted microarrays and oligonucleotide microarrays [122; 123]. In spotted DNA chips, complementary DNAs (cDNA), oligonucleotide chains or short PCR products of the interested genes are immobilized on the chip [123]. In oligonucleotide chips, oligonucleotides are built or produced *in situ* on an array by, e.g., photolithographic manufacturing (Affymetrix chips) or ink-jet synthesizer (Agilent chips) [123]. DNA microarray experiments can be carried out in a double-channel or single-channel manner [122; 123], depending on the chip type (e.g., Affymetrix and cDNA chips are single- and double-channel microarrays, respectively) and the experimental purpose (e.g., intensity ratios are measured by double-channel experiments, and raw intensities are obtained from single-channel experiments) [123]. Various microarray experimental designs are developed to serve different purposes. Time series, a series of samples following each other in time, is typically used to study a development over time [124]. Conditions, generally referred to as time points or states and accompanied by several replicate hybridizations to ensure adequate information for the results' significance assessment, are commonly used to, e.g., infer gene functions or genes' relationships [124].

The quality of raw DNA microarray data is subject to errors induced during the experiment, both systematic errors (e.g., asynchronous cells are hybridized together, cross-hybridization, microarray's surface characteristics) and random mistakes (e.g., human errors) [123]. Thus, proper data preprocessing and normalization are needed before further data analysis [125]. Besides, microarray quality control in image analysis is also shown to be important in obtaining good quality data [126]. In practice, filtered log-transformed DNA microarray data is often assumed to be of Gaussian distribution [127]. One common application of gene expression data is to group

genes with similar expression magnitudes and/or dynamics shapes together, assuming that they share similar functions or occur simultaneously [128]. Extensive research has been devoted to this area [129–133], involving many clustering techniques such as hierarchical methods [134; 135], partitioning methods [30; 135], and model based methods [136].

The gene expression data used in the real case application of this thesis contains 1960 genes measured from 95 conditions in mouse. These conditions consist of 23 treatments, each is a time series with an average of four time points, and the treatments are the combinations of six Toll-like receptor (TLR) agonists and four gene knock-out mutants. Different conditions and genes are used in [Publication **II**] and [Publication **III**], depending on each particular problem.

Protein-DNA binding data

Protein-DNA interactions play a central role in many biological processes, including transcriptional regulation [11; 12]. Thus, many experimental [13; 14] and computational [19; 79; 137–140] efforts are devoted to investigating protein-DNA physical interactions and their binding mechanisms and affinities.

Experimentally, ChIP (chromatin immunoprecipitation) related techniques, ChIP-chip [13] or ChIP-Seq [14], are normally used to obtain protein-DNA binding data. Specifically, ChIP requires cross-linking of living cells with formaldehyde, shearing of chromatin into short fragments via sonication, immunoprecipitating protein-bound DNA fragments by an antibody specific to the interested protein(s), reversing the protein-DNA cross-links, purifying and determining the DNA sequence(s) [141]. ChIP-chip is the technique that combines ChIP with DNA microarrays [13], which is currently the most widely applied method to map the protein binding sites on DNA sequences genome-wide. ChIP-seq, the technique that combines ChIP with the next generation massively parallel sequencing [14], is taking over the dominance of ChIP-chip by its high resolution [14].

Computationally, many methods are developed to discover protein-DNA binding sites and/or predict their binding affinities [19; 79; 137–140]. This can be achieved by analyzing protein-DNA complexes or studying DNA sequences alone. From the first aspect, the identities of amino acids at protein-DNA interface were first used to reveal TFBSs [137]. Later, the protein structural information is also taken into account [138]. Alternatively, binding affinities are also predictable via modeling protein-DNA complexes at an all-atom level, where the protein-DNA binding energies are evaluated and used [139]. From the second perspective, the patterns or motifs of protein-DNA binding sites can be discovered by analyzing DNA sequences that contain the binding sites (TFBS discovery algorithm) [79], which can be used to

further predict the binding sites of the unknown sequences (TFBS prediction algorithm) [19]. With enormous efforts being devoted to analyzing protein-DNA binding sites, finding functional binding sites has been acknowledged and explored, e.g., Beyer et al. jointly assess multiple evidences to predict the regulatory TFBSs [140].

In this thesis, the protein-DNA binding data used for real case analysis is computed from a probabilistic TFBS prediction algorithm, namely ProbTF [19]. The data is composed of the probabilities (within the closed region $[0, 1]$) of 266 TFs binding to 20397 sequences in mouse, which is assumed to be of beta distribution. The TFs and genes selected from this data set differ in each application (see each clustering event in [Publication **II**] and [Publication **III**]), depending on the purpose of the analysis and the genes available in the other data sets, e.g., the conditions and the genes in the gene expression data.

PPI data

Most proteins need to form polymers (either homo- or hetero-polymers) to exert their functions, rendering PPIs a fundamental regulatory mechanism of GRNs [11; 12]. There are two widely applied techniques to experimentally detect PPIs, i.e., the yeast two-hybrid (Y2H) system [142; 143], and affinity purification followed by mass spectrometry (AP-MS) [143; 144]. In Y2H system, if a pair of proteins (namely 'bait' and 'prey') interact, the reporter gene expresses [142; 143]. In an AP-MS experiment, a tagged protein ('bait'), which is expressed in the cell of interest, is first extracted together with the associated proteins ('prey') from the cell by co-immunoprecipitation or tandem affinity purification; and then the extracted proteins are identified by MS [143; 144]. Y2H and AP-MS mainly differ in two aspects. First, Y2H detects binary interactions, while AP-MS finds one-to-many relationships [143]. Second, the interactions found by Y2H may not exist *in vivo* due to the non-physiological conditions under which it is conducted, while those detected by AP-MS are not restricted to hypothesis [143]. Also, there are several complementary features between these two techniques [143]. For example, Y2H can not detect weak interactions such as those that are stabilized by other subunits of a complex which AP-MS is capable of, AP-MS is not able to find transient interactions while Y2H can, and AP-MS is biased towards abundant proteins whereas Y2H is not [143]. Both of these two techniques can be conducted in a high throughput manner. For example, large-scale Y2H can be carried out in a colony-array format [145–147], and AP-MS has been systematically applied to large sets of yeast proteins [148–150].

There are many databases to store PPIs, such as MINT [151], IntAct [152], DIP [153; 154], BioGRID [155], HPRD [156] and MIPS/MPact [157]. These

databases differ in their scope, type and coverage of data [143]. To retrieve a more complete data set, one can integrate information from multiple databases. Platforms such as PINA [158] can be used for this purpose.

3.2.3 Gene clustering methods

Gene clustering is a typical unsupervised machine learning problem, for which many methods are developed [30; 134; 136]. The most commonly used approaches can be roughly classified into three categories, the hierarchical methods, the partitioning methods, and the model-based methods [159].

Hierarchical clustering can be either agglomerative or divisive, which proceeds by recursively fusing or separating the objects into greater or finer groups to optimize a certain criterion [134]. Different criteria are developed to serve this purpose, among which single linkage, complete linkage, average linkage and group average linkage are widely applied [135]. Distances such as Euclidean distance [160], Mahalanobis distance [161], Manhattan distance [122], and Hamming distance [162] are generally adopted in these criteria to measure the cluster dissimilarity. Thus, the accuracy of hierarchical clustering highly depends on the distance measurement, which requires expert domain knowledge especially for complex data types. For example, Euclidean distance, which is commonly used when data is representable in vector space, is not appropriate for high-dimensional text clustering [163]; and semantic similarity measurements, such as graph-structure based distances and information content based methods, are especially applicable to gene ontology (GO) based clustering [10]. Further, hierarchical clustering is computationally inefficient, given that computing distances among all observation pairs requires a complexity of $O(n^2)$, where n is the number of observations [164]. Also, at what granularity should the algorithm stop is an important issue and could not be naturally determined without prior knowledge or estimation of the number of clusters [159].

Partitioning, also called iterative partitioning or iterative relocation, is another class of commonly used clustering methods, where data points are moved across groups until no further improvement could be obtained based on a certain criterion [30]. Many well-known algorithms belong to this category, such as K-means clustering [30], fuzzy C-means (FCM) [165] and partitioning around medoids (PAM) [166], among which K-means is the most representative. In particular, it partitions the objects into K clusters such that each object belongs to its closest group that is represented by the group mean, where K needs to be pre-specified.

Hierarchical methods and partitioning methods are also called 'heuristic methods', since both of them rely on some heuristics and follow intuitively reasonable procedures [159]. Although considerable research has been done on these methods, still little associated systematic guidance is available for

solving some practical issues [159], including how to specify the number of clusters, how to handle the outliers, and how to choose or define a good distance for a particular clustering problem.

Clustering algorithms of the third category are called model based methods, which try to fit the given data to certain mathematical models, assuming that data are generated by a mixture of the underlying probability distributions [136]. Model based methods can naturally solve the problems generically inherited by heuristic methods [159] which, e.g., often determine the number of clusters by casting it as the model selection problem (some commonly used model selection criteria are discussed later in this chapter) and group the outliers as separate clusters [136; 159]. Further, model based methods outweigh heuristic methods in their statistical nature [136; 159]. There are different types of methods within this category, such as finite mixture models [136], infinite mixture models [167], model based hierarchical clustering [168], and specialized model-based partitioning clustering algorithm [164] (e.g., Self Organizing Map (SOM) [169]). Among these alternatives, finite model based clustering is the main focus of this thesis, base on which all the joint mixture models presented in [Publication **II**] to [Publication **IV**] and [Publication **VI**] are built.

In finite mixture model gene clustering, each observation $\mathbf{x}_j$ ($j = 1, \ldots, n$ and n is the number of genes) is assumed to be drawn from finite mixture distributions with the prior probability π_δ, component-specific distribution f_δ and its parameters θ_δ [136]. The formula is shown in Equation 3.22 [136], where $\theta = \{(\pi_\delta, \theta_\delta) : \delta = 1, \ldots, g\}$ represents all the unknown parameters, $0 < \pi_\delta \leq 1$ for any δ, and $\sum_{\delta=1}^{g} \pi_\delta = 1$.

$$f(\mathbf{x}_j|\theta) = \sum_{\delta=1}^{g} \pi_\delta f_\delta(\mathbf{x}_j|\theta_\delta) \tag{3.22}$$

3.2.4 Joint finite mixture models

In [Publication **II**] to [Publication **IV**] and [Publication **VI**], four joint finite mixture models, i.e., beta-Gaussian mixture model (BGMM), stratified beta-Gaussian mixture model (sBGMM), Gaussian-Bernoulli mixture model (GBMM) and gamma-Bernoulli mixture model (ΓBMM), are constructed to improve the clustering accuracy via data fusion. To explain the modeling framework, assume the model has g components, there are n genes and N data types, and the second dimension and the parameters of data type i ($i \in \{1, \ldots, N\}$) are p_i and θ_i, respectively. Also, define $\pi = [\pi_1, \ldots, \pi_g]^T$ and $\theta = [\pi, \theta_1, \ldots, \theta_N]^T$. If the distribution of data type i has ρ parameters, then

$$\theta_i = [\vartheta_{1,11}, \ldots, \vartheta_{1,gp_i}, \vartheta_{2,11}, \ldots, \vartheta_{2,gp_i}, \ldots, \vartheta_{\rho,11}, \vartheta_{\rho,gp_i}]^T . \tag{3.23}$$

where ϑ_{ρ,gp_i} represents the ρ^{th} parameter of the data (in the i^{th} data set) belonging to the g^{th} group in the p_i^{th} dimension. Further, denote X_i ($i \in \{1,\ldots,N\}$) as the observations of data type i, and function f_i of $\mathbf{x}_i$ as the density function of the corresponding distribution ($f_i \neq f_j$ if $i \neq j$). Specifically, in [Publication **II**] to [Publication **IV**] and [Publication **VI**], four types of data are used, which are assumed to be of beta, Gaussian, gamma and Bernoulli distributions, and denoted as X_1 to X_4, respectively.

A joint finite mixture model is built from multiple mixture models that are of different distributions, assuming that data of different distributions are independent. In this thesis, four component models are developed or used to build joint models, which are beta mixture model (BMM), Gaussian mixture model (GMM), gamma mixture model (ΓMM), and Bernoulli mixture model (BerMM). Each component model is assumed to be the product of p_i ($i \in \{1,2,3,4\}$) independent distributions of data type X_i, whose probability density functions are defined in Equations 3.24 to 3.27, respectively. Note that α and β denote the two shape parameters in Equation 3.24, and the shape and scale parameters, respectively, in Equation 3.26. $|V|$ in Equation 3.25 is the determinant of the diagonal covariance matrix of the Gaussian distribution in the mixture model, i.e., $|V| = \prod_{u=1}^{p_2} \sigma_u^2$, where $V = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_{p_2}^2)$ and σ is the standard deviation. Also, μ and q represent the mean in Equations 3.25 and 3.27, respectively.

$$f_{1,\delta}(\mathbf{x}_1|\theta_{1,\delta}) = \prod_{u=1}^{p_1} \frac{x_{1,u}^{\alpha_{\delta u}-1}(1-x_{1,u})^{\beta_{\delta u}-1}}{B(\alpha_{\delta u},\beta_{\delta u})} \tag{3.24}$$

$$f_{2,\delta}(\mathbf{x}_2|\theta_{2,\delta}) = \frac{1}{(2\pi)^{\frac{p_2}{2}}|V|^{\frac{1}{2}}} \exp\big(-\frac{1}{2}(\mathbf{x}_2-\mu_\delta)^T V^{-1}(\mathbf{x}_2-\mu_\delta)\big) \tag{3.25}$$

$$f_{3,\delta}(\mathbf{x}_3|\theta_{3,\delta}) = \prod_{u=1}^{p_3} \frac{\mathbf{x}_{3,u}^{\alpha_{\delta u}-1}\exp(-\mathbf{x}_{3,u}\beta_{\delta u}^{-1})}{\Gamma(\alpha_{\delta u})\beta_{\delta u}^{\alpha_{\delta u}}} \tag{3.26}$$

$$f_{4,\delta}(\mathbf{x}_4|\theta_{4,\delta}) = \prod_{u=1}^{p_4} q_{\delta u}^{\mathbf{x}_{4,u}}(1-q_{\delta u})^{(1-\mathbf{x}_{4,u})} \tag{3.27}$$

EM algorithm

Expectation maximization (EM) algorithm [170], a general technique to obtain the maximum likelihood estimates (MLE) from incomplete data, is widely used for parameter estimation in model based methods [136]. EM is an iterative method alternating between performing an expectation (E) step and a maximization (M) step [170]. In the E step, an expectation of the data log-likelihood with respect to the current estimate of the distribution for the latent variables is computed [170]. In the M step, the parameters that maximize the expected data log-likelihood calculated in the E step are found, which are then used as the inputs of the next E step [170].

Specifically, for a specific data type i ($i \in \{1,2,\ldots,N\}$), MLEs are determined by the marginal likelihood of the observed data [136] (as represented

by Equation 3.28), which, however, is often intractable.

$$\log L(\theta_i) = \sum_{j=1}^{n} \log(\left[\sum_{\delta=1}^{g} \pi_\delta f_{i,\delta}(\mathbf{x}_{i,j}|\theta_{i,\delta})\right]) \tag{3.28}$$

In practice, this problem is generally casted in the framework of incomplete data (L_c is used to represent the complete data likelihood), and the EM algorithm iteratively applies the following two steps [170].

- E step: calculate the expectation of the complete data log-likelihood function, $Q(\theta_i|\theta_i^{(m)})$, with respect to the conditional distribution of the latent variables given the observations under the current estimate of the parameters $\theta_i^{(m)}$, as shown below, where $\mathbf{c} = \{c_1, c_2, \ldots, c_n\}$ and c_j ($j \in \{1, 2, \ldots, n\}$) is an indicator function of gene j.

$$Q(\theta_i|\theta_i^{(m)}) = E_{\mathbf{c}|X_i,\theta_i^{(m)}}\left[\log(L_c|X_i, \theta_i)\right].$$

- M step: find the parameters that maximize $Q(\theta|\theta^{(m)})$, i.e.,

$$\theta_i^{(m+1)} = \underset{\theta_i}{\text{argmax}} Q(\theta_i|\theta_i^{(m)}).$$

In the joint finite mixture model for clustering genes from data sources of N distinct distributions, L_c can be factorized as $L_c(\theta) = \prod_{i=1}^{N} f(X_i|\mathbf{c}, \theta)$, and $Q(\theta|\theta^{(m)})$ is obtained from Equation 3.29, where $\tau_{j\delta}^{(m)}$ is computed from Equation 3.30 and is the estimated posterior probability of a gene belonging to a group, i.e., $\tau_{j\delta} = f(c_j = \delta|\mathbf{x}_{1,j}, \ldots, \mathbf{x}_{N,j}, \theta_\delta^{(m)})$ denotes the posterior probability of gene j belonging to group δ.

$$Q(\theta|\theta^{(m)}) = \sum_{j=1}^{n}\sum_{\delta=1}^{g} \tau_{j\delta}^{(m)} \log(\pi_\delta \prod_{i=1}^{N} f_{i,\delta}(\mathbf{x}_{i,j}|\theta_{i,\delta})) \tag{3.29}$$

$$\begin{aligned} \tau_{j\delta}^{(m)} &= f(c_j = \delta|\mathbf{x}_{1,j}, \ldots, \mathbf{x}_{N,j}, \theta_\delta^{(m)}) \\ &= \frac{\pi_\delta^{(m)} \prod_{i=1}^{N} f_{i,\delta}(\mathbf{x}_{i,j}|\theta_{i,\delta}^{(m)})}{\sum_{\delta'=1}^{g} \pi_{\delta'}^{(m)} \prod_{i=1}^{N} f_{i,\delta'}(\mathbf{x}_{i,j}|\theta_{i,\delta'}^{(m)})} \end{aligned} \tag{3.30}$$

If the stratified mixture modeling strategy is used ([Publication **III**]), i.e., genes are partitioned into K groups based on the priors, then Equations 3.29 and 3.30 become Equations 3.31 and 3.32, where $\pi_{\delta,(k)}$ is the

estimated posterior probability of a gene belonging to group δ while being in the k^{th} ($k \in \{1, 2, \ldots, K\}$) layer according to the prior.

$$Q(\theta|\theta^{(m)}) = \sum_{j=1}^{n}\sum_{\delta=1}^{g} \tau_{j\delta}^{(m)} \log(\pi_{\delta,(k)} \prod_{i=1}^{N} f_{i,\delta}(\mathbf{x}_{i,j}|\theta_{i,\delta})) \tag{3.31}$$

$$\tau_{j\delta}^{(m)} = \frac{\pi_{\delta,(k)}^{(m)} \prod_{i=1}^{N} f_{i,\delta}(\mathbf{x}_{i,j}|\theta_{i,\delta}^{(m)})}{\sum_{\delta'=1}^{g} \pi_{\delta',(k)}^{(m)} \prod_{i=1}^{N} f_{i,\delta'}(\mathbf{x}_{i,j}|\theta_{i,\delta'}^{(m)})} \tag{3.32}$$

Estimation of the parameters in the M step depends highly on the probability density function of the data distribution and varies greatly among models [136]. Since Equation 3.29 can be reformulated as Equation 3.33, parameter estimation of the joint finite mixture model can be reduced to estimating the parameters of each component mixture model and the parameter π.

$$Q(\theta|\theta^{(m)}) = \sum_{i=1}^{N}(Q(\theta_i|\theta_i^{(m)})) + Q(\pi) \tag{3.33}$$

For the four component models, i.e., BMM, GMM, ΓMM, and BerMM which are developed or employed in this thesis, the methods used to update their parameters are listed below.

- Parameters of **beta** distribution: two versions, i.e, the standard version and the approximated version.
  - Standard version: the parameters are estimated by maximizing $Q(\theta_1|\theta_1^{(m)})$ using Newton-Raphson method. Let $\theta_{1,\delta} = (\alpha_\delta, \beta_\delta)$, then the parameters are updated by Equation 3.34 with the constraint $\theta_{1,\delta} \geq \mathbf{1}$, where $H^{-1}(\theta_{1,\delta}^{(m)})$ is the Hessian matrix evaluated at $\theta_{1,\delta}^{(m)}$ and $\nabla_{\theta_{1,\delta}}\mathcal{L}(\theta^{(m)})$ is the partial derivatives of the Lagrangian function with respect to $\theta_{1,\delta}^{(m)}$. This version is used in [Publication **II**] and [Publication **III**], with the derivations presented in the appendix of [Publication **II**].

$$\hat{\theta}_{1,\delta}^{(m+1)} = \theta_{1,\delta}^{(m)} - H^{-1}(\theta_{1,\delta}^{(m)})\nabla_{\theta_{1,\delta}}\mathcal{L}(\theta^{(m)}) \tag{3.34}$$

  - Approximated version: the parameters are estimated by maximizing $\log(L_c|X_1, \theta_1)$ instead of its expectation using a numerical optimization method, i.e., the 'betafit' function in Matlab. This version is presented in [Publication **II**].

- Parameters of **Gaussian** distribution: two versions, i.e., the standard version and the approximated version. Both versions employ the diagonal covariance matrix in the probability density function of the Gaussian distribution.

– Standard version: the parameters are estimated by maximizing $Q(\theta_2|\theta_2^{(m)})$, with the update formulas shown in Equations 3.35 and 3.36. This version is used in [Publication **II**] to [Publication **IV**]. The derivations are referenced from [136] and can be found in the appendix of [Publication **II**].

$$\hat{\mu}_{\delta u}^{(m+1)} = \sum_{j=1}^{n} \tau_{j\delta}^{(m)} x_{2,ju} / \sum_{j=1}^{n} \tau_{j\delta}^{(m)} \tag{3.35}$$

$$\hat{\sigma}_u^{2,(m+1)} = \sum_{j=1}^{n} \sum_{\delta=1}^{g} \tau_{j\delta}^{(m)} (x_{2,ju} - \mu_{\delta u}^{(m)})^2 / n \tag{3.36}$$

– Approximated version: the parameters are estimated to maximize $\log(L_c|X_2, \theta_2)$. The update formulas are shown in Equations 3.37 and 3.38, where $I_\delta^{(m)}$ is composed of all the genes in cluster δ estimated from the E step, and $n_\delta^{(m)} = |I_\delta^{(m)}|$, i.e., $n_\delta^{(m)}$ is the number of genes in cluster δ estimated at the m^{th} iteration. This version is presented and implemented in [Publication **II**].

$$\hat{\mu}_\delta^{(m+1)} = \sum_{j \in I_\delta^{(m)}} x_{2,ju}^{(m)} / n_\delta^{(m)} \tag{3.37}$$

$$\hat{\sigma}_u^{2,(m+1)} = \sum_{j \in I_\delta^{(m)}} \sum_{\delta=1}^{g} (x_{2,ju} - \mu_{\delta u}^{(m)})^2 / n \tag{3.38}$$

- Parameters of **gamma** distribution: the parameters are estimated to maximize $Q(\theta_3|\theta_3^{(m)})$ using Newton-Raphson method. Let $\theta_{3,\delta} = (\alpha_\delta, \beta_\delta)$, then the parameters are updated by Equation 3.39, with the constraint $\theta_{3,\delta} \geq \mathbf{1}$, where $H^{-1}(\theta_{3,\delta}^{(m)})$ is the Hessian matrix evaluated at $\theta_{3,\delta}^{(m)}$, and $\nabla_{\theta_{3,\delta}} \mathcal{L}(\theta^{(m)})$ is the partial derivatives of Lagrangian function with respect to $\theta_{3,\delta}$. The derivations are presented in [Publication **VI**].

$$\hat{\theta}_{3,\delta}^{(m+1)} = \theta_{3,\delta}^{(m)} - H^{-1}(\theta_{3,\delta}^{(m)}) \nabla_{\theta_{3,\delta}} \mathcal{L}(\theta^{(m)}) \tag{3.39}$$

- Parameters of **Bernoulli** distribution: the parameters are estimated to maximize $Q(\theta_4|\theta_4^{(m)})$. The update formula, as shown in Equation 3.40, is used in [Publication **IV**] and [Publication **VI**], whose derivations are presented in [Publication **VI**].

$$\hat{\mathbf{q}}_\delta^{(m+1)} = \frac{\sum_{j=1}^{n} \tau_{j\delta}^{(m)} \mathbf{x}_{4,j}}{\sum_{j=1}^{n} \tau_{j\delta}^{(m)}} \tag{3.40}$$

- π: two versions, i.e., the non-stratified version and the stratified version.

– Non-stratified version: π's are updated according to Equation 3.41. This version is used in non-stratified models such as BGMM, GBMM and ΓBMM, whose derivations are available in the appendix of [Publication **II**].

$$\hat{\pi}_{\delta}^{(m+1)} = \sum_{j=1}^{n} \tau_{j\delta}^{(m)} / n \tag{3.41}$$

– Stratified version: π's are updated according to Equation 3.42, where n_k is the number of genes in the k^{th} layer according to the prior. This version is used in sBGMM, whose derivations are shown in [Publication **III**].

$$\hat{\pi}_{\delta,(k)}^{(m+1)} = \sum_{j \in G_k} \tau_{j\delta}^{(m)} / n_k \tag{3.42}$$

Model selection criteria

There are two kinds of commonly used model selection criteria, i.e., likelihood-based methods [171–173] and approximation-based methods [174–176].

Likelihood-based methods include bootstrap [171; 172] and cross-validation [173], where cross-validation methods can be further divided into many different alternatives depending on how the partitions are chosen [173]. Bootstrap method, although reported to be able to solve problems such as those with small data size and simple underlying structure [171; 172], is shown to slightly bias the number of clusters towards the null hypothesis [172]. Cross-validation method can solve the model selection problem in finite mixture model clustering by utilizing any scoring function that measures the fitness between the data and the model [173]. However, it is inefficient in data usage in the sense that the log-likelihood is estimated based on models that are trained from partial instead of the whole data set [173]. Further, likelihood-based methods are generally computationally expensive [173] which, thus, although applaudable in some clustering applications [171–173], are not widely used in this field.

Approximation-based methods, computationally efficient and simple, are preferred by most people, although the qualities of the results are subject to the underlying approximations, theoretically [173]. These methods include closed-form approximations to the Bayesian solution, Monte Carlo sampling of the Bayesian solution, and penalized likelihood methods [173]. The first two classes both adopt Bayesian approach which treats the number of components, g, as a parameter and obtains its posterior distribution based on the data and the model [173]. Since this posterior is often difficult to be obtained in the closed form, it either has to be approximated analytically (closed-form approximations to the Bayesian solution) or estimated via sampling techniques such as Markov chain Monte Carlo (MCMC) method [173].

Both of these two methods have been successfully applied in solving model selection problems [177; 178]. Penalized likelihood methods are typically derived from approximations based on asymptotic arguments as the data size goes to infinity which, strictly speaking, also approximate to the Bayesian solution [173; 179]. As stated by their names, penalized likelihood methods simply penalize the log-likelihood by an additive factor, which makes them much easier to be implemented compared with other methods [173]. There are many different penalization methods, such as Bayesian information criterion (BIC) [131; 175], integrated classification likelihood-BIC (ICL-BIC, simplified as ICL in this thesis) [174], Akaike information criterion (AIC) [180], and modified AIC (such as AIC3 [181]), all of which are reported to work well in certain applications.

In [Publication **II**] to [Publication **IV**] and [Publication **VI**], four well-known approximation-based model selection criteria, i.e., BIC [131; 175], ICL [174], AIC [176; 180], and AIC3 [176; 181] are compared in each finite mixture model. The formulas are given in Equations 3.43 to 3.46, where $-2\sum_{j=1}^{n}\sum_{\delta=1}^{g}\tau_{j\delta}\log(\tau_{j\delta})$ is the estimated entropy of the fuzzy classification matrix $\mathbf{C} = ((\tau_{j\delta}))$ [174], d is the number of free parameters, and M (in equations 3.43 and 3.44) is the total amount of the data. Recall that $M = \sum_{i=1}^{N} M_i$, where M_i is the size of data set i and N is the number of input data sets.

$$\mathrm{BIC} = -2\log L(\hat{\theta}) + d\log(nM) \tag{3.43}$$

$$\mathrm{ICL} = -2\log L(\hat{\theta}) + d\log(nM) - 2\sum_{j=1}^{n}\sum_{\delta=1}^{g}\tau_{j\delta}\log(\tau_{j\delta}) \tag{3.44}$$

$$\mathrm{AIC} = -2\log L(\hat{\theta}) + 2d \tag{3.45}$$

$$\mathrm{AIC3} = -2\log L(\hat{\theta}) + 3d \tag{3.46}$$

The number of free parameters, d, is distinct in different models, with the ones used in this thesis listed in Equations 3.47 to 3.54. Recall that p_1 to p_4 each represents the second dimension of beta, Gaussian, gamma, Bernoulli distributed data, respectively, g is the number of clusters, and K stands for the number of stratified layers in sBGMM. Specifically, Equation 3.47 is used in [Publication **II**], Equation 3.48 is adopted in [Publication **II**] to [Publication **IV**], Equation 3.49 is employed in [Publication **VI**], Equation 3.50 is utilized in [Publication **IV**] and [Publication **VI**], Equation 3.51 is implemented in [Publication **II**] and [Publication **III**], Equation 3.52 is applied in [Publication **III**], Equation 3.53 is employed in [Publication **IV**], and Equation 3.54 is used in [Publication **VI**].

$$\begin{aligned}
\text{BMM}: \quad d &= 2p_1 g + g - 1, && (3.47)\\
\text{GMM}: \quad d &= p_2 + p_2 g + g - 1, && (3.48)\\
\Gamma\ \text{MM}: \quad d &= 2p_3 g + g - 1, && (3.49)\\
\text{BerMM}: \quad d &= p_4 g + g - 1, && (3.50)\\
\text{BGMM}: \quad d &= 2p_1 g + p_2 + p_2 g + g - 1, && (3.51)\\
\text{sBGMM}: \quad d &= 2p_1 g + p_2 + p_2 g + K(g-1), && (3.52)\\
\text{GBMM}: \quad d &= p_2 + p_2 g + p_4 g + g - 1, && (3.53)\\
\Gamma\ \text{BMM}: \quad d &= 2p_3 g + p_4 g + g - 1, && (3.54)
\end{aligned}$$

Chapter 4

Application

A gene regulatory network (GRN) is a dynamical ensemble of multiple cellular components. Thus, monitoring a GRN's dynamics allows us observing subtle intracellular variations during gene expression, facilitating our understanding of the GRN and its regulatory mechanisms.

[Publication **VI**] extends and applies the clustering framework introduced in Chapter 3 to study the dynamics of delayed stochastic GRN via detecting their noisy attractors. [Publication **V**] and [Publication **VII**] are the supportive materials to [Publication **VI**], where [Publication **V**] provides the biological and validation basis of [Publication **VI**], and [Publication **VII**] points out one future direction in this research.

This chapter presents the background of the problem studied in [Publication **V**] to [Publication **VII**] by introducing the key concepts appeared, data sources explored, the algorithm used to drive the simulation, and the GRNs that are investigated. The validation results and biological findings of these publications are summarized in Chapter 5 and available in the publications.

4.1 Key concepts

The joint clustering framework is applied for noisy attractor detection from multiple data sources in [Publication **VI**]. [Publication **V**] provides the biological and validation bases for this study by exploring the stochasticity of a GRN and the variability of the distribution of cell differentiation pathway choices affected by parameter tuning. By investigating the important regulatory roles played by proteins' functionalities in gene regulation, [Publication **VII**] points out one future direction of studies in [Publication **VI**], i.e., applying the current noisy attractor detection technique to GRNs where proteins with various degrees of functionalities are involved.

The key concept used in all the three publications is 'cell differentiation', and the ones only appeared in [Publication **VI**] are 'ergodic set' and 'noisy

attractor', which are defined below.

Definition 7 (Cell differentiation)**.** *Cell differentiation can be viewed from cell level and population level. The cell level involves progressive morphological and biochemical changes such that a cell becomes functionally specialized with the passage of time. The population level refers to specialization of two or more cells in a multicellular organism which may trend different paths. In the population level, cells undergo nuclear, cytoplasmic and biochemical changes to establish different developmental patterns [182].*

Definition 8 (Ergodic Set)**.** *Ergodic set is a closed set of state cycles such that each state cycle can be reached by a single or more gene(s)' mistakes caused by internal noise from any other state cycle of the same ergodic set [29].*

Definition 9 (Noisy Attractor)**.** *Noisy attractor is the set of states where a gene regulatory network spends most of the time when on that ergodic set [29].*

4.2 Data sources

Three types of data sources are used when applying the data fusion method to noisy attractor detection, which are the time series of RNA, protein and promoter states (the 'state' here is defined as being or not bound to a transcription factor). The data are generated from *SGNSim* [183], whose dynamics is driven by the delayed Stochastic Simulation Algorithm ('delayed SSA') [184].

4.2.1 RNA time series

Transcriptional regulation is one of the most important steps in gene regulation [40; 185]. Since RNAs can be translated into proteins, it is assumed that the RNA abundance of a gene is predictive of the corresponding protein's activity. Thus, the RNAs of the genes encoding transcriptional regulatory proteins can be used to model the regulatory mechanisms within a GRN [185]. With this assumption, gene expression data is widely used for this purpose [129; 186; 187] and computational models at different granularity levels are developed, ranging from coarse-grained clusterings of co-regulated genes [129–133], to Boolean networks' binary representation of gene relationships [188; 189], and further to fully parameterized stochastic models of biochemical kinetics [186; 187].

4.2.2 Protein time series

More and more evidence suggest that it is inefficient to analyze regulatory molecules' activities within a GRN using RNA time series. For example, only around 20% transcription factor (TF) RNAs' profiles correlate with the expression levels of their targets in *Escherichia coli* and *Saccharomyces cerevisiae* [190]. Thus, protein time series has been used as the default data source in many explorations on GRNs' regulatory mechanism [4; 29; 191; 192], with the rational that proteins are the final products of genetic information and control the cellular matter and energy flows [3]. Experimentally, protein time series can be obtained from, e.g., protein microarrays [193].

It is reported that proteins may not be functional even after maturation [194]. Alternatively, depending on how 'function' is defined, non-functional proteins can be viewed as proteins with lower degrees of functionalities. According to [194], functions of a matured protein, i.e., after post-translational modification, can be classified as 'specific', 'conditional' and 'general'. In particular, specific functions are those that a protein is specifically adapted to, conditional functions require certain conditions to induce their activities, and general functions are characteristic of their general features which can be ascribed to all proteins, e.g., maintaining the Donnan equilibrium [194]. Thereby, it may be important to distinguish proteins with different degrees of functionalities, especially when cells deliberately regulate gene expression via tuning proteins' functionalities (see studies in [Publication **VII**]). Further, the emergence of functional protein microarrays [193] makes the detection of functional proteins experimentally feasible, facilitating studies at this granularity level.

4.2.3 Promoter states

Although RNA and protein time series cover information from the key cellular regulatory steps, i.e., transcription and translation, neither of them offers the explicit evidence whether or not two genes (interact via their products) directly interact. This can be achieved by complementing RNA and/or protein time series with information of promoter states, i.e., data containing binary values that indicate whether a gene's promoter is bounded by its TF(s) or not ('1' represents 'bind', '0' stands for 'non-bind'). This information source, although rarely available, can be obtained indirectly *in vivo* via, e.g., cloning a reporter gene (such as fluorescent protein encoding gene) into the expression vector of the target gene [195; 196].

4.3 Algorithms, strategies and GRN modules

In [Publication **V**] to [Publication **VII**], a modified version of the stochastic simulation algorithm (SSA) [197], i.e., the delayed SSA [184], is used to drive the dynamics of GRNs that are studied, where the translations and transcriptions are all modeled as time delayed reactions. In the following subsections, the SSA, the delayed SSA, the modeling strategies, and the delayed stochastic GRNs explored in [Publication **V**] to [Publication **VII**] are described in detail, respectively.

4.3.1 Stochastic simulation algorithm

The SSA [78; 197], a dynamic Monte Carlo simulation of the chemical master equations, is a computational algorithm that numerically simulates chemical reactions. It attains temporal stochastic dynamics by calculating the probability of each possible chemical reaction event and the resulting changes in the number of each molecular species at a certain moment [78; 197]. The mathematical derivation and formulation of the SSA are described below [198; 199].

Define τ ($\tau \in [0, \infty)$) and i as the 'time to the next reaction' and the 'index of the next reaction', respectively, and let $\mathbf{x}$ be the current system state. Then, the probability that the next reaction in the system will occur in the infinitesimal time interval $[t+\tau, t+\tau+\mathrm{d}\tau)$ and will be an R_i reaction is $p(\tau, i|\mathbf{x}, t)\mathrm{d}\tau$. It can be factorized as Equation 4.1,

$$p(\tau, i|\mathbf{x}, t)\mathrm{d}\tau = P_0(\tau|\mathbf{x}, t)a_i(\mathbf{x})\mathrm{d}\tau, \tag{4.1}$$

where $P_0(\tau|\mathbf{x}, t)$ is the probability that no reactions occur during $[t, t+\tau)$, and $a_i(\mathbf{x})\mathrm{d}\tau$ represents the probability that reaction R_i occurs in the infinitesimal time interval $[t+\tau, t+\tau+\mathrm{d}\tau)$. Note that $a_i(\mathbf{x})$ is the propensity function of reaction R_i when the system state is $\mathbf{x}$, which is usually expressed as the product of the probability reaction constant and a certain combination of the available reactants of R_i.

Now the problem is how to solve $P_0(\tau|\mathbf{x}, t)$. Following the above definitions and logic, the probability that no reaction occurs in $[t, t+\tau+\mathrm{d}\tau)$, $P_0(\tau+\mathrm{d}\tau|\mathbf{x}, t)$, can be factorized as Equation 4.2, where the two terms on the righthand side represent that no reaction occurs in $[t, t+\tau)$ and $[t+\tau, t+\tau+\mathrm{d}\tau)$, respectively, and n is the number of reactions.

$$P_0(\tau+\mathrm{d}\tau|\mathbf{x}, t) = P_0(\tau|\mathbf{x}, t)(1 - \sum_{i=1}^{n} a_i(\mathbf{x})\mathrm{d}\tau) \tag{4.2}$$

After simple algebraic rearrangements, Equation 4.2 becomes Equation 4.3, which can be further reformulated as Equation 4.4 by taking the limit $\mathrm{d}\tau \rightarrow$

0. Note that $a_0(\mathbf{x}) = \sum_{i=1}^{n} a_i(\mathbf{x})$.

$$\begin{aligned} \frac{P_0(\tau + \mathrm{d}\tau|\mathbf{x}, t) - P_0(\tau|\mathbf{x}, t)}{\mathrm{d}\tau} &= -a_0(\mathbf{x}) P_0(\tau|\mathbf{x}, t) \qquad (4.3) \\ \lim_{\mathrm{d}\tau \to 0} \frac{P_0(\tau + \mathrm{d}\tau|\mathbf{x}, t) - P_0(\tau|\mathbf{x}, t)}{\mathrm{d}\tau} &= -a_0(\mathbf{x}) P_0(\tau|\mathbf{x}, t) \\ \frac{\mathrm{d}P_0(\tau|\mathbf{x}, t)}{\mathrm{d}\tau} &= -a_0(\mathbf{x}) P_0(\tau|\mathbf{x}, t) \qquad (4.4) \end{aligned}$$

Thus, $P_0(\tau|\mathbf{x}, t) = \exp(-a_0(\mathbf{x})\tau)$, given the initial condition $P_0(\tau = 0|\mathbf{x}, t) = 1$. Finally, the probability defined in Equation 4.1 becomes

$$p(\tau, i|\mathbf{x}, t) = a_i(\mathbf{x}) \exp(-a_0(\mathbf{x})\tau). \qquad (4.5)$$

Equation 4.5, on the other hand, can be written in the form of Equation 4.7, where $p_1(\tau|\mathbf{x}, t)$ and $p_2(i|\tau, \mathbf{x}, t)$ are the probabilities that the next reaction occurs in $[t+\tau, t+\tau+\mathrm{d}\tau)$ regardless of which reaction it is, and that the next reaction occurring during this time interval is R_i, respectively.

$$\begin{aligned} p(\tau, i|\mathbf{x}, t) &= a_0(\mathbf{x}) \exp(-a_0(\mathbf{x})\tau) \times \frac{a_i(\mathbf{x})}{a_0(\mathbf{x})} \qquad (4.6) \\ &= p_1(\tau|\mathbf{x}, t) p_2(i|\tau, \mathbf{x}, t) \qquad (4.7) \end{aligned}$$

Also, it is seen from Equation 4.6 that τ is an exponential random variable with mean and standard deviation $\frac{1}{a_0(\mathbf{x})}$, and i is a statistically independent integer random variable with point probabilities $\frac{a_i(\mathbf{x})}{a_0(\mathbf{x})}$.

Gillespie developed two different but equivalent formulations to implement the exact SSA, i.e., the direct method and the first reaction method [78]. Also, many other methods are developed to implement the exact SSA, including, e.g., the next reaction method [200], the optimized direct method [201] and the sorting direct method [202]. Among others, the direct method is found to be effective in most cases [201], whose derivations and implementation procedures are described below [198].

From Equations 4.6 and 4.7, the probability distribution functions of τ and i are derived as Equations 4.8 and 4.9, respectively.

$$\begin{aligned} F_1(\tau|\mathbf{x}, t) &= \int_0^{\tau} p_1(\tau'|\mathbf{x}, t)\mathrm{d}\tau' \\ &= 1 - \exp(-a_0(\mathbf{x})\tau) \qquad (4.8) \\ F_2(i|\tau, \mathbf{x}, t) &= \sum_{i'=1}^{i} p_2(i'|\tau, \mathbf{x}, t) \\ &= \frac{1}{a_0(\mathbf{x})} \sum_{i'=1}^{i} a_{i'}(\mathbf{x}) \qquad (4.9) \end{aligned}$$

Equation 4.8 becomes Equation 4.10 after reformulation, and from Equation 4.9, there exists a largest r_2 that satisfies Formula 4.11,

$$\begin{aligned} \tau &= \frac{1}{a_0(\mathbf{x})} \ln(\frac{1}{1-F_1}) \\ &= \frac{1}{a_0(\mathbf{x})} \ln(\frac{1}{r_1}), \qquad (4.10) \\ r_2 &\leq \frac{\sum_{i'=1}^{i} a_{i'}(\mathbf{x})}{a_0(\mathbf{x})}, \qquad (4.11) \end{aligned}$$

where $r_i \in [0,1]$ $(i \in \{1,2\})$. Equation 4.11 can be further written as Equation 4.12.

$$i = \text{the smallest integer satisfying} \sum_{i'=1}^{i} a_{i'}(\mathbf{x}) > r_2 a_0(\mathbf{x}) \qquad (4.12)$$

Thus, the time interval $[t, t+\tau)$ and the reaction R_i can be determined according to Equations 4.10 and 4.12, respectively, by drawing random numbers r_1 and r_2 from the uniform distribution in the unit-interval.

The procedure for constructing a numerical realization of the SSA process with the direct method is shown below [198].

- Initialization Step:
  - Set $t = t_0$ (t_{stop} can also be specified);
  - Set $\mathbf{x} = \mathbf{x}_0$ (includes, e.g., the number of each molecule in the system and the reaction constants).
- Monte Carlo Step:
  - With the system at time t and in state $\mathbf{x}$, evaluate all the $a_i(\mathbf{x})$'s and their sum $a_0(\mathbf{x})$;
  - Generate values for τ and i using Equations 4.10 and 4.12.
- Update Step:
  - Set $t \leftarrow t + \tau$;
  - Set $\mathbf{x} \leftarrow \mathbf{x} + \nu_i$ (update the number of molecules according to ν_i, which is the state-change vector of reaction R_i).
- Iteration Step:
  - If the number of reactants is not zero or the time has not been exceeded when t_{stop} is set, record $(\mathbf{x}, t)$ and iterate from 'Monte Carlo Step';

  - Otherwise, end the simulation.

Note that different implementation methods share similar simulation steps, but differ in the 'Monte Carlo Step'.

Although the Gillespie exact SSA can simulate the reactions of chemical or biochemical systems efficiently, it is computationally expensive since only one reaction is allowed in each step which can be very small for large-scale systems [199]. Many adapted techniques are developed to overcome this problem, which generally obtain implementations at large timescales via compromising the exactitude of the theorem behind the algorithm. These adapted techniques can be roughly classified into two categories based on the strategy each method adopts [199]. Methods of the first category usually use larger step size to allow several reactions take place, such as Poisson τ-leap method [203] and the binomial leap methods [204]. In the second category, chemical reaction systems are partitioned into different subsystems (e.g., slow, intermediate, and fast subsystems); while using SSA in the slow subsystem, different approximations and techniques such as various leap methods, chemical Langevin equations, and reaction rate equations are applied to the other subsystem(s) [199].

4.3.2 Delayed SSA

While the SSA is proven to be reasonable for modeling discrete molecular events, the delayed SSA [184], which introduces time delays in these reactions, is developed to model multi-step reactions [184]. The delayed SSA uses a waiting list, i.e., a list of elements and the time intervals needed for them to be released, to account for the time delay, whose procedure is shown below [192].

- Initialization Step:
  - Set $t = t_0$ (t_{stop} can also be specified);
  - Set $\mathbf{x} = \mathbf{x}_0$ ($\mathbf{x}$ includes, e.g., the number of each molecule in the system and the reaction constants);
  - Form a group of input events and a separate group of output events from the list of reactions;
  - Create an empty waiting list L for delayed output events.
- Monte Carlo Step:
  - Do a 'Monte Carlo Step' of SSA for the input events to get the next reacting event R_i and the corresponding occurrence time t_i.
- Update Step:

  - Compare t_i with the least time, τ_{min}, in L.
    * If $t_i < \tau_{min}$:
      · Generate the delay τ_i for R_i (note that several delayed output events may exist for one input event, and how τ_i is generated depends on R_i and G_i);
      · Set $t \leftarrow t + t_i$;
      · Set $\mathbf{x} \leftarrow \mathbf{x} + \nu_i$ (ν_i is formed by performing R_i);
      · If $\tau_i = 0$: perform the output event G_i;
      · Decrement the delays in L by t_i;
      · If $\tau_i \neq 0$: add $\{G_i, \tau_i\}$ into L.
    * If $t_i > \tau_{min}$:
      · Set $t \leftarrow t + \tau_{min}$;
      · Set $\mathbf{x} \leftarrow \mathbf{x} + \nu_{min}$ (ν_{min} is formed by performing the output event G_{min}, which is associated with τ_{min});
      · Delete $\{G_{min}, \tau_{min}\}$ from L;
      · Decrement the delays in L by τ_{min}.

- Iteration Step:
  - If the number of reactants is not zero or the time has not been exceeded when t_{stop} is set, record $(\mathbf{x}, t)$ and iterate from 'Monte Carlo Step';
  - Otherwise, end the simulation.

The validity of modeling GRNs with delayed SSA has been verified by matching the model with the measurements at the single molecule (such as RNA, protein) level [192; 205]. It is a better modeling approach compared with the delayed differential equations (DDE) due to its stochasticity [206], and can more precisely model a GRN's noise than stochastic differential equations (SDE) where the distribution of the noise term has never been fixed and validated (e.g., the noise term of protein time series does not necessarily follow Gaussian distribution [205]). Further, the parameters used in delayed SSA can be measured directly or indirectly from the real biological systems, while most parameters of methods such as DDE and SDE are obtained by model fitting and could not be physically measured. This makes delayed SSA currently a more realistic modeling approach compared with its alternatives.

4.3.3 Modeling strategies for delayed stochastic GRNs

GRNs are complex dynamical systems. Besides comprising many complex cellular processes such as transcription and translation, various regulatory

events and molecular interactions occur in a GRN [4; 12]. Given this complexity, appropriate approximation methods are required to allow studying GRNs' dynamics within a single system [4]. In [Publication **V**] to [Publication **VII**], the GRNs are all modeled with the modeling strategies proposed in [4] which, while keeping the model as realistic as possible, can diminish the model complexity dramatically.

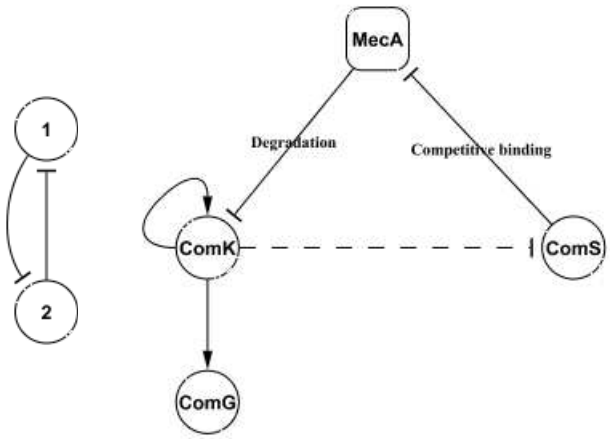
Specifically, these strategies [4] are summarized below:

- Consider reactions' delays and GRNs' noise:
  - Using the delayed SSA to drive a system's dynamics, which takes into account the time delays and large fluctuations caused by high molecular noise.
- Reduce GRNs' complexity:
  - Skipping the whole cascade of events, and assuming the gene products are fed back into the GRN.
  - Skipping the multiple steps in transcription and translation, and modeling them as single delayed reactions.
- Construct genes' relationships:
  - Each gene product is bound to another gene's promoter, playing either an active or inhibitive role. A regulatory function is then assigned to each of such connections, determining the gene's next state based on its inputs' states.
    - Allowing reactions among gene products, assigning the created complex molecules to a random gene promoter binding site, and choosing what regulatory effect this complex has on the expression level of its binding gene.
    - Allowing each gene to have multiple binding sites, and randomly assigning the regulatory effects to all the possible combinations of the binding events.
    - Allowing different gene products to bind to a single binding site, each with its own regulatory effect.

4.3.4 Delayed stochastic models of GRNs

In this thesis, three gene regulatory modules, the toggle switch (TS), the MeKS module from *Bacillus subtilis* [76], and a modified TS model are used or developed under the delayed stochastic framework using strategies presented in [4]. Specifically, TSs are used in [Publication **V**] and [Publication **VI**], the MeKS module is employed in [Publication **VI**], and the modified TS model is presented and utilized in [Publication **VII**]. The structures

of the TS (or the modified TS model) and the MeKS module are shown in Fig. 4.1.



(a) (b)

Figure 4.1: The structures of (a) TS (or modified TS) and (b) MeKS module. The round nodes are proteins (or dimers), and the round rectangle node represents protein complex. An arrow head represents activation and a 'T' shape head stands for repression. The solid and dotted lines represent the direct and indirect reactions, respectively. Both GRNs are drawn using Cytoscape [34]. The MeKS module is drawn based on [76].

The reactants and reaction rates that are used in most or all the studied GRNs are listed in Table 4.1. Specifically, each TS or modified TS is assumed to have n nodes. P_i, R_i and Pro_i each represents the protein, RNA and promoter of gene i, where $i \in \{1, 2, \ldots, n\}$ in the (modified) TSs and is a string representing a gene/protein name in the MeKS module. Pro_iP_j is the binding complex of protein j (short for 'being encoded by gene j') and the promoter of gene i ($i \neq j$). Also, the basal transcription rate and protein decay rate are symbolized as k_t and k_d, respectively, with $k_{t,genename}$ and $k_{d,proteinname}$ employed if they differ in their values for different genes or gene products in a GRN. The translation rate is denoted as k_{tr}. $rbsd$ is used to represent the RNA decay rate. k_{rep} and k_{unrep} are defined as the binding and unbinding rates of a repressor to a gene, respectively. In the differentiation models, X denotes the essential molecule required for differentiation, and is the parameter controlling differentiation time. $k_{x,modelname}$ and $k_{diff,modelname}$ are used to model the rate for a differentiation event to occur and the rate of the differentiation process, respectively. RNA poly-

merase and ribosome, represented as $RNAp$ and Rib, respectively, are explicitly modeled in (modified) TSs. This is because according to [184; 192], variations of the copy numbers of key molecules, such as genes, RNA polymerase and ribosome, can cause stochastic fluctuations and consequently the change of a GRN's dynamics if the amount of the molecules is limited. Also, it is pointed out that, e.g., in *Escherichia coli*, the free RNA polymerase concentration is found to be 28 molecules per cell at any given moment under normal conditions [207]. Thus, small deviations from the normal conditions, such as increased transcription due to stress, in systems as small as such could cause non-negligible variations in the networks' dynamics. For the sake of simplicity and accuracy, RNA polymerase and ribosome are both modeled in (modified) TSs where only two genes are involved.

Symbol	Meaning
P_i	Protein of gene i
R_i	RNA of gene i
Pro_i	Promoter of gene i
Pro_iP_j	Binding complex of gene j's protein with gene i's promoter
$RNAp$	RNA polymerase
Rib	Ribosome
k_t	Transcription rate
k_d	Protein decay rate
$k_{t,genename}$	Transcription rate of gene 'genename'
$k_{d,proteinname}$	Protein decay rate of protein 'proteinname'
k_{tr}	Translation rate
$rbsd$	RNA decay rate
k_{rep}	Binding rate of a repressor to a gene
k_{unrep}	Unbinding rate of a repressor to a gene
$Cell_0$	Stem cell or undifferentiated cell
$Cell_i$	Differentiated cell types ($i \neq 0$)
X	Essential molecule for cell differentiation to occur
$k_{x,modelname}$	Rate for a differentiation event to occur in model 'modelname'
$k_{diff,modelname}$	Rate of a differentiation process in model 'modelname'

Table 4.1: Reactants and reaction rates that are present in all the studied GRNs or differentiation models. Each GRN is assumed to have n nodes. $i \in \{1, 2, \ldots, n\}$ in the (modified) TSs and is a string representing a gene/protein name in the MeKS module. The symbols listed in the upper and lower boxes are present in the GRNs and differentiation models, respectively.

The reactants and reaction rates that only appear in the MeKS module or the modified TS are shown in Table 4.2. In the MeKS module, due to the involvement of the protein complex MecA, Pro_KMecA is used to represent the binding complex of MecA and gene $comK$'s promoter, $k_{deg,MecA}$ is used to symbolize the degradation rate of MecA induced by protein ComS's competitive binding, and $k_{deg,K}$ is adopted to stand for the degradation rate of

protein ComK caused by MecA. In the modified TS, due to the discrimination of various degrees of gene products' functionalities, one more subscript indicating functional or not is added to proteins and dimers. In particular, $P_{f,i}$, $P_{nf,i}$, $D_{f,i}$ and $D_{nf,i}$ represent the functional and non-functional proteins and dimers of gene i ('non-functional dimers' here refer to dimers that can not repress gene expression, but may or may not preserve the other functions). Unlike in the regular TSs, P_i represents pre-mature proteins in the modified TSs, which can become functional or non-functional proteins after processes such as post-translational modification. Also, k_{trans} is used to denote the transformation rate for a nascent protein to become mature, either functional or non-functional, and k_{dimer} and $k_{undimer}$ are used to represent the dimerization and dedimerization rates. Further, $Pro_iD_{f,j}$ and $Pro_iD_{nf,j}$ each stands for the binding complex of the i^{th} gene's promoter with the j^{th} gene's functional and non-functional dimer.

Symbol	Meaning
MecA	Protein complex MecA
ComI	Protein ComI, I$\in$ {K, S, C, E, F, G}
$comI$	Gene $comI$, $I \in \{K, S\}$
$k_{deg,K}$	Degradation rate of ComK caused by MecA
Pro_KMecA	Binding complex of MecA with the promoter of gene $comK$
$k_{deg,MecA}$	Degradation rate of MecA caused by competitive binding of ComS
$k_{deg,K}$	Degradation rate of ComK caused by MecA
P_i	Pre-mature protein of gene i
$P_{f,i}$	Functional protein of gene i
$P_{nf,i}$	Non-functional protein of gene i
$D_{f,i}$	Functional protein dimer of gene i
$D_{nf,i}$	Non-functional protein dimer of gene i
$Pro_iD_{f,j}$	Binding complex of gene j's functional dimer with gene i's promoter
$Pro_iD_{nf,j}$	Binding complex of gene j's non-functional dimer with gene i's promoter
k_{trans}	Transformation rate for a pre-mature protein to become (non-)functional
k_{dimer}	Dimerization rate
$k_{undimer}$	Dedimerization rate

Table 4.2: Reactants and reaction rates that appear only in the MeKS module or the modified TSs or have different meanings with those in the regular TSs. The modified TS is assumed to have n nodes, and $i \in \{1, 2, \ldots, n\}$ with $i \neq j$ if both subindexes are present. The symbols listed in the upper and lower boxes are present in the MeKS module and the modified TSs, respectively.

The reaction rates, if present in more than one models, differ among different models, which are set together with the system initiation state

(i.e., the amount of each molecule in the system when the simulation starts) to reproduce the dynamics of each GRN *in vivo*. The concentrations of the substrates are assumed to be invariable, whose effects are taken into account in the corresponding reaction rates. The parameter setting of each model is available in each publication.

The time delays for the molecules to be present after the corresponding reactions have occurred are represented by τ_js ($j \in \{1, 2, \ldots, 5, 5std\}$) in the modeled GRNs, as shown in Table 4.3. The delay parameters are set to match the experimental measurements in *Escherichia coli* [205], and are fixed in this thesis due to their low variability between transcriptions of single genes [208]. Specifically, τ_1 accounts for the time needed for a promoter to turn on its open state, and is set to 40s according to [208]. τ_2 is the time for the RNA polymerase to be detached from the gene, which is considered as τ_1 plus the transcriptional elongation time [192]. Since gene *tsr-venus* has 2500 nucleic acids [205] and the average elongation rate of transcription in *Escherichia coli* is approximately $50nt/s$ [192], $\tau_2 = \tau_1 + 2500/50 = 90s$. Similarly, τ_3, the time for the clearance of ribosome binding site in mRNA during translation initiation, is set to 2s according to [209]; and τ_4, the time for the ribosome to be released into the cytosol and become free, is set to $\tau_4 = \tau_3 + 2500/45 = 58s$, given $15nt/s$ as the average translation rate [192]. The time of a protein's assembly process after its translation follows a Gaussian process [192], whose mean and standard deviation are set to $\tau_5 = 420s$ and $\tau_{5std} = 140s$, respectively, according to the experiments [205].

Symbol	Meaning
τ_1	Time delay for the closed promoter complex to become open
τ_2	Time delay for the RNA polymerase to be detached from the gene
τ_3	Time delay for ribosome binding site clearance during translation initiation
τ_4	Time delay for ribosome to be released into the cytosol
τ_5	The mean time of a protein to be assembled after its production
τ_{5std}	The standard deviation of the time for a protein to be assembled after its production

Table 4.3: Time delays in delayed stochastic GRNs.

Each model is composed of a set of chemical reactions, which are shown, separately, below for each module.

TS model

Cell differentiation has long been hypothesized to be regulated by bistable genetic sub-circuits controlling many downstream genes [210]. During this process, a stem cell turns into a stable cell type which, by hypothesis, corresponds to stable states of the GRN [2]. Such genetic differentiation deci-

sion sub-circuits must be at least coinstantaneously bistable to allow cells branching into distinct cell types and make them reliably acting as cellular memory units [191]. It is shown that the TS, a GRN of two genes mutually repressing each other (Fig. 4.1 (a)), can be used by cells to adopt different phenotypes [186; 191; 211] and as decision circuits of differentiation pathways [212].

In the delayed stochastic framework, TSs are modeled by Reactions 4.13 to 4.18 [192], where $i \in \{1, 2\}$ when only the sub-index i is present, or $i, j \in \{1, 2\}$ $(i \neq j)$ when both sub-indices are present.

$$Pro_i + RNAp \xrightarrow{k_t} Pro_i(\tau_1) + RNAp(\tau_2) + R_i(\tau_1) \tag{4.13}$$

$$R_i + Rib \xrightarrow{k_{tr}} R_i(\tau_3) + Rib(\tau_4) + P_i(\tau_5, \tau_{5std}) \tag{4.14}$$

$$Pro_i + P_j \underset{k_{unrep}}{\overset{k_{rep}}{\rightleftharpoons}} Pro_iP_j \tag{4.15}$$

$$Pro_iP_j \xrightarrow{k_d} Pro_i \tag{4.16}$$

$$R_i \xrightarrow{rbsd} \emptyset \tag{4.17}$$

$$P_i \xrightarrow{k_d} \emptyset \tag{4.18}$$

In a TS, each promoter Pro_i controls the expression of an RNA (R_i), which can then be translated by a ribosome (Rib) into a protein (P_i) as shown by Reactions 4.13 and 4.14. The binding and unbinding of the repressor to its target gene's promoter, which defines the TS, is shown by Reaction 4.15. The degradation of the protein-promoter complexes, RNAs, and proteins are modeled via Reactions 4.16 to 4.18, respectively. Note that in this model, a protein decays both in its free form and when it binds to a promoter, with the same rate.

The differentiation model used to study a GRN's response to its internal dynamic change ([Publication **V**] to [Publication **VII**]) or to validate the noisy attractor detection results ([Publication **VI**]) is given as Reactions 4.19 to 4.23, where $*$ means no consumption of the reactant [183].

$$\emptyset \xrightarrow{k_{x,TS}} X, \tag{4.19}$$

$$X + *P_1 + Cell_0 \xrightarrow{k_{dif,TS}} Cell_1 \tag{4.20}$$

$$X + *P_2 + Cell_0 \xrightarrow{k_{dif,TS}} Cell_2 \tag{4.21}$$

$$X + *P_1 + *P_2 + Cell_0 \xrightarrow{0.5 \times k_{dif,TS}} Cell_3 \tag{4.22}$$

$$X + Cell_0 \xrightarrow{0.5 \times k_{dif,TS}} Cell_4 \tag{4.23}$$

The scheme of the differentiation model is illustrated by Fig. 4.2, where the cell can choose among four pathways depending on the expression levels

of both genes. The rational here is that the destiny of a stem cell after differentiation is governed by the proteins that determine the noisy attractor the cell arrives at. This is supported, e.g., by the evidence that two protein families, Polycomb and Trithorax, work antagonistically to control the genome programming during differentiation [42]. In the context of neurogenesis and astrogenesis by which neurons and glias are generated, separately, the presence of the Trithorax group member, Mll1, and the shutting down of Polycomb group members are required for neurogenesis [213; 214], and the dominance of Polycomb group members over Trithorax proteins is needed for cells to differentiate into glias [214].

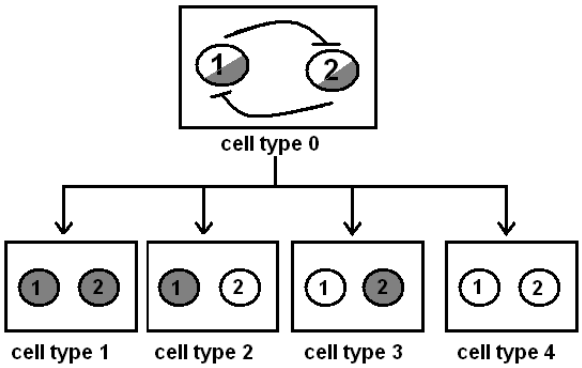


Figure 4.2: Schematic figure of the stochastic TS with four possible differentiation pathways (differentiate into cell type 1 to cell type 4), depending on the expression levels of the proteins during differentiation. Gray and white balls each represent genes with high and low expression levels, respectively. The half-white and half-gray balls show the toggling behavior of the genes' expression levels in the stem or mother cell (cell type 0). This figure is reproduced from [Publication **V**] with permission.

MeKS module

In *Bacillus subtilis*, while most cells sporulate, a minority becomes competent for gene uptake from the environment under nutrient limitation [76]. With extensive research and effort devoted to the exploration of this phenomenon, a detailed picture of the regulatory mechanism of this GRN is unveiled. It is found that a TF, namely ComK, plays a major role in this GRN. Specifically, it activates the expression of a series of genes that control competence, such as *comC*, *comE*, *comF* and *comG* [215–217], among which *comG* is of the most importance [218]. The gene *comK* maintains a basal expression level once entering its stationary phase [219], and can activate itself when sufficient amount of ComK is accumulated [220; 221]. MecA is a protein complex that degrades ComK, which prevents cells from

being competent [219]. Protein ComS can competitively bind to MecA, reducing the amount of MecA that can degrade ComK [219; 222]. Further, over-expression of *comK* suppresses the expression of *comS* [215], indicating a negative feedback loop and an attempt to escape from the competence state [76]. Thus, this GRN is mainly governed by three components, i.e., MecA, *comK*, and *comS*, which is, thereby, called the 'MeKS' module [76]. The structure of this MeKS module is depicted in Fig. 4.1 (b).

The MeKS module is modeled under the delayed stochastic framework by Reactions 4.24 to 4.36 in [Publication **VI**]. Specifically, Reactions 4.24 to 4.26 model the binding of the protein complex MecA to the promoter of gene *comK*, and the disassociation and decay of their binding complex, respectively.

$$Pro_K + MecA \xrightarrow{k_{rep}} Pro_K MecA \tag{4.24}$$

$$Pro_K MecA \xrightarrow{k_{unrep}} Pro_K + MecA \tag{4.25}$$

$$Pro_K MecA \xrightarrow{k_{d,MecA}} Pro_K \tag{4.26}$$

Reaction 4.27 models the competitive binding of protein ComS to MecA, which is simplified as the degradation of MecA by ComS. Reaction 4.28 shows the degradation of protein ComK by MecA.

$$P_S + MecA \xrightarrow{k_{deg,MecA}} P_S \tag{4.27}$$

$$P_K + MecA \xrightarrow{k_{deg,K}} MecA \tag{4.28}$$

Reaction 4.29 models the indirect repression of protein ComK to gene *comS*, where '$(invhill : 1, 2)$' represents the inverse of the hill function $\frac{X^b}{a^b+X^b}$ (X is P_K in this case) with $a = 1$ and $b = 2$ [183].

$$Pro_S + *P_K(invhill : 1, 2) \xrightarrow{k_{t,S}} Pro_S(\tau_1) + R_S(\tau_1) \tag{4.29}$$

Reaction 4.30 shows the transcription of *comK*, which maintains a basal expression level at the stationary state and whose auto-activation is inversely correlated with MecA. Note that '$(max : 1, 1)$' stands for the function $b \times \max(a, X)$ with parameters $a = b = 1$ [183], where X is P_K in this context.

$$Pro_K + *P_K(max : 1, 1) + *MecA(invhill : 1, 5) \xrightarrow{k_{t,K}} Pro_K(\tau_1) + R_K(\tau_1) \tag{4.30}$$

The translation of proteins ComS, ComK and ComG are modeled by Reactions 4.31 and 4.32, where the translation of proteins ComK and ComG are simplified by a single reaction since they share similar expression profiles

and differ only in delay.

$$R_S \xrightarrow{k_{tr}} R_S(\tau_3) + P_S(\tau_5, \tau_{5std}) \tag{4.31}$$

$$R_K \xrightarrow{k_{tr}} R_K(\tau_3) + P_K(\tau_5, \tau_{5std}) + P_G(4\tau_5, \tau_{5std}) \tag{4.32}$$

Reactions 4.33 and 4.34 each models the production and decay of the protein complex MecA. The decays of RNAs and proteins are modeled by Reactions 4.35 and 4.36, respectively, where 'i' represents 'S' and 'K' in Reaction 4.35, and stands for 'S', 'K' and 'G' in Reaction 4.36.

$$\emptyset \xrightarrow{k_{t,MecA}} MecA \tag{4.33}$$

$$MecA \xrightarrow{k_{d,MecA}} \emptyset \tag{4.34}$$

$$R_i \xrightarrow{rbsd} \emptyset \tag{4.35}$$

$$P_i \xrightarrow{k_d} \emptyset \tag{4.36}$$

Cell differentiation is modeled by Reactions 4.37 to 4.39, with a similar scheme as illustrated for the TS model (Fig. 4.2 can be referenced). Specifically, Reaction 4.38 represents the competent pathway since the activity of *comG* is reported to increase when cells become competent [76], and Reaction 4.39 is used to represent the sporulation pathway because, according to [76], as sporulation begins the expression of *comS* starts to increase.

$$\emptyset \xrightarrow{k_{x,MeKS}} X \tag{4.37}$$

$$*P_G + Cell_0 \xrightarrow{k_{dif,MeKS}} Cell_G \tag{4.38}$$

$$*P_S + Cell_0 \xrightarrow{k_{dif,MeKS}} Cell_S \tag{4.39}$$

Modified TS model

The modified TS model presented in [Publication **VII**] distinguishes from the regular TS by two important features. First, proteins need to dimerize, i.e., to form D, before any of their functionalities to be activated. Second, there exist non-functional dimers, referring to dimers that have lost their specific functions (i.e., gene repression) but may preserve several or all the other functions in this context, whose preserved functions are regulatable. Specifically, the modified TS model consists of Reactions 4.40 to 4.53.

$$Pro_i + RNAp \xrightarrow{k_t} Pro_i(\tau_1) + RNAp(\tau_2) + R_i(\tau_1) \quad (4.40)$$

$$R_i + Rib \xrightarrow{k_{tr}} R_i(\tau_3) + Rib(\tau_4) + P_i(\tau_5, \tau_{5std}) \quad (4.41)$$

$$Pro_i + D_{f,j} \underset{k_{unrep}}{\overset{k_{rep}}{\rightleftharpoons}} Pro_i D_{f,j} \quad (4.42)$$

$$Pro_i D_{f,j} \xrightarrow{k_d} Pro_i \quad (4.43)$$

$$R_i \xrightarrow{rbsd} \emptyset \quad (4.44)$$

$$P_i \xrightarrow{k_d} \emptyset \quad (4.45)$$

$$P_i \xrightarrow{k_{trans}} P_{f,i} \quad (4.46)$$

$$P_i \xrightarrow{k_{trans} \times f_i} P_{nf,i} \quad (4.47)$$

$$P_{f,i} \xrightarrow{k_d} \emptyset \quad (4.48)$$

$$P_{nf,i} \xrightarrow{k_d} \emptyset \quad (4.49)$$

$$2P_{f,j} \xrightarrow{k_{dimer}} D_{f,j} \quad (4.50)$$

$$D_{f,j} \xrightarrow{k_{undimer}} 2P_{f,j} \quad (4.51)$$

$$P_{f,j} + P_{nf,j} \xrightarrow{k_{dimer}} D_{nf,j} \quad (4.52)$$

$$D_{nf,j} \xrightarrow{k_{undimer}} P_{f,j} + P_{nf,j} \quad (4.53)$$

Further, to analyze how proteins' functionalities affect cell differentiation at cell population level, Reactions 4.54 to 4.58 are added to control the differentiation process (scheme references Fig. 4.2).

$$\emptyset \xrightarrow{k_{x,TS}} X \quad (4.54)$$

$$X + *D_{f,1} + Cell_0 \xrightarrow{k_{dif,TS}} Cell_1 \quad (4.55)$$

$$X + *D_{f,2} + Cell_0 \xrightarrow{k_{dif,TS}} Cell_2 \quad (4.56)$$

$$X + *D_{f,1} + *D_{f,2} + Cell_0 \xrightarrow{0.5 \times k_{dif,TS}} Cell_3 \quad (4.57)$$

$$X + Cell_0 \xrightarrow{0.5 \times k_{dif,TS}} Cell_4 \quad (4.58)$$

Notice the two differences in the basic modified TS model compared with the regular TS, i.e., Reactions 4.46 to 4.53 are added and proteins are replaced with functional protein dimers. In particular, Reactions 4.46 and 4.47 model the conversion of the protein P_i into its functional ($P_{f,i}$) and non-functional ($P_{nf,i}$) forms, respectively, where f_i controls the fraction of non-functional proteins produced by gene i. Reactions 4.48 and 4.49 model

the degradation of both forms of proteins. Reactions 4.50 to 4.53 model the composing and decomposing processes of functional and non-functional protein dimers.

Functional protein dimers are assumed to comprise three functions, i.e., 'recognizing and binding to the regulatory element', 'repressing gene expression', and 'being involved in the differentiation process'. Non-functional dimers here refer to dimers that can not repress gene expression, but may or may not preserve the other functions. Thus, the tuning of the degrees of protein dimers' functionalities is done by allowing non-functional dimers to have zero to two of the rest of the functions, which are 'recognizing sequence' and 'being involved in cell differentiation'. Let Reactions 4.59 to 4.61 be 'function set 1' and Reactions 4.62 to 4.66 be 'function set 2', each modeling the function of 'recognizing sequence' and 'being involved in cell differentiation' of non-functional dimers, respectively. Also, consider Reactions 4.40 to 4.58 as the basic reaction set. Then, four cases are analyzed, which are modeled by adding none, either or both of the two function sets to the basic reaction set, according to the preserved functions of the non-functional dimers in each case.

Function set 1:

$$Pro_i + D_{nf,j} \underset{k_{unrep}}{\overset{k_{rep}}{\rightleftharpoons}} Pro_i D_{nf,j} \tag{4.59}$$

$$Pro_i D_{nf,j} + RNAp \xrightarrow{k_t} Pro_i D_{nf,j}(\tau_1) + RNAp(\tau_2) + R_i(\tau_1) \tag{4.60}$$

$$Pro_i D_{nf,j} \xrightarrow{k_d} Pro_i \tag{4.61}$$

Function set 2:

$$X + *D_{nf,1} + Cell_0 \xrightarrow{k_{dif,TS}} Cell_1 \tag{4.62}$$

$$X + *D_{nf,2} + Cell_0 \xrightarrow{k_{dif,TS}} Cell_2 \tag{4.63}$$

$$X + *D_{f,1} + *D_{nf,2} + Cell_0 \xrightarrow{0.5\times k_{dif,TS}} Cell_3 \tag{4.64}$$

$$X + *D_{nf,1} + *D_{f,2} + Cell_0 \xrightarrow{0.5\times k_{dif,TS}} Cell_3 \tag{4.65}$$

$$X + *D_{nf,2} + *D_{nf,2} + Cell_0 \xrightarrow{0.5\times k_{dif,TS}} Cell_3 \tag{4.66}$$

Chapter 5

Discussion

5.1 Summary

With the aim of understanding the regulatory mechanisms of a gene regulatory network (GRN), this thesis explores its topology by developing data fusion methods to improve the accuracy of TFBS prediction and gene clustering, and investigates its dynamics via applying the developed clustering framework to detect such networks' noisy attractors. The structure of this thesis is illustrated by Table 5.1, with the details of each publication summarized, organized by chapters, below.

Method	*[Publication **I**]*	TFBS prediction
	*[Publication **II**]*	Gene clustering: BGMM
	[Publication **III**]	Gene clustering: sBGMM
	[Publication **IV**]	Gene clustering: GBMM
Application	[Publication **V**]	Theoretical and validation bases of [Publication **VI**]
	*[Publication **VI**]*	Noisy attractor detection
	[Publication **VII**]	Future direction of [Publication **VI**]

Table 5.1: Summary of the results. The backbone articles are shown in italic face.

5.1.1 Methods

[Publication **I**] to [Publication **IV**] explore the structure of a GRN at the DNA sequence level and gene level. In particular, [Publication **I**] develops a new data fusion method and, at the meanwhile, explores two novel information sources to improve TFBS prediction accuracy (TFBS prediction is defined in Definition 4 in Section 3.1.1). [Publication **II**] to [Publication **IV**] build unified joint model based clustering framework to group genes from multiple data sources. The results of [Publication **I**], i.e., protein-DNA bind-

ing probabilities, can be used as one input of the gene clustering problems shown in [Publication **II**] and [Publication **III**].

TFBS prediction

Publication I This publication presents a new data fusion method for combining multiple genome-level data sources, aiming at improving the accuracy of TFBS prediction. Also, it explores the extent to which two novel information sources, i.e., DNA duplex stability and nucleosome positioning data, can improve TFBS predictions, either alone or in combination with other data sources. By testing on a carefully constructed data set of verified binding sites from the mouse genome, three key conclusions are drawn. First, the new data fusion method outweighs its traditional data integration alternative in significantly reduced false positive rates. Second, DNA duplex stability and nucleosome occupation data are informative on improving TFBS prediction accuracy, especially when combined with other genome-level data sources, such as evolutionary conservation. Third, integrating non-redundant informative data sources can provide the most efficient data fusion.

Gene clustering

[Publication **II**] to [Publication **IV**] present methods of model based gene clustering (gene clustering is defined in Definition 6 in Section 3.2.1) from multiple data sources. This series of methods together show a flexible unified probabilistic modeling framework, which can be extended to integrate data of any parametric distribution in principle. One advantage of model based clustering algorithms is their automatic selection of the number of clusters, which is achieved, generally, by casting it as the model selection problem [136]. Comparison of four well-known model selection criteria, i.e., Akaike information criterion (AIC) [180], a modified AIC (AIC3) [176], Bayesian information criterion (BIC) [131; 175] and the integrated classification likelihood-BIC (ICL-BIC) [174], is done for each model, based on which the best criterion is selected for each clustering algorithm. All the results demonstrate the superiority of incorporating multiple data sources in gene clustering.

Besides the aforementioned 'points-of-parity', the 'points-of-difference' among the presented models are described below.

Publication II This publication presents a beta-Gaussian mixture model, namely BGMM, for clustering genes from beta and Gaussian distributed data. One typical application of this model is to cluster gene expression data and protein-DNA binding probabilities, assuming that genes clustered

together share similar expression profiles and their regulatory regions are bounded by the same or similar set of TFs. As the backbone article of the work done on gene clustering, [Publication **II**] presents the joint modeling framework systematically. Particularly, it introduces three versions of BGMM, i.e., standard, approximated and hybrid, based on their expectation maximization (EM) algorithm. In the E step of the EM algorithm, the standard method maximizes the expectation of the complete data log-likelihood, and the approximated method maximizes the complete data log-likelihood directly. In the standard and the approximated versions, all the parameters are estimated using the standard or the approximated EM method, and in the hybrid method, the parameters of the beta distribution are estimated approximately and those of the Gaussian distribution are updated following the standard procedure. All the models use the diagonal covariance matrix in the probability density function of the Gaussian distribution to reduce the number of estimated parameters, rendering the models much more efficient in dealing with large dimensional data. No significant performance difference is found among the three versions based on the simulation test. However, their best model selection criteria are different, i.e., ICL is selected for the standard version and AIC is chosen for the approximated and the hybrid versions. In [Publication **II**], the performance of BGMM is tested using the standard version via simulations and a real case application. Both of the results demonstrate the superiority of this joint clustering framework compared with its component models (i.e., beta mixture model and Gaussian mixture model). The real data comes from mouse, where gene expression data and protein-DNA binding data (obtained from TFBS prediction) are assumed to be Gaussian and beta distributed, respectively. The results not only show the performance improvement of BGMM but also identify three groups of synchronously regulated genes involved in the Myd88-dependent Toll-like receptor (TLR)-3/4 signaling cascade. Also, the mathematical derivations of the standard BGMM are presented in the appendix of [Publication **II**].

Publication III This publication presents a stratified beta-Gaussian mixture model, i.e., sBGMM. Besides integrating beta and Gaussian distributed data, this model also utilizes a third information source. This third information is used as the prior of the model, which can come from any sources such as protein-protein interaction (PPI) data. The parameters of sBGMM are estimated using the standard EM algorithm, and the covariance matrix employed in the probability density function of the Gaussian distribution is diagonal. Based on the simulation results, ICL is recommended as the safer choice in model selection. Besides presenting the algorithm itself, much effort of [Publication **III**] has been devoted to the performance test and real case application. Specifically, the algorithm is first tested by simulations, with

the results demonstrating the advantage of using the prior to guide the clustering procedure. Then, it is applied to a set of mouse data, assuming that gene expression data and protein-DNA binding probabilities are of Gaussian and beta distributions, respectively. Two real case studies are carried out, with one for performance test and the other for biological information finding. Two sets of priors are compared when testing the performance with the real data, i.e., the priors obtained from the network structures stored in the database TRED [223] and those from the online classification tool DAVID [224]. Besides showing the performance improvement of sBGMM compared with BGMM, the results also indicate that the best performance of sBGMM is achieved when the information used for prior construction is consistent with those stored in the other data sources. The prior used for biological information finding is derived from a cancer related network, i.e., NFKB network, from TRED. The results reveal two sets of genes that are oppositely regulated by eight TFs, i.e., the genes within each group are either activated by TFs 'E2f6', 'E2f7', 'Foxm1', 'Nfatc1' and repressed by TFs 'Rest', 'Rfx5', 'Mxd1', 'Stat1', or goes the other way around. Further, all the genes and TFs are found to be responsive to Myd88-dependent TLR-3/4 signaling. The mathematical derivations of the standard sBGMM are shown in the 'Methods' section of [Publication **III**].

Publication IV This publication presents a model for clustering Gaussian and Bernoulli distributed data, i.e., GBMM. One typical application of this model is to cluster genes with gene expression data and PPI data, assuming that genes within the same cluster share similar expression profiles and have on average more PPIs with a set of genes than genes from different clusters. The standard EM is used to update all the parameters, and the diagonal covariance matrix is employed in the probability density function of the Gaussian distribution. Simulation tests show that the clustering performance is highly improved after jointly utilizing two data sources, and the more known PPIs the better the results are. Moreover, AIC and AIC3 are shown to perform similarly and are both recommended for GBMM.

5.1.2 Application

[Publication **VI**] applies the multiple data fusion clustering framework to a biological problem in another research domain, i.e., noisy attractor detection, to study the dynamics of delayed stochastic GRNs. [Publication **V**] provides the theoretical and validation bases for this application. [Publication **VII**] investigates further the complex dynamic nature of a GRN, pointing out one future direction of [Publication **VI**]. In all these publications, *SGNSim* [183] is used to simulate the GRN models, where the dynamics is driven by the 'delayed SSA' [184].

Publication V This publication studies the plasticity of a delayed stochastic model of a toggle switch (TS) as a multipotent differentiation pathway switch, at the single cell and cell population levels. This study is done by varying the mean, noise, and bias of proteins' expression levels (via tuning the proteins' expression level and degradation rates) and observing distributions of differentiation pathways choices of genetically homogeneous cell populations. The results show that small changes in each of these dynamical features significantly and distinctively affect the dynamics of a single cell and the differentiation pattern of cell population.

[Publication **V**] shows that the stochastic TS has high plasticity regarding differentiation pathway choice regulation, providing the theoretical and validation bases for [Publication **VI**].

Publication VI This publication develops a model, called ΓBMM, under the joint clustering framework presented in [Publication **II**] to [Publication **IV**] to integrate gamma and Bernoulli distributed random variables, and applies it to a problem in a different research domain, i.e., determining the noisy attractors (noisy attractor is defined in Definition 9 in Section 4.1) of the delayed stochastic GRNs. Specifically, ΓBMM novels in its immunity to any empirical pre-assumptions of the number of state regimes and the consideration of multiple perspectives of a GRN when determining its noisy attractors which, compared with the conventional method (K-means) used in [29], can provide a much richer spectrum of the possible noisy attractors and capture the real dynamic variations of a GRN under a delayed stochastic setting. The parameters of the gamma distribution and the Bernoulli distribution in ΓBMM are both estimated by the standard EM algorithms. After computationally testing its performance, ΓBMM is applied to detect the noisy attractors of a TS and an excitable circuit from *Bacillus subtilis*, i.e., MeKS module [76]. The results are validated by cells' differentiation pattern(s) obtained from each corresponding differentiation model. Besides showing the accuracy of applying ΓBMM in noisy attractor detection, the results also reveal three transition states in the TS, i.e., bistable, tristable and monostable, toggling the long cherished belief that the TS has only two noisy attractors [191].

Publication VII This publication investigates how the regulation of proteins' functionalities affect the dynamics of a delayed stochastic GRN, i.e., a modified TS model, and the cell differentiation pattern it regulates at the level of cell population. This study is carried out by tuning the degrees of protein dimers' functionalities. Each protein that forms a dimer can be either functional or non-functional, leading to zero to two possible functions of a protein dimer, i.e., 'promoter binding', 'gene repression' and 'cell differentiation involvement'. Further, three factors are investigated to study

the effect of proteins' differential functionalities on cells' dynamics and their phenotypical variations after differentiation, which are 'the rate at which a protein becomes functional or non-functional', 'the fraction of non-functional proteins', and 'the bias towards producing more non-functional proteins in one sub-system of the TS' (a sub-system is defined as a gene and its products). Among others, the results show that altering the degree of proteins' functionalities is an important regulator of GRNs.

[Publication **VII**], exploring the dynamical consequences of the variation of the degree of proteins' functionalities in delayed stochastic GRNs, shows the necessity of discriminating the same kind of proteins based on their active function(s). It points out one possible future direction of [Publication **VI**], i.e., applying the joint clustering framework to detect noisy attractors of GRNs where the amount of proteins with low degrees of functionalities is not neglectable.

5.2 Conclusions

The topology of a GRN is studied at both the sequence and the gene levels, as represented by the studies in TFBS prediction and gene clustering in Chapter 3, respectively. Besides the proposed data fusion principle for TFBS prediction, which is characteristic of low false positive rate, the utilization of DNA duplex stabilities in facilitating TFBS prediction forms a valuable contribution to this field. In the area of gene clustering, despite the huge body of existing clustering methods, the proposed approach novels in its ability of jointly utilizing multiple information sources to solve the clustering problem. Thus, the objects can be grouped in a more reasonable way regarding the specific problem studied, and the results are more robust to the high level of genomic noise compared with using single data sources alone. For example, genes share similar expression profiles, being regulated by the same TFs and whose products interact with each other, are more likely to be involved in the same pathways than those that only share similar expression patterns due to, e.g., the coincidence of simultaneously expressed genes. Another example would be its application in noisy attractor detection as discussed in Chapter 4. With the proposed method, multiple aspects of the cellular system are simultaneously monitored to detect the noisy attractors of a GRN, via which the third noisy attractor is observed in TSs under certain noise level, toggling the long cherished belief that TSs have only two noisy attractors.

The purpose of the work presented in Chapter 4 is to study the dynamics of a GRN, with the approach of detecting its noisy attractors. The key paper is [Publication **VI**], which proposes a method to detect the noisy attractors of a GRN. The method itself is an application of the clustering method developed in Chapter 3. The results demonstrate the applicability

of the proposed method in studying GRNs' dynamics, and the biological finding that TSs have three noisy attractors other than two indicates that cells governed by real genetic TSs may differentiate into three distinct cell types under certain conditions. [Publication **V**] demonstrates that the differentiation pattern of cells governed by a genetic circuit can be viewed as a function of its noise level which can be controlled by its biological parameters. Based on [Publication **V**], it is possible to verify how well the noisy attractors are detected in [Publication **VI**], i.e., via observing cells' differentiation patterns. [Publication **VII**] further studies how changes in proteins' functionalities affect the dynamics of delayed stochastic GRNs, indicating an important regulatory role of proteins' functionalities in gene expression and pointing out a possible future direction for [Publication **VI**].

Through the current studies, it is found that fusing multiple information sources can facilitate or solve many problems regarding gene regulatory mechanisms. Given the complexity and stochasticity of a GRN, observing information from one source renders the results subject to its inherent noise (e.g., [Publication **I**] to [Publication **IV**]) and, in some applications, insufficient to reveal the underlying ground truth (e.g., [Publication **VI**]). Moreover, as explored in [Publication **I**] which may be extendable to other applications, the most efficient data fusion requires the data sources to be informative but not redundant.

5.3 Future directions

In the study of improving TFBS prediction via fusing genomic data sources ([Publication **I**]), two novel information sources, i.e., nucleosome positioning and DNA duplex stability, are investigated besides evolutionary conservation and regulatory potential. However, the two new data sources are obtained from computational predictions, limiting the prediction accuracy. Further study in this area may involve, on one hand, utilizing experimental data when available and, on the other hand, exploring other information sources, such as ChIP-chip data, to further improve TFBS prediction accuracy.

In gene clustering ([Publication **II**] to [Publication **IV**]), one basic assumption of the developed models is that the ground truth clusterings for data of different distributions are the same. The model sBGMM, which employs three data sources, is tested by the case where heterogeneous information sources do not accord in their underlying structures (see [Publication **III**]). The results show that the joint finite mixture modeling framework is tolerant to this inconsistency when priors are used to guide the clustering process. An alternative to solve this problem is to employ other modeling strategies, such as hierarchical Bayes modeling [225] which models the true clustering structure while allowing the existence of the individual structure for each data type.

As shown in [Publication **VII**], the degree of proteins' functionalities significantly affects the dynamics of delayed stochastic GRNs and their cell differentiation patterns at cell population level. Further, noisy attractors have so far been only detected in genetic switches, i.e., TS and an excitable genetic circuit in *Bacillus subtilis* ([Publication **VI**]), due to their straight connections to cell differentiation for the validation purpose. Thus, given its validated performance, it is feasible and promising to apply the noisy attractor detection method to GRNs where proteins with low degrees of functionalities are non-neglectable, and/or explore more complicated GRNs in the future.

Finally, data fusion is applicable to many biological problems other than those concerned here, which needs further investigation. It is also believed that with more data sources becoming experimentally and/or computationally available and more advanced analysis techniques being developed, people will get a more and more precise view of GRNs' regulatory mechanisms.

Bibliography

[1] H. Pearson. Genetics: what is a gene? *Nature*, 441(7092):3980–401, May 2006.

[2] S.A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.*, 22(3):437–467, Mar. 1969.

[3] B.H. Junker and F. Schreiber. *Analysis of biological networks*. John Wiley & Sons, Hoboken, New Jersey, 2008.

[4] A.S. Ribeiro, R. Zhu, and S.A. Kauffman. A general modeling strategy for gene regulatory networks with stochastic dynamics. *J. Comp. Biol.*, 13(9):1630–1639, Nov. 2006.

[5] A. Gimelbrant, J.N. Hutchinson, B.R. Thompson, and A. Chess. Widespread monoallelic expression on human autosomes. *Science*, 318(5853):1136–1140, Nov. 2007.

[6] C. Janeway. *Immunobiology (6th edition)*. Garland Science, New York, USA, 2005.

[7] A. Beyer, J. Hollunder, H.P. Nasheuer, and T. Wilhelm. Post-transcriptional expression regulation in the yeast *Saccharomyces cerevisiae* on a genomic scale. *Mol. Cell Proteomics.*, 3(11):1083–1092, Nov. 2004.

[8] S. Hautaniemi, S. Kharait, A. Iwabu, A. Wells, and D.A. Lauffenburger. Modeling of signal-response cascades using decision tree analysis. *Bioinformatics*, 21(9):2027–2035, May 2005.

[9] A. Wolf-Yadlin, S. Hautaniemi, D.A. Lauffenburger, and F.M. White. Multiple reaction monitoring for robust quantitative analysis of cellular signaling networks. *Proc. Natl. Acad. Sci. USA*, 104(14):5860–5865, Apr. 2007.

[10] K. Ovaska, M. Laakso, and S. Hautaniemi. Fast gene ontology based clustering for microarray experiments. *BioData Min.*, 1(1):11, Nov. 2008.

[11] B. Lewin. *Genes VIII*. Pearson Education (US), Upper Saddle River, 2004.

[12] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular biology of the cell (Fifth edition)*. Garland Science, New York, USA, 2008.

[13] B. Ren, F. Robert, J.J. Wyrick, O. Aparicio, E.G. Jennings, I. Simon, and et al. Genome-wide location and function of DNA binding proteins. *Science*, 290(5500):2306–2309, Dec. 2000.

[14] J.P. Peter. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, 10(10):669–680, Oct. 2009.

[15] M.F. Berger, A.A. Philippakis, A.M. Qureshi, F.S. He, P.W. Estep III, and M.L. Bulyk. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, 24(11):1429–1435, Nov. 2006.

[16] L. Zhang, S. Kasif, and C.R. Cantor. Quantifying DNA-protein binding specificities by using oligonucleotide mass tags and mass spectroscopy. *Proc. Natl. Acad. Sci. USA*, 104(9):3061–3066, Feb. 2007.

[17] R. Staden. Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.*, 12(1):505–519, Jan. 1984.

[18] W.W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, 5(4):276–287, Apr. 2004.

[19] H. Lähedsmäki, A.G. Rust, and I. Shmulevich. Probabilistic inference of transcription factor binding from multiple data sources. *PLoS One*, 3(3):e1820, Mar. 2008.

[20] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I.K. Moore, and et al. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–778, Aug. 2006.

[21] D.X. Jiang, C. Tang, and A.D. Zhang. Cluster analysis for gene expression data: a survey. *IEEE Trans. Knowl. Data Eng.*, 16(11):1370–1386, 2004.

[22] H. Shatkay, S. Edwards, W.J. Wilbur, and M. Boguski. *Genes, themes and microarrays: using information retrieval for large-scale gene analysis. In* Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB). AAAI, USA, Sep. 2000.

[23] H. Midelfart, A. Lægreid, and J. Komorowski. *Classification of Gene Expression Data in an Ontology. In* Medical data analysis, volume 2199. Springer Berlin /Heidelberg, Berlin, Jan. 2001.

[24] E. Segal, H. Wang, and D. Koller. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, 19(Suppl 1):i264–i272, 2003.

[25] K. Tu, H. Yu, and Y.X. Li. Combing gene expression profiles and protein-protein interaction data to infer gene functions. *J. Biotechnol.*, 124(3):475–485, Jul. 2006.

[26] T. Strachan and A.P. Read. *Human molecular genetics 3.* Garland Publishing, New York, USA, 2004.

[27] K.M. Vickaryous and B.K. Hall. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol. Rev.*, 81(3):425–455, Aug. 2006.

[28] L. Kadanoff, S. Coppersmith, and M. Aldana. Boolean dynamics with random couplings. In *Perspectives and problems in nonlinear science (springer applied mathematical sciences series).* Apr. 2003.

[29] A.S. Ribeiro and S.A. Kauffman. Noisy attractors and ergodic sets in models of genetic regulatory networks. *J. Theor. Biol.*, 247(4):743–755, Aug. 2007.

[30] J. MacQueen. *Some methods for classification and analysis of multivariate observations. In* Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, volume 1. University of California Press, California, USA, 1967.

[31] B. Zhang, Q. Wang, and X. Pan. MicroRNAs and their regulatory roles in animals and plants. *J. Cell. Physiol.*, 210(2):279–289, Feb. 2007.

[32] F.H.C. Crick. On protein synthesis. *Symp. Soc. Exp. Biol. XII*, 12:139–163, Oct. 1958.

[33] F. Crick. Central dogma of molecular biology. *Nature*, 227:561–563, Aug. 1970.

[34] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, and et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–2504, Nov. 2003.

[35] J. Tooze and R. Weiss. *RNA tumor viruses. In* Molecular biology of tumor viruses (2nd edition). Cold Spring Harbor Laboratory, New York, USA, 1985.

[36] P. Ahlquist. RNA-dependent RNA polymerases, viruses, and RNA silencing. *Science*, 296(5571):1270–1273, May 2002.

[37] B.J. McCarthy and J.J. Holland. Denatured DNA as a direct template for *in vitro* protein synthesis. *Proc. Natl. Acad. Sci. USA*, 54(3):880–886, Sep. 1965.

[38] T. Uzawa, A. Yamagishi, and T. Oshima. Polypeptide synthesis directed by DNA as a messenger in cell-free polypeptide synthesis by extreme thermophiles, *Thermus thermophilus* HB27 and *Sulfolobus tokodaii* strain 7. *J. Biochem.*, 131(6):849–853, Jun. 2002.

[39] N.J. Trun and J.E. Trempy. *Fundamental bacterial genetics*. Blackwell Science, UK, 2004.

[40] B.G. Faitsch and G.E. Moulton. *The complete idiot's guide to biology.* Penguin Group, New York, USA, 2004.

[41] V.L. Davidson and D.B. Sittman. *Biochemistry (4th edition).* Lippincott Williams & Wilkins, Maryland, USA, 1999.

[42] L. Ringrose. Polycomb comes of age: genome-wide profiling of target sites. *Curr. Opin. Cell Biol.*, 19(3):290–297, May 2007.

[43] M. Karin. Too many transcription factors: positive and negative interactions. *New Biol.*, 2(2):126–131, Feb. 1990.

[44] D.S. Latchman. Transcription factors: an overview. *Int. J. Biochem. Cell Biol.*, 29(12):1305–1312, Dec. 1997.

[45] P.J. Mitchell and R. Tjian. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science*, 245(4916):371–378, Jul. 1989.

[46] M. Ptashne and A. Gann. Transcriptional activation by recruitment. *Nature*, 386(6625):569–577, Apr. 1997.

[47] S.E. Halford and J.F. Marko. How do site-specific DNA-binding proteins find their targets? *Nucleic Acids Res.*, 32(10):3040–3052, Jun. 2004.

[48] R.G. Roeder. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.*, 21(9):327–335, Sep. 1996.

[49] D.B. Nikolov and S.K. Burley. RNA polymerase II transcription initiation: a structural view. *Proc. Natl. Acad. Sci. USA*, 94(1):15–22, Jan. 1997.

[50] T.I. Lee and R.A. Young. Transcription of eukaryotic protein-coding genes. *Annu. Rev. Genet.*, 34:77–137, Dec. 2000.

[51] R.F. Weaver. *Molecular biology.* McGraw Hill Higher Education, Boston, 2007.

[52] S. Baumberg. *Prokaryotic Gene Expression.* Oxford University Press, New York, USA, 2002.

[53] Q. Tian, S.B. Stepaniants, M. Mao, L. Weng, M.C. Feetham, M.J. Doyle, and et al. Integrated genomic and proteomic analyses of gene expression in mammalian cells. *Mol. Cell Prot.*, 3:960–969, Jul. 2004.

[54] L. Nie, G. Wu, and W. Zhang. Correlation between mRNA and protein abundance in *desulfovibrio vulgaris*: a multiple regression to identify of variations. *Biochem. Biophys. Res. Commun.*, 339(2):603–610, Jan. 2006.

[55] R. Brockmann, A. Beyer, J.J. Heinisch, and T. Wilhelm. Posttranscriptional expression regulation: what determines translation rates? *PLoS Comput. Biol.*, 3(3):e57, Mar. 2007.

[56] S. Hammond, E. Bernstein, D. Beach, and G. Hannon. An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. *Nature*, 404(6775):293–296, Mar. 2000.

[57] M. Kozak. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene*, 21(361):13–37, Nov. 2005.

[58] A.S. Spirin. *Ribosomes.* Plenum, New York, USA, 1999.

[59] A.A.M. Thomas, R. Rijnbrand, and H.O. Voorma. Recognition of the initiatio codon for protein synthesis in foot-and-mouth disease virus RNA. *J. Gen. Virol.*, 77(Pt 2):265–272, Feb. 1996.

[60] W.C. Merrick. Mechanism and regulation of eukaryotic protein synthesis. *Microbiol. Rev.*, 56(2):291–315, Jun. 1992.

[61] M. Knöfler, C. Waltner, E. Wintersberger, and E.W. Müllner. Translational repression of endogenous thymidine kinase mRNA in differentiating and arresting mouse cells. *J. Biol. Chem.*, 268(15):11409–11416, May 1993.

[62] M. Wickens. Messenger RNA. Springtime in the desert. *Nature*, 363(6427):305–306, May 1993.

[63] J.A. Garcia-Sanz, W. Mikulits, A. Livingstone, I. Lefkovits, and E.W. Müllner. Translational control: a general mechanism for gene regulation during T cell activation. *FASEB J.*, 12(3):299–306, Mar. 1998.

[64] V. Shastri. *Encyclopaedic dictionary of biotechnology.* Isha Books, Adarsh Nagar, Delhi, 2005.

[65] X.J. Yang and E. Seto. Lysine acetylation: codified crosstalk with other posttranslational modifications. *Mol. Cell*, 31(4):449–461, Aug. 2008.

[66] G. Leroy, J.T. Weston, B.M. Zee, N.L. Young, M.D. Plazas-Mayorca, and B.A. Garcia. Heterochromatin protein 1 is extensively decorated with histone code-like post-translational modifications. *Mol. Cell Proteomics*, Epub ahead of print, Jun. 2009.

[67] R. Mironova, T. Niwa, R. Dimitrova, M. Boyanova, and I. Ivanov. Glycation and post-translational processing of human interferon-γ expressed in *Escherichia coli. J. Biol. Chem.*, 278(51):51068–51074, Dec. 2003.

[68] M. Sato. Citrullination, a novel post-translational modification of histone. *Seikagaku.*, 79(7):686–690, Jul. 2007.

[69] C. Vanbelle, F. Halgand, T. Cedervall, E. Thulin, K.S. Åkerfeldt, O. Laprévote, and et al. Deamidation and disulfide bridge formation in human calbindin D_{28k} with effects on calcium binding. *Protein Sci.*, 14(4):968–979, Apr. 2005.

[70] T. Takata, L.G. Woodbury, and K.J. Lampi. Deamidation alters interactions of β-crystallins in hetero-oligomers. *Mol. Vis.*, 15:241–249, Jan. 2009.

[71] D.F. Brennan and D. Barford. Eliminylation: a post-translational modification catalyzed by phosphothreonine lyases. *Trends Biochem Sci.*, 34(3):108–114, Feb. 2009.

[72] J.A. Gatehouse, G.W. Lycett, A.J. Delauney, R.R. Croy, and D. Boulter. Sequence specificity of the post-translational proteolytic cleavage of vicilin, a seed storage protein of pea (*Pisum sativum L.*). *Biochem J.*, 212(2):427–432, May 1983.

[73] U.S. Department of Energy Office of Science. *Genomes to Life Program Roadmap.* http://genomicsgtl.energy.gov/, Apr. 2001.

[74] M.A. Westenberg, S.A.F.T. van Hijum, A.T. Lulko, O.P. Kuipers, and J.B.T.M. Roerdink. *Interactive visualization of gene regulatory networks with associated gene expression time series data. In* Visualization in medicine and life sciences. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[75] S.A. Kauffman. *The origins of order*. Oxford University Press, New York, USA, 1993.

[76] G.M. Süel, J. Garcia-Ojalvo, L.M. Liberman, and M.B. Elowitz. An excitable gene regulatory circuit induces transient cellular differentiation. *Nature*, 440(23):545–550, Mar. 1984.

[77] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, 34(2):166–176, Jun. 2003.

[78] D.T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. phys.*, 22(4):403–434, Dec. 1976.

[79] K.D. MacIsaac and E. Fraenkel. Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Comput. Biol.*, 2(4):e36, Apr. 2006.

[80] A. Siepel, G. Bejerano, J.S. Pedersen, A.S. Hinrichs, M. Hou, K. Rosenbloom, and et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, 15(8):1034–1050, Aug. 2005.

[81] A. Woolfe, M. Goodson, D. Goode, P. Snell, G. McEwen, T. Vavouri, and et al. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.*, 3(1):e7, Jan. 2005.

[82] K.A. Frazer, H. Tao, K. Osoegawa, P.J. de Jong, X. Chen, M.F. Doherty, and et al. Noncoding sequences conserved in a limited number of mammals in the SIM2 interval are frequently functional. *Genome Res.*, 14(3):367–372, Mar. 2004.

[83] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E.S. Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–254, May 2003.

[84] R. Siddharthan, E.D. Siggia, and E. van Nimwegen. PhyloGibbs: A Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol.*, 1(7):e67, Dec. 2005.

[85] A.C. Wilson, S.S. Carlson, and T.J. White. Biochemical evolution. *Annu. Rev. Biochem.*, 46:573–639, 1977.

[86] M. Brudno, C. Do, G. Cooper, M.F. Kim, E. Davydov, E.D. Green, and et al. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.*, 13(4):721–731, Apr. 2003.

[87] M. Blanchette, W.J. Kent, C. Riemer, L. Elnitski, A.F.A. Smit, K.M. Roskin, and et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, 14(4):708–715, Apr. 2004.

[88] N. Bray and L. Pachter. MAVID: constrained ancestral alignment of multiple sequences. *Genome Res.*, 14(4):693–699, Apr. 2004.

[89] R.C. Hardison, J. Oeltjen, and W. Miller. Long human-mouse sequence alignments reveal novel regulatory elements: reasons to sequence the mouse genome. *Genome Res.*, 7(10):959–966, Oct. 1997.

[90] M.A. Nobrega, I. Ovcharenko, V. Afzal, and E.M. Rubin. Scanning human gene deserts for long-range enhancers. *Science*, 302(5644):413, Oct. 2003.

[91] E.T. Dermitzakia and A.G. Clark. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.*, 19(7):1114–1121, Jul. 2002.

[92] G.A. Wray, M.W. Hahn, E. Abouheif, J.P. Balhoff, M. Pizer, M.V. Rockman, and et al. The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.*, 20(9):1377–1419, Sep. 2003.

[93] S. Schwarz, Z. Zhang, K. Frazer, A. Smit, C. Riemer, J. Bouck, and et al. PipMaker: a web server for aligning two genomic DNA sequences. *Genome Res.*, 10(4):577–586, Apr. 2000.

[94] E.H. Margulies, M. Blanchette, NISC Comparative Sequencing Program, D. Haussler, and E.D. Green. Identification and characterization of multi-species conserved sequences. *Genome Res.*, 13(12):2507–2518, Dec. 2003.

[95] N. Stojanovic, L. Florea, C. Riemer, D. Gumucio, J. Slightom, M. Goodman, and et al. Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. *Nucleic Acids Res.*, 27(19):3899–3910, Oct. 1999.

[96] M.A. Chapman, I.J. Donaldson, J. Gilbert, D. Grafham, J. Rogers, A.R. Green, and et al. Analysis of multiple genomic sequence alignments: a web resource, online tools, and lessons learned from analysis of mammalian SCL loci. *Genome Res.*, 14(2):313–318, Jan. 2004.

[97] I. Ovcharenko, G.G. Loots, B.M. Giardine, M. Hou, J. Ma, R.C. Hardison, and et al. Mulan: multiple-sequence local alignment and visualization for studying function and evolution. *Genome Res.*, 15(1):184–194, Jan. 2005.

[98] G. Bejerano, D. Haussler, and M. Blanchette. Into the heart of darkness: Large-scale clustering of human non-coding DNA. *Bioinformatics*, 20(Suppl. 1):i40–i48, Aug. 2004.

[99] J. Taylor, S. Tyekucheva, D.C. King, R.C. Hardison, W. Miller, and F. Chiaromonte. ESPERR: learning strong and weak signals in genomic sequence alignments to identify functional elements. *Genome Res.*, 16(12):1596–1604, Dec. 2006.

[100] K.E. van Holde. *Chromatin*. Springer, New York, USA, 1989.

[101] J. Widom. Role of DNA sequence in nucleosome stability and dynamics. *Rev. Biophys.*, 34(3):269–324, Aug. 2001.

[102] B.E. Bernstein, C.L. Liu, E.L. Humphrey, E.O. Perlstein, and S.L. Schreiber. Global nucleosome occupancy in yeast. *Genome Biol.*, 5(9):R62, 2004.

[103] G.C. Yuan, Y.J. Liu, M.F. Dion, M.D. Slack, L.F. Wu, S.J. Altschuler, and et al. Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science*, 309(5734):626–630, Jul. 2005.

[104] I. Albert, T.N. Mavrich, L.P. Tomsho, J. Qi, S.J. Zanton, S.C. Schuster, and et al. Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature*, 446(7135):572–576, Mar. 2007.

[105] S.M. Johnson, F.J. Tan, H.L. McCullough, D.P. Riordan, and A.Z. Fire. Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin. *Genome Res.*, 16(12):1505–1516, Dec. 2006.

[106] Y. Mito, J.G. Henikoff, and S. Henikoff. Genome-scale profiling of histone H3.3 replacement patterns. *Nat. Genet.*, 37(10):1090–1097, Oct. 2005.

[107] N.D. Heintzman, R.K. Stuart, G. Hon, Y. Fu, C.W. Ching, R.D. Hawkins, and et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, 39(3):311–318, Mar. 2007.

[108] F. Ozsolak, J.S. Song, X.S. Liu, and D.E. Fisher. High-throughput mapping of the chromatin structure of human promoters. *Nat. Biotechnol.*, 25(2):244–248, Feb. 2007.

[109] G.C. Yuan and J.S. Liu. Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput. Biol.*, 4(1):e13, Jan. 2008.

[110] I.P. Ioshikhes, I. Albert, S.J. Zanton, and B.F. Pugh. Nucleosome positions predicted through comparative genomics. *Nat. Genet.*, 38(10):1210–1215, Oct. 2006.

[111] H.E. Peckham, R.E. Thurman, Y. Fu, J.A. Stamatoyannopoulos, W.S. Noble, K. Struhl, and et al. Nucleosome positioning signals in genomic DNA. *Genome Res.*, 17(8):1170–1177, Aug. 2007.

[112] H. Lodish, A. Berk, and P. Matsudaira. *Molecular cell biology (5th edition)*. W.H. Freeman & Company, New York, USA, 2003.

[113] C.P. Bi and C.J. Benham. WebSIDD: server for prediction of the stress-induced duplex destabilized sites in superhelical DNA. *Bioinformatics*, 20(9):1477–1479, Jun. 2004.

[114] C.J. Benham. Sites of predicted stress-induced DNA duplex destabilization occur preferentially at regulatory loci. *Proc. Natl. Acad. Sci. USA*, 90(7):2999–3003, Apr. 1993.

[115] D. Kowalski, D. Natale, and M. Eddy. Stable DNA unwinding, not "breathing", accounts for single-stranded specific nuclease hypersensitivity of specific A+T-rich regions. *Proc. Natl. Acad. Sci. USA*, 85(24):9464–9468, Dec. 1988.

[116] S.D. Sheridan, C.J. Benham, and G.W. Hatfield. Activation of gene expression by a novel DNA structural transmission mechanism that requires supercoiling-induced DNA duplex destabilization in an upstream activating sequence. *J. Biol. Chem.*, 273(33):21298–21308, Aug. 1998.

[117] L.B. Rothman-Denes, X. Dai, E. Davydova, R. Carter, and K. Kazmierczak. Transcriptional regulation by DNA structural transitions and single-stranded DNA-binding proteins. *Cold Spring Harbor Symp. Quant. Biol.*, 63:63–73, 1998.

[118] C.J. Benham, T. Kohwi-Shigematsu, and J. Bode. Stress-induced duplex destabilization in chromosomal scaffold/matrix attachment regions. *J. Mol. Biol.*, 274:181–196, 1997.

[119] G. Thijs, M. Lescot, K. Marchal, S. Rombauts, B.D. Moor, P. Pouzé, and et al. A higher order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, 17(12):1113–1122, Nov. 2001.

[120] J.M. Claverie and S. Audic. The statistical significance of nucleotide position-weight matrix matches. *Bioinformatics*, 12(5):431–439, Oct. 1996.

[121] X. Xiao, E.R. Dow, R. Eberhart, Z.B. Miled, and R.J. Oppelt. *Gene clustering using self-organizing maps and particle swarm optimization. In* Proceedings of the 17th International Parallel and Distributed Processing Symposium (IPDPS'03). IEEE Computer Society, Apr. 2003.

[122] H.C. Causton, J. Quackenbush, and A. Brazma. *Microarray/gene expressions data analysis: a beginner's guide.* Blackwell Science, UK, 2003.

[123] R. Autio. *Computational methods for high-throughput data analysis in cancer research.* Tampere University of Technology, Doctoral Thesis, Sep. 2008.

[124] H.C. Causton, J. Quackenbush, and A. Brazma. *Analysis of microarray gene expression data. In* Handbook of statistical genetics (3rd edition). John Wiley & Sons, England, UK, 2007.

[125] J. Tuimala and M.M. Laine. *DNA microarray data analysis.* CSC, the Finnish IT center for Science, Helsinki, Finland, Dec. 2005.

[126] S. Hautaniemi, H. Edgren, P. Vesanen, M. Wolf, A.K. Järvinen, O. Yli-Harja, and et al. A novel strategy for microarray quality control using Bayesian networks. *Bioinformatics*, 19(16):2031–2038, Nov. 2003.

[127] G.K. Smyth, Y.H. Yang, and T.Speed. *Statistical issues in cDNA microarray data analysis. In* Functional genomics: methods and protocols. Humana Press, Totowa, NJ, Jun. 2002.

[128] T.J. Hestilow and Y. Huang. Clustering of gene expression data based on shape similarity. *EURASIP J. Bioinform. Syst. Biol.*, 2009:195712, Apr. 2009.

[129] J. Monod and F. Jacob. Clustering of unevenly sampled gene expression time-series data. *Fuzzy Sets Sys.*, 152(1):49–66, May 2005.

[130] P. Ma, C.I. Castillo-Davis, W. Zhong, and J.S. Liu. A data-driven clustering method for time course gene expression data. *Nucleic Acids Res.*, 34(4):1261–1269, Mar. 2006.

[131] W. Pan. Incorporating gene functions as priors in model-based clustering of microarray gene expression data. *Bioinformatics*, 22(7):795–801, Jan. 2006.

[132] Y. Yuan and C-T. Li. *Unsupervised clustering of gene expression time series with conditional random fields. In* Proceedings of the Inaugural IEEE International Conference on Digital EcoSystems and Technologies (DEST'07). IEEE Computer Society, Feb. 2007.

[133] L. Rueda, A. Bari, and A. Ngom. *Clustering time-series gene expression data with unequal time intervals. In* Transactions on computational system biology X. Springer-Verlag, Berlin, Germany, 2008.

[134] S.C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, Sep. 1967.

[135] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction.* Springer Science + Business Media, New York, USA, 2009.

[136] G.J. McLachlan and D. Peel. *Finite mixture model.* John Wiley & Sons, New York, USA, 2000.

[137] N.C. Seeman, J.M. Rosenberg, and A. Rich. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. USA*, 73(3):804–808, Mar. 1976.

[138] N.M. Luscombe, R.A. Laskowski, and J.M. Thornton. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res.*, 29(13):2860–2874, Jul. 2001.

[139] R.G. Endres, T.C. Schulthess, and N.S. Wingreen. Toward an atomistic model for predicting transcription factor binding sites. *Proteins*, 57(2):262–268, Nov. 2004.

[140] A. Beyer, C. Workman, J. Hollunder, D. Radke, U. Möller, T. Wilhelm, and et al. Integrated assessment and prediction of transcription factor binding. *PLoS Comput. Biol.*, 2(6):e70, Jun. 2006.

[141] O. Aparicio, J.V. Geisberg, and S. Kevin. *Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. In* Current protocols in cell biology. California, USA, Jun. 2004.

[142] S. Fields and O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246, Jul. 1989.

[143] A. Panchenko and T. Przytycka. *Protein-protein interactions and networks: identification, computer, analysis, and prediction.* Springer-Verlag, London, UK.

[144] A. Kumar and M. Snyder. Protein complexes take the bait. *Nature*, 415(6868):123–124, Jan 2002.

[145] P. Uetz, L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, J.R. Knight, and et al. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403(6770):623–627, Feb. 2000.

[146] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA*, 98(8):4569–4574, Apr. 2001.

[147] P. Uetz. Two-hybrid arrays. *Curr. Opin. Chem. Biol.*, 6(1):57–62, Feb. 2002.

[148] Y. Ho, A. Gruhler, A. Heilbut, G.D. Bader, L. Moore, S.L. Adams, and et al. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415(6868):123–124, Jan. 2002.

[149] N.J. Krogan, W.T. Peng, G. Cagney, M.D. Robinson, R. Haw, G. Zhong, and et al. High-definition macromolecular composition of yeast RNA-processing complexes. *Mol. Cell*, 13(2):225–239, Jan. 2004.

[150] A.C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, and et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):637–643, Mar. 2006.

[151] A. Chatraryamontri, A. Ceol, L.M. Palazzi, G. Nardelli, M.V. Schneider, L.Castagnoli, and et al. MINT: the Molecular INTeraction database. *Nucleic Acids Res.*, 35(Database issue):D572–D574, Jan. 2006.

[152] S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, and et al. IntAct-open source resource for molecular interaction data. *Nuleic Acids Res.*, 35(Database issue):D561–D565, Feb. 2006.

[153] I. Xenarios, D.W. Rice, L. Salwinski, M.K. Baron, E.M. Marcotte, and D. Eisenberg. DIP: the database of interacting proteins. *Nucleic Acids Res.*, 28(1):289–291, Jan. 2000.

[154] L. Salwinski, C.S. Miller, A.J. Smith, F.K. Pettit, J.U. Bowie, and D. Eisenberg. The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, 32(Database issue), Jan. 2004.

[155] D. Fiedler, H. Braberg, M. Mehta, G. Chechik, G. Cagney, P. Mukherjee, and et al. Functional organization of the *S. cerevisiae* phosphorylation network. *Cell*, 136(5):952–963, Mar. 2009.

[156] T.S.K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, and et al. Human protein reference database-2009 update. *Nuleic Acids Res.*, 37(Database issue):D767–D772, Jan. 2009.

[157] U. Güldener, M. Münsterkötter, M. Oesterheld, P. Pagel, A. Ruepp, H.W. Mewes, and et al. MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res.*, 34(Database issue):D436–D441, Jan. 2006.

[158] J. Wu, T. Vallenius, K. Ovaska, J. Westermarck, T.P. Makela, and S. Hautaniemi. Integrated network analysis platform for protein-protein interactions. *Nat. Methods*, 6(1):75–77, Jan. 2009.

[159] C. Fraley and A.E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.*, 97(458):611–631, Jun. 2002.

[160] D.W. Mount. *Bioinformatics: sequence and genome analysis (2nd edition).* John Inglis, New York, USA, 2004.

[161] K. McGarigal, S. Cushman, and S.G. Stafford. *Multivariate statistics for wildlife and ecology research.* Springer-Verlag, New York, USA, 2000.

[162] D. Russell. *The principles of computer networking.* Cambridge University Press, Cambridge, UK, 1989.

[163] M. Steinbach, G. Karypis, and V. Kumar. *A comparison of document clustering techniques. In* KDD Workshop on Text Mining. Boston MA, Aug. 2000.

[164] S. Zhong and J. Ghosh. A unified framework for model-based clustering. *J. Mach. Learn. Res.*, 4:1001–1037, Nov. 2003.

[165] P. Melin and O. Castillo. *Hybrid intelligent systems for pattern recognition using soft computing: an evolutionary approach for neural networks and fuzzy systems.* Springer-Verlag, Berlin, Germany, 2005.

[166] L. Kaufman and P.J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis.* Wiley, New York, USA, 1990.

[167] M. Medvedovic and S. Sivaganesan. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18(9):1194–1206, Sep. 2002.

[168] S. Vaithyanathan and B. Dom. *Model-based hierarchical clustering. In* Proceeding of the 16th Conference on Uncertainty in Artificial Intelligence (IPDPS'03). Morgan Kaufmann Publishers, San Francisco, CA, USA, Jul. 2003.

[169] T. Kohonen. *Self-organizing map.* Springer-Verlag, New York, USA, 1997.

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

[171] G.J. McLachlan. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Appl. Statist.*, 36(3):318–324, 1987.

[172] G.J. McLachlan and D. Peel. *On a resampling approach to choosing the number of components in normal mixture models. In* Computing science and statistics, volume 28. Interface Foundation of North America, Fairfax Station, Virginia, 1997.

[173] P. Smyth. Model selection for probabilistic clustering using cross-validated likelihood. *Stat. Comp.*, 10(1):63–72, Jan. 2000.

[174] Y. Ji, C. Wu, P. Liu, J. Wang, and R.K. Coombes. Applications of beta-mixture models in bioinformatics. *Bioinformatics*, 21(9):2118–2122, May 2005.

[175] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, Mar. 1978.

[176] C. Biernacki and G. Govaert. Choosing models in model-based clustering and discriminant analysis. *J. Statis. Comput. Simul.*, 64(1):49–71, 1999.

[177] T.R. Kiehl, R.M. Mattheyses, and M.K. Simmons. A Bayesian method for classification and discrimination. *Can. J. Statist.*, 20(4):451–461, Dec. 1992.

[178] J. Diebolt and C.P. Robert. Bayesian estimation of finite mixture distributions through Bayesian sampling. *J. R. Statist. Soc.*, 56(2):363–375, 1994.

[179] D.M. Chickering and D. Heckerman. Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. *Mach. Learn.*, 29(2/3):181–244, Mar. 1997.

[180] H. Akaike. A new look at the statistical identification model. *IEEE Trans. Automat. Contr.*, 19(6):716–723, 1974.

[181] H. Bozdogan. Model selection and akaike information criterion (AIC): the general theory and its analytic extensions. *Psychometrika*, 52(3):345–370, Sep. 1987.

[182] S.C. Rastogi. *Cell and molecular biology.* New Age International, New Delhi, 2003.

[183] A.S. Ribeiro and J. Lloyd-Price. SGN Sim, a stochastic genetic networks simulator. *Bioinformatics*, 23(6):777–779, Mar. 2007.

[184] M. Roussel and R. Zhu. Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression. *Phys. Biol.*, 3(4):274–284, Dec. 2006.

[185] A.A. Margolin and A. Califano. Theory and limitations of genetic network inference from microarray data. *Ann. N.Y. Acad. Sci.*, 1115:51–72, Dec. 2007.

[186] M. Acar, J. Mettetal, and A. van Oudenaarden. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected *Escherichia coli* cells. *Genetics*, 149(4):1633–1648, Aug. 1998.

[187] W.J. Mo, X.P. Fu, X.T. Han, G.Y. Yang, J.G. Zhang, F.H. Guo, and et al. A stochastic model for identifying differential gene pair co-expression patterns in prostate cancer progression. *BMC Genomics*, 10(340):doi:10.1186/1471–2164–10–340, Jul. 2009.

[188] S. Liang, S. Fuhrman, and R. Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.*, 3:18–29, 1998.

[189] T. Akutsu, S. Miyano, and S. Kuhara. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pac. Symp. Biocomput.*, 4:17–28, 1999.

[190] M.J. Herrgard, M.W. Covert, and B.O. Palsson. Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Res.*, 13(11):2423–2434, Oct. 2003.

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

[192] R. Zhu, A.S. Ribeiro, D. Salahub, and S.A. Kauffman. Studying genetic regulatory networks at the molecular level: delayed reaction stochatic models. *J. Theor. Biol.*, 246(4):725–745, Jun. 2007.

[193] D.A. Hall, J. Ptacek, and M. Snyder. Protein microarray technology. *Mech. Ageing Dev.*, 128(1):161–167, Jan. 2007.

[194] S.C. Lovell. Are non-functional, unfolded proteins ('junk proteins') common in the genome? *FEBS Letters*, 554(3):237–239, Nov. 2003.

[195] D.M. Thompson, K.R. King, K.J. Wieder, W. Toner, M.L. Yarmush, and A. Jayaraman. Dynamic gene expression profiling using a microfabricated living cell array. *Anal. Chem.*, 76(14):4098–4103, Jun. 2004.

[196] K.R. King, S. Wang, D. Irimia, A. Jayaraman, M. Toner, and M.L. Yarmush. A high-throughput microfluidic real-time gene expression living cell array. *Lab Chip*, 7(1):77–85, Jan. 2007.

[197] D.T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81(25):2340–2361, May 1977.

[198] D.T. Gillespie. *Stochastic chemical kinetics. In* Handbook of Materials Modeling*, (Yip, S., ed.).* Springer, Dordrecht, 1998.

[199] M. Tiina. *Stochastic methods for modeling intracellular signaling.* Tampere University of Technology, Doctoral Thesis, Dec. 2007.

[200] M.A. Gibson and J. Bruck. Effieicent exact stochastic simulation of chemical systems with many species and many channels. *J. Phys. Chem.*, 104(9):1876–1889, Mar. 2000.

[201] Y. Cao, H. Li, and L. Petzold. Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. *J. Chem. Phys.*, 121(9):4059–4067, Sep. 2004.

[202] J.M. McCollum, G.D. Peterson, C.D. Cox, M.L. Simpson, and N.F. Samatova. The sorting direct method for stochastic simulation of biochemical systems with varying reaction execution behavior. *Comput. Biol. Chem.*, 30(1):39–49, Feb. 2006.

[203] D.T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *J. Chem. Phys.*, 115(4):1716–1733, Jul. 2001.

[204] T. Tian and K. Burrage. Binomial leap methods for simulating stochastic chemical kinetics. *J. Chem. Phys.*, 121(21):10356–10364, Dec. 2004.

[205] J. Yu, J. Xiao, X. Ren, K. Lao, and X.S. Xie. Probing gene expression in live cells, one protein molecule at a time. *Science*, 311(5767):1600–1603, Mar. 2006.

[206] A. Loinger, A. Lipshtat, N.Q. Balaban, and O. Biham. Stochastic simulations of genetic switch systems. *Phys. Rev. E*, 75(2):021904, Feb. 2007.

[207] H. Bremer, P. Dennis, and M. Ehrenberg. Free RNA polymerase and modeling global transcription in *Escherichia coli*. *Biochimie*, 85(6):597–609, Jun. 2003.

[208] W.R. McClure. Rate-limiting steps in RNA chain initiation. *Proc. Natl. Acad. Sci. USA*, 77(10):5634–5638, Oct. 1980.

[209] D.E. Draper. *Translation initiation. In Escherichia coli* and *Salmonella*. ASM Press, Washington D.C., 1996.

[210] J. Monod and F. Jacob. Teleonomic mechanisms in cellular metabolism, growth, and differentiation. *Cold Spring Harb. Symp. Quant. Biol.*, 26:389–401, 1961.

[211] Z. Neubauerz and E. Calef. Immunity phase shift in defective lysogens: nonmutational hereditary change of early regulation of lambda prophage. *J. Mol. Biol.*, 51(1):1–13, Jul. 1970.

[212] S. Huang, G. Eichler, Y. Bar-Yam, and D. Ingber. Cell fates as a high-dimensional attractor states of a complex gene regulatory network. *Phys. Rev. Lett.*, 94(12):128701, Apr. 2005.

[213] D.A. Lim, Y.-C. Huang, T. Swigut, A.L. Mirick, J.M. Garcia-Verdugo, J. Wysocka, and et al. Chromatin remodelling factor Mll1 is essential for neurogenesis from postnatal neural stem cells. *Nature*, 458(7237):529–533, Mar. 2009.

[214] Y. Hirabayashi, N. Suzki, M. Tsuboi, T.A. Endo, T. Toyoda, J. Shinga, and et al. Polycomb limits the neurogenic competence of neural precursor cells to promote astrogenic fate transition. *Cell*, 63(5):600–613, Sep. 2009.

[215] J. Hahn, L. Kong, and D. Dubnau. The regulation of competence transcription factor synthesis constitutes a critical control point in the regulation of competence in *Bacillus subtilis*. *J. Bacteriol.*, 176(18):5753–5761, Sep. 1994.

[216] D. van Sinderen, A. Luttinger, L. Kong, D. Dubnau, G. Venema, and L. Hamoen. *comK* encodes the competence transcription factor, the key regulatory protein for competence in *Bacillus subtilis*. *J. Bacteriol.*, 15(3):455–462, Feb. 1995.

[217] D. Dubnau. DNA uptake in bacteria. *Annu. Rev. Microbiol.*, 53:217–244, 1999.

[218] M. Albano, R. Breitling, and D. Dubnau. Nucleotide sequence and genetic organization of the *Bacillus subtilis comG* operon. *J. Bacteriol.*, 171(10):5386–5404, Oct. 1989.

[219] K. Turgay, J. Hahn, J. Burghoorn, and D. Dubnau. Competence in *Bacillus subtilis* is controlled by regulated proteolysis of a transcription factor. *EMBO J.*, 17(22):6730–6738, Nov. 1998.

[220] H. Maamar and D. Dubnau. Bistability in the *Bacillus subtilis* K-state (competence) system requires a positive feedback loop. *Mol. Microbiol.*, 56(3):615–624, May 2005.

[221] W.K. Smits, C.C. Eschevins, K.A. Susanna, S. Bron, O.P. Kuipers, and L.W. Hamoen. Stripping *Bacillus*: ComK auto-stimulation is responsible for the bistable response in competence development. *Mol. Microbiol.*, 56(3):604–614, May 2005.

[222] M. Ogura, L. Liu, M. Lacelle, M.M. Nakano, and P. Zuber. Mutational analysis of ComS: evidence for the interaction of ComS and MecA in the regulation of competence development in *Bacillus subtilis*. *Mol. Microbiol.*, 32(4):799–812, May 1999.

[223] F. Zhao, Z. Xuan, L. Liu, and M.Q. Zhang. TRED: a transcriptional regulatory element database and a platform *in silico* gene regulation studies. *Nucleic Acids Res.*, 33(Database issue):D103–D107, Jan. 2005.

[224] G.Jr. Dennis, B.T. Sherman, D.A. Hosack, J. Yang, W. Gao, H.C. Lane, and et al. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.*, 4(5):P3, Apr. 2003.

[225] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian data analysis (2nd edition)*. Chapman & Hall/CRC, Boca Raton, F.L., USA, 2004.