STATISTICAL MECHANICAL MODELING OF EUKARYOTIC GENE REGULATION

BY

CHIEH-CHUN CHEN

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Bioengineering in the Graduate College of the University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Doctoral Committee:
Associate Professor Sheng Zhong, Chair
Professor Eric Jakobsson
Professor Huimin Zhao
Assistant Professor Jian Ma
Gene expression patterns are regulated by gene regulatory networks. Central to transcriptional regulation of gene expression is the regulation of the quantities of transcription factors (TFs) bound to genomic regulatory sequences. This thesis work is built on statistical mechanics to study the stochastic interactions of TFs and regulatory sequences. We present a predictive model to learn how TFs interact with cis-regulatory sequences and with each other. By analyzing large scale TF-DNA binding data, the model can discover cooperative interactions among TFs and predict the strength of TF-DNA binding.

Less clear is how the genome and the epigenome jointly instruct TFs binding. We present an epigenome-sensitive model to systematically analyze the epigenomic functions in modulating transcription factor-DNA binding. We discovered preferences of TFs for specific combinations of epigenomic modifications, termed as epigenomic motifs. Epigenomic motifs explain why some TFs appear to have different DNA binding motifs derived from in vivo and in vitro experiments. The data suggest that the epigenome can modulate transcriptional noise and boost the cooperativity of weak TF binding sites. We also show that the epigenome might suppress the TF binding differences on SNP-containing binding sites in two people, in theory and in real data.

To identify regulatory relationships between TFs and target genes is another major topic in gene regulation. We developed an analytical method to identify a statistical thermodynamic model that best describes the form of TF-TF interaction among a set of TFs for every target gene. Based on this method, we developed a computational framework to infer regulatory relationships from multiple time course gene expression datasets. RNA interference data and large scale TF-DNA binding data independently validated a statistically significant fraction of these regulatory relationships. Moreover, this framework has the flexibility to incorporate other independent datasets to increase prediction accuracy.
To my family, for their love and support.
I would like to express my sincere gratitude to people for helping and supporting me during my PhD study.

I owe my deepest gratitude to my advisor, Sheng Zhong, for his constant support and invaluable guidance. I have always been amazed by his broad knowledge and his passions about science, which inspired me to do my research. I am very grateful for the opportunity to closely work with him and I have learned a lot from him not only about research but also about his positive attitude and determination.

I would also like to thank the members of my thesis defense committee: Eric Jakobsson, Jian Ma and Huimin Zhao. Eric provided me with the great opportunity to work on Molecular Science Student Workbench during my first two years of study, which was a wonderful experience to cooperate with fellows from different areas and he always offered me generous feedback and different perspectives on my work. Jian has always supported me with his kind advice on both research and career. I thank Dr. Zhao for his valuable comments and insightful discussions about my research.

I want to give special thanks to my research colleagues, in particular Dan Xie and Xiaoyi Cao. We have had wonderful experiences working together on several research projects and I will always miss the days that we spent together in our office. I am also grateful to Xin He, Feng Hong, Shu Xiao and Pengfei Yu for the pleasant collaborations we had.

Finally, I would like to thank my family. My parents, Shun-Fa Chen and Fu-Ying Chiu, and my sister, Yen-Jen Chen, have always unconditionally supported me to do whatever I love. My fiance, Yen-Wei Huang, came to University of Illinois with me to pursue his PhD. We have had each other’s company in the journey of our PhDs, with all those joy, excitement and frustrations. This dissertation could not have been possible without their support.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>viii</td>
</tr>
<tr>
<td>CHAPTER 1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>CHAPTER 2 BACKGROUND</td>
<td>6</td>
</tr>
<tr>
<td>2.1 Statistical Thermodynamic Modeling on Transcription Factor Binding</td>
<td>6</td>
</tr>
<tr>
<td>2.2 Embryonic Stem Cells</td>
<td>9</td>
</tr>
<tr>
<td>CHAPTER 3 ANALYZING LARGE-SCALE TRANSCRIPTION FACTOR-DNA BINDING DATA THROUGH STATISTICAL MECHANICAL MODELING</td>
<td>11</td>
</tr>
<tr>
<td>3.1 Related Works</td>
<td>12</td>
</tr>
<tr>
<td>3.2 A Statistical Mechanical Model of Transcription Factor-DNA Interaction</td>
<td>12</td>
</tr>
<tr>
<td>3.3 Cooperative Interactions Among Transcription Factors for DNA Binding in Mouse Embryonic Stem Cells</td>
<td>16</td>
</tr>
<tr>
<td>3.4 Discussions</td>
<td>23</td>
</tr>
<tr>
<td>CHAPTER 4 ON THREE-WAY INTERACTIONS OF TRANSCRIPTION FACTOR, DNA AND EPIGENOME</td>
<td>25</td>
</tr>
<tr>
<td>4.1 Related Works</td>
<td>26</td>
</tr>
<tr>
<td>4.2 An Epigenome-Sensitive Transcription Factor-DNA Model</td>
<td>28</td>
</tr>
<tr>
<td>4.3 Simulation Studies Reflect Epigenomic Regulation on Transcriptional Noise and Epigenomic Boost on Weak Transcription Factor Binding Sites</td>
<td>33</td>
</tr>
<tr>
<td>4.4 Transcription Factor-Specific Epigenomic Motifs in Mouse Embryonic Stem Cells</td>
<td>39</td>
</tr>
<tr>
<td>4.5 Epigenome Might Attenuate Personal Variations on Transcription Factor Binding</td>
<td>50</td>
</tr>
<tr>
<td>4.6 Discussions</td>
<td>57</td>
</tr>
<tr>
<td>CHAPTER 5</td>
<td>IDENTIFY REGULATORY RELATIONSHIPS THROUGH STATISTICAL THERMODYNAMIC MODELING</td>
</tr>
<tr>
<td>-----------</td>
<td>--------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>5.1</td>
<td>Related Works</td>
</tr>
<tr>
<td>5.2</td>
<td>Selection of Statistical Thermodynamic Models for Combinatorial Control of Multiple Transcription Factors in Early Differentiation of Embryonic Stem Cells</td>
</tr>
<tr>
<td>5.3</td>
<td>Inferring Gene Regulatory Networks by Statistical Thermodynamic Modeling</td>
</tr>
<tr>
<td>5.4</td>
<td>Discussions</td>
</tr>
</tbody>
</table>

| CHAPTER 6 | CONCLUSIONS | 84 |

| REFERENCES | 87 |
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Importance of cooperative interactions</td>
<td>20</td>
</tr>
<tr>
<td>4.1</td>
<td>Statistical thermodynamic models of TF-DNA binding</td>
<td>27</td>
</tr>
<tr>
<td>4.2</td>
<td>Lack of association between weak TFBSs and promoters.</td>
<td>39</td>
</tr>
<tr>
<td>4.3</td>
<td>Comparison of model performances with and without epigenomic data.</td>
<td>42</td>
</tr>
<tr>
<td>4.4</td>
<td>Distribution of SNP-containing NFκB binding sites.</td>
<td>54</td>
</tr>
<tr>
<td>5.1</td>
<td>The Boltzmann distribution for the two states of a TFBS</td>
<td>62</td>
</tr>
<tr>
<td>5.2</td>
<td>The Boltzmann distribution of a promoter with one TF and one RNAP</td>
<td>64</td>
</tr>
<tr>
<td>5.3</td>
<td>The Boltzmann distribution of a promoter with its RNAP and two TFs</td>
<td>64</td>
</tr>
<tr>
<td>5.4</td>
<td>Sensitivity test for $K_{TF}$, $K_g$, $q_p$, $K_d$ and $H$.</td>
<td>73</td>
</tr>
<tr>
<td>5.5</td>
<td>Validation by ChIP-chip data</td>
<td>79</td>
</tr>
<tr>
<td>5.6</td>
<td>Validation by RNA interference data</td>
<td>79</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

2.1 Molecular configurations and the statistical weights of configurations ........................................ 7

3.1 A toy example of modeling TF binding intensities ................................................................. 14
3.2 Three motifs of Nanog ........................................................................................................... 18
3.3 The comparison of ROC curves from three classification methods ........................................... 22

4.1 Modeling the epigenome and the genome as a physical system .............................................. 29
4.2 Epigenomic regulation of transcriptional noise ........................................................................ 35
4.3 Epigenomic boost of cooperativity of weak binding sites ..................................................... 38
4.4 Comparison of different cutoffs on calling strong and weak binding sites ............................... 40
4.5 Transcription factor-specific epigenomic codes ...................................................................... 41
4.6 Epigenomic marks improve model predictions of Nanog binding ........................................... 44
4.7 Comparison of different metrics of prediction power .............................................................. 45
4.8 Comparison of window sizes on model predictions ................................................................. 46
4.9 Comparing model predictions with different sequence motifs .............................................. 48
4.10 H3K9ac and H3K4me2 dampen personal variation of NFκB binding ....................................... 53
4.11 Variations of the strengths of NFκB binding regions ............................................................. 54
4.12 Interactions of NFκB and epigenomic marks ......................................................................... 55
4.13 H3K9ac and H3K4me2 are associated with small variation of NFκB binding ......................... 56

5.1 Flowchart of the Interaction-Identifier method ................................................................. 62
5.2 Forms of one TF-RNAP interactions and their corresponding parameters for modeling the probability of RNAP binding ................................................................. 64
5.3 Forms of two TF-RNAP interactions and their corresponding parameters for modeling the probability of RNAP binding ................................................................. 66
5.4 Results from synthetic data using the Interaction-Identifier algorithm ................................ 71
5.5 The identified regulatory network among Oct4, Sox2 and Nanog ........................................... 74
5.6 Activation regulations of Oct4 and Nanog on target genes identified using Interaction-Identifier ................................................................. 75
5.7 Repression regulations of Oct4 and Nanog on target genes identified using Interaction-Identifier. .......................... 76
5.8 Flowchart of the Network-Identifier algorithm. .................. 77
5.9 The gene regulatory network identified by Network-Identifier . 81
CHAPTER 1
INTRODUCTION

Gene Regulatory Networks

Gene regulatory networks play a pivotal role in controlling the expression of thousands of genes in developmental processes. The complex regulatory systems are implemented by regulatory codes in genomic sequences, which specify the sets of genes that must be expressed in precise spatial and temporal manners. Serving as input signals of the control systems, multiple regulatory proteins, also known as transcription factors (TFs), would recognize and bind to specific regulatory sequences, and exert transcriptional control of the associated genes. Thus, the gene expression patterns of a cell type can be maintained. The interactions among regulatory proteins, their target sequences and associated genes are the main components of a transcriptional network.

Challenges

Several fundamental and important questions are raised in the studies of transcriptional regulation. First, how genomic regulatory sequences attract TFs? In the presence of only one TF, it has been known that a TF generally binds on its specific set of DNA sequences [1][2]. However, the interactions between TFs and how such interactions affect TF-DNA binding are not well understood. It is not clear, at a quantitative level, what are the roles of such TF interactions, and what are the participating TFs. Moreover, the information used to quantitatively control TF-DNA binding not only is encoded in the genomic sequences, but likely is also embedded in the chemical modifications, commonly referred as epigenomic modifications, on the genomic sequences or the nearby histones [3]. These modifications can alter the chromatin structure and function by changing the charge of the nucleosome or
directly interacting with TFs. Such epigenomic influences on TF-DNA binding have not yet being quantitatively modeled.

Second, how do regulatory sequences integrate the signals from TFs and generate the precise expression patterns? It has been recognized that changes in cis-regulatory sequences could alter TF binding events and further lead to different expression patterns of target genes [4]. Some association rule-based methods [5][6] have been proposed to identify the relationship between sequence and expression, but how to quantitatively solve the sequence-expression function is still an open question. Some simplifications are usually assumed, such as prior knowledge about the set of the TFs contributing to the regulation of the expression, or promoter regions as the only regulatory sequences. Several quantitative methods [7][8] have been developed to model the transcription process in two steps: 1) inferring the binding of TF molecules on DNA based on the TF concentration and the binding affinity of the sequence; and 2) predicting gene expression from the combination of TFs bound to DNA.

Third, another major challenge in the study of gene regulation is to identify the regulatory relationships and to reconstruct gene regulatory networks. A number of analytical methods [9][10][11][12][13][14] have been proposed to reconstruct gene regulatory networks based on gene expression data and protein-DNA binding data. However, most of these methods primarily rely on the correlation between the expression of transcription regulators and that of target genes. They fail to reveal the underlying biophysical properties of combinatorial TF-DNA binding.

**New Technologies**

The completion of reference genomes for model organisms has laid the foundation of studying functional sequence elements for gene regulation. The advent of chromatin immunoprecipitation (ChIP) coupled with high-throughput microarray (ChIP-chip) or massively parallel DNA sequencing (ChIP-seq) technologies have enhanced the capabilities of exploring the unknown of regulatory mechanisms. These technologies generate genome-wide TF binding data, which provide new opportunities to closely examine the interaction of TFs and their target genes. ChIP technologies have also been utilized to identify genome-wide patterns of histone modifications. Together with methy-
lated DNA immunoprecipitation sequencing (MeDIP-Seq) and methylation-sensitive restriction enzyme sequencing (MRE-seq) to detect DNA methylation across the genome, epigenomes have become measurable. The availability of epigenomic data makes it possible to explore complex regulatory functions of epigenomes on transcriptional regulation. The most direct output is the readout of gene expression data, usually generated by gene expression microarrays. Recently, microarray technology is gradually replaced by RNA-seq technology due to its high accuracy and affordable costs. These technologies allow us to measure the expression levels of thousands of genes simultaneously.

The Need of Quantitative Frameworks

The different forms of genome-wide data including TF binding ChIP-seq data, epigenomic modification data, gene expression and genomic sequence data present distinct pieces of the puzzle of transcriptional regulation. The huge amount of data urges quantitative methodologies that can bring in all information to study gene regulation. Moreover, the transcription system itself is essentially a quantitative system. The concentrations of TFs as input to the transcription system are quantitatively related to the transcripts as output. The fluctuation of TF concentrations would induce the changes of TF binding intensities, resulting in changes of gene expression levels. Finally, some important biophysical properties, such as synergistic effects of two TFs helping each other bind onto DNA are specific to the interacting TFs. Such specificity can only be appreciated in quantitative models.

Quantitative Models Incorporating Statistical Mechanics

Quantitative models incorporating the underlying mechanisms of transcriptional regulation offer a unique advantage: the capability of gaining mechanistic understandings, such as interaction among TFs and DNA and the epigenomic effects on TF-DNA binding. By fitting real data to such models, it becomes possible to learn more about the regulatory mechanisms, such as the interaction rules of TFs.

Statistical mechanics, or statistical thermodynamics, provides a powerful framework to learn the statistical distribution of molecular-level microstates.
in a physical system characterized and defined by thermodynamic properties. Statistical mechanics makes it possible to understand and to interpret the macroscopic properties of a system in terms of the underlying constituent molecules and the interactions among them. Following the work pioneered by Shea and Ackers [15], this thesis work is built on statistical mechanics to study the stochastic interactions of TFs and regulatory sequences, which is briefly introduced in Chapter 2.

Chapter 3 presents a computational method to predict TF binding affinity from DNA sequences by incorporating both TF-DNA and TF-TF interactions. The method can efficiently calculate TF-DNA binding affinities for a long stretch (several hundred basepairs) of genomic sequence, taking into account interactions between strong and weak as well as homotypical and heterotypic TF binding sites. The method helps us understand how the interactions among TFs contribute to binding affinities.

Both the genomic sequences and the epigenomic modifications define the regional diversity of the regulatory genome. Less clear is how the genome and the epigenome jointly encode and deliver the information to TFs. In Chapter 4, we present an epigenome-sensitive approach to systematically analyze the epigenomic functions in modulating transcription factor-DNA binding. We discovered preferences of TFs for specific combinations of epigenomic modifications, termed as epigenomic motifs. Epigenomic motifs explain why some TFs appear to have different DNA binding motifs derived from in vivo and in vitro experiments. The data suggest that the epigenome can modulate transcriptional noise and boost the cooperativity of weak TF binding sites. We also show that the epigenome might suppress the TF binding differences on SNP-containing binding sites in two people, in theory and in real data.

To identify TF-target regulatory relationships is one major goal in studying gene regulation. In Chapter 5, we developed an analytical method to explore the effects of combinatorial control of TFs on gene expressions. The method enables us to identify a statistical thermodynamic model that best describes the form of TF-TF interaction among a set of TFs for every target gene. Based on this method, we developed a computational framework to infer regulatory relationships from multiple time course gene expression datasets. RNA interference data and large scale TF-DNA binding data independently validated a statistically significant fraction of these regulatory relationships. Moreover, this framework has the flexibility to incorporate other independent
datasets to increase prediction accuracy. Finally, we give a brief summary and future direction of this thesis work in Chapter 6.
In this chapter, I introduce some background materials, covering relevant aspects of statistical thermodynamic modeling of regulatory sequences, and a brief biological background about embryonic stem cells, our research subject, throughout the thesis.

2.1 Statistical Thermodynamic Modeling on Transcription Factor Binding

Thermodynamics was first introduced in physics to study the conversion of energy into work or heat of a system from a macroscopic point of view. Statistical mechanics incorporating statistical tools with thermodynamic principles provides a powerful framework to model and further to predict the collective motion of molecules at the microscopic level on the basis of known characteristics and interactions of a system.

The statistical thermodynamic concept was first adopted on the study of molecular mechanism for gene regulation in Bacteriophage Lambda [15]. Later it was further utilized on modeling TF-DNA and TF-RNA polymerase (RNAP) interactions in bacteria [16][17][18], based on the assumption that the level of gene expression is proportional to the equilibrium probability that RNAP is bound to the promoter of interested gene; and these probabilities can be computed in a statistical mechanics framework. These models brought the stochastic interactions of TFs, regulatory sequences and RNAP together, and enabled a quantitative model for the transcription rate in prokaryotes. Following Buchler et al [18] and Bintu et al [16][17], we begin with an overview of the statistical thermodynamic framework on transcription factor binding onto DNA.
Figure 2.1: **Molecular configurations and the statistical weights of configurations.** The sequence contains three TFBSs for two TFs (A and B). The first TFBS for A (red box on the left) and the TFBS for B are close enough for their bound TFs to interact (arrows in States 5 and 8). \( c \): a physical state; \( W(c) \): Boltzmann weight for state \( c \), which is proportional to the probability that the system visits this state; \( q \): the binding affinity between a transcription factor and the TFBSs.

### 2.1.1 Model Assumptions

First, a DNA sequence is associated with a physical state, which is defined by the combination of transcription factors (TFs) bound to the sequence. When we consider one piece of genomic sequence a time, the physical state of a sequence can be regarded as the physical state of a cell. Second, TF-DNA binding has reached thermodynamic equilibrium, which implies the proportion of cells at each physical state does not change over time. Third, the binding affinity between a TF to any genomic location is a joint effect of multiple TFBSs in the “neighborhood” of this genomic location. Each TFBS has its own binding strength, and they may cooperate if within a certain distance.

### 2.1.2 Model Formulation

A genomic sequence \( (S) \) is modeled as a physical system. Every transcription factor binding site (TFBS) in \( S \) can exist in one of the two physical states, occupied or not occupied by a TF. Thus, a sequence containing \( n \) TFBSs can exist in any of the total of \( 2^n \) states (Figure 2.1 shows the \( 2^3 \) states for a sequence containing 3 TFBSs). Let \( c \) denote a state, and \( C \) denote all states. There is a certain probability associated with every state of the system, denoted as \( P(c) \). Such a probabilistic distribution is called a Boltzmann
distribution [18].

From the perspective of a particular TF (named A), the event that A is bound to sequence S is equivalent to the union of some of states of S. In the example in Figure 2.1, the event “A is bound” is the union of States 2, 4, 5, 6, 7, and 8. We call these states the occupied states (O). Obviously, \( C = O + \bar{O} \). The probability that A is bound to S is

\[
P_A(O) = \frac{\sum_{c \in O} P(c)}{\sum_{c \in C} P(c)}.
\]  

(2.1)

We introduce the Boltzmann weight, \( W(c) \), for every state \( c \). \( W(c) \) is proportional to \( P(c) \) in the way that

\[
P(c) = \frac{W(c)}{\sum_{c \in C} W(c)}.
\]  

(2.2)

Thus, the probability that A is bound to S is

\[
P_A(O) = \frac{\sum_{c \in O} W(c)}{\sum_{c \in C} W(c)}.
\]  

(2.3)

Two factors contribute to \( W(c) \). The first factor is the binding affinity between the TF, A, and every TFBS. We denote this factor as \( q \). The second factor is the cooperativity between TFBSs, denoted as \( \omega \). The Boltzmann weight \( W(c) \) is modeled as

\[
W(c) = q \cdot \omega = \prod_{i=1}^{n} (q_i)^{O_i} \cdot \prod_{i<j} \omega(i, j)^{O_i O_j},
\]  

(2.4)

where \( i \) and \( j \) are the indices of the TFBSs on \( S \); and \( o_i \) is the indicator of whether the \( i \)th TFBS is occupied (\( o_i = 1 \), if occupied; \( o_i = 0 \), otherwise). This formulation implies that the state with no TFBSs bound (\( o_i = 0 \) for every \( i \)) has a Boltzmann weight of 1 (State 1 in Figure 2.1). Suppose the \( i \)th and the \( j \)th TFBSs are bound by TFs A and B, respectively, \( \omega(i, j) \) is modeled as

\[
\omega(i, j) = \begin{cases} 
> 1 & \text{cooperative binding} \\
1 & \text{independent binding} \\
< 1 & \text{competitive binding}
\end{cases}
\]  

(2.5)
The binding affinity $q_i$ between TF $A$ and the $i$th TFBS (denoted by $S_i$) is contributed by two factors: the TF concentration $[A]$ and the preference of the TF to bind onto the binding site sequence $S_i$, denoted as $K(S_i)$. These are modeled as

$$q_i = [A] \cdot K(S_i)$$  \hspace{1cm} (2.6)

$$= [A] \cdot K(S_{\text{con}}) e^{-\Delta E(S_i)}$$  \hspace{1cm} (2.7)

$$= [A] \cdot \eta K(S_{\text{con}})(\text{LLR}(S_{\text{con}}) - \text{LLR}(S)),$$  \hspace{1cm} (2.8)

where $K(S_i)$ is the association constant of the binding site $S_i$. We note that $S_{\text{con}}$ is the strongest binding site (consensus binding site) of this specific TF $A$ and $\Delta E(S_i)$ is the “mismatch energy” of $S_i$ relative to $S_{\text{con}}$. In the other word, $\Delta E(S_i)$ denotes the extra energy needed to bind onto a non-consensus sequence. Following the existing works [1][19], the extra energy is proportional to the log likelihood ratio (LLR) score of a site:

$$e^{-\Delta E(S_i)} = \eta(\text{LLR}(S_{\text{con}}) - \text{LLR}(S)),$$  \hspace{1cm} (2.9)

where $\eta$ is a free parameter, and the LLR score could be computed based on the position specific weight matrix (PSWM) of the TF and the nucleotides background distribution.

### 2.2 Embryonic Stem Cells

Embryonic stem cells (ESCs) are derived from early mammalian embryos and can be propagated through apparently unlimited, undifferentiated proliferation (self-renewal) in cultured cell lines (mouse: [20] [21], human: [22]). ESCs possess several notable properties that account for their exceptional scientific and medical importance. ESCs have remarkable potential to develop into many different cell types in the body (known as “pluripotency” [23]) and therefore they may be used to study body development, both normal and abnormal. A major challenge in the study of ESCs is to explain how the complex gene network is wired to control their properties of pluripotency and self-renewal. Transcriptional control is thought to be a key control
mechanism for ESCs to maintain their undifferentiated state [24] [25] [26] [27] [28] [29] [30] [31]. Regulatory proteins and relevant genomic sequences work together to precisely tune the expression levels of thousands of target genes in ESCs. The interactions among these regulatory proteins and their interactions with particular genomic sequences collectively define a transcription network. Understanding of the part of the network at work in ESCs, i.e. the functional state of the transcription network in ESCs, can reveal how the undifferentiated state of ESCs is maintained, and how it can be disrupted to initiate different routes of differentiation.
CHAPTE_3

ANALYZING LARGE-SCALE
TRANSCRIPTION FACTOR-DNA
BINDING DATA THROUGH STATISTICAL
MECHANICAL MODELING

Transcriptional regulation, as the first step of controlling gene expression, is of fundamental importance in many cellular processes. The transcriptional level of a gene can be controlled by a combination of transcription factors (TFs) that interact with one or several genomic (cis) regulatory regions, including promoter and enhancers. Thus, quantitatively modeling the binding of multiple TFs to a cis-regulatory sequence has been a focal question of understanding transcriptional regulation.

The blossom of chromatin immunoprecipitation followed by either genome tiling array (ChIP-chip) or sequencing (ChIP-Seq) technologies generated genome-wide TF binding data, which enables us to explore the principle of TF-DNA interactions, which further serves as a bridge to understanding the complex transcriptional regulation. The importance of cooperative interactions has been reported in previous studies [32][33]. However, the interactions among TF molecules that promote or inhibit their binding affinities on DNA binding are not well understood.

In this chapter, we propose a quantitative model incorporating TF-DNA and TF-TF interactions to predict TF binding affinity from DNA sequence, called STAP (Sequence To Affinity Prediction). STAP utilizes dynamic programming technique to efficiently compute the binding energy of a long piece of DNA. By analyzing genome-wide data, the model is able to learn how cooperative interactions among TFs contribute to DNA binding. STAP was applied to analyze the ChIP-seq data of 12 TFs in mouse embryonic stem cells [34]. The results demonstrated the strength of TF-DNA binding could be significantly modulated by cooperative interactions among TFs with adjacent binding sites. This chapter is based on [35].
3.1 Related Works

Various computational methods from statistical learning [5][36] and machine learning [6][37][38] were proposed to extract sequence features and learn combinatorial and interaction rules of the sequence features. However, these methods were unable to reflect the underlying biophysical principle of TF-DNA interactions. Moreover, certain important sequence features, such as possible interactions among adjacent TF molecules, are usually neglected in these approaches.

The statistical mechanical models enjoyed an advantage of having physical interpretations of their model parameters, and thus the fitted models provide direct insights to regulatory mechanisms. Shea et al. [15] and Buchler et al. [18] pioneered the biophysical-based approaches on gene regulation. Numerous works [7][39][40][41] following the statistical mechanical principles provide one step forward to model expression data. However, these methods have not addressed the interaction among multiple transcription factors and DNA. Most of these methods were only conducted on simulation studies or certain regulatory sequences. Thus, a statistical mechanical method incorporating combinatorial interactions among TFs and DNA, which enables to analyze genome-wide TF binding data, is still missing.

3.2 A Statistical Mechanical Model of Transcription Factor-DNA Interaction

3.2.1 Prediction of Binding Intensities from Sequences

Given a sequence $S$, our goal is to predict the binding intensity of a TF of the interest. We first scan the sequence with position-specific weight matrices (PSWM) of TFs under a very-low-threshold constraint to identify putative binding sites. Note that in theory every position should be considered a possible binding site (i.e. able to be bound by any TF). Here we adopt a very-low-threshold constraint to filter out those very weak sites for computational purpose. However, the constraint still allow us to consider both strong and weak binding sites in the sequence.

The assumption is that: not only the TF of our interest ($A$, following
the notation in Section 2.1) is attracted by its own TFBS but other adjacent binding sites of the same TF (homotypic cooperativity) or different TFs (heterotypic cooperativity) could interact with A and help stabilize A’s binding. Based on the previous model [18] reviewed in Section 2.1, the weight of a state is related to the interaction of TF and a TFBS as well as the possible TF-TF interactions (Equation 2.4, \( W(c) \) in Figure 3.1). To think from a statistical point of view, we further assume that the binding intensity of the TF of our interest, A, on the sequence \( S \) is proportional to the expected number of the TF A molecules bound on their binding sites:

\[
\bar{N}_A = \frac{\sum_{c \in \mathcal{O}} N_A(c) W(c)}{\sum_{c \in \mathcal{C}} W(c)}, \tag{3.1}
\]

where \( N_A(c) \) is the number of A molecules bound in state \( c \) (Figure 3.1). Note that for two TF molecules to interact their distance should be close enough. We simply use a maximum distance threshold to filter out impossible interactions.

Because the number of states is exponential to the number of sites in a sequence, the brute-force computation of the above quantity is expensive. The computation of the partition function in Equation 3.1 (the denominator) follows the transfer matrix method in statistical mechanics and is similar to the dynamic programming algorithms in other related work [7]. We show that dynamic programming can also be applied to compute the summation in the numerator. Let \( c[i] \) be one state up to the site \( i \), where \( i \) is bound by its cognate TF \( f_i \), we define:

\[
Z(i) = \sum_{c[i]} W(c[i]) \tag{3.2}
\]

and

\[
Y_A(i) = \sum_{c[i]} W(c[i]) N_A(c[i]). \tag{3.3}
\]

We have the following recurrence equations:

\[
Z(i) = q(i) \left[ \sum_{j \in \Phi(i)} \omega(i, j) Z(j) + 1 \right], \tag{3.4}
\]

\[
Y_A(i) = q(i) \left\{ \sum_{j \in \Phi(i)} \omega(i, j) [Y_A(j) + I(f_i, A) Z(j)] + I(f_i, A) \right\}, \tag{3.5}
\]
Figure 3.1: A toy example of modeling TF binding intensities. The sequence contains three TFBSs for two TFs (A and B). The first TFBS for A (red box on the left) and the TFBS for B are close enough for their bound TFs to interact (arrows in States 5 and 8). c: a physical state; W(c): Boltzmann weight for state c, which is proportional to the probability that the system visits this state; q: the binding affinity between a transcription factor and the TFBSs. N_A(c): the number of molecule A bound on the sequence.

where Φ(i) is the set of sites before i and I(f_i, k) is the indicator variable of whether f_i is equal to k. Finally we have

\[
\sum_c W(c) = 1 + \sum_{i=1}^{n} Z(i) \tag{3.6}
\]

and

\[
\sum_c W(c) N_A(c) = \sum_{i=1}^{n} Y_A(i). \tag{3.7}
\]

The alternative way to predict the binding intensity of the TF of our interest A on the sequence S is simply using the probability that A is bound to S by summing over the statistical weights of configurations with A binding on DNA against the summation of statistical weights from all configurations.

\[
P_A(O) = \frac{\sum_{c\in O} W(c)}{\sum_{c\in C} W(c)}. \tag{3.8}
\]

The latter one might be a better approximation to TF binding intensities in ChIP-seq experiments. The reason is that it is unlikely for multiple TF molecules bound simultaneously on the regions around hundreds of base pairs long under ChIP-seq experiment settings.
3.2.2 Model Inferences

The model takes inputs as a set of sequences and their measured binding intensities to the TF of the interest (also called the primary factor), a set of TF motifs (including the TF of the interest) and learns TF-binding models that can be used to predict binding affinity of any new sequence. A TF-binding model consists of two parts: the set of cooperative factors, and the free parameters, which include \( [TF] \cdot \eta K(S_{con}) \) (Equation 2.8 in Section 2.1) for each factor, and the interaction parameters (\( \omega \) in Equation 2.4 in Section 2.1) between the primary factor and any cooperative factors (including self-cooperative interactions). We note that when there are more than one cooperative factor, we do not allow interactions among these factors, as doing so will greatly increase the number of parameters (quadratic to the number of factors), and we may not be able to estimate them since we only have binding data for the primary factor.

3.2.3 Statistical Learning and Computational Strategy

The STAP model is fitted by maximizing the performance in the form of Pearson correlation coefficient between the predicted binding affinities and the overlapping ChIP-seq counts (or ChIP-chip intensities). At the first step of creating the binding model, we identify the motifs from the input motif collection that are cooperative to the primary factor. For each of these motifs, we calculate the predictive power learned from the model using this motif as well as the primary factor. We estimate the statistical significance of this predictive power by comparing with a null distribution of predictive powers constructed from 1000 randomized motifs. Specifically, we randomly choose a motif from JASPAR [42], which could be different from the input motif collection, and then randomly shuffle the columns of this motif. The predictive power (referred to Pearson correlation coefficient) of the model using this random motif and the primary factor will be estimated. We use p-value 0.05 as the threshold for significance judgment. After learning all significant motifs, we combine them into a single model and estimate the model parameters.

For parameter estimation, we use the combination of the Nelder-Mead simplex method and the quasi-Newton method (the BFGS algorithm), both
provided in the GNU Scientific Library [43]. We alternate the two optimization methods until the solutions converge (as defined by the respective criterion of the two methods) or a specified number of alternations are reached. This approach is not guaranteed to find the global optimum, but we find through simulation that it usually produces reasonable solutions, while the global optimization method we tested, Simulated Annealing, is too slow for our purpose.

When running STAP on a TF ChIP-seq binding dataset, we generally need to use only a subset of data for training the binding model, while the rest can be used as testing data. In our experiments with mouse embryonic stem cell ChIP-seq data, we first identify the peak positions of the strongest bound regions (provided in both cases from our data sources) and extract the surrounding sequences, defined as 250 bp upstream/downstream of the peaks. Since these sequences only represent regions bound by TFs, we also add an equal number of sequences that do not show significant binding, chosen randomly from the genome. The binding affinities of these negative sequences are not always available, so we use some value below the lowest binding affinity among all bound sequences, as the substitute of measurements. In our experiments, the size of the training data is 1000 sequences (500 for both positive and negative sets). Our construction of testing data is similar: we choose the next 500 bound sequences and 500 random unbound sequences.

3.3 Cooperative Interactions Among Transcription Factors for DNA Binding in Mouse Embryonic Stem Cells

3.3.1 A Novel Characterization of Nanog Binding Specificity

We studied ChIP-seq data on 12 TFs active in embryonic stems cells [30]: cMyc, CTCF, E2f1, Esrrb, Klf4, Nanog, nMyc, Oct4, Sox2, STAT3, Tcfcp2l1 and Zfx. At the first step, we identified the motifs of the 12 TFs. For each factor, we ran the MEME program [44] on the top 100 regions (ranked by tag counts) detected in the ChIP-seq experiments. These motifs are by and large similar to those reported in the original ChIP-seq paper [34]. However,
we noted that the motifs of Oct4, Sox2 and Nanog, learned by [34] were remarkably similar to each other. We hypothesized that this similarity was due to co-localization of the factors, which resulted in similar collections of genomic regions being used for enrichment-based motif finding. To test this hypotheses, we used sequences bound exclusively by each of these three factors and performed MEME analysis again (NestedMICA [45] and Gibbs sampler [46] gave similar results). The resulting Oct4 and Sox2 motifs are similar to the corresponding parts of the previously identified Oct4-Sox2 joint motif, while the Nanog motif is different (Figure 3.2, Nanog1). We also noted that several other DNA binding profiles of Nanog were reported from previous studies [34][47], but they do not resemble each other. Inspired by the importance of Nanog as an essential regulator in ESC proliferation and self-renewal [47], we set out to characterize the binding specificity of Nanog using a combination of computational and experimental approaches.

Even though STAP was not designed for de novo motif finding, it is applicable to compare multiple motifs of the same factor. By setting these motifs as alternative inputs and comparing the model fit to genome-wide binding data, the best motif can be recognized. We applied this strategy to the new Nanog motif as well as two previously published ones (Nanog2 [47] and Nanog3 [28], Figure 3.2A) to test if the new motif better explains the ChIP-seq data. The new Nanog motif resulted in a much higher correlation than the other two in the sequences bound only by Nanog, but not Oct4 and Sox2 (Figure 3.2B, Nanog-only), providing initial support to the novel Nanog motif. In a second test, we utilized STAP’s capability of analyzing cases where multiple factors are bound. As discussed before, the enrichment of Oct4 and Sox2 binding sites in the Nanog-bound sequences tend to confuse the motif discovery tools. This obstacle was resolved by setting Oct4 and Sox2 as cooperative factors, and varying the candidate primary motif. In this way, the difference of results was attributed to the different Nanog motifs, with the effects of Oct4 and Sox2 sites automatically disentangled. Again, the new Nanog motif provided a significantly better fit to the ChIP-seq counts of the Nanog bound sequences than the other motifs (Figure 3.2B, Nanog-500). In addition, the fitting of observations with the new Nanog motif is highly significant under a test using randomized motifs.
Figure 3.2: **Three motifs of Nanog.** (A) Nanog1 found in this work; Nanog2 from [47]; Nanog3 from [28]. (B) The performance of our model using different motifs. Two sequence sets are used in the analysis. Nanog-only: the Nanog ChIP-seq binding sequences excluding those with Oct4 and Sox2 TFBSs. Nanog-500: The top 500 enriched Nanog binding regions from ChIP-seq experiments.
3.3.2 Cooperativity Among TFs is Frequently Associated with DNA Binding

We next identified cooperative interactions among TFs for DNA binding. For each ChIP-seq experiment, we created training and testing data sets, each consisting of 500 bound and 500 randomly chosen unbound sequences. STAP was applied to learn the significant cooperative factors (among all eleven possible candidates) for each experiment in the training data, following the procedure described above (Table 3.1). This analysis reproduced some known (functional or physical) interactions, including Sox2-Oct4 [27] and cMyc-E2f1 [48]. In addition, the pairs Nanog-Esrrb and Oct4-Esrrb, which were reported to interact in ESCs [49][50], exhibited small p-values (0.06 and 0.08 respectively). The results also suggested that Klf4 may cooperate with a number of other factors, i.e. Oct4, Sox2, Nanog and STAT3. Klf4 facilitates self-renewal of ESCs and promotes the efficiency of inducing pluripotency [51], through mechanisms that are not completely clear. The predicted cooperative interactions between Klf4 and other key TFs may underlie the function of Klf4. Using the independent testing data, we were able to confirm most of predicted interactions. All cooperative pairs, except CTCF as a co-factor of Klf4, improved the basic models where only the primary factor was used, in the testing data, suggesting that the results were not due to model overfitting. These results seem to suggest that even though eleven motifs were tested simultaneously at each experiment, the significance threshold (p-value = 0.05) is stringent in practice. We therefore chose not to further correct multiple hypothesis testing.

After training a single binding model for each factor using all its significant cooperative factors, we compared the effectiveness of this cooperative model with the “non-cooperative model” where no cooperative interaction (not even self-cooperativity) is allowed, in the independent testing data. For most factors, incorporating TF interactions substantially improved the predictive ability of the models (Table 3.1). These results were consistent with our initial intuition that incorporating TF-TF interactions may improve the predictive model, and hence we recommend the final trained model for predictive purposes (to classify a new sequence as being bound to the TF or not). Interestingly, for CTCF and to a small extent Zfx, the cooperative model outperformed the non-cooperative one, even though no significant co-
Table 3.1: Importance of cooperative interactions. Non-cooperative (non-coop.) model: only the motif of the primary factor; Cooperative (coop.) model: allow cooperative interactions. Significance of a cooperative factor is determined through comparison with a large number of randomized motifs.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Non-coop.</th>
<th>Coop.</th>
<th>Improvement</th>
<th>Significant Coop. Factor (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cMyc</td>
<td>0.57</td>
<td>0.82</td>
<td>44%</td>
<td>E2f1(0.004), Klf4(0.04), Zfx(0.033)</td>
</tr>
<tr>
<td>CTCF</td>
<td>0.75</td>
<td>0.81</td>
<td>7%</td>
<td></td>
</tr>
<tr>
<td>E2f1</td>
<td>0.50</td>
<td>0.66</td>
<td>31%</td>
<td>Nanog(0.048)</td>
</tr>
<tr>
<td>Esrrb</td>
<td>0.62</td>
<td>0.78</td>
<td>26%</td>
<td>Zfx(0.003)</td>
</tr>
<tr>
<td>Klf4</td>
<td>0.58</td>
<td>0.74</td>
<td>28%</td>
<td>CTCF(0)</td>
</tr>
<tr>
<td>Nanog</td>
<td>0.24</td>
<td>0.50</td>
<td>107%</td>
<td>Sox2(0), Klf4(0.012), Zfx(0.05)</td>
</tr>
<tr>
<td>nMyc</td>
<td>0.67</td>
<td>0.83</td>
<td>23%</td>
<td>E2f1(0.005)</td>
</tr>
<tr>
<td>Oct4</td>
<td>0.45</td>
<td>0.56</td>
<td>22%</td>
<td>E2f1(0.029), Klf4(0.032), Zfx(0.017)</td>
</tr>
<tr>
<td>Sox2</td>
<td>0.50</td>
<td>0.62</td>
<td>24%</td>
<td>Klf4(0.014), Oct4(0.039), Zfx(0.045)</td>
</tr>
<tr>
<td>STAT3</td>
<td>0.52</td>
<td>0.65</td>
<td>24%</td>
<td>Klf4(0.004), E2f1(0.049), Zfx(0.039)</td>
</tr>
<tr>
<td>Tcfc2l1</td>
<td>0.74</td>
<td>0.76</td>
<td>3%</td>
<td>Esrrb(0.121)</td>
</tr>
<tr>
<td>Zfx</td>
<td>0.70</td>
<td>0.71</td>
<td>1%</td>
<td></td>
</tr>
</tbody>
</table>

operative factor was found, suggesting that self-cooperativity may play a role in these factors.

To explore other interacting factors that did not have genome-wide binding data, we repeated the above analysis using motifs from the JASPAR database [42], in addition to the motifs in this dataset. We found several cooperative pairs involving factors not in the original TF list in ChIP-seq experiments, including for example, Elk1-Klf4, SP1-Nanog, Zfx-TFAP2A and GABPA-Oct4. The most interesting pair seems to be GABPA-Oct4. GABPA expression is known to be induced in undifferentiated ES cells and its expression decreases during differentiation [52]. Moreover, GABPA has been shown to regulate the expression of Oct4 in mouse ESCs [53]. Thus, it would be interesting to test experimentally how GABPA is related to the function of Oct4. This is an example where our method can be utilized to automatically discover biologically plausible hypothesis from existing resources of DNA binding and motif data.
3.3.3 STAP Improves Prediction of TF Targets over Existing Methods

An intended application of STAP is to use the learned binding model to predict affinities of unseen sequences to a set of TFs. An initial support to this application came from the results above showing incorporating cooperative interactions were more predictive than simple models without interactions (Table 3.1). We then compared STAP with the existing methods that are also capable of predicting TF target sequences. Two popular programs were chosen for this purpose, Cluster-Buster [54] and Stubb [55]. Both programs take a set of TF motifs as input, and predict if some binding site clusters appear in a test sequence. To use these programs to predict the targets of some TF, it was necessary to obtain the relevant motifs (in addition to the motif of this TF). Neither program provides such capabilities, and therefore we used another program Clover for this purpose [56]. In summary, the executed procedure of applying these two programs was: first learn all over-represented motifs using Clover from TF-bound sequences in the training data, and then classify all sequences in the test data using Cluster-Buster or Stubb (the same training and testing data as used in the previous section). We evaluated the classification performance with the standard ROC curves, which quantifies the tradeoff of specificity and sensitivity as the classification threshold varies.

Clover identified a number of overrepresented motifs from the collection of 12 motifs of the 12 assayed TFs. These results were similar to STAP’s predictions in some aspects: both predicted few interacting factors for CTCF, E2f1 and Esrrb, and some pairs were predicted by both including Nanog-Sox2 and Tcfcp2l1-Esrrb. But Clover and STAP generated quite different results on other factors. We noticed that Clover results were largely parallel to the co-localization results in [34], with Oct4, Sox2, Nanog and Esrrb forming a cluster of mutually interacting factors. Clover effectively identified motifs whose presence in the training sequences and the motif set was simply applied to predict TF targets by Cluster-Buster and Stubb. In almost all cases, STAP better classified the sequences in the testing data than the other two programs (see Figure 3.3).
Figure 3.3: The comparison of ROC curves from three classification methods.
3.4 Discussions

In this work, we adapted the theoretical models pioneered by Shea-Ackers [15] and formulated by Buchler et al. [18] to the analysis of large-scale TF binding data. Different from these previous works, we developed a dynamic programming algorithm that efficiently computes the binding affinity of a long piece of DNA (in several hundred base pairs). STAP enables to automatically learn interaction parameters of TFs from the binding data. Through extensive evaluations, we demonstrated that this is an effective computational framework to extract information from TF-DNA binding data.

STAP was applied to several important analysis tasks, including comparison of TF binding profiles, identification of TF interactions, and prediction of TF target sequences. These tasks are commonly encountered in analysis of genome-wide data, and we believe STAP offers key benefits over existing methods. First, STAP was applied to compare several putative Nanog motifs. Such functionality can be useful, for example, when one needs to compare outputs from multiple motif-finding programs or from different experiments. Furthermore, when multiple factors access the same target regions, STAP is able to disentangle the effects of confounding factors. This was demonstrated in the analysis of Nanog-bound sequences, which are often bound by Oct4 and Sox2 as well. Second, we took advantage of the new method to predict TF-TF interactions. Similar analyses were done previously by first predicting the binding sites of the pair of motifs, and then analyzing the co-occurrence pattern of two types of sites [57][58]. Co-occurrence based analysis does not utilize the measured TF-binding intensities, sacrificing a significant amount of available information. Co-occurrence based analysis also requires the explicit annotation of binding sites, a task known for its inaccuracy. Weak binding sites were shown to contribute significantly to TF binding [7], making a binary demarcation of sites and non-sites more problematic. Finally, we demonstrated that STAP is able to make more accurate predictions of TF targets in new sequences than other state-of-the-art programs.

Combinatorial gene regulation by definition involves the relationship among different transcription factors. However, how such relationships should be defined and inferred is not clear in practice. We assumed that cooperative interactions are due to protein-protein interactions, but this may not always be true. For example, the transcription factor B may stimulate DNA-
binding of the transcription factor A through chromatin modification that makes DNA more accessible. It is difficult to distinguish different mechanisms of cooperative interactions when only DNA binding data is available. This is important for interpreting the results, as the predictions may not be confirmable through protein-protein interaction assays. In addition, this suggests that the cooperative interactions, as defined by stimulated effects of DNA binding on another factor, may not be symmetric.
CHAPTER 4

ON THREE-WAY INTERACTIONS OF TRANSCRIPTION FACTOR, DNA AND EPIGENOME

Central to transcriptional regulation of gene expression is the regulation of the quantities of transcription factors (TFs) bound to genomic regulatory sequences. The information used to quantitatively control TF-DNA binding is not only encoded in the genomic sequences, but likely is also embedded in the chemical modifications to the genomic sequences and the nearby histones [3]. The chemical modifications (called epigenomic modifications) include the addition of a methyl group or a hydroxymethyl group to the 5th carbon of cytosine (5-mC and 5-hmC) and a number of posttranslational modifications to the histone proteins [59]. These modifications can alter the chromatin structure and function by changing the charge of the nucleosome or directly interacting with TFs [60]. In turn, TFs can tether DNA modification enzymes and histone modification enzymes to change the epigenomic (epi-) modifications around the TF binding region. Hence, both the genomic sequences and the epi-modifications define the regional diversity of the regulatory genome.

Less clear is how the genome and the epigenome jointly encode and deliver the information to TFs. Such information instructs the TFs which parts of the genome they should interact with and the intensities of every interaction.

In this chapter, we present a statistical thermodynamic model that incorporates epigenomic modifications. This model can learn synergistic and antagonistic interactions between specific TFs and epigenomic modifications from genome-wide TF binding and epigenomic data. We used this model together with new experiments to explore a few central questions on the mechanisms of TF-DNA binding. First, to what extent does an epigenomic modification change the binding strength between a TF and a genomic sequence, which is composed of multiple strong and weak binding sites? Second, is the epigenomic influence to TF-DNA binding invariant to the nucleotide composition of the genomic sequence? Third, many TFs have preferred DNA recognition codes (a.k.a. motifs); are there TF-specific epigenomic recogni-
tion codes? Fourth, does the epigenome modulate the variability (noise) of gene expression in an isogenic cell population? Finally, what is the role of the epigenome in modulating individual variation of TF-binding among humans?

We used two complementary experimental systems to study the above questions. The first system is mouse embryonic stem (mES) cells. We experimentally assayed genome-wide distributions of 5-methylcytosine (5-mC), 5-hydroxymethylcytosine (5-hmC), histone variant H2A.Z, and acetylation of histone 3 lysine 27 (H3K27ac). We combined these data with published chromatin immunoprecipitation followed by sequencing (ChIP-seq) datasets of five other epigenomic modifications [59][61][62] and nine TFs [34] from mES cells. This combined dataset allowed us to study TF-epigenome-DNA interactions relatively comprehensively. The second system is the white blood cells of seven people, which allowed us to explore individual differences in humans.

4.1 Related Works

Genome-wide distributions of TF-binding and epigenomic modifications can now be obtained by high-throughput sequencing methods [63]. The explosive growth of data urges the methodological developments that can achieve mechanistic understanding of gene regulation. In particular, quantitative models are needed to learn the regulatory rules implemented by epigenomic modifications. Two classes of methods were developed to study transcriptional regulation with different goals and mathematical foundations. The first class of methods aims at deriving regulator-target relationships or finding regulatory sequences and motifs. These methods were built upon statistical associations among sequence patterns, TF binding, and gene expression [64][5][65][38][66][36][37]. An advantage of this class of methods is that it is easy to incorporate new data types including epigenomic modifications. Indeed, using statistical enrichment and machine learning ideas, recent efforts have incorporated nucleosome positions [67] and epigenomic modifications to identify TFBSs [68] and regulatory genomic sequences [67][69][70].

The second class of methods aims at deriving molecular mechanisms of TF-DNA interactions, using a statistical thermodynamic framework (reviewed in [71]). The intensity of TF-DNA binding was modeled as the equilibrium
Table 4.1: Statistical thermodynamic models of TF-DNA binding

<table>
<thead>
<tr>
<th>Methods</th>
<th>Model assumptions</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Var. TFBS affinities</td>
<td>TF coop.</td>
</tr>
<tr>
<td>Shea et al [15]</td>
<td>N</td>
<td>$P^1$</td>
</tr>
<tr>
<td>Bulcher et al [18]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Janssens et al [40]</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Segal et al [7]</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Gertz et al [39]</td>
<td>N</td>
<td>$P^1$</td>
</tr>
<tr>
<td>He et al [35]</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Fakhouri et al [74]</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Raveh-Sadka et al [72]</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Mirny [73]</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>This work</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>


$P^1$: TF-TF interactions are modeled as invariant to any participating TFs.
$P^2$: Nucleosome positions are considered but histone/DNA modifications are not.
Var. TFBS affinities: TFBS affinities are modeled as specific to the TFBS sequence.
TF coop.: Cooperativity between TFBSs.
Epi-data: Modeling epigenomic environment.
HT data: Applicability to analyzing high-throughput genomic data.

output of input sequences and TFs [15][18]. Partially due to a huge computational burden, this class of methods was originally restricted to analyze a few selected regulatory sequences in single-cell organisms, where a few simplified assumptions can be made [15][18][39] (Section 2.1.1, Table 4.1). These models were extended to analyze nucleosome positions [72][73], gene expression in drosophila embryonic development [7][40][74], and genome-wide TF binding data [35] (Chapter 3). The latter development offered a unique advantage, which is the capability of gaining mechanistic understanding of TF-TF interaction and TF-DNA binding from genome-wide binding data. However, this class of models cannot easily take into account epigenomic modifications, which are argued to be more influential to TF-DNA binding than cooperative interactions between TFs [75][76].

27
4.2 An Epigenome-Sensitive Transcription Factor-DNA Model

We developed a quantitative model for TF-DNA binding in a given epigenomic context. The goal of this model is to predict the binding intensity of a TF in any genomic region in any cell type, using the genomic sequence and the epigenomic modifications (cell-type-specific data). This model incorporates four types of biophysical information: the active concentrations of the TFs, the binding preferences of these TFs to DNA (motif), the nucleotide composition of the genomic sequence, and the epigenomic modifications. Given input data including position-specific weight matrices (PSWM), ChIP-seq derived TF binding sequences and binding intensities, and genome-wide distribution of epigenomic modifications, this model can learn cooperativity among TFBSs (any number of strong and weak, homotypic and heterotypic TFBSs). More importantly, it can learn synergistic and antagonistic interactions between a specific TF and every assayed epigenomic modification. The learning process involves two steps (Figure 4.1B). First, the model scans each epigenomic mark independently to identify those that interact with the transcription factor of interest and modulate its binding affinities to genomic sequences. Second, these identified epigenomic marks are combined into one unified model to predict the binding affinity of any genomic region. The model quantifies the improvements of predicted binding affinities by using the identified epigenomic marks.

4.2.1 Epigenome Effects on Transcription Factor Binding Affinities

TF-DNA binding is not only contributed by genomic sequences, but also affected by epigenomic modifications. Thus, we introduce the binding affinity between a TF and a TFBS under the epigenomic context, denoted as $q^{\text{epi}}$. The Boltzmann weight described in Equation 2.4 is rewritten by

$$W(c) = q^{\text{epi}} \cdot \omega = \prod_{i=1}^{n} (q_i^{\text{epi} i})^{O_i} \cdot \prod_{i<j} \omega(i, j)^{O_i O_j},$$

(4.1)
Figure 4.1: **Modeling the epigenome and the genome as a physical system.** (A) The states of the system and their probabilities. As an example, a hypothetical genomic sequence is occupied by two epigenomic modifications (orange and gray shades), which partially overlap. The sequence contains three TFBSs for two TFs (A and B). The two TFBSs for A (red boxes) are each occupied by one epigenomic modification, and the TFBS for B (green box) is located in the overlapping region of the two modifications. The first TFBS for A (red box on the left) and the TFBS for B are close enough for their bound TFs to interact (arrows in States 5 and 8). Because each of the three TFBSs can reside in either the bound or the unbound state, the whole sequence can reside in a total of 23 physical states (listed in the State column). c: a physical state; W(c): Boltzmann weight for state c, which is proportional to the probability that the system visits this state; q^{epi}_{c}: the binding affinity between a transcription factor and the sequence under the epigenomic context. (B) The workflow for inferring epi-marks that influence the binding of a TF. Central to this workflow is our epigenome-sensitive TF-DNA binding model (the Epi-sensitive TF-DNA model). Inputs to this model are TF binding data (ChIP-seq), PSWM of the TF and epigenomic modification data (ChIP-seq, 5-hmC-seq, MeDIP-seq, and MRE-seq). Outputs of the model include the influences of epigenomic marks to the binding of each transcription factor and the cooperativities between TFBSs.
where \(i\) and \(j\) are the indices of the TFBSs on the sequence \(S\); and \(o_i\) is the indicator of whether the \(i\)th TFBS is occupied (\(o_i = 1\), if occupied; \(o_i = 0\), otherwise).

The binding affinity \(q_{epi}^i\) between a TF (\(A\), following the notion in Section 2.1) and the \(i\)th TFBS on \(S\) denoted as \(S_i\) is related to three factors: the TF concentration \([A]\), the preference of the TF to bind onto the binding site sequence \(S_i\), denoted as \(K(S_i)\), and the epigenomic influence \((\omega_A^{epi}(S_i))\). The binding affinity is modeled as

\[
q_{epi}^i = [A] \cdot K(S_i) \cdot \omega_A^{epi}(S_i) \tag{4.2}
\]

\[
q_{epi}^i = [A] \cdot K(S_i) \cdot \prod_k \omega_k^{A}(S_i), \tag{4.3}
\]

where \(K(S_i)\) is the association constant of the binding site \(S_i\), \(k\) is an index for each type of epigenomic modification and \(\omega_k^{A}(S_i)\) represents the influence of the \(k\)th epigenomic modification on the binding intensity on \(S_i\). We model the TFBS-specific epigenomic influence \(\omega_k^{A}(S_i)\) as follows. Let \(\omega_k^{A}\) be the overall effect of the \(k\)th epigenomic modification to transcription factor \(A\) binding,

\[
\omega_k^{A} = \begin{cases} 
> 1 & \text{epigenomic modification} \ k \ \text{promotes binding} \\
1 & \text{no influence} \\
< 1 & \text{epigenomic modification} \ k \ \text{suppresses binding} 
\end{cases} \tag{4.4}
\]

The TFBS-specific effect \(\omega_k^{A}(S_i)\) is a joint effect of the overall effect \((\omega_k^{A}(S))\) and the intensity of the \(k\)th epigenomic modification on \(S_i\), denoted as \(I_k(S_i)\). Taking the ChIP-seq data for a histone modification for example, \(I_k(S_i)\) is measured by the ratio of the number of sequencing reads between the experimental sample and the control sample. This study used the number of extended sequencing reads falling on \(S_i\). We model the joint effect as

\[
\omega_k^{A}(S_i) = \begin{cases} 
(\omega_k^{A})^{I_k(S_i)} & \text{if } I_k(S_i) \geq \sigma \\
1 & \text{otherwise} 
\end{cases} \tag{4.5}
\]

where \(\sigma\) is a threshold determining whether the measured intensity is beyond noise level. We note that \(\omega_k^{A}(S_i) = 1\) implies either there is no detectable \(k\)th modification or the \(k\)th modification has no influence to the binding. Figure 4.1A illustrates how this model works for a sequence with three TFBSs and
two partially overlapping epigenomic modifications. We call this model an epigenome-sensitive TF-DNA binding model.

4.2.2 Model Inferences

This model has two major applications. One is to predict the binding intensities of a TF throughout the genome in any cell type. The other application is to learn genomic-location-specific epigenomic influences on TF binding, i.e. $\omega_k^A(S_i)$. A third and relatively minor application is to learn the cooperativity between TFBSs in different epigenomic contexts. The required inputs are the genome sequence, the PSWM of the TF, and the epigenomic data. Epigenomic data are often generated by ChIP-seq, MeDIP-seq, and other sequencing based experiments. Standard analysis packages, including sequence mapping [77] and mapped reads postprocessing [78] can process each dataset into a genome-wide distribution of the intensity of an epigenomic modification. Our model takes such a distribution as an input through $I_k(S_i)$, the intensity of the $k$th epigenomic modification on $S_i$.

4.2.3 Statistical Learning with the Model

The model has two sets of model parameters, which are the cooperativity between TFBSs ($\omega_{A,B}$) and the influence of each epigenomic modification ($\omega_k^A$). To train these model parameters, four inputs are required. These include the genome sequence, the PSWM of the TF, the epigenomic data (ChIP-seq and other forms), and the ChIP-seq data of the TF of interest. Let $I(A)$ be the genome-wide distribution of binding intensities of transcription factor $A$. For example, if we segregate the human genome into 15 million 200bp long windows, then $I(A)$ is a vector of 15 million elements. Each element represents the ChIP-seq measured binding intensity in the corresponding window. Following previous notations, we use $P_A(O)$ in Equation 2.3 to denote the model predicted binding probability of $A$ in every window. We propose to learn the model parameters by maximizing the following target function

$$f(\omega_k^A, \omega_{A,B}) = \text{corr}(P_A(O), I(A))$$  \hspace{1cm} (4.6)
where \( corr(\cdot) \) is the Pearson Correlation and \( P_A(O) \) is a function of \( \omega^A_k \) and \( \omega_{A,B} \).

### 4.2.4 Computational Strategy

We implement a maximization strategy to maximize \( f(\omega^A_k, \omega_{A,B}) \). We maximize it by the Nelder-Mead simplex method provided in the GNU Scientific Library [43]. We start with random initial parameters and repeat it 500 times to avoid local minima. In applications where the cooperativity among TFBSs is not of interest, we propose to ignore the cooperativity term (set \( \omega_{A,B} = 1 \)) and only maximize with respect to \( \omega^A_k \).

We adapted the same way to generate the training and testing data from TF ChIP-seq datasets just like in Chapter 3. In general, we only need to use a subset of ChIP-seq data for training the model, while the rest of the data could be used as testing data. In the mouse embryonic stem cell experiments, we used 1000 sequences, including 500 for positive and 500 for negative sets, as our training dataset. The construction of testing data was in the same fashion, 500 bound regions and 500 random unbound regions. For human lymphocyte cell data, we extracted the surrounding 1000bp sequences centered on the centers of those peak regions and replaced the sequences with individual SNPs as individual sequence files. In the comparison between GM12878 and GM18505, we identified 91 peak regions as the DSDB set and 1035 peak regions as the DSNDB set. We collected 300 sequences that did not show binding by randomly chosen from the genome as negative data. In our DSDB experiment, we used 61 and 30 positive data along with 200 and 100 negative data as the training and testing data, respectively. In the DSNDB experiment we used 240 and 120 positive data with 200 and 100 negative data as the training and testing data, respectively. The four-fold cross validation method was performed to avoid the overfitting problem.

### 4.2.5 Identification of Transcription Factor-Specific Epigenomic Interactions

We identify an epigenomic modification \( k \) as influencing the binding of TF \( A \) when \( \omega^A_k \gg 1 \) (promoting) or \( \omega^A_k \ll 1 \) (suppressing). To test for the null
hypothesis that $\omega_k^A = 1$, we shuffle the intensities of epigenomic modification $k$ on the genome to obtain 200 random epigenomic profiles. We subsequently compute 200 $\omega_k^A$ values from the shuffled data and use them as the empirical null distribution. For each epigenomic modification $k$, we test $\omega_k^A = 1$ using the empirical null distribution and reject the null hypothesis using a multiple-hypothesis-adjusted p-value [79].

4.3 Simulation Studies Reflect Epigenomic Regulation on Transcriptional Noise and Epigenomic Boost on Weak Transcription Factor Binding Sites

4.3.1 Parameter Settings in Simulation Studies

To better understand the nature of our quantitative model for TF-DNA binding in a given epigenomic context, we explored the dynamic aspects of the model framework. The computation depends on the TF-specific constant (i.e. $[TF] \cdot \eta K(S_{con})$ in Equation 2.8), the site strength, the effect of the epi-mark on TF-DNA binding, and the intensity of the epi-mark. Here we assigned the parameters plausible but arbitrary values. The TF-specific constant was set to be 2.0, learned by our model from Nanog data in mouse embryonic stem cells. The strong, middle, and weak site strengths were set to be 0.9, 0.6, and 0.3, respectively. The effect terms of an epi-mark for promoting and suppressing TF-DNA binding were set to be 1.5 and 0.5, respectively. The intensity of the epi-mark was set to 2. Unless otherwise stated, we used these parameter set in the simulation study.

4.3.2 Epigenomic Regulation of Transcriptional Noise

We asked how epigenomic modifications may modulate transcriptional noise [80] and the cooperativity of TFBSs. To address this question, we used constraint-based simulation studies [81], with the constraints being the physical and empirical limits of TF concentrations and epigenomic states in eukaryotic cells. Transcriptional noise is the variability of gene expression among cells in an isogenic population [80][82][83]. We asked whether the
epigenome can modulate the level of transcriptional noise. We studied simple transcription systems with one TFBS, by examining the change in binding probability as a function of the concentration of the TF and the presence of epigenomic marks. Following the main assumption of thermodynamic models of gene expression, every cell in an isogenic cell population has the same probability of producing a transcript, denoted as \( p = c \cdot P_A(O) \), where \( P_A(O) \) is defined in Equation 2.3 and \( c \) is a constant. The expected number of transcripts is proportional to \( p \), therefore the variability of \( p \) reflects transcriptional noise [72].

Without any epigenomic marks, the binding probability increased as the concentration of the TF increased, forming a sigmoid curve (green curve, Figure 4.2A-B). In the presence of an activation mark, the sigmoid curve shifted to the left (red curve, Figure 4.2A-B) with no overlap to the original curve. Similarly, in the presence of a repression mark, the curve shifted to the right (blue curve, Figure 4.2A-B). In a transcriptional system with one strong TFBS, the binding probability reached a half of the maximum binding probability when the TF concentration reached 10,000 molecules per cell (green curve, Figure 4.2A). In another system containing a weak TFBS, half of the maximum binding was reached at the TF concentration of 300,000 molecules per cell (green curve, Figure 4.2B). These simulated data were consistent with a recent survey that found the range of TF concentrations to be generally between 10,000 and 300,000 molecules per cell in fruit fly, mouse, and human cells [76]. The dynamic range of TF binding probabilities, constrained by the range of TF concentrations, is a major indicator of transcriptional noise [72]. These constraint-based simulations provided a theoretical prediction that in the presence of a strong binding site, an activation mark decreases the dynamic range of binding probabilities and thus suppresses transcriptional noise, whereas a repression mark enhances transcriptional noise (Figure 4.2A). However, in a transcriptional system with a single weak binding site, both activation and repression marks tend to suppress transcriptional noise (Figure 4.2B). The key assumption to these predictions is that the half of total binding probability of a weak (strong) TFBS is reached at about the upper (lower) bound of the available concentrations of the TF.
Figure 4.2: Epigenomic regulation of transcriptional noise.
Transcriptional noise is introduced when the binding probability (y axis) between a TF and its target TFBS falls into a particular range (horizontal yellow band). There is nearly no noise above or below this range, because almost all cells would uniformly have this target TFBS in the bound or the unbound state, respectively. The binding probabilities are constrained by the realistic range (vertical blue band) of TF concentrations in eukaryotic cells (x axis). (A) In the presence of a strong binding site (S), the binding probabilities are shown as functions of the TF concentration and the presence of epigenomic marks (Red curve: activation mark, green: no epigenomic marks, blue: repression mark). Activation marks suppress transcriptional noise by reducing the range of feasible binding probabilities, whereas repression marks enhance transcriptional noise. (B) In the presence of a weak binding site (W), both activation (red) and repression (blue) marks tend to suppress transcriptional noise.
4.3.3 The Epigenome may Boost the Cooperativity of Weak Binding Sites

We asked whether the epigenome could modulate the cooperativity of adjacent TFBSs. To obtain a baseline (no cooperativity) for this analysis, in a simulation study, we fixed the TF concentration ([A] in Equation 4.2) and compared the binding affinities between a strong TFBS and a weak TFBS in various epigenomic conditions. As expected, in the presence of an activation mark, the binding affinity increases as the intensity of this activation modification increases (solid curves, Figure 4.3A), and the reverse is true in the presence of a repression mark (dashed curves, Figure 4.3A). Moreover, an increase of epigenomic intensity produces a smaller difference in the binding affinities of the two TFBSs (solid and dashed curves become closer as epigenomic intensity increases, Figure 4.3A). However, the binding affinity of a weak TFBS cannot surpass the affinity of a strong TFBS in any levels of an epigenomic modification (neither the solid curves nor the dashed curves crossed, Figure 4.3A). In other words, when there is no cooperativity between TFBSs, under the same epigenomic condition, the order of binding strengths among different genomic sequences is fixed. Because TF concentration, [A], is a multiplicative factor that is separate from the rest in the calculation of the binding affinity ($q_{epi}$ in Equation 4.2), changing TF concentration would not change the contributions from other factors to the binding affinity ($q_{epi}^{pr}$). Thus, the analyses above hold for any TF concentrations.

Next, we examined the cooperativity of adjacent TFBSs. With nearly no epigenomic modifications, a simulated genomic sequence containing two weak TFBSs exhibited a binding affinity larger than that of another sequence containing one weak TFBS (dashed and solid blue curves at epigenomic intensity $= 10^{-2}$, Figure 4.3B), but smaller than that of a medium-strength TFBS and a strong TFBS (green and red curves at epigenomic intensity $= 10^{-2}$, Figure 4.3B). As the intensity of an activation mark increased, the binding affinity of the two-weak-TFBS sequence first surpassed that of the medium-strength TFBS and later superseded the strong TFBS to become the sequence with the largest binding strength (Figure 4.3B). This suggests that in the presence of the epigenome, the binding affinities of different genomic sequences may not always be monotonic. Considering that without cooperativity, the binding affinities of different sequences are strictly monotonic (Figure 4.3A),
these data suggest that epigenomic modifications are not only capable of increasing the binding affinity of each of the two weak TFBSs, but also can increase the cooperativity between the two TFBSs.

Finally, we examined whether the binding affinity of the two weak TFBSs could surpass that of the medium-strength TFBS within the range of typical intensities of epigenomic modifications measured by ChIP-seq experiments. The dashed curve and the green curve crossed at the epigenomic intensity of $10^{0.12} (= 1.32)$, corresponding to the enrichment ratio of $e^{1.32} (= 3.74)$ between the number of sequence reads in the input and the control samples. Because the enrichment ratio of these two numbers is typically between 1 and 40 [84], the change of order of the binding affinities of these two simulated genomic sequences can happen in typical epigenomic conditions.

4.3.4 Epigenomic Boost of Weak TFBSs is Potentially a Regulatory Mechanism

With the theoretical understanding that epigenomic modifications can boost the cooperativity of weak TFBSs, we hypothesized that this is a general mechanism of quantitative regulation of gene expression. We explored this hypothesis with tri-methylation of Histone 3 Lysine 4 (H3K4me3) and the transcription factor Oct4, which is essential for maintaining undifferentiation [21][85] of mES cells. Using Oct4 PSWM, we scanned all Oct4 binding regions, which were defined by the peaks in ChIP-seq data in mES cells [34]. We categorized the Oct4 TFBSs into two sets, strong TFBSs (2055 regions) and weak TFBSs (1921 regions). The average H3K4me3 intensity on weak-TFBSs was larger than 150% of that on strong-TFBSs (P-value $< 10^{-20}$, Figure 4.3C). The largest difference of H3K4me3 intensities between the two sets appeared at the center of Oct4 binding regions (Position = 0, Figure 4.3C). This suggests that on Oct4 binding regions throughout the genome, H3K4me3 is more concentrated on those containing only weak sequence motifs. We ruled out promoters as a confounding factor to the association of strong H3K4me3 to weak TFBSs, because weak TFBSs do not preferentially locate in promoters (Chi-square test p-value = 0.907, Table 4.2).

We then asked if these weak-TFBS-only sequences could obtain a larger boost of binding affinity than the other sequences. Our simulation analysis
Figure 4.3: **Epigenomic boost of cooperativity of weak binding sites.** (A) Monotonicity of binding probability for single binding sites. Given the transcription factor concentration, the binding probability (y axis) is shown as functions of the intensities (Epi-intensity, x axis) and types (solid: activation, dashed: repression) of epigenomic modifications and the strengths of the binding sites (red: strong, blue: weak). For a single binding site, the binding probability is monotonic to the strength of the binding site for all intensities of epigenomic modifications (red curves are always above blue curves). (B) Epigenomic boost of binding-site cooperativity. In the presence of an activation mark, the binding probabilities are monotonic for single strong (red), medium-strength (green), and weak (blue) binding sites. A pair of two weak binding sites has a smaller binding probability in the absence of the activation mark (dotted blue line at Epi-intensity = $10^{-2}$). While the intensity of the activation mark increases, the binding probability of this pair of weak sites gradually surpasses that of the medium-strength site and the strong binding site, breaking the monotonicity of single binding sites. (C) Activation mark H3K4me3 has larger average intensities in weak-TFBS regions (blue) than in strong-TFBS-containing regions (red). SD: standard deviation. (D) The difference of model-predicted binding probabilities with and without the epigenome (y axis) is larger in weak-TFBS-only regions (right column) than in the regions containing both strong and weak sites (mixed, middle column), which in turn is larger than in the strong-TFBS-only regions (left column).
Table 4.2: Lack of association between weak TFBSs and promoters. The distribution of strong and weak TFBSs in promoters and other regions are summarized. Chi-square test p-value = 0.907.

<table>
<thead>
<tr>
<th></th>
<th>Strong TFBSs</th>
<th>Weak TFBSs</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promoters</td>
<td>353</td>
<td>2004</td>
<td>2357</td>
</tr>
<tr>
<td>Other regions</td>
<td>1360</td>
<td>6578</td>
<td>7938</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1713</strong></td>
<td><strong>8582</strong></td>
<td><strong>10295</strong></td>
</tr>
</tbody>
</table>

suggested this was the case in theory (Figure 4.3B). We now test it with the measured epigenomic and TF binding intensities in mES cells. We classified the ChIP-seq peaks into three sets, those only containing strong TFBSs, those containing both strong and weak TFBSs (mixed), and those only containing weak TFBSs. We computed the change in Oct4 binding affinities on these sequence sets from not using H3K4me3 ChIP-seq data to using H3K4me3 ChIP-seq data. The weak-TFBS-only set exhibited a larger increase in binding affinities than the mixed set, which in turn had a larger increase than the strong-TFBS-only set (Figure 4.3D, Figure 4.4). These data suggest that the endogenous levels of H3K4me3 in mES cells are sufficient to boost the binding affinity of adjacent weak TFBSs. Thus, epigenomic boost of the binding affinity of adjacent weak TFBSs is not only a theoretical possibility, but also is likely to a wide-spread regulatory mechanism.

4.4 Transcription Factor-Specific Epigenomic Motifs in Mouse Embryonic Stem Cells

4.4.1 Identification of Transcription Factor-Specific Epigenomic Motifs

Even though some epigenomic modifications are assumed to take some general roles in promoting or suppressing TF-DNA binding, little is known whether such epigenomic functions are specific to certain TFs or are general to every TF. To explore this question, we applied our new model to genome-wide distribution data of nine TFs and nine types of epigenomic modifications in mES cells (assayed by ChIP-seq, MeDIP-seq, MRE-seq, and 5-hmC-seq). Thirty interactions between TFs and epigenomic modifications were identi-
Figure 4.4: Comparison of different cutoffs on calling strong and weak binding sites. The difference of model-predicted binding probabilities with and without the epigenome (y axis) is larger in weak-TFBS-only regions (right column) than in the regions containing both strong and weak sites (mixed, middle column), which in turn is larger than in the strong-TFBS-only regions (left column). The thresholds for calling strong sites and weak sites are $K(S_{con}) - 3.5$ and $K(S_{con}) - 7.0$, respectively, where $K(S_{con})$ is the consensus score. These thresholds are different from those used in Figure 4.3D, where the thresholds for calling strong sites and weak sites are $K(S_{con}) - 2.5$ and $K(S_{con}) - 5.0$, respectively.

Fied, forming an interaction network (Figure 4.5A, Table 4.3). Here, “interaction” refers to the positive or negative influence of an epigenome modification to the binding between a TF and DNA. Among the nine epigenome modifications, H3K4me3, H3K27ac, and 5-mC each interacts with a large number of TFs, forming hubs in the interaction network. Among the five epigenome modifications that exhibited repressive roles, only 5-mC represses the mES cell-specific regulators Oct4, Sox2, Nanog, and Stat3. Compared to the hubs, H3K4me1 is more specific. It promotes the binding of Nanog, Sox2 and Stat3. Even more specific are H2A.Z, 5-hmC, H3K9me3, and H3K36me3, which suppress the binding of cMyc and nMyc. These data suggest that not all epigenomic modifications “uniformly” influence every TF. Some epigenome modifications may selectively promote or suppress the binding of specific TFs.

Considering TFs often have recognition preferences to certain short genomic sequences (motifs), we hypothesized that there are TF-specific epigenomic motifs. By an epigenomic motif we refer to a specific combination of epigenome modifications that is characteristic to the in vivo binding sites of a TF. To test this hypothesis, we estimated the influence of every epige-
4.4.2 Epigenomic Motif Improves Predictions of Transcription Factor Binding Intensities

We hypothesized that the predictive power of TF binding intensities should be increased by incorporating the information of epigenomic motifs. In other
Table 4.3: Comparison of model performances with and without epigenomic data. Transcription factor binding and epigenomic data in mES cells were used as inputs. Model-inferred interacting epigenomic marks of each transcription factor (row) are reported (2nd column). Model performances were evaluated with Pearson correlation using both sequence data and epigenomic data (3rd column) and using sequence data alone (4th column). The improvement was quantified as the difference of the correlations divided by the correlation without epigenomic data (5th column). ω_k^A, the overall effect of the kth epigenomic modification to transcription factor A, as defined in Equation 4.4.

<table>
<thead>
<tr>
<th>TF</th>
<th>Interacting epigenomic marks (ω_k^A, training/testing p-value)</th>
<th>Corr. with epi- marks</th>
<th>Corr. w/o epi- marks</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>cMyc</td>
<td>H3K4me3(1.78,0/0), H3K9me3(0.06, 0/0), H3K36me3(0.1, 0/0.005), 5-mC(MRE)(1.82, 0/0), 5-mC(MeDIP)(0.06, 0/0), 5-hmC(0.06, 0/0), H2AZ(0.06, 0/0), H3K27ac(2.53, 0/0)</td>
<td>0.77</td>
<td>0.64</td>
<td>20.3%</td>
</tr>
<tr>
<td>Nanog</td>
<td>H3K4me1(5.25, 0/0), H3K4me3(11.75, 0/0), 5-mC(MRE)(2.92, 0.01/0.045), 5-mC(MeDIP)(0.06, 0.005/0), H3K27ac(9.90, 0/0)</td>
<td>0.63</td>
<td>0.21</td>
<td>200%</td>
</tr>
<tr>
<td>nMyc</td>
<td>H3K4me3(1.49, 0.01/0), H3K9me3(0.16, 0.03/0), 5-mC(MRE)(15.8, 0.01/0), 5-mC(MeDIP)(0.06, 0.01/0), 5-hmC(0.06, 0/0), H2AZ(0.06, 0/0), H3K27ac(1.80, 0/0)</td>
<td>0.78</td>
<td>0.74</td>
<td>5.41%</td>
</tr>
<tr>
<td>Oct4</td>
<td>H3K4me3(1.67, 0/0), H3K27ac(2.15, 0/0), 5-mC(MeDIP)(0.06, 0.015/0.05)</td>
<td>0.57</td>
<td>0.43</td>
<td>32.6%</td>
</tr>
<tr>
<td>Sox2</td>
<td>H3K4me1(2.20, 0/0), H3K4me3(2.57, 0/0), 5-mC(MeDIP)(0.06, 0.005/0.025), H3K27ac(4.21, 0/0)</td>
<td>0.67</td>
<td>0.53</td>
<td>26.4%</td>
</tr>
<tr>
<td>STAT3</td>
<td>H3K4me1(1.40, 0.05/0), H3K27ac(2.41, 0/0), 5-mC(MeDIP)(0.06, 0.015/0.005), H3K4me3(1.94, 0/0.015)</td>
<td>0.59</td>
<td>0.49</td>
<td>20.4%</td>
</tr>
<tr>
<td>Esrrb</td>
<td>No significant marks</td>
<td>N/A</td>
<td>0.81</td>
<td>N/A</td>
</tr>
<tr>
<td>Klf4</td>
<td>H3K27ac(2.38, 0/0)</td>
<td>0.82</td>
<td>0.80</td>
<td>2.5%</td>
</tr>
<tr>
<td>Zfx</td>
<td>H3K27ac(2.19, 0/0)</td>
<td>0.84</td>
<td>0.82</td>
<td>2.43%</td>
</tr>
</tbody>
</table>
words, if epigenomic motifs exist, they should help to better predict TF binding intensities than using DNA sequences alone. Three computational experiments were done to test this hypothesis. We chose the Nanog TF for these experiments, mostly because Nanog is an essential TF in ES cells and Nanog’s DNA recognition motif is not well understood. In the first experiment, we removed the epigenomic data and fed our model with genomic sequences only. Without epigenomic data, our model degenerates into the STAP model (Chapter 3). STAP uses the sequences (500bp) and the TF-specific PSWM to predict TF binding affinities, taking into account all possible interactions among strong and weak TFBSs. To quantify the model’s predictive power, we used Pearson correlation between the ChIP-seq signals (as observed binding intensities) and the model-predicted binding intensities. Pearson correlations were 21.1% and 21.2% in the training and the testing datasets, respectively, providing a baseline predictive power (Control-1 in red, Figure 4.6). We then applied the model to test each epigenomic modification. H3K4me1, H3K27ac and H3K4me3 largely increased the model’s predictive power of Nanog binding intensities from the baseline (red bars, Figure 4.6). These three epigenomic marks were thus inferred as interacting with Nanog. To test the robustness of model inference, we changed the metric for quantifying prediction power into Spearman’s rank correlation (Figure 4.7) and varied window sizes (Figure 4.8). Neither of these changes affected the inferred interacting epigenomic marks.

In the second experiment, we randomly shuffled the genomic positions of the observed epi-modification intensities, generating 200 permuted datasets. Feeding the permuted datasets to the model, we obtained a background distribution of predictive power (Control-2 in red, Figure 4.6). Using this background distribution, we identified three epi-modifications with which the model can significantly better predict TF binding intensities (red bars with * in Figure 4.6, permutation p-value = 0). These three epi-modifications were identified as interacting with Nanog. This permutation experiment used the same number of model parameters and the same amount of data (PSWM, sequence, and epi-data) as the experiment using the original data. It rules out the possibility that the increased predictive power was due to increased model complexities.
Figure 4.6: Epigenomic marks improve model predictions of Nanog binding. Model predicted binding intensities are correlated to ChIP-seq reported binding intensities (y axis: Pearson correlation). The model predictions are based on sequence data alone (Control-1), sequence data plus randomized epigenomic data (Control-2), or sequence data plus one epigenomic mark (other columns). Results on both training data (shaded bars) and testing data (hollow bars) are plotted. Epigenomic marks that significantly improve the predictions of Nanog binding (marked by *) are identified by using the standard deviations of the control experiments (error bars). Combined with the Nanog motif (PSWM) derived from in vivo experiments (red bars), several epigenomic marks can increase the accuracy of predicted binding intensities, achieving Pearson Correlations above 0.47 (H3K4me3 and H3K27ac, red bars). However, combined with the Nanog motif derived from in vitro experiments (blue bars), no epi- mark except H3K27ac can improve the predictions of ChIP-seq measurements. Even for H3K27ac, the Pearson Correlation obtained from the in vitro motif (0.29) is much smaller than the Pearson correlation obtained from the in vivo motif (0.47). None of the four measured epigenomic marks in adipose cells help to better predict Nanog binding in stem cells (green bars), suggesting that cell-type-specific epi- data are required for increasing the prediction accuracy.
Figure 4.7: **Comparison of different metrics of prediction power.** Model predicted binding intensities are correlated to ChIP-seq reported binding intensities (y axis). The model predictions were based on sequence data alone (Control-1), sequence data plus randomized epigenomic data (Control-2), or sequence data plus one epigenomic mark (other columns). Results on training data (shaded bars) and testing data (hollow bars) using Spearman correlation (red bars) and Pearson correlation (blue bars) are plotted. The model inferred influence of each epigenomic mark to Nanog binding ($\omega^A_{k,Nanog}$ in Equation 4.4) is given in the brackets following each mark.

4.4.3 Transcription Factor-Specific Epigenomic Motif is Cell-Type Specific

As the 3rd control experiment, we replaced the epi- modifications in mES cells with the epi- modifications of mouse adipose cells [78] and kept the other data intact. None of the four epi- modifications in mouse adipose cells significantly increased the predictive power of Nanog binding in mES cells (green bars vs. Control-1 and Control-2 in red, Figure 4.6), suggesting our learned TFBS-epigenomic interactions were cell-type specific.
Figure 4.8: **Comparison of window sizes on model predictions.** Model predicted binding intensities are correlated to ChIP-seq reported binding intensities (y axis). The model predictions were based on sequence data alone (Control-1), sequence data plus randomized epigenomic data (Control-2), or sequence data plus one epigenomic mark (other columns). Results on training data (shaded bars) and testing data (hollow bars) with the window sizes of 350bp (red bars) and 500bp (blue bars) are plotted. The model inferred influence of each epigenomic mark to Nanog binding ($\omega_k^{A=Nanog}$ in Equation 4.4) is given in the brackets following each mark.
4.4.4 Epigenome Alone is Less Predictive of Transcription Factor Binding than Epigenome and Genome Combined

We asked to what extent the epigenome can predict TF binding without using the genomic sequences. Two control datasets were generated. First, each epigenomic mark was fed to our model without sequence data ($S_i$ becomes invariant to $i$ in Equation 4.2, solid red bars, Figure 4.9). The enhancer and open chromatin marks H3K4me1 and H3K27ac were most strongly predictive of Nanog binding, followed by the promoter mark H3K4me3. These data are consistent with the idea that open chromatins and hypersensitivity sites are predictive of transcription factor binding regions [87]. Interestingly, H2A.Z is the fourth epigenomic mark that is predictive of Nanog binding. The regulatory function of H2A.Z in mammalian cells remains controversial. While H2A.Z is generally thought as an active mark of transcription, it is negatively correlated with gene expression in a mES cell differentiation process [88]. The positive association of H2A.Z with Nanog binding suggests that H2A.Z may facilitate Nanog binding in undifferentiated mES cells. Second, we collected all (214) PSWMs from the JASPAR database [42] as background motifs. These background PSWMs were fed to the model with each epigenomic mark. The mean and standard deviation of the model predicted binding intensities from these background PSWMs were derived (hollow red bars and error bars, Figure 4.9). The predictive powers of these control datasets were compared to the predictive powers using both epigenomic and PSWM information (blue bars, Figure 4.9). The in vivo Nanog motif combined with epigenomic data (solid blue bars) increased the accuracy of predicted Nanog binding affinities than using epigenomic data alone (red bars). More than 20% increases of predictive power were observed using Nanog motif and H3K4me1 or H3K27ac than using H3K4me1 or H3K27ac alone. Even larger increases were found in comparing in vivo Nanog motif (solid blue bars) with background PSWMs (hollow red bars). The latter comparison used models with the same number of model parameters. It rules out the possibility that the increased predictive power was due to increased model complexities.
Figure 4.9: **Epigenomic marks improve model predictions of Nanog binding.** Using the Nanog dataset, we compared model predictions in four scenarios. In each scenario, the model predictions were correlated to ChIP-seq measured Nanog binding intensities (y axis). These scenarios are: 1, using each epigenomic mark without DNA sequence data (solid pink bars); 2, using each epigenomic mark with all the (214) PSWMs from the JASPAR database (hollow pink bars: the mean of the 214 correlations, error bar: standard deviation of the mean); 3, using each epigenomic modification with the in vivo Nanog motif (solid blue bars); 4, using each epigenomic modification with the in vitro Nanog motif (hollow blue bars).
4.4.5 In vitro Derived Transcription Factor-DNA Binding Motifs do not Interact with Epigenomic Motif

The TF-DNA binding motifs derived from the enriched sequence patterns using in vitro binding assays do not always agree with the enriched motifs from in vivo binding assays [89]. Depending on the TFs, the differences in motifs derived from in vitro and in vivo experiments can be small [90] or large (Section 3.3.1). The causes of such differences are unknown. We hypothesized that some epigenomic modifications can synergize with DNA to produce a somewhat different binding preference of a TF than the binding preference of this TF to naked DNA. To test this hypothesis, we chose to further analyze the Nanog motifs derived in vitro [47] and in vivo (Section 3.3.1). We used the in vitro Nanog motif together with all epigenomic data to learn and predict in vivo binding affinities (blue bars, Figure 4.6) and compared to the results from the in vivo motif (red bars, Figure 4.6). Without considering epigenomic data, the in vitro and in vivo motifs had similar predictive powers of ChIP-seq signals (Control-1 in red vs. Control-1 in blue, Figure 4.6). However, except for H3K27ac, adding epi- modifications to the in vitro motif did not increase the predictive power of Nanog binding. Even for H3K27ac, its contribution to predicting Nanog binding was much larger when combined with the in vivo motif than when combined with the in vitro motif (red and blue H3K27ac bars, Figure 4.6). This means the model failed to identify clear TFBS-epigenomic interactions with the in vitro Nanog motif, suggesting that the epigenomic motif is specific to the in vivo Nanog DNA binding motif. In several cases, including H3K4me3, H3K27me3, H3K36me3, and 5-mC (both MRE and MeDIP), feeding the model with epi- data together with the in vitro motif even slightly decreased its predictive power as compared to not using epi- data at all (blue bars vs. Control-1 in blue, Figure 4.6). This is because the model allowing for TFBS-epigenomic interactions is more complex than that without epi- data. However, there is no extra information added due to the lack of interaction between the in vitro motif and the epi- marks. These data explain the difference between the TF-DNA binding motifs derived in vivo and in vitro: although the Nanog sequence motifs derived in vitro and in vivo have similar binding affinities to the Nanog protein in vitro [35], only the in vivo motif may interact with epi- modifications. The in vivo binding intensities are determined by TFBS-epigenomic interactions and cannot be
faithfully reproduced with the sequence motif (either in vitro or in vivo) alone.

4.5 Epigenome Might Attenuate Personal Variations on Transcription Factor Binding

4.5.1 Individual Variation of Transcription Factor Binding is Greater than Epigenomic Variation Among Humans

From evolutionary and population genomics perspectives, we are interested in comparing the magnitudes of epigenomic and TF binding variations, both between species [88] and within a species. We hypothesized that the individual differences in TF binding are greater than the differences in the epigenome. To test this hypothesis, we used published nuclear factor kappaB (NFκB) binding data [91] and deoxyribonucleaseI (DNaseI) hypersensitivity data [92] from the lymphocytes of human individuals whose genomes were sequenced [77]. DNaseI hypersensitivity levels are correlated with open chromatin and histone acetylations [93]. DNaseI hypersensitivity data were used as a surrogate to epi- modification data, because genome-wide epi-modification measurements in multiple sequenced-determined individuals are not yet available. We first compared two European individuals, GM12878 and GM12892. These were the only two individuals with both NFκB binding data and DNaseI hypersensitivity data available. We quantified the differences of NFκB binding and DNaseI hypersensitivity on all SNP-containing genomic regions. The individual differences of DNaseI hypersensitivity showed a unimodal distribution ranging from 0 to 1 (Figure 4.10A). Here, 0 corresponds to no individual difference (equal digestion levels) and 1 reflects a one fold difference (one person’s digestion level is twice as the other’s). The individual differences in NFκB binding showed a bimodal distribution, with the first mode between 0 and 1 (similar to DNaseI) and the second mode above 1 (unique). The second mode suggested a specific feature of TF binding. Although individual differences are moderate (less than one fold) for both epi-modification and TF binding on the majority of genomic regions, there is a subset of genomic regions where TF binding differences are greater than...
usual (one fold).

To test whether this feature was sensitive to the two chosen individuals, we compared two groups of people. We quantified the variation of DNaseI hypersensitivity levels among 6 people [92], and the variation of NFκB binding among 3 people [91]. The epigenomic variation showed a unimodal distribution. The TF binding variation again showed a bimodal distribution, with the first mode in similar range as epi- variation and a unique second mode above the range of epi- variation (Figure 4.10B). This result is consistent with the direct comparison of two individuals. In summary, while there is a “general level of variation” in TF binding and epi- modification, there is a subset of genomic regions where TF binding variation is greater than this general level.

4.5.2 H3K9ac and H3K4me2 may Dampen the Variation of Transcription Factor Binding Across Human Individuals

Genomic variations including single nucleotide polymorphisms (SNPs) can result in phenotypic variation. Still unknown is how epigenomes modulate the correlation of genotypes and phenotypes among humans. We chose TF binding intensities as a molecular phenotype to study this question.

To study how epi- variation can interact with genomic variation, we did three between-individual comparisons across different ethnic groups. We first compared a European (NIGMS catalog ID: GM12878) and a Nigerian (GM18505). We categorized NFκB binding regions with the TFBSs containing SNPs into two sets. The first set had differences in NFκB binding intensities between these two individuals. This set was called Different Sequence Different Binding (DSDB) (Figure 4.11A). The second SNP-containing set had similar NFκB binding levels in the two individuals, and were termed the Different Sequence No Difference in Binding (DSNDB) set. The first set (DSDB) was consistent with the theory that nucleotide changes in the TFBS should change the binding affinity of this TFBS; however, the second set (DSNDB) appeared to be inconsistent with such a theory. We hypothesized that the epi- marks on DSNDB stabilized the binding affinities of these binding sites. In other words, the epi- modifications on the TFBSs buffered sequence changes (SNPs) from changing binding intensities.
Theoretically, the difference in binding affinities between two TFBSs is the largest without any epi-marks (y-intercept, Figure 4.3A). When epimodification intensities increase, the binding difference in the two TFBSs decreases (from left to right, Figure 4.3A). This is true for any two TFBSs of the same TF. Thus, we have derived a theoretical mechanism for the epigenome to attenuate the TF binding differences on SNP-containing TFBSs in two individuals.

We proceeded to examine whether the theoretical mechanism could have been implemented in humans. We first used our model to learn epi-marks that help to explain the binding intensities in all SNP-containing TFBSs (Table 4.4). Four epi-marks were identified by the model, which were H3K4me1/2, H3K9ac, and H3K27ac (Figure 4.12). Among them, H3K4me2 and H3K9ac were identified as marks that better explain the binding intensities in DSNDB sites. If H3K4me2 and H3K9ac were used to attenuate binding differences between two people, there should be higher intensities of H3K4me2 and H3K9ac in DSNDB sites than in DSDB sites. Indeed, the intensities of H3K4me2 and H3K9ac were much higher in DSNDB sites than in DSDB sites (P-values $< 10^{-20}$, Figure 4.10C). To assess whether these results were specific to the chosen individuals in our analysis, we did two more comparisons. The second comparison was between a European (GM12878) and a Nigerian (GM19099), and the third comparison was between a European descendant (GM12878) and a Japanese (GM18951). Each comparison identified its own DSDB and DSNDB sites. However, all comparisons found significantly higher H3K4me2 and H3K9ac intensities in DSNDB sites than in DSDB sites (Figure 4.13). The NFκB binding intensities in DSDB and DSNDB of GM12878 had similar distributions, and therefore are unlikely to contribute to explain the differences of H3K4me2 and H3K9ac intensities in GM12878 (Figure 4.11B-C). As a control, adding H3K36me3 data to the model did not increase the correlation of model predicted binding intensities to NFκB ChIP-seq data (Figure 4.12). Accordingly, the difference in H3K36me3 levels between DSDB and DSNDB sets was not clear and not consistent in these comparisons (Figure 4.13). These data suggested a mechanistic explanation to the SNPs in TFBSs that do not produce between-individual differences in TF binding: epi-modifications on these TFBSs attenuated the binding differences.
Figure 4.10: H3K9ac and H3K4me2 dampen personal variation of NFκB binding. (A) Histograms of individual differences (x axis) of NFκB binding (green) and DNase I hypersensitivity (orange). Individual difference (D) is quantified by $D = |A - B|/A$, where A and B are normalized sequencing reads on the same genomic region from two people (GM12878 and GM12892). Nearly all genomic regions have moderate individual differences in DNase I hypersensitivity ($D \leq 1$). However, a subset of genomic regions has large individual differences in NFκB binding ($D > 1$). (B) Histograms of average individual differences ($\bar{D}$, x axis) of NFκB binding (green) and DNA hypersensitivity (orange). $\bar{D}$ measures the average individual differences in a group of people (shown in circles). Although very few genomic regions have large average individual differences ($\bar{D} > 1$) of DNase I hypersensitivity, a subset of genomic regions have large average individual differences ($\bar{D} > 1$) of NFκB binding. (C) The average intensities of H3K9ac and H3K4me2 are higher in DSNDB regions (blue) than in DSDB regions (red). The centers of all NFκB ChIP-seq peaks are superimposed to “Position 0” on the x axis. DSNDB: different sequence no difference in binding. DSDB: different sequence different binding. SD: standard deviation.
Figure 4.11: Variations of the strengths of NFκB binding regions. (A) The inter-individual variation of the strengths of NFκB binding regions are quantified by Difference Ratio (DR, y axis), which is defined as $DR = \frac{|I(S_i) - I(S_j)|}{\min(I(S_i), I(S_j))}$, where $I(S_i)$ and $I(S_j)$ are the binding strengths of sequences $S_i$ and $S_j$ in individuals $i$ and $j$ measured by ChIP-seq experiments. The mean (each bar) and standard error (error bars) of DRs in DSDB (left) and DSNDB sequence sets (right) are shown. The distribution of GM12878 NFκB binding in DSDB (left) and DSNDB (right) sequence sets, where DSDB and DSNDB were identified from the comparison of GM12878 and GM18505 (B) and from the comparison of GM12878 and GM12892 (C). CEU: Northern and western Europe. YRI: Nigeria. SE: standard error.

Table 4.4: Distribution of SNP-containing NFκB binding sites.

<table>
<thead>
<tr>
<th></th>
<th>CEU vs. CEU</th>
<th>CEU vs. YRI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GM12878 vs. GM12892</td>
<td>GM12878 vs. GM18505</td>
</tr>
<tr>
<td>DSDB</td>
<td>60</td>
<td>91</td>
</tr>
<tr>
<td>DSNDB</td>
<td>360</td>
<td>1035</td>
</tr>
</tbody>
</table>
Figure 4.12: Interactions of NFκB and epigenomic marks. The Pearson correlation between model-predicted and ChIP-seq measured binding intensities (x axis) is used to identify the epigenomic marks interacting with NFκB. The genomic sequence and ChIP-seq data of GM12878 were used to fit the model. DSDB and DSNDB sequences were identified from comparing sequence and epigenomic data of GM12878 and GM18505. The results from four-fold cross validations are shown. Shaded bars: training data. Hollow bars: testing data. Length of each bar: the average value from four-fold cross validations. A total of 200 randomized epi- datasets were used as controls (Control). Four-fold cross validations were performed on each randomized dataset. The mean correlation from these four-fold cross validations of 200 random datasets is represented by the length of each Control bar. Error bars: standard deviations of the mean. The epigenomic marks that significantly increase the Pearson correlation from the control experiments are identified (*, p-value < 0.01).
Figure 4.13: **H3K9ac and H3K4me2 are associated with small variation of NFκB binding.** Two other comparisons (GM12878 vs. GM19099 and GM12878 vs. GM18951) confirm that the average intensities of H3K9ac and H3K4me2 are higher in DSNDB regions (blue) than in DSDB regions (red). As a control, no reproducible differences between DSNDB and DSDB regions are found for H3K36me3. The centers of all NFκB ChIP-seq peaks are superimposed to “Position 0” on the x axis. DSNDB: different sequence no difference in binding. DSDB: different sequence different binding. SD: standard deviation. CEU: Northern and western Europe. YRI: Nigeria. JPT: Japan.
4.6 Discussions

The overarching tenet of this work is obtaining mechanistic insights from high-throughput genomic data. Towards this goal, we forfeited commonly used “statistical enrichment” methods that look for large overlaps of two or more genomic features. Instead, we developed a statistical mechanical model for the three-way interactions among the genomic sequence, the epimodifications, and TF binding. The model is specified as a physical system, and every model parameter has a biophysical interpretation. This allows the analytical results obtained from this model to have mechanistic interpretations.

Several epimodifications were previously assumed to promote or repress TF binding in a ubiquitous manner. For example, mono-, di-, and tri-methylations on histone lysine 4 (H3K4me1/2/3) were thought to promote binding of any TF. Our data suggested that some TFs tend to preferentially recognize TF-specific epigenomic codes. This implies that rather than ubiquitously promoting or suppressing TF binding, some epimarks can specifically interact with some TFs. This is conceivable because the maintenance of epimarks often require histone or DNA modification enzymes to be brought to a genomic sequence by specific transcription factors [94][95]. In addition, epimodifications are strongly associated with the three-dimensional architectures of the local chromatin [96]. It is also conceivable that some TFs would have different binding preferences to the same DNA sequence but different 3D chromatin conformations.

We showed that epimodifications can boost the cooperativity of adjacent weak TFBSs. Thus, there is a functional advantage of coding a cis-regulatory sequence with a cluster of weak TFBSs rather than one strong binding site. The advantage is that the binding affinity of a cluster of weak TFBSs has a larger tunable range than a strong TFBS, in the presence of the epigenome. Thus, clusters of weak TFBSs offer the epigenome larger “controllability”. This may explain why weak TFBSs tend to cluster in the mammalian genomes [97]. Indeed, H3K4me3 was strongly enriched in Oct4 binding regions that only contained weak TFBSs. Moreover, H3K4me3 generated larger enhancements of binding affinities in the weak-TFBS-only binding regions than in other Oct4 binding regions. Thus, the “epigenomic boost” of TFBS cooperativity is a mechanism implemented in mammalian cells. This provides an alternative view on the evolutionary origin of TFBS
clusters, in which the presence of the epigenome was previously ignored [98]. A central question in personalized medicine is how genomic variation generates phenotypic variation. This is a challenging question because genomic variation was only partially correlated with TF-binding variation [91]. In particular, a set of SNPs in TFBSs does not introduce differences to TF binding as predicted by available TF-DNA binding models. Incorporating the epigenome into the TF-DNA binding model, we can now appreciate that some epimarks can buffer genomic changes from generating changes in TF binding intensities. A case in point is that H3K4me2 and H3K9ac attenuate the personal variation of NFκB binding on SNP-containing binding sites in human lymphocytes. These results highlight the importance of considering the epigenome when analyzing the functional consequences of genomic variations.
Transcriptional control is a key regulatory mechanism for cells to direct their destinies. A large number of transcription factors (TFs) could simultaneously bind to a regulatory sequence. With the constellation of TFs bound, the expression level of a target gene is usually determined by the combinatorial control of a number of TFs. The interactions among regulatory proteins and their regulatory sequences collectively form a regulatory network. A major challenge in the study of gene regulation is to identify the interaction relationships within a regulatory network.

Quantitative models describing gene expression in terms of quantity, speed, timing, and environmental context are essential for the study of many biological processes. Thermodynamic models are based on the assumption that the level of gene expression is proportional to the equilibrium probability that RNA polymerase (RNAP) is bound to the promoter of interested gene; and these probabilities can be computed in a statistical mechanics framework. In prokaryotes under well studied assumptions, a function is available to translate each particular way of interaction among transcription factors (TFs) and RNAP into the level of the expression of the target gene [16][17]. Such functions are termed “regulation factors” [16][18]. There is to date few discussions on the extent to which these regulation factors hold for eukaryotes [18].

In this chapter, we proposed a method, Interaction-Identifier, based on statistical thermodynamic model principles to select the best fit interaction forms (i.e. infer the form of TF-TF and TF-RNAP interactions) for each target gene from time course microarray data. Interaction-Identifier enables the investigation of regulation factors from empirical data in eukaryotic systems. Applying this method to a time course microarray dataset of retinoid acid (RA) induced differentiation of mouse embryonic stem cells(ESCs), we clearly
distinguished different interaction forms among Oct4, Sox2 and Nanog, and their roles of as an activator, a repressor and a helper on each target gene. The detailed characterization of interaction forms among multiple transcription factors allow us to build a core transcription network in ESCs using a bottom-up approach. Along with the same line, We further developed a computational framework, called Network-Identifier, for inferring gene regulatory networks from time course gene expression data. Applying to the analysis of five datasets of differentiation of mouse ESCs, we identified a transcription network composed of 34 TF-TF interactions and 185 TF-target relationships. Data from RNAi [20] and chromatin immunoprecipitation coupled with microarray (ChIP-chip) data [27][51] independently validated a statistically highly significant fraction of these regulatory relationships. This chapter is based on [8] [99].

5.1 Related Works

A number of analytical methods have been proposed to reconstruct gene regulatory networks from gene expression and protein-DNA binding data. Association rule mining [9], Boolean Network [100], temporal models [101],[10], ARACNE [11] and Bayesian networks [12] [13] [14] are among the most popular routes. For example, the Module Networks approach built a probabilistic model for the gene expression correlations between regulators and target genes and iteratively searched for the most compatible partition of targets genes to their respective regulators [102]. The correlation of gene expression patterns of regulators and the target genes is often the essential piece of information utilized by the current procedures. It is widely recognized that the statistical correlation of the regulators and the targets is often an inaccurate representation of the regulator-target relationship [103] [104]. This is because the quantity of a TF’s mRNA does not necessarily correlate to its active protein concentration, and even the active protein concentration does not necessarily correlate to its transcriptional efficiency on every target gene. Using correlation, or some transformed version of correlation measure as the basis for reconstructing regulatory networks is an approximation made for convenience of modeling and analysis, with a sacrifice of making spurious findings (see examples in [102]). A network reconstruction method based
on quantities that closely represent the biophysical properties of TF-DNA binding, transcription activation and repression is still missing.

5.2 Selection of Statistical Thermodynamic Models for Combinatorial Control of Multiple Transcription Factors in Early Differentiation of Embryonic Stem Cells

5.2.1 Identify Interactions Among Transcription Factors and RNA Polymerase

We propose a computational framework, called Interaction-Identifier, to identify the interaction form among the TFs and RNA polymerase (RNAP) on the promoter of a target gene at steady state. This method begins by using a statistical thermodynamic model to predict the equilibrium probability that RNAP binds to the promoter of its targeted gene ($P_{RNAP}$) based on concentrations of associated TFs and interaction forms among TFs and RNAP. Then, a kinetic model is used to simulate the dynamics of expression of target genes, assuming: a) the transcription rate is proportional to the $P_{RNAP}$; b) mRNA degradation rate is linearly dependent on the RNA concentration; c) the concentration changes of TF factor can be inferred from the changes in the mRNA levels of TFs. By searching the space of different TF interaction forms, Interaction-Identifier identifies the underlying TF interaction form of each target gene, which minimizes the difference between the model-derived expression profile and the observed expression data (Figure 5.1).

Statistical Thermodynamic Models for RNAP Binding

Cells receive a wide variety of cellular and environmental signals, which are often processed combinatorially to generate specific genetic responses. We follow Buchler et al [18] and Bintu et al [16][17] to integrate combinatorial signal at the level of cis-regulatory transcription control in bacteria through the statistical thermodynamics of TF-DNA and TF-RNAP-DNA interactions. These interactions can be quantified by several tunable parameters.
Table 5.1: The Boltzmann distribution for the two states of a TFBS

<table>
<thead>
<tr>
<th>State</th>
<th>TF</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>free</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>attached</td>
<td>1</td>
<td>$q_{TF}$</td>
</tr>
</tbody>
</table>

based on different selections and placements of various protein-binding DNA sequences. Under the same theoretical framework as Section 2.1, the model of TF-RNAP-DNA interaction is briefed in the following Section.

**TF-DNA Interactions**

At a given time in a cell, there are only two states for a transcription factor binding site (TFBS): attached with or free of a TF. Let $q_{TF}$ denote as the ratio of the probability of a TFBS in the attached state to that in the free state (Table 5.1).

The probability that the TFBS of a target gene is bound with a TF could
be denoted as

\[ P(TF_{\text{binding}}) = \frac{q_{TF}}{1 + q_{TF}}. \] (5.1)

RNAP-promoter binding (without any TF present) can be described by the same form

\[ P(RNAP_{\text{binding}}) = \frac{q_p}{1 + q_p}. \] (5.2)

**TF-RNAP-DNA Interactions**

Let us look at two different cases of TF-RNAP-DNA interactions in the following.

- **One TF**

  If we consider the case of a TF interacting with a RNAP, there are four possible states for a promoter: (1) bound by both the TF and the RNAP; (2) bound by the RNAP only; (3) bound by the TF only; (4) free from either the TF or the RNAP (Table 5.2). The probability of the promoter of the target gene bound with a RNAP could be represented as

\[ P(RNAP_{\text{binding}}) = \frac{q_p + \omega_{TFp} q_{TF} q_p}{1 + q_p + q_{TF} + \omega_{TFp}}, \] (5.3)

where

\[ \omega_{TFp} = \begin{cases} 
1 & \text{no interaction} \\
10 - 100 & \text{activation} \\
0 & \text{repression}
\end{cases} \] (5.4)

Different settings of \( \omega \) reflect different roles a TF could play. If \( \omega \) is set to 1, it represents that there is no interaction between the RNAP and the TF. They bind independently to the promoter. If \( \omega \) is set to 10-100, it represents that the TF helps recruit the RNAP binding to the promoter. The larger \( \omega \) is, the larger the synergism is. If \( \omega \) is set to 0 or close to 0, it represents that the TF blocks the RNAP binding to the promoter, and thus the TF serves as a repressor (Figure 5.2).

- **Two TFs**

  The case of two TFs capable of binding to a promoter together with a RNAP could be represented in the same fashion (Table 5.3).
Table 5.2: The Boltzmann distribution of a promoter with one TF and one RNAP

<table>
<thead>
<tr>
<th>State</th>
<th>TF</th>
<th>RNAP</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>$q_p$</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>$q_{TF}$</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>$\omega_{TFpq_pq_{TF}}$</td>
</tr>
</tbody>
</table>

1) Simple Repressor

![Diagram of Simple Repressor]

$W_{RP} = 0$

2) Simple Activator

![Diagram of Simple Activator]

$W_{AP} = 10\sim100$

Figure 5.2: Forms of one TF-RNAP interactions and their corresponding parameters for modeling the probability of RNAP binding. A is a transcription factor acting as an activator of genes. R is a transcription factor acting as a repressor of genes. P represents RNAP. The curve with a dot at the end represents a repression effect; the one with an arrow in the end indicates either cooperation between transcription factors or activation of genes by the transcription factor.

Table 5.3: The Boltzmann distribution of a promoter with its RNAP and two TFs

<table>
<thead>
<tr>
<th>$(TF_1, TF_2)$</th>
<th>(0, 0)</th>
<th>(1, 0)</th>
<th>(0, 1)</th>
<th>(1, 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>$q_{TF1}$</td>
<td>$q_{TF2}$</td>
<td>$\omega_{TF1TF2q_pq_{TF1}q_{TF2}}$</td>
</tr>
<tr>
<td>1</td>
<td>$q_p$</td>
<td>$\omega_{TF1p}q_pq_{TF1}$</td>
<td>$\omega_{TF2p}q_pq_{TF2}$</td>
<td>$(\omega_{TF1p} + \omega_{TF2p})\omega_{TF1TF2q_pq_{TF1}q_{TF2}q_p}$</td>
</tr>
</tbody>
</table>
The probability of RNAP binding to the promoter could be denoted as

\[
P(\text{RNAP}_{\text{binding}}) = \frac{\sum_{j} \sum_{k} P(1,j,k)}{\sum_{i,j,k \in \{0,1\}} P(i,j,k)},
\]

where \( P(i, j, k) = P(\text{RNAP} = i, \text{TF}_1 = j, \text{TF}_2 = k) \).

The parameters \( \omega \) could be set differently to reflect the nature of these interactions between two TFs or the interactions between one TF and one RNAP. The parameter \( w_{\text{TF}_1\text{TF}_2} \) is used to simulate the interaction between the two TFs. A large \( w_{\text{TF}_1\text{TF}_2} \) (10-100) represents that the two TFs stabilize each other onto the promoter. If the two TFs have no interaction, \( w_{\text{TF}_1\text{TF}_2} \) should be set to 1. If the two TFs compete for the binding, \( w_{\text{TF}_1\text{TF}_2} \) should be set to 0 or close to 0. The other two parameters, \( w_{\text{TF}_1p} \) and \( w_{\text{TF}_2p} \), represent the interaction between each TF and RNAP, respectively. They can be set to reflect different interactions similar to \( w_{\text{TF}_1\text{TF}_2} \). By adjusting the parameters \( w_{\text{TF}_1p} \), \( w_{\text{TF}_2p} \) and \( w_{\text{TF}_1\text{TF}_2} \), we can obtain an analytical form for the probability of RNAP binding under different forms of interactions among RNAP and the two TFs. Figure 5.2 and Figure 5.3 summarize the parameter choices for two forms of simple interactions and five forms of three-way interactions.

**Linking TF Concentration to the Probability of Promoter Occupancy**

Besides the forms of RNAP binding probability, we describe the influence of TF concentration on the probability of TF binding to the promoter of its target gene in the following.

Let \([\text{TF} - \text{DNA}]\) represent the cellular concentration of the promoter bound by the TF. The binding process can be denoted as

\[
[\text{TF}] + [\text{DNA}] \rightarrow [\text{TF} - \text{DNA}]
\]

Then the probability that the TFBS of a target gene is bound with a TF could be formulated as

\[
P(\text{TF}_{\text{binding}}) = \frac{[\text{TF} - \text{DNA}]}{[\text{DNA}] + [\text{TF} - \text{DNA}]}.
\]

At equilibrium state, the concentrations of the substrates could be de-
Figure 5.3: Forms of two TF-RNAP interactions and their corresponding parameters for modeling the probability of RNAP binding. $A_1$ and $A_2$ are transcription factors acting as activators of genes. $R_1$ and $R_2$ are transcription factors acting as repressors of genes. P represents RNAP. The curve with a dot at the end represents a repression effect; the one with an arrow in the end indicates either cooperativity between transcription factors or activation of genes by the transcription factor.
scribed using the Hill equation

\[
P(\text{TF}_{\text{binding}}) = \frac{[\text{TF}]^H}{[\text{TF}]^H + [K_{\text{TF}}]^H} = \frac{[\text{TF}]^H}{[\text{TF}]^H + 1} = \frac{1}{1 + \frac{[\text{TF}]}{K_{\text{TF}}}}, \quad (5.8)
\]

where \([\text{TF}]\) is the cellular concentration of the activated TF targeted by this site, \(K_{\text{TF}}\) is the effective dissociation constant (relative to the genomic background) representing the concentration required for half of the TF binding to the promoter, and \(H\) is the Hill coefficient. If \(H > 1\), transcription factor binding is positively cooperative; if \(H = 1\), the transcription factor binding is not cooperative; if \(H < 1\), the transcription factor binding is negatively cooperative.

Recall the percentage of promoters bound by TFs can also be described using \(q_{\text{TF}}\), the ratio of the probabilities of the promoter in the bound and free states,

\[
P(\text{TF}_{\text{binding}}) = \frac{[\text{TF}]^H}{[\text{TF}]^H + 1} = \frac{q_{\text{TF}}}{q_{\text{TF}} + 1} = \frac{1}{1 + \frac{[\text{TF}]}{K_{\text{TF}}}}, \quad (5.9)
\]

Thus, we can obtain

\[
q_{\text{TF}} = \left(\frac{[\text{TF}]}{K_{\text{TF}}}\right)^H. \quad (5.10)
\]

We use the unit of \([\text{TF}]\) and \(K_{\text{TF}}\) as the number of TFs per cell. There have been a few efforts to estimate \(K_{\text{TF}}\) from empirical data [2]. In this study, we assume at each time point in the time course, \([\text{TF}]\) is linearly related to the expression level of the TF, as did in earlier module network studies [102]. It follows that \([\text{TF}]\) peaks at the same time as its gene expression peaks. We further assume \(q_{\text{TF}}\) is maximized at the maximum \([\text{TF}]\) (see sensitivity analysis in Section 5.2.2 for further discussion on this assumption). We adopt the value 1/20 for \(q_p\) from [16][17][18].

A Kinetic Model for the Quantity of the mRNA of the Target Gene

With the above statistical thermodynamic model of TF-DNA interactions, we enable to quantify the equilibrium binding probability of the RNAP to the promoter, given the cellular concentrations of all the TFs. However, the bridge connecting from the binding probability of RNAP to the gene
expression levels is still missing. Thus, we further use a kinetic model to analyze the dynamics of gene expression over times.

Assume that the changes of TF concentrations can be inferred from the changes of mRNA levels of TFs, and the mRNA degradation rate are linearly dependent on the mRNA concentration. Thus, based on the principle of statistical thermodynamic models that the transcription rate is proportional to the binding probability of RNAP, an ordinary differential equation was proposed to mimic the dynamics of gene expressions in the following.

\[
\frac{dG}{dt} = K_g(P(RNAP_{binding})) - K_d\left(\frac{G}{G_{max}}\right),
\]

(5.11)

where \(G\) denotes as the transcript concentration (number per cell); \(G_{max}\) denotes as the maximum concentration of the transcript (number per cell); \(K_g\) represents the maximal synthesized rate of transcripts (per minute per cell) and \(K_d\) is the degradation rate of transcripts (per minute per cell).

The maximum rate of mRNA synthesis rate has been estimated to be about one mRNA per 6 – 8 seconds [105]. Following [106] [107], we assume that the rate of degradation around 1/6 of the maximum transcription rate. Therefore, we use \(K_g =10\) counts per minute and \(K_d = 10/6\) counts per minute in this study.

Although gene expressions should be continuous signals throughout the time, an assumption should be made that gene expressions are measured when the transcriptional system is in its equilibrium state at each time point, which is satisfied by all time course microarray data. Under this circumstance, the expression could be represented by

\[
G = \alpha P(RNAP_{binding}),
\]

(5.12)

where

\[
\alpha = G_{max} \frac{K_g}{K_d}.
\]

(5.13)

### Computational Strategy

By combining the statistical thermodynamics models and the kinetic model, we are able to derive the expression profiles from the interesting models shown in Figure 5.2 and Figure 5.3. With measured time course gene expres-
sion data from microarray experiments, we compute the Pearson correlation coefficient between the observed expression pattern and the model-derived expression patterns. Since different combinations of TFs and different interaction forms will lead to different expression patterns, we search the space of TF interaction forms to find the fittest interaction form. Finally, the interaction form that predicts an expression pattern with the highest correlation to the observed expression pattern is identified as the most plausible interaction form that TFs take to regulate this target gene (Figure 5.1). Note that if a gene has all Pearson correlations between the observed expression and model-derived expression patterns not over a user-defined threshold, it might suggest the real TF interaction form is not included in our search space. Thus, Interaction-Identifier would return no interaction form for that particular gene.

We first generate synthetic data to check the practicability of Interaction-Identifier. To further test the robustness of the model, we conduct the sensitivity analysis to explore the effects of choices of parameter settings on method performance. Then we apply Interaction-Identifier to mouse embryonic stem cells (ESCs) data. We infer five interaction patterns among three regulators: Oct4, Sox2 and Nanog on ten target genes.

5.2.2 Simulation Studies

Generation of Simulation Data

As a proof of principle, we first use synthetic data to show the validity of method. We choose three commonly seen regulatory patterns (Figure 5.4). These regulatory patterns are: 1. a target gene is activated by one TF (The simple activator model in Figure 5.2); 2. RNAP is blocked by a TF (repressor), and this TF is stabilized to DNA by a helper TF (Repressor recruited by a helper in Figure 5.3); 3. a target gene is regulated by two interacting activators (Dual activators interacting in Figure 5.3), and one of the two activators is transcriptionally repressed by a third TF.

For each of these three regulatory patterns, we do simulations as follows. First, we simulate the concentration change of each TF over time, which we call \textit{realTFExp} using equations of the format or its variants: \( E_A = \)
\[ a_A + b_A \log T + \epsilon, \] where \( a_A \) and \( b_A \) are background gene expression index and coefficient describing changes of expression index with time \( T \). The \( \epsilon \) represents the variability of expression for gene \( A \). Different patterns of transcription factor expression can be obtained by using different parameters of \( a_A, b_A \) and \( \epsilon \). Assuming that the concentration of TF is a linear transformation of \( E_A \), we feed these simulated concentrations of the TFs into a chosen regulatory pattern described in Figure 5.4 and derive the expression pattern of the target gene (\( realTargetExp \)) according to the statistical thermodynamics models and the kinetic model. Noises following \( N(0, 1) \) are added to all the real expression patterns for both TFs and the target gene. We assume only the noise-added expression patterns are observed, and we denote the observed expression values as \( obsTFExp \) and \( obsTargetExp \). The \( obsTFExp \) for all TFs in consideration are used to derive expression pattern for the target gene under each model in Figure 5.2 and Figure 5.3. The model derived expression patterns are termed \( modelTargetExp \). For each model, \( obsTargetExp \) is compared to \( modelTargetExp \) in terms of Pearson correlation.

The parameters we use in the study are followed the literatures, where \( K_g = 10 \) counts per minute, \( K_d = 10/6 \) counts per minute and \( q_p = 1/20 \). We further assume that \( K_{TF} = \) the maximum \([TF]\) and \( H = 2 \).

**Simulation Data Analysis**

We use three regulatory patterns to test our new algorithm. Under the first regulatory pattern, two simulations are conducted. First, TF’s expression increases linearly over time. \( realTFexp = 500 + 500T, \) where \( T = 2, 4, 8, 16, 32, 64 \) and 128. In the second simulation, TF’s expression increases exponentially over time. \( realTFExp = 500 + 200\log T, \) where \( T = 2, 4, 8, 16, 32, 64 \) and 128. Because there is only one TF in consideration, there are only two candidate regulatory models, either repression (The simple repressor model in Figure 5.2) or activation (The simple activator model in Figure 5.2). In both simulations our method correctly picked our Model 2 (Row 1, Figure 5.4). Two simulations are performed under the second regulatory pattern. For each simulation, our method consistently identifies the correct regulatory model out of five candidate models (Row 2, Figure 5.4). Under the third regulatory pattern, we conduct a two-step analysis. In the first step, we apply the method to judge the regulatory relationship between
Figure 5.4: Results from synthetic data using the **Interaction-Identifier algorithm**. The concentration of A was simulated using either a linear function: \([TF] = 500 + 500T\) or an exponential function: \([TF] = 500 + 200\log T\), where \(T\) represents the time.
TFs A and B (Row 3, Figure 5.4), i.e. one TF is controlling the expression of another TF. After a regulatory model is determined between A and B, we use the expression pattern of B derived from the Step 1 to identify the interaction pattern between TFs B and C. There are two candidate models for Step 1 and five candidate models for Step 2. Altogether 10 potential regulatory models exist among the four genes. In two independent simulations, our method identifies both the correct regulatory models (Row 3, Figure 5.4).

Sensitivity Analysis

We check to what extent the choices of parameters affect the method performance. Regulatory model 7 (the regulatory pattern between B, C, D in Row 3, Figure 5.4) is chosen to perform the sensitivity analysis. We vary $K_{TF}$, $K_g$, $K_d$ and $q_p$ in very wide ranges, for example a 10000 fold range for $K_{TF}$, and re-run our algorithm. Results in Table 5.4 shows that the method can robustly identify the correct regulatory model even if the parameters are offset by 100 fold. The only exceptions are the cases where the synthesis rates of mRNA were set to be too slow: below 1 mRNA molecule every 10 minutes, as compared to the default of 10 mRNA per minute from empirical data. We therefore do not suggest using a very small synthesis rate.

5.2.3 Applications on Mouse Embryonic Stem Cells

Oct4, Sox2 and Nanog are key transcription factors to maintain pluripotency of embryonic stem cells (ESCs). Nanog is known to be jointly regulated by Oct4 and Sox2. For other target genes, we identified the TFs from either literature survey or ChIP-chip data. In this study, we focus on genes regulated by two key transcription factors in embryonic stem cell: Oct4 and Nanog [28].

Time course microarray data have been generated for retinoid acid induced differentiation of mouse embryonic stem cells [20]. Genes that are jointly regulated by Oct4 and Nanog have been reliably identified [28]. Among these target genes, nine genes (Jarid2, Sall4, Rif1, Gbx2, REST, Zin3, Foxc1, Smarcad1 and Atbf1) are represented on the Affymetrix U133 microarray and therefore their time course data are available.
Table 5.4: Sensitivity test for $K_{TF}$, $K_g$, $q_p$, $K_d$ and $H$. The results indicate that the correct model can be identified even with drastic variations in parameters used in the algorithm.

<table>
<thead>
<tr>
<th>Model</th>
<th>$K_{TF}$</th>
<th>Pearson</th>
<th>Kg Pearson</th>
<th>Kd Pearson</th>
<th>$q_p$ Pearson</th>
<th>H Pearson</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.01</td>
<td>0.9500</td>
<td>1/60</td>
<td>0.9671</td>
<td>60/24</td>
<td>0.9671</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.9505</td>
<td>1/6</td>
<td>0.9671</td>
<td>60/30</td>
<td>0.9671</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.9671</td>
<td>10</td>
<td>0.9671</td>
<td>60/36</td>
<td>0.9671</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.9571</td>
<td>600</td>
<td>0.9671</td>
<td>60/42</td>
<td>0.9671</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.9562</td>
<td>1000</td>
<td>0.9671</td>
<td>60/48</td>
<td>0.9671</td>
</tr>
<tr>
<td>4</td>
<td>0.01</td>
<td>-0.9514</td>
<td>1/60</td>
<td>-0.9695</td>
<td>60/4</td>
<td>-0.9695</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>-0.9517</td>
<td>1/6</td>
<td>-0.9695</td>
<td>60/30</td>
<td>-0.9695</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>-0.9695</td>
<td>10</td>
<td>-0.9695</td>
<td>60/36</td>
<td>-0.9695</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>-0.9567</td>
<td>600</td>
<td>-0.9695</td>
<td>60/42</td>
<td>-0.9695</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>-0.9562</td>
<td>1000</td>
<td>-0.9695</td>
<td>60/48</td>
<td>-0.9695</td>
</tr>
<tr>
<td>5</td>
<td>0.01</td>
<td>-0.8822</td>
<td>1/60</td>
<td>-0.7953</td>
<td>60/24</td>
<td>-0.7953</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>-0.9720</td>
<td>1/6</td>
<td>-0.7953</td>
<td>60/30</td>
<td>-0.7953</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>-0.7953</td>
<td>10</td>
<td>-0.7952</td>
<td>60/36</td>
<td>-0.7953</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>-0.6160</td>
<td>600</td>
<td>-0.7952</td>
<td>60/42</td>
<td>-0.7953</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>-0.6125</td>
<td>1000</td>
<td>-0.7952</td>
<td>60/48</td>
<td>-0.7953</td>
</tr>
<tr>
<td>6</td>
<td>0.01</td>
<td>0</td>
<td>1/60</td>
<td>-0.9668</td>
<td>60/24</td>
<td>-0.9668</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>-0.9720</td>
<td>1/6</td>
<td>-0.9668</td>
<td>60/30</td>
<td>-0.9668</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>-0.9668</td>
<td>10</td>
<td>-0.9668</td>
<td>60/36</td>
<td>-0.9668</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>-0.6654</td>
<td>600</td>
<td>-0.9668</td>
<td>60/42</td>
<td>-0.9668</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>-0.6138</td>
<td>1000</td>
<td>-0.9668</td>
<td>60/48</td>
<td>-0.9668</td>
</tr>
<tr>
<td>7</td>
<td>0.01</td>
<td>0.9608</td>
<td>1/60</td>
<td>0.96031</td>
<td>60/24</td>
<td>0.9693</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.9616</td>
<td>1/6</td>
<td>0.96932</td>
<td>60/30</td>
<td>0.9693</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.9693</td>
<td>10</td>
<td>0.96932</td>
<td>60/36</td>
<td>0.9693</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.7092</td>
<td>600</td>
<td>0.96932</td>
<td>60/42</td>
<td>0.9693</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.6143</td>
<td>1000</td>
<td>0.96932</td>
<td>60/48</td>
<td>0.9693</td>
</tr>
</tbody>
</table>
Figure 5.5: The identified regulatory network among Oct4, Sox2 and Nanog. The Interaction-Identifier algorithm is applied to the expression data of Oct4, Sox2, and Nanog from time course microarray data.

For each interaction form in Figure 5.2 and 5.3, we use the differential equation to derive the steady state level of mRNA expression level using the estimated TF concentration $[TF]$ and the specific TF dissociation constant $K_{TF}$ based on measured mRNA levels. We derive a series of steady state mRNA concentrations corresponding to measured expression profile of the target gene. We then compute the Pearson correlation between the derived concentrations of target genes over time and the observed concentrations from the time course microarray data. The interaction form that predicts a concentration dynamics with a largest correlation to the measured expression level is identified as the most plausible interaction form.

Interaction Models for Oct4, Sox2 and Nanog in Mouse ESCs

We apply the Interaction-Identifier method to the regulatory model for Nanog. The time course expression data suggest that Oct4 and Sox2 help each other to stabilize onto the regulatory sequence and attract the RNAP (Figure 5.5).

We then identify the regulatory models for the Oct4 and Nanog regulated genes. Although these nine genes are all regulated by Oct4 and Nanog in ESCs, they are not regulated under the same mechanism. Jarid2, Sall4, Rif1, Zic3, Gbx2, emoes and REST is regulated under model 3, with one TF as an activator and the other as a helper (Figure 5.6). Atbf1 is regulated under model 5 where Oct4 and Nanog are independent repressors (Figure 5.7 (a)). Foxc1 is regulated under model 4 where Nanog is a helper and Oct4 is a repressor (Figure 5.7 (b)). These results suggest that Atbf1 and Foxc1 are probably involved in lineage differentiation and therefore need to be repressed.
Figure 5.6: Activation regulations of Oct4 and Nanog on target genes identified using Interaction-Identifier. The directed arrows represent activation and the dotted line represents the function of the helper. The relationship between Nanog and Oct4 with these target genes follows model 3 in Figure 5.3.

by key transcription factors in ESC. Interestingly, Foxc1 has been shown to be involved in ocular development [108] and Abf1 mRNA is found to be abundant in prostate [109]. Finally, none of the models being considered derives an expression pattern similar to the observed expression pattern of Smarcad1 (All Pearson correlations are smaller than 0.5). This may suggest that besides Oct4 and Nanog, there are other mechanisms responsible for the transcriptional control of Smarcad1.

5.3 Inferring Gene Regulatory Networks by Statistical Thermodynamic Modeling

The interactions among regulatory proteins and their regulatory sequences collectively form a regulatory network, which controls the fate of cells. A major challenge in the study of gene regulation is to identify the interaction relationships within a regulatory network. Based on Interaction-Identifier to select for the statistical thermodynamic model that best describes the TF-TF and TF-RNAP interaction for each target gene, we further develop a computational framework, Network-Identifier, for inferring gene regulatory
Figure 5.7: Repression regulations of Oct4 and Nanog on target genes identified using Interaction-Identifier. (a) model 5 (Figure 5.3) (b) model 4 (Figure 5.3), where the dotted line represents the function of the helper, a line with an arrow in the end represents the effect of activator; a line with a solid dot in the end represents the effect of repressor.

networks from multiple time course gene expression data.

5.3.1 Incorporate Multiple Datasets to Infer Gene Regulatory Networks

Network-Identifier utilizes Interaction-Identifier to find common TF interaction forms of target genes across multiple time course microarray datasets, and then incorporates those predicted regulatory relationships supported by independent datasets into a regulatory network. The method has three components: 1) Interaction-Identifier (See Section 5.2.1), 2) Evidence merger and 3) Verification component, shown in Figure 5.8. In the following section, we will describe each component in details.

Interaction-Identifier Component

Network-Identifier requires more than one time course microarray experiments for the same biological process as input datasets. For each time course dataset, Network-Identifier enumerates all possible regulatory forms on each
target gene. These interaction forms include the activation or repression by a single TF (Figure 5.2), and the five interaction forms between any two TFs (Figure 5.3). For each gene, Network-Identifier evaluates the fitness of each interaction form with Interaction-Identifier (See Section 5.2.1) and ranks them according to their fitness. The ten most likely interaction forms of TFs (i.e. ten interaction forms with the highest Pearson correlation coefficient) on a target gene are recorded in the Top-10 List. A built-in user-defined threshold (default = 0.8) for Interaction-Identifier eliminates any interaction that is not well supported by data. It is therefore possible for a target gene to have less than 10 candidate TF interaction forms in its Top-10 List.

**Evidence Merger Component**

The Top-10 Lists from every dataset are passed onto Evidence merger, which searches for the most frequently appeared interaction form in the Top-10 Lists of a target gene. This most frequently identified interaction form is passed onto the verification component.
Verification Component

The verification component groups target genes according to their TF interaction forms. For each regulator-target relationship, for example TF-1 represses gene a, the target genes grouped into this relationship are subject to statistical tests. Chi-square tests are used to test whether the identified TF-target relationships are enriched with regulatory relationships identified from independent experimental data, such as ChIP-chip and RNA interference (RNAi) data. Finally, if the tests are all insignificant, Network-Identifier will fail to report any regulatory network. If some of these tests are significant, suggesting there is consistency between the expression-derived regulatory relationships and those found by independent methods, Network-Identifier will invoke a compromise algorithm to report the regulatory relationships that are confirmed by at least two independent data sources. Currently the implemented compromise algorithm is to require the regulatory relationship identified by expression data to be reproduced in at least one of the two other experiments: ChIP-chip and RNAi. It is easy to substitute this algorithm with more sophisticated algorithms [110] or when some of the independent data are not available.

5.3.2 Applications on Mouse Embryonic Stem Cells

We employ five time series microarray datasets of mouse ESCs in this study, including a dataset for retinoid acid induced differentiation [20] and four datasets for spontaneous differentiation of four ESC lines (three lines from [52]; one unpublished, S.Z. and W.H.W). We restrict the analysis to the regulatory relationships among 747 genes that are annotated by Gene Ontology term, Transcription Regulator Activity, and are present on the Affymetrix U72av2 array. We designate six known TFs, Oct4, Sox2, Nanog, Klf4, Esrrb and Tcl1 as regulators of this system, due to their previously characterized role in ESCs.

A Gene Regulatory Network of Mouse ESCs

Interaction-Identifier is first applied to each time course microarray dataset. A list of common TF Interaction forms across datasets is then generated.
Table 5.5: Validation by ChIP-chip data

<table>
<thead>
<tr>
<th>Role</th>
<th>TF</th>
<th># of target genes</th>
<th># of genes verified</th>
<th>Chi-Square</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activation</td>
<td>Nanog</td>
<td>39</td>
<td>12</td>
<td>10.46986</td>
<td>0.00121</td>
</tr>
<tr>
<td></td>
<td>Sox2</td>
<td>121</td>
<td>21</td>
<td>12.437</td>
<td>0.00042</td>
</tr>
<tr>
<td></td>
<td>Oct4</td>
<td>67</td>
<td>8</td>
<td>2.436113</td>
<td>0.11857</td>
</tr>
<tr>
<td></td>
<td>Klf4</td>
<td>49</td>
<td>18</td>
<td>13.90787</td>
<td>0.00019</td>
</tr>
<tr>
<td>Repression</td>
<td>Nanog</td>
<td>47</td>
<td>11</td>
<td>4.190152</td>
<td>0.04066</td>
</tr>
<tr>
<td></td>
<td>Sox2</td>
<td>132</td>
<td>19</td>
<td>5.778738</td>
<td>0.01622</td>
</tr>
<tr>
<td></td>
<td>Oct4</td>
<td>103</td>
<td>11</td>
<td>2.121288</td>
<td>0.145264</td>
</tr>
<tr>
<td></td>
<td>Klf4</td>
<td>62</td>
<td>14</td>
<td>1.335151</td>
<td>0.247891</td>
</tr>
</tbody>
</table>

Table 5.6: Validation by RNA interference data

<table>
<thead>
<tr>
<th>Role</th>
<th>TF</th>
<th># of target genes</th>
<th># of genes verified</th>
<th>Chi-Square</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activation</td>
<td>Nanog</td>
<td>39</td>
<td>9</td>
<td>9.710604</td>
<td>0.00183</td>
</tr>
<tr>
<td></td>
<td>Sox2</td>
<td>121</td>
<td>20</td>
<td>16.22083</td>
<td>5.6E-05</td>
</tr>
<tr>
<td></td>
<td>Oct4</td>
<td>67</td>
<td>13</td>
<td>25.26604</td>
<td>5E-07</td>
</tr>
<tr>
<td></td>
<td>Esrrb</td>
<td>95</td>
<td>6</td>
<td>2.966206</td>
<td>0.085021</td>
</tr>
<tr>
<td></td>
<td>Tcl1</td>
<td>21</td>
<td>2</td>
<td>4.650429</td>
<td>0.03105</td>
</tr>
<tr>
<td></td>
<td>Klf4</td>
<td>49</td>
<td>16</td>
<td>25.21262</td>
<td>5.1E-07</td>
</tr>
<tr>
<td>Repression</td>
<td>Nanog</td>
<td>47</td>
<td>2</td>
<td>0.018713</td>
<td>0.891192</td>
</tr>
<tr>
<td></td>
<td>Sox2</td>
<td>132</td>
<td>12</td>
<td>9.035917</td>
<td>0.00265</td>
</tr>
<tr>
<td></td>
<td>Oct4</td>
<td>103</td>
<td>7</td>
<td>3.909397</td>
<td>0.04802</td>
</tr>
<tr>
<td></td>
<td>Esrrb</td>
<td>73</td>
<td>6</td>
<td>5.407721</td>
<td>0.02005</td>
</tr>
<tr>
<td></td>
<td>Tcl1</td>
<td>27</td>
<td>2</td>
<td>4.663594</td>
<td>0.03081</td>
</tr>
<tr>
<td></td>
<td>Klf4</td>
<td>62</td>
<td>10</td>
<td>0.537394</td>
<td>0.463515</td>
</tr>
</tbody>
</table>

by Evidence merger. Genes are then grouped by their predicted regulators as well as their roles of regulation, i.e. activators and repressors. Twelve gene groups are formed. ChIP-chip data are available for Oct4, Sox2, Nanog and Klf4. Five out of eight regulatory-target relationships involving these four regulators are significantly enriched with ChIP-chip verified relationships (Table 5.5). RNA knock-out experiments are performed for all the six regulators [20] [51]. Nine out of twelve target gene groups involving these six regulators are enriched with RNAi verified regulatory relationships (Table 5.6). Note that when using RNAi data for testing the predicted regulatory role of a TF, we only count the target genes whose changes of expression are in the consistent direction to the predicted role of its TF, but not counting all targets genes with any changes to both directions. These tests demonstrate that the predicted regulatory relationships were in general consistent to those derived from independent experiments.
Network-Identifier identifies the regulatory relationships that are predicted by expression data and had consistent evidence from either RNAi or ChIP-chip data. We use Cytoscape [111] to display the final reported regulatory relationships (Figure 5.9).

87 regulators and target genes are reported in the ESC transcription network (Figure 5.9). In particular, the mutual regulation of Klf2 and Klf4 were recently shown to be an important module for maintaining the undifferentiated state of ESCs [51]. Utf1 and Myc are known to be key ESC transcription factors. The result that they are under the control of Oct4 and Klf4 underscores the importance of Klf4 in promoting self-renewal. Mtf2 has only recently been implied to inhibit differentiation by recruiting the polycomb group of transcription repressors [58]. This analysis indicates that Klf4 and Sox2 could synergistically activate Mtf2 in ESCs. The regulatory relationships for a number of genes involved in lineage specific differentiation are also identified. These include Gata6, Gata3, Sox17 and FoxA2. Inhibiting these lineage specific differentiation genes in ESCs is critical to maintain an undifferentiated state. Among the predicted network, there are a number of transcription repressors, including Ctpb2 and Rest. Ctpb2 is predicted to be activated by Oct4. Rest is predicted to be jointly regulated by Oct4 and Sox2. These results suggest that Oct4 and Sox2 could indirectly inhibit differentiation genes by activating transcription repressors such as Ctpb2 and Rest.

5.4 Discussions

New algorithms combining the strengths of both physical and influence approaches to identify genetic regulatory network are highly preferable. Interaction-Identifier integrates three piece of information together to inferring genetic regulatory interactions: a) mechanistic models of transcriptional factor binding and RNA transcription [18], b) prior knowledge of network components based on ChIP-chip data, c) time series expression data. Furthermore, Interaction-Identifier combines two methodologies together, kinetic modeling and correlation analysis. We further develop Network-Identifier based on Interaction-Identifier to reconstruct gene regulatory networks. Multiple temporal gene expression datasets are used as inputs to Network-Identifier.
Figure 5.9: The gene regulatory network identified by Network-Identifier. Yellow nodes represent regulators. Green nodes represent genes promoting self-renewal and pluripotency. Red nodes represent genes used for differentiation. Sharp and blunt arrows represent activation and repression effects, respectively. Red and green lines represent activation and repression activities with RNAi evidence, respectively. Blue and black lines denote regulatory relationships with ChIP-chip evidence.
ChIP-chip and RNAi data can also be utilized by Network-Identifier as independent validation datasets to further improve the predicted networks. Moreover, Network-Identifier has great flexibility in incorporating independent datasets other than ChIP-chip or RNAi data to reinforce the strength of validation.

In both methods, we choose to represent the expression level as continuous instead of using discretized expression levels. Previously, reverse engineering approaches have been developed to infer boolean network underlying changes in the gene expression level assuming that expression levels of different genes can be categorized into different states [112]. In reality, gene expression levels tend to be continuous rather than discrete. Furthermore, continuous signals have much greater capacity over discrete signals in implementing different control functions, such as signal transformation and transduction, precise feedback and feed forward and maintaining homeostasis [113]. An implicit assumption of using continuous concentrations of the chemical species (mRNA and protein) is that the stochastic fluctuations due to single molecules are ignored. In both prokaryotic and eukaryotic cells, noises in gene expression levels has been observed and suggested to be an evolvable trait, which possible plays a role in cellular phenotypic variation and cellular differentiation [82] [114] [115] [116]. Both stochasticity inherent in the biochemical process of gene expression (intrinsic noise) and fluctuations in other cellular components (extrinsic noise) contribute substantially to overall phenotypic variation [115]. The mRNA signals obtained were effectively averages of pooled populations of cells; where the influence of stochastic noise of single molecules on chemical concentration (mRNA and protein) were presumably effectively decreased.

Some assumptions are made in the methodological frameworks. First, the form of the interaction among the TFs and RNAP are assumed to be invariant under the multiple conditions from which the gene expression data are obtained. This assumption can be violated when the experimental conditions are dramatically different from each other, for example, different stress conditions. This assumption is better satisfied by using data from the biological process, for example, a developmental process. For this reason, we suggest using time course gene expression data rather than data generated from different experimental conditions. Even for time course data, the users should exercise caution, because the regulation factor can still change in some cir-
cumstances, such as when the cell goes through different phases of the cell cycle [117] [118]. The second assumption is that the transcriptional system is at equilibrium state in each time point when the gene expression is measured. This assumption is satisfied by all the time course microarray data. The third and the biggest assumption is that the statistical thermodynamic models derived and tested for prokaryotes can be applied to eukaryote systems. This is essentially ignoring a number of transcriptional regulatory mechanisms that eukaryotes and especially high level eukaryotes utilize, such as chromatin modification and long range regulation. As a first-order approximation, the Interaction-Identifier method is still useful to analyze the biophysical properties of the known TFs. Another point in favour of the validity of this method is that the absolute value of the model-derived gene expression level does not influence the correlation calculation. Only the pattern of change of the expression levels over time influence the correlation calculation. Many of the eukaryotic specific regulatory features, such as the distance between the enhancer and the promoter, are invariant for the target gene over the time course, and therefore such features should not affect the selection of the corrected model. Most important of all, Our methods, together with Segal et al [7] and Gertz et al’s attempts [39], have shown that statistical thermodynamic models are a reasonable route to capture the underlying relationship between regulatory sequence and gene expression in either prokaryotes and eukaryote systems. Future work that takes the molecular features and events into account will potentially provide us with a thorough understanding of combinatorial gene regulation.
With genome-wide distributions of TF-binding, epigenomic modifications and gene expression data available by high-throughput methods, the explosive growth of data urges the methodological developments that can achieve mechanistic understanding of transcription regulation. The thesis has focused on modeling different aspects of transcription regulatory mechanisms. To be more specific, we developed statistical mechanical models to learn the regulatory rules in transcription networks.

How transcription factors (TFs) interact with cis-regulatory sequences and interact with each other is a fundamental, but not well understood, aspect of gene regulation. We present a computational method to address this question, relying on the established biophysical principles. It takes into account all combinations and configurations of strong and weak binding sites to analyze large scale TF-DNA binding data to discover cooperative interactions among TFs, infer sequence rules of interaction and predict TF target genes. The distinctions between our model and other statistical approaches for analyzing cis-regulatory sequences include the utility of physical principles and the treatment of the DNA binding data as quantitative representation of binding strengths. Applying this method to the ChIP-seq data of 12 TFs in mouse embryonic stem cells (ESCs), we found that the strength of TF-DNA binding could be significantly modulated by cooperative interactions among TFs with adjacent binding sites. Our model showed that a novel Nanog motif could better explain the ChIP-seq data than previously published ones. A series of comparisons showed that our model has more predictive power than several state-of-the-art methods for cis-regulatory sequence analysis.

Not only the nucleotide composition but also likely the chemical modifications to the genomic sequences and the nearby histones could quantitatively affect TF-DNA binding. However, less clear is how the genome and the epigenome jointly encode and deliver the information to TFs. Such infor-
information instructs the TFs which parts of the genome they should interact with and the intensities of every interaction. We developed a model-based approach to systematically analyze the epigenomic functions in modulating transcription factor-DNA binding. We discovered TF-specific epigenomic motifs, which explained why some TFs appeared to have different DNA binding motifs derived from in vivo and in vitro experiments. The theoretical results suggested that the epigenome can modulate transcriptional noise and boost the cooperativity of weak TF binding sites. We observed that epigenomic boost of binding affinities in weak TF binding sites is a widespread regulatory mechanism in mouse ESCs. Moreover, using personal data, we identified strong associations between H3K4me2/H3K9ac and the degree of individual differences in NFκB binding in SNP-containing binding sites, suggesting the theoretical mechanism for epigenome to attenuate the TF binding differences on SNP-containing binding sites in two individuals have been implemented in human cells. Thus, this model presents a powerful approach to analyze the functions of epigenomic modifications.

To explore the effects of combinatorial control of TFs on gene expressions, we developed an analytical method to identify a statistical thermodynamic model that best describes the form of TF-TF interaction among a set of TFs for every target gene. We applied it to infer the combinatorial control of the key transcription factors in mouse ESCs. In particular, it identified that Oct4 and Sox2 help each other to stabilize onto DNA and attract the RNAP. This indicates that the DNA-bound Oct4 will be less in Sox2 knock-down ESCs, and vice versa. This is in line with the fact that the knock-down of either of the two transcription factors will decrease the expression levels of the mutual target genes and start the differentiation process [20]. We have subsequently categorized the mutual targets of Oct4 and Nanog according to the pattern of their combinatorial effect. Although Oct4 and Nanog often serve as activators for maintaining the expression of ESC specific genes, they also inhibit genes for lineage specific differentiation. Little is known about how Oct4 and Nanog switch their tasks between activators and repressors. The model does provide us a way to learn the possible changes of interaction forms of TFs for different target genes.

One ultimate goal of studying gene regulation is to identify gene regulatory networks. Based on the above work able to learn TF interaction forms, we developed a computational framework to infer regulatory relationships based
on biophysical principles of transcription regulation. Multiple temporal gene expression datasets are used as inputs, and ChIP-chip and RNAi data can also be utilized as independent validation datasets to further improve the predicted networks. Applying to the analysis of five datasets of differentiation of mouse ESCs, we identified a transcription network composed of 34 TF-TF interactions and 185 TF-target relationships. Independent data validated a statistically highly significant fraction of these regulatory relationships. Moreover, the framework has flexibility in incorporating independent datasets other than ChIP-chip or RNAi data to increase predictive accuracy.

In summary, statistical mechanical models based on depicting the interactions between TF-TF and TF-DNA to predict binding probability of a regulator protein, have shown its applicability to capture the underlying relationship among transcription factors, regulatory sequences and gene expressions. However, it should be recognized that there are still a number of simplifications made in the modeling of the biophysical properties of gene regulation. A number of molecular events are not included in the models, such as DNA looping: long range interaction of enhancer binding TFs and RNAP, and short-range repression: a “quenching” mechanism where a bound repressor molecule shuts off activator binding within a limited distance. Future work that takes these molecular features and events into account will potentially provide us with a thorough understanding of combinatorial gene regulation.
REFERENCES


87


91


