ANALYSIS OF COMBINATORIAL GENE REGULATION WITH THERMODYNAMIC MODELS

BY

CHIEH-CHUN CHEN

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Bioengineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2009

Urbana, Illinois

Adviser:

Professor Sheng Zhong

Abstract

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 and further to reconstruct gene regulatory networks.

In this thesis, we developed an analytical method, Interaction-Identifier, to identify a thermodynamic model that best describes the form of TF-TF interaction among a set of TFs for every target gene. Applying this approach to time-course microarray data in mouse embryonic stem cells, we have inferred five interaction patterns among three regulators: Oct4, Sox2 and Nanog on ten target genes. We further proposed a computational framework, Network-Identifier, utilizing Interaction-Identifier, to reconstruct gene regulatory networks. Applied to five datasets of differentiating embryonic stem cells, Network-Identifier identified a gene regulatory network among 87 transcription regulator genes. This network suggests that Oct4, Sox2 and Klf4 indirectly repress lineage specific differentiation genes by activating transcriptional repressors of Ctbp2, Rest and Mtf2.

Acknowledgments

This project would not have been possible without the support of many people. Many thanks to my adviser, Dr. Sheng Zhong, who offered valuable guidance and support. Thanks to my family, my significant other, Yen-Wei Huang, and numerous friends who endured this long process with me, always offering support and love.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

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 and further to reconstruct gene regulatory networks.

A number of analytical methods have been proposed to reconstruct gene regulatory networks from gene expression and protein-DNA binding data. Association rule mining [18] , Boolean Network [19], temporal models [16] [55], ARACNE [34] and Bayesian networks [26] [27] [31] 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 [42]. 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 [17] [49]. 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 [42]). A network reconstruction method based on quantities that closely represent the biophysical properties of TF-DNA binding, transcription activation and repression is still missing.

Thermodynamic models are based on the assumption that the level of gene expression is proportional to the equilibrium probability that RNAP binds to the promoter of interest; and these probabilities can be computed in a statistical mechanics framework.

In this thesis, we proposed a method, Interaction-Identifier, based on 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 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 [28] and chromatin immunoprecipitation coupled with microarray (ChIP-chip) data [10] [30] independently validated a statistically highly significant fraction of these regulatory relationships.

The remainder of this thesis is organized as follows. In Chapter 2, we make a thorough review of the related work on thermodynamic modelling. We then introduce Interaction-Identifier and Network-Identifier in Chapter 3. Based on the methods, we conduct experiments on synthetic data and mouse embryonic stem cell data. The performance analysis and the evaluation are described in Chapter 4. In Chapter 5, we present the concluding remarks and future work.

Chapter 2

Literature Review

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. Statistic 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 statistic thermodynamic concept [45] was first adopted on the study of molecular mechanism for gene regulation in Bacteriophage Lambda. Later it was further utilized on modeling TF-DNA and TF-RNA polymerase (RNAP) interactions in bacteria [7][8][12], 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.

More recently, Gertz et al [24] and Segal et al [41] applied thermodynamic principles to model expression data on large numbers of promoters or enhancers in Eukaryote system. Gertz et al showed that a thermodynamic model could successfully convey the relationship between promoter sequence and gene expression in Yeast system. Under a fixed time point in drosophila development, Segal et al demonstrated a thermodynamic model could predict the spatial expression patterns of segmentation genes in Drosophila [41]. Both works have shown promising results of thermodynamic models on gene regulation in Eukaryote system. In the following sections, we introduce how these two works unravelled the effects of *cis*-regulatory transcription control on gene expression (i.e. to find the relationships between sequence and gene expression) based on thermodynamic modeling.

2.1 Analysis of Combinatorial *cis*-regulation in Synthetic Promoter in Yeast

Although the fundamental theory of gene regulation has been studied and defined, the connections between regulatory information(*cis*-motifs and transcription factors) and gene expression profiles is still unclear [53]. Several studies developed *in silico* promoter models [22][54] demonstrated the associations between promoter modules and gene expressions. A ground-breaking study in Yeast [4] achieved the relatively high accuracy of prediction from conserved *cis*-motif logics to expression. This made it tempting to design synthetic promoters that allow refined and targeted modifications of promoter architecture. Through synthetic promoter engineering [2], *cis*-motif logic, including orientation, binding energy and position, could be clearly elucidated and served as control variables to study gene expression and gain insights of regulatory complexity.

Gertz et al applied a thermodynamic framework based on the assumption that gene expression is regulated by the interaction of TF-DNA and TF-TF stabilizing RNAP binding. It is specified by the changes in free energies of different TF binding events and the concentrations of TFs under different conditions. Thus, through a thermodynamic model, differential expression could be explained by changes of the TF concentraions and the events of TF-DNA binding and TF-TF intractions.

In order to learn how *cis*-regulatory mechanisms affecting gene expression in yeast, a strategy of combinatorial engineering was utilized to construct the synthetic promoter libraries [24]. All random combinations of three or four TF-BSs as building blocks were placed upstream of a core-promoter attached with yellow fluorescent protein. Then those synthetic promoters were integrated into the yeast genome. By quantifying florescent intensities, the level of gene expression could be observed. The results of applying a thermodynamic approach suggests modeling the biophysical principle of TF-DNA and TF-TF interactions can generally depict the expression driven by different combinations of TFBSs.

2.2 Predicting Spatial Expression Patterns from Sequence in Drosophila Segmentation

Drosophila melanogaster is a model organism for genetics research since its short life cycle, the relatively small genome and easily manipulation in laboratory. Moreover, since its embryos grow outside the body, it provides an excellent means of studying embryonic development in eukaryotes. Studying its notable segmentation network has helped accumulate most of our knowledge about the mechanisms of segmentation in arthropods [37]. Its well-characterized segmentation gene network involves a cascade of gene regulation. It consists of a four-tiered hierarchy of maternal and zygotic factors that define the antero-posterior body axis in a stepwise refinement of expression patterns. The maternal factors form gradients spanning the entire antero-posterior axis; they are translated into broad, non-periodic domains of zygotic gap gene expression and subsequently into periodic patterns of seven pair-rule and finally fourteen segmental stripes that prefigure the fourteen segments of the larva. Regulation within the network is highly combinatorial and, in the top tiers, almost transcriptional.

Segal et al developed a statistical thermodynamic framework to predict the expression of a target DNA sequence. The main idea is to sum over expression levels predicted by a logistic model under all possible configurations of TFs on a given sequence. First, they computed the probability of any configuration determined by the concentration of TFs and the binding affinities of the transcription factor binding sites in the configuration. The homo-cooperativity interactions between TFs were also considered as an essential factor to help TFs binding onto DNA. Second, for each configuration, they further used a logistic model to infer its ability of recruiting RNAP. The contribution of each TF on a configuration is assumed to be independent to the expression outcome, where activators contribute positively and repressors contribute negatively. With the unique saturation property of the logistic model, maximal or minimal transcription is achieved beyond a certain number of bound activators and repressors, respectively. Thus, the whole probability of RNAP binding is the weighted sum of RNAP binding probability for every configuration, where the weight of each configuration is the probability of the configuration. Under the thermodynamic principle, the degree of gene transcription is assumed to be proportional to the binding probability of the RNAP to the promoter.

With the spatial expression patterns for known key transcription factors and their binding-site preferences as inputs, the method was applied to model the process of transcriptional regulation and further to predict the spatial expression. The results show that expression patterns predicted by the model exhibit remarkable agreements with the observed pattern for most modules.

These successes of applying thermodynamic models onto Eukaryote system make it tempting to experiment novel methods for reconstructing gene regulatory networks based on more biophysically appropriate method.

Chapter 3

Proposed Method

Our proposed method consists of two components. The first component is Interaction-Identifier[15], which is developed to identify a thermodynamic model that best describes the form of TF-TF interaction among a set of TFs for every target gene. The second component is Network Identifier[14], which enables to further infer gene regulatory networks from multiple time course gene expression data, based on Interaction-Identifier. Next, we will describe each component in details in the following sections.

3.1 Interaction-Identifier

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 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 underlining TF interaction form of each target gene, which minimizes the difference between the model-derived expression profile and the observed expression data (Figure 3.1).

In the following, we first introduce the basic thermodynamic models for RNAP binding pioneered by Buchler et al [12] in Section 3.1.1. Next, we describe the kinetic model which derives the gene expression profile across times for each TF interaction form in Section 3.1.2. We then introduce the process of identifying the underlying TF interaction form for each target gene in Section 3.1.3.

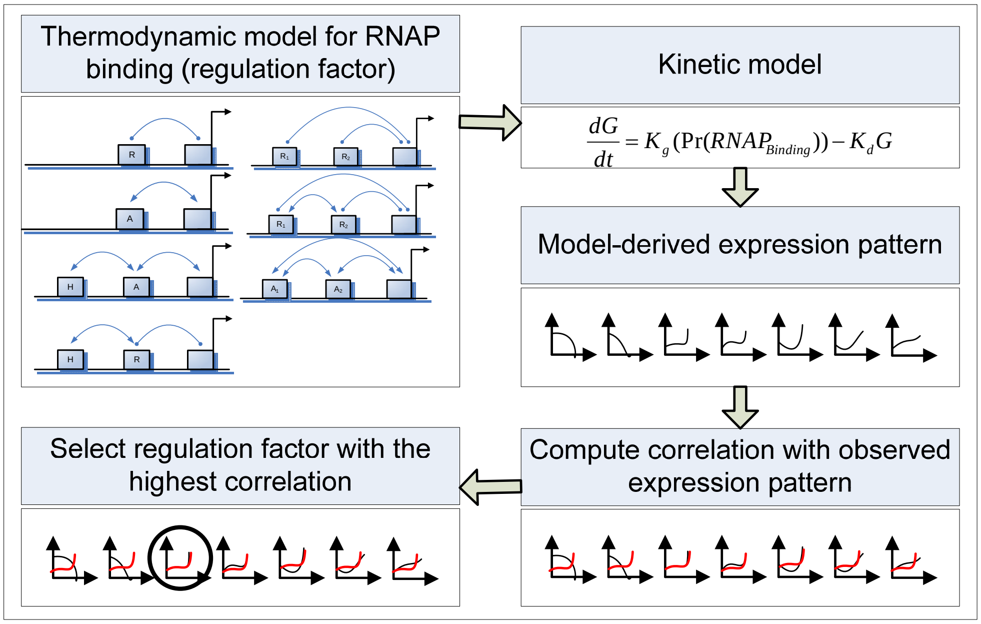

Thermodynamic model for RNAP binding (regulation factor)
R
A
H
R
R₁
R₂
A₁
A₂
Kinetic model
$\frac{dG}{dt} = K_g(\Pr(RNAP_{Binding})) - K_dG$
Model-derived expression pattern
Compute correlation with observed expression pattern
Select regulation factor with the highest correlation

Figure 3.1: Flowchart of the Interaction-Identifier method [14]

Table 3.1: The Boltzmann distribution for the two states of a TFBS

State	TF	Weight
free	0	1
attached	1	q_{TF}

3.1.1 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 [12] and Bintu et al[7][8] to integrate combinatorial signal at the level of *cis*-regulatory transcription control in bacteria through the thermodynamics of TF-DNA and TF-RNAP-DNA interactions. These interactions can be quantified by several tunable parameters based on different selections and placements of various protein-binding DNA sequences. In this section, this theoretical framework is briefed.

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 3.1).

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

$$P(TF_{binding}) = \frac{q_{TF}}{1 + q_{TF}}.$$

Table 3.2: The Boltzmann distribution of a promoter with one TF and one RNAP

State	TF	RNAP	Weight
1	0	0	1
2	0	1	q_p
3	1	0	q_{TF}
4	1	1	$\omega_{TFp} q_p q_{TF}$

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

$$P(RNAP_{bindig}) = \frac{q_p}{1 + q_p}.$$

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 promotor: (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 3.2).

The probability of the promoter of the target gene bound with a RNAP could be represented as

$$P(RNAP_{binding}) = \frac{q_p + \omega_{TFp} q_{TF} q_p}{1 + q_p + q_{TF} + \omega_{TFp}},$$

where

$$\omega_{TFp} = \begin{cases} 1 & \text{no iteraction} \\ 10 - 100 & \text{activation} \\ 0 & \text{repression} \end{cases}$$

Different settings of ω reflect different roles a TF could play. If ω is set to 1, it represents that there is no interaction between the RNAP and the TF. They bind independently to the promoter. If ω is set to 10-100, it represents that the TF helps recruit the RNAP binding to the promoter. The larger ω is, the larger the synergism is. If ω 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 3.2, model 2 and model 1, respectively).

- 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 3.3).

The probability of RNAP binding to the promoter could be denoted as

Table 3.3: The Boltzmann distribution of a promoter with its RNAP and two TFs

(TF_1, TF_2)	(0, 0)	(1, 0)	(0, 1)	(1, 1)
RNAP				
0	1	q_{TF1}	q_{TF2}	$\omega_{TF1TF2}q_{TF1}q_{TF2}$
1	q_p	$\omega_{TF1p}q_pq_{TF1}$	$\omega_{TF2p}q_pq_{TF2}$	$(\omega_{TF1p}+\omega_{TF2p})\omega_{TF1TF2}q_{TF1}q_{TF2}q_p$

$$P(RNAP_{binding}) = \frac{\sum_j \sum_k P(1, j, k)}{\sum_{i,j,k\in\{0,1\}} P(i, j, k)},$$

where $P(i,\ j,\ k) = P(RNAP = i,\ TF_1 = j,\ TF_2 = k)$.

The parameters ω 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_{TF1TF2} is used to simulate the interaction between the two TFs. A large w_{TF1TF2} (10-100) represents that the two TFs stabilize each other onto the promoter. If the two TFs have no interaction, w_{TF1TF2} should be set to 1. If the two TFs compete for the binding, w_{TF1TF2} should be set to 0 or close to 0. The other two parameters, w_{TF_1p} and w_{TF_2p}, represent the interaction between each TF and RNAP, respectively. They can be set to reflect different interactions similar to w_{TF1TF2}. By adjusting the parameters w_{TF_1p} , w_{TF_2p} and w_{TF1TF2}, we can obtain an analytical form for the probability of RNAP binding under different forms of interactions among RNAP and the two TFs. Figure 3.2 summarizes 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 $[TF-DNA]$ represent the cellular concentration of the promoter bound by the TF. The binding process can be denoted as

$$[TF] + [DNA] \rightarrow [TF - DNA]$$

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

$$P(TF_{binding}) = \frac{[TF - DNA]}{[DNA] + [TF - DNA]}$$

At equilibrium state, the concentrations of the substrates could be described

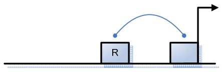
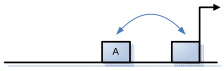
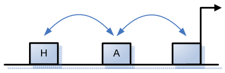
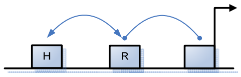
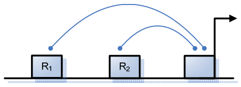
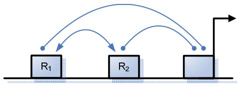
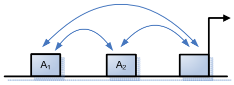
Model / Promoter state	Parameter
1. Simple Repressor 	W_{RP}=0
2. Simple Activator 	W_{AP}=10~100
3. Activator recruited by a helper (H) 	W_{AP}=10~100, W_{AH}=10~100, W_{HP}=1
4. Repressor recruited by a helper (H) 	W_{RP}=0, W_{HP}=1, W_{RH}=10~100
5. Dual repressors 	W_{R1P}=0, W_{R2P}=0, W_{R1R2}=1
6. Dual repressors interacting 	W_{R1P}=0, W_{R2P}=0, W_{R1R2}=10~100
7. Dual activators interacting 	W_{A1P}=10~100, W_{A2P}=10~100, W_{A1A2}=10~100

Figure 3.2: Forms of TF-RNAP interactions and their corresponding parameters for modeling the probability of RNAP binding. A_1 and A_2 are activators. R_1 and R_2 are repressors. P represents RNAP. The line with a dot at the end represents an repression effect; the line with an arrow at the end indicates either cooperation between two TFs or activation of a gene by a TF.

using the Hill equation

$$P(TF_{binding}) = \frac{[TF]^H}{[TF]^H + [K_{TF}]^H} = \frac{(\frac{[TF]}{K_{TF}})^H}{\frac{[TF]}{K_{TF}}^H + 1}$$

, where $[TF]$ is the cellular concentration of the activated TF targeted by this site, K_{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_{TF}, the ratio of the probabilities of the promoter in the bound and free states,

$$P(TF_{binding}) = \frac{(\frac{[TF]}{K_{TF}})^H}{\frac{[TF]}{K_{TF}}^H + 1} = \frac{q_{TF}}{q_{TF} + 1}.$$

Thus, we can obtain

$$q_{TF} = (\frac{[TF]}{K_{TF}})^H.$$

We use the unit of $[TF]$ and K_{TF} as the number of TFs per cell. There have been a few efforts to estimate K_{TF} from empirical data [48]. In this study, we assume at each time point in the time course, $[TF]$ is linearly related to the expression level of the TF, as did in earlier module network studies [42]. It follows that $[TF]$ peaks at the same time as its gene expression peaks. We further assume q_{TF} is maximized at the maximum $[TF]$ (see sensitivity analysis in Section 4.1.3 for further discussion on this assumption). We adopt the value 1/20 for q_p from [12][7][8].

3.1.2 Kinetic Model

With the above 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 [15].

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 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 [15].

$$\frac{dG}{dt} = K_g(P(RNAP_{binding})) - K_d(\frac{G}{G_{max}}),$$

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 [29]. Following [32] [11], we assume that the rate of degradation around 1/6 of the maximum transcription rate. Therefore, we use Kg =10 counts per minute and Kd =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

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

where

$$\alpha = G_{max}\frac{K_g}{K_d}.$$

3.1.3 Identification of the Underlining TF Interaction Forms

By combining the thermodynamics models and the kinetic model, we are able to derive the expression profiles from the interesting models shown in Figure 3.2. With measured time course gene expression data from microarray experiments, we compute the Pearson correlation coefficient between the observed expression pattern and the model-derived expression patterns. Since different combination 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 3.1). 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.

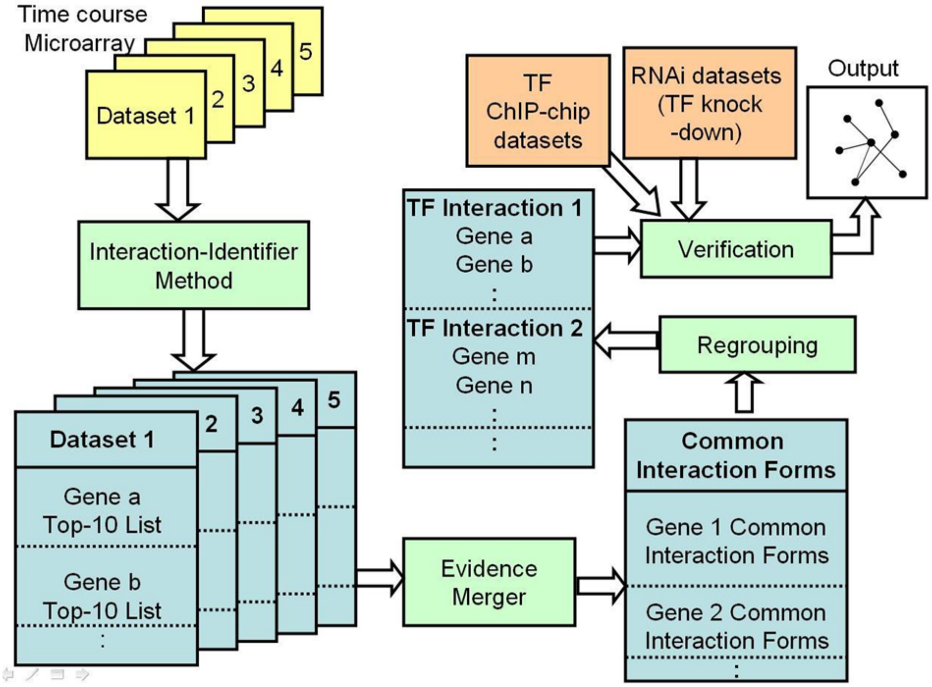


Figure 3.3: Flowchart of the Network-Identifier algorithm [14].

3.2 Network-Identifier

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 thermodynamic model that best describes the TF-TF and TF-RNAP interaction for each target gene, we further develop a computational framework, Network-Identifier [14], for inferring gene regulatory networks from multiple time course gene expression data.

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 3.1, [15]), 2) Evidence merger and 3) Verification component, shown in Figure 3.3. In the following section, we will describe each component in details.

3.2.1 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, and the five interaction forms between any two TFs (Figure 3.2). For each gene, Network-Identifier evaluates the fitness of each interaction form with Interaction-Identifier (See Section 3.1) 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.

3.2.2 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.

3.2.3 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 (RANi) 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 [30] or when some of the independent data are not available.

Chapter 4

Results

In this chapter, 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). We infer five interaction patterns among three regulators: Oct4, Sox2 and Nanog on ten target genes. We further apply Network-Identifier onto five datasets of mouse differentiating ESCs and identify a gene regulatory network among 87 transcription regulator genes.

4.1 Simulation Study on Interaction-Identifier

4.1.1 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 4.1). These regulatory patterns are: 1. a target gene is activated by one TF (Model 2 in Figure 3.2); 2. RNAP is blocked by a TF (repressor), and this TF is stabilized to DNA by a helper TF (Model 4 in Figure 3.2); 3 a target gene is regulated by two interacting activators (Model 7 in Figure 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 *realTFExp* using equations of the format or its variants: $E_A = a_A + b_A logT + \epsilon$, where a_A and b_A are background gene expression index and coefficient describing changes of expression index with time T. The ϵ 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 ϵ. 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 4.1 and derive the expression pattern of the target gene (*realTargetExp*) according to the 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 3.2. The model derived expression patterns are termed *modelTargetExp*. For each model, *obsTargetExp* is compared to *modelTargetExp* in terms of Pearson correlation.

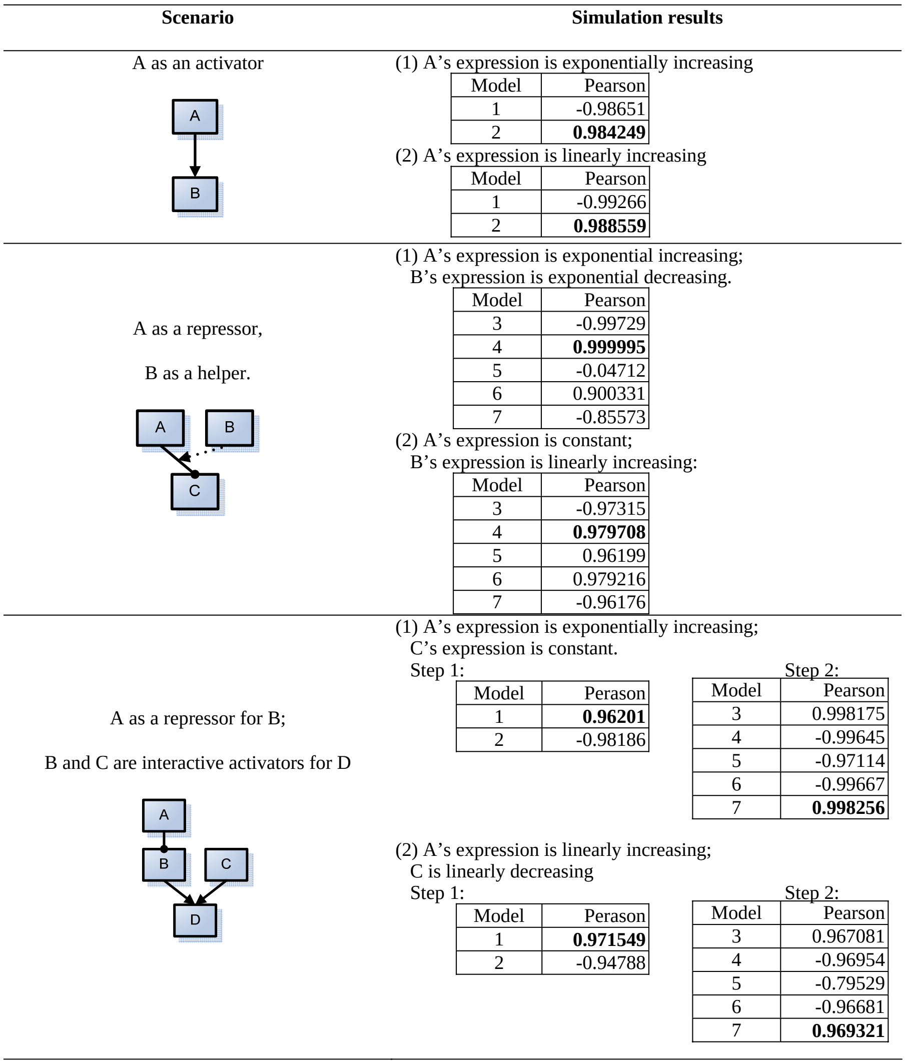
The parameters we use in the study are followed the literatures, where $K_g =$ 10 counts per minute, $K_d = 10/6$ counts per minute and $q_p = 1/20$. We further assume that $K_{TF} =$ the maximum $[TF]$ and $H = 2$.

4.1.2 Simulation Data Analysis

We use three regulatory patterns to test our new algorithm. Under the first regulatory pattern, two simulations are conducted. First, TFs expression increases linearly over time. $realTFexp = 500 + 500T$, where $T = 2, 4, 8, 16, 32, 64$ and 128. In the second simulation, TFs expression increases exponentially over time. $realTFExp = 500 + 200logT$, 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 (Model 1 in Figure 3.2) or activation (Model 2 in Figure 3.2). In both simulations our method correctly picked our Model 2 (Row 1, Figure 4.1). 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 4.1). 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 TFs A and B (Row 3, Figure 4.1), 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 4.1).

4.1.3 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 4.1) is chosen to perform the sensitivity analysis. We vary K_{TF}, K_g, K_d and q_p in very wide ranges, for example an 10000 fold range for K_{TF}, and rerun our algorithm. Results in Table 4.1 shows that the method can robustly identify the correct regulatory model even if the parameters are off-set by 100 fold. The only exceptions are the cases where the synthesis rates of mRNA were set to be too slow V 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.



Scenario — **Simulation results**

A as an activator

(1) A’s expression is exponentially increasing

Model	Pearson
1	-0.98651
2	**0.984249**

(2) A’s expression is linearly increasing

Model	Pearson
1	-0.99266
2	**0.988559**

A as a repressor,

B as a helper.

(1) A’s expression is exponential increasing;
B’s expression is exponential decreasing.

Model	Pearson
3	-0.99729
4	**0.999995**
5	-0.04712
6	0.900331
7	-0.85573

(2) A’s expression is constant;
B’s expression is linearly increasing:

Model	Pearson
3	-0.97315
4	**0.979708**
5	0.96199
6	0.979216
7	-0.96176

A as a repressor for B;

B and C are interactive activators for D

(1) A’s expression is exponentially increasing;
C’s expression is constant.

Step 1:

Model	Perason
1	**0.96201**
2	-0.98186

Step 2:

Model	Pearson
3	0.998175
4	-0.99645
5	-0.97114
6	-0.99667
7	**0.998256**

(2) A’s expression is linearly increasing;
C is linearly decreasing

Step 1:

Model	Perason
1	**0.971549**
2	-0.94788

Step 2:

Model	Pearson
3	0.967081
4	-0.96954
5	-0.79529
6	-0.96681
7	**0.969321**

Figure 4.1: Results from synthetic data using the Interaction-Identifier algorithm. [14] The concentration of A was simulated using either a linear function: $[TF] = 500 + 500T$ or an exponential function:$[TF] = 500 + 200logT$, where T represents the time.

Table 4.1: Sensitivity test for K_{TF} , K_g , q_p, K_d and H [14]. Numbers in bold represent the highest correlations under each parameter set. The results indicate that the correct model can be identified even with drastic variation in parameters used in the algorithm.

Model	K_{TF}	Pearson	Kg	Pearson	Kd	Pearson	qp	Pearson	H	Pearson
3	0.01	0.9500	1/60	0.9671	60/24	0.9671	1/35	0.9677	1	0.9711
	0.1	0.9505	1/6	0.9671	60/30	0.9671	0.05	0.9671	2	0.9671
	1	0.9671	10	0.9671	60/36	0.9671	0.10	0.9662	3	0.9637
	10	0.9571	600	0.9671	60/42	0.9671	1	0.9642	4	0.9641
	100	0.9562	1000	0.9671	60/48	0.9671	10	0.9639	5	0.9688
4	0.01	-0.9514	1/60	-0.9697	60/4	-0.9695	1/35	-0.9695	1	-0.9715
	0.1	-0.9517	1/6	-0.9695	60/30	-0.9695	0.05	-0.9695	2	-0.9695
	1	-0.9695	10	-0.9695	60/36	-0.9695	0.10	-0.9697	3	-0.9693
	10	-0.9567	600	-0.9695	60/42	-0.9695	1	-0.9715	4	-0.9719
	100	-0.9562	1000	-0.9695	60/48	-0.9695	10	-0.9719	5	-0.9748
5	0.01	-0.8822	1/60	-0.7953	60/24	-0.7953	1/35	-0.7953	1	-0.9242
	0.1	-0.9720	1/6	-0.7953	60/30	-0.7953	0.05	-0.7953	2	-0.7953
	1	-0.7953	10	-0.7952	60/36	-0.7953	0.10	-0.7952	3	-0.5936
	10	-0.6160	600	-0.7952	60/42	-0.7953	1	-0.7936	4	-0.4017
	100	-0.6125	1000	-0.7952	60/48	-0.7953	10	-0.7898	5	-0.2617
6	0.01	0	1/60	-0.9668	60/24	-0.9668	1/35	-0.9668	1	-0.9678
	0.1	-0.9720	1/6	-0.9668	60/30	-0.9668	0.05	-0.9668	2	-0.9668
	1	-0.9668	10	-0.9668	60/36	-0.9668	0.10	-0.9667	3	-0.9598
	10	-0.6654	600	-0.9668	60/42	-0.9668	1	-0.9654	4	-0.9354
	100	-0.6138	1000	-0.9668	60/48	-0.9668	10	-0.9579	5	-0.8747
7	0.01	0.9608	1/60	0.96931	60/24	0.9693	1/35	0.9696	1	0.9716
	0.1	0.9616	1/6	0.96932	60/30	0.9693	0.05	0.9693	2	0.9693
	1	0.9693	10	0.96932	60/36	0.9693	0.10	0.9690	3	0.9690
	10	0.7092	600	0.96932	60/42	0.9693	1	0.9686	4	0.9739
	100	0.6143	1000	0.96932	60/48	0.9693	10	0.9686	5	0.9801

4.2 Applications on Mouse Embryonic Stem Cells

In this section, we first introduce the background knowledge for embryonic stem cells (ESCs). Next, we present the Interaction forms of important ESCs target genes identified by Interaction-Identifier. Finally, we present the gene regulatory network found by Network-Identifier.

4.2.1 Background Knowledge for 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: [28] [21], human: [51]). 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" [36]) 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 [1] [6] [13] [10] [33] [10] [25] [46] [5]. 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.

4.2.2 Interaction Models for Oct4, Sox2 and Nanog in mouse ESCs

Dataset

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 embroyotic stem cell (ESC): Oct4 and Nanog [33].

Time course microarray data have been generated for retinoid acid induced differentiation of mouse embryonic stem cells [28]. Genes that are jointly regulated by Oct4 and Nanog have been reliably identified [33]. 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.

For each interaction form in Figure 3.2, we use the differential equation to derive the steady state level of mRNA expression level using the estimated $[TF]$ and 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.

Five Interaction Models Among Regulators: Oct4, Sox2 and Nanog

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 4.2).

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 and emoes, are regulated under model 7 (Figure 4.3 (a)), where Oct4 and Nanog synergistic activators. REST is regulated under model 3, with one TF as an activator and the other as a helper (Figure 4.3 (b)). Atbf1 is regulated under model 5 where Oct4 and Nanog are independent repressors (Figure 4.4 (a)). Foxc1 is regulated under model 4 where Nanog is a helper and Oct4 is a repressor (Figure 4.4 (b)). These results suggest that Atbf1 and Foxc1 are probably involved in lineage differentiation and therefore need to be repressed by key transcription factors in ESC. Interestingly, Foxc1 has been shown to be involved in ocular development [35] and Abf1 mRNA is found to be abundant in prostate [50]. 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.

4.2.3 Gene Regulatory Network in Mouse ESCs

Dataset

We employ five time series microarray datasets of mouse ESCs in this study, including a dataset for retinoid acid induced differentiation [28] and four datasets for spontaneous differentiation of four ESC lines (three lines from [43]; 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

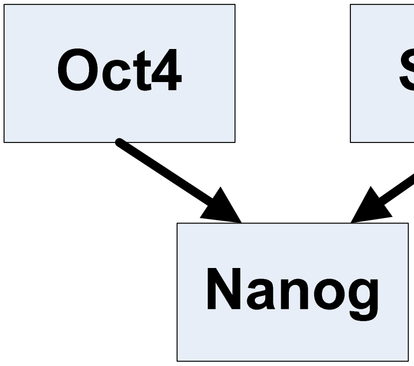


Figure 4.2: The identified regulatory network among Oct4, Sox2 and Nanog [15]. Apply the Interaction-identification algorithm to the expression data of Oct4, Sox2, and Nanog from time course microarray data.

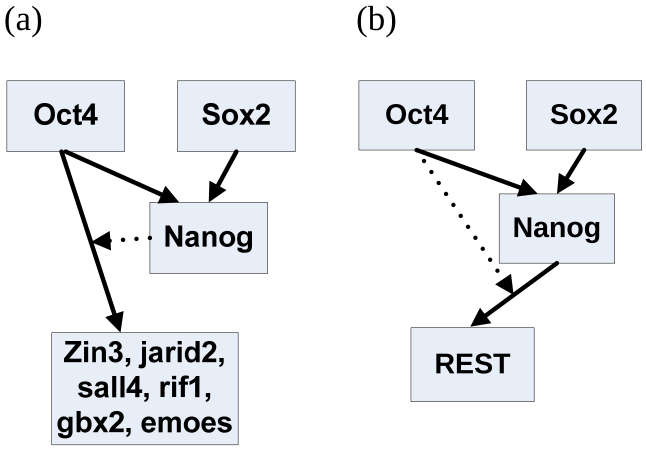


Figure 4.3: Activation regulations of Oct4 and Nanog on target genes identified using Interaction-identification algorithm. [15]. The directed arrows represent activation and the dotted line represents the function of the helper. The relationship between Nanog and Oct 4 with these target genes follow the model 3 in Figure 3.2.

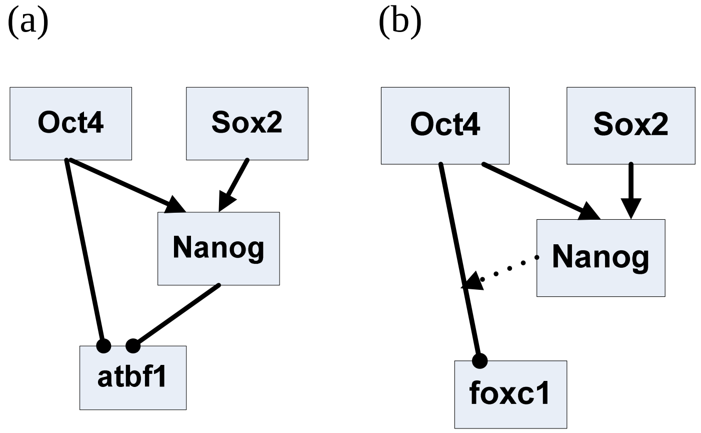


Figure 4.4: Repression regulations of Oct4 and Nanog on target genes identified using Interaction-identification algorithm. [15]. (a) model 5 (Figure 3.2) (b) model 4 (Figure 3.2), where dotted line represents the function of the helper, a line with an arrow in the end represents the effect of activator; a line with a solid dot in the end represents the effect of repressor.

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.

Gene Regulatory Network

Interaction-Identifier is first applied to each time course microarray dataset. A list of common TF Interaction forms across datasets is then generated 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 4.2). RNA knock-out experiments are performed for all the six regulators [28] [30]. Nine out of twelve target gene groups involving these six regulators are enriched with RNAi verified regulatory relationships (Table 4.3). 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

Table 4.2: Validation by ChIP-chip data

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

Table 4.3: Validation by RNA interference data

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

data. We use Cytoscape [44] to display the final reported regulatory relationships (Figure 4.5).

87 regulators and target genes are reported in the ESC transcription network (Figure 4.5). In particular, the mutual regulation of Klf2 and Klf4 were recently shown to be an important module for maintaining the undifferentiated state of ESCs [30]. 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 [56]. 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.

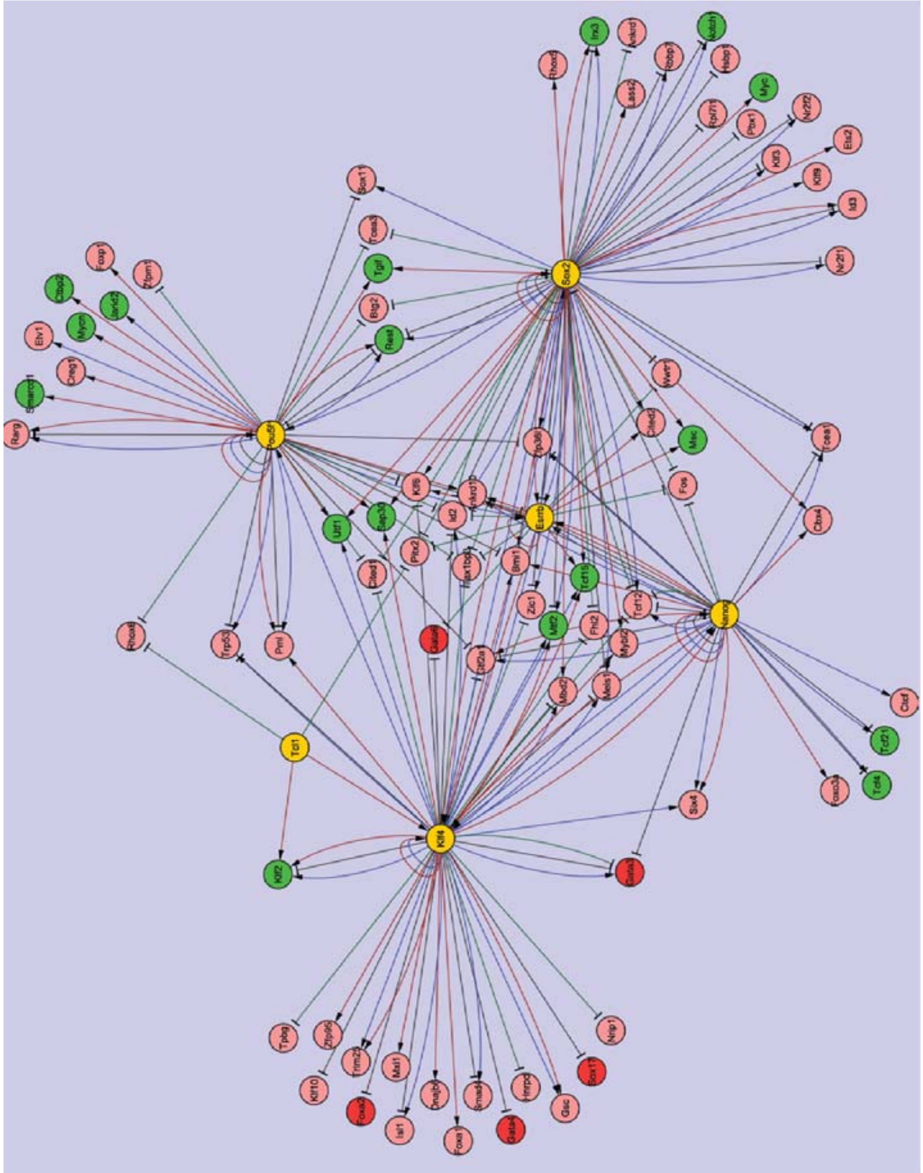


Figure 4.5: The gene regulatory network identified by Network-Identifier [14]. Yellow nodes represent regulators. Green nodes represent genes promoting self-renewal and pluripotency. Red nodes represent genes used for differentiation. Sharp and blunt arrows represent activation and repression effects, 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.

Chapter 5

Discussions and Conclusions

5.1 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 [12], b) prior knowledge of network components based on ChIP-chip data, c) time series expression data. Furthermore, Interaction-Identifier combines two methodologies together, kinetic modelling and correlation analysis. We further develop Network-Identifier based on Interaction-Identifier to reconstruct gene regulatory networks.

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 [**?**]. In reality, gene expression levels tend to be continuous rather than discrete. Furthermore, continuous signals have much great capacity over discrete signals in implementing different control functions, such as signal transformation and transduction, precise feedback and feed forward and maintaining homeostasis [52]. 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 [20] [39] [38] [9]. 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 [38]. 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 circumstances, such as when the cell goes through different phases of the cell cycle [47] [3]. 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 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 [41] and Gertz et al's attempts [24], have shown that thermodynamic models are a reasonable route to capture the underlying relationship between regulatory sequence and gene expression in either prokaryotes and eukaryote systems.

5.2 Conclusions

Thermodynamic models based on depicting the interactions between TF-TF and TF-DNA to predict RNAP binding probability, have shown its applicability to capture the underlying relationship between regulatory sequence and gene expression.

Interaction-Identifier is developed to identify interaction forms of TFs for target genes. We apply it to infer the combinatorial control of the key transcription factors in mouse ESCs. In particular, Interaction-Identifier method identifies 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 knockdown 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 [28]. 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. Interaction-Identifier does provide us a way to learn the possible changes of interaction forms of TFs for different target genes.

Along this line, Network-Identifier is proposed to reconstruct transcription network based on biophysical models of transcription regulation. Multiple temporal gene expression datasets are used as inputs to Network-Identifier. 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.

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 model. These include: 1) the interactions of more than two TFs, 2) long range interaction of enhancer binding TFs and RNAP, 3) DNA methylation and 4) chromatin structure and state. Future work that takes these molecular features and events into account will potentially provide us with a thorough understanding of combinatorial gene regulation.

References

[1] M. J. Abeyta, A. T. Clark, R. T. Rodriguez, M. S. Bodnar, R. A. Pera, and M. T. Firpo. Unique gene expression signatures of independently-derived human embryonic stem cell lines. *Hum Mol Genet*, 13(6):601–608, Mar 2004.

[2] E. Andrianantoandro, S. Basu, D. K. Karig, and R. Weiss. Synthetic biology: new engineering rules for an emerging discipline. *Mol Syst Biol*, 2:2006–2006, 2006.

[3] N. Banerjee and M. Q. Zhang. Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res*, 31(23):7024–7031, Dec 2003.

[4] M. A. Beer and S. Tavazoie. Predicting gene expression from sequence. *Cell*, 117(2):185–198, Apr 2004.

[5] B. E. Bernstein, T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, R. Jaenisch, A. Wagschal, R. Feil, S. L. Schreiber, and E. S. Lander. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2):315–326, Apr 2006.

[6] B. Bhattacharya, T. Miura, R. Brandenberger, J. Mejido, Y. Luo, A. X. Yang, B. H. Joshi, I. Ginis, R. S. Thies, M. Amit, I. Lyons, B. G. Condie, J. Itskovitz-Eldor, M. S. Rao, and R. K. Puri. Gene expression in human embryonic stem cell lines: unique molecular signature. *Blood*, 103(8):2956–2964, Apr 2004.

[7] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips. Transcriptional regulation by the numbers: applications. *Current Opinion in Genetics & Development*, 15(2):125 – 135, 2005. Chromosomes and expression mechanisms.

[8] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips. Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev*, 15(2):116–124, Apr 2005.

[9] W. J. Blake, M. KAErn, C. R. Cantor, and J. J. Collins. Noise in eukaryotic gene expression. *Nature*, 422(6932):633–637, Apr 2003.

[10] L. A. Boyer, T. I. Lee, M. F. Cole, S. E. Johnstone, S. S. Levine, J. P. Zucker, M. G. Guenther, R. M. Kumar, H. L. Murray, R. G. Jenner, D. K. Gifford, D. A. Melton, R. Jaenisch, and R. A. Young. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, 122(6):947–956, Sep 2005.

[11] O. Brandman, J. E. Ferrell, R. Li, and T. Meyer. Interlinked fast and slow positive feedback loops drive reliable cell decisions. *Science*, 310(5747):496–498, Oct 2005.

[12] N. E. Buchler, U. Gerland, and T. Hwa. On schemes of combinatorial transcription logic. *Proc Natl Acad Sci U S A*, 100(9):5136–5141, Apr 2003.

[13] R. Catena, C. Tiveron, A. Ronchi, S. Porta, A. Ferri, L. Tatangelo, M. Cavallaro, R. Favaro, S. Ottolenghi, R. Reinbold, H. Schöler, and S. K. Nicolis. Conserved pou binding dna sites in the sox2 upstream enhancer regulate gene expression in embryonic and neural stem cells. *J Biol Chem*, 279(40):41846–41857, Oct 2004.

[14] C. C. Chen and S. Zhong. Inferring gene regulatory networks by thermodynamic modeling. *BMC Genomics*, 9 Suppl 2, 2008.

[15] C. C. Chen, X. G. Zhu, and S. Zhong. Selection of thermodynamic models for combinatorial control of multiple transcription factors in early differentiation of embryonic stem cells. *BMC Genomics*, 9 Suppl 1, 2008.

[16] C. T. Chen, J. C. Wang, and B. A. Cohen. The strength of selection on ultraconserved elements in the human genome. *Am J Hum Genet*, 80(4):692–704, Apr 2007.

[17] K. H. Cho, J. R. Kim, S. Baek, H. S. Choi, and S. M. Choo. Inferring biomolecular regulatory networks from phase portraits of time-series expression profiles. *FEBS Lett*, 580(14):3511–3518, Jun 2006.

[18] C. Creighton and S. Hanash. Mining gene expression databases for association rules. *Bioinformatics*, 19(1):79–86, Jan 2003.

[19] M. I. Davidich and S. Bornholdt. Boolean network model predicts cell cycle sequence of fission yeast. *PLoS One*, 3(2), 2008.

[20] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, Aug 2002.

[21] M. J. Evans and M. H. Kaufman. Establishment in culture of pluripotential cells from mouse embryos. *Nature*, 292(5819):154–156, Jul 1981.

[22] S. Fessele, H. Maier, C. Zischek, P. J. Nelson, and T. Werner. Regulatory context is a crucial part of gene function. *Trends in Genetics*, 18(2):60 – 63, 2002.

[23] T. S. Gardner and J. J. Faith. Reverse-engineering transcription control networks. *Physics of Life Reviews*, 2(1):65 – 88, 2005.

[24] J. Gertz, E. D. Siggia, and B. A. Cohen. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature*, 457(7226):215–218, Jan. 2009.

[25] M. Golan-Mashiach, J. E. Dazard, S. Gerecht-Nir, N. Amariglio, T. Fisher, J. Jacob-Hirsch, B. Bielorai, S. Osenberg, O. Barad, G. Getz, A. Toren, G. Rechavi, J. Itskovitz-Eldor, E. Domany, and D. Givol. Design principle of gene expression used by human stem cells: implication for pluripotency. *FASEB J*, 19(1):147–149, Jan 2005.

[26] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, Jul 2000.

[27] D. Husmeier. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic bayesian networks. *Bioinformatics*, 19(17):2271–2282, Nov 2003.

[28] N. Ivanova, R. Dobrin, R. Lu, I. Kotenko, J. Levorse, C. DeCoste, X. Schafer, Y. Lun, and I. R. Lemischka. Dissecting self-renewal in stem cells with rna interference. *Nature*, 442(7102):533–538, Aug 2006.

[29] V. Iyer and K. Struhl. Absolute mrna levels and transcriptional initiation rates in saccharomyces cerevisiae. *Proc Natl Acad Sci U S A*, 93(11):5208–5212, May 1996.

[30] J. Jiang, Y. S. Chan, Y. H. Loh, J. Cai, G. Q. Tong, C. A. Lim, P. Robson, S. Zhong, and H. H. Ng. A core klf circuitry regulates self-renewal of embryonic stem cells. *Nat Cell Biol*, 10(3):353–360, Mar 2008.

[31] S. Y. Kim, S. Imoto, and S. Miyano. Inferring gene networks from time series microarray data using dynamic bayesian networks. *Brief Bioinform*, 4(3):228–235, Sep 2003.

[32] J. Lewis. Autoinhibition with transcriptional delay: a simple mechanism for the zebrafish somitogenesis oscillator. *Curr Biol*, 13(16):1398–1408, Aug 2003.

[33] Y. H. Loh, Q. Wu, J. L. Chew, V. B. Vega, W. Zhang, X. Chen, G. Bourque, J. George, B. Leong, J. Liu, K. Y. Wong, K. W. Sung, C. W. Lee, X. D. Zhao, K. P. Chiu, L. Lipovich, V. A. Kuznetsov, P. Robson, L. W. Stanton, C. L. Wei, Y. Ruan, B. Lim, and H. H. Ng. The oct4 and nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet*, 38(4):431–440, Apr 2006.

[34] A. A. Margolin, K. Wang, W. K. Lim, M. Kustagi, I. Nemenman, and A. Califano. Reverse engineering cellular networks. *Nat Protoc*, 1(2):662–671, 2006.

[35] D. Y. Nishimura, C. C. Searby, W. L. Alward, D. Walton, J. E. Craig, D. A. Mackey, K. Kawase, A. B. Kanis, S. R. Patil, E. M. Stone, and V. C. Sheffield. A spectrum of foxc1 mutations suggests gene dosage as a mechanism for developmental defects of the anterior chamber of the eye. *Am J Hum Genet*, 68(2):364–372, Feb 2001.

[36] S. Pease, P. Braghetta, D. Gearing, D. Grail, and R. L. Williams. Isolation of embryonic stem (es) cells in media supplemented with recombinant leukemia inhibitory factor (lif). *Dev Biol*, 141(2):344–352, Oct 1990.

[37] A. D. Peel, A. D. Chipman, and M. Akam. Arthropod segmentation: beyond the drosophila paradigm. *Nat Rev Genet*, 6(12):905–916, Dec. 2005.

[38] J. M. Raser and E. K. O'Shea. Noise in gene expression: origins, consequences, and control. *Science*, 309(5743):2010–2013, Sep 2005.

[39] N. Rosenfeld, J. W. Young, U. Alon, P. S. Swain, and M. B. Elowitz. Gene regulation at the single-cell level. *Science*, 307(5717):1962–1965, Mar 2005.

[40] W. A. Schmitt, R. M. Raab, and G. Stephanopoulos. Elucidation of gene interaction networks through time-lagged correlation analysis of transcriptional data. *Genome Res*, 14(8):1654–1663, Aug 2004.

[41] E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul. Predicting expression patterns from regulatory sequence in drosophila segmentation. *Nature*, 451(7178):535–540, Jan 2008.

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

[43] K. H. Sene, C. J. Porter, G. Palidwor, C. Perez-Iratxeta, E. M. Muro, P. A. Campbell, M. A. Rudnicki, and M. A. Andrade-Navarro. Gene function in early mouse embryonic stem cell differentiation. *BMC Genomics*, 8:85–85, 2007.

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

[45] M. A. Shea and G. K. Ackers. The or control system of bacteriophage lambda. a physical-chemical model for gene regulation. *J Mol Biol*, 181(2):211–230, Jan 1985.

[46] H. Skottman, M. Mikkola, K. Lundin, C. Olsson, A. M. Strömberg, T. Tuuri, T. Otonkoski, O. Hovatta, and R. Lahesmaa. Gene expression signatures of seven individual human embryonic stem cell lines. *Stem Cells*, 23(9):1343–1356, Oct 2005.

[47] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9(12):3273–3297, Dec 1998.

[48] G. D. Stormo. Dna binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, Jan 2000.

[49] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255, Oct 2003.

[50] X. Sun, H. F. Frierson, C. Chen, C. Li, Q. Ran, K. B. Otto, B. L. Cantarel, B. M. Cantarel, R. L. Vessella, A. C. Gao, J. Petros, Y. Miura, J. W. Simons, and J. T. Dong. Frequent somatic mutations of the transcription factor atbf1 in human prostate cancer. *Nat Genet*, 37(4):407–412, Apr 2005.

[51] J. A. Thomson, J. Itskovitz-Eldor, S. S. Shapiro, M. A. Waknitz, J. J. Swiergiel, V. S. Marshall, and J. M. Jones. Embryonic stem cell lines derived from human blastocysts. *Science*, 282(5391):1145–1147, Nov 1998.

[52] J. J. Tyson, K. Chen, and B. Novak. Network dynamics and cell physiology. *Nat Rev Mol Cell Biol*, 2(12):908–916, Dec 2001.

[53] M. Venter. Synthetic promoters: genetic control through cis engineering. *Trends in Plant Science*, 12(3):118 – 124, 2007.

[54] T. WERNER, S. FESSELE, H. MAIER, and P. J. NELSON. Computer modeling of promoter organization as a tool to study transcriptional coregulation. *FASEB J.*, 17(10):1228–1237, 2003.

[55] F. X. Wu. Stability analysis of genetic regulatory networks with multiple time delays. *Conf Proc IEEE Eng Med Biol Soc*, 2007:1387–1390, 2007.

[56] Q. Zhou, H. Chipperfield, D. A. Melton, and W. H. Wong. A gene regulatory network in mouse embryonic stem cells. *Proc Natl Acad Sci U S A*, 104(42):16438–16443, Oct 2007.