Modeling Gene Regulatory Networks Using a State-Space Model with Time Delays

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Chu Shin Koh

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building
110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5C9

Abstract

Computational gene regulation models provide a means for scientists to draw biological inferences from large-scale gene expression data. The expression data used in the models usually are obtained in a time series in response to an initial perturbation. The common objective is to reverse engineer the internal structure and function of the genetic network from observing and analyzing its output in a time-based fashion. In many studies (Wang [39], Resendis-Antonio [31]), each gene is considered to have a regulatory effect on another gene. A network association is created based on the correlation of expression data. Highly correlated genes are thought to be co-regulated by similar (if not the same) mechanism. Gene co-regulation network models disregard the cascading effects of regulatory genes such as transcription factors, which could be missing in the expression data or are expressed at very low concentrations and thus undetectable by the instrument. As an alternative to the former methods, some authors (Wu et al. [40], Rangel et al. [28], Li et al. [20]) have proposed treating expression data solely as observation values of a state-space system and derive conceptual internal regulatory elements, i.e. the state-variables, from these measurements. This approach allows one to model unknown biological factors as hidden variables and therefore can potentially reveal more complex regulatory relations.

In a preliminary portion of this work, two state-space models developed by Rangel et al. and Wu et al. respectively were compared. The Rangel model provides a means for constructing a statistically reliable regulatory network. The model is demonstrated on highly replicated T-cell activation data [28]. On the other hand, Wu et al. develop a time-delay module that takes transcriptional delay dynamics into consideration. The model is demonstrated on non-replicated yeast cell-cycle data [40]. Both models presume time-invariant expression data. Our attempt to use the Wu model to infer small gene regulatory network in yeast was not successful. Thus we develop a new modeling tool incorporating a time-lag module and a novel method for constructing regulatory networks from non-replicated data. The latter involves an alternative scheme for determining network connectivity. Finally, we evaluate the networks generated from the original and extended models based on a priori biological knowledge.

Acknowledgements

I would like to thank my supervisors, Dr. Anthony Kusalik, for his guidance and patience throughout my graduate studies and Dr. Gopalan Selvaraj, for his inspiration and financial support. I also like to acknowledge Dr. FangXiang Wu for many useful discussions on state-space modeling and much advice on MatLab programming. My gratitude extends to Dr. Mark Eramian for his great help with LaTeX formatting. My family and my wife, Mariatta Wijaya, gave me consistent support through the years. This thesis would not be possible without them.

Contents

List of Tables

List of Figures

List of Abbreviations

AIC	Akaike's Information Criterion
BIC	Bayesian Information Criterion
ChIP	Chromatin immunoprecipitation
ChIP-on-Chip	Chromatin immunoprecipitation on a microarray chip
FFL	Feed-forward loop
GNWD	Gene Network With Delay
MISO	Multiple input, single output
PCR	Polymerase chain reaction
SISO	Single input, single output
TF	Transcription factor
TG	Transcription factor targeted or regulated gene
LOF	List of Figures
LOT	List of Tables

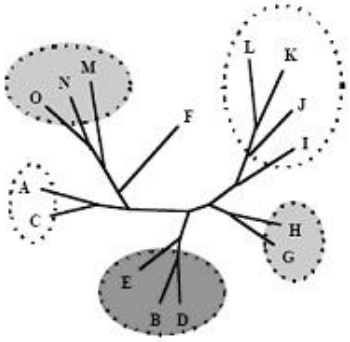
Chapter 1

Introduction

Microarray technology allows researchers to study expression profiles of thousands of genes simultaneously. One of the ultimate goals for measuring expression data is to reverse engineer the internal structure and function of a transcriptional regulation network. This is usually achieved by measuring changes in gene expression levels through time in response to an initial stimulation such as environmental pressure or drug addition.

The data collected from time-course experiments are subject to cluster analysis to identify patterns of expression triggered by the perturbation [13, 27]. A fundamental assumption is that genes sharing similar expression patterns are commonly regulated, and that the genes are involved in related biological functions. Biologists refer to this as "guilty by association". Some frequently used clustering methods for finding co-regulated genes are hierarchical clustering, trajectory clustering, k-means clustering, principal component analysis (PCA), and self-organizing maps (SOM). A general review of these clustering techniques is described by Belacel et al. [3]. A gene network derived from the above clustering methods is often represented as a wiring diagram. Figure 1.1 is an example of a discretized trajectory clustering result. Trajectory cluster analysis groups genes with similar time-based expression patterns (i.e. trajectories) and infers shared regulatory control of the genes. The clustering result is represented as trajectories in multidimensional space (Figure 1.1 (top)), which allows one to find the part-to-part correspondences between two trajectories. The extent of gene-gene interactions are captured by heuristic distances generated from the cluster analysis. The output networked diagram from the analysis (Figure 1.1 (top)) provides insights toward the underlying molecular interaction network structure (Figure 1.1 (bottom)).

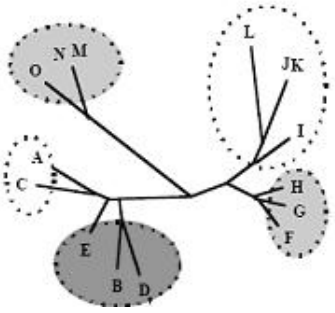
Two major limitations of the conventional clustering methods are that: (1) they cannot capture the effects of regulatory genes that are not included in the microarray; (2) they do not account for transcriptional time delay which occurs in cells. For example, the rate at which a gene is transcribed is a function of the abundance of its transcribing complex, and that complex typically contains several proteins. Some of these are core proteins that catalyze mRNA synthesis and others are factors that modulate mRNA synthesis according to the genetic and environmental specifications for a given gene. Consequently, transcription of such genes is delayed due to the time needed for the production of the corresponding transcription factors.



Trajectory (Gene Expression) Clusters

trajectory	I											II				III				IV		
time	1	2	3	4	5	6	7	8	9	10	11	2	3	4	1	2	3	4	1	2	3	4
A	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1	0	0	0	1	0	0	0
B	0	0	0	0	0	0	0	1	1	1	1	1	0	0	1	0	0	0	1	0	0	0
C	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1	1	0	0	1	0	0	0
D	0	0	0	0	0	0	0	1	1	1	1	1	0	0	1	1	0	0	1	1	0	0
E	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
G	0	0	0	0	0	0	1	1	1	1	1	0	0	0	1	0	0	0	1	0	0	0
H	0	0	0	0	0	0	1	1	1	1	1	0	0	0	1	0	0	0	1	0	0	0
I	0	0	0	0	0	1	1	1	1	1	1	0	0	0	1	0	0	0	1	0	0	0
J	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
M	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
O	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Wiring (Molecular Interaction) Clusters



gene	Boolean rule
A	F and H and J
B	G and H and J
C	F and H and I
D	G and H and I
E	H and I and J
F	I and J and K and L and (not G)
G	I and J and K and L and (not O)
H	I and J and K and L
I	J and K and L
J	K and L
K	K or L
L	L or M
M	N or O
N	N and O
O	N and O and (not E)

Figure 1.1: Inference of shared control through cluster analysis of the time-course gene expression data (top); an example of molecular interaction network represented by a wiring diagram (bottom) (diagram modified from Somogyi [35]).

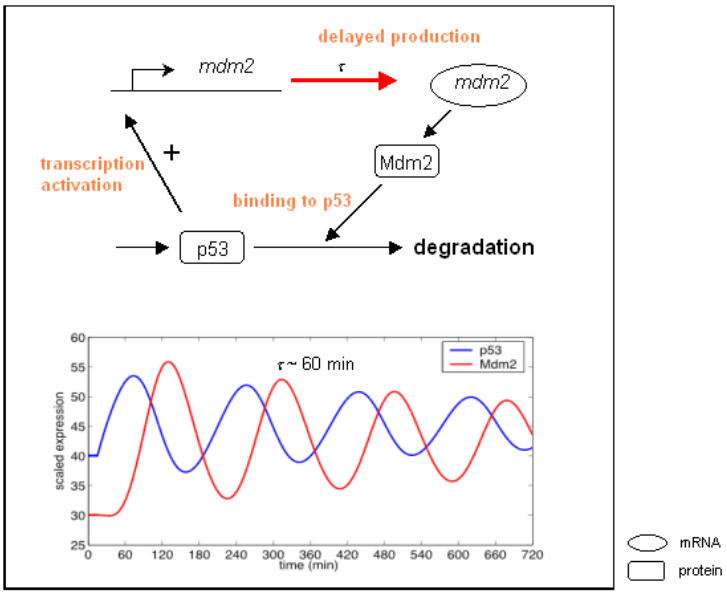


Figure 1.2: A simplified example of p53-mdm2 transcriptional delay feedback loop (diagram modified from Bar-Or [2]).

Figure 1.2 illustrates a simplified example of transcriptional delay in human cells. The transcriptional delay relation of p53 and mdm2 has been demonstrated by Bar-Or et al. [2]. Over-expression of p53 triggers a negative feed-back mechanism. First, p53 stimulates expression of mdm2 gene. The production of mdm2 protein in turn represses the transcriptional functions of p53 and promotes p53 proteolytic degradation. Under the stress conditions, p53 and mdm2 proteins undergo damped oscillations where the mdm2 peaks with a delay of $\tau \approx 60$min relative to p53. Ota et al. [24] have conducted a comprehensive analysis of delay in transcriptional regulation by using gene expression profiles in yeast.

As an alternative to the previous models, some authors (Wu et al. [40], Rangel et al. [28], and Li et al. [20]) have proposed state-space models in an attempt to account for the effects of missing data and complex time-delayed relationships. These methods will be described in Section 3.2. Our study reveals limitations of the Wu and the Rangel models. The results are described in Section 3.3 and 3.4. To complement the existing methods, we have developed a new modeling tool called GNWD. This program allows one to model gene regulatory network with time delays and provides cross-validation to the ChIP-on-chip data. The results are described in Chapters 4 and 5, respectively.

Chapter 2

Background

2.1 Microarrays

2.1.1 What is a DNA microarray?

A DNA microarray or chip is a collection of synthetic oligonucleotides or polymerase-amplified DNA (amplicons) immobilized onto a solid surface such as a glass or silicon slide. These DNA segments are called probes. The sequence identity for each probe is known. The circular surface occupied by a unique probe in high concentration is called a spot. The complementary strand of DNAs, which the probes hybridize with, are called targets. In a gene expression experiment, targets are complementary DNA (cDNA) pools that are reverse transcribed from total or messenger RNA (mRNA) or synthetic oligonucleotides. The targets are labelled by a fluorescent dye such as Cyanine (Cy3 or Cy5) for signal detection. Ideally, the signal intensity of a spot reflects the amount of labelled target bound to the probe and therefore the level of RNA expression in the original biological sample. DNA microarray technology provides a platform for researchers to study expression profiles of thousands of genes simultaneously because of miniaturization.

2.1.2 Types of DNA microarrays

There are two types of DNA microarrays: spotted and *in situ* synthesis arrays. The two are differentiated by how DNA probes are immobilized on the chip. For spotted arrays, the amplified DNA samples are transferred from micro-well plates and immobilized onto a chip using programmable micro-pipettes or pins that deliver a set volume. For *in situ* synthesis arrays such as Affymetrix arrays, the oligonucleotides are synthesized directly on the chip using a photolitographic method.

Depending on the type of probes printed, spotted arrays can be further broken down into cDNA and oligonucleotide arrays. The probes on a cDNA array are products of polymerase chain reaction (PCR) amplification. They are double stranded DNA and are greater than 150 bases long. These probes are PCR amplified based on collected expressed sequence tags (ESTs) or full-length cDNA clones. Therefore, each probe represents one gene. On the other hand, the probes on oligonucleotide arrays are products of chemical synthesis. They are single stranded DNA and less than 100 bases

long. These shorter probes are computationally designed using whole genome information. Multiple probes can be designed for each gene in order to increase the extent of gene coverage and therefore attribution of hybridization to a given gene sequence.

A whole-genome promoter array is an example of a spotted oligonucleotide microarray. The main application of this special chip is to help researchers gain an understanding of the orchestration of patterns of expression. Unlike the others, the probes on these arrays are based on the gene promoter regions. Hence, the targets (complementary to probes) are not directly expressed in the living cells. These special arrays are developed for ChIP-on-chip experiments (see Section 2.2). For probe design and array production protocols, refer to Young (*Saccharomyces cerevisiae* array [23]) and Thibaud-Nissen (*Arabidopsis thaliana* array [38]).

2.2 ChIP-on-Chip

ChIP-on-Chip is an acronym for "Chromatin Immunoprecipitation on a glass slide microarray (chip)". The technology is also known as genome-wide location analysis. Chromatin Immunoprecipitation, or ChIP, refers to a laboratory technique used to determine whether a protein binds to a specific DNA sequence *in vivo*. This technique has been combined with whole-genome promoter microarrays to determine the genomic locations (DNA fragments) where a transcription factor (TF, a class of regulatory protein) would bind.

Figure 2.1 illustrates a simplified ChIP-on-Chip procedure. A protein binds to various promoter regions (the blue solid rectangle in Figure 2.1) in vivo in order to promote or inhibit gene transcriptions. The TF-DNA complex is cross-linked by formaldehyde and the chromosome is fragmented by sonication. A TF-specific antibody is used to isolate the TF-DNA complex from the vast pose of DNA fragments. Once the isolation is complete, a reverse cross-linking is performed to release the DNA fragments. Finally, the DNA are amplified, labeled (by fluorescing cyanine), and hybridized to a genomic array. Further details on the experiment procedure and protocols are available at Young's website [23].

The genomic array data is collected and analyzed. Since the experimental procedure pre-selects for the binding DNAs, the signal intensities of the corresponding spots are enhanced. The stronger the signal intensity of a spot, the greater the significance of the binding. Consequently, a rank-statistics based analysis can be applied to determine the binding sites. These binding sites are potential cis-elements that are directly responsible for the transcription regulation of the target gene.

An important application of the location analysis is to help scientists understand the regulatory circuitry in the biological systems. In yeast, Lee et al. [19], Iyer et al. [16], and Ren et al. [30] have used ChIP microarrays to connect most transcription factors listed in the Yeast Proteome

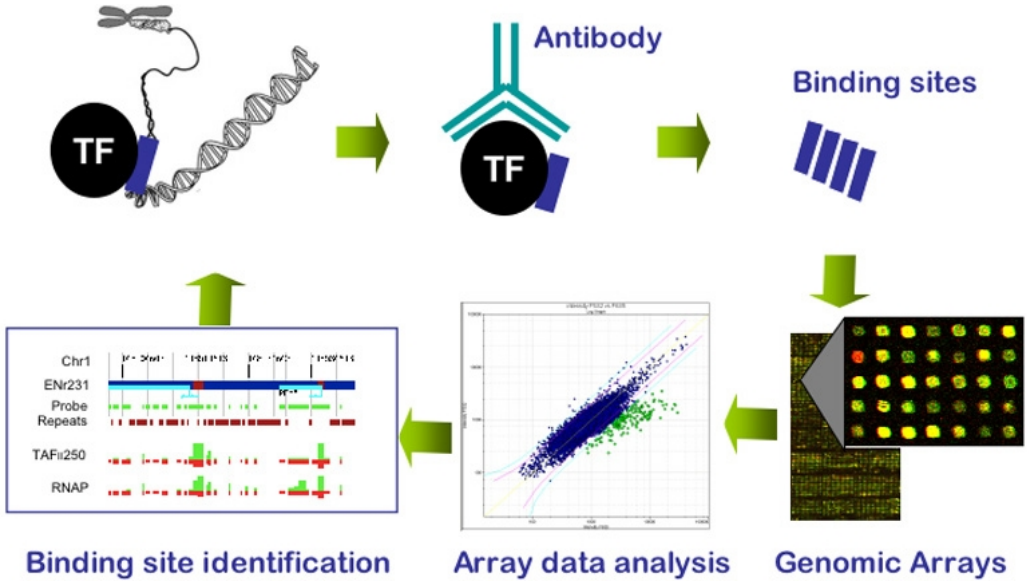


Figure 2.1: ChIP-on-chip procedure. Figure is taken from http://www.chiponchip.org/

Database (YPD) [8] to a large number of target genes.

2.3 Yeast Cell Cycle

A cell cycle consists of two main activities: DNA duplication and cell division. The two activities are separated by cell growth and preparation for mitosis. Thus the yeast cell cycle can be broken down into four phases:

G1	:	Cell growth and preparation for DNA replication
S	:	DNA synthesis/duplication
G2	:	Preparation for mitosis
M	:	Mitosis

A group of proteins in the cytoplasm control the cell cycle process. In budding yeast, the main regulators are cyclins (CLN1-3, CLB1-6) and a cyclin dependent kinase (CDC28). The cyclins are regulated, at the transcription level, by a set of transcription factors. The expression of these transcription factors and cyclins are thought to be phase specific. In Figure 2.2, the stages of the cell cycle are depicted together with yeast cell morphology (yellow), transcription factors (blue), and the cyclins that regulate Cdc28 activity (green). The transcription factors and cyclins are positioned to represent the stage during which they are thought to function [5, 22].

Figure 2.3 is a heatmap representation of the cyclin gene expressions obtained from time-course microarray experiments. The color-coded rectangles capture the extent of up- (red) or down-

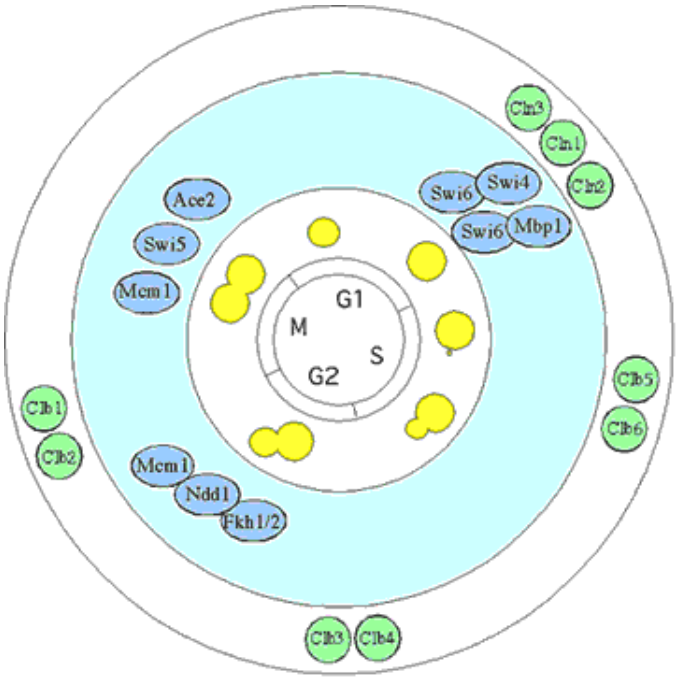


Figure 2.2: Yeast cell cycle showing morphology (yellow), the phase specific transcription factors (blue), and cyclins (green). The transcription factors and cyclins are positioned to represent the stage during which they are thought to function. Figure is taken from http://web.wi.mit.edu/young/cellcycle.

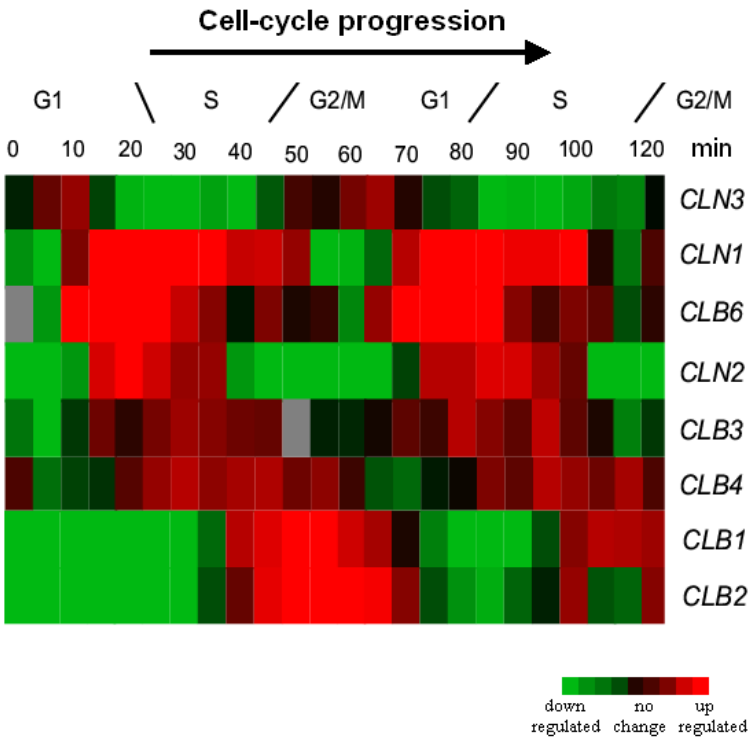


Figure 2.3: Heatmap for cyclin gene expression levels relative to a common reference at $t = 0$. The gradient represents the extent of up- (red) or down- (green) regulation of the cyclins relative to the reference. A grey rectangle represents a flagged/filtered observation. Figure modified from SGD [10].

(green) regulation of the cyclins relative to a common reference (refer to Spellman's experiment protocol [36] for details). The expression levels of these cyclins are used as landmarks for specific phases of the yeast cell cycle. For example, CLN1 cyclin is known to participate in the regulation of transition of G1 phase to S phase [5]. The expression of CLN1 is expected to peak at the G1 $\rightarrow$ S transition point. Using this information, the period of $t = 0$ to $t = 25$ is considered as the G1 phase. Based on these landmarks, one could roughly partition the expression data shown in Figure 2.3 into two cell cycles of approximately 60 minutes each.

Pramila et al. [26] applied a permutation based statistical method (PBM5 [12]) to rank the significance of the cell cycle regulated genes based on their periodicity and magnitude of oscillation among five cell-cycle data sets. Using the boundary information defined by the landmarks mentioned above, the duration each of the periodically expressed gene takes to reach its peak expression is estimated. Peak time is represented in percentage, where the M/G1 boundary is taken as 0%.

While a ChIP-on-chip experiment helps answer the question: "What genes does the transcription factor regulate?", peak-time analysis provides insights as to when these genes are maximally expressed. State-space modeling provides a means to investigate how the genes might be regulated. In this study, we focus on state-space modeling approaches for determining such a gene regulatory network. Existing techniques will be described in Chapter 3.

Chapter 3

Related Research

3.1 Overview

State-space representation is often used in control engineering to model the states of a complex physical system. The internal state variables represent the smallest subset of system variables that can characterize the entire state of the system at any given time. The state of a system is defined by the internal state variables, which together with the initial states ($t = t_0$) and inputs $u_1, ..., u_n$, determine the state at any future time $t \geq t_0$.

There are three major advantages of using a state-space method for modeling a complex system such as a gene regulatory network. First, the system analysis is accomplished by solving a set of first-order equations rather than an equivalent, higher-order equation. This greatly simplifies the mathematical notations and the process of solving the equations. Secondly, the approach uses latent variables (i.e. hidden variables) to relate explicitly the timing of the stimulus inputs and the past system behavior to the observed expression activity. This means that at any point in time, the model can predict how the system responds to the input stimulation. Finally, the goodness-of-fit can be assessed by a well-established method such as the Kolmogorov-Smirnov (K-S) test [6]. This allows researchers to assess the degree of agreement between the model and the experimental data, or between the model derived from one biological system and another. For example, a question of interest may be: do all *E. coli* strains utilize the same regulatory mechanism for glucose metabolism? The question may be answered by comparing the gene regulation models generated for different *E. coli* strains.

3.2 Modeling Gene Network Using a State-Space Approach

Wu et al. [40] develop a state-space model with time-delay to model yeast cell-cycle data. The model is demonstrated on non-replicated data. In the same vein, Sung et al. [37] develop a discretized Bayesian network model to contruct a multiple time-delay gene network using the same dataset. Both methods attempt to capture the effects of transcriptional delays in the yeast cell-cycle. The former method emphasizes identification of a set of internal state variables that govern the cell-

cycle process. The later approach focuses on finding regulatory relationships and associating the regulatory time delay with every "parent-child" (i.e. regulator-target) pair. The Wu model assumes that one gene does not directly regulate another. Therefore, the method does not partition the dataset. The drawbacks of this model are that it is not clear how a network can be derived from the modeling tool, and there is no validation against biological knowledge of the effects of time-delay modeling. The Sung model suggests a new network structure learning algorithm, "Learning By Modification" (LBM), to identify potential regulators and then associates them with target genes. The dataset is partitioned into parent set (the regulators) and child set (the targets).

Rangel et al. [28, 29] present another state-space model, applied to T-cell activation data. The model provides a means for constructing reliable gene regulatory networks based on bootstrap statistical analysis. The method is applied to highly replicated data. The confidence intervals of gene-gene interaction matrix elements are estimated by resampling with replacement the replicates X times, where X is a large number (e.g. 200). This approach however has a severe limitation for application in microarray data analysis because most currently available time course microarray data are either replicated over few time points (< 5) or not replicated at all.

Recently, Li et al. [20] publish their work on infering transcription factor activies using a discretized state-space modeling technique. The Li model incorporates the results of ChIP-on-chip experiments into the model building. The network structure is pre-determined by the genome-wide binding assay data. The transcription factor activities are then inferred with mathematical modeling using time-course experiments.

A preliminary portion of this work studied the state-space models developed by Rangel et al. [28] and Wu et al. [40] respectively. The results will be presented in the following section. A difference in data requirements and the lack of publicly available datasets make a detailed, result-based comparison of the two state-space models almost impossible. Table 3.1 provides a feature comparison of the above four models with our own model, GNWD. The details of GNWD will be described in Chapter 4.

In Table 3.1, five state-space models (Wu et al. [40], Rangel et al. [28], Li et al. [20], Sung et al. [37], and GNWD) are examined for: (1) the method used for the network structure determination; (2) the type (continuous or discrete) of the modeling; (3) the method used for the model realization; (4) whether they handle multiple time-delays (i.e. for $\tau > 1$) in the biological system; and (5) whether the model works with non-replicated data. Among the five models, those of Li and Sung are discrete models. Gene expression data are converted into a series of "on" and "off" states over time. It is however, a non-trivial task to determine the presence or absence of TF activity based on a spectrum of continuous expression data. Similar to the Li model, GNWD utilizes ChIP-on-chip data to assist network structure determination. Unlike the Li model, GNWD is a continous model that is designed to capture complex time-delay relationships. The Wu, Sung, and GNWD models

Model	Network Structure	Type	System Identification	Multiple Time-delays	Non-replicated Data
Wu	N/A	Continuous	Maximum Likelihood Factor Analysis	Yes	Yes
Rangel	BootStrap	Continuous	Expectation Maximization	No	No
Li	ChIP	Discrete	Expectation Maximization	No	Yes
Sung	Learning by Modification	Discrete	Expectation Maximization	Yes	Yes
GNWD	ChIP	Continuous	Subspace Identification	Yes	Yes

Table 3.1: Overview of state-space methods for modeling gene networks.

handle multiple time-delays in the biological system and can work with non-replicated data. With the Sung being a discrete model, and it not being clear how the network structure can be derived from the Wu model, the GNWD model provides a solution that complements these existing models.

3.3 Study of the Wu and the Rangel State-Space Models

In Wu's model, gene expression of n genes at any time t, $x_1..x_n$, are view as the output of a regulatory network system. The gene expression dynamics are governed by the linear combinations of a set of internal state variables, $z_1..z_m$, as illustrated in Figure 3.1.

The model can be described mathematically by the following equations:

$$z_{t+1} = \sum_{\tau=0}^{\tau_{max}} Y_\tau \circ A_\tau z_{t-\tau} + w_t \tag{3.1}$$

$$x_t = Cz_t + v_t \tag{3.2}$$

where A_τ is the state transition matrix for τ time-delays, and $Y_\tau(t) = [y_{ij\tau}]_{p\times p}$ ($\tau = [0, \tau_{max}]$) are boolean matrices which capture the time-delayed regulatory relationships. The value in each element of the boolean matrix Y_τ can be either "1" or "0", which corresponds to the presence or absence of a time-delay between internal variables i and j with τ number of time-delays. The symbol "$\circ$" denotes the Hadamard (element-wise) multiplication of Y_τ and A. C captures the influence of internal state variables on gene expression level at each time point and w_t and v_t are uncorrelated white noise sequences. The identification of the number of internal state variables is estimated either by using the Bayesian Information Criterion (BIC) (Schwarz [32]) or Akaike's Information Criterion (AIC). For further details, refer to Wu et al. [40]. A Bayesian network representation of Wu's model is shown in Figure 3.2.

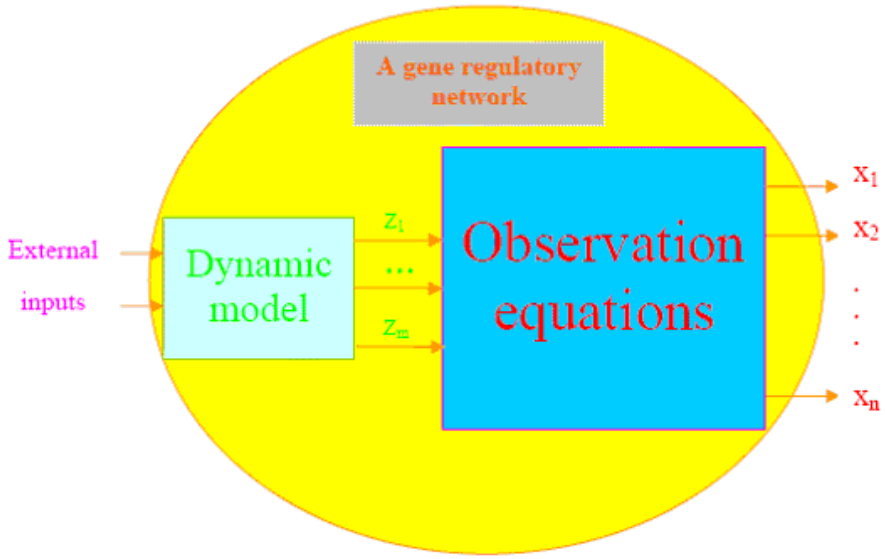


Figure 3.1: A state-space model for a gene regulatory network (from Wu et al. [40]).

Using the similar notation, Rangel's method can be described by the following equations:

$$z_{t+1} = Az_t + Bx_t + w_t \tag{3.3}$$

$$x_t = Cz_t + Dx_{t-1} + v_t \tag{3.4}$$

Two additional input matrices B (input to state matrix) and D (input to observation matrix) are added in Rangel's model. Input to the state matrix describes the influence of gene expression values from previous time points on the hidden states. This models the biological phenomenon of feedback regulation of gene expression to the regulatory elements, i.e. internal state variables. Input to the observation matrix captures the influences of gene-gene expression levels at consecutive time points. A Bayesian network representation of this model is shown in Figure 3.3.

The abstraction of regulatory elements allows one to make predictions of how the system will respond to an input stimulation. This leads to a testable mechanism for gene regulation. The caveat is that it makes associating biological meanings and functions to these conceptual regulatory elements difficult. Each internal variable could represent one regulatory element, or a cascading effect of multiple regulatory elements in the network. It should be noted that the internal variables are not necessarily system outputs. Therefore, they may not always be directly accessible, measurable or controllable. Without biological knowledge of what the regulators are and how they work together in the system, the model creates a black box effect (the boxed area in Figure 3.2 and 3.3) for understanding the full structure of the regulatory network.

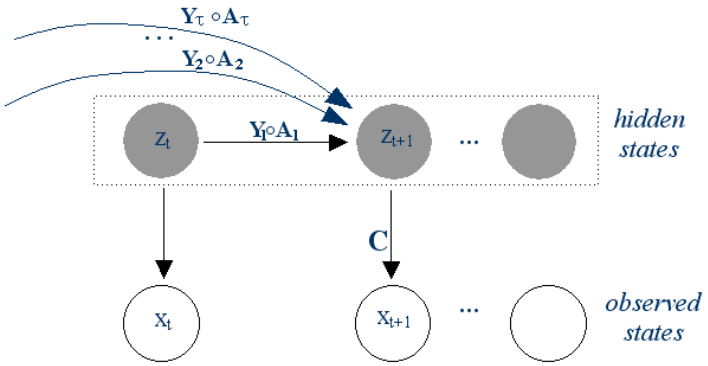


Figure 3.2: Bayesian network representation of the Wu model for gene expression. At time $t+1$, the hidden state Z is determined by the sum of element-wise matrix multiplication between Y and A over τ time-delays.

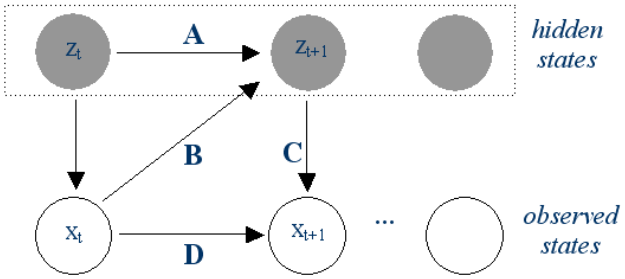


Figure 3.3: Bayesian network representation of the Rangel model for gene expression. At time $t+1$, the hidden state Z is determined by the state matrix A and the input matrix B.

3.3.1 Internal Variables

In Wu's state-space model, the gene expression levels (observation values) are governed by the linear combination of the internal state variables. It has been shown using the Spellman data [36] that there are 6 and 4 internal variables for the 701 alpha factor-synchronized (ALP) and elutriation-synchronized (ELU) cell-cycle regulated genes, respectively [40]. It is not clear however, what biological meanings these internal variables have. Each variable could correspond to one key regulatory enzyme in a pathway, one key process in a cellular activity, or even one dominant pathway in cells. As suggested by Spellman [36], many of the cell-cycle regulated genes do not play any role in cell-cycle regulation but participate in other known biochemical activities. Therefore, it is possible that some of the state variables correspond to those genes among the 701 which are

Pathway	Code	Number of input cell-cycle regulated genes (pn)	Number of internal variables (k)
Glycolysis / Gluconeogenesis	sce00010	5	4
Fructose and mannose metabolism	sce00051	5	4
Galactose metabolism	sce00052	6	5
Oxidative phosphorylation	sce00190	6	5
Purine metabolism	sce00230	14	13
Pyrimidine metabolism	sce00240	12	11
Selenoamino acid metabolism	sce00450	5	4
Starch and sucrose metabolism	sce00500	9	8
Aminosugars metabolism	sce00530	6	5
Nicotinate and nicotinamide metabolism	sce00760	5	4
DNA polymerase	sce03030	8	7
MAPK signaling pathway	sce04010	14	13

Table 3.2: The number cell-cycle regulated genes and the number of internal variables identified by the Wu model for the selected biochemical pathways.

involved in other cellular activities.

In this section, we examine the possibility of breaking down the number of genes (i.e. the 701 cell-cycle regulated genes) into smaller, biologically meaningful groups in an attempt to unveil the correlation between the number of internal variables identified by the Wu model and the grouping of the genes. Ideally, by providing the model a smaller subset of genes which function in a single biological process or pathway, one could expect the number of internal variables to reflect the regulation complexity of that particular process or pathway. Superimposing the information with the knowledge of the process or pathway such as the number of rate-determining steps, kinetic models or regulation mechanism of the constituent enzymes found in the literature, this could provide leads to deciphering the biological meanings (if there are any) of the internal variables.

With the above mentioned objective in mind, we searched the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database [9] for the biochemical pathways associatied with the 701 cell-cycle regulated genes. 156 of these genes are assigned to at least one biochemical pathway in the database. In order for the pathway information to be useful, the number of genes mapped to each pathway (pn) must at least be greater than or equal to 2 (i.e. A regulates B or vice versa). An arbitrary cutoff of $pn >= 5$ is applied. Using these criteria, we selected 13 biochemical pathways, apart from the cell-cycle biochemical pathway itself. The gene expression data for each pathway

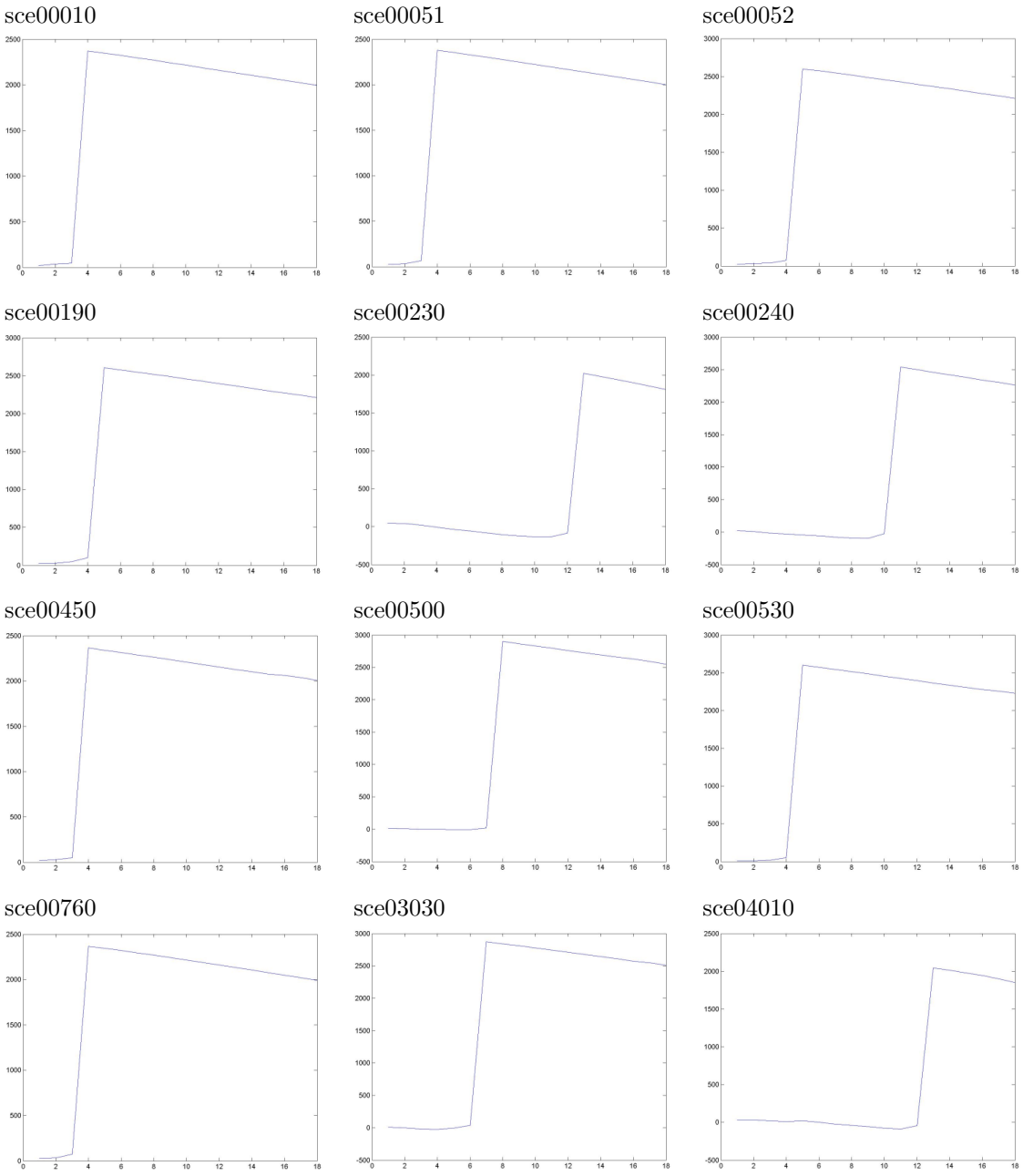


Table 3.3: BIC for each pathway (y-axis) against the number of internal variables, k (x-axis).

are queried from the Spellman dataset. For each pathway, we run the Wu state-space model on the associated gene expression data. Table 3.2 outlines the pathways used in this study, the number of input cell-cycle regulated genes and the number of internal variables identified by Wu's model. It appears that when the number of genes (n) is less than the number of time points (m, $m = 18$ in this case), the number of internal variables (k) is always one less than the number of input genes for the pathways. Figure 3.3 illustrates the corresponding BIC values for each pathway under study. Note that the model consistently reaches a peak at $k = n - 1$ for all pathways.

One could hypothesize that the expression of any gene in a pathway can be expressed as a function of the rest of genes in the pathway. Therefore, the number of internal regulatory elements for a pathway involving a small number of genes is always one less than the total number. This information, however, is not at all useful in finding the biological meaning of the internal variables. The negative results suggest that the model does not support small gene networks. Therefore, our attempt to decipher the internal variables using biochemical pathway data was unsuccessful.

3.3.2 Time-Delay and Noise

Rangel's model expects highly replicated data, which has a large number of timepoints for control analysis and is sufficiently replicated for the bootstrap analysis. To the best of our knowledge, currently there is a very limited amount of microarray data that fit the requirements. Due to this lack of actual experimental data, we have generated a set of artificial data as described in Table 3.4 in order to evaluate Rangel's model. The data consists of one regulator U and 5 genes (G1 - G5). It is assumed that the expression of the regulatory is periodic, described by a sine function. G1, G2, and G3 are expressed according to sine functions with time-delay of $\tau = 0, 1, 2$ respectively. G4 and G5 are random, uniformly distributed data with the range of -0.1 to 0.1 (i.e. one tenth of the range of a sine function, -1 to 1). G4 and G5 provide noise to test how the model handles uncorrelated data. The artificial data consists of 18 timepoints and 16 replicates. Each replicate has uniformly distributed noise in the range of -0.05 to 0.05 (i.e. one twentieth of the range of a sine function) assigned to each timepoint. Figure 3.4 is a graphical representation of the artificial data. The model is tested in two aspects: the ability to analyse time-delay data and the ability to handle noise.

Default parameters for the number of cycles, number of bootstrap samples, tolerence, etc. are used for the test. Figure 3.5 illustrates the log likelihood scores of the training and validation sets. Based on the likelihood scores, the number of internal variables is set to 3. Figure 3.6 is the network output of Rangel's model generated from the artificial data. It seems that the model can correctly identify network connection when the input delay equals to 1 ($\tau = 1$). These relations are the U $\rightarrow$ G2, G1 $\rightarrow$ G2, and G2 $\rightarrow$ G3 in this test. The model also detects connectivities between U to G1 and U to G3. However it did not predict the correct regulatory relationships. The model

Names	Function	Delay (τ)	Regulated by
Regulator U	$2 \times sin(x)$	N/A	N/A
Gene G1	$sin(x) + v$	0	U
Gene G2	$sin(x + \tau) + v$	1	R1
Gene G3	$sin(x + 2\tau) + v$	2	R1
Gene G4	$random()$	N/A	N/A
Gene G5	$random()$	N/A	N/A

Table 3.4: Artificial data for evaluating the Rangel model.

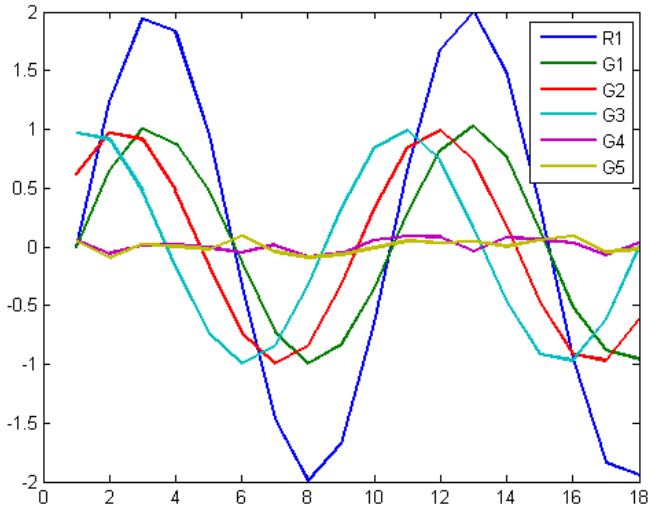


Figure 3.4: Inputs to Rangel's Model.

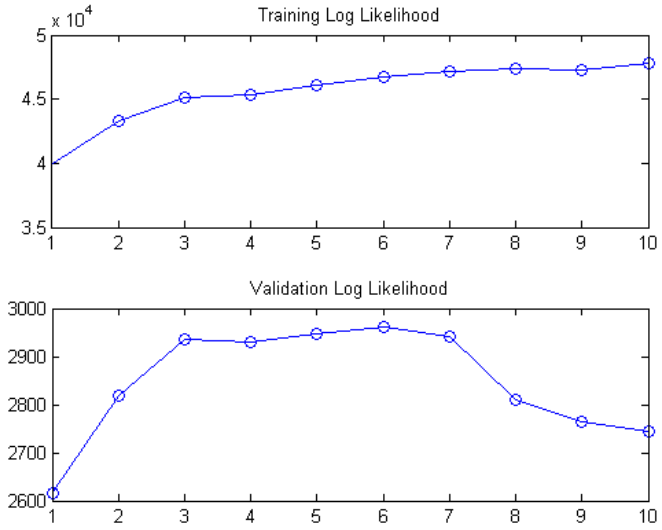


Figure 3.5: Number of internal variables determination.

miss-characterized random data G5 as a regulator for G2 and G3. It should be noted that since G4 and G5 are random data, the connections between G5 → G2 and G5 → G3 are erroneous. The results suggest that one should apply filtering steps to remove noisy data (in this case G4 and G5) prior to the data analysis.

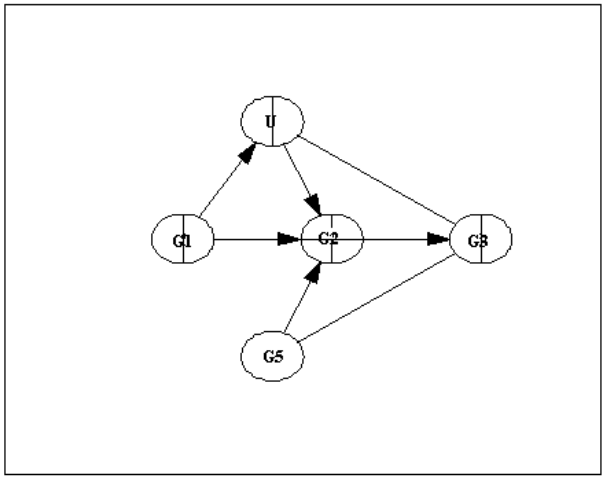


Figure 3.6: Output network from Rangel's model for the artificlal data. Each node represents a gene, each directed edge represents the direction of regulation, each undirected edge indicates connectivity in both directions, and finally, each vertical line that partition a node represents self-regulation.

Chapter 4

Data and Methodology

This thesis proposes a new technique for determining prospective gene regulation models from time-series gene-expression data. The presentation of that technique consists of three parts. First, we implement a state-space model that incorporates multiple time delays. The model is described in Section 4.3. Secondly, we develop an alternative means for determining network connectivity for both non-replicated and replicated data (Section 4.4). This involves replacing Rangel's bootstrap confidence intervals (derived from highly replicated data) for identifying gene-gene interaction with a substitute. Finally, the networks generated from the new model are visualized using techniques from the literature. The results of the modeling are presented in Chapter 5.

4.1 Data

Two datasets are used in this study. First, an artificial dataset is created to validate the model. The artificial data is created in such way that it: (1) mimics the periodic property of cell-cycle microarray data; (2) simulates the systematic errors in microarray experiments; (3) contains multiple time delay relations between regulators and targets. Secondly, we applied our model to analyze the yeast cell-cycle microarray data published by Spellman et al. [36]. Details of both datasets are described in the following subsections.

4.1.1 Artificial data

Yeast cell-cycle regulated genes demonstrate a periodic pattern [36]. Gene expression data are reported as $log_2(\frac{sample\_expression}{reference\_expression})$. That is, one measures the changes in expression with respect to a common reference instead of an absolute expression. A 2-fold change, i.e. $log_2(ratio) = \pm 1$, is generally considered significant. To simulate the periodicity and to mimic the oscillation magnitude, we have create a set of artifical data using $sine()$ and $cosine()$ functions. Refer to Table 4.1 for details.

The artificial data consists of data streams of 2 regulators, R1 and R2, and 9 target genes, G1, G2, ..., G9. G1 to G3 are associated with R1 with delays $\tau = 0, 1, 2$, respectively. G4 to G6 are associated with R2 with delays $\tau = 0, 1, 2$, respectively. These relatively simple cases will test the

Names	Function	Delay (τ)	Regulated by
Regulator R1	$sin(x)$	N/A	N/A
Regulator R2	$cos(x)$	N/A	N/A
Gene G1	$sin(x) + v$	0	R1
Gene G2	$sin(x + \tau) + v$	1	R1
Gene G3	$sin(x + 2\tau) + v$	2	R1
Gene G4	$cos(x) + v$	0	R2
Gene G5	$cos(x + \tau) + v$	1	R2
Gene G6	$cos(x + 2\tau) + v$	2	R2
Gene G7	$sin(x) + cos(x) + v$	0	R1+R2
Gene G8	$sin(x + \tau) + cos(x + \tau) + v$	1	R1+R2
Gene G9	$sin(x + 2\tau) + cos(x + 2\tau) + v$	2	R1+R2

Table 4.1: Artificial data consists of 2 regulators (R1,R2) and 9 genes (G1-G9).

ability of the model to associate the target genes to their regulators, and to predict the number of the delays. G7 to G9 are associated with both R1 and R2 with delays $\tau = 0, 1, 2$, respectively. In these more complex cases, we test the ability of the model to connect the target genes to the multiple regulators, and to predict the number of the delays. Each data stream has a uniformly distributed noise, v, in the range of -0.05 to 0.05 (i.e. one twentieth of the range of sine and cosine functions), assigned to each timepoint.

4.1.2 *Saccharomyces cerevisiae* cell-cycle data

The second dataset used in this project consists of 800 alpha factor-based yeast cell-cycle regulated genes identified by Spellman et al. [36]. The microarray hybridizations were done on asynchronous yeast cell samples at every 7 min for 18 time points. Normalized expression data were downloaded from the Stanford Microarray Database (SMD [11]). No further pre-processing was done. The *knninpute*() function from MatLab's Bioinformatics toolbox was used to impute the missing data using a nearest-neighbor method.

4.2 Assumption

We assume that (1) there exist effects of hidden variables in the biological system that cannot be measured in a gene expression profiling experiment, e.g. missing data or mRNA degradation, and (2) the experiment time points capture the significant physiological changes of the biological system.

4.3 Time Delay Model

We consider the expression vector of a regulator (e.g. a transcription factor) as an input function to the system. Therefore, the time period, τ, from the over-expression of the regulator to the over or suppressed expression of the targeted gene is represented as an input-delay function. Assuming a state-space system with p regulators, q target genes and n state variables, the model can be described using the following equations:

$$z_{t+1} = Az_t + Bu_{t-\tau} + w_t \tag{4.1}$$

$$x_t = Cz_t + v_t \tag{4.2}$$

where A is the $n \times n$ state transition matrix. B is the $n \times p$ input matrix. It captures the impacts of the expression of p regulators on the system. $u(t-\tau)$ is a $p \times 1$ input vector. C is the $q \times n$ output matrix that represents the influence of internal state variables on the output gene expression level at each time point. w_t and v_t are uncorrelated white noise sequences. We adopt Wu's model (see Section 3.3) which removes the feed-through matrix, D, assuming that gene-gene regulation can be captured by indirect regulation through internal variables instead of direct gene regulation from one time point to the next. As described by Rangel et al. [28], the product of $C \times B$ produces a $q \times p$ matrix that depicts the regulatory relationships between p regulators and q target genes. The possible values for the time delay for each of the p regulators, τ_i, where $i = 1, ..., p$, is estimated by scanning a range of positive integers, with the minimum time delay of zero. The best fit is determined by minimizing the Akaike's AIC criterion score for the residual variance (in %). A Bayesian network representation of the model is shown in Figure 4.1.

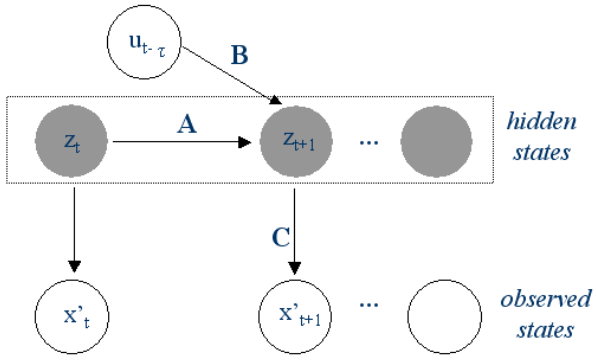


Figure 4.1: Bayesian network representation of the new model for gene expression.

The model was implemented as a MatLab program, Gene Network with Delay (GNWD). GNWD uses various functions from MatLab's Control System and System Identification toolboxes. The

$n4sid()$ and $aic()$ functions are used for system identification, system stability, and delay analysis. The $n4sid()$ function implements a variant of state-space modeling using the sub-space method (N4SID). The N4SID algorithm was developed by De Moor et al. [25], and it computes the parameterization of the model (i.e. solving the matrices A, B, and C). The $aic()$ function returns the Akaike's Information Criterion (AIC) score [1] of the model. The AIC score is used to determine the order of the system and best-fitted number of input time delays as described by Wu et al. [40]. AIC was developed by Akaike and Hirotugu [1]. The method penalizes the complexity of an estimated model in order to avoid overfitting the data. In our case, the best-fitted model is determined by one that has the lowest AIC score. Finally, the $compare()$ function is used to determine the overall model fitness to the data. The model fitness is represented as a percentage, estimated as follows:

$$Fitness = (1 - \frac{norm(Yh - Y)}{norm(Y - \overline{Y})}) \times 100\% \tag{4.3}$$

where $Y(t) = (y_0, y_1, ..., y_n)$ is the actual gene expression vector, $\overline{Y}$ is the mean of Y, and $Yh(t) = (yh_0, yh_1, ..., yh_n)$ is the predicted expression vector from the model. n is the total number of time points. $norm(Yh - Y)$ and $norm(Y - \overline{Y})$ are the Euclidean distances between predicted and the actual expression vectors, and between the actual expression vector and mean expression, respectively. Ideally, if the distance between the predicted and the actual expression vectors is zero, the function returns a 100% fitness.

GNWD supports two modeling tools: single input and multiple input delay models. The single input delay model captures the simple one-to-one regulatory relations. The multiple input delay model works for complex many-to-one regulatory relations.

4.3.1 Single Input Delay Model

In a simple one-to-one regulatory relation, the regulation of a gene is highly related to its transcription factor (TF). In other words, residual regulation by other factors is relatively insignificant and can be treated as hidden variables, i.e. the missing data. Therefore, a single-input and single-output (SISO) model (TF vs. gene or TF vs. TF) can be used to describe the input and output signals. The SISO modeling can be applied to identify network motifs such as feed-forward loops, multi-component loops, and single input motifs as described by Lee et al. [19]. Figure 4.2 illustrates how GNWD is used to model two such network motifs. The network motifs are shown on the left and the corresponding state-space models on the right. According to Lee et al. [19], two anaerobic condition related transcription factors in yeast, Rox1 and Yap6, form a regulatory circuit in which they regulate each other. The regulation circuit is represented as a multi-component loop motif as shown in Figure 4.2 A), where the over- or under- expression of one TF regulates the gene expression of another (i.e. $p = q = 2$). In the GNWD state-space representation, the mRNA expression levels of ROX1 and YAP6 (orange boxes) over time are the observed values. The TF protein expression

levels, Rox1 and Yap6 (purple ellipses), and possibly other hidden factors (purple ellipse labeled with a "?" mark) are the hidden variables. At time t, the protein expression levels are affected by gene expression of ROX1 and YAP6 with τ_1 and τ_2 input time delays, respectively. The hidden variables in turn dictate the output gene expressions of ROX1 and YAP6 genes at time *t+1*. The multiple time-delay relationships can be expressed as a 2×2 matrix as follows:

$$\begin{bmatrix} 0 & \tau_1 \\ \tau_2 & 0 \end{bmatrix}$$

Another example of a network motif is the regulation of CLB2, a G2/M-cyclin, and SWI4 transcription factor by MCM1. It is illustrated by Lee et al. [19] as an example of a feed-forward motif. The MCM1 gene regulates CLB2 as well as the SWI4 transcription factor, which also regulates CLB2 cyclin. In this network motif, there are two regulators, two target genes (i.e. $p = q = 2$), and three possible input time delays and each corresponds to a regulatory relation (refer to Figure 4.2 B). The multiple time-delay relationships is expressed as a 2×2 matrix as follows:

$$\begin{bmatrix} \tau_2 & \tau_3 \\ 0 & \tau_1 \end{bmatrix}$$

The time delay, τ, is estimated by scanning a range of possible integers, with the minimum time delay of zero. In case of yeast cell cycle data, the maximum number of delay should not exceed the time for a complete cell cycle (G1 $\rightarrow$ S $\rightarrow$ G2 $\rightarrow$ M), which is estimated to be $\approx$60 minutes [37]. For Spellman's time-course microarray data, since each sampling interval is 7 minutes, the maximum delay should never exceed 8 sampling intervals (i.e. $60min \times 1sample/7min$). Similar to Li et al. [20] but unlike Ota and Sung [24, 37], we believe that the actual time delay between binding and transcription is on the order of minutes. This is based on an assumption that gene transcriptional regulations are most likely to occur within the same phase or at the transition point from one phase to another. Since the longest cell-cycle phase, G1, takes $\approx$25 minutes (refer to Chapter 2), which means the maximal reasonable delays is less than 3 sampling intervals (i.e. $25min \times 1sample/7min$). Hence, the default maximal delay for yeast cell cycle is set at 2 sampling intervals, i.e. 14 minutes, for Spellman's data [36]. Note that the default value may not be applicable to other biological systems.

4.3.2 Multiple Input Delay Model

SISO may not work well when multiple regulators show significant regulation of a target gene. The presence of a second regulator increases the model complexity. In addition, some studies have shown that different gene pairs have different time delays for gene regulation [7, 37]. Therefore, the multiple time-delay issue should also be addressed. As a result, we present a multiple input delay

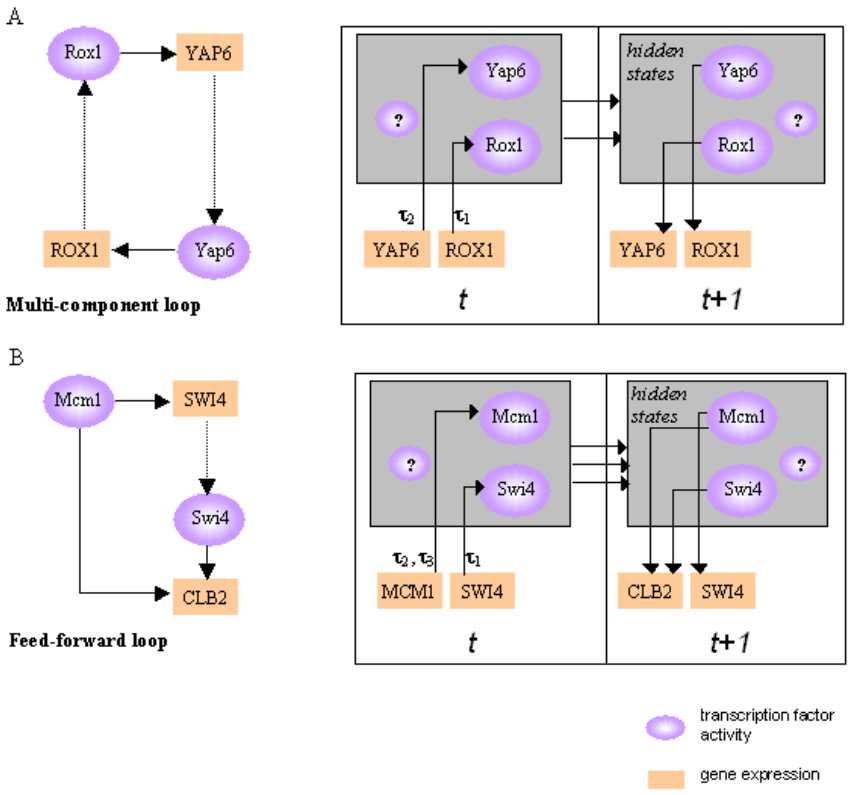


Figure 4.2: An example of SISO state-space representation of the gene regulatory network motifs described by Lee et al. [19]: A) Multi-component loop, and B) Feed-forward loop. The network motifs are shown on the left and the corresponding state-space models on the right.

model. In it, the transcription profiles of all known regulators, if available, are provided as inputs to the system. The input delays are estimated individually for each regulator. The MISO model can be used to determine multi-input and regulator cascade network motifs, as described by Lee et al. [19].

Figure 4.3 illustrates how GNWD is used to model a multi-input network motif. In this example, the protein component of yeast large (60S) ribosomal subunit, RPL16, is transcriptionally regulated by three transcription factors: FHL1, RAP1, and YAP5 (i.e. $p = 3$, $q = 1$). Assuming that each TF has zero or some input delay to the regulation of RPL16, the multiple time-delay relationship can be described as follows:

$$\begin{bmatrix} \tau_1 & \tau_2 & \tau_3 \end{bmatrix}$$

The maximum number of input channels allowed in the model depends on the complexity of the motif structure and the time delay of each input channel. A higher number of available time points is required to model a more complicated network structure. In the case of Spellman's yeast

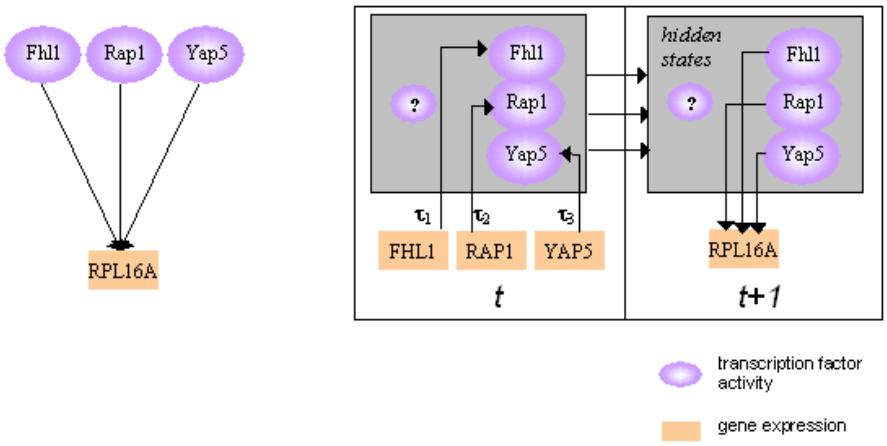


Figure 4.3: An example of MISO state-space representation of a multi-input gene regulatory network motif described by Lee et al. [19]. The network motif are shown on the left and the corresponding state-space model on the right.

microarray data (18 time points), GNWD can compute a stable system for a maximum of four input and input delays. This number is determined by trial and error. It may vary depending on the complexity of the network.

4.4 Network Connectivity

Rangel et al. [28] construct reliable gene regulatory networks based on bootstrap statistical analysis. The method is applied to highly replicated data. This approach has a severe limitation, however, because most currently available time course microarray data are either replicated few times (e.g. less than 5) or not replicated at all. Li et al. [20] use genome-wide location analysis results to construct a network structure and then infer the transcription factor activities with mathematical modeling. The latter approach significantly reduces the false positive node connections since the network connectivity is pre-determined. In addition, the method can be used to model gene regulatory networks from non-replicated data. The limitation of Li's approach is that it removes the power to uncover new connections that are not identified by ChIP-on-chip data.

To complement the existing approaches, we present a three-step solution incorporating GNWD such that network connectivity is based on, but not limited by, genome-wide location analysis results. First, the data is partitioned into two groups: transcription factors (TFs) and target genes (TGs). Each TF is a possible regulator of another TF and/or TG. Secondly, GNWD creates an initial set of network connections based on the location analysis results. At this stage, all location-analysis-derived TF versus TF and TF versus TG regulatory relations are screened for corresponding state-space models. Only the confirmed regulatory relations are recorded and subject

to the next round of analysis. For each TF, GNWD records the optimized parameters (initial state, number of time-delays, number of state variables) that reflect the complexity of the regulations. Finally, GNWD performs the second round of network connection screening based on the regulation parameters generated in the second step. For example, if a transcription factor A regulates n TGs with time delay τ_1, k state variables, and initial state $z(0) = 0$, the GNWD program will attempt to recruit other genes that have not been identified as targets of A but possess regulatory relations with A that resemble the existing ones. This is based on a common assumption that genes with high correlation in expression profiles are likely to be co-regulated. The approach is implemented by MatLab's *pem*() functions which is an alternative to the N4SID algorithm that uses prediction error model (PEM) for parameterization. According to Favoreel et al. [14], the latter algorithm is relatively more sensitive compared to N4SID once the intial parameters are determined.

4.5 Network Visualization

GNWD generates a network output file format that can be directly imported into Cytoscape [33] for network visualization, integration, and analysis.

Chapter 5

Results

This chapter presents the results of modeling using GNWD. First, we describe the output of modeling the artificial data (as described in Chapter 4) and the lessons learned in the modeling process. In Section 5.2, we show the results of modeling yeast cell-cycle expression data. The global regulatory network diagram is presented as well as detailed analysis of G1- and B-type cyclins. Finally, we illustrate the capability of GNWD in selecting the most feasible regulatory mechanism from multiple models.

5.1 Modeling a Gene Network using Artificial Data

To demonstrate the difference between the SISO and MISO models, we first apply only the SISO one to network prediction of the artificial data. The two regulators, R1 and R2, are expected to connect to the target genes, G1 to G9, as described in the data section (see Table 4.1) of the previous chapter. Figure 5.1 is a graphical representation of the SISO network. The network visualization is generated using Cytoscape where each node represents a gene and each directed edge represents a predicted regulatory relationship between a regulator and the target gene. Each edge is labelled with the predicted number of input time delays. Eleven out of twelve edges are identified by GNWD-SISO. Among the eleven, 9 edges are annotated with the correct time-delays. The text output is tabulated in Table 5.1. The "Order" column gives the order of the system that reflects the model complexity. "Fitness (%)" (percent of fitness) reflects the goodness-of-fit of the state-space model to the data. The "AIC" column contains the Akaike's Information Criterion score. The best-fitted model is selected by minimizing the AIC score. Refer to Section 4.3 for more information regarding the columns.

The results show that the SISO model can predict 100% correctly the one-to-one regulations but not the many-to-one regulations. For many-to-one regulations, the SISO model detects 5 out of 6 ($\approx$83%) of them, but only 3 out of 6 are predicted with correct delays. As expected, almost all predicted connections (4 out of 5) from the many-to-one regulation are in higher order state-space systems (i.e. second order state-space systems) compared to the rest. GNWD-SISO predicts a more complex regulation mechanism in these systems and produces poorer scores for the percent of fitness

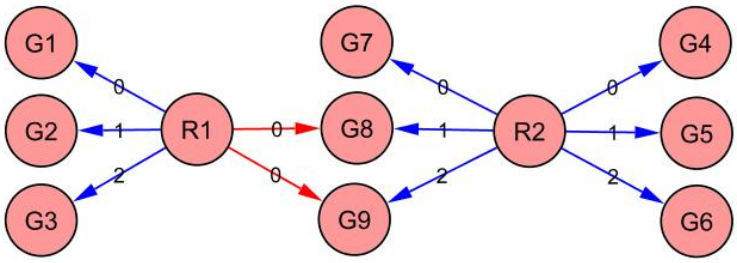


Figure 5.1: SISO output for artificial data. All edges are labeled with the predicted time delays. A blue edge represents a correct interaction; a red edge represents an incorrect one. relationship.

and AIC (see Table 5.1). The fact that the SISO model can identify most of the regulatory relations in our simulation suggests that in the absence of *a priori* knowledge of the network structure, the single-input and single-output model may be used to detect more complex network connections but the number of time-delays and the order of the system may need to be re-assessed by the MISO model.

Regulator	Target	Order	Delay (τ)	Fitness(%)	AIC
R1	G1	1	0	98.76	-9.0735
R1	G2	1	1	98.76	-9.1076
R1	G3	1	2	98.68	-8.9094
R1	G8	2	0	82.40	-3.9379
R1	G9	1	0	81.44	-3.2183
R2	G4	1	0	98.70	-8.8321
R2	G5	1	1	98.66	-8.9243
R2	G6	1	2	98.82	-9.1675
R2	G7	2	0	85.69	-5.0339
R2	G8	2	1	82.82	-4.3840
R2	G9	2	2	83.31	-4.2520

Table 5.1: SISO output for the artificial data.

We then applied the GNWD-MISO model for network prediction of the G7 to G9 genes. Given the knowledge of R1 and R2 co-regulates G7, G8, and G9, GNWD-MISO can correctly predict 6 out of 6 edges and the corresponding number of time-delays. Figure 5.2 is a graphical representation of the results. The text output is shown in Table 5.2. Note that the GNWD can produce much better models (better than 99% fitness, and much lower AIC scores) in latter case. The results illustrate the advantage of incorporating *a priori* network structure knowledge in the modeling process.

Regulator	Target	Order	Delay (t1,t2)	Fitness(%)	AIC
R1,R2	G7	1	0,0	99.27	-8.0843
R1,R2	G8	1	1,1	99.18	-8.6052
R1,R2	G9	1	2,2	99.15	-8.4814

Table 5.2: MISO output for artificial data.

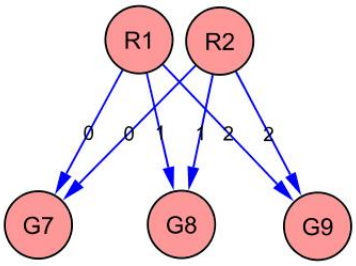


Figure 5.2: MISO output for artificial data.

5.2 Modeling the Gene Network in *Saccharomyces cerevisiae*

5.2.1 Learning the Network Structure

The genome-wide location analysis results of nine known cell-cycle related transcription factors (SWI4, SWI6, MBP1, MCM1, ACE2, SWI5, FKH1, FKH2, and NDD1) were downloaded from Young's website [23]. The results are reported as p-values that reflect the significance of the binding between TFs and the corresponding promoter regions. We considered a p-value less than or equal to 0.01 as being significant. This cut-off is less stringent than the 0.001 cut-off proposed by Lee et al. [19]. A relaxed threshold was selected to reduce the number of false negatives in location analysis. Complementarily, the number of false positives is controlled by providing cross-validation evidence from the modeling of time-series gene expression data. Based on the location analysis results and the selected cut-off, we identified 301 out of 800 cell-cycle regulated genes reported by Spellman et al. [36] which bound to least one of the nine TFs. Refer to Appendix A for the list of the 301 genes and the binding map to the nine TFs. In that table, a "+" sign in a cell represents a significant binding ($p \leq 0.01$).

5.2.2 Modeling Gene Network

We applied our modeling tool to the 301 cell-cycle regulated genes identified above. GNWD predicted the regulation models of 93 genes or approximately 31% of the total input genes. The results

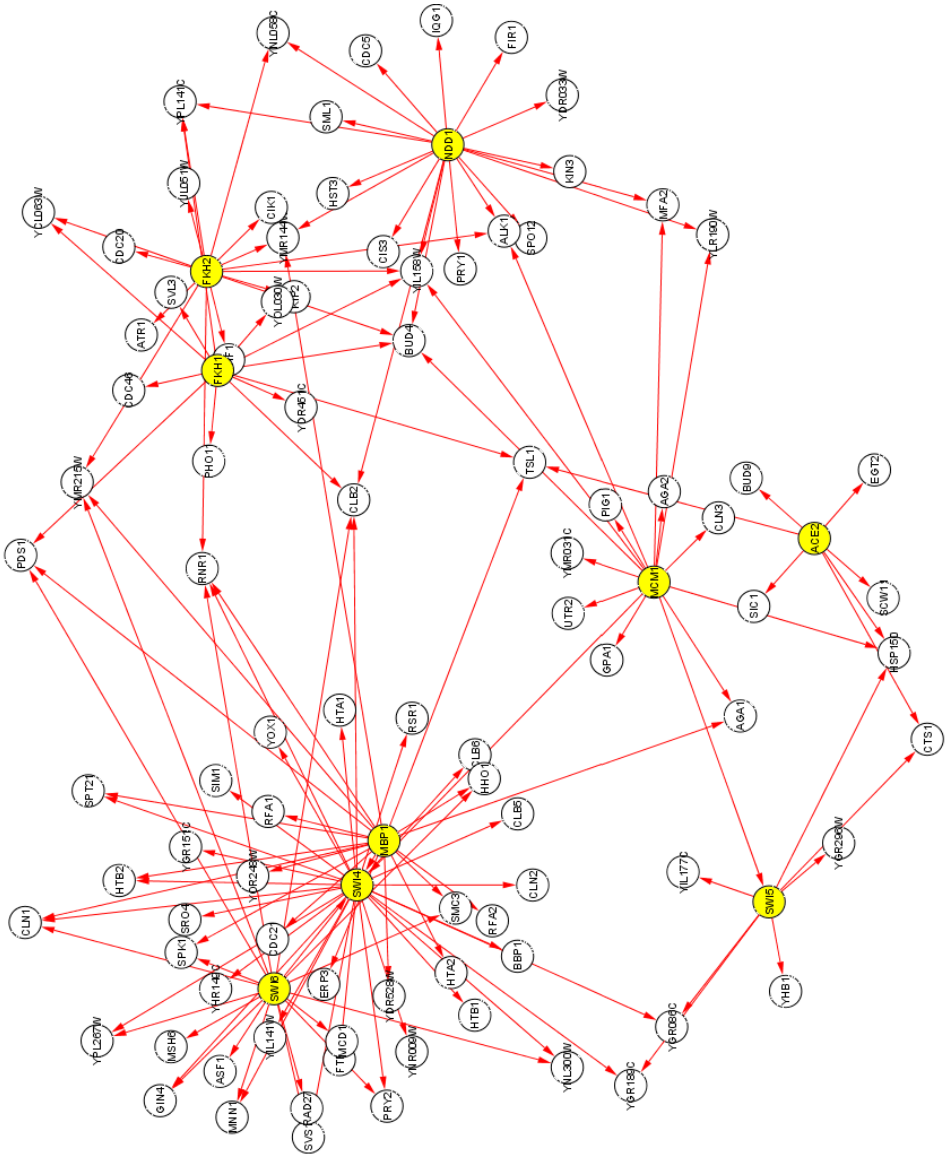


Figure 5.3: Gene regulatory network of 93 cell-cycle regulated genes. Each node represents a gene and a yellow node represents a transcription factors.

are tabulated and shown in Appendix B. On a Pentium III 800MHz computer, the total run time for GNWD to analyze the 301 genes is approximately 90 minutes.

Almost half of the 93 genes are regulated in the G1 phase and about 25% are regulated in the G2/M phase. Compared to the 301 input genes, this represents a minor increase in percentage of genes regulated in G1 phase (36% to 44%), and a slight decrease for M/G1 phase (17% to 12%). Refer to Figure 5.4 for the percentages of the 93 correctly-modelled genes (in dark red) and the 301 input genes (in dark blue) in different phases of cell-cycle. The differential success rates in modeling G1- and M/G1-regulated genes may due to the differences in the number of the TFs from each phase. There was no M/G1 specific transcription factor used in this study. On the other hand, there were three (SWI4, SWI6, MBP1) G1-activated TFs.

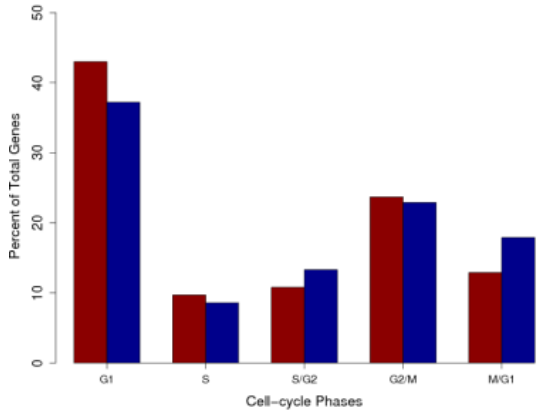


Figure 5.4: Distributions of total input genes and GNWD modelled genes to different cell-cycle phases.

Among the nine transcription factors, SWI4, SWI6, and MBP1 are known for their important roles in G1 and late G1 phase gene regulation [16, 34]. The three TFs encode for two transcription factor complexes: SBF (SWI4 and SWI6), and MBF (SWI6 and MBP1). SBF and MBF control over 50% of the total detected regulatory relations in our model. Refer to Figure 5.3 for the graphical representation of the modelled network. In this network diagram, each yellow node represents a TF and each white node represents a target gene. A directed arrow between a TF and a target gene node represents a detected regulatory relation. Figure 5.3 reveals a large cluster of target genes regulated by combinations of SBF and MBF (left side of Figure 5.3). The forkhead transcription factors FKH1 and FKH2, and NDD1 regulate a smaller cluster of G2/M-phase expressed genes on the right of the network diagram. Among the modelled genes in the two most abundant phases, the regulation of G1 phase's G1-cyclins (CLN1, CLN2, and CLN3) and G2/M phase's B-type cyclins (CLB2, CLB5, and CLB6) are identified. The modelled regulatory mechanisms of the cyclins were

further investigated. The results are discussed in the following subsection.

5.2.3 Regulations of G1- and B-type cyclins

We examined more closely the regulation models of 3 G1-cyclins (CLN1, CLN2, and CLN3) and 5 B-type G2/M-cyclins (CLB1, CLB2, CLB4, CLB5, and CLB6). These two sets comprise all the CLN and CLB cyclins in the dataset (CLB3 was not present). The CLN and CLB cyclins were selected due to their important roles in cell-cycle regulation and relatively well-studied regulatory mechanisms. Figure 5.5 is a diagram produced by GNWD which features the selected genes. Each node represents a gene or a transcription factor, each directed edge represents a regulatory relation, and each edge label denotes the regulatory delay between two nodes. For example, SWI6 → CLN2 has a delay of 2 samples (i.e. 2×7 min/sample = 14 min). The network edges are color-coded such that a red edge represents known interaction based on location analysis and a blue edge represents an unknown relationship.

GNWD uncovers a network of 15 nodes with 30 edges. 21 out of the 30 edges (i.e 70%) have known regulatory relationships. The average model fitness is 67%. A tabulated output is provided in Table 5.3. In that table, the column "Order" means the order of the system which reflects the model complexity. The percent of fitness reflects the goodness-of-fit of the state-space model to the data. AIC is the Akaike's Information Criterion score. Among the novel regulatory relations determined, there is evidence to support SWI6 → CLN2 [16], FKH2 → CLB1 [17], NDD1 → FKH2 [18] regulation in the literature.

Regulation of CLN2

The modeling tool uncovered the regulatory relationship between SWI6 and CLN2 (with order=2 and delay=2) that is not reported in the location analysis results (see Appendix A). As mentioned in the previous section, SWI4 and SWI6 encode a heterodimer complex, SBF. It has been shown that SBF induces CLN2 transcription in the late G1 phase [16]. In our modeling, we detected the regulatory relations of SWI4 → CLN2 with a first order system (AIC score=-1.36), and SWI6 → CLN2 with a second order system (AIC score=-0.47) (see Table 5.3). The difference in the AIC score indicates that although both TFs contribute to the regulation of CLN2, SWI4 represents a better model to control CLN2 regulation than SWI6. One may postulate that SWI4 is the DNA binding component of the SBF complex and therefore it is the rate determining factor for the transcription of CLN2.

Using the SISO modeling, we demonstrated that SWI4 and SWI6 regulate CLN2 with input delays of 0 and 2, respectively. The corresponding model fitnesses are 65% and 61%. The log_2 predicted and measured CLN2 expression over 18 time-points is shown in Figure 5.6 a). We applied the MISO modeling tool in GNWD to this data in an attempt to improve the modeling of the CLN2

Regulator	Target	Order	Delay (τ)	Fitness(%)	AIC	Binding Evidence
FKH1	CLB2	2	0	73.465076	-2.510095	Y
FKH1	SWI4	2	0	71.941152	-3.016292	N
FKH2	CLB2	2	0	69.618154	-2.677518	Y
FKH2	SWI4	2	0	71.553277	-2.617587	N
FKH2	CLB1	2	0	71.09556	-2.485664	N
MBP1	SWI4	2	2	62.060827	-1.539307	Y
MBP1	CLB2	2	2	64.177229	-2.604613	Y
MBP1	CLN2	2	1	61.249674	-0.918211	Y
MBP1	CLB1	2	2	60.996793	-2.303097	N
MCM1	SWI4	2	1	67.744846	-2.430281	Y
MCM1	CLB2	2	2	64.483538	-2.042422	Y
NDD1	CLB2	2	2	73.178501	-1.628935	Y
NDD1	CLN1	2	2	60.400341	-1.758735	Y
NDD1	CLB6	2	0	71.960519	-1.687905	Y
NDD1	CLB5	2	0	65.499866	-2.475658	Y
NDD1	FKH2	2	2	72.445931	-3.800203	N
NDD1	CLN3	2	2	65.957132	-2.655417	N
SWI4	CLB2	2	2	68.538231	-1.872905	Y
SWI4	CLN3	2	1	60.289651	-3.029056	Y
SWI4	CLN2	1	0	64.906209	-1.360941	Y
SWI4	CLN1	1	0	65.22902	-2.237727	Y
SWI4	CLB6	2	0	73.243562	-2.039319	Y
SWI4	CLB5	2	0	75.557682	-2.937196	Y
SWI4	FKH2	2	1	68.484476	-3.41278	N
SWI4	CLB1	2	2	77.002957	-2.119164	N
SWI6	SWI4	2	0	70.893228	-1.993188	Y
SWI6	CLB2	2	2	63.223062	-1.786548	Y
SWI6	CLN2	2	2	61.366645	-0.473000	Y
SWI6	CLN1	2	2	61.383819	-1.329583	Y
SWI6	ACE2	2	0	62.457434	-1.956277	N

Table 5.3: GNWD output for yeast cyclins regulatory network

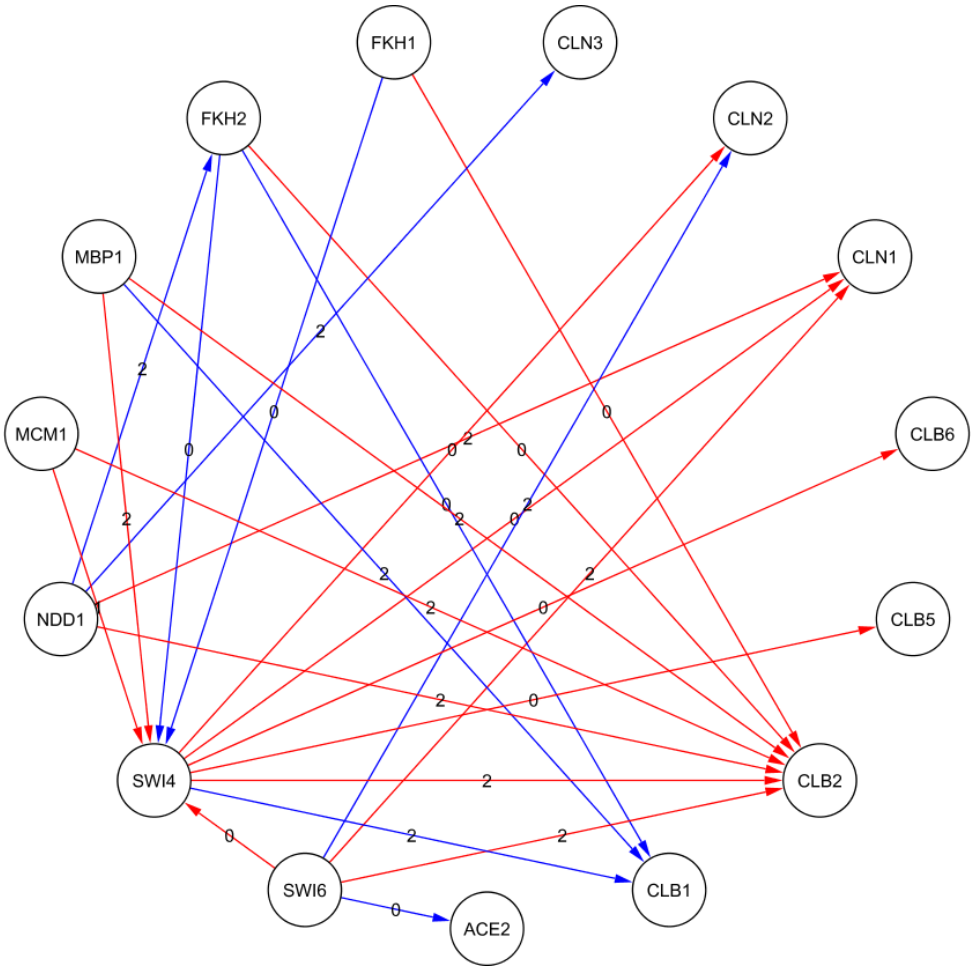


Figure 5.5: Gene regulatory network for the G1- and G2/M-cyclins. A red edge represents a known interaction based on location analysis and literature search; a blue edge represents an unknown relationship.

gene expression. Figure 5.6 b) illustrates a MISO model for SWI4+SWI6 → CLN2. The MISO modeling produces 3 possible models (see Table 5.4). The best-fitted model based on AIC score (noted with an asterisk) is a first-order system with fitness equal to 67%, delays $[\tau_{swi4}, \tau_{swi6}] = [0, 2]$. Compared to the previously mentioned 2 SISO models, the MISO model is relatively more stable (i.e. lower AIC score). However, the overall percent fitness did not increase as much as one would expected. These results suggest that the SWI4 and SWI6 do not regulate CLN2 transcription in a combined manner. In other words, the unexplained output variation of one regulator cannot be attributed to the other regulator.

Regulators (R1,R2) → Target	Order	R1 Delay	R2 Delay	Fitness(%)	AIC	Best Fit
(SWI4, SWI6) → CLN2	1	0	1	64.798724	-2.566028	
	1	0	2	66.922105	-2.889358	*
	2	0	0	67.240317	-0.892564	
	2	0	1	65.115198	-0.894492	

Table 5.4: MISO output for the CLN2 regulation.

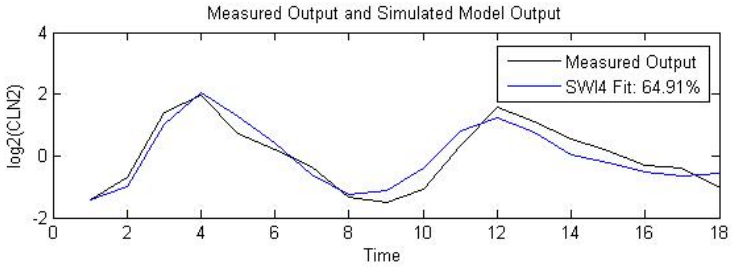
Regulation of CLB2

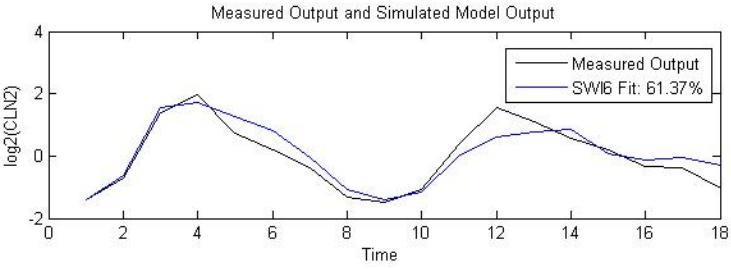
CLB2 encodes a B-type cyclin that activates the cyclin-dependent kinase, CDC28, to promote the transition from G2 to M phase of the cell cycle. The promoter region of CLB2 gene contains cis-element binding sites to 10 different transcription factors (refer to Figure 5.7) according to Harbison et al. [15]. The binding motifs are also confirmed by the ChIP-on-chip results (see Appendix A). Using the $p \leq 0.01$ cutoff, seven out of nine TFs (i.e. FKH1, FKH2, NDD1, MCM1, MBP1, SWI4, SWI6) show significant *in vivo* binding to CLB2.

The transcription factors that are found at the CLB2 promoter regions are known to regulate genes at different cell-cycle phases. For example, the SBF (SWI4,SWI6) and MBF (SWI6,MBP1) complexes promote G1 to S phase transition, MCM1 regulates late G2 and some M/G1 genes, and NDD1 functions at the G2/M phase [34]. Hence, it is unlikely that all binding factors are functional and are active at the same time. In our modeling (see Table 5.3), we detected regulatory relationships of the seven TFs to CLB2. Furthermore, a closer look at the regulation of CLB2 reveals four feed-forward-loop (FFL) network motifs (see Figure 5.5). A network motif is a biochemical wiring pattern that recurs throughout the transcriptional network. The feed-forward-loop is one of the most common network motifs found in the bacterium *Escherichia coli* and the yeast *Saccharomyces cerevisiae* [21]. A feed-forward-loop is a three-gene motif composed of two input transcription factors: a master and a secondary regulator. The master regulator regulates the secondary regulator and they both jointly regulate a target gene. We present the four FFLs found by the SISO modeling in Figure 5.5. The top-left node is the master node of the FFLs. They are FKH1, NDD1, MBP1 and MCM1. The top-right node is the secondary regulator and this is SWI4 except when the master node is NDD1 in which case the secondary regulator is FKH2. The average SISO model fitness for each TF → CLB2 regulation is 68%. All TFs except the forkhead TFs, FKH1 and FKH2, have delay of 2 sampling intervals. Among the four FFLs, MCM1+SWI4→CLB2 is also reported by Young et al. [23] as a feed-forward-loop using only the location analysis data with $p \leq 0.001$.

Mangan et al. [21] suggest that one important function of FFLs is to speed up the response time of the transcription networks. That is, although positive gene regulation can be efficiently achieved by increasing the concentration of the TF's protein product, the response time is governed

a)





b)

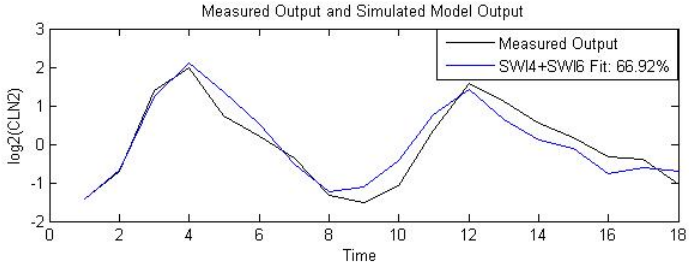


Figure 5.6: a) Two SISO models for SWI4 → CLN2 (fitness=65%) and SWI6 → CLN2 (fitness=61%); b) MISO model SWI4+SWI6 → CLN2 (fitness=67%).

by the lifetime of the protein product, which is often much longer. Therefore, one way to speed up the response is to increase the degradation rate of the protein product through a second regulator and perhaps to block access to the target gene's binding site by the first TF's protein product. Since the later regulator controls the expression of the former TF (the secondary regulator) and the target gene, it is called the master regulator. At the transcript level, one would expect the target gene expression level to be a function of the expression of both regulators should the FFL mechanism be functional.

We applied MISO modeling to the four FFL motifs identified by the SISO model for CLB2 regulation. We hypothesize that if a FFL is present, one would expect the master and secondary regulators to work in a collaborative manner. That is, the unexplained variation seen in the principal TF's regulation can be elucidated by the feed-forward regulation of the secondary TF, and vice versa. On the other hand, if the FFL is inactive or if only one of the two regulators works, then the modeling will not be improved by MISO modeling and the percent fitness of the model will remain roughly the same or be worse.

The output of the MISO modeling is tabulated in Table 5.6. The best-fitting model is marked

Figure 5.7: Promoter regions for CLB2 gene with CLB5 in close proximity. The cis-element binding sites are identified by Harbison et al. The figure is the results of a query given at the YeastGenome website [10].

with an asterisk in the rightmost column. The best model for FKH1+SWI4→CLB2 at ≈80% fitness is a first order system with zero time-delay for FKH1 and 2 time-delays for SWI4. The best model for MBP1+SWI4→CLB2 is a second order system with zero time-delay for MBP1 and 2 time-delays for SWI4. The fitness is ≈82%. We did not observe significant improvements in terms of percent fitness for the NDD1+FKH2 and MCM1+SWI4 models. This suggests that only the former two out of the four possible FFLs are likely to control CLB2 regulation. The log_2 of the predicted and measured CLB2 expression over 18 time-points by the four best-fitting FFL models are shown in Figure 5.8. There are obvious improvements in model fitness for the MBP1+SWI4 and FKH1+SWI4 models over the NDD1+FKH2 and MCM1+SWI4 models, which supports our hypothesis.

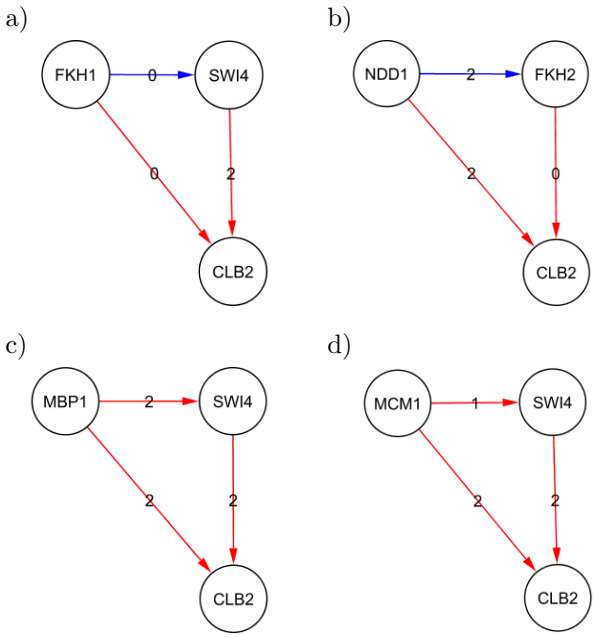


Table 5.5: Feed-forward-loop network motifs in the regulation of CLB2 found by GNWD. Each edge is labeled with the value of time delay. A red edge represents a known interaction based on location analysis and literature search; a blue edge represents an unknown relationship.

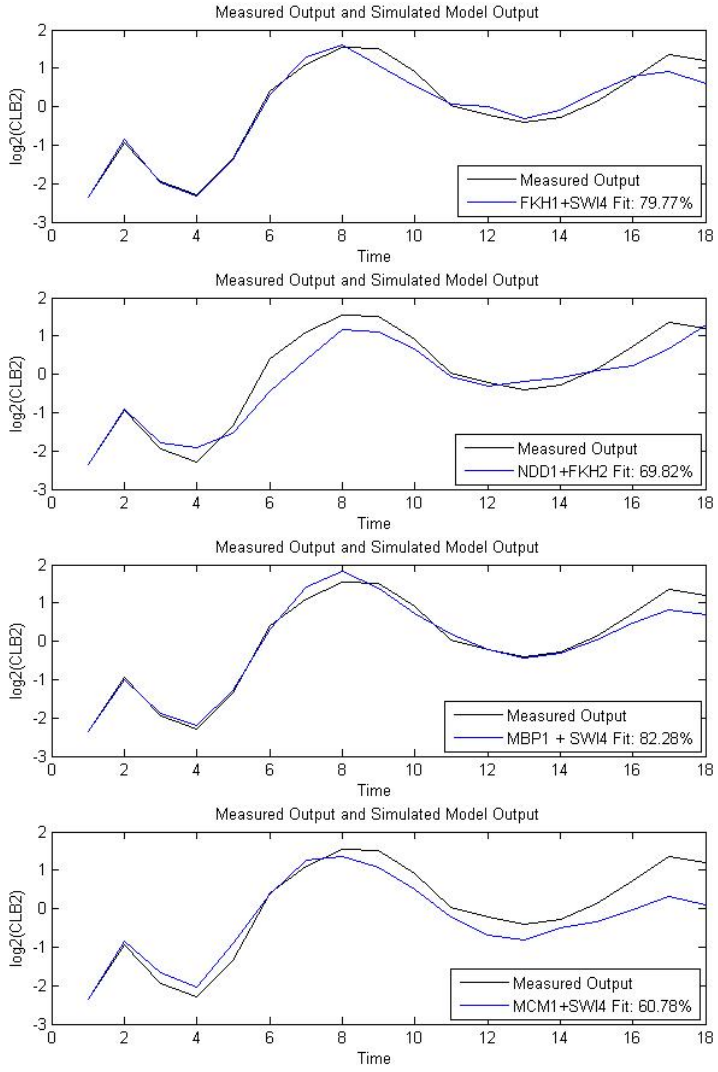


Figure 5.8: Comparisons of the predicted CLB2 expression over 18 time-points by the four best-fitting FFL models to the measured log_2 values.

Regulators (R1,R2) → Target	Order	R1 Delay	R2 Delay	Fitness(%)	AIC	Best Fit
FKH1,SWI4→CLB2	1	0	1	63.644575	-1.537761	
	1	0	2	79.773793	-3.229595	*
	1	1	0	61.814689	-1.510308	
	1	1	2	70.844626	-1.853543	
	2	0	0	71.974595	-1.759781	
	2	0	1	76.75773	-2.140303	
	2	0	2	75.445502	-2.533515	
	2	1	0	74.935231	-3.046253	
	2	1	1	74.816372	-1.912746	
NDD1,FKH2→CLB2	1	0	1	61.692141	-0.711095	
	1	0	2	69.823501	-2.268478	*
	1	1	0	60.652689	-0.637223	
	1	1	2	74.517609	-1.745955	
	1	2	0	72.64509	-1.793645	
	1	2	1	69.148994	-1.644347	
	2	0	0	74.392399	-0.98651	
	2	0	1	75.908715	-1.714605	
	2	0	2	78.764701	-1.941071	
	2	1	0	71.297965	-0.877083	
	2	1	1	74.106334	-1.081118	
	2	1	2	69.865951	-2.514543	
	2	2	0	70.654615	-1.453532	
	2	2	1	70.375696	-2.480809	
MBP1,SWI4→CLB2	1	1	0	65.281567	-1.531062	
	2	0	2	82.284591	-2.228897	*
MCM1,SWI4→CLB2	2	0	0	76.835382	-1.5779	
	2	0	1	78.146638	-1.707728	
	2	1	1	76.606035	-1.755346	
	2	1	2	60.319422	-1.722575	
	2	2	0	63.90925	-1.804213	
	2	2	1	65.281011	-1.704606	
	2	2	2	60.778085	-2.139843	*

Table 5.6: Multiple inputs, single output regulatory relations for CLB2.

Chapter 6

Discussion

6.1 Discrete versus Continuous Models

Yeast cell-cycle regulated genes demonstrate a periodic pattern [36]. The gene expressions are known to be phase specific. The expression data are reported as $log_2(\frac{sample\_expression}{reference\_expression})$. That is, one measures the changes in expression with respect to a common reference instead of absolute expression. A 2-fold change in expression, i.e. $log_2(ratio) = \pm 1$, is generally considered significant. It is important to note that a negative $log_2(ratio)$ does not imply inactivity of a regulator. Instead, it means that the gene expression level is relatively lower (by the fold change) compared to the control sample, e.g. the time zero sample.

Among the five state-space or Bayesian network solutions described in this work, the models published by Li et al. [20] and Sung et al. [37] are discrete. The challenge in discretization is to find a reasonable threshold to define the active and inactive states of gene expression. As described above, a negative measurement does not necessarily imply inactivity of a gene. Consider the following scenario: at time $t = 0$, gene A is expressed at 40% capacity, gene B is expressed at 10% capacity. At time $t = 1$, both genes A and B are expressed at 20% capacity. In this case, the time $t = 1$ measurements for gene A are $log_2(20\%/40\%)$ equals -1, and for gene B are $log_2(20\%/10\%)$ equals +1, whereas the actual levels of expression for both genes are the same at time $t = 1$. Therefore, it is a non-trivial task to define the "on" and "off" status of a gene based on gene expression fold-change alone. Soinov et al. [4] have proposed an alternative method (see Section 2.2) to bypass the assumption of arbitrary discretization thresholds for the regulators. The states of a "predicted gene" (i.e. a target gene) are determined by the quantitative expression levels (or changes in the expression with respect to a control sample) of the "explaining genes" (i.e. the regulators). The results are presented in the form of a rooted decision tree such that the states (up-/down- regulated, or expressed/not expressed) of a target gene (leaf node) is determined by the combinatorial decision rules of the regulators (non-leaf nodes). The Soinov approach can potentially improve the performance of discrete network modeling.

On the other hand, the biggest challenge in quantitative modeling is the inherent noise in the expression data. Especially when a gene is expressed at a low level, a low signal-to-noise ratio causes

an inaccurate measurement of fold-change. This will in turn affect the ability of quantitative models in learning the network structure and in getting good model fitness. In this study, the average model fitness for yeast expression data is 67%.

6.2 Gene Regulatory Network: what, when and how

A ChIP-on-chip experiment helps answer the question: what genes does a transcription factor regulate. The evidence of *in vivo* protein-DNA interactions can help biologists to uncover regulatory network structure [34, 19, 20]. However, the existence of a binding site does not mean that the regulation mechanism is triggered under a certain experiment condition. The location analysis results do not provide insights into which transcription factor(s) regulates a gene transcription under a perturbation or treatment when multiple cis-element binding sites are available at the promoter region. In yeast, a B-type cyclin, CLB2, is known to have cis-element binding sites for 10 different transcription factors (see Figure 5.7) according to Harbison et al. [15]. Many of these transcription factors are known to regulate genes at different cell-cycle phases. It is unlikely that all binding factors are functional at the same time. Our modeling tool provides a way to model the gene regulation based on time-course expression data. In this document, we analyzed 301 cell-cycle regulated genes with possible regulatory relationships to at least one of the nine known transcription factors. Among these, we are able to identify and model the regulation mechanisms of 93 ($\approx 31\%$) genes.

Peak time analysis provides insights into when a gene is maximally expressed during the cell-cycle. An understanding of the gene expression timelines is useful for associating a time factor to the physiological changes in cells. However, the duration for a gene to reach its peak expression (peak time) in a cell-cycle alone is not enough to constitute the full picture for gene regulation. For example, the transcription factor complex, SBF (SWI4 + SWI6), regulates CLN1 and CLN2 transcription in the late G1 phase and drives the transition into S phase. The peak times for the heterodimer SWI4 and SWI6 are 13% and 37%, respectively. The peak times for the SBF regulated genes CLN1 and CLN2 are 25% and 23%, respectively. One component of the SBF regulator, SWI6 reaches the peak time later than both CLN1 and CLN2. This shows that the peak time analysis does not convey information on how genes are regulated. One may hypothesize that SWI4 is the rate determining factor in the regulation of the cyclins and that the G1 cyclins will quickly reach their peak expressions at 25% after SWI4 reaches its peak at 13%. Our modeling results support the above mentioned assumption (refer to Section 5.2.3). The SWI4 and SWI6 transcription factors seem to regulate CLN2 transcription in a combinatorial manner. The percent fitness of the SWI4+SWI6→CLN2 model is no better than two separated single-input and single-output models. Interestingly, our modeling results also suggest that CLN2 is regulated by both

SWI4 and SWI6, and CLN1 is regulated only by SWI4. This could be the result of relatively weaker role of SWI6 in cyclin regulation.

6.3 Model Overfitting

GNWD uses location analysis results to help identify the TF and target gene pairs. This significantly reduces the risk of overfitting by filtering out the unrelated inputs (i.e. unwanted noise). In addition, Akaikes information criterion [1] scoring system is applied to the model selection process. The AIC method discourages the selection of a higher order system by imposing a penalty for the complexity of the estimated model. It attempts to find the best goodness-of-fit with a minimum system complexity. Therefore, this provides another guard against overfitting the data.

Chapter 7

Conclusions and Future Work

We have developed a new technique, GNWD, for determining prospective gene regulation models from time-series gene-expression data. The modeling tool is demonstrated on an artifical data and yeast cell-cycle gene-expression data. Using the yeast microarray data, we illustrated that our model can help identifying regulatory relations with multiple time delays. The model complements ChIP-on-chip results by predicting the most probable gene regulatory mechanisms between transcription factors and their target genes.

GNWD uses genome-wide location analysis data to reveal the primary network structure. Additional regulatory relationships can be determined by goodness-of-fit of the model based on alternate models. It will be interesting to compare this method to the learning-by-modification method developed by Sung et al. [37] where the network structure is based on a backward elimination mechanism. The current version of GNWD supports a command line console with no graphical user interface. Some features can be implemented to increase user friendliness. An example includes an interface to load multiple experiments. This involves implementing application logic for checking consistency of all input files. Another important facet of future work would be a systematic study of the effect of noise on state-space modeling.

References

[1] Akaike and Hirotugu. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] R.L. Bar-Or, R. Maya, L.A. Segel, U. Alon, et al. Generation of oscillations by the p53-mdm2 feedback loop: A theoretical and experimental study. *Proc. Natl. Acad. Sci. USA*, 97:11250–11255, 2000.

[3] N. Belacel, Q. Wang, and M. Cuperlovic-Culf. Clustering methods for microarray gene expression data. *OMICS: A Journal of Integrative Biology*, 10(4), 2006.

[4] A. Brazma, I. Jonassen, I. Eidhammer, and D.R. Gilbert. Approaches to the automatic discovery of patterns in biosequences. *Journal of Computational Biology*, 5(2):279–305, 1997.

[5] L.L. Breeden. Cyclin transcription: Timing is everything. *Current Biology*, 10:586–588, 2000.

[6] I. Chakravarti, R. Laha, and J. Roy. Handbook of methods of applied statistics. *John Wiley and Sons*, pages 392–394, 1967.

[7] T. Chen, H. He, and G. Church. Modeling gene expression with differential equations. *Pacific Symposium on Biocomputing*, 4:29–40, 1999.

[8] M.C. Costanzo, M.E. Crawford, J.E. Hirschman, J.E. Kranz, et al. Ypd, pombepd and wormpd: model organism volumes of the bioknowledge library, an integrated resource for protein information. *Nucleic Acids Research*, 29:75–79, 2001.

[9] KEGG Yeast Pathway Database. (http://www.kegg.org), February 2006.

[10] Saccharomyces Genome Database. (http://yeastgenome.org), March 2006.

[11] Stanford Microarray Database. (http://genome-www5.stanford.edu), June 2005.

[12] U. de Lichtenberg, L.J. Jensen, A. Fausboll, T.S. Jensen, P. Bork, and S. Brunak. Comparison of computational methods for the identification of cell cycle-regulated genes. *Bioinformatics*, 21(7):1164–1171, 2005.

[13] M.B. Eisen, P.T. Spellman, P.O. Browndagger, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95(25):14863–14868, 1998.

[14] W. Favoreel, B. De Moor, and P. Van Overschee. Subspace state space system identification for industrial processes. *Journal of Process Control*, 10(2), 2000.

[15] C.T. Harbison, D.B. Gordon, T.I. Lee, N.J. Rinaldi, K.D. Macisaac, T.W. Danford, N.M. Hannett, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431:99–104, 2004.

[16] V.R. Iyer, C.E. Horak, C.S. Scafe, D. Botstein, M. Snyder, and P.O. Brown. Genomic binding sites of the yeast cell-cycle transcription factors sbf and mbf. *Nature*, 409:533–538, 2001.

[17] P. Jorgensen and M. Tyers. The fork'ed path to mitosis. *Genome Biology*, 1(3):10221–10224, 2000.

[18] M. Koranda, A. Schleiffer1, L. Endler, and G. Ammerer. Forkhead-like transcription factors recruit ndd1 to the chromatin of g2/m-specific promoters. *Nature*, 406:94–98, 2000.

[19] T. Lee, N.J. Rinaldi, F. Robert, D.T. Odom, et al. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(25):799–804, 2002.

[20] Z. Li, S. M. Shaw, M.J. Yedwabrick, and C. Chan. Using a state-space model with hidden variables to infer transcription factor activities. *Bioinformatics*, 22(6):747–754, 2006.

[21] S. Mangan and U. Alon. Structure and function of the feed-forward loop network motif. *PNAS*, 100(21):11980–11985, 2003.

[22] M.D. Mendenhall and A.E. Hodge. Regulation of cdc28 cyclin-dependent protein kinase activity during the cell cycle of the yeast saccharomyces cerevisiae. *Microbiol Molecular Biology*, 62:1191–1243, 1998.

[23] ChIP on-chip experiment protocol at Richard A. Young's Lab. (http://staffa.wi.mit.edu/cgi-bin/young_public/navframe.cgi?s=14&f=chiparray), December 2006.

[24] K. Ota, T. Yamada, Y.Yamanishi, S. Goto, and M. Kanehisa. Comprehensive analysis of delay in transcriptional regulation using expression profiles. *Genome Informatics*, 14:302–303, 2003.

[25] P. Van Overschee and B. De Moor. N4sid:subspace algorithms for the identication of combined deteministic-stochastic systems. *Automatica*, 30(1):75–93, 1994.

[26] T. Pramila, W. Wu, S. Miles, W.S. Noble, and L.L. Breeden. The forkhead transcription factor hcm1 regulates chromosome segregation genes and fills the s-phase gap in the transcriptional circuitry of the cell cycle. *Genes and Development*, 20:2266–2278, 2006.

[27] M.F. Ramoni, P. Sebastianidagger, and I.S. Kohane. Cluster analysis of gene expression dynamics. *Proc. Natl. Acad. Sci. USA*, 99(14):2266–2278, 2002.

[28] C. Rangel, J. Angus, F. Falciani, Z. Ghahramani, et al. Modelling t-cell activation using gene expression profiling and state space models. *Bioinformatics*, 20:1361–1372, 2004.

[29] C. Rangel, J. Angus, Z. Ghahramani, and D.L. Wild. *Probabilistic Modelling in Bioinformatics and Medical Informatics*, chapter Modeling Genetic Regulatory Networks using Gene Expression Profiling and State Space Models. Springer-Verlag, in press.

[30] B. Ren, F. Robert, J.J. Wyrick, et al. Genome-wide location and function of dna binding proteins. *Science*, 290:2306–2309, 2000.

[31] O. Resendis-Antonio, J.A. Freyre-Gonzlez, R. Menchaca-Mndez, R.M. Gutirrez-Ros, et al. Modular analysis of the transcriptional regulatory network of e. coli. *Trends in Genetics*, 21:16–20, 2005.

[32] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[33] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga1, et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13:2498–2504, 2003.

[34] I. Simon, J. Barnett, N. Hannett, C.T. Harbison, et al. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, 106(6):697–708, 2001.

[35] R. Somogyi, S. Fuhrman, M. Askenazi, and A. Wuensche. The gene expression matrix: Towards the extraction of genetic network architectures. In *Proc. of Second World Cong. of Nonlinear Analysts (WCNA96)*, volume 30, pages 1815–1824, 1997.

[36] P.T. Spellman, G. Sherlock, M.Q. Zhang, V. R. Iyer, et al. Comprehensive identification of cell-cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297, 1998.

[37] W. Sung, T. Liu, and A. Mittal. Learning multi-time delay gene network using bayesian network framework. In *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence(ICTAI)*, 2004.

[38] F. Thibaud-Nissen, H. Wu, T. Richmond, J.C. Redman, et al. Development of arabidiopsis whole-genome microarrays and their application to the discovery of binding sites for the tga2 transcription factor in salicylic acid-treated plants. *The Plant Journal*, 47:152–162, 2006.

[39] E. Wang and E. Purisima. Network motifs are enriched with transcription factors whose transcripts have short half-lives. *Trends in Genetics*, 21(9):492–495, 2005.

[40] F. Wu, W. Zhang, and A. Kusalik. State-space model with time delays for genetic regulatory networks. *Journal of Biological Systems*, 2004.

Appendix A

ChIP-on-chip binding map for cell-cycle genes (p-value cut-off=0.01). "+" represents a significant binding of $p \leq 0.01$.

ORF	Symbol	Fkh1	Fkh2	Ndd1	Mcm1	Ace2	Swi5	Mbp1	Swi4	Swi6
YLR131C	ACE2	+	+	+	+					
YNR044W	AGA1				+			+	+	+
YGL032C	AGA2				+					
YCL025C	AGP1							+	+	
YGL021W	ALK1		+		+					
YNL172W	APC1	+								
YPR034W	ARP7	+								
YJL115W	ASF1							+		+
YKL185W	ASH1						+			
YML116W	ATR1		+	+						
YJR148W	BAT2		+	+		+	+		+	+
YPL255W	BBP1								+	+
YJR092W	BUD4	+	+	+	+					
YLR353W	BUD8	+								
YGR041W	BUD9	+	+		+	+	+	+	+	+
YML102W	CAC2							+		+
YPL111W	CAR1		+							
YLR438W	CAR2	+	+	+				+	+	+
YGR140W	CBF2								+	+
YGL116W	CDC20	+	+	+	+					
YOR074C	CDC21							+		+
YLR103C	CDC45		+					+		+
YLR274W	CDC46				+					
YMR001C	CDC5			+						
YJL194W	CDC6		+	+	+			+	+	+
YNL192W	CHS1						+			
YBR038W	CHS2			+						
YMR198W	CIK1	+	+							
YJL158C	CIS3	+	+	+		+		+	+	+
YPR119W	CLB2	+	+	+	+			+	+	+
YLR210W	CLB4	+								
YPR120C	CLB5							+	+	+
YGR109C	CLB6		+	+	+			+	+	+
YMR199W	CLN1	+	+	+		+		+	+	+
YPL256C	CLN2								+	+
YAL040C	CLN3				+	+	+		+	+
YMR078C	CTF18	+								
YLR286C	CTS1	+	+			+	+	+		+
YKL096W	CWP1		+			+			+	+
YJR048W	CYC1								+	
YGR092W	DBF2	+		+	+					
YML110C	DBI56				+					

ORF	Symbol	Fkh1	Fkh2	Ndd1	Mcm1	Ace2	Swi5	Mbp1	Swi4	Swi6
YDL101C	DUN1							+		+
YNL327W	EGT2					+	+			
YJL196C	ELO1		+	+	+			+	+	+
YLR056W	ERG3							+	+	+
YMR015C	ERG5	+	+						+	+
YDL018C	ERP3	+	+					+		+
YLR300W	EXG1	+	+	+		+	+	+	+	+
YDR261C	EXG2								+	
YOR317W	FAA1					+			+	
YIL009W	FAA3						+			
YJL157C	FAR1				+					
YKL182W	FAS1					+			+	
YER032W	FIR1	+	+							
YER145C	FTR1					+			+	+
YEL042W	GDA1				+					
YHR061C	GIC1	+	+					+	+	+
YDR309C	GIC2		+	+				+	+	+
YDR507C	GIN4				+			+	+	+
YLR342W	GLS1								+	+
YHR005C	GPA1				+					
YCR065W	HCM1		+	+	+			+	+	
YBR138C	HDR1	+	+	+						
YBR009C	HHF1		+						+	
YPL127C	HHO1							+	+	+
YBR010W	HHT1		+						+	
YDL227C	HO								+	+
YMR032W	HOF1				+					
YPL116W	HOS3	+								
YJL092W	HPR5									+
YBR133C	HSL7		+							
YJL159W	HSP150					+	+			+
YOR025W	HST3			+						
YDR191W	HST4				+					
YDR225W	HTA1								+	+
YBL003C	HTA2							+	+	
YOL012C	HTA3				+				+	+
YDR224C	HTB1								+	+
YBL002W	HTB2							+	+	
YHR094C	HXT1	+								
YDR342C	HXT7						+			
YPL242C	IQG1			+						
YIL026C	IRR1							+		+
YJL073W	JEM1							+		
YAR018C	KIN3		+	+	+					
YBL063W	KIP1	+								
YPL155C	KIP2		+							
YGL216W	KIP3				+					
YPR159W	KRE6							+	+	+
YKL103C	LAP4								+	
YJL134W	LCB3						+			
YDL003W	MCD1							+		+

ORF	Symbol	Fkh1	Fkh2	Ndd1	Mcm1	Ace2	Swi5	Mbp1	Swi4	Swi6
YKL165C	MCD4								+	
YBL023C	MCM2							+		
YEL032W	MCM3			+	+					
YGL201C	MCM6				+					
YNL173C	MDG1	+								
YNL328C	MDJ2					+	+			
YDR461W	MFA1				+					
YNL145W	MFA2	+	+	+	+		+			
YER001W	MNN1							+	+	+
YGR014W	MSB2								+	+
YDR097C	MSH6							+		
YPR149W	NCE102		+	+					+	+
YPL124W	NIP29								+	
YDR150W	NUM1			+	+					
YGL038C	OCH1							+	+	+
YGL055W	OLE1						+			
YPR075C	OPY2		+					+		+
YNL289W	PCL1		+	+	+	+		+	+	+
YDL127W	PCL2					+	+		+	+
IL050W	PCL7						+			
YDL179W	PCL9						+			
YNL231C	PDR16		+					+	+	+
YDR113C	PDS1	+	+					+		+
YMR076C	PDS5	+	+					+	+	+
YAR071W	PHO11	+								
YBR092C	PHO3		+		+					
YLR273C	PIG1				+					
YKL164C	PIR1				+		+			
YKL163W	PIR3				+		+			
YGL008C	PMA1		+	+	+	+			+	
YCR024C-A	PMP1	+	+	+						
YEL060C	PRB1					+				
YJL079C	PRY1		+	+	+	+			+	+
YKR013W	PRY2	+							+	+
YJL078C	PRY3	+	+			+	+	+		+
YDL055C	PSA1		+			+			+	+
YLR142W	PUT1								+	
YKL113C	RAD27							+		+
YER095W	RAD51	+	+					+		+
YNL312W	RFA2							+		
YER070W	RNR1	+	+					+	+	+
YGR152C	RSR1								+	
YBR070C	SAT2	+					+	+	+	
YHR205W	SCH9								+	
YMR305C	SCW10	+			+			+	+	+
YGL028C	SCW11	+	+			+	+			
YGR279C	SCW4		+	+					+	+
YIL076W	SEC28	+			+					
YHR098C	SFB3				+					
YLR079W	SIC1						+			
YIL123W	SIM1		+	+	+			+	+	+

ORF	Symbol	Fkh1	Fkh2	Ndd1	Mcm1	Ace2	Swi5	Mbp1	Swi4	Swi6
YGR143W	SKN1		+	+	+	+	+	+		+
YJL074C	SMC3							+		
YML058W	SML1			+						
YDR011W	SNQ2		+			+	+	+	+	+
YHR152W	SPO12		+	+	+	+		+		+
YMR179W	SPT21							+	+	+
YOR247W	SRL1		+	+	+	+	+	+	+	+
YIL140W	SRO4								+	
YMR183C	SSO2	+	+							
YFL026W	STE2				+					
YKL209C	STE6				+					
YDR297W	SUR2									+
YML052W	SUR7	+	+	+	+				+	
YGL162W	SUT1		+							
YPL032C	SVL3	+								
YPL163C	SVS1	+	+						+	+
YJL187C	SWE1	+	+					+	+	+
YER111C	SWI4				+			+	+	+
YDR146C	SWI5		+	+	+					
YBR083W	TEC1					+	+		+	
YGR099W	TEL2	+								
YML064C	TEM1	+	+				+			
YBR067C	TIP1				+					
YNL273W	TOF1							+		+
YML100W	TSL1	+	+	+		+	+	+	+	+
YOR075W	UFE1							+		+
YKR042W	UTH1		+	+	+	+	+	+	+	+
YEL040W	UTR2	+	+		+				+	+
YPL253C	VIK1	+	+		+	+	+	+	+	
YHL028W	WSC4	+		+		+	+			+
YAL022C	YAL022C		+			+	+	+	+	
YBL064C	YBL064C	+								
YBL111C	YBL111C						+		+	
YBL112C	YBL112C						+		+	
YBL113C	YBL113C						+		+	
YBR071W	YBR071W	+					+	+	+	
YBR139W	YBR139W	+	+	+						
YBR157C	YBR157C					+	+			
YBR158W	YBR158W	+	+	+		+	+		+	
YBR161W	YBR161W								+	
YCL024W	YCL024W							+	+	
YCL063W	YCL063W	+	+							+
YDR033W	YDR033W		+	+						
YDR055W	YDR055W						+			
YDR157W	YDR157W					+	+		+	
YDR190C	YDR190C				+					
YDR247W	YDR247W	+								
YDR307W	YDR307W		+	+	+			+		
YDR451C	YDR451C	+	+	+	+			+	+	+
YDR501W	YDR501W	+	+					+	+	+
YDR528W	YDR528W				+					

ORF	Symbol	Fkh1	Fkh2	Ndd1	Mcm1	Ace2	Swi5	Mbp1	Swi4	Swi6
YDR545W	YDR545W				+	+	+		+	+
YEL017W	YEL017W	+							+	+
YEL047C	YEL047C							+		
YEL077C	YEL077C						+	+		
YER124C	YER124C	+	+			+				
YER152C	YER152C					+				
YER189W	YER189W		+	+	+	+	+	+	+	+
YER190W	YER190W		+	+	+	+	+	+	+	+
YFL064C	YFL064C				+	+	+		+	+
YFL065C	YFL065C				+	+	+		+	+
YNL160W	YGP1						+			
YGR086C	YGR086C						+		+	
YGR151C	YGR151C								+	
YGR153W	YGR153W								+	
YGR189C	YGR189C				+	+	+	+	+	+
YGR221C	YGR221C								+	+
YGR230W	YGR230W			+						
YGR296W	YGR296W		+	+			+	+	+	+
YGR234W	YHB1					+	+		+	
YHR143W	YHR143W	+	+			+		+		
YHR149C	YHR149C				+			+	+	+
YHR151C	YHR151C		+	+	+	+		+		+
YIL056W	YIL056W		+	+		+		+	+	+
YIL121W	YIL121W								+	+
YIL122W	YIL122W				+				+	
YIL129C	YIL129C					+	+			
YIL141W	YIL141W								+	
YIL158W	YIL158W	+	+	+	+					
YIL177C	YIL177C						+			
YJL051W	YJL051W		+	+	+					
YJL225C	YJL225C		+			+	+		+	+
YJR030C	YJR030C							+		
YJR054W	YJR054W								+	+
YJR110W	YJR110W		+							
YKL008C	YKL008C		+						+	+
YKL044W	YKL044W		+				+		+	+
YKL052C	YKL052C		+						+	
YKL069W	YKL069W	+								
YKL151C	YKL151C					+		+		
YKR041W	YKR041W				+					+
YLL012W	YLL012W								+	
YLR013W	YLR013W						+			
YLR034C	YLR034C				+					
YLR049C	YLR049C						+	+		
YLR057W	YLR057W						+			
YLR084C	YLR084C		+	+	+				+	+
YLR154C	YLR154C		+							
YLR190W	YLR190W		+	+	+					
YLR194C	YLR194C						+			
YLR209C	YLR209C	+								
YLR302C	YLR302C				+					

ORF	Symbol	Fkh1	Fkh2	Ndd1	Mcm1	Ace2	Swi5	Mbp1	Swi4	Swi6
YLR380W	YLR380W							+		+
YLR437C	YLR437C	+	+	+				+	+	+
YLR462W	YLR462W			+		+	+		+	+
YLR463C	YLR463C					+	+	+	+	+
YLR464W	YLR464W			+		+	+		+	+
YLR465C	YLR465C					+	+	+	+	+
YLR466W	YLR466W			+		+	+		+	+
YLR467W	YLR467W					+	+	+	+	+
YML050W	YML050W		+	+	+					
YML125C	YML125C	+	+	+		+	+			
YML133C	YML133C						+			
YMR002W	YMR002W			+						
YMR031C	YMR031C				+					
YMR144W	YMR144W	+	+	+	+		+	+	+	+
YMR145C	YMR145C		+	+					+	
YMR163C	YMR163C		+							
YMR215W	YMR215W	+	+					+		+
YMR253C	YMR253C				+					
YNL056W	YNL056W			+	+					
YNL058C	YNL058C			+	+					
YNL078W	YNL078W					+	+	+		
YNL134C	YNL134C			+						
YNL176C	YNL176C	+	+							
YNL300W	YNL300W								+	+
YNL339C	YNL339C	+		+	+	+	+		+	+
YNR009W	YNR009W		+					+	+	+
YOL011W	YOL011W				+				+	+
YOL019W	YOL019W								+	
YOL030W	YOL030W	+								
YOL114C	YOL114C		+						+	+
YOR023C	YOR023C			+						
YOR066W	YOR066W				+			+		
YOR073W	YOR073W		+							
YOR114W	YOR114W		+							
YOR248W	YOR248W		+	+	+	+	+	+	+	+
YOR264W	YOR264W						+			
YOR273C	YOR273C		+			+	+	+	+	
YOR283W	YOR283W	+								
YOR315W	YOR315W	+	+	+	+	+			+	+
YOR372C	YOR372C		+					+	+	+
YML027W	YOX1		+					+	+	+
YPL025C	YPL025C		+			+	+		+	+
YPL141C	YPL141C	+	+	+						
YPL158C	YPL158C						+			
YPL250C	YPL250C		+							
YPL267W	YPL267W								+	
YPL283C	YPL283C	+	+	+		+	+	+	+	+
YPR013C	YPR013C	+							+	
YPR202W	YPR202W						+			
YPR203W	YPR203W						+			
YLR121C	YPS4								+	

ORF	Symbol	Fkh1	Fkh2	Ndd1	Mcm1	Ace2	Swi5	Mbp1	Swi4	Swi6
YML109W	ZDS2				+					

Table A.1: ChIP-on-chip binding map for cell-cycle genes (p-value cut-off=0.01). "+" represents a significant binding of $p \leq 0.01$.

Appendix B

GNWD output for the 301 yeast cell-cycle regulated genes.

Regulator	Target	Order	Delay (τ)	Fitness(%)	AIC	Target Gene Annotation
ACE2	EGT2	2	0	60.048369	-0.222038	cell-cycle regulation protein
ACE2	CTS1	2	1	74.057228	-1.51966	endochitinase
ACE2	SCW11	2	2	77.470499	-1.543409	similarity to glucanase
ACE2	SIC1	2	0	60.496078	-1.796579	p40 inhibitor of CDC28P-Clb protein kinase complex
ACE2	BUD9	2	0	62.995657	-2.164896	budding protein
ACE2	TSL1	2	0	74.038731	-2.612199	alpha,alpha-trehalose-phosphate synthase, 123 KD subunit
ACE2	HSP150	2	0	82.06412	-2.98018	member of the PIR1P/HSP150P/PIR3P family
FKH1	CLB2	2	0	72.39239	-2.204547	cyclin, G2/M-specific
FKH1	CLB2	2	0	72.39239	-2.204547	cyclin, G2/M-specific
FKH1	YDR451C	2	0	60.357144	-2.205495	strong similarity to YOX1P
FKH1	PHO11	2	2	71.621892	-2.324851	secreted acid phosphatase
FKH1	TSL1	2	2	73.857862	-2.711535	alpha,alpha-trehalose-phosphate synthase, 123 KD subunit
FKH1	CDC46	2	2	73.977152	-2.889637	cell division control protein
FKH1	BUD4	1	0	69.379495	-2.895427	budding protein
FKH1	PDS1	2	2	62.352019	-3.04957	cell cycle regulator
FKH1	YPL141C	2	2	65.449936	-3.133143	strong similarity to protein kinase KIN4P
FKH1	YIL158W	2	0	75.123807	-3.288392	similarity to hypothetical protein YKR100c
FKH1	YOL030W	2	0	69.174669	-3.40947	strong similarity to glycoprotein GAS1P
FKH1	SVL3	2	0	75.388873	-3.916309	strong similarity to PAM1P
FKH1	YCL063W	2	0	80.547973	-5.480048	weak similarity to yeast translation regulator GCD6P
FKH2	RNR1	2	0	65.5414	-1.784932	ribonucleoside-diphosphate reductase, large subunit
FKH2	YMR215W	1	0	64.573384	-1.810287	similarity to GAS1 protein
FKH2	HHF1	2	0	69.398648	-1.877042	histone H4
FKH2	YNL058C	2	0	71.347029	-2.07159	similarity to YIL117c
FKH2	YJL051W	2	0	60.72833	-2.142167	hypothetical protein
FKH2	YPL141C	1	0	61.797678	-2.309275	strong similarity to protein kinase KIN4P
FKH2	BUD4	1	0	63.867755	-2.763157	budding protein
FKH2	CIK1	1	0	65.480881	-2.793112	spindle pole body associated protein
FKH2	CDC20	2	0	63.715931	-3.07063	cell division control protein
FKH2	ALK1	2	0	76.58263	-3.407463	DNA damage-responsive protein
FKH2	KIP2	1	0	70.840209	-3.616939	kinesin-related protein

Regulator	Target	Order	Delay (τ)	Fitness(%)	AIC	Target Gene Annotation
FKH2	YIL158W	2	0	66.465175	-4.027983	similarity to hypothetical protein YKR100c
FKH2	YMR144W	2	2	64.549466	-4.156344	weak similarity to MLP1P
FKH2	YCL063W	2	0	69.377645	-4.412188	weak similarity to yeast translation regulator GCD6P
FKH2	ATR1	2	0	71.899359	-4.63104	aminotriazole and 4-nitroquinoline resistance protein
MBP1	CLN1	2	0	62.303104	-1.216541	cyclin, G1/S-specific
MBP1	CLN1	2	0	62.303104	-1.216541	cyclin, G1/S-specific
MBP1	AGA1	2	0	71.835769	-1.318709	a-agglutinin anchor subunit
MBP1	YOR248W	2	0	60.578606	-1.510189	hypothetical protein
MBP1	HTB2	2	2	76.535451	-1.635653	histone H2B.2
MBP1	HTA2	2	2	74.292535	-1.734255	histone H2A.2
MBP1	SPT21	2	0	60.781543	-1.776206	required for normal transcription at a number of loci
MBP1	CDC21	2	0	61.171301	-1.973526	thymidylate synthase
MBP1	YMR215W	2	1	73.521772	-2.12566	similarity to GAS1 protein
MBP1	SWI4	2	1	61.982663	-2.203049	transcription factor
MBP1	HHO1	2	0	79.438238	-2.284701	histone H1 protein
MBP1	RNR1	2	0	64.277429	-2.399333	ribonucleoside-diphosphate reductase, large subunit
MBP1	SPK1	2	0	68.07085	-2.439243	ser/thr/tyr protein kinase
MBP1	RFA2	2	0	66.838408	-2.499096	DNA replication factor A, 36 kDa subunit
MBP1	YDR528W	2	2	64.196024	-2.685099	similarity to LRE1P
MBP1	RFA1	2	2	66.192852	-2.76145	DNA replication factor A, 69 KD subunit
MBP1	SMC3	2	2	73.012981	-2.821601	required for structural maintenance of chromosomes
MBP1	ERP3	2	0	70.477764	-3.042157	weak similarity to DEP1P
MBP1	TSL1	2	2	75.47332	-3.365967	alpha,alpha-trehalose-phosphate synthase, 123 KD subunit
MBP1	PDS1	2	0	71.810516	-3.499292	cell cycle regulator
MBP1	YMR144W	2	0	65.516481	-3.702367	weak similarity to MLP1P
MCM1	MFA2	2	0	62.25064	-1.016734	mating pheromone a-factor 2
MCM1	AGA1	2	0	82.52144	-2.040395	a-agglutinin anchor subunit
MCM1	YLR190W	2	2	65.741133	-2.076357	hypothetical protein
MCM1	UTR2	2	2	62.777288	-2.307552	cell wall protein
MCM1	ALK1	2	2	69.804709	-2.527855	DNA damage-responsive protein
MCM1	SWI5	2	2	70.769996	-2.54073	transcription factor
MCM1	SWI5	2	2	70.769996	-2.54073	transcription factor
MCM1	HSP150	2	2	64.039383	-2.559155	member of the PIR1P/HSP150P/PIR3P family
MCM1	AGA2	2	1	71.048026	-2.665176	a-agglutinin binding subunit
MCM1	GPA1	2	0	68.608262	-2.76755	GTP-binding protein alpha subunit of the pheromone pathway
MCM1	SWI4	2	0	70.097117	-2.844173	transcription factor
MCM1	YIL158W	2	0	65.630147	-2.874086	similarity to hypothetical protein YKR100c
MCM1	CLN3	2	0	61.53676	-2.910408	cyclin, G1/S-specific

Regulator	Target	Order	Delay (τ)	Fitness(%)	AIC	Target Gene Annotation
MCM1	PIG1	2	1	65.193977	-2.985317	putative type 1 phosphatase regulatory subunit
MCM1	BUD4	2	2	66.609154	-3.047091	budding protein
MCM1	YMR031C	2	2	63.178041	-3.592597	similarity to YKL050c and human restin
NDD1	YDR033W	2	1	70.387661	-1.786533	membrane protein related to HSP30P
NDD1	CIS3	2	1	60.729477	-1.845925	strong similarity to PIR1P/HSP150P/PIR3P family
NDD1	YNL058C	2	2	70.272817	-2.07121	similarity to YIL117c
NDD1	MFA2	2	2	67.18557	-2.219689	mating pheromone a-factor 2
NDD1	SML1	2	2	64.75619	-2.313915	protein inhibitor of ribonucleotide reductase
NDD1	CDC5	2	0	74.117685	-2.39871	involved in regulation of DNA replication
NDD1	HST3	2	2	66.239645	-2.429964	silencing protein
NDD1	YIL158W	2	2	66.569236	-2.510331	similarity to hypothetical protein YKR100c
NDD1	KIN3	2	0	62.392821	-2.543515	ser/thr protein kinase
NDD1	CLB2	2	0	75.143033	-2.689671	cyclin, G2/M-specific
NDD1	CLB2	2	0	75.143033	-2.689671	cyclin, G2/M-specific
NDD1	YLR190W	2	0	71.227838	-2.784225	hypothetical protein
NDD1	YMR144W	2	2	60.885597	-2.849805	weak similarity to MLP1P
NDD1	ALK1	2	0	78.889112	-2.91632	DNA damage-responsive protein
NDD1	PRY1	2	0	81.049047	-2.988546	strong similarity to the plant PR-1 class of pathogen related proteins
NDD1	SPO12	2	2	62.737082	-3.051609	sporulation protein
NDD1	IQG1	2	0	76.440366	-3.219517	involved in cytokinesis, has similarity to mammalian IQGAP proteins
NDD1	BUD4	2	0	78.065775	-3.647724	budding protein
NDD1	YPL141C	2	0	78.498739	-3.723794	strong similarity to protein kinase KIN4P
NDD1	FIR1	2	1	77.823827	-3.805727	interacts with the poly(A) polymerase in the two hybrid system
SWI4	SVS1	1	0	70.537326	-1.451141	vanadate sensitive suppressor
SWI4	MNN1	1	0	62.313402	-1.47927	alpha-1,3-mannosyltransferase
SWI4	CLB2	2	2	62.034039	-1.705976	cyclin, G2/M-specific
SWI4	CLB2	2	2	62.034039	-1.705976	cyclin, G2/M-specific
SWI4	HTB2	1	0	60.978756	-1.767544	histone H2B.2
SWI4	HTA1	1	0	71.809078	-1.845717	histone H2A
SWI4	YNL300W	1	0	60.412651	-1.978269	similarity to MID2P
SWI4	YOX1	2	0	70.553591	-2.048365	homoeodomain protein
SWI4	CLN1	1	0	69.937617	-2.048885	cyclin, G1/S-specific
SWI4	CLN1	1	0	69.937617	-2.048885	cyclin, G1/S-specific
SWI4	YGR189C	1	0	64.685005	-2.06787	family of putative glycosidases might exert a common role in cell wall organization
SWI4	YNR009W	2	2	61.806464	-2.108983	hypothetical protein
SWI4	CLB6	2	0	74.38724	-2.122882	cyclin, B-type
SWI4	CLB6	2	0	74.38724	-2.122882	cyclin, B-type

Regulator	Target	Order	Delay (τ)	Fitness(%)	AIC	Target Gene Annotation
SWI4	HTB1	1	1	69.795229	-2.14667	histone H2B
SWI4	PRY2	2	0	74.653417	-2.171415	similarity to the plant PR-1 class of pathogen related proteins
SWI4	SRO4	1	0	61.569785	-2.297586	required for axial pattern of budding
SWI4	GIN4	1	0	64.394668	-2.387945	ser/thr protein kinase
SWI4	YPL267W	2	0	64.490964	-2.436754	weak similarity to C.elegans transcription factor unc-86
SWI4	YGR086C	2	2	69.700726	-2.450202	strong similarity to hypothetical protein YPL004c
SWI4	CLN2	1	0	64.809099	-2.451147	cyclin, G1/S-specific
SWI4	CLN2	1	0	64.809099	-2.451147	cyclin, G1/S-specific
SWI4	SPT21	1	0	68.100673	-2.588911	required for normal transcription at a number of loci
SWI4	YIL141W	2	1	66.841842	-2.6399	questionable ORF
SWI4	RNR1	1	0	73.66253	-2.6472	ribonucleoside-diphosphate reductase, large subunit
SWI4	SIM1	2	0	61.646892	-3.004911	involved in cell cycle regulation and aging
SWI4	YHR149C	1	0	65.645714	-3.156839	similarity to hypothetical protein YGR221c
SWI4	CLB5	1	0	67.09113	-3.160326	cyclin, B-type
SWI4	CLB5	1	0	67.09113	-3.160326	cyclin, B-type
SWI4	BBP1	2	2	70.427343	-3.219215	cell division control protein
SWI4	YGR151C	1	0	67.022668	-3.284654	questionable ORF
SWI4	RSR1	1	0	68.464644	-3.67413	GTP-binding protein
SWI4	FTR1	2	2	78.334865	-4.277765	iron permease that mediates high-affinity iron uptake
SWI5	CTS1	2	0	76.022162	-1.394755	endochitinase
SWI5	YGR086C	2	1	60.292148	-2.182385	strong similarity to hypothetical protein YPL004c
SWI5	YGR189C	2	1	63.120445	-2.193886	family of putative glycosidases might exert a common role in cell wall organization
SWI5	HSP150	2	2	72.965477	-2.245656	member of the PIR1P/HSP150P/PIR3P family
SWI5	YIL177C	2	2	62.49735	-2.872546	strong similarity to subtelomeric encoded proteins
SWI5	YGR296W	2	1	66.368044	-2.917391	strong similarity to YPL283c; YNL339c and other Y' encoded proteins
SWI5	YHB1	2	1	68.790802	-3.062183	flavohemoglobin
SWI6	MNN1	2	2	67.001767	-1.269696	alpha-1,3-mannosyltransferase
SWI6	YNL300W	2	1	66.307379	-1.507283	similarity to MID2P
SWI6	CLN1	2	2	62.272008	-1.548075	cyclin, G1/S-specific
SWI6	CLN1	2	2	62.272008	-1.548075	cyclin, G1/S-specific
SWI6	YMR215W	2	2	67.580958	-1.913423	similarity to GAS1 protein
SWI6	MCD1	2	2	63.575917	-1.939164	Mitotic Chromosome Determinant
SWI6	GIN4	2	0	61.960682	-1.960046	ser/thr protein kinase
SWI6	SWI4	2	0	70.893228	-1.993188	transcription factor

Regulator	Target	Order	Delay (τ)	Fitness(%)	AIC	Target Gene Annotation
SWI6	PRY2	2	2	69.158446	-2.001919	similarity to the plant PR-1 class of pathogen related proteins
SWI6	RNR1	2	0	68.161451	-2.009178	ribonucleoside-diphosphate reductase, large subunit
SWI6	MSH6	2	0	60.965749	-2.022091	DNA mismatch repair protein
SWI6	SVS1	2	2	71.71493	-2.060424	vanadate sensitive suppressor
SWI6	RAD27	2	0	61.087315	-2.342807	ssDNA endonuclease and 5'-3'exonuclease
SWI6	ASF1	2	2	64.062118	-2.493437	anti-silencing protein
SWI6	CLB2	2	2	62.297831	-2.496261	cyclin, G2/M-specific
SWI6	CLB2	2	2	62.297831	-2.496261	cyclin, G2/M-specific
SWI6	SPK1	2	0	71.56099	-2.579225	ser/thr/tyr protein kinase
SWI6	SMC3	2	2	70.020246	-2.742052	required for structural maintenance of chromosomes
SWI6	YPL267W	2	2	61.365374	-2.781722	weak similarity to C.elegans transcription factor unc-86
SWI6	HHO1	2	2	78.707774	-2.933763	histone H1 protein
SWI6	PDS1	2	0	70.702625	-3.974437	cell cycle regulator

Table B.1: GNWD output for the 301 yeast cell-cycle regulated genes.