BMC Bioinformatics

Methodology article

Dynamic modeling of cis-regulatory circuits and gene expression prediction via cross-gene identification

Li-Hsieh Lin†1, Hsiao-Ching Lee†2, Wen-Hsiung Li3,4 and Bor-Sen Chen*1

Address: 1Lab. of System Biology, National Tsing Hua University, 101, Sec 2, Kuang Fu Road, Hsinchu, 300, Taiwan., 2Department of Life Science & Institute of Bioinformatics and Structural Biology, National Tsing Hua University, Hsinchu, 300, Taiwan., 3Department of Ecology and Evolution, University of Chicago, USA. and 4Genomics Research Center, Academia Sinica, Taipei, Taiwan.

Email: Li-Hsieh Lin - LHlin@moti.ee.nthu.edu.tw; Hsiao-Ching Lee - d884234@life.nthu.edu.tw; Wen-Hsiung Li - whli@uchicago.edu; Bor-Sen Chen* - bschen@ee.nthu.edu.tw

* Corresponding author †Equal contributors

Published: 18 October 2005

BMC Bioinformatics 2005, **6**:258 doi:10.1186/1471-2105-6-258

Received: 19 March 2005
Accepted: 18 October 2005

This article is available from: http://www.biomedcentral.com/1471-2105/6/258

Abstract

Background: Gene expression programs depend on recognition of *cis* elements in promoter region of target genes by transcription factors (TFs), but how TFs regulate gene expression via recognition of *cis* elements is still not clear. To study this issue, we define the *cis*-regulatory circuit of a gene as a system that consists of its *cis* elements and the interactions among their recognizing TFs and develop a dynamic model to study the functional architecture and dynamics of the circuit. This is in contrast to traditional approaches where a *cis*-regulatory circuit is constructed by a mutagenesis or motif-deletion scheme. We estimate the regulatory functions of *cis*-regulatory circuits using microarray data.

Results: A novel cross-gene identification scheme is proposed to infer how multiple TFs coordinate to regulate gene transcription in the yeast cell cycle and to uncover hidden regulatory functions of a *cis*-regulatory circuit. Some advantages of this approach over most current methods are that it is based on data obtained from intact *cis*-regulatory circuits and that a dynamic model can quantitatively characterize the regulatory function of each TF and the interactions among the TFs. Our method may also be applicable to other genes if their expression profiles have been examined for a sufficiently long time.

Conclusion: In this study, we have developed a dynamic model to reconstruct *cis*-regulatory circuits and a cross-gene identification scheme to estimate the regulatory functions of the TFs that control the regulation of the genes under study. We have applied this method to cell cycle genes because the available expression profiles for these genes are long enough. Our method not only can quantify the regulatory strengths and synergy of the TFs but also can predict the expression profile of any gene having a subset of the *cis* elements studied.

Background

Cis-regulatory circuits have been applied to model cell cycle control and other developmental processes [1,2]. Recently, the genetic regulatory and transcriptional networks of yeast have been studied [3-8]. However, in order to understand the regulation of any particular gene in the network, one should study carefully how the genes in the network collectively operate with remarkable precision in

response to environmental cues and the structure and function of the *cis*-regulatory circuit of the gene [7]. The *cis*-regulatory circuit of a gene consists of its *cis* elements, i.e., binding motifs of transcription factor (TF), and the interactions among their recognizing TFs. The *cis* elements of a gene can be considered as the information processing units in the regulatory circuit; they receive multiple inputs from the TFs that bind the *cis* elements of the gene. The output is the instruction for the transcription apparatus to determine whether the gene is to be expressed at a specific rate or to be repressed [9,10].

A *cis*-regulatory circuit may be regarded as a control device that is called into play by the TFs that have target sites inside the promoter [14-20]. A well-known example is the promoter region of the developmentally regulated *endo* 16 gene of the sea urchin [12-14]. It is about ~2300 bp in length and consists of several clusters of target sites for distinct functions. Yuh *et al.* [12-14] have explored the function of each subregion of the *endo* 16 system and every target site within each subregion, using a *cis*-regulatory logic model.

However, a drawback of most current methods for inferring *cis*-regulatory circuits is that they rely on changing or deleting some binding site sequences (e.g., [12-14]), which may not provide intact functional information for reconstructing the *cis*-regulatory circuit. The deletion or mutation experiments may change or destroy the original *cis*-regulatory circuit structure. Using such data, one may lose significant interactions among transcription factors (TFs). Obviously, it is appealing to develop a method that can infer intact *cis*-regulatory circuits. Recently, there are some statistical and system level approaches to study the genome-wide transcription regulation and address cooperativity among TFs ([7-11,15,16]). Important advances have been made toward understanding transcriptional regulatory networks. One strategy infers global networks directly from whole genome microarray data, and another strategy focuses on the identification of shared *cis* elements in the promoters of co-regulated genes, signified by similar expression profiles [8]. In this paper we develop a new method to combine microarray data and TF-binding location data by chromatin immunoprecipitation [5,18] to study the regulatory and interaction functions of various *cis* elements with regard to the target gene. The data of Lee *et al.* 2002 [5] can reveal the *in vivo* physical interactions of TFs with their *cis* elements on the promoter and therefore can provide a more reliable view of functional interactions between TFs and *cis* elements. Combining these types of data and microarray data, we propose a novel cross-gene identification scheme to infer how multiple TFs coordinate to regulate gene transcription. Our approach is rather different from most existing statistical and system level methods for analyzing gene expression data. Our results show that this novel method is suitable for deciphering the complex TFs interactions and for predicting the gene expression. In addition, we also identify the dynamic regulatory functions of TFs interaction in the yeast cell cycle, which cannot be achieved by current methods.

Results

Characterizing the cis-regulatory circuit of a gene

In this study, there are two steps for characterizing the cis-regulatory circuit of a gene. The first is to find a cluster of genes that includes the gene of interest and a number of other genes each of which shares a subset of cis elements with the gene of interest. Assuming a certain regulatory function for each of the TFs that recognize the cis elements of genes in the cluster and certain interaction functions among the TFs of a circuit, we set up dynamic equations for the cis-regulatory circuits of the genes in the cluster to describe their expression profiles. Since each gene in the cluster shares some cis elements with the gene of interest, a matrix of cis elements for the cluster of genes can be constructed. In this model, the regulatory functions of individual TFs and the interactions among TFs can be estimated from microarray data. In the second step, a cross-gene identification scheme is developed with an array of expression profiles of genes in the cluster (e.g., Spellman et al., 1998 [18]) to identify regulatory functions of the TFs and their possible interactions for each gene in the cluster; the parameters are estimated by the least square estimation algorithm.

Finally, plugging these estimated regulatory functions and interactions into the dynamic equations, one can explicitly describe the *cis*-regulatory circuit of the gene of interest.

Choice of a cluster of genes

As an illustration, suppose some genes of the yeast cell cycle are of interest. We find a cluster of genes for each gene of interest according to their cis elements found in Simon et al. 2001 [4]. To simplify the analysis, we consider only the nine TFs that are currently known to be important cell cycle TFs of the yeast (i.e., Mbp1, Swi4, Swi6, Mcm1, Fkh1, Fkh2, Ndd1, Swi5, and Ace2). The cluster of genes for the gene of interest is called the reference gene cluster (RGC). In an RGC, we assume that each gene shares some cis elements of the gene of interest. Furthermore, the regulatory functions and the interactions of the TFs recognizing the same cis elements are assumed to be the same for all genes in the RGC. For example, in Figure 1a gene MFA2 is the gene of interest; it causes cell cycle arrest and is essential for mating in yeast [19]. This gene has three main cis elements, Ndd1, Mcm1 and Swi5 [4], from which we want to reconstruct the cis-regulatory cir-

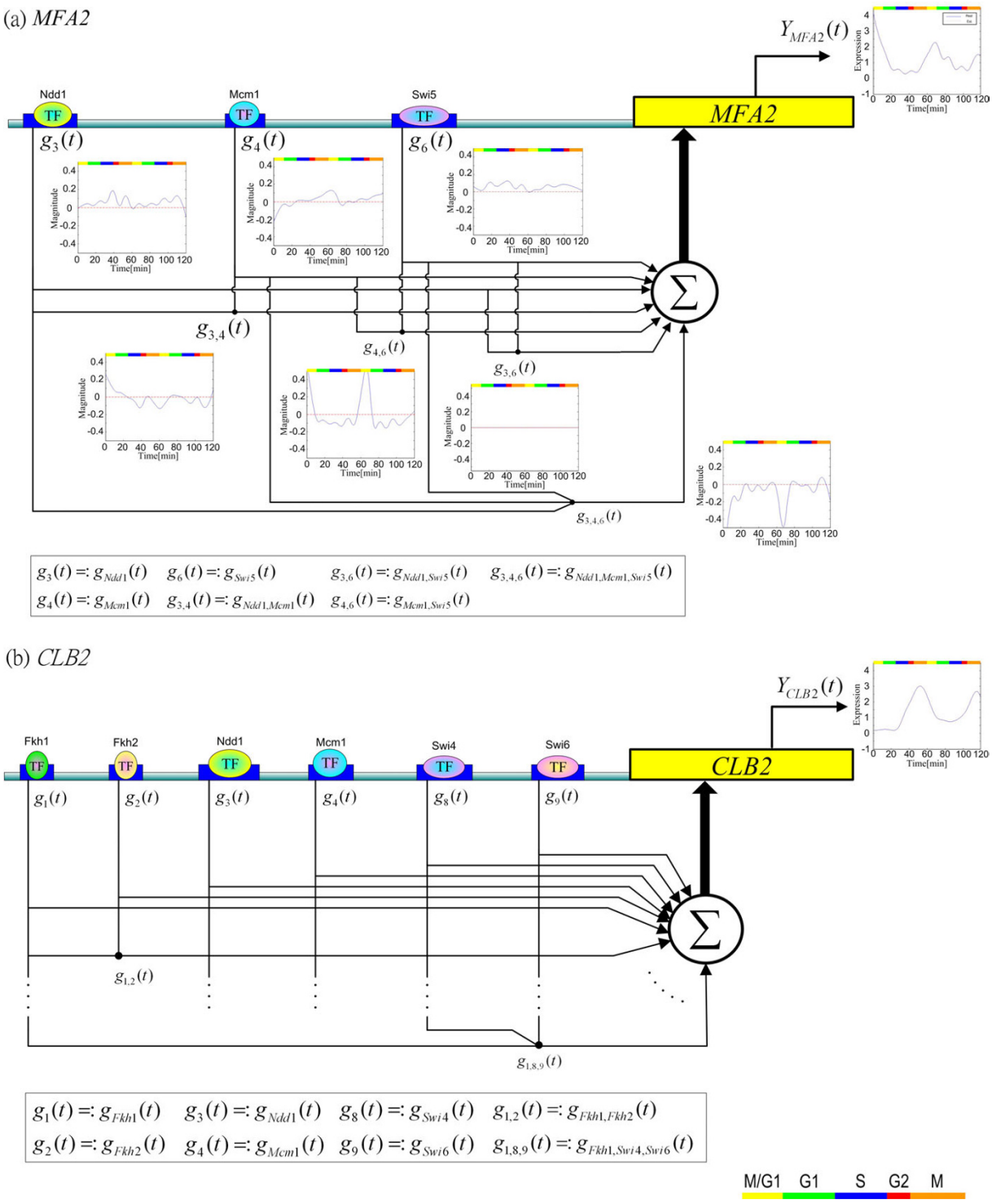

Figure 1
Dynamic model of the *cis*-regulatory circuit of gene *MFA2* (a) and of gene *CLB2* (b). The genome-wide TF-binding location data obtained using chromatin immunoprecipitation [4] is used to identify the transcription factor binding motifs (*cis* elements). A binding transcription factor p has a regulatory function $g_p(t)$ and interacts with other recognizing TFs to produce the regulatory functions $g_{p,q}(t)$ and $g_{p,q,r}(t)$. These regulatory functions are the inputs of the *cis*-regulatory circuit and generate the dynamic output (i.e., the expression profile) of the target gene. Different phases of the cell cycle are indicated by the colored bar at the right lower corner.

Table 1: The reference gene clusters (RGCs) of *MFA2*. Target Gene *MFA2* and the connectivities to *cis* elements.

	Fkh1	Fkh2	Ndd1	Mcm1	Ace2	Swi5	Mbp1	Swi4	Swi6
MFA2 (YNL145W)	0	0	1	1	0	1	0	0	0

Reference genes cluster (RGC) and their connectivities to *cis* elements

ORF	Fkh1	Fkh2	Ndd1	Mcm1	Ace2	Swi5	Mbp1	Swi4	Swi6	ORF	Fkh1	Fkh2	Ndd1	Mcm1	Ace2	Swi5	Mbp1	Swi4	Swi6
YAL022C	0	0	0	0	0	1	0	0	0	YKL209C	0	0	0	1	0	0	0	0	0
YAR018C	0	0	1	1	0	0	0	0	0	YLR274W	0	0	0	1	0	0	0	0	0
YDR150W	0	0	1	0	0	0	0	0	0	YML050W	0	0	1	1	0	0	0	0	0
YFL026W	0	0	0	1	0	0	0	0	0	YML125C	0	0	0	0	0	1	0	0	0
YGL032C	0	0	0	1	0	0	0	0	0	YMR001C	0	0	1	0	0	0	0	0	0
YIL050W	0	0	0	0	0	1	0	0	0	YMR002W	0	0	1	0	0	0	0	0	0
YIL129C	0	0	0	0	0	1	0	0	0	YMR253C	0	0	0	1	0	0	0	0	0
YJL079C	0	0	1	1	0	0	0	0	0	YNL056W	0	0	1	1	0	0	0	0	0
YJL157C	0	0	0	1	0	0	0	0	0	YNL058C	0	0	1	1	0	0	0	0	0
YKL163W	0	0	0	1	0	1	0	0	0	YNL145W	0	0	1	1	0	1	0	0	0
YKL164C	0	0	0	1	0	1	0	0	0	YOR066W	0	0	0	1	0	0	0	0	0

Table 2: The reference gene clusters (RGCs) of *CLB2*. Target Gene *CLB2* and the connectivities to *cis* elements.

	Fkh1	Fkh2	Ndd1	Mcm1	Ace2	Swi5	Mbp1	Swi4	Swi6
CLB2 (YPR119W)	1	1	1	1	0	0	0	1	1

Reference genes cluster (RGC) and their connectivities to *cis* elements

ORF	Fkh1	Fkh2	Ndd1	Mcm1	Ace2	Swi5	Mbp1	Swi4	Swi6	ORF	Fkh1	Fkh2	Ndd1	Mcm1	Ace2	Swi5	Mbp1	Swi4	Swi6
YAR018C	0	0	1	1	0	0	0	0	0	YKR013W	0	0	0	0	0	0	0	1	1
YBR133C	0	1	0	0	0	0	0	0	0	YLR056W	0	0	0	0	0	0	0	1	1
YBR138C	1	0	1	0	0	0	0	0	0	YLR084C	0	1	1	1	0	0	0	1	0
YBR139W	1	0	1	0	0	0	0	0	0	YLR131C	0	1	1	1	0	0	0	0	0
YCL063W	1	1	0	0	0	0	0	0	0	YLR190W	0	1	1	1	0	0	0	0	0
YDL227C	0	0	0	0	0	0	0	1	1	YLR209C	1	0	0	0	0	0	0	0	0
YDR033W	0	1	1	0	0	0	0	0	0	YLR210W	1	0	0	0	0	0	0	0	0
YDR146C	0	1	1	1	0	0	0	0	0	YLR274W	0	0	0	1	0	0	0	0	0
YDR150W	0	0	1	0	0	0	0	0	0	YLR342W	0	0	0	0	0	0	0	1	1
YDR224C	0	0	0	0	0	0	0	1	1	YML050W	0	0	1	1	0	0	0	0	0

Table 2: The reference gene clusters (RGCs) of *CLB2*. Target Gene *CLB2* and the connectivities to *cis* elements. *(Continued)*

YDR225W	0	0	0	0	0	0	0	1	1	YML064C	1	1	0	0	0	0	0	0	0
YDR507C	0	0	0	1	0	0	0	1	1	YMR001C	0	0	1	0	0	0	0	0	0
YEL017W	1	0	0	0	0	0	0	0	0	YMR002W	0	0	1	0	0	0	0	0	0
YEL040W	0	1	0	1	0	0	0	1	1	YMR015C	1	0	0	0	0	0	0	1	0
YER001W	0	0	0	0	0	0	0	1	1	YMR183C	1	0	0	0	0	0	0	0	0
YFL026W	0	0	0	1	0	0	0	0	0	YMR253C	0	0	0	1	0	0	0	0	0
YGL032C	0	0	0	1	0	0	0	0	0	YMR305C	0	0	0	0	0	0	0	1	1
YGL038C	0	0	0	0	0	0	0	1	1	YMR307W	0	0	0	0	0	0	0	1	1
YGL116W	0	1	1	1	0	0	0	0	0	YNL056W	0	0	1	1	0	0	0	0	0
YGR014W	0	0	0	0	0	0	0	1	1	YNL058C	0	0	1	1	0	0	0	0	0
YGR099W	1	0	0	0	0	0	0	0	0	YNL231C	0	1	0	0	0	0	0	1	1
YGR151C	0	0	0	0	0	0	0	1	0	YNL300W	0	0	0	0	0	0	0	1	1
YGR152C	0	0	0	0	0	0	0	1	0	YOL011W	0	0	0	0	0	0	0	1	0
YGR153W	0	0	0	0	0	0	0	1	0	YOL030W	1	0	0	0	0	0	0	0	0
YGR221C	0	0	0	0	0	0	0	1	1	YOL114C	0	0	0	0	0	0	0	1	1
YGR279C	0	0	0	0	0	0	0	1	1	YOR066W	0	0	0	1	0	0	0	0	0
YHR061C	0	1	0	0	0	0	0	1	1	YOR073W	0	1	0	0	0	0	0	0	0
YIL056W	0	1	1	0	0	0	0	1	1	YOR372C	0	0	0	0	0	0	0	1	1
YIL121W	0	0	0	0	0	0	0	1	0	YPL032C	1	0	0	0	0	0	0	0	0
YIL123W	0	1	0	1	0	0	0	1	1	YPL116W	1	0	0	0	0	0	0	0	0
YIL158W	0	1	1	1	0	0	0	0	0	YPL127C	0	0	0	0	0	0	0	1	1
YJL051W	0	1	1	1	0	0	0	0	0	YPL141C	1	0	0	0	0	0	0	0	0
YJL079C	0	0	1	1	0	0	0	0	0	YPL155C	0	1	0	0	0	0	0	0	0
YJL157C	0	0	0	1	0	0	0	0	0	YPL163C	0	0	0	0	0	0	0	1	1
YJL158C	0	1	1	0	0	0	0	1	1	YPL255W	0	0	0	0	0	0	0	0	1
YJR054W	0	0	0	0	0	0	0	1	1	YPL256C	0	0	0	0	0	0	0	0	1
YJR092W	1	1	1	1	0	0	0	0	0	YPR013C	1	0	0	0	0	0	0	1	0
YJR110W	0	1	0	0	0	0	0	0	0	YPR119W	1	1	1	1	0	0	0	1	1
YKL096W	0	1	0	0	0	0	0	1	1	YPR149W	0	1	1	0	0	0	0	1	0
YKL103C	0	0	0	0	0	0	0	1	0	YPR159W	0	0	0	0	0	0	0	1	1
YKL209C	0	0	0	1	0	0	0	0	0										

cuit of MFA2 from yeast microarray data. The cis elements of MFA2 are denoted as follows:

MFA2:{Ndd1, Mcm1, Swi5}. (1)

Some genes chosen for the cluster and their *cis* elements are:

YAL022C:{Swi5}, *YAR018C*:{Ndd1, Mcm1}, *YKL163W*:{Mcm1, Swi5} ....

The cluster of genes can be represented by a connectivity matrix of *cis* elements as shown in Table 1 in which "1" denotes the connection of a *cis* element with a gene, while "0" means no connection. Similarly, the RGC for gene *CLB2* can be represented by the connectivity matrix in Table 2.

Dynamic modeling of cis-regulatory circuits

Figure 1 illustrates the leaky integrator-based dynamic models of the cis-regulatory circuits of two yeast genes (MFA2 and CLB2) [1,2]. The dynamics of gene expression can be modeled by a simple first-order nonlinear differential equation that is well established and analyzed in [17]; each model includes the possible regulatory functions of the individual TFs and possible interactions among the TFs. For the target gene MFA2 (Figure 1a), the cis-regulatory circuit is modeled by the following dynamic equation

$$\dot{Y}_{MFA2}(t) = g_{\mathrm{Ndd1}}(t) + g_{\mathrm{Mcm1}}(t) + g_{\mathrm{Swi5}}(t) + g_{\mathrm{Ndd1,Mcm1}}(t) + g_{\mathrm{Ndd1,Swi5}}(t) + g_{\mathrm{Mcm1,Swi5}}(t) + g_{\mathrm{Ndd1,Mcm1,Swi5}}(t) - \lambda_{MFA2} Y_{MFA2}(t) + \varepsilon_{MFA2}(t), \quad (2)$$

where $\varepsilon_{MFA2}(t)$ denotes the noise (data uncertainty), $g_{\mathrm{Ndd1}}(t)$, $g_{\mathrm{Mcm1}}(t)$ and $g_{\mathrm{Swi5}}(t)$ are the regulatory functions of transcription factors Ndd1, Mcm1, and Swi5 or the incident transcriptional regulations at the Ndd1, Mcm1, and Swi5 *cis* elements, respectively, λ_{MFA2} represents the mRNA decay rate of the target gene and we used the degradation rate measured by Wang *et al*. 2002 [20]. The $g_{\mathrm{Ndd1,Mcm1}}(t)$, $g_{\mathrm{Ndd1,Swi5}}(t)$, $g_{\mathrm{Mcm1,Swi5}}(t)$ and $g_{\mathrm{Ndd1,Mcm1,Swi5}}(t)$ denote the following nonlinear interactions among the three TFs:

$$g_{\mathrm{Ndd1,Mcm1}}(t) =: f(g_{\mathrm{Ndd1}}(t), g_{\mathrm{Mcm1}}(t)),$$

$$g_{\mathrm{Ndd1,Swi5}}(t) =: f(g_{\mathrm{Ndd1}}(t), g_{\mathrm{Swi5}}(t)), \quad (3)$$

$$g_{\mathrm{Mcm1,Swi5}}(t) =: f(g_{\mathrm{Mcm1}}(t), g_{\mathrm{Swi5}}(t)),$$

and

$$g_{\mathrm{Ndd1,Mcm1,Swi5}}(t) =: f(g_{\mathrm{Ndd1}}(t), g_{\mathrm{Mcm1}}(t), g_{\mathrm{Swi5}}(t)). \quad (4)$$

The biological meaning of Equation (2) is that the change in the mRNA expression level of gene *MFA2* is due to the productions of regulatory functions of individual TFs and interactions among the TFs, i.e., $g_{\mathrm{Ndd1}}(t)$ + $g_{\mathrm{Mcm1}}(t)$ + $g_{\mathrm{Swi5}}(t)$ + $g_{\mathrm{Ndd1,Mcm1}}(t)$ + $g_{\mathrm{Ndd1,Swi5}}(t)$ + $g_{\mathrm{Mcm1,Swi5}}(t)$ + $g_{\mathrm{Ndd1,Mcm1,Swi5}}(t)$, and $-\lambda_{MFA2} Y_{MFA2}(t)$, which is the degradation of mRNA. Similarly, the *cis*-regulatory circuit of the target gene *CLB2* in Figure 1b is modeled by

$$\dot{Y}_{CLB2}(t) = g_{\mathrm{Fkh1}}(t) + g_{\mathrm{Fkh2}}(t) + \cdots + g_{\mathrm{Fkh1,Fkh2}}(t) + \cdots + g_{\mathrm{Fhk1,Fkh2,Ndd1}}(t) + \cdots - \lambda_{CLB2} Y_{CLB2}(t) + \varepsilon_{CLB2}(t). \quad (5)$$

For simplicity, the indices of target genes and *cis* elements are denoted by numerical notation, so that the *cis*-regulatory circuit of gene i can be written as

$$\dot{Y}_i(t) = \sum_{p}^{v} g_p(t) + \sum_{pq} g_{p,q}(t) + \sum_{pqr} g_{p,q,r}(t) + \cdots - \lambda_i Y_i(t) + \varepsilon_i(t), \quad (6)$$

where v is the total number of *cis* elements in gene i and its corresponding degradation rate λ_i measured by Wang *et al.* 2002 [20]. If λ_i is unavailable, it should be estimated together with the parameters $g_p(t)$, $g_{p,q}(t)$, $g_{p,q,r}(t)$ (see Methods). For the *cis*-regulatory circuits in Equations (2), (5) and (6), one obviously cannot estimate the multiple unknowns $g_p(t)$, $g_{p,q}(t)$, $g_{p,q,r}(t)$, … by only the expression profile (i.e., $Y_i(t)$) of the i^{th} target gene. However, since the functions $g_p(t)$, $g_{p,q}(t)$, $g_{p,q,r}(t)$, … are assumed to be the same for all genes in the RGC and since there are overlaps of *cis* elements among genes in this RGC, one can estimate these functions from an array of expression profiles $Y_1(t)$, $Y_2(t)$, …, $Y_N(t)$ of the genes in the RGC simultaneously, taking advantage of cross information enhancement. The RGCs for *MFA2* and *CLB2* are shown in Table 1 and Table 2, respectively. In this situation, a cross-gene identification method is proposed as follows. By integrating the dynamic equations of *cis*-regulatory circuits for N genes in the RGC of the gene of interest, we obtain the following array of dynamic equations

$$\begin{pmatrix} \dot{Y}_1(t) \\ \dot{Y}_2(t) \\ \vdots \\ \dot{Y}_i(t) \\ \vdots \\ \dot{Y}_N(t) \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{0} & \cdots & 0 & 1 & \cdots & 0 & \cdots & 1 \\ 1 & 1 & \cdots & 1 & 1 & \cdots & 1 & \cdots & 0 \\ \vdots & & & \vdots & & \vdots & & & \vdots \\ 0 & 1 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 1 \\ \vdots & & & \vdots & & \vdots & & & \vdots \\ 1 & 1 & \cdots & 1 & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix} \bullet \begin{pmatrix} g_1(t) \\ g_2(t) \\ \vdots \\ g_{1,2}(t) \\ g_{1,3}(t) \\ \vdots \\ g_{1,2,3}(t) \\ \vdots \\ g_{p,q,r}(t) \end{pmatrix} - \begin{pmatrix} \lambda_1 Y_1(t) \\ \lambda_2 Y_2(t) \\ \vdots \\ \lambda_i Y_i(t) \\ \vdots \\ \lambda_N Y_N(t) \end{pmatrix} + \begin{pmatrix} \varepsilon_1(t) \\ \varepsilon_2(t) \\ \vdots \\ \varepsilon_i(t) \\ \vdots \\ \varepsilon_N(t) \end{pmatrix}. \quad (7)$$

In the cross-gene dynamic equations in (7), the process to identify regulatory functions $g_p(t)$, $g_{p,q}(t)$ and $g_{p,q,r}(t)$ from microarray data $Y_i(t)$, i = 1, 2,..., N *is* called the *cross-gene identification* approach, in which the regulatory functions $g_p(t)$, $g_{p,q}(t)$ and $g_{p,q,r}(t)$ are shared by genes in the RGC. Therefore, the estimation of the regulatory functions of

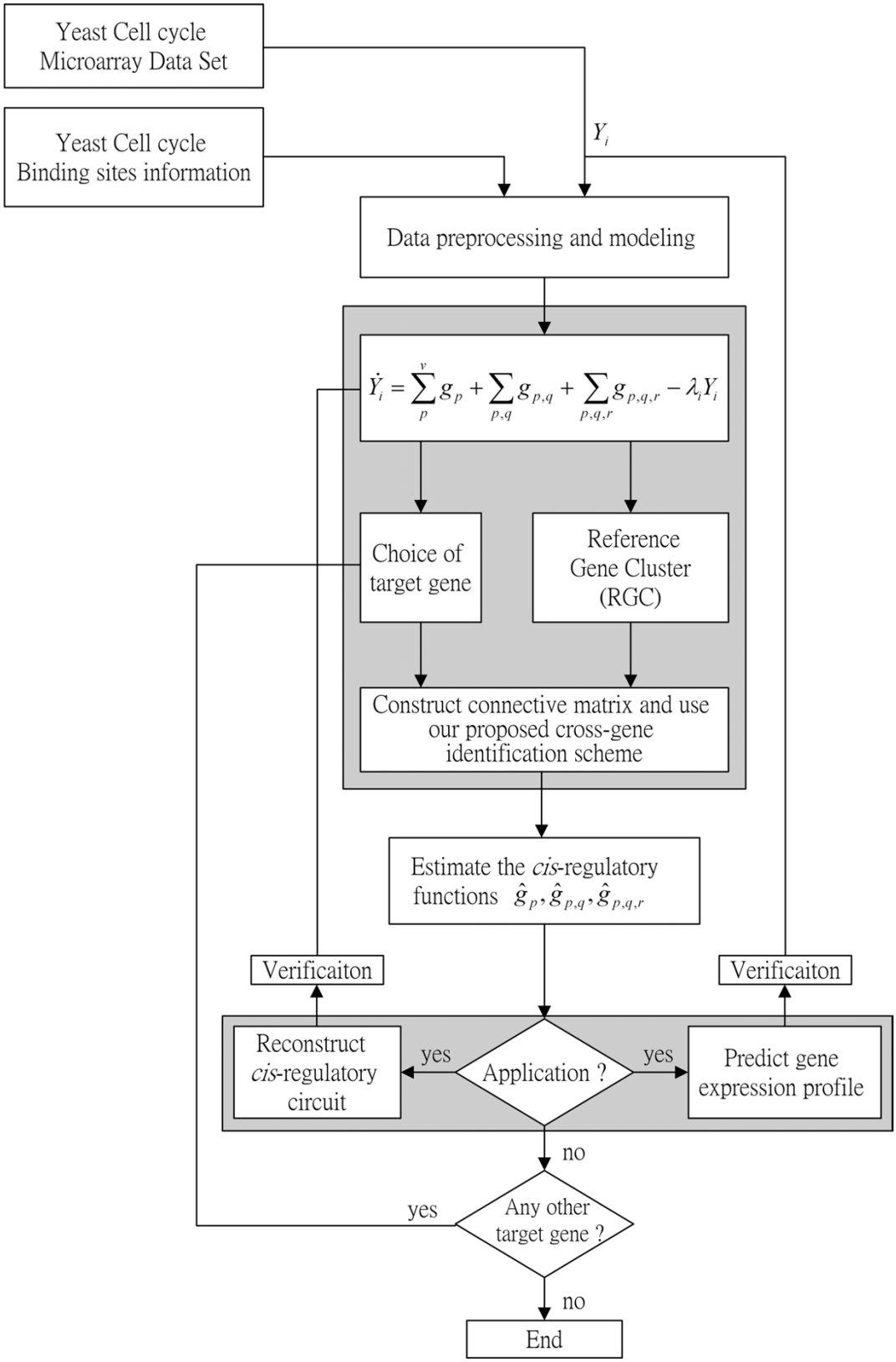

Figure 2
The overall flowchart of the modeling, identification and prediction of a *cis*-regulatory circuit.

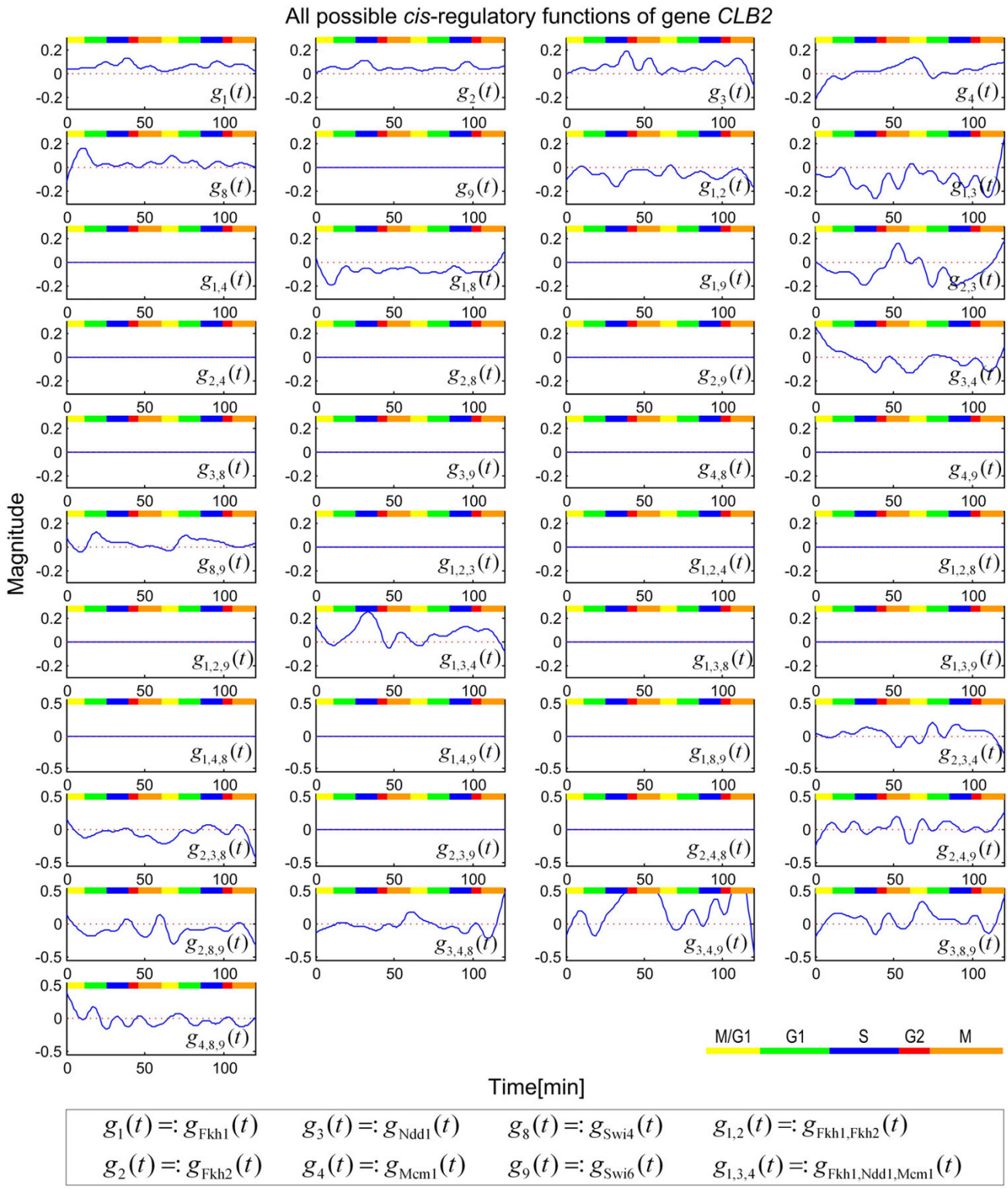

Figure 3
All estimated *cis*-regulatory functions, including the regulatory function $g_p(t)$ of each individual TF and the interactions $g_{p,q}(t)$ and $g_{p,q,r}(t)$ among the TFs that recognize the *cis* elements of the *CLB2* gene. The numerical notation of regulatory functions is given in the box at the bottom of the figure. Different phases in the cell cycle are indicated by the colored bar near the right lower corner.

one gene can also use the information from the profiles $Y_1(t)$, $Y_2(t)$, …, $Y_N(t)$ of other genes in RGC to improve the identification ability of the regulatory functions to reconstruct the *cis*-regulatory circuit of the gene of interest, which is called *cross information enhancement*.

Remark 1 : Suppose that the gene of interest in Equation (6) has *cis* elements $p = 1,..., v$. Then all genes whose *cis* elements are subsets of these v *cis* elements are included in the same RGC of the gene of interest.

Cross-gene identification scheme

Since the number of functions $g_p(t)$, $g_{p,q}(t)$, $g_{p,q,r}(t)$, … is finite, we can estimate these functions if the number N of dynamic equations in Equation (7) is large enough. Equation (7) can be rewritten in an algebraic form

$$X(t) = A \cdot G(t) + E(t), \quad (8)$$

where

$$X(t) = \begin{pmatrix} \dot{Y}_1(t) + \lambda_1 Y_1(t) \\ \dot{Y}_2(t) + \lambda_2 Y_2(t) \\ \vdots \\ \vdots \\ \dot{Y}_i(t) + \lambda_i Y_i(t) \\ \vdots \\ \vdots \\ \dot{Y}_N(t) + \lambda_N Y_N(t) \end{pmatrix}, \quad G(t) = \begin{pmatrix} g_1(t) \\ g_2(t) \\ \vdots \\ g_{1,2}(t) \\ g_{1,3}(t) \\ \vdots \\ g_{1,2,3}(t) \\ \vdots \\ g_{p,q,r}(t) \end{pmatrix}, \quad E(t) = \begin{pmatrix} \varepsilon_1(t) \\ \varepsilon_2(t) \\ \vdots \\ \vdots \\ \varepsilon_i(t) \\ \vdots \\ \vdots \\ \varepsilon_N(t) \end{pmatrix}$$

Remark 2 : In order to calculate the derivatives in $X(t)$ from undersampled data, a cubic spline interpolation method [21,22] is employed for curve fitting to obtain more accurate differential values and to learn more reliable models.

In Equation (8), $X(t)$ can be calculated from microarray data for the RGC of the gene of interest, which can then be used to estimate the vector $G(t)$ by the least squares method, leading to the following solution:

$$\hat{G}(t) = (A^T A)^{-1} A^T X(t) \quad (9)$$

for all t. After $\hat{G}(t)$ is estimated from Equation (9), the regulatory functions $g_1(t)$, …, $g_{1,2}(t)$, … and $g_{p,q,r}(t)$ in Equation (8) can be reconstructed for the genes in the RGC at every time point. If a *cis*-regulatory circuit is free of any function $g_p(t)$, $g_{p,q}(t)$, or $g_{p,q,r}(t)$, the value of the estimated function should be very small or zero. After the functions are estimated, they can be plugged into Equation (6) and the reconstruction of the *cis*-regulatory circuit of the gene of interest is completed. The flowchart for modeling, identifying and predicting a *cis*-regulatory circuit is shown in Figure 2.

In order to obtain more accurate *cis*-regulatory circuits, the model should include the triple interactions among the recognizing TFs; i.e., the vector $G(t)$ in Equation (7) should include $g_{p,q,r}(t)$.

Remark 3: If the degradation parameters λ_i in (6) are unavailable, the estimation procedure of $G(t)$ and λ_i from equation (7) to equation (9) should be modified as in Methods.

Two application examples

I. The cis*-regulatory circuit of* MFA2

Suppose that the cis-regulatory circuit of the MFA2 gene is of interest. We construct a dynamic model of the cis-regulatory circuits of the genes in the RGC of MFA2 in Table 1 and then estimate the regulatory functions and interactions by the cross-gene identification scheme. The estimated transcriptional regulatory functions $g_p(t)$ and interactions $g_{p,q}(t)$ and $g_{p,q,r}(t)$ are shown as the insets in Figure 1a. These insets indicate that cis elements Ndd1, Mcm1 and Swi5 in MFA2 all have cell cycle regulatory abilities in the late G1 phase. In addition, every individual cis element has a positive regulatory function on the MFA2 gene; for example, the function $g_4(t)$ for Mcm1 has an obvious peak value in the transition phase late G1 of the cell cycle. Michaelis and Herskowitz [19] found that the MFA2 gene causes the cell cycle arrest at the G1 phase and is required for mating in yeast. Note that the interaction $g_{3,6}(t)$ between TFs Ndd1 and Swi5 is very weak or absent in the cell cycle. In contrast, the interaction $g_{4,6}(t)$ between TFs Mcm1 and Swi5 is dynamic; it has a high positive peak value in the late G1 phase, which coincides with MFA2's activity phase. This interaction seems to play an important role of positive regulation in this cis-regulatory circuit. On the other hand, the regulatory function $g_{3,4,6}(t)$ of the interaction among TFs Ndd1, Mcm1 and Swi5 is negative on gene MFA2. This regulation may repress the expression of gene MFA2 to make the expression decay to the steady state. If there is no repression function such as $g_{3,4,6}(t)$, the expression of MFA2 will increase and may disrupt in the cell cycle.

II. The cis-regulatory circuit of CLB2

Clb proteins are crucial cyclins for completing the G2/M transition of the mitotic cell cycle and the most typical one is the B-type mitotic cyclin Clb2, which is required for entry into mitosis [23]. Suppose that the cis-regulatory circuit of gene CLB2 is of interest. Fkh1, Fkh2, Ndd1, Mcm1, Swi4 and Swi6 have been identified as the TFs that bind to the promoter sequence of CLB2[3,4,24,25]. As shown in

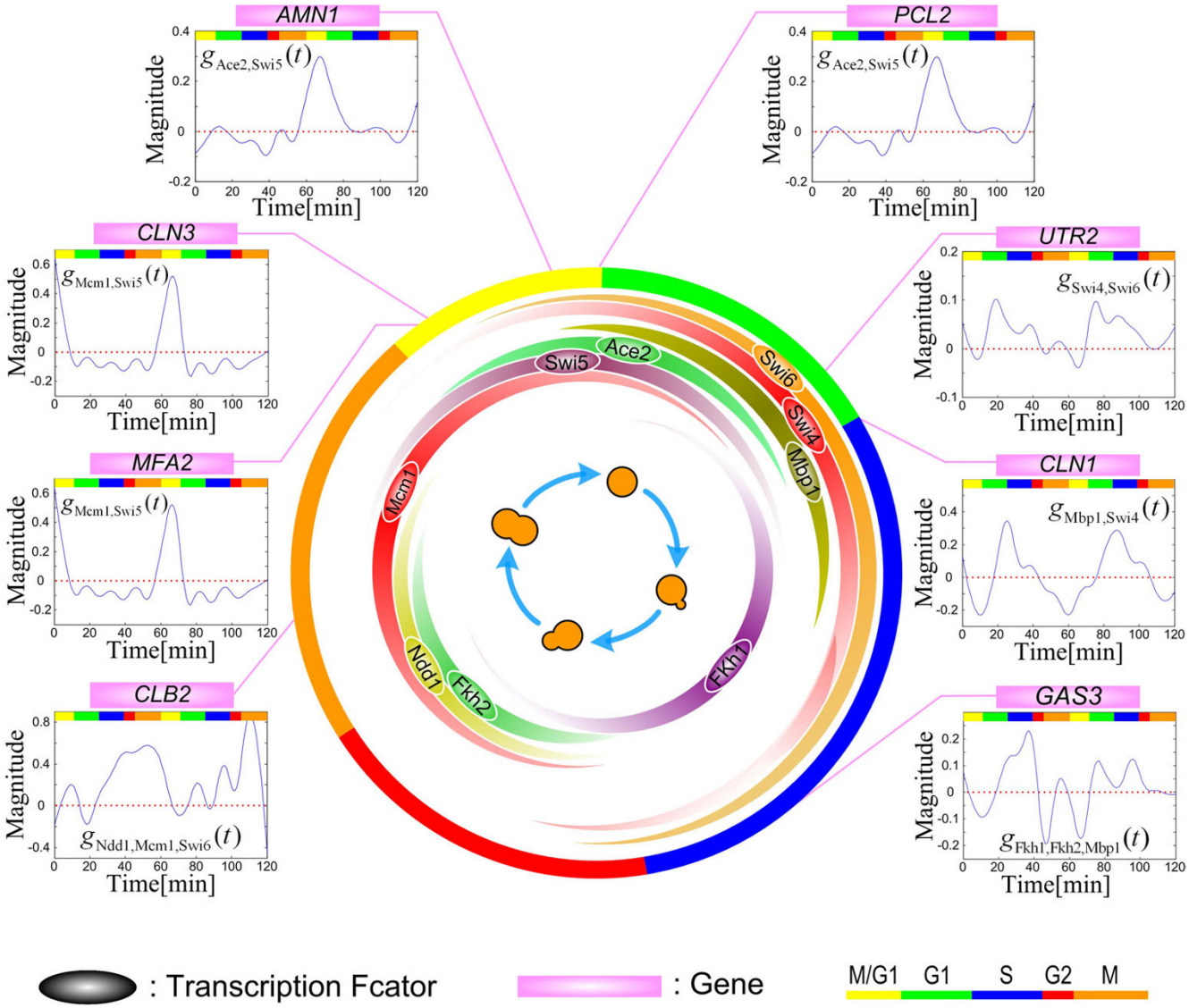

Figure 4

The gene expression phases match the main regulatory functions of TFs. It is seen that the main interaction functions of TFs have a peak value and always occur during or soon after the mRNA expression phases of the corresponding genes. For example, the gene *MFA2* has the main interaction regulatory function $g_{Mcm1,Swi5}(t)$ which has a peak during the expression phases between the TFs Mcm1 and Swi5 by identifying the *cis*-regulatory circuit. As another example, the gene *UTR2* has the main interaction regulatory function $g_{Swi4,Swi6}(t)$, which has a peak during the expression phases between the TFs Swi4 and Swi6 by identifying the *cis*-regulatory circuit. These results indicate that the main regulatory functions have a peak value phase to match the gene expression phase. Therefore, we can estimate the gene expression phase by identifying the main regulatory function. The width of a colored band in the inner circle is approximately proportional to the expression level of the TF gene of interest in the cell cycle. A pink line points to the main expression phase of a target gene in the pink box. Different phases in the cell cycle are indicated by the colored bar at the right lower corner.

Figure 1b, the possible cis-regulatory circuit of CLB2 is very complex. Using the cross-gene identification scheme, we reconstructed the cis-regulatory functions shown in Figure 3. Although there are 41 possible regulatory functions, including $g_p(t)$, $g_{p,q}(t)$ and $g_{p,q,r}(t)$, only 20 regulatory functions are found to have nonzero values (Figure 3).

The two *cis* elements Fkh1 and Fkh2 are found to have very similar regulatory functions $g_1(t)$ and $g_2(t)$, in agree-

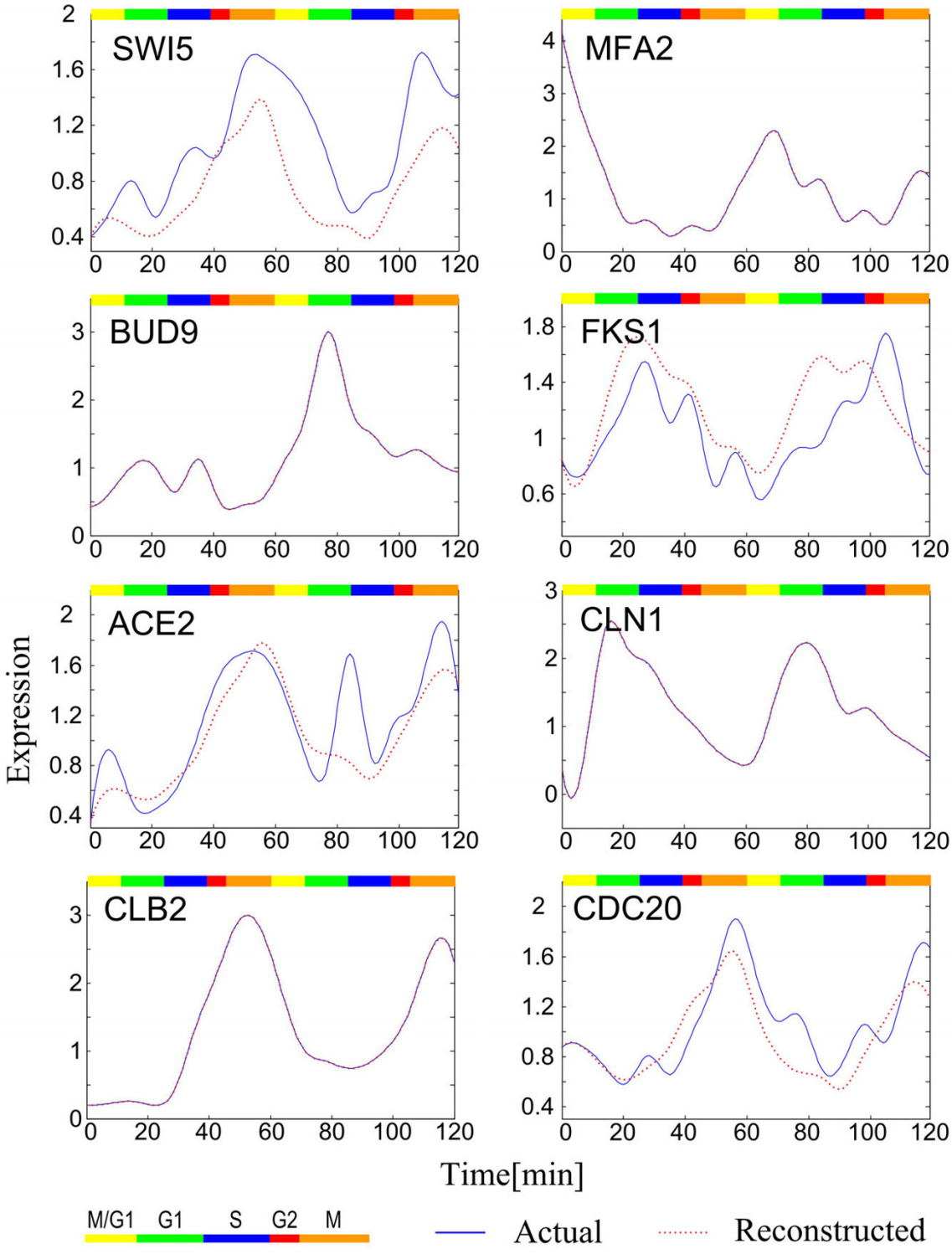

Figure 5
Comparison between the actual gene expression profiles (with cubic spline) and the reconstructed gene expression profiles. The examples shown were randomly chosen. The reconstructed gene expression profiles were obtained by integrating the estimated *cis*-regulatory functions and the chromatin immunoprecipitation data. When there is only one blue line in a figure it means that the reconstructed function is very close to the actual gene expression profile. Different phases in the cell cycle are indicated by the colored bar.

ment with the experimental evidence that the forkhead family members Fkh1 and Fkh2 of transcription factors have overlapping roles in the control of the G2/M transition [25,26]. The regulatory function $g_2(t)$ of Fkh2 has a distinct cell cycle regulatory ability (Figure 3) and especially, the interaction function $g_{2,3}(t)$ between Fkh2 and Ndd1 has a strong regulatory contribution to the gene expression profile in the M/G1 phase. The regulatory function $g_4(t)$ of Mcm1 has two peaks. There is experimental evidence that Mcm1 is a member of an evolutionarily conserved class of transcription factors that have related to DNA binding sequences and dimerization domains. In addition, Mcm1 binds the early cell cycle box (ECB) that contains a Mcm1 *cis* element in the *SWI4*, *CLN3*, *CDC6*, and *CDC47* promoters and activates M/G1-specific transcription [27].

The cell cycle genes that are activated during the late G1 or S phase have SBF or MBF sequence-specific transcription factors that bind the *cis* elements in their promoter region. SBF (the Swi4-Swi6 cell cycle box binding factor) is a heterodimer of Swi4 and Swi6 [3,28,29]. The regulatory functions $g_8(t)$ and $g_9(t)$ of Swi4 and Swi6 and their interaction function $g_{8,9}(t)$ are estimated using the dynamic expression model (Figure 3). It is well-known that neither Swi4 nor Swi6 alone has obvious cell cycle regulation ability, and indeed we found that $g_8(t)$ has only one peak and so shows no cycle and that $g_9(t)$ shows no capability of cell cycle regulation (Figure 3). In contrast, the combination of Swi6 and Swi4 to make the complex SBF enables the *cis* elements Swi6 and Swi4 to provide cell cycle regulation capacity; that is, the interaction function $g_{8,9}(t)$ of Swi6 and Swi4 has a peak during the G1/S phase of the cell cycle. Ndd1 and Fkh2 are bound to identical promoters throughout the cell cycle and their interaction $g_{2,3}(t)$ is an important transcriptional process targeted by the Cdk activity [24,30]. In addition, there is another obvious positive interaction $g_{3,4,9}(t)$ among Ndd1, Mcm1, and Swi6 (Figure 3). It has a large regulatory ability in the G1/S phase, which almost dominates the expression profile of *CLB2*. In contrast, $g_{3,4}(t)$ has a negative regulation contribution. We therefore propose that the key factor Swi6 in the interaction $g_{3,4,9}(t)$ is similar to its role in SBF and MBF. At any rate, our model suggests that Swi6 plays a key role in the interaction among Ndd1, Mcm1, and Swi6. This is a new finding in the *cis*-regulatory circuit of *CLB2*.

We also confirm the well-known interaction among Ndd1, Fkh2, and Mcm1 in the *cis*-regulatory circuits of the *CLB2* and *SWI5* genes because the interaction function $g_{2,3,4}(t)$ has a distinct regulatory ability in the G2/M phase of the cell cycle (Figure 3). Interestingly, the interaction function $g_{3,4,9}(t)$ among Ndd1, Mcm1 and Swi6 is about two times higher than any of the other functions in Figure 3. In addition, the interaction $g_{1,3,4}(t)$ among Fkh1, Ndd1 and Mcm1 is highly positive, the interaction $g_{3,4,8}(t)$ among Ndd1, Mcm1 and Swi4 is mildly positive, while the interaction $g_{3,8,9}(t)$ among Ndd1, Swi4 and Swi6 is negative. These observations are in agreement with the fact that the regulatory ability of an interaction among TFs is usually much stronger than that of an individual TF; in other words, there is synergy among TFs.

In summary, there are many experimental observations that support the *cis*-regulatory functions identified by the dynamic model and our model provides novel insights into the quantitative regulation of the *cis*-regulatory circuit of a gene of interest.

Support from expression phases of TFs in the cell cycle

In this paper, the question of why the strengths of regulatory functions in the cis-regulatory circuits are different in different phases of the cell cycle is investigated. Based on the mRNA expression profiles of transcription factor genes from experiments, the distribution of the expressions of the nine TF genes in the different phases of the cell cycle is shown in Figure 4. In support of our results, the large positive interaction functions (peaks) among a set of TFs always occur during the expression phases of the genes of the interacting TFs. For example, for the cis-regulatory circuit of MFA2 (Figure 1a), there is an obvious peak for the function $g_{4,6}(t)$ of the interaction between TFs Mcm1 and Swi5 during the M phase, and in Figure 4, this peak indeed occurs during the expression phases of the two TF genes [18,27]. As another example, for the cis-regulatory circuit of CLB2, there is a strong interaction ($g_{3,4,9}(t)$) among Ndd1, Mcm1 and Swi6 starting from the G2 phase (Figure 3). We can therefore infer that the expression of a cell cycle gene in a specific phase of the cell cycle needs a specific inducing signal, which is mainly from the interactions of certain specific TFs that bind the cis elements of the gene.

Accuracy of reconstructed cis-regulatory circuits

The accuracy of the reconstructed cis-regulatory circuit of a gene can be evaluated by reconstructing the expression profile of the gene using the reconstructed cis-regulatory circuit

$$\dot{\hat{Y}}_i(t) = \sum_p^v \hat{g}_p(t) + \sum_{pq} \hat{g}_{p,q}(t) + \sum_{pqr} \hat{g}_{p,q,r}(t) + \cdots - \lambda_i \hat{Y}_i(t), \quad (10)$$

where $\hat{g}_p(t)$, $\hat{g}_{p,q}(t)$ and $\hat{g}_{p,q,r}(t)$ have been estimated by the cross-gene identification scheme. The reconstructed profile and the observed profile are compared in Figure 5. We find that if the number of the *cis* elements of a gene is large enough, the reconstructed expression profile is very accurate; otherwise, it may be inaccurate. Fortunately, although the reconstructed expression profile is not accu-

a

Number of RGC	Predicted gene accession number	Mean square error (MSE)
RGC_1	YGL028C	1.2694
RGC_2	YJL157C	0.3482
RGC_3	YHR152W	0.1259
RGC_4	YHR143W	0.695
RGC_5	YMR305C	0.0855
RGC_6	YIL121W	0.1153
RGC_7	YOL011W	0.0869
RGC_8	YHR149C	0.311
RGC_9	YGR152C	0.4089
RGC_10	YPR120C	0.4474
RGC_11	YNR009W	0.4067
RGC_12	YPL255W	0.3916
RGC_13	YML064C	0.0736
RGC_14	YML125C	0.0939
RGC_15	YPR013C	0.0926
RGC_16	YKL151C	0.1541
RGC_17	YKL044W	0.9212
RGC_18	YNL339C	2.0549
RGC_19	YMR145C	0.1306
RGC_20	YMR015C	0.1359
RGC_21	YPL155C	0.6567
RGC_22	YJL051W	0.7856
RGC_23	YOL030W	0.4845
RGC_24	YJR110W	0.2617
RGC_25	YIL056W	0.8052
RGC_26	YNL056W	0.4925
RGC_27	YLR013W	0.4122
RGC_28	YLR209C	0.5594
RGC_29	YLR084C	0.68
RGC_30	YIL009W	0.2891
RGC_31	YGR086C	0.3596
RGC_32	YAL022C	0.1352
RGC_33	YNL058C	0.41
RGC_34	YIL158W	0.0466
RGC_35	YLR190W	0.3882
RGC_36	YPL242C	0.1649
RGC_37	YCR024C-A	0.1783
RGC_38	YHR151C	0.1891
RGC_39	YGL008C	0.7552
RGC_40	YIL050W	0.1061
RGC_41	YOR073W	0.2387
RGC_42	YEL017W	0.1188
RGC_43	YPL116W	0.1665
RGC_44	YCL063W	0.113
RGC_45	YMR183C	0.0296
RGC_46	YIL129C	0.1597
RGC_47	YPL032C	0.0551
RGC_48	YKL008C	0.117
RGC_49	YLR342W	0.0814
RGC_50	YDL055C	0.2077
RGC_51	YBL002W	0.316
RGC_52	YPL127C	0.2238
RGC_53	YDL018C	0.4313
RGC_54	YJR030C	0.3781
RGC_55	YMR078C	0.5847
RGC_56	YNL300W	0.7581
RGC_57	YGR153W	0.4001
RGC_58	YPR075C	0.1955
RGC_59	YGR189C	0.5323
RGC_60	YGR234W	0.8625
RGC_61	YJL159W	0.8619
RGC_62	YDL127W	1.5604
RGC_63	YER190W	0.0661
RGC_64	YOR074C	0.1065
RGC_65	YML027W	0.1176
RGC_66	YDL101C	0.0321
RGC_67	YMR179W	0.1823
RGC_68	YER189W	0.0655
RGC_69	YGR109C	1.2195
RGC_70	YER070W	0.3758
RGC_71	YDL003W	0.429
RGC_72	YBR070C	0.4805
RGC_73	YLR463C	0.5441
RGC_74	YLR465C	0.3625
RGC_75	YPR202W	0.3572
RGC_76	YLR464W	0.3935
RGC_77	YFL065C	0.369
RGC_78	YLR462W	0.3527
RGC_79	YBL111C	0.2549
RGC_80	YLR466W	0.253
RGC_81	YOR372C	0.3655
RGC_82	YJL115W	0.3465
RGC_83	YJL079C	0.4326
RGC_84	YER145C	0.253
RGC_85	YOL012C	0.4529
RGC_86	YDR225W	1.2079
RGC_87	YDR224C	0.7269
RGC_88	YMR215W	0.6068
RGC_89	YLR056W	0.2487
RGC_90	YLR131C	0.1322
RGC_91	YMR001C	0.3859
RGC_92	YMR002W	0.0965
RGC_93	YBL003C	0.1892
RGC_94	YMR307W	0.0919
RGC_95	YOR247W	0.0256
RGC_96	YOR248W	0.0249
RGC_97	YCR065W	0.492
RGC_98	YPL163C	1.1383
RGC_99	YPL256C	1.0978
RGC_100	YGR014W	0.1084

b

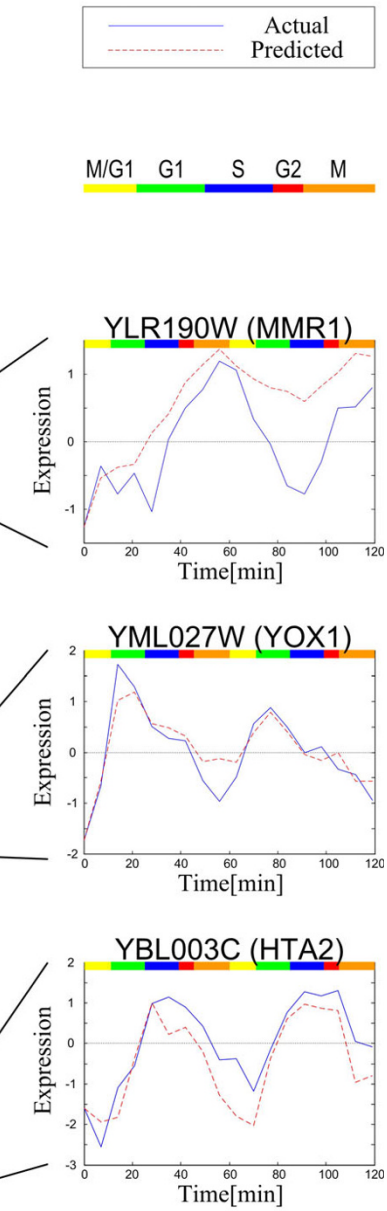

Figure 6
The regulatory functions integrated with ChIP-chip data to predict gene expression profiles. To test the prediction performance of our model, (a) 100 yeast cell cycle genes that have not been employed in the reconstruction of the *cis*-regulatory circuits are randomly chosen from their corresponding reference gene clusters (RGCs). The maximum mean square error (MSE) of prediction results is 2.055 and the minimum is 0.025. (b) Three examples of the comparison between the actual (blue) and the predicted (red) gene expression profiles. Different phases in the cell cycle are indicated by the colored bar.

rate in some genes, the trend of the expression profile for a gene is always correct.

Prediction of gene expression profile

In the above, each regulatory circuit was identified using 95% of genes in its RGC, and the remaining 5% of genes in the RGC will now be used for predicting expression profiles, i.e., for cross validation. Our cross-gene identification scheme assumes that the regulatory functions of TFs are universal in the cluster of genes with similar functions. Under this assumption, our method should be able to predict the expression profiles of other genes in the same cluster that have not been employed in Equation (7) to reconstruct the cis-regulatory circuits. This is one way to validate our model.

In Figure 6, we randomly chose 100 RGCs to test the prediction accuracy; that is, for each RGC we show the prediction result for one of the unused genes. Figure 6a shows the predicted target gene and the mean square error (MSE) of prediction results which has the maximum of 2.055 and the minimum of 0.025. In addition, three examples of the detailed comparison between the actual and the predicted gene expression profiles are shown in Figure 6b. In general, the predicted profiles are satisfactory approximations of the observed expression profiles. We found that the smaller the number of the *cis* elements, the less accurate the prediction results. However, if some *cis* elements of a gene have strong regulatory functions, the expression profiles of this gene can be predicted accurately even when the number of *cis* elements is small. If some genes in the RGC have the same *cis* elements but have different observed gene expression profiles, these expression profiles will lead to poor estimation of parameters. This is the main cause of prediction error. Why does this situation arise? It may be that some *cis* elements of these genes have not yet been identified or there are some errors in the inference of the *cis* elements. For example, *MMR1* may have another *cis* element Gcr2 [5] and this may be why the predicted profile is quite different from the observed profile (Figure 6b).

Discussion

In contrast to current methods, our method uses all possible expression profile information from the cluster of genes to reconstruct the *cis*-regulatory circuit of a target gene. In particular, our method is capable of extracting dynamic interactions among TFs. For this reason, the analysis and interpretation of output expression profiles become straightforward. Therefore, our method has a high potential for applications such as studying variations of the *cis*-regulatory circuit of the same gene in different yeast strains to investigate the regulatory evolution of the gene.

The contributions of this study include: (1) a nonlinear dynamical model is developed for *cis*-regulatory circuits in terms of regulatory functions and interactions among TFs, (2) a cross-gene identification scheme is proposed to estimate many parameters involved in the dynamical model of *cis*-regulatory circuits from the expression profiles of genes in the reference gene cluster, (3) a detailed identification of the dynamic *cis*-regulatory abilities of TFs, which vary with time, and (4) a gene expression prediction method is developed by the proposed dynamic *cis*-regulatory circuit, assuming that the *cis*-regulatory functions of the same TFs in different circuits are the same. Three advantages of our method over current methods are that the *cis*-regulatory circuit is constructed with the circuit structure intact, that it uses the expression profiles of many genes simultaneously to obtain extra information, and that it is dynamic and quantitative.

Significantly, our model not only can confirm known regulations but also can provide conjectures for experimental verification. Consider the key positive *cis*-regulatory function $g_{4,6}(t)$ in Figure 1a. We propose that during the expression of gene *MFA2*, the transcription factor Ndd1 (the G2/M phase) communicates with the transcription factor Mcm1 (the M phase) to transmit a specific signal to induce the expression of the *MFA2* gene. Such conjectures from the reconstructed *cis*-regulatory circuits may be useful for studying the regulatory evolution of genes by comparing the *cis*-regulatory functions of different strains, or for predicting the gene expression behavior before conducting an experiment.

However, we found poor results in some cases. For example, in Figure 3, we were unable to find the basal regulatory function $g_9(t)$ of individual Swi6 or the interaction $g_{2,4}(t)$ between Fkh2 and Mcm1 for gene *CLB2* [31]. These regulatory functions have not been identified by our scheme because they have no obvious specific phase regulatory ability. Besides, from the *cis*-regulatory functions in Figures 1 and 2, several *cis*-regulatory function profiles did not show a clear periodicity. A possible reason may be that the original microarray data were noisy and the use of cubic spline interpolation and linear transformation of microarray data in our scheme had introduced new noise and distortions. Most probably, the *cis* element information used to construct the *cis*-regulatory circuits under yeast cell cycle is not complete; only nine significant *cis* elements were considered in this study to reduce the complexity of the mathematical model. Another possible source of error is that we have not considered the order of the *cis* elements on the promoter region, which may affect the strength of the interaction between TFs [36]. Such differences, however, can be incorporated by putting, say both $g_{p,q}(t)$ and $g_{p,q,r}(t)$, into the model.

In view of the facts that there are uncertainties about the *cis* elements of some of the genes studied and that microarray data are noisy, it is remarkable that our method gave accurate results for the expression profiles of the majority of the cell cycle genes studied and also gave fairly accurate predictions of the expression profiles of other cell cycle genes. In the future, if better *cis* elements data and more accurate and longer gene expression profile data become available, we should be able to improve the reconstruction of *cis*-regulatory circuits. Also, our approach may be extended to reconstruct *cis*-regulatory circuits in more diverse conditions and more complex eukaryotes. After *cis*-regulatory circuits are accurately described by explicit dynamical equations, some applications will be straightforward.

Conclusion

In this study, we assume that the regulatory functions of the same *cis* elements and the interaction functions among their TFs are similar across genes within the cluster of genes with overlapping *cis* elements; i.e., the regulatory functions and interaction functions are universal in this cluster of genes. Under this assumption, the cross-gene identification scheme takes advantage of cross-information enhancement to improve the accuracy of parameter estimation. The number of genes used should be large enough, so that we have a large number of outputs (i.e., their microarray data) for parameter estimation.

After the parameters of the *cis*-regulatory circuits of interest have been estimated, the circuits can be explicitly described by plugging these parameters into their corresponding dynamic equations. Moreover, these estimated functions and interactions can be used to predict the expression profiles of other genes that share the same *cis*-regulatory elements but have not been used to identify the *cis*-regulatory circuits. In this manner, we can evaluate the performance of the proposed dynamic model of *cis*-regulatory circuits. From a number of examples, we have found that the predicted results are in most cases satisfactory, confirming the validity of the proposed dynamic model of *cis*-regulatory circuits. Our modeling represents a new approach to studying *cis*-regulatory circuits from cross-gene expression data. It is a systems biology approach because we consider the regulatory circuits of many genes and many TFs at the same time and we use system identification techniques to estimate the parameters of the circuits. The results of expression prediction from experimental data suggest that our novel approach is suitable for deciphering the regulatory functions and the cooperativity of the TFs that regulate the expression of a gene.

Methods

Experimental data

To identify the cis-regulatory circuit of a gene of interest in the yeast cell cycle, we apply our approach to the data of Spellman et al. 1998 [18], which contains expression profiles of 6178 open reading frames (ORFs) in the yeast Saccharomyces cerevisiae during the cell cycle [33]. Our analysis was applied to the α-factor arrest data set. The raw data were transformed into a linear scale from the original log ratio carried out by Spellman et al. 1998 [18]. To reduce the effect of noise and to overcome undersampled microarray data in the estimation of cis-regulatory circuits, the cubic spline was used for data interpolation and smoothing to obtain a less sensitive first derivative of the expression pattern and to learn a more reliable model. Furthermore, the noise is modeled in the noise term $\varepsilon_i(t)$ in Equation (7).

From the RGC of the gene of interest, the cross-gene identification method from Equations (8) to (9) is employed to reconstruct the *cis*-regulatory circuit. The connectivity information between TFs and their target genes was obtained from the yeast cell cycle analysis [4]. We focused on the nine transcription factors that have been identified to play important roles in the transcription regulation of a set of yeast genes whose expressions are cell-cycle dependent; these nine transcription factors are Mbp1, Swi4, Swi6, Mcm1, Fkh1, Fkh2, Ndd1, Swi5, and Ace2 [24,26,27,34] (See additional file 1: Table for the original data used to perform this analysis).

Estimation of degradation rate

If the mRNA degradation rate λ_i in Equation (6) has not been estimated experimentally, it should be estimated together with the parameters $g_p(t)$, $g_{p,q}(t)$, $g_{p,q,r}(t)$,. The algorithm to estimate the λ_i is described as follows. First, Equation (6) is changed to

$$\dot{Y}_i(t) = \sum_p^{\nu} g_p(t) + \sum_{pq} g_{p,q}(t) + \sum_{pqr} g_{p,q,r}(t) + \cdots - \lambda_i(t) Y_i(t) + \varepsilon_i(t), \quad (11)$$

where ν is the total number of *cis* elements in gene *i*, and the degradation rate is substituted as $\lambda_i(t)$. Similarly, since the functions $g_p(t)$, $g_{p,q}(t)$, $g_{p,q,r}(t)$, … are assumed to be the same for all genes in the RGC and since there are overlaps of *cis* elements among genes in this RGC, one can estimate these functions from an array of expression profiles $Y_1(t)$, $Y_2(t)$, …, $Y_N(t)$ of the genes in the RGC simultaneously, taking advantage of cross information enhancement. The RGCs for *MFA2* and *CLB2* are shown in Table 1 and Table 2, respectively.

Second, by integrating the dynamic equations of *cis*-regulatory circuits for N genes in the RGC of the gene of interest, we obtain the following array of dynamic equations

$$\begin{pmatrix} \dot{Y}_1(t) \\ \dot{Y}_2(t) \\ \vdots \\ \dot{Y}_i(t) \\ \vdots \\ \dot{Y}_N(t) \end{pmatrix} = \left(\begin{array}{cccccccccc|ccccc} 1 & \mathbf{0} & \cdots & 0 & 1 & \cdots & 0 & \cdots & 1 & -Y_1(t) & 0 & & \leftrightarrow & & 0 \\ 1 & 1 & \cdots & 1 & 1 & \cdots & 1 & \cdots & 0 & 0 & -\mathbf{Y_2(t)} & & & & \\ & \vdots & & & \vdots & & & & \vdots & & & \ddots & & & \\ & \vdots & & & \vdots & & & & \vdots & & & & & & \\ 0 & 1 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 1 & \updownarrow & & & -Y_1(t) & & \updownarrow \\ & \vdots & & & \vdots & & & & \vdots & & & & & \ddots & \\ & \vdots & & & \vdots & & & & \vdots & & & & & & 0 \\ 1 & 1 & \cdots & 1 & 0 & \cdots & 0 & & 1 & 0 & & \leftrightarrow & & 0 & -Y_N(t) \end{array}\right) \cdot \begin{pmatrix} g_1(t) \\ g_2(t) \\ \vdots \\ g_{1,2}(t) \\ g_{1,3}(t) \\ \vdots \\ g_{1,2,3}(t) \\ \vdots \\ g_{p,q,r}(t) \\ \lambda_1(t) \\ \lambda_2(t) \\ \vdots \\ \vdots \\ \lambda_i(t) \\ \vdots \\ \vdots \\ \lambda_N(t) \end{pmatrix} + \begin{pmatrix} \varepsilon_1(t) \\ \varepsilon_2(t) \\ \vdots \\ \varepsilon_i(t) \\ \vdots \\ \varepsilon_N(t) \end{pmatrix} \quad (12)$$

In the dynamic equations in (12), the regulatory functions $g_p(t)$, $g_{p,q}(t)$ and $g_{p,q,r}(t)$ are shared by genes in the RGC. Therefore, the estimation of these functions from expression profiles $Y_1(t)$, $Y_2(t)$, …, $Y_N(t)$ can use also information from other genes to enhance our ability to reconstruct the *cis*-regulatory circuit of the gene of interest.

Finally, since the number of functions $g_p(t)$, $g_{p,q}(t)$, $g_{p,q,r}(t)$, … is finite, we can estimate these functions and the decay rates $\lambda_1(t)$, $\lambda_2(t)$, …, $\lambda_N(t)$ if the number N of dynamic equations in Equation (12) is large enough. Equation (12) can be rewritten in an algebraic form

$$\tilde{\mathrm{X}}(t) = \tilde{\mathrm{A}}(t) \cdot \tilde{\mathrm{G}}(t) + \mathrm{E}(t), \quad (13)$$

where

$$\tilde{\mathrm{X}}(t) = \begin{pmatrix} \dot{Y}_1(t) \\ \dot{Y}_2(t) \\ \vdots \\ \dot{Y}_i(t) \\ \vdots \\ \dot{Y}_N(t) \end{pmatrix}, \quad \tilde{\mathrm{G}}(t) = \begin{pmatrix} g_1(t) \\ g_2(t) \\ \vdots \\ g_{1,2}(t) \\ \vdots \\ g_{p,q,r}(t) \\ \lambda_1(t) \\ \lambda_2(t) \\ \vdots \\ \vdots \\ \lambda_N(t) \end{pmatrix}, \quad \mathrm{E}(t) = \begin{pmatrix} \varepsilon_1(t) \\ \varepsilon_2(t) \\ \vdots \\ \vdots \\ \varepsilon_i(t) \\ \vdots \\ \vdots \\ \varepsilon_N(t) \end{pmatrix}.$$

Then, we have the following dynamic equations for all time profiles

$$\begin{pmatrix} \tilde{\mathrm{X}}(t_1) \\ \tilde{\mathrm{X}}(t_2) \\ \vdots \\ \tilde{\mathrm{X}}(t_m) \end{pmatrix} = \begin{pmatrix} \tilde{\mathrm{A}}(t_1) & 0 & \leftrightarrow & 0 \\ 0 & \tilde{\mathrm{A}}(t_2) & & \updownarrow \\ \updownarrow & & \ddots & 0 \\ 0 & \leftrightarrow & 0 & \tilde{\mathrm{A}}(t_m) \end{pmatrix} \cdot \begin{pmatrix} \tilde{\mathrm{G}}(t_1) \\ \tilde{\mathrm{G}}(t_2) \\ \vdots \\ \tilde{\mathrm{G}}(t_m) \end{pmatrix} + \begin{pmatrix} \mathrm{E}(t_1) \\ \mathrm{E}(t_2) \\ \vdots \\ \mathrm{E}(t_m) \end{pmatrix}. \quad (14)$$

Let us denote the above equations in the following simple algebraic form

$$\tilde{\mathrm{X}} = \tilde{\Phi} \cdot \tilde{\Theta} + \mathrm{M}. \quad (15)$$

In Equation (15), $\tilde{\mathrm{X}}$ and $\tilde{\Phi}$ can be calculated from microarray data for the RGC of the gene of interest, which can then be used to estimate the vector $\tilde{\Theta}$ by the least squares method, leading to the following solution:

$$\hat{\tilde{\Theta}} = (\tilde{\Phi}^T \tilde{\Phi})^{-1} \tilde{\Phi} \tilde{\mathrm{X}}. \quad (16)$$

After $\hat{\tilde{\Theta}}$ is estimated from Equation (16), the regulatory functions $\tilde{\mathrm{G}}(t)$ in Equation (13) can be reconstructed for the genes in the RGC at every time point. However, in Equation (12), the degradation rate $\lambda(t)$ is a time-varying function and is affected by both the error terms and experimental data. Therefore, in order to reduce the influence on degradation rate, we simplify Equation (12) and average the negative gradients of $\lambda(t)$ to obtain the constant value $\hat{\lambda}$. Then the estimated $\hat{\lambda}$ is substituted into Equation (7) to re-identify the *cis*-regulatory functions to derive the final regulatory functions. Using this procedure, we can avoid the effects of the time-varying function $\lambda(t)$ on the identification process and reduce the influence by different experimental data. Hence, the degradation rate can be estimated. After the functions are estimated, they can be plugged into Equation (11) and then the reconstruction of the *cis*-regulatory circuit of the gene of interest is completed.

Authors' contributions

L.H. Lin carried out the model design and computation of this study, and drafted the manuscript. H.C. Lee participated in the design of the study and drafted the manuscript. W.H. Li amended and improved the design and the presentation of the study. B.S. Chen gave the topic and suggestions and was responsible for the entire study. All authors read and approved the final manuscript.

Additional material

Additional File 1

The cell cycle genes and their connectivities to cis *elements. 769 cell cycle genes defined by Spellman* et al. *[18] were selected from a total of 6178 genes in the data set. "1" denotes the connection of a* cis *element with a gene, while "0" means no connection. The main* cis *element data were compiled from the data set of Simon* et al. *2001 [4] by choosing a* P *value (significance level) ≤ 0.0015. Under this threshold, many interactions among* cis *elements for genes confirmed by the conventional data are included [35,36]. Additionally, we modified some* cis *element data, using well-known experimental information to correct false negatives [35,36].*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-6-258-S1.xls]

Acknowledgements

We thank Ming-Che Shih, Geoff Morris, Jake Byrnes, Josh Rest and Ya-Wen Chang for helpful comments. This study was supported by an NSC grant NSC 91-2321-B-007-002, by Academia Sinica, Taiwan and by NIH grants.

References

1. Shea M, Ackers G: The OR control system of bacteriophage lambda. A physical-chemical model for gene-regulation. *J Molec Biol* 1985, 181:211-230.
2. Mjolsness E, Sharp DH, Reinitz J: **A connectionist model of development.** *J Theor Biol* 1991, **152:**429-453.
3. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO: **Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF.** *Nature* 2001, **409:**533-538.
4. Simon I, Barnett J, Hannett N, Harbison CT, Rinaldi NJ, Volkert TL, Wyrick JJ, Zeitlinger J, Gifford DK, Jaakkola TS, Young RA: **Serial regulation of transcriptional regulators in the yeast cell cycle.** *Cell* 2001, **106:**697-708.
5. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA: **Transcriptional regulatory networks in *Saccharomyces cerevisiae*.** *Science* 2002, **298:**799-804.
6. Gao F, Foat BC, Bussemaker HJ: **Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data.** *BMC Bioinformatics* 2004, **5:**31.
7. Wei H, Kaznessis Y: **Inferring gene regulatory relationships by combining target-target pattern recognition and regulator-specific motif examination.** *Biotechnol Bioeng* 2005, **89:**53-77.
8. Haverty PM, Hansen U, Weng Z: **Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification.** *Nucleic Acids Res* 2004, **32:**179-188.
9. Kato M, Hata N, Banerjee N, Futcher B, Zhang MQ: **Identifying combinatorial regulation of transcription factors and binding motifs.** *Genome Biol* 2004, **5:**R56.
10. Nachman I, Regev A, Friedman N: **Inferring quantitative models of regulatory networks from expression data.** *Bioinformatics* 2004, **20:**I248-I256.
11. Wang W, Cherry JM, Nochomovitz Y, Jolly E, Botstein D, Li H: **Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation.** *Proc Natl Acad Sci USA* 2005, **102:**1998-2003.
12. Yuh CH, Moore JG, Davidson EH: **Quantitative functional interrelations within the *cis*-regulatory system of the *S. purpuratus Endo16* gene.** *Development* 1996, **122:**4045-4056.
13. Yuh CH, Bolouri H, Davidson EH: **Genomic *cis*-regulatory logic: experimental and computational analysis of a sea urchin gene.** *Science* 1998, **279:**1896-1902.
14. Yuh CH, Bolouri H, Davidson EH: ***Cis*-regulatory logic in the *endo16* gene: switching from a specification to a differentiation mode of control.** *Development* 2001, **128:**617-629.
15. Banerjee N, Zhang MQ: **Identifying cooperativity among transcription factors controlling the cell cycle in yeast.** *Nucleic Acids Res* 2003, **31:**7024-7031.
16. Das D, Banerjee N, Zhang MQ: **Interacting models of cooperative gene regulation.** *Proc Natl Acad Sci USA* 2004, **101:**16234-16239.
17. Chen HC, Lee HC, Lin TY, Li WH, Chen BS: **Quantitative characterization of the transcriptional regulatory network in the yeast cell cycle.** *Bioinformatics* 2004, **20:**1914-1927.
18. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PQ, Bostein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization.** *Mol Biol Cell* 1998, **9:**3273-3297.
19. Michaelis S, Herskowitz I: **The a-factor pheromone of *Saccharomyces cerevisiae* is essential for mating.** *Mol Cell Biol* 1988, **8:**1309-1318.
20. Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO: **Precision and functional specificity in mRNA decay.** *Proc Natl Acad Sci USA* 2002, **99:**5860-6865.
21. de Boor C: *A Practical Guide to Splines* New York: Springer-Verlag Press; 1978.
22. Bar-Joseph Z, Gerber G, Gifford DK, Jaakkola TS, Simon I: **A new approach to analyzing gene expression time series data.** In *The Sixth Annual International Conference on Research in Computational Molecular Biology (RECOMB)* Washington DC; 2002:39-48.
23. Maher M, Cong F, Kindelberger D, Nasmyth K, Dalton S: **Cell cycle-regulated transcription of the CLB2 gene is dependent on Mcm1 and a ternary complex factor.** *Mol Cell Biol* 1995, **15:**3129-3137.
24. Koranda M, Schleiffer A, Endler L, Ammerer G: **Fork-head-like transcription factors recruit Ndd1 to the chromatin of G2/M-specific promoters.** *Nature* 2000, **406:**94-98.
25. Zhu G, Spellman PT, Volpe T, Brown PO, Botstein D, Davis TN, Futcher B: **Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth.** *Nature* 2000, **406:**90-94.
26. Kumar R, Reynolds DM, Shevchenko A, Goldstone SD, Dalton S: **Forkhead transcription factors, Fkh1p and Fkh2p, collaborate with Mcm1p to control transcription required for M-phase.** *Curr Biol* 2000, **10:**896-906.
27. McInerny CJ, Partridge JF, Mikesell GE, Creemer DP, Breeden LL: **A novel Mcm1-dependent element in the SWI4, CLN3, CDC6, and CDC47 promoters activates M/G1-specific transcription.** *Genes Dev* 1997, **11:**1277-1288.
28. Dirick L, Moll T, Auer H, Nasmyth K: **A central role for SWI6 in modulating cell cycle start-specific transcription in yeast.** *Nature* 1992, **357:**508-513.
29. Ho Y, Costanzo M, Moore L, Kobayashi R, Andrews BJ: **Regulation of transcription at the *Saccharomyces cerevisiae* start transition by Stb1, a Swi6-binding protein.** *Mol Cell Biol* 1999, **19:**5267-5278.
30. Loy CJ, Lydall D, Surana U: **Ndd1, a high-dosage suppressor of cdc28-1N, is essential for expression of a subset of late-S-phase-specific genes in *Saccharomyces cerevisiae*.** *Mol Cell Biol* 1999, **19:**3312-3327.
31. Lim FL, Hayes A, West AG, Pic-Taylor A, Darieva Z, Morgan BA, Oliver SG, Sharrocks AD: **Mcm1p-induced DNA bending regulates the formation of ternary transcription factor complexes.** *Mol Cell Biol* 2003, **23:**450-461.
32. Beer MA, Tavazoie S: **Predicting gene expression from sequence.** *Cell* 2004, **117:**185-198.
33. **Yeast Cell Cycle Analysis Project** [http://cellcycle-www.stanford.edu/]
34. McBride HJ, Yu Y, Stillman DJ: **Distinct regions of the Swi5 and Ace2 transcription factors are required for specific gene activation.** *J Biol Chem* 1999, **274:**21029-21036.
35. Cross FR, Hoek M, McKinney JD, Tinkelenberg AH: **Role of Swi4 in cell cycle regulation of CLN2 expression.** *Mol Cell Biol* 1994, **14:**4779-4787.
36. Costanzo M, Schub O, Andrews B: **G1 Transcription factors are differentially regulated in *Saccharomyces cerevisiae* by the Swi6-binding protein Stb1.** *Mol Cell Biol* 2003, **23:**5064-5077.