BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 18 2006, pages 2276–2282
doi:10.1093/bioinformatics/btl380

Systems biology

Identification of transcription factor cooperativity via stochastic system model

Yu-Hsiang Chang†, Yu-Chao Wang† and Bor-Sen Chen*

Laboratory of Control and Systems Biology, Department of Electrical Engineering, National Tsing Hua University, Hsinchu 300, Taiwan

Received on May 8, 2006; revised on June 30, 2006; accepted on July 6, 2006
Advance Access publication July 14, 2006
Associate Editor: Martin Bishop

ABSTRACT

Motivation: Transcription factor binding sites are known to co-occur in the same gene owing to cooperativity of the transcription factors (TFs) that bind to them. Genome-wide location data can help us understand how an individual TF regulates its target gene. Nevertheless, how TFs cooperate to regulate their target genes still needs further study. In this study, genome-wide location data and expression profiles are integrated to reveal how TFs cooperate to regulate their target genes from the stochastic system perspective.

Results: Based on a stochastic dynamic model, a new measurement of TF cooperativity is developed according to the regulatory abilities of cooperative TF pairs and the number of their occurrences. Our method is employed to the yeast cell cycle and reveals successfully many cooperative TF pairs confirmed by previous experiments, e.g. Swi4-Swi6 in G1/S phase and Ndd1-Fkh2 in G2/M phase. Other TF pairs with potential cooperativity mentioned in our results can provide new directions for future experiments. Finally, a cooperative TF network of cell cycle is constructed from significant cooperative TF pairs.

Contact: bschen@ee.nthu.edu.tw

Supplementary information: http://www.ee.nthu.edu.tw/~bschen/cooperativity/

1 INTRODUCTION

Precise transcriptional control is one of the major reasons for different gene expression and regulation. Owing to advances in DNA microarray technology and genome sequencing, measuring gene expression levels on a genomic scale has become possible. Measuring gene expression time profiles can help us understand mechanisms of transcriptional regulation, including functional significance of cell cycle regulation and response of environmental changes (Spellman *et al*., 1998; Gasch *et al*., 2000). However, time profiles alone are not enough to identify precisely the whole transcriptional regulatory network. More precise method also needs the binding information of transcription factors (TFs) to promoters in DNA. Genes are always regulated by a number of upstream regulatory genes through binding of TFs to specific sites in the DNA promoter region. To construct a gene regulatory network, it is important to understand the binding relationship between target genes and TFs. In recent studies, the genome-wide location (ChIP-chip) analysis is employed to obtain the binding information of TFs to promoters in DNA (Iyer *et al*., 2001; Simon *et al*., 2001; Lee *et al*., 2002; Harbison *et al*., 2004; Chang *et al*., 2005; Lin *et al*., 2005). However, these studies did not shed light on the interactions or cooperativity between TFs.

With advances in experimental approaches and abundant data sources, functional genomics has begun to investigate the more complex, cooperative TF interactions to regulate properly gene expression. In order to find cooperative TF interactions, a statistical technique was employed to identify significant homotypic or heterotypic TF binding site clusters (Wagner, 1999). Moreover, the correlation of genome-wide expression profiles was employed to uncover functional motif combinations in the promoters (Pilpel *et al*., 2001). Furthermore, genome-wide location data (ChIP-chip) and gene expression profiles were integrated to assess TF cooperativity rigorously (Banerjee and Zhang, 2003; Kato *et al*., 2004; Tsai *et al*., 2005). However, these studies only use expression correlation score in the view of statistics to determine TF cooperativity. We are not aware of the previous attempts to construct an overall transcriptional regulatory system to uncover TF cooperativity from the dynamic system perspective. In this study, we assess TF cooperativity not only in dynamic system perspective, but also in statistics. We could get more insight into the regulatory mechanism by constructing its dynamic system model and could also make more precise prediction of TF cooperativity.

A dynamic model is often employed to describe a complex and kinetic system in many fields. Systems biology and computational biology methods have recently been widely employed to describe the biological functions from the dynamic system point of view (Hasty *et al*., 2002; Davidson *et al*., 2003; Hood, 2003; Tegner *et al*., 2003; Chen *et al*., 2004). In this article, we exploit genome-wide location data (Harbison *et al*., 2004) and gene expression profiles (Spellman *et al*., 1998) to construct a dynamic regulatory model for the genes of interest. The dynamic regulatory model is developed to describe how the upstream regulatory genes control a target gene to produce the output mRNA expression through its regulatory network. In order to reduce computational complexity, we estimate TF cooperativities one gene by one gene. In the dynamic regulatory model, there are two groups of input regulations affecting expression profiles of the target gene. The first group is due to individual TF regulation, and the other is due to TF cooperativity regulation. After estimating the regulatory abilities of these two groups, we can get all TF abilities for regulating the transcriptional process of the target gene. It is more possible to have TF cooperativity if binding

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

sites for specific pairs of regulators co-occur more frequently within the same promoter regions of target genes (Harbison *et al*., 2004). Therefore, according to the estimated cooperative TF abilities and the number of TF pairs co-binding to the target genes, a statistical measurement method is specified to assess TF cooperativity rigorously. The statistical measurement method is proposed to reduce the possible estimation error owing to few data points. The new measurement of TF cooperativity is the multiplicative cooperative *p*-value, which can be obtained by multiplying all individual *p*-values of different target genes to estimate efficiently the cooperativity of any cooperative TF pairs on a genomic scale. Further, we use the shuffling method to avoid overfitting to confirm the reliability of the proposed method.

Our results show that many cooperative TF pairs that were previously characterized through experiments indeed have a high cooperativity in our analysis, thus validating the method we proposed as a TF cooperativity predictor and providing valuable information for further analysis. A cooperative network is also constructed with significant cooperative TF pairs for the yeast cell cycle. Finally, the results reveal several novel possible cooperative TF pairs not found in previous studies, thus providing new directions for future experiments.

2 METHODS

2.1 Selecting and processing experimental data

We used the genome-wide yeast microarray hybridization data of Spellman *et al*. (1998) as our mRNA expression profiles. They proposed many experimental methods for resetting the yeast cell cycle to measure mRNA expression profiles for the whole genome comprehensively. Here, we used the 768 cell cycle-related genes selected by Simon *et al*. (2001) from 'α factor' experimental cell cycle data as the target genes. 'α factor' is one experimental technique to synchronize cell cycle. The genome-wide location data are taken from Harbison *et al*. (2004), in which the genomic occupancy of 203 DNA-binding TFs is determined in yeast. We selected the significant binding set using a binding *p*-value $p < 0.0015$ as the inputs in the dynamic gene transcriptional regulatory model.

In genome-wide expression profiles of Spellman *et al*. (1998), there are many missing data. We used the cubic spline interpolation method (Faires and Burden, 1998), which employs piecewise third-order polynomials to fit data points, for compensating these missing values in the profiles. Furthermore, in order to avoid the overfitting estimation of the profiles, we also use the cubic spline method to interpolate the data points when there are more than six TFs binding to the target gene, which implies that the number of parameters to be estimated is 23. Only 5% of the target genes are bound by more than six TFs (Supplementary Material). In order to fit the gene dynamic model in the linear scale, the microarray data are returned from the $\log_2$ scale to the linear scale.

2.2 Dynamic model of gene regulatory networks

First, we consider a gene regulatory network as a system block with several regulatory genes as inputs and a target gene as output. Owing to random noise and fluctuation at the molecular level, the transcriptional behavior of the gene regulatory network is described by a stochastic discrete dynamic equation, and the general form of transcriptional regulation for a target gene is written as follows:

$$y[t+1] = a \cdot y[t] + \sum_{i=1}^{N} b_i \cdot x_i[t] + G[t] + k + \varepsilon[t], \qquad (1)$$

where $y[t]$ represents mRNA expression level of the target gene at time point t, and the parameter a indicates the effect of the present state value $y[t]$ to the next state value $y[t+1]$. $x_i[t]$ $i \in \{1, 2, \ldots, N\}$ represent the regulation functions of N TFs binding to the target gene and b_i indicates the regulatory ability of the i-th TF. $G[t]$ is the possible regulatory function of cooperative TFs. k represents the basal level from other factors, and $\varepsilon[t]$ denotes a stochastic noise owing to model uncertainty and fluctuation of mRNA microarray data in the target gene.

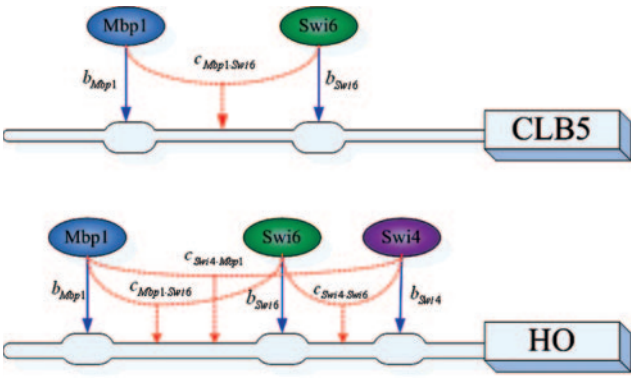

Fig. 1. Examples of dynamic transcriptional regulatory models of target genes CLB5 and HO. The candidates of regulatory TFs of target gene were obtained from the genome-wide location data of Harbison *et al*. (2004) when *p*-value was chosen as $p < 0.0015$. In the examples, the regulatory TFs of CLB5 are Mbp1 and Swi6, and the regulatory TFs of HO are Mbp1, Swi6 and Swi4. The regulations of individual TF are represented by solid lines and the regulations of possible TF cooperativities are represented by dotted lines.

For system identification of gene regulatory network, it is more practical to consider the biochemical reaction between the TF regulation functions $x_i[t]$ at the motif binding sites and their relevant mRNA expression profiles $y_i[t]$ of the upstream regulatory gene (Goldbeter and Koshland, 1981; Mestl *et al*., 1995). For this purpose, we describe the binding regulation function $x_i[t]$ of the TF as the following sigmoid function

$$x_i[t] = f_i(y_i[t]) = \frac{1}{1 + \exp[-r(y_i[t] - M_i)]}, \qquad (2)$$

where r denotes the transition rate of the sigmoid function and M_i denotes the mean of mRNA expression level of the regulatory gene i. $y_i[t]$ and $x_i[t]$ represent the mRNA expression profiles of the i-th regulatory gene and the binding regulation function of the corresponding TF, respectively. The sigmoid function can also be considered as a method for normalizing the expression profiles of regulatory genes between 0 and 1, which has been successfully employed to describe the binding of regulatory genes (Chen *et al*., 2004).

The regulatory function $G[t]$ is combined by all possible TF cooperativities of the target gene. We describe the regulatory function $G[t]$ in Equation (1) as the following:

$$G[t] = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} c_{i \cdot j} \cdot x_{i \cdot j}[t], \qquad (3)$$

where $x_{i\cdot j}[t] \equiv f_{i\cdot j}(y_i[t] \cdot y_j[t])$ is a sigmoid function of product $y_i[t] \cdot y_j[t]$ to denote the binding function of cooperative TFs i and j, and $c_{i\cdot j}$ denotes the regulatory ability (or kinetic activity) of the cooperative TFs i and j. N denotes the number of TFs binding to the target gene.

Substituting Equation (3) into Equation (1), we get the following dynamic transcriptional regulatory equation:

$$y[t+1] = a \cdot y[t] + \sum_{i=1}^{N} b_i \cdot x_i[t] + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} c_{i \cdot j} \cdot x_{i \cdot j}[t] + k + \varepsilon[t]. \qquad (4)$$

In Figure 1, there are examples of constructing the discrete dynamic models for target genes. After describing the general stochastic dynamic model of

transcriptional regulation, we could identify the models of gene expression from the available microarray data.

2.3 Identifying gene regulatory networks

After constructing the discrete stochastic dynamic model, we use the method of maximum likelihood to estimate the parameters. The details are shown in Supplementary Material.

After identifying the parameters, the transcriptional regulatory network of target genes could be expressed as the following dynamic equation:

$$y[t+1] = \hat{a} \cdot y[t] + \sum_{i=1}^{N} \hat{b}_i \cdot x_i[t] + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \hat{c}_{i \cdot j} \cdot x_{i \cdot j}[t] + \hat{k}, \qquad (5)$$

where $\hat{a}$, $\hat{b}_i$, $\hat{c}_{i \cdot j}$, $\hat{k}$ are the estimated parameters of a, b_i, $c_{i,j}$, k, respectively. Therefore, the TF cooperativity at each target gene could be evaluated by $\hat{c}_{i \cdot j}$, the regulatory ability of the cooperative TF pair, in Equation (5). However, $\hat{c}_{i \cdot j}$ in (5) is only from one target gene point of view. A statistical method should be developed to measure the cooperativity of TFs to combine all $\hat{c}_{i \cdot j}$ of the TF i and j for all target genes (i.e. on a genomic scale). It would be discussed in the following section.

2.4 Measurement of TF cooperativity

In order to measure the possible TF cooperativity, we define a new measurement, the multiplicative cooperative p-value P_C, according to the statistics of regulatory abilities $\hat{c}_{i \cdot j}$ of all target genes in the dynamic model of Equation (5) and the number of TF pairs appearing in all target genes. For any TF pair, if the binding sites of specific TF pairs co-occur more frequently in the target genes, then it is more likely to have cooperativity between these two TFs (Harbison *et al.*, 2004). In addition, the magnitude of the regulatory ability $\hat{c}_{i \cdot j}$ of TF cooperation in our dynamic model can be also useful to help us determine the significant TF cooperativity because the magnitude represents its importance in the transcriptional regulation of the target gene.

To define the multiplicative cooperative p-value P_C on a genomic scale, we first calculate the individual p-value for each regulatory ability of a cooperative TF pair. The p-value is defined as $(P_B)_m$ when a cooperative TF pair binds to the m-th target gene. For the m-th target gene bound by the cooperative TF pair i and j, we rewrote its regulatory ability $\hat{c}_{i \cdot j}$ as $\hat{c}^m_{i \cdot j}$. After constructing 1000 random permutations of $x_{i \cdot j}[t]$ (Supplementary Material) in Equation (4) for the m-th target gene bound by TFs pair i and j, we used these random permutations to estimate 1000 different $\hat{c}^m_{i \cdot j}$ rewritten as $\tilde{c}^m_{i \cdot j}$ and then computed a probability density $p_{|\tilde{c}^m_{i \cdot j}|}(x)$ of all absolute values $|\tilde{c}^m_{i \cdot j}|$ according to their magnitudes by normalization. The probability density distribution $p_{|\tilde{c}^m_{i \cdot j}|}(x)$ of cooperativity is shown in Figure 2. Then, using $p_{|\tilde{c}^m_{i \cdot j}|}(x)$, we calculated the individual p-value $(P_B)_m$ of TFs i and j at the m-th target gene by

$$(P_B)_m = \int_{|\hat{c}^m_{i \cdot j}|}^{\infty} p_{|\tilde{c}^m_{i \cdot j}|}(x)\mathrm{d}x, \qquad (6)$$

where $|\hat{c}^m_{i \cdot j}|$ denotes the regulatory ability of the cooperative TF pair i and j at the m-th target gene without random permutation.

When calculating individual p-value for each regulatory ability $(P_B)_m$, we use 1000 random permutations to infer the probability density function $p_{|\tilde{c}^m_{i \cdot j}|}(x)$. Theoretically, the continuous probability density function is obtained when the number of shuffling is infinite, which is not applicable in practice. Thus, we use the discrete normalized histogram to approximate the probability density function instead. In Figure 2, $(P_B)_m$ should be the area under the smooth curve (pdf), but as we can calculate only the approximation of the pdf, $(P_B)_m$ would be in fact the numbers of $|\tilde{c}^m_{i \cdot j}|$, which are greater than $|\hat{c}^m_{i \cdot j}|$ normalized by the number of shuffling, i.e. the sum of heights of corresponding bars. In reality, the minimal value of $(P_B)_m$ is equal to the inverse number of shuffling, i.e. 0.001 in the case of 1000 shuffling. Thus, taking a greater number of shuffling would make quite a different $(P_B)_m$ value. However, choosing the number of shuffling would not affect the final

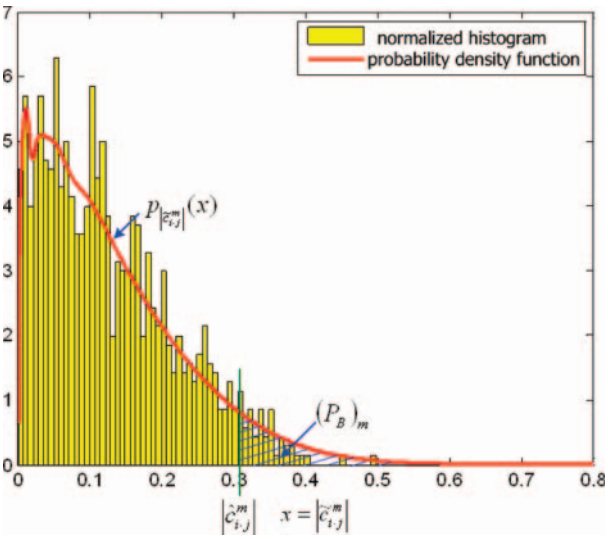

Fig. 2. Probability density function $p_{|\tilde{c}^m_{i \cdot j}|}(x)$ and normalized histogram for $|\tilde{c}^m_{i \cdot j}|$. The pdf $p_{|\tilde{c}^m_{i \cdot j}|}(x)$ of all $|\tilde{c}^m_{i \cdot j}|$ is obtained according to the number of $|\tilde{c}^m_{i \cdot j}|$ occurrence estimated by random permuting $x_{i \cdot j}[t]$ for the m-th target gene bound by TF pair i and j. The individual p-value $(P_B)_m$ of the m-th target gene is the sum of heights of corresponding bars in practice.

significant cooperative TF pairs, as long as the number is large enough. So in order to reduce the computation complexity, we choose the number of shuffling to be 1000.

A TF pair may co-occur to bind many target genes, so we could obtain many individual p-values $(P_B)_m$, $m \in \{1, 2, \ldots, L\}$, where L is the number of the target genes bound by the possible cooperative TF pair i and j. In Figure 1, for example, TFs Mbp1 and Swi6 both bind to HO and CLB5. Therefore, the gene HO has its $c_{Mbp1 \cdot Swi6}$ and the corresponding $(P_B)_{HO}$, and so does CLB5 with its $c_{Mbp1 \cdot Swi6}$ and the corresponding $(P_B)_{CLB5}$. After computing all the individual p-value $(P_B)_m$ for the TF pair i and j for all target genes, i.e. $(P_B)_m$ for $m \in \{1, 2, \ldots, L\}$, we define the multiplicative cooperative p-value P_C for the TF pair i and j by multiplying all of its $(P_B)_m$ as follows:

$$P_C = \prod_{m=1}^{L} (P_B)_m \qquad (7)$$

where L is the number of target genes bound by the possible cooperative TF pair i and j, and $(P_B)_m$ denotes the individual p-value of the cooperative TF pair i and j binding to the m-th target gene.

In Equations (6) and (7), the more likely the TF cooperativity is, the smaller P_C of the TF pair is. Therefore, there are two situations that make P_C smaller. One is the larger number L of the TF pairs co-occurring in all target genes, i.e. with a larger number of individual p-values for the TF pair. This situation is the same with the result that the TF pair has a greater possibility of cooperation if binding sites of specific TF pairs co-occur more frequently in target genes (Harbison *et al.*, 2004). Another situation is that the regulatory ability $|\hat{c}_{i \cdot j}|$ of the TF pair i and j is so significant at some target genes that the individual p-value $(P_B)_m$ of the TF pair is small at the target genes. In this way, the multiplicative cooperative p-value P_C also becomes small. Therefore, it is reasonable to use the multiplicative cooperative p-value P_C to evaluate the TF cooperativity.

2.5 Determination of significant cooperative TF pairs

After calculation of the multiplicative cooperative p-value P_C, we have a list of cooperative TF pairs. The list has $\binom{N}{2}$ pairs with their P_C values,

respectively, where N represents the total number of TFs. Sorted by the P_C values (the smaller P_C means the more likely cooperative TF pair), it can be presented as a list in which we can find the most possible cooperative TF pairs to the least possible ones. However, which TF pairs are considered as significant cooperative TF pairs should be determined.

Because the measurement of TF cooperativity depends on the individual p-values P_B and the number of the target genes bound by the possible cooperative TF pair L, we determine the threshold of significant cooperative pairs by P_B and L as well. The idea is that we determine the threshold of P_B and L, respectively, and then calculate the significance threshold as

$$P_{C,\,\text{Threshold}} = P_{B,\,\text{Threshold}}^{L_{\text{Threshold}}} \tag{8}$$

$P_{C,\text{Threshold}}$ is defined as $P_{B,\text{Threshold}}$ to the power of $L_{\text{Threshold}}$, where $P_{B,\text{Threshold}}$ and $L_{\text{Threshold}}$ denote the threshold of P_B and L, respectively. How to determine $P_{B,\text{Threshold}}$ and $L_{\text{Threshold}}$ is as follows. We first construct P_B distribution and L distribution according to all P_B values and all L values. Then at the significance level 0.05 of P_B distribution and L distribution, we can determine the significance thresholds of P_B and L, respectively. Because a smaller P_B means that the P_B is more significant, so $P_{B,\text{Threshold}}$ is determined as the fifth percentile of the P_B distribution. For $L_{\text{Threshold}}$, a larger L means that the L is more significant, so $L_{\text{Threshold}}$ is determined as the 95th percentile of the L distribution. By $P_{B,\text{Threshold}}$, $L_{\text{Threshold}}$ and (8), we can calculate the significance threshold $P_{C,\text{Threshold}}$ and determine the significant cooperative TF pairs, which P_C are smaller than $P_{C,\text{Threshold}}$. Once the significance threshold $P_{C,\text{Threshold}}$ is determined, whether or not P_C is significant at that level is binary.

3 RESULTS

3.1 Analysis of TF cooperativity

We integrated the expression profiles of the yeast cell cycle taken from Spellman *et al*. (1998) and genome-wide location data of Harbison *et al*. (2004) to identify the cooperativity of all possible TF interaction pairs among 203 TFs. After constructing and identifying parameters of gene transcriptional regulatory networks of yeast, we calculated the multiplicative cooperative p-value P_C of each possible TF interaction pair. Sorted by P_C values, all possible TF cooperative pairs are ranked according to their cooperative possibilities. We then calculate the significance threshold $P_{C,\text{Threshold}}$ and determine the significant cooperative TF pairs. In this study, the significance threshold $P_{C,\text{Threshold}} = 10^{-21}$ and we identified 55 significant cooperative TF pairs, which are shown in Table 1. According to these results, the cooperative network of significant cooperative TF pairs is given in Figure 3. Of the results listed, 72.73% are confirmed by evidence in many studies. Most of these TF cooperative pairs confirmed are related to the cell cycle of yeast. It may be because our source data are the yeast cell cycle expression profiles so that most of these TF cooperative pairs found in the list are related to cell cycle. Moreover, the TF cooperative pairs that have not yet been proved but are found by the method we proposed can provide a direction for future experiments. The details of the TF cooperativities we found are discussed below.

3.2 Cell cycle

3.2.1 G_1/S Phase (Swi4-Swi6, Mbp1-Swi6, Mbp1-Swi4, Stb1-Swi6) The G_1 to S phase transition of the eukaryotic cell cycle is a critical point for the coordination of cell cycle progression with cellular growth. SBF and MBF are sequence-specific TFs that activate gene expression to mediate the G_1/S transition of the cell cycle in yeast (Iyer *et al*., 2001). SBF is a protein complex of Swi4 and Swi6, and MBF is a protein complex of Mbp1 and Swi6 (Koch *et al*., 1993). The related Swi4 and Mbp1 proteins are the DNA-binding components of the respective factors, and Swi6 may have a regulatory function (Dirick *et al*., 1992; Primig *et al*., 1992). The cooperativity between Swi4 and Swi6 is included in our results, and so is the cooperativity between Mbp1 and Swi6. In addition, Mbp1 and Swi4 sharing 50% identity in their DNA binding domains are found with the probable cooperativity in our results (Koch *et al*., 1993).

In the late G_1 phase, Stb1 is an important regulator controlling the timing of start transcription that is revealed in the absence of the G_1 regulator Cln3 and binds to Swi6 *in vivo* (Ho *et al*., 1999). The interaction of Stb1 and Swi6 confirmed by Ho *et al*. (1999) and Costanzo *et al*. (2003) is also found in our results.

3.2.2 G_2/M Phase (Ndd1-Fkh1, Ndd1-Fkh2, Fkh1-Fkh2, Mcm1-Fkh2, Ndd1-Mcm1) In the G_2/M phase, the important regulator is SFF, which is a larger transcription factor containing Ndd1, Fkh1 and Fkh2 (Koranda *et al*., 2000; Kumar *et al*., 2000; Pic *et al*., 2000; Futcher, 2002). Our results strongly indicate the cooperativity between Ndd1 and Fkh2, which are components of SFF. Another cooperativity between Ndd1 and Fkh1, which are also components of SFF, is found in our results, too. In addition, Fkh1 and Fkh2 are found with cooperativity in our results. Fkh1 and Fkh2 share 72% identical DNA binding domain, and the double mutant of Fkh1 and Fkh2 displays obvious morphological change (Kumar *et al*., 2000).

SFF is thought to regulate a program of mitotic transcription in conjunction with the transcription factor Mcm1. Moreover, Fkh2, a component of SFF, assembles into a ternary complex with Mcm1 on both the SWI5 and CLB2 cell cycle genes (Koranda *et al*., 2000; Kumar *et al*., 2000). The fact that Fkh2 shows cooperativity with Mcm1 is also listed in our result. In addition to Ndd1 and Fkh2, Mcm1 and Fkh2, another cooperative TF pair in the G_2/M phase, Ndd1 and Mcm1, is found in our results. For the above cooperative TF pairs found in the G_2/M phase, it is not surprising to find experimentally that Mcm1, Fkh2 and Ndd1 also form a complex to regulate the CLB2 gene and other genes (Kumar *et al*., 2000; Zhu *et al*., 2000).

3.2.3 M/G_1 Phase (Ace2-Swi5) In the M/G_1 phase, we found that the TF pair, Ace2 and Swi5, is cooperative. Ace2 and Swi5 is a pair of TFs of yeast that regulates the expression of many cell cycle-specific genes (Doolin *et al*., 2001). In recent studies, Ace2 and Swi5 cooperate to induce the expressions of a subset of genes, but the antagonistic interaction between Ace2 and Swi5 has been found (Doolin *et al*., 2001). With 82% identical DNA binding domains, Ace2 and Swi5 bind to the same DNA sequence (McBride *et al*., 1999), and it is possible that proteins compete for access to these promoters, but only one activates transcription (Doolin *et al*., 2001). Therefore, one partner of Swi5 and Ace2 sometimes can have a stronger contribution towards regulation, and the finding of antagonistic interaction of Ace2 and Swi5 is not surprising.

3.3 Mating (Ste12–Dig1)

The TF Ste12 is responsible for activating genes in response to MAP kinase cascades controlling mating and filamentous growth. Two inhibitors Dig1 and Dig2 regulate Ste12 negatively (Olson *et al*., 2000). It was found that Dig1 and Dig2 do not function through redundant mechanisms, but rather inhibit pheromone-responsive transcription through interactions with separate regions

Table 1. Significant cooperative TF pairs of cell cycle target genes

Ranking	TF1	TF2	P_C*	Literature evidences
1	Swi4	Swi6	2.68E−162	Kumar *et al*. (2000); Koch *et al*. (1993); Manke *et al*. (2003); Banerjee and Zhang (2003); Tsai *et al*. (2005)[a]
2	Mbp1	Swi6	7.32E−125	Koch *et al*. (1993); Manke *et al*. (2003); Banerjee and Zhang (2003); Tsai *et al*. (2005)[a]
3	Fkh2	Ndd1	1.55E−88	Koranda *et al*. (2000); Manke *et al*. (2003); Banerjee and Zhang (2003); Tsai *et al*. (2005)[a]
4	Mbp1	Swi4	4.07E−83	Koch *et al*. (1993); Manke *et al*. (2003)
5	Gat3	Yap5	1.40E−76	Manke *et al*. (2003)
6	Fkh2	Swi6	7.30E−59	Tsai *et al*. (2005)[a]
7	Fkh1	Fkh2	7.25E−58	Kumar *et al*. (2000); Zhu *et al*. (2000); Pic *et al*. (2000); Manke *et al*. (2003); Banerjee and Zhang (2003)
8	Pdr1	Yap5	2.69E−49	Manke *et al*. (2003)
9	Fkh2	Swi4	1.00E−48	Manke *et al*. (2003); Tsai *et al*. (2005)[a]
10	Mcm1	Ndd1	2.46E−46	Kumar *et al*. (2000); Koranda *et al*. (2000); Manke *et al*. (2003); Banerjee and Zhang (2003); Tsai *et al*. (2005)[a]
11	Fkh2	Mbp1	8.94E−46	Manke *et al*. (2003); Tsai *et al*. (2005)[b]
12	Fkh1	Swi6	4.00E−45	
13	Gat3	Pdr1	3.42E−44	Manke *et al*. (2003); Banerjee and Zhang (2003); Tsai *et al*. (2005)[b]
14	Fkh1	Ndd1	8.97E−43	Koranda *et al*. (2000); Banerjee and Zhang (2003); Tsai *et al*. (2005)[a]
15	Ndd1	Swi6	1.00E−42	
16	Msn4	Yap5	1.76E−41	Banerjee and Zhang (2003); Tsai *et al*. (2005)[b]
17	Gat3	Msn4	1.00E−39	Banerjee and Zhang (2003); Tsai *et al*. (2005)[b]
18	Fkh2	Mcm1	1.41E−39	Kumar *et al*. (2000); Pic *et al*. (2000); Manke *et al*. (2003); Banerjee and Zhang (2003); Tsai *et al*. (2005)[b]
19	Stb1	Swi6	1.58E−38	Ho *et al*.; Costanzo *et al*. (2003)
20	Cin5	Yap6	2.21E−37	Manke *et al*. (2003); Banerjee and Zhang (2003)
21	Ndd1	Swi4	7.19E−37	Manke *et al*. (2003); Tsai *et al*. (2005)[a]
22	Gat3	Hap4	1.00E−36	
23	Hap4	Yap5	1.00E−36	Tsai *et al*. (2005)[b]
24	Fkh1	Mbp1	1.95E−36	Tsai *et al*. (2005)[a]
25	Stb1	Swi4	4.80E−35	Costanzo *et al*. (2003); Manke *et al*. (2003); Banerjee and Zhang (2003); Tsai *et al*. (2005)[b]
26	Ste12	Swi6	1.54E−34	
27	Ste12	Swi4	3.06E−34	
28	Gat3	Rap1	2.27E−33	
29	Msn4	Pdr1	6.00E−33	Tsai *et al*. (2005)[b]
30	Rap1	Yap5	6.34E−33	
31	Dig1	Ste12	2.28E−32	Olson *et al*. (2000); Manke *et al*. (2003); Tsai *et al*. (2005)[b]
32	Mbp1	Ndd1	4.90E−32	Manke *et al*. (2003)
33	Dat1	Yap5	1.85E−31	
34	Ace2	Swi5	7.77E−31	Doolin *et al*. (2001); Manke *et al*. (2003); Banerjee and Zhang (2003)
35	Skn7	Swi6	1.61E−29	
36	Dat1	Gat3	1.00E−27	
37	Dat1	Hap1	1.00E−27	Tsai *et al*. (2005)[a]
38	Dat1	Hap4	1.00E−27	
39	Rlm1	Swi4	1.00E−27	
40	Rlm1	Swi6	1.00E−27	
41	Mcm1	Swi6	1.94E−27	
42	Fkh1	Mcm1	6.80E−26	Kumar *et al*. (2000); Tsai *et al*. (2005)[b]
43	Mbp1	Stb1	9.80E−26	Costanzo *et al*. (2003)
44	Skn7	Swi4	1.99E−25	Manke *et al*. (2003)
45	Gat3	Rgm1	1.00E−24	Manke *et al*. (2003)
46	Pdr1	Rgm1	1.00E−24	Manke *et al*. (2003)
47	Hap1	Msn4	2.00E−24	Tsai *et al*. (2005)[b]
48	Hap4	Pdr1	3.00E−24	Tsai *et al*. (2005)[b]
49	Pdr1	Smp1	3.00E−24	Manke *et al*. (2003); Banerjee and Zhang (2003)
50	Dat1	Msn4	4.00E−24	Tsai *et al*. (2005)[b]
51	Hap1	Rap1	9.50E−23	Tsai *et al*. (2005)[a]
52	Hap1	Yap5	2.40E−22	Tsai *et al*. (2005)[b]
53	Rgm1	Yap5	3.41E−22	Manke *et al*. (2003)
54	Swi5	Yap5	3.60E−22	Tsai *et al*. (2005)[b]
55	Hap1	Hap4	5.66E−22	

The TF pairs are sorted by P_C value.
*List of the results with P_C smaller than $P_{C,\text{Threshold}} = 10^{-21}$.
[a]Confident synergistic TF pairs as stated in Tsai *et al*.'s study (2005).
[b]Plausible or doubtful synergistic TF pairs as stated in Tsai *et al*.'s study (2005).

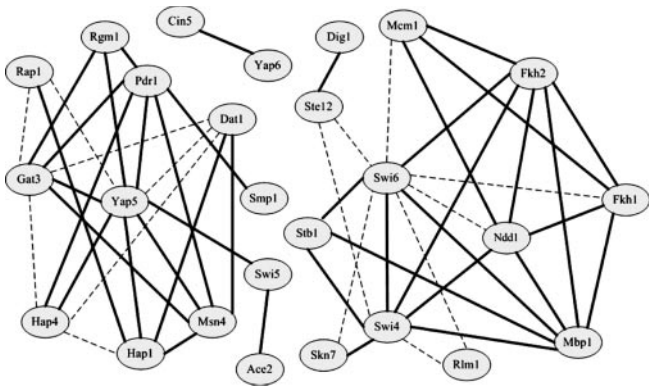

Fig. 3. The significant cooperative TF network of cell cycle target genes. The cooperative TF pairs confirmed by literature evidences are shown in solid lines, and those still to be confirmed are expressed in dotted lines.

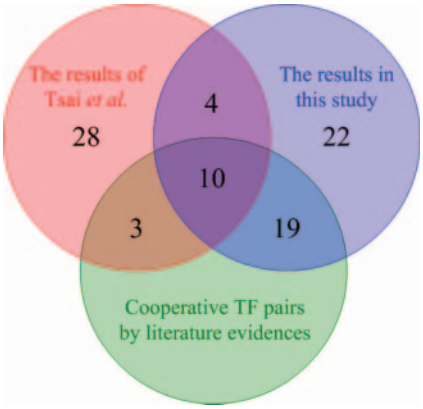

Fig. 4. Overlap of our results, Tsai *et al*.'s (2005) results, and cooperative TF pairs supported by literature evidences. There are total 45 and 55 cooperative pairs in Tsai *et al*.'s results (confident synergistic TF pairs as stated in their study) and in our results, respectively. By comparing both results, we can easily find that the confirmed rate of our method is higher than that obtained by Tsai *et al*.'s method.

of Ste12 (Olson *et al*., 2000). In our results, we indeed found that Ste12 cooperates with Dig1. However, in the location data of Harbison *et al*., Dig2 is not listed in the 203 TF. That is why another protein Dig2 was not found to show cooperativity with Ste12 in our results. If the genome-wide location data are comprehensive for more TFs, we believe that our method can predict more cooperative TF pairs, like Dig2 and Ste12.

3.4 Comparison with results of other methods

Pilpel *et al*. (2001) uncover functional motif combinations in the promoters of *Saccharomyces cerevsiae* using microarray data. They uncover not only cell cycle-related motifs, but also sporulation and various stress responses. Focusing on cell cycle, they identified only 10 motif pairs. Comparing our results with the cell cycle results from Pilpel *et al*. (2001), there are only three TF pairs in common (Supplementary Material). This may be because they only use microarray data but not use ChIP-chip data to infer combinatorial motifs.

Banerjee and Zhang (2003) integrated genome-wide location data from Lee *et al*. (2002) and gene expression data from Cho *et al*. (1998) to infer cooperativity among transcription factors by expression correlation. Comparing the TF cooperativities we found with Banerjee and Zhang's results show that many cooperative TF pairs confirmed by literature evidences are found in both results even though their study based on different dataset (Supplementary Material). Furthermore, our results indicate more cooperative TF pairs confirmed by literature evidences but not found by Banerjee and Zhang's methods. However, there are still four pairs, namely Arg80–Arg81, Ace2–Hsf1, Hsf1–Skn7 and Hir1–Hir2, found in Banerjee and Zhang's results but not in our results. The details can be seen in Supplementary Material.

Tsai *et al*. (2005) use statistical methods (ANOVA) to identify synergistic pairs of yeast cell cycle TFs. They combined ChIP-chip data from Harbison *et al*. (2004) and microarray data from Spellman *et al*. (1998) as we did. Comparing our results with Tsai *et al*.'s results (confident synergistic TF pairs as stated in their study), we find that many cooperative TF pairs confirmed by literature evidences are found in both results. Once more, our method finds more real cooperative TF pairs than Tsai *et al*.'s method. The overlap of our results, Tsai *et al*.'s results and cooperative TF pairs supported by literature evidences is shown in Figure 4. In addition, there are four pairs, Fkh2-Swi6, Fkh1-Mbp1, Dat1-Hap1 and Hap1-Rap1, which are still not confirmed by literature evidences, are also found by both results. These TF pairs provide directions for future experiments. From Figure 4, we also find that there are three pairs, Hir1–Hir2, Hir1–Hir3 and Hir2–Hir3, found in Tsai *et al*.'s results but not in ours. These three TF pairs could be considered as false negative TF pairs of our method. After checking these three cooperative TF pairs in detail, we find that Hir1 and Hir2 co-occurred to bind only six target genes of the 768 cell cycle genes from the location data of Harbison *et al*. (2004) at p-value $p < 0.0015$. For this cooperative TF pair, the number L in Equation (7) is only 6, and that is why its P_C is not small enough to be considered significant. The situations of Hir1–Hir3 and Hir2–Hir3 are the same, thus these three pairs are considered as false negatives in our results. Finally, we calculated the confirmed rates, which are the rates of both results compared with literature evidences, of the TF pairs listed in our findings and Tsai *et al*.'s results (confident synergistic TF pairs as stated in their study). The confirmed rate of our method, 52.73%, is higher than that of Tsai *et al*.'s, 28.89%.

4 DISCUSSION

In this study, we successfully identified cooperative TF pairs by integrating genome-wide location data and expression profiles. Our method is based on transcriptional regulatory mechanisms of TFs and their cooperativities at binding sites from dynamic system perspective. Unlike others' studies using statistical correlation (Pilpel *et al*., 2001; Banerjee and Zhang, 2003; Tsai *et al*., 2005), our method provides the view of dynamic regulatory systems to mimic the transcriptional procedure. The measure of TF cooperativity of TFs is developed by a multiplicative cooperative p-value according to the statistics of estimated regulatory ability of cooperative genes in the dynamic regulatory model. We developed a method to determine significant cooperative TF pairs among all possible TF cooperativities as well. Our results showed that many cooperative TF pairs identified by our method are confirmed by literature evidences. We also indicated several possible TF pairs with cooperativity that are listed in our results but not confirmed yet.

These provide new directions for future experiments. Moreover, it was shown that the confirmed rate of our method is higher than other methods using statistical correlation (Pilpel *et al*., 2001; Banerjee and Zhang, 2003; Tsai *et al*., 2005).

From the cooperative network of significant TF cooperative pairs shown in Figure 3, we can find that the interaction links among several TFs (Swi4, Swi6, Fkh1, Fkh2, Ndd1 and Mbp1) are more compact than others (compactness means that there are more interaction links with the TFs). All of these TFs have important roles in the yeast cell cycle, so it is likely that there exist TF cooperativities among these TFs. However, some of these TFs may co-bind to many target genes because of their importance for the yeast cell cycle but nevertheless they show no cooperativity among them. In our analysis, this situation will lead to a false positive (the false positive pairs are listed in the Supplementary Material). It may be another possible reason why the interaction links among these important TFs are more compact. We also list the TF cooperative pairs that are confirmed by literature evidences but not shown up in our results, i.e., the false negatives. The false negative rate is 27.27% (Supplementary Material).

In order to confirm the reliability of the proposed method, the shuffling method was used to test the overfitting. In the expression data shuffling case, only 60% of TF cooperativities are found, which may be due to the correct ChIP-chip data and the uncertainty reduction by *p*-value detection method. In the location data shuffling, no cooperativities of TFs are found due to the same significance threshold $P_{C,\text{Threshold}}$ as the original result without shuffling, i.e. 10^{-21}. When comparing the results, we found that there is no overfitting by the proposed method (Supplementary Material).

With the microarray data of the yeast cell cycle of Spellman *et al*. (1998) and the ChIP-chip data of Harbison *et al*. (2004), it is shown that our method can successfully predict the cooperative TF pairs of the yeast cell cycle. If the gene expression profiles and genome-wide location data of other process for yeast or other species, such as the heat shock stress, are available, our method could be employed to identify the cooperative TF pairs of different stress conditions or other species (Supplementary Material). Furthermore, our method can be extended to identify the cooperativities among more than two TFs by adding extra terms generated by multiplications of their expression profiles through the sigmoid function. Moreover, we can integrate protein-protein interaction data to determine interactions not only among TFs but also among TFs and other proteins in our model (Supplementary Material). That is, by modifying the method we proposed, we can construct an integrated cellular network of transcription regulation and protein–protein interaction.

ACKNOWLEDGEMENTS

The authors thank Professor Wen-Ping Hsieh for her valuable discussions and comments.

Conflict of Interest: none declared.

REFERENCES

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

Chang,W.C. *et al.* (2005) Quantitative inference of dynamic regulatory pathways via microarray data. *BMC Bioinformatics*, **6**, 44.

Chen,H.C. *et al.* (2004) Quantitative characterization of the transcriptional regulatory network in the yeast cell cycle. *Bioinformatics*, **20**, 1914–1927.

Cho,R.J. *et al*. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.

Costanzo,M. *et al.* (2003) G1 transcription factors are differentially regulated in *Saccharomyces cerevisiae* by the Swi6-binding protein Stb1. *Mol. Cell Biol*., **23**, 5064–5077.

Davidson,E.H. *et al*. (2003) Regulatory gene networks and the properties of the developmental process. *Proc. Natl. Acad. Sci. USA*, **100**, 1475–1480.

Dirick,L. *et al*. (1992) A central role for SWI6 in modulating cell cycle start-specific transcription in yeast. *Nature*, **357**, 508–13.

Doolin,M.T. *et al*. (2001) Overlapping and distinct roles of the duplicated yeast transcription factors Ace2p and Swi5p. *Mol. Microbiol*., **40**, 422–432.

Faires,J.D. and Burden,R. (1998) *Numerical Methods*, 2nd edn, Brooks/Cole Publishing Company, Pacific Grove, CA, pp. 87–91.

Futcher,B. (2002) Transcriptional regulatory networks and the yeast cell cycle. *Curr. Opin. Cell Biol*., **14**, 676–683.

Gasch,A.P. *et al*. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.

Goldbeter,A. and Koshland,D.E.,Jr. (1981) An amplified sensitivity arising from covalent modification in biological systems. *Proc. Natl. Acad. Sci. USA*, **78**, 6840–6844.

Harbison,C.T. *et al*. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Hasty,J. *et al*. (2002) Engineered gene circuits. *Nature*, **420**, 224–230.

Ho,Y. *et al*. (1999) Regulation of transcription at the *Saccharomyces cerevisiae* start transition by Stb1, a Swi6-binding protein. *Mol. Cell Biol*., **19**, 5267–5278.

Hood,L. (2003) Systems biology: integrating technology, biology, and computation. *Mech. Ageing. Dev*., **124**, 9–16.

Iyer,V.R. *et al*. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533–538.

Kato,M. *et al*. (2004) Identifying combinatorial regulation of transcription factors and binding motifs. *Genome Biol*., **5**, R56.

Koch,C. *et al*. (1993) A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase. *Science*, **261**, 1551–1557.

Koranda,M. *et al*. (2000) Forkhead-like transcription factors recruit Ndd1 to the chromatin of G2/M-specific promoters. *Nature*, **406**, 94–98.

Kumar,R. *et al*. (2000) Forkhead transcription factors, Fkh1p and Fkh2p, collaborate with Mcm1p to control transcription required for M-phase. *Curr. Biol*., **10**, 896–906.

Lee,T.I. *et al*. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.

Lin,L.H. *et al*. (2005) Dynamic modeling of *cis*-regulatory circuits and gene expression prediction via cross-gene identification. *BMC Bioinformatics*, **6**, 258.

Manke,T. *et al*. (2003) Correlating protein–DNA and protein–protein interaction networks. *J. Mol. Biol*., **333**, 75–85.

McBride,H.J. *et al*. (1999) Distinct regions of the Swi5 and Ace2 transcription factors are required for specific gene activation. *J. Biol. Chem*., **274**, 21029–21036.

Mestl,T. *et al*. (1995) A mathematical framework for describing and analyzing gene regulatory networks. *J. Theor. Biol*., **176**, 291–300.

Olson,K.A. *et al*. (2000) Two regulators of Ste12p inhibit pheromone-responsive transcription by separate mechanisms. *Mol. Cell Biol*., **20**, 4199–4209.

Pic,A. *et al*. (2000) The forkhead protein Fkh2 is a component of the yeast cell cycle transcription factor SFF. *EMBO J*., **19**, 3750–3761.

Pilpel,Y. *et al*. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet*., **29**, 153–159.

Primig,M. *et al*. (1992) Anatomy of a transcription factor important for the start of the cell cycle in *Saccharomyces cerevisiae*. *Nature*, **358**, 593–597.

Simon,I. *et al*. (2001) Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, **106**, 697–708.

Spellman,P.T. *et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cererisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Tegner,J. *et al*. (2003) Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proc. Natl. Acad. Sci. USA*, **100**, 5944–5949.

Tsai,H.K. *et al*. (2005) Statistical methods for identifying yeast cell cycle transcription factors. *Proc. Natl. Acad. Sci. USA*, **102**, 13532–13537.

Wagner,A. (1999) Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, **15**, 776–784.

Zhu,G. *et al*. (2000) Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. *Nature*, **406**, 90–94.