

國立交通大學

統計學研究所

碩士論文

利用相對性 R squared 方法

辨認酵母菌轉錄因子



Yeast Cell Cycle Transcription Factors

Identification by the Relative R squared Method

研究生：王郁涵

指導教授：王秀瑛 教授

吳謂勝 教授

中華民國九十九年六月

Yeast Cell Cycle Transcription Factors

Identification by the Relative R squared Method

研究生：王郁涵

Student：Yu-Han Wang

指導教授：王秀瑛教授

Advisor：Dr. Hsiuying Wang

吳謂勝教授

Dr. Wei-Sheng Wu

國立交通大學
統計學研究所
碩士論文



National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Statistics

June 2010

Hsinchu, Taiwan, Republic of China

中華民國九十九年六月

利用相對性 R squared 方法

辨認酵母菌轉錄因子

研究生：王郁涵

指導教授：王秀瑛 教授

吳謂勝 教授

國立交通大學理學院

統計學研究所



轉錄因子在調控基因表現上扮演著重要的角色，為了更加瞭解細胞裡的轉錄機制，辨認出相關的轉錄因子是很重要的。本篇研究結合染色質免疫沉澱片和基因表現片兩種資料，並以一種新的統計方法-- relative R squared 來辨認細胞週期裡的轉錄因子。研究結果辨認出 15 個轉錄因子，其中 12 個為已知和細胞週期有關的轉錄因子，其餘 3 個 (Hap4, Reb1 and Tye7) 則是我們新發現的轉錄因子且分別有四項證據支持這些轉錄因子的正確性。此外，在這 15 個轉錄因子中，我們可以辨認出其中 7 個運作的時間點位於細胞週期的哪個階段，且這些辨認出來的結果大多有相關的文獻來驗證。由於類似的辨認方法很多，故我們以 Jaccard similarity score 來評斷各方法的優劣，並發現我們的方法優於現存的其他方法。最後，我們將此方法應用於另一筆有關細胞週期的基因表現晶片資料來證明我們的方法具有穩健性。

關鍵詞：相對性 R squared 方法；轉錄因子；基因調控

Yeast Cell Cycle Transcription Factors

Identification by the Relative R squared Method

Student : Yu-Han Wang

Advisor : Dr. Hsiuying Wang

Dr. Wei-Sheng Wu

Institute of statistics
National Chiao Tung University
Hsinchu, Taiwan

Abstract

Transcription factors (TFs) play critical roles in controlling gene expressions. To understand how the cell cycle-regulated genes can be transcribed just before they are needed, it is essential to identify their transcriptional regulators. We developed a novel relative R squared method to identify cell cycle TFs in yeast by integrating the ChIP-chip and cell cycle gene expression data. Our method identified 15 cell cycle TFs, 12 of which are known cell cycle TFs, while the remaining three (Hap4, Reb1 and Tye7) are putative novel cell cycle TFs. Four lines of evidence are provided to show the biological significance of our prediction. Besides, for seven of the 15 identified cell cycle TFs, we can further assign a specific cell cycle phase in which the TFs function. Most of our predictions are supported by previous experimental or computational studies. Furthermore, we show that our method performs better than five existing methods for identifying yeast cell cycle TFs. Finally, an application of our method to different cell cycle gene expression datasets suggests that our method is robust.

Keywords: relative R squared method; transcription factor; gene regulation

誌謝

憶起剛進交大的時候，對未知的學業旅程充滿著惶恐與不安。幸好所上的教授都非常親切，在認真指導我們的課業之餘，也會處處關心我們的生活，告訴我們未來在就業上會遇到的困難與該注意的事項，讓人覺得整個統研所就像一個大家庭一樣，而教授們就是我們的爸爸媽媽。

上了碩二以後，很幸運的能跟到所上最親切的教授---王秀瑛老師，每個禮拜一次的咪聽日總是在和氣的氣氛中渡過。由於是第一次最論文研究，很多東西其實都不甚了解，這時候老師都會詳細的解說給我聽，遇到瓶頸時，老師也會設法幫我找到解決的辦法。尤其為了彌補我生物知識上的不足，特別找來了吳謂勝教授一起指導，很多生物學上不會的問題，幸好有吳教授幫忙解決，可以說要是沒有兩位教授，我的論文一定沒有辦法那麼順利的完成。

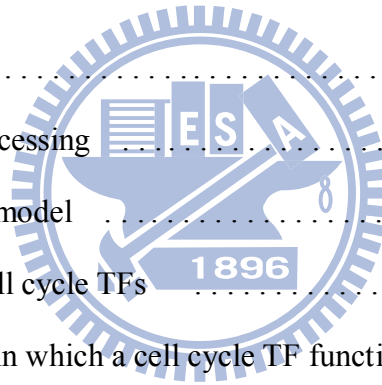
不管是在課業還是論文研究上，陪伴在我身邊的同學們都幫了我不少忙，我們一起討論作業、一起解決程式的錯誤、一起為彼此加油打氣、一起分享人生的酸甜苦辣。謝謝室友玲玲載著我到處趴趴走，謝謝吳剛每次都把我氣得又好氣又好笑，謝謝嚶像個大哥一般耐心傾聽我的煩惱，謝謝老大總在我心情不好時適時給予安慰，謝謝大家總是包容我的記性差跟吵鬧，我想我再也沒辦法遇到比你們更貼心的好朋友了吧。

最後，家一直是支持我的最佳動力，不管遇到什麼挫折，父母永遠都會在背後支持我、鼓勵我，也是因為有他們，才能讓我在完成學業的過程中沒有後顧之憂，謝謝所有在我身邊幫助我的人，因為有你們，我才能在這兩年的研究所生活裡擁有很多幸福的回憶。

王郁涵 謹誌于
國立交通大學統計學研究所
中華民國九十九年六月

Contents

Contents	iv
List of Tables	v
List of Figures	v
1. Introduction	1
2. Methods	3
2.1 The regression model	3
2.2 Identification of cell cycle TFs	6
2.3 Identification of the cell cycle phase in which a cell cycle TF functions	7
3. Data Analysis	8
3.1 Datasets and data preprocessing	8
3.2 Fitting the regression model	9
3.3 Identification of 15 cell cycle TFs	10
3.4 The cell cycle phases in which a cell cycle TF functions	11
4. Discussion	12
4.1 Performance comparison with existing methods	12
4.2 Robustness against different cell cycle gene expression datasets	13
4.3 Threshold setting	14
5. Conclusions	15
Figures	16
Tables	19
Reference	24



List of Figures

1. Flowchart of the procedure of our method.	16
2. Interactions between a novel cell cycle TF and the other identified cell cycle TFs. . . .	17
3. The results of using different cell cycle gene expression datasets.	18

List of Tables

1. The 15 identified cell cycle TFs.	19
2. Known cell cycle genes and proteins that have genetic or physical interactions with the three novel cell cycle TFs (Reb1, Tye7, and Hap4).	20
3. Performance comparison of six cell cycle TF identification methods to retrieve the known cell cycle TFs annotated in the MIPS database.	21
4. Jaccard similarity score with different values of p_0 and s	21
5. List of 203 TFs from Harbison et al.	22
6. Cell cycle TFs which are identified by the six methods.	23

1. Introduction

A transcription factor (TF) is a protein that binds to specific DNA sequences and controls the transfer genetic information from DNA to mRNA. Transcription factors perform this function alone or with other proteins in a complex, by promoting, or blocking the recruitment of RNA polymerase (the enzyme that performs the transcription of genetic information from DNA to RNA) to specific genes. A defining feature of transcription factors is that they contain one or more DNA-binding domains, which attach to specific sequences of DNA adjacent to the genes that they regulate.

Cell cycle transcription factors (TFs) are the genes that regulate the expression of cell cycle-regulated genes. We regard a TF as a cell cycle TF if a statistically significant portion of its regulatory targets are in the set of 800 cell cycle-regulated genes identified by Spellman et al. [23]. Eukaryotic cell cycle is a complex process, which consists of four main phases: DNA replication (S-phase) and mitosis (M-phase), separated by two gap phases (G1 and G2) [3]. Proper regulation of the cell cycle process is crucial to the growth and development of all organisms. Therefore, understanding this regulation is central to the study of many diseases, most notably cancer [28]. Since cell cycle is an important biological process which is regulated at many levels, identifying the cell cycle-regulated genes and their transcriptional regulators are two essential issues for the study of the cell cycle regulation process at the transcriptional level.

The major approach developed to identify TFs is ChIP-chip technique. It is a technique that combines chromatin immunoprecipitation ("ChIP") with microarray technology ("chip"). The main technique is to isolate and identify the DNA sequences occupied by specific DNA binding proteins in cells. These binding sites may indicate functions of various transcriptional

regulators and help identify their target genes during animal development and disease progression. The identified binding sites may also be used as a basis for annotating functional elements in genomes. The types of functional elements that one can identify using ChIP-on-chip include promoters, enhancers, repressor and silencing elements, insulators, boundary elements, and sequences that control DNA replication. So we can identify physical interactions between TFs and promoters. Although many studies utilize ChIP-chip data to accomplish the regulation process of yeast cell cycle and build the network of TF-promoter interactions [22], ChIP-chip data alone cannot tell whether a TF is an activator or a repressor. In order to solve this problem, we combine time course gene expression data. Typically, time course gene expression data are collected by microarray experiments in which gene expression levels of thousands of genes are measured across a number of time points across the cell cycle [6,23,19]. Many computational methods have been developed to identify cell cycle-regulated genes using the time course gene expression data. These methods include Fourier analysis [23], partial least square regression [13], single pulse modeling [36], k-means clustering [24], QT-clustering [12], singular value decomposition [1], correspondence analysis [10], and wavelet analysis [14].

Transcription factors play critical roles in controlling gene expressions. To understand how the cell cycle-regulated genes can be transcribed just before they are needed, it is essential to identify their transcriptional regulators. Several computational methods have been developed to identify yeast cell cycle TFs [2,5,7,26,30-35], including statistical methods (ANOVA analysis [26] and Fisher's G test [5]), linear regression [7], network component analysis [35], rule-based modeling [2], and dynamic system modeling [33]. In this paper, we propose a relative R^2 method to identify cell cycle TFs that regulate the expression of cell cycle-regulated genes. The performance of our method is shown to be better than these previous approaches (see Discussion section).

2. Methods

2.1 The regression model

The relative R^2 method is first proposed by Wang and Li [27] to select the true regulation relations of miRNAs. We extend it to the selection of regulation relations of TFs here. Suppose we have microarray expression data of n TFs, z_1, \dots, z_n , across ℓ time points and B is a matrix which shows all the TFs that bind to the promoter of a gene of interest and is obtained from ChIP-chip data. For each gene in the binding matrix B , we can find the TFs, say z_1, \dots, z_N , such that each of the TFs has this gene as its target. We fit the microarray expression data of the gene in terms of the microarray expression of the N TFs using the regression model that is written as

$$y_t = \beta_0 + \beta_1 z_{1t} + \beta_2 z_{2t} \dots + \beta_N z_{Nt} + \varepsilon_t \quad (1)$$

where y_t represents the target gene's expression profile at time point t , β_0 represents the target gene's basal expression level induced by RNA polymerase II, β_i indicates the regulatory ability of TF i , z_{it} represents the expression profile of TF i at time point t and ε_t denotes the stochastic noise due to the modeling error and the measuring error of the target gene's expression profile. Here ε_t is assumed to be a Gaussian noise with mean zero and unknown standard deviation σ .

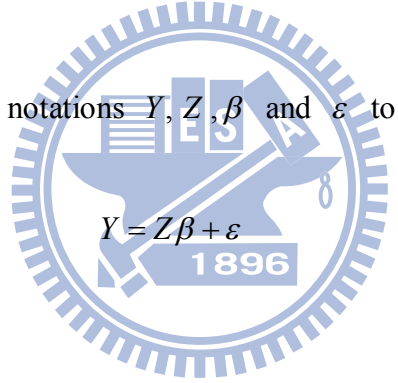
After writing down the linear regression model of gene regulation, the next step is to estimate the unknown parameters in the model. We rewrite Equation (1) into the following regression form:

$$y_t = [1 \quad z_{1t} \quad \cdots \quad z_{Nt}] \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_N \end{bmatrix} + \varepsilon_t \quad (2)$$

Using the yeast cell cycle gene expression data from Pramila *et al.* [19], we can get the values of $\{z_{it}, y_t\}$ for $i \in \{1, 2, \dots, N\}, t \in \{1, 2, \dots, \ell\}$. Equation (2) at different time points can be put together as follows

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_\ell \end{bmatrix} = \begin{bmatrix} 1 & z_{11} & \cdots & z_{N1} \\ 1 & z_{12} & \cdots & z_{N2} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & z_{1\ell} & \cdots & z_{N\ell} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_N \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_\ell \end{bmatrix} \quad (3)$$

For simplicity, we define the notations Y, Z, β and ε to represent Equation (3) as follows



$$Y = Z\beta + \varepsilon \quad (4)$$

where $Y = [y_1 \quad \cdots \quad y_\ell]^T$, Z is the system matrix, $\beta = [\beta_0 \quad \beta_1 \quad \cdots \quad \beta_N]^T$ is the unknown parameter vector, and ε is the error vector. The parameter vector β can be estimated by the best linear unbiased estimator as follows [17]

$$\hat{\beta} = (Z^T Z)^{-1} Z^T Y = [\hat{\beta}_0 \quad \hat{\beta}_1 \quad \cdots \quad \hat{\beta}_N]^T \quad (5)$$

Let $\hat{y}_i = (Z\hat{\beta})_i$. Define $SS_{total} = \sum_{i=1}^N (y_i - \bar{y})^2$ and $SS_{reg} = \sum_{i=1}^N (\hat{y}_i - \bar{y})^2$, where

$\bar{y} = \sum_{i=1}^N y_i / N$ is the mean of y_1, y_2, \dots, y_ℓ . The R^2 is defined as SS_{reg} / SS_{total} , which is

used as an indication of the fitness of the linear regression model. The value of R^2 lies

between 0 and 1 and the larger the value means the model fits better. However, we do not directly use R^2 in this study, but use the relative R^2 values as a criterion to choose high-confidence TFs of a gene. The definition of relative R^2 is given later.

Now we want to select TFs that significantly affect the level of the target gene among the N TFs. We rank the TFs according to their p-values---the smaller the p-value, the higher the rank. The p-value of TF z_i is defined as the probability

$$P(|W| \geq \frac{|\hat{\beta}_i|}{\sqrt{Var(\hat{\beta}_i)}}) \quad (6)$$

which is the p-value used to test $H_0 : \beta_i = 0$, where W denotes the standard normal random variable. The more detailed explanation is mentioned as follows : Note that the expected value of $\hat{\beta}$ equals β , the variance of $\hat{\beta}$ equals $(Z^T Z)^{-1} \sigma^2$ and σ^2 can be estimated by sample variance $\hat{\sigma}^2 = \sum_{i=1}^{\ell} (y_i - \hat{y}_i)^2 / \ell - r$, where r denotes the rank of Z . Thus, $E(\hat{\beta}_i) = \beta_i$ and $Var(\hat{\beta}_i)$ can be approximated by the i th diagonal element of $(Z^T Z)^{-1} \hat{\sigma}^2$. By the Central Limit Theorem, it is easy to show that

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{Var(\hat{\beta}_i)}} \sim N(0,1) \quad (7)$$

when ℓ is large. So the test statistic which is used to test $H_0 : \beta_i = 0$ is $\hat{\beta}_i / \sqrt{Var(\hat{\beta}_i)}$ and $P(|W| \geq |\hat{\beta}_i| / \sqrt{Var(\hat{\beta}_i)})$ is the p-value. Note that if ℓ is small, for obtaining more accurate probability approximation, we may use the T statistic to replace the standard normal random variable W , where the T statistic follows a t distribution.

Rank a TF as the j th significant TF if its p-value is the j th smallest p-value. Consider the TFs that have a p-value less than a critical value, say p_0 . Assume that there are M TFs, z_1, \dots, z_M , whose p-values are less than p_0 . We can use the M TFs to fit the microarray expression data of the target gene. The model is

$$y_t = \beta_0 + \beta_1 z_{1t} + \beta_2 z_{2t} \dots + \beta_M z_{Mt} + \varepsilon_t \quad (8)$$

Denote the R^2 for the regression model (1) as g_N and the R^2 for the regression model (8) as g_M . If $g_M / g_N \geq s$, these M TFs are regarded as the high-confidence TFs of the target gene, where s is a given threshold. The value g_M / g_N is defined as the relative R^2 value in Wang and Li [27]. We use the relative R^2 values to evaluate the fitness of model (8) instead of using the standard R^2 . It is because that even if g_N is not high, it is still possible that the gene is the true target for some TFs among these N TFs. In our study, the smaller of M value means the better of the results because we want to find small proportion of the high-confidence TFs from the potential TFs.

From the above analysis, we can refine the TF-promoter binding matrix $B = [b_{i,j}]$ into a TF-gene regulatory matrix $C = [c_{i,j}]$. In this matrix, $c_{i,j} = 1$ if $b_{i,j} = 1$ and if TF j is shown by the relative R^2 method to exert a significant regulatory effect on the expression of gene i . Otherwise, $c_{i,j} = 0$.

2.2 Identification of cell cycle TFs

From the high-confidence TF-gene regulatory matrix, the regulatory targets of each of the TFs in yeast can be inferred. Then a TF is said to be a cell cycle TF if a statistically significant portion of its regulatory targets are in the set of 800 cell cycle-regulated genes identified by

Spellman *et al.* [23]. The hypergeometric distribution is used to test the statistical significance [17,29]. The procedure for checking whether TF j is a cell cycle TF is as follows. Let F be the set of genes that are bound by TF j (inferred from the TF-promoter binding matrix), G be the set of genes that are regulated by TF j (inferred from the TF-gene regulatory matrix), V be the set of cell cycle-regulated genes that are also bound by TF j , and T be the set of cell cycle-regulated genes that are also regulated by TF j . Then the p -value for rejecting the null hypothesis (H_0 : TF j is not a cell cycle TF) is calculated as

$$p = P(x \geq |T|) = \sum_{x \geq |T|} \frac{\binom{|V|}{x} \binom{|F| - |V|}{|G| - x}}{\binom{|F|}{|G|}} \quad (9)$$

where $|G|$ means the number of genes in set G . TF j is said to be a cell cycle TF if its p -value is less than 0.05. This procedure is applied to each of the TFs under study.

2.3 Identification of the cell cycle phase in which a cell cycle TF functions

For each of the identified cell cycle TFs in section 2.2, we want to determine in which cell cycle phase it functions. We regard that a cell cycle TF functions in the X phase ($X = \text{MG}_1, \text{G}_1, \text{S}, \text{SG}_2, \text{G}_2\text{M}$) if a statistically significant portion of its regulatory targets belong to the X phase cell cycle-regulated genes identified by Spellman *et al.* [23]. Equation (9) is again used to test the statistical significance. While G and F are defined as before, V now denotes the set of X phase cell cycle-regulated genes that are also bound by the cell cycle TF j under study and T now denotes the set of X phase cell cycle-regulated genes that are also regulated by the cell cycle TF under study. We say that a cell cycle TF functions in the X phase ($X = \text{MG}_1, \text{G}_1, \text{S}, \text{SG}_2, \text{G}_2\text{M}$) if its p -value is less than 0.05.

3. Data Analysis

The flowchart of our method is as follows (see Figure 1). Using the ChIP-chip data of Harbison *et al.* [11], we derived a TF-promoter binding matrix. From this binding matrix, we can know all the TFs that bind to the promoter of a gene of interest. These TFs are regarded as the potential transcriptional regulators of the gene of interest. However, binding of a TF to the promoter of a gene does not necessarily imply regulation. A TF may bind to the promoter of a gene but has no regulatory effect on that gene's expression. Hence, additional information is required to solve this ambiguity inherent in the TF-promoter binding matrix. In this study, we use the additional information provided by the yeast cell cycle gene expression data [19] to solve this problem. We use a linear regression model to describe how the target gene's expression during cell cycle is controlled by the TFs that bind to its promoter (inferred from the TF-promoter binding matrix). Among these bound TFs, those that have significant regulatory effects on the target gene's expression can be extracted (see Methods). From this procedure, we can refine the TF-promoter binding matrix into a high-confidence TF-gene regulatory matrix. Each TF-gene regulatory relationship in this matrix is supported by the ChIP-chip and gene expression data. From the high-confidence TF-gene regulatory matrix, the regulatory targets of each of the 203 TFs in yeast can be inferred. Finally, a TF is said to be a cell cycle TF if a statistically significant portion of its regulatory targets are cell cycle-regulated genes.

3.1 Datasets and data preprocessing

We use two data sources in this study. First, the ChIP-chip data are from Harbison *et al.* [11]. They used genome-wide location analysis to determine the genomic occupancy of 203 TFs (See Table 5) in rich media conditions. Second, the yeast cell cycle gene expression data are from Pramila *et al.* [19]. The alpha30 data set is used because it has the largest number of time

points. Samples for all genes in the yeast genome are collected with a sampling interval of 5 minutes and a total of 25 time points, which cover two cell cycles. That is, each gene has a 25 time points gene expression profile.

Using the ChIP-chip data from Harbison *et al.*'s paper [11], we can construct a TF-promoter binding matrix $B = [b_{i,j}]$, where $b_{i,j} = 1$ if the p -value for TF j to bind the promoter of gene i is ≤ 0.001 . Otherwise, $b_{i,j} = 0$. We have a matrix B that includes 4305 binding relationships between the promoters of genes and TFs. However, binding of a TF to the promoter of a gene does not necessarily imply regulation. Hence, additional information is required to solve this ambiguity inherent in the TF-promoter binding matrix. Using our relative R^2 method, we can refine the TF-promoter binding matrix B into a high-confidence TF-gene regulatory matrix C .

3.2 Fit the regression model

We now apply relative R^2 method to the alpha30 data set. The data set includes 4774 genes across 25 time points. The microarray expressions of the 4774 genes across 25 time points can be represented by a 4774×25 matrix, and the 203 TFs across 25 time points can be represented by a 203×25 matrix. To apply the relative R^2 method to an gene in the 4774 genes, we first normalize the expression data of the genes using the 4774 expression data for each time point. The normalization method is first to calculate the mean and standard deviation of the 4774 expression values for each time point. Then, for each time point, the normalized expression data is the original expression data minus the mean and then divided by the standard deviation. The procedure can reduce systematic biases. We fit the regression model with the expression profiles of several TFs (inferred from the TF-promoter binding matrix B) as the inputs and the gene expression profile of the target gene as the output. Using relative R^2 method with

$p_0 = 0.72$ and $s = 0.97$, we obtain a high-confidence TF-gene regulatory matrix C and there are 2494 elements of C whose values are equal to 1.

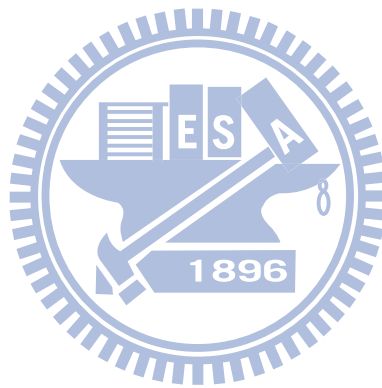
3.3 Identification of 15 cell cycle TFs

From the high-confidence TF-gene regulatory matrix C , the regulatory targets of each of the TFs in yeast can be inferred. Using the p -value in (9), our method identified 15 cell cycle TFs (see Table 1). Among them, 12 are known cell cycle TFs according to the MIPS database [18], including the eight well-known major cell cycle TFs (Ace2, Fkh1, Fkh2, Mbp1, Mcm1, Swi4, Swi5, and Swi6), Abf1, Hir3, Stb1, and Yox1.

The remaining three predicted novel cell cycle TFs (Hap4, Reb1 and Tye7) are supported by four lines of evidence. First, Hap4, Reb1 and Tye7 are shown in literature [25,34] to have physical or genetic interactions with some known cell cycle TFs (see Figure 2), suggesting that these three TFs may play a role in the yeast cell cycle. Second, Hap4, Reb1 and Tye7 are shown in literature [25,34] to regulate some known cell cycle-regulated genes or have protein-protein interactions with some known cell cycle proteins (see Table 2), indicating that our prediction is biologically meaningful. Third, Hap4, Reb1 and Tye7 are also predicted as novel cell cycle TFs by previous computational studies [5,26,35]. Since the same results are predicted by different computational methods, it indicates that our predictions are not happened by chance and may represent novel findings. Fourth, Hap4 and Reb1 were predicted to be cell cycle-regulated by previous studies [8,19]. Being cell cycle regulated themselves, these TFs may play a role in the cell cycle process. Since we provided many lines of evidence to justify our prediction, our results are worthy of further experimental investigation by molecular biologists.

3.4 The cell cycle phases in which a cell cycle TF functions

After identifying the cell cycle TFs, it is desirable to determine in which cell cycle phase a cell cycle TF functions. We regard that a cell cycle TF functions in the X phase ($X = M G_1, G_1, S, S G_2, G_2 M$) if a statistically significant portion of its regulatory targets belong to the X phase cell cycle-regulated genes defined by Spellman *et al.* [23] (see Methods). For seven of the 15 identified cell cycle TFs, we are able to determine the cell cycle phase in which they exert their functions (see Table 1). On average, 86% of our predictions have literature support (57% with experimental evidence and 29% with computational evidence), suggesting that our results may have biological meaning.



4. Discussion

4.1 Performance comparison with existing methods

Five previous studies also tried to identify the yeast cell cycle TFs. Tsai *et al.* [26] identified 30 cell cycle TFs by applying a statistical method (ANOVA analysis) and Cheng *et al.* [5] identified 40 cell cycle TFs by applying another statistical method (Fisher's G test). Cokus *et al.* [7] identified 12 cell cycle TFs by applying linear regression analysis. Andersson *et al.* [2] identified 15 cell cycle TFs by applying rule-based modeling. Wu *et al.* [33] identified 17 cell cycle TFs by using a time-lagged dynamic model of gene regulation (See Table 6). Since these five approaches are different from ours, a performance comparison should be done. As suggested by de Lichtenberg *et al.* [8], we tested the ability of each of these five methods to retrieve the known cell cycle TFs annotated in the MIPS database [18]. Performance comparison was based on the Jaccard similarity score [21], which scores the overlaps between a method's output and the list of known cell cycle TFs (i.e., the true answers). The definition of Jaccard similarity score is given later. Therefore, the higher the Jaccard similarity score, the better the ability of a method to retrieve the known cell cycle TFs. As shown in Table 3, our method has the highest Jaccard similarity score among the six methods. Therefore, our method outperforms the five existing methods.

Before giving definition of Jaccard similarity score, we first describe the origin of Jaccard similarity score. It is evolved from the Jaccard coefficient, which measures similarity between sample set A and sample set B. The Jaccard coefficient is defined as the size of the intersection of the sample sets divided by the size of the union of the sample sets and can be written as $J(A,B)=|A \cap B|/|A \cup B|$. Given two objects, A and B, each with n binary attributes, the Jaccard coefficient is a useful measure of the overlap that A and B share with their

attributes. Each attribute of A and B can either be 0 or 1. The total number of each combination of attributes for both A and B are specified as follows: M_{11} represents the total number of attributes where A and B both have a value of 1. M_{01} represents the total number of attributes where the attribute of A is 0 and the attribute of B is 1. M_{10} represents the total number of attributes where the attribute of A is 1 and the attribute of B is 0. M_{00} represents the total number of attributes where A and B both have a value of 0. Each attribute must fall into one of these four categories, meaning that $M_{11} + M_{01} + M_{10} + M_{00} = \Omega$, The Jaccard similarity coefficient, J, is given as $J = M_{11} / (M_{01} + M_{10} + M_{11})$. In our study, suppose each attribute of A and B represents the number of positives deriving from fact and our method, respectively. Then M_{11} means true positives and was renamed as TP, M_{01} means false positives and was renamed as FP and M_{10} means false negatives and was renamed as FN. Therefore, the Jaccard similarity coefficient, J, is given as

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} = \frac{TP}{TP + FP + FN} \quad (10)$$

and was renamed as the Jaccard similarity score.

4.2 Robustness against different cell cycle gene expression datasets

Besides the above analysis, we also apply the relative R^2 method to another cell cycle gene expression dataset: alpha38 dataset [19]. This dataset has a sampling interval of 5 minutes and a total of 25 data points. In our method, we identified 18 cell cycle TFs. Among them, 13 (Ace2, Cin5, Fkh1, Fkh2, Hir3, Mbp1, Mcm1, Rap1, Swi, Swi5, Swi6, Ume6, Yox1) are known cell cycle TFs according to MIPS database [18] and the remaining five cell cycle TFs are Fhl1, Ino2, Leu3, Met32 and Yap1. In this analysis, the relative R^2 method also leads to high Jaccard similarity score 0.317. Besides, we found that among the 15 cell cycle TFs identified in this study which uses alpha30 dataset [19], 10 TFs are also identified using the

alpha38 dataset (see Figure 3). This suggests that our method is robust against different cell cycle gene expression datasets.

4.3 Threshold setting

There are three thresholds that we need to decide in the above analysis, p_0 , s and α . In the relative R square method, we first use the criterion involving p_0 to select TFs that have significant effect on a gene, then use the criterion involving s to check whether the TFs left are able to account for the dynamics of the target gene's expression (see Methods for details). Since a p-value indicates the significance of regulation of a TF on the gene, it is reasonable to require that p_0 can not be too large. As mentioned in Wang and Li [27] that the selection of p_0 should be more relaxed, while the selection of s can be more strict because the selection of s value is the main criterion. We suggest choosing s more than 0.9 to ensure the accuracy of results. To achieve highest Jaccard similarity score, we conduct simulations for different cases by varying the values of p_0 and s (see Table 4). Finally, p_0 is selected as 0.72 and s is selected as 0.97.

For the hypergeometric significant level α selection, α is commonly selected as 0.05. In this case, we identified 15 cell cycle TFs which included 12 true positive and 3 false positive for $p_0 = 0.72$ and $s = 0.97$ and obtained Jaccard similarity score 0.308. But if we relax the significant level value to 0.15, under the same p_0 and s , we identified 28 cell cycle TFs which included 16 true positive (Abf1, Ace2, Fkh1, Fkh2, Hir3, Mbp1, Mcm1, Ndd1, Rfx1, Stb1, Swi4, Swi5, Swi6, Ume6, Yhp1, Yox1) and 12 false positive (Dal81, Dat1, Fhl1, Gal4, Hap4, Msn4, Pdr1, Phd1, Reb1, Tye7, Yap1, Yap5). This Jaccard similarity score for this case is 0.333.

5. Conclusions

We developed a method to identify cell cycle TFs in yeast by integrating the ChIP-chip [11] and cell cycle gene expression data [19]. We identified 15 cell cycle TFs, 12 of which are known cell cycle TFs. The remaining three TFs (Hap4, Reb1 and Tye7) are putative novel cell cycle TFs. Our predictions are supported by the interaction (physical or genetic) data and previous studies. In addition, for seven of the 15 identified cell cycle TFs, our method can assign a specific cell cycle phase in which the TFs function. On average, 86% of our predictions have literature support (57% with experimental evidence and 29% with computational evidence). Besides, a high-confidence TF-gene regulatory matrix is derived as a byproduct of our method. Each TF-gene regulatory relationship in this matrix is supported by the ChIP-chip and gene expression data. Moreover, we compared the performance of our method with five existing methods and showed that our method has a better ability to retrieve the known cell cycle TFs. Finally, applying our method to different cell cycle gene expression datasets, we identify similar sets of TFs, suggesting that our method is robust.

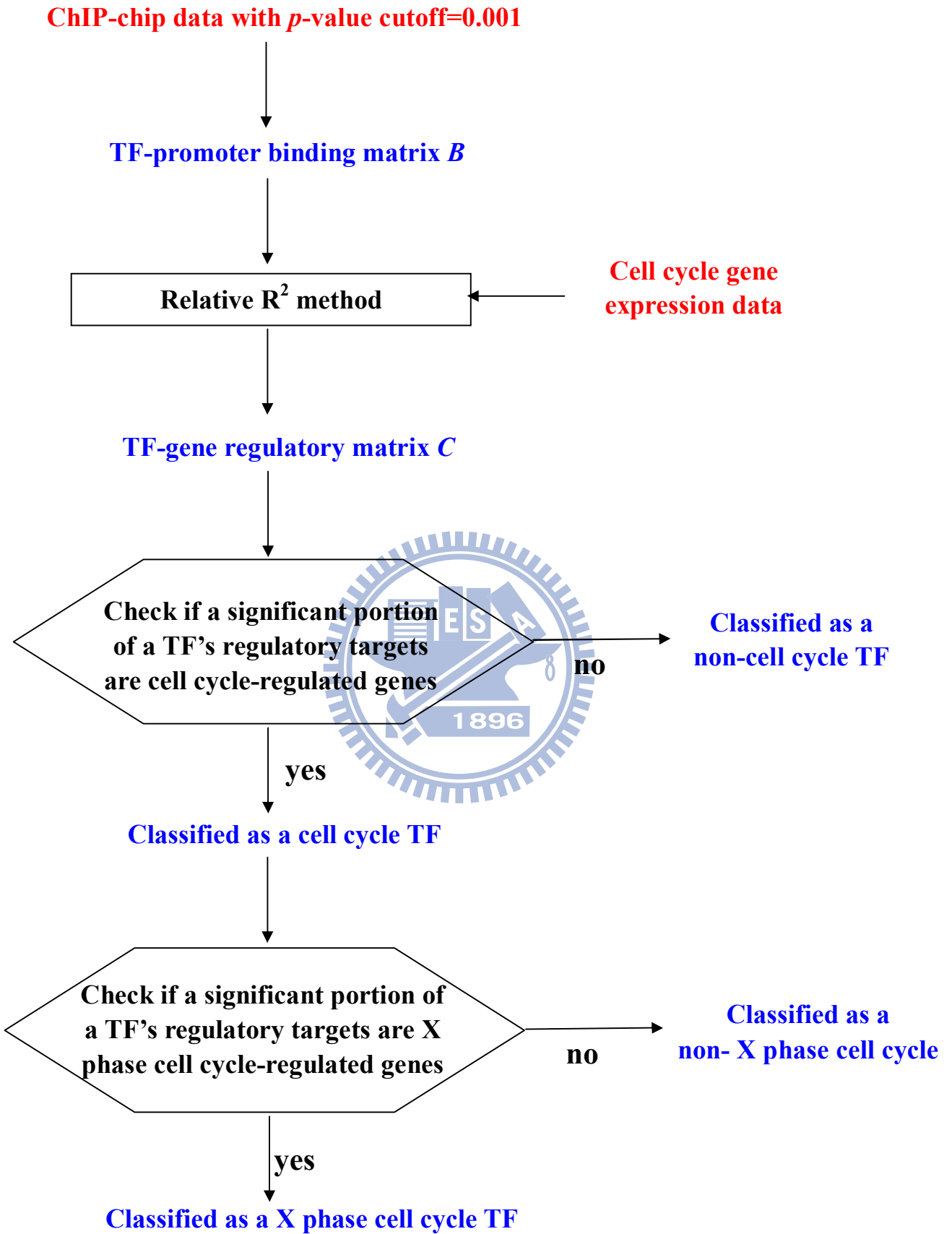


Figure 1: Flowchart of the procedure of our method

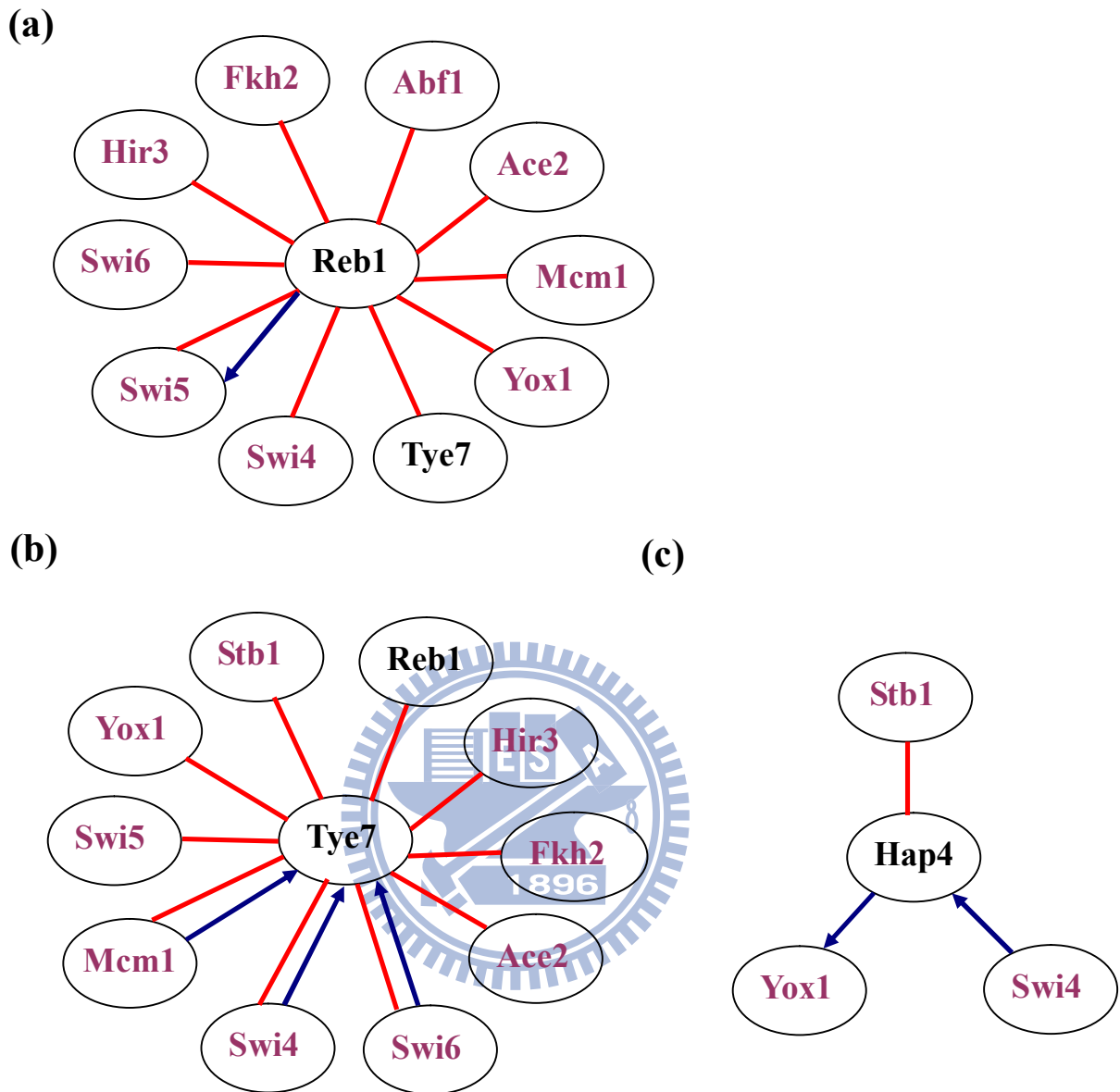


Figure 2: Interactions between a novel cell cycle TF and the other identified cell cycle

TFs

The physical or genetic interactions between a novel cell cycle TF ((a) Reb1, (b) Tye7, and (c) Hap4) and the other identified cell cycle TFs are shown. Each oval indicates an identified cell cycle TF. A TF name is colored purple if it is a known cell cycle TF [18] but black otherwise. Two ovals are connected by an undirected red line if these two TFs have physical interactions indicated by the current protein-protein interaction data [34]. Two ovals are connected by a directed blue line if the two TFs have genetic interactions indicated by ChIP-chip or/and mutant data [25]. For example, Reb1 → Swi5 means that either TF Reb1 binds to the promoter of gene *SWI5* or the disruption of TF Reb1 results in a significant change of the expression of gene *SWI5*.

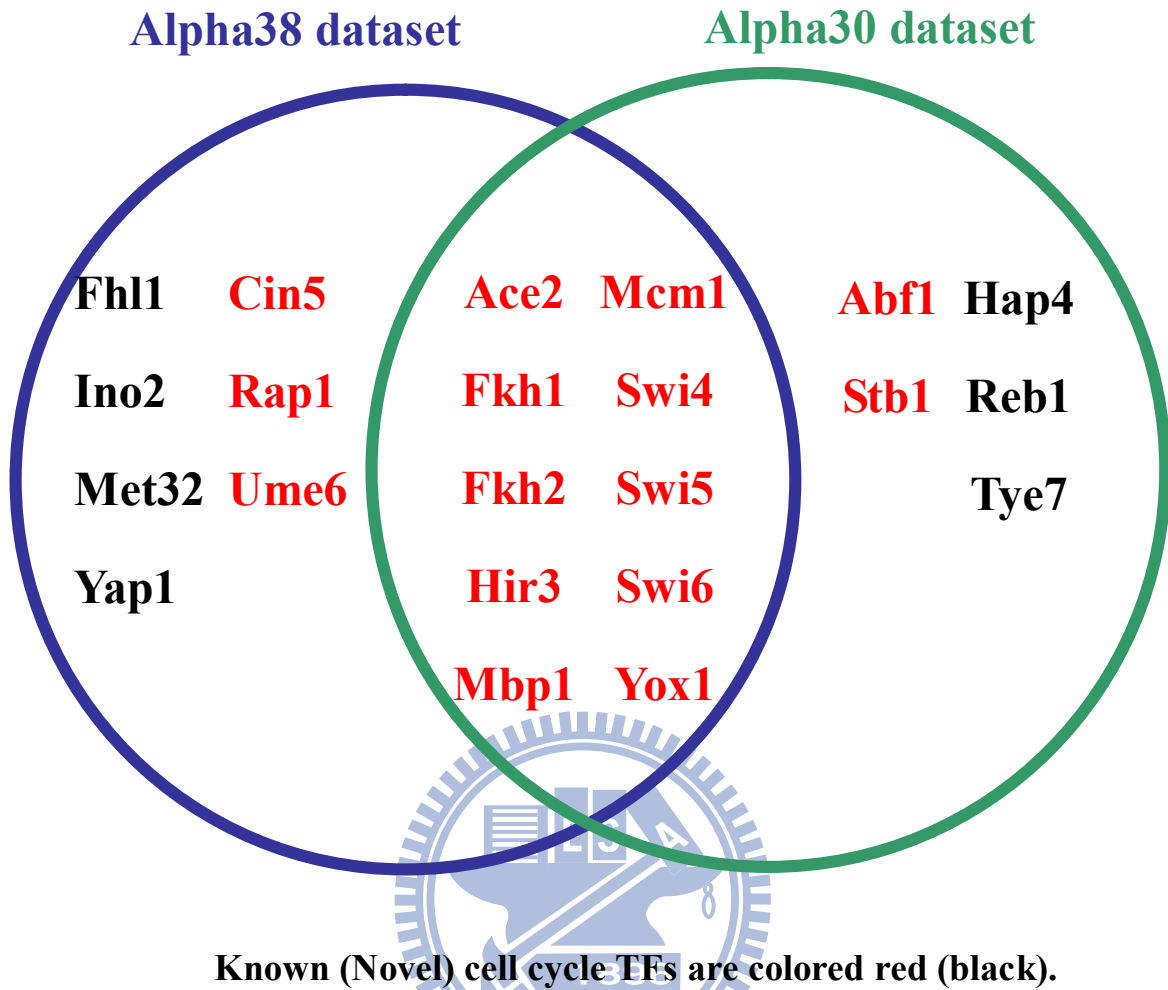


Figure 3: The results of using different cell cycle gene expression datasets

Our method identified 15 and 18 cell cycle TFs using Pramila *et al.*'s alpha30 dataset and alpha38 dataset [19]. Both datasets have a sampling interval of 5 minutes and a total of 25 data points for each gene in the yeast genome. We found that among the 15 cell cycle TFs identified using alpha30 dataset, 10 TFs are also identified using alpha38 dataset. This suggests that our method is robust against different cell cycle gene expression datasets.

Table 1: The 15 identified cell cycle TFs

The twelve known cell cycle TFs (according to the MIPS database [18]) are bold-faced and colored blue. The 15 identified TFs are ordered by the confidence of being cell cycle TFs (according to the hypergeometric p -value calculated using Equation (9)). For seven of the 15 identified cell cycle TFs, the cell cycle phase in which the TFs function are shown. “E” means that the prediction is supported by experimental evidence, “C” means that the prediction is supported by previous computational studies, and “N” stands for our novel prediction.

TF name	Hypergeometric p -value	MG ₁	G ₁	S	SG ₂	G ₂ M
Mbp1	<10 ⁻¹¹	C [30]				
Reb1	<10 ⁻¹¹			N		
Swi4	0.003					
Fkh1	0.004					C[26,30] E[4,22]
Fkh2	0.004				C[26,30] E[16,22]	
Swi5	0.006		C[26,30] E[4,9,22]			
Swi6	0.008	C[26,30,5]				
Ace2	0.012		C[26,30] E[4,15,22]			
Abf1	0.014					
Hir3	0.016					
Stb1	0.021					
Yox1	0.039					
Mcm1	0.041					
Tye7	0.043					
Hap4	0.047					

Table 2: Known cell cycle genes and proteins that have genetic or physical interactions with the three novel cell cycle TFs (Reb1, Tye7, and Hap4)

Known cell cycle genes which are regulated by Reb1 [25]	<i>CDC5, CDC9, CDC21, CDC39, CDC50, CLB2, CLB3, SWI5</i>
Known cell cycle proteins which have protein-protein interaction with Reb1 [34]	Abf1, Ace2, Cdc28, Fkh2, Hcm1, Hir1, Hir2, Hir3, Mcm1, Mec1, Paf1, Swi4, Swi5, Swi6, Tos4, Tos8, Yox1
Known cell cycle genes which are regulated by Tye7 [25]	<i>CDC19, HIR2</i>
Known cell cycle proteins which have protein-protein interaction with Tye7 [34]	Ace2, Cdc28, Cdc37, Clb5, Cln3, Fkh2, Gts1, Hcm1, Hir1, Hir2, Hir3, Mcm1, Met30, Paf1, Reb1, Sis2, Stb1, Swi4, Swi5, Swi6, Tds4, Tds8, Yox1, Yrb1
Known cell cycle genes which are regulated by Hap4 [25]	<i>CDC31, CDC36, CDC50, YOX1</i>
Known cell cycle proteins which have protein-protein interaction with Hap4 [34]	Bub1, Stb1

Table 3: Performance comparison of six cell cycle TF identification methods to retrieve the known cell cycle TFs annotated in the MIPS database

Performance comparison was based on the Jaccard similarity score [21], which scores the overlaps between a method's output and the list of known cell cycle TFs. Specifically, the Jaccard similarity score is defined as $TP/(TP+FP+FN)$, where TP stands for true positives, FP for false positives, and FN for false negatives. Note that the higher the Jaccard similarity score, the better the ability of a method to retrieve the known cell cycle TFs.

	TP	FP	FN	Jaccard similarity score
Our method	12	3	24	0.308
Wu <i>et al.</i> 's method	12	5	24	0.293
Tsai <i>et al.</i> 's method	13	17	23	0.245
Anderson <i>et al.</i> 's method	10	5	26	0.244
Cokus <i>et al.</i> 's method	9	3	27	0.231
Cheng <i>et al.</i> 's method	13	29	23	0.200

Table 4: Jaccard similarity score with different values of p_0 and s

We conduct simulations for different cases by varying the values of p_0 and s and had the highest Jaccard similarity score when p_0 is selected as 0.72 and s is selected as 0.97.

$p_0 \backslash s$	0.995	0.99	0.97	0.95	0.85	0.7
0.9	0.028	0.028	0.027	0.027	0.027	0.027
0.8	0.195	0.184	0.216	0.211	0.105	0.053
0.72	0.154	0.180	0.308	0.275	0.237	0.189
0.7	0.205	0.205	0.293	0.256	0.237	0.189
0.6	0.209	0.209	0.238	0.196	0.225	0.175
0.5	0.163	0.140	0.159	0.196	0.149	0.174
0.4	0.167	0.171	0.179	0.233	0.190	0.182
0.3	0.048	0.048	0.095	0.114	0.143	0.190

Table 5 : List of 203 TF from Harbison et al.

MATA1	ABF1	ABT1	ACA1	ACE2	ADR1	AFT2
ARG80	ARG81	ARO80	ARR1	ASH1	ASK10	AZF1
BAS1	BYE1	CAD1	CBF1	CHA4	CIN5	CRZ1
CST6	CUP9	DAL80	DAL81	DAL82	DAT1	DIG1
DOT6	ECM22	EDS1	FAP7	FHL1	FKH1	FKH2
FZF1	GAL3	GAL4	GAL80	GAT1	GAT3	GCN4
GCR1	GCR2	GLN3	GTS1	GZF3	HAA1	HAC1
HAL9	HAP1	HAP2	HAP3	HAP4	HAP5	HIR1
HIR2	HIR3	HMS1	HMS2	HOG1	HSF1	IFH1
IME1	IME4	INO2	INO4	IXR1	KRE33	KSS1
LEU3	MAC1	MAL13	MAL33	MBF1	MBP1	MCM1
MDS3	MET18	MET28	MET31	MET32	MET4	MGA1
MIG1	MIG2	MIG3	MOT3	MSN1	MSN2	MSN4
MSS11	MTH1	NDD1	NDT80	NNF2	NRG1	OAF1
OPI1	PDC2	PDR1	PDR3	PHD1	PHO2	PHO4
PIP2	PPR1	PUT3	RAP1	RCO1	RCS1	RDR1
RDS1	REB1	RFX1	RGM1	RGT1	RIM101	RLM1
RLR1	RME1	ROX1	RPH1896	RPI1	RPN4	RTG1
RTG3	RTS2	SFL1	SFP1	SIG1	SIP3	SIP4
SKN7	SKO1	SMK1	SMP1	SNF1	SNT2	SOK2
SPT10	SPT2	SPT23	SRD1	STB1	STB2	STB4
STB5	STB6	STE12	STP1	STP2	STP4	SUM1
SUT1	SUT2	SWI4	SWI5	SWI6	TBS1	TEC1
THI2	TOS8	TYE7	UGA3	UME6	UPC2	USV1
WAR1	WTM1	WTM2	XBP1	YAP1	YAP3	YAP5
YAP6	YAP7	TOD6	YBR239c	REI1	YDR026c	YDR049w
YDR266c	URC2	JHD1	YER130c	YER184c	OTU1	YFL052w
YGR067c	YHP1	YJL206c	YKL222c	OAF3	YLR278c	YML081w
YNR063w	YOX1	YPR022c	YPR196w	YRR1	ZAP1	ZMS1

Table 6 : Cell cycle TFs which is identified by the six methods.

The number of true positives, false positives, and false negatives are expressed as (TP, FP, FN). The known cell cycle TFs (according to the MIPS database) are colored red.

Our method (12,3,24)	Wu (12,5,24)	Tsai (13,17,23)	Anderson (10,5,26)	Cokus (9,3,27)	Cheng (13,29,23)	
Abf1	Ace2	Ace2	Ace2	Ace2	Abf1	Swi4
Ace2	Ash1	Bas1	Azf1	Bas1	Ash1	Swi5
Fkh1	Cin5	Dig1	Dig1	Fkh2	Bas1	Swi6
Fkh2	Cst6	Fhl1	Fkh1	Mbp1	Dal80	Tbs1
Hap4	Fkh1	Fkh1	Fkh2	Mcm1	Fkh2	Tye7
Hir3	Fkh2	Fkh2	Mcm1	Ndd1	Fzf1	Uga3
Mbp1	Mbp1	Gal4	Mpb1	Spt2	Gat3	Upc2
Mcm1	Mcm1	Gat3	Ndd1	Ste12	Gcr2	Yap7
Reb1	Ndd1	Haa1	Stb1	Swi4	Hap2	YER184C
Stb1	Rlm1	Hap4	Ste12	Swi5	Hap3	YGR067C
Swi4	Stb1	Hir1	Swi4	Swi6	Hir1	Yox1
Swi5	Ste12	Hir2	Swi6	Yox1	Hir2	YPR196W
Swi6	Stp1	Mbp1	Tec1		Hir3	
Tye7	Swi4	Mcm1	Xbp1		Ihf1	
Yox1	Swi5	Met31	Yox1		Kss1	
	Swi6	Met4			Mbp1	
	Tec1	Met18			Mcm1	
		Mig1			Met32	
		Mig2			Met4	
		Msn2			Msn1	
		Msn4			Ndd1	
		Ndd1			Nrg1	
		Stb1			Otu1	
		Ste12			Pho2	
		Swi4			Reb1	
		Swi5			Rgm1	
		Swi6			Rme1	
		Tec1			Spt2	
		Yap5			Srd1	
		Yox1			Stp4	

References

1. Alter O, Brown PO, Botstein D: **Singular value decomposition for genome-wide expression data processing and modeling.** *Proc Natl Acad Sci USA* 2000, **97**:10101-10106.
2. Andersson CR, Hvidsten TR, Isaksson A, Gustafsson MG, Komorowski J: **Revealing cell cycle control by combining model-based detection of periodic expression with novel cis-regulatory descriptors.** *BMC Syst Biol* 2007, **1**:45.
3. Bähler J: **Cell-cycle control of gene expression in budding and fission yeast.** *Annu Rev Genet* 2005, **39**:69-94.
4. Breeden LL: **Periodic transcription: a cycle within a cycle.** *Curr Biol* 2003, **13**(1):R31-38.
5. Cheng C, Li LM: **Systematic identification of cell cycle regulated transcription factors from microarray time series data.** *BMC Genomics* 2008, **9**:116.
6. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW: **A genome-wide transcriptional analysis of the mitotic cell cycle.** *Mol Cell* 1998, **2**(1):65-73.
7. Cokus S, Rose S, Haynor D, Grønbech-Jensen N, Pellegrini M: **Modelling the network of cell cycle transcription factors in the yeast *Saccharomyces cerevisiae*.** *BMC Bioinformatics* 2006, **7**:381.
8. de Lichtenberg U, Jensen LJ, Fausbøll A, Jensen TS, Bork P, Brunak S: **Comparison of computational methods for the identification of cell cycle-regulated genes.** *Bioinformatics* 2005, **21**(7):1164-1171.
9. Dohrmann PR, Butler G, Tamai K, Dorland S, Greene JR, Thiele DJ, Stillman DJ: **Parallel pathways of gene regulation: homologous regulators SWI5 and ACE2 differentially control transcription of HO and chitinase.** *Genes Dev* 1992,

- 6(1):93-104.
10. Fellenberg K, Hauser NC, Brors B, Neutzner A, Hoheisel JD, Vingron M: **Correspondence analysis applied to microarray data.** *Proc Natl Acad Sci USA* 2001, **98**:10781-10786.
 11. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA: **Transcriptional regulatory code of a eukaryotic genome.** *Nature* 2004, **431**:99-104.
 12. Heyer LJ, Kruglyak S, Yooseph S: **Exploring expression data: identification and analysis of coexpressed genes.** *Genome Res* 1999, **9**:1106-1115.
 13. Johansson D, Lindgren P, Berglund A: **A Multivariate Approach Applied to Microarray data for Identification of Genes with Cell-Cycle Coupled Transcription.** *Bioinformatics* 2003, **19(4)**:467-473.
 14. Klevecz RR: **Dynamic architecture of the yeast cell cycle uncovered by wavelet decomposition of expressionmicroarray data.** *Funct Integr Genomics* 2000, **1**:186-192.
 15. Laabs TL, Markwardt DD, Slattery MG, Newcomb LL, Stillman DJ, Heideman W: **ACE2 is required for daughter cell-specific G1 delay in Saccharomyces cerevisiae.** *Proc Natl Acad Sci USA* 2003, **100(18)**:10275-10280.
 16. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CR, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne J, Volkert TL, Fraenkel E, Gifford DK, Young RA: **Transcriptional Regulatory Networks in Saccharomyces cerevisiae.** *Science* 2002, **298**:799-804.
 17. Mendenhall W, Sincich T: *Statistics for Engineering and the Sciences*, 4th edition. Englewood Cliffs: Prentice-Hall; 1995.

18. Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S, Weil B: **MIPS: a database for genomes and protein sequences.** *Nucleic Acids Res* 2002, **30**:31-34.
19. Pramila T, Wu W, Miles S, Noble WS, Breeden LL: **The Forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle.** *Genes Dev* 2006, **20(16)**:2266-2278.
20. Rowicka M, Kudlicki A, Tu BP, Otwinowski Z: **High-resolution timing of cell cycle-regulated gene expression.** *Proc Natl Acad Sci USA* 2007, **104(43)**:16892-16897.
21. Shakhnovich BE, Reddy TE, Galinsky K, Mellor J, Delisi C: **Comparisons of predicted genetic modules: identification of co-expressed genes through module gene flow.** *Genome Inform Ser Workshop Genome Inform* 2004, **15**:221-228.
22. Simon I, Barnett J, Hannett N, Harbison C, Rinaldi N, Volkert T, Wyrick J, Zeitlinger J, Gifford DK, Jaakkola TS, Young RA: **Serial regulation of transcriptional regulators in the yeast cell cycle.** *Cell* 2001, **106**:697-708.
23. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein 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.
24. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture.** *Nature Genet* 1999, **22**:281-285.
25. Teixeira MC, Monteiro P, Jain P, Tenreiro S, Fernandes AR, Mira NP, Alenquer M, Freitas AT, Oliveira AL, Sá-Correia I: **The YEASTRACT database: a tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae*.** *Nucl Acids Res* 2006, **34**:D446-D451.
26. Tsai HK, Lu HH, Li WH: **Statistical methods for identifying yeast cell cycle transcription factors.** *Proc Natl Acad Sci USA* 2005, **102**:12532-12537.

27. Wang H, Li WH: **Increasing MicroRNA Target Prediction Confidence by the Relative R-squared Method.** *J Theor Biol* 2009, **259**:793-798.
28. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, Matese JC, Perou CM, Hurt MM, Brown PO, David Botstein D: **Identification of gene periodically expressed in the human cell cycle and their expression in tumors.** *Mol Biol Cell* 2002, **13(6)**:1977-2000.
29. Wu WS, Chen BS: **Identifying stress transcription factors using gene expression and TF-gene association data.** *Bioinformatics and Biology Insights* 2007, 1:9-17
30. Wu WS, Li WH, Chen BS: **Computational reconstruction of transcriptional regulatory modules of the yeast cell cycle.** *BMC Bioinformatics* 2006, **7**:421.
31. Wu WS, Li WH, Chen BS: **Identifying regulatory targets of cell cycle transcription factors using gene expression and CHIP-chip data.** *BMC Bioinformatics* 2007, **8**:188.
32. Wu WS, Li WH: **Identifying gene regulatory modules of heat shock response in yeast.** *BMC Genomics* 2008, **9**:439.
33. Wu WS, Li WH: **Systematic identification of yeast cell cycle transcription factors using multiple data sources.** *BMC Bioinformatics* 2008, **9**:522.
34. Wu X, Zhu L, Guo J, Fu C, Zhou H, Dong D, Li Z, Zhang DY, Lin K: **SPIDER: Saccharomyces protein-protein interaction database.** *BMC Bioinformatics* 2006, **7**:S16.
35. Yang YL, Suen J, Brynildsen MP, Galbraith SJ, Liao JC: **Inferring yeast cell cycle regulators and interactions using transcription factor activities.** *BMC Genomics* 2005, **6(1)**:90.
36. Zhao LP, Prentice R, Breeden L: **Statistical modeling of large microarray data sets to identify stimulus response profiles.** *Proc Natl Acad Sci USA* 2001, **98**:5631-5636.