Gene 485 (2011) 172–176

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Methods paper

Yeast cell cycle transcription factors identification by variable selection criteria

Hsiuying Wang [a], Yu-Han Wang [a], Wei-Sheng Wu [b,*]

[a] Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan
[b] Lab of Computational Systems Biology, Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan

ARTICLE INFO

Article history:
Accepted 3 June 2011
Available online 16 June 2011

Received by A.J. van Wijnen

Keywords:
Transcription factors (TFs)
Yeast
Cell cycle
Linear regression

ABSTRACT

Identifying cell cycle transcription factors (TFs) is important for understanding the transcriptional regulation of the cell cycle process which controls the growth and development of all organisms. Existing computational approaches for identifying cell cycle TFs are mainly based on methods with a fixed selection criterion. That is, the same criterion was applied to each TF to determine whether it is a cell cycle TF or not. Since the characteristic of each TF may be quite different, it is not suitable to use a fixed selection criterion in identifying cell cycle TFs. Instead of using a fixed selection criterion, we propose a method with variable selection criteria to identify cell cycle TFs in yeast by integrating the ChIP-chip and cell cycle gene expression data. Our method is shown to outperform five existing methods which used the same ChIP-chip dataset as we did. Fifteen cell cycle TFs were identified by our approach, 12 of which are known cell cycle TFs, while the remaining three (Hap4, Reb1 and Tye7) are novel cell cycle TFs. The biological significance of our predictions is shown by four lines of indirect evidence derived from the protein–protein interaction data, TF mutant data, ChIP-chip data and the results of the previous computational studies.

1. Introduction

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) (Bähler, 2005). 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 (Whitfield et al., 2002). The cell cycle process is precisely regulated at many levels and one important aspect of this regulation is at the transcriptional level. Many genes specific to the cell cycle are transcribed just before they are needed (Rowicka et al., 2007). To have a good understanding of the cell cycle, it is essential to identify the cell cycle-regulated genes and their transcriptional regulators.

DNA microarray technologies have been performed to identify cell cycle-regulated genes. Typically, time course gene expression data are collected by micoarray experiments in which gene expression levels of thousands of genes are measured at a number of time points across the cell cycle (Cho et al., 1998; Spellman et al., 1998; Pramila et al., 2006). Many computational methods have been developed to identify cell cycle-regulated genes using the time course gene expression data. These methods include Fourier analysis (Spellman et al., 1998), partial least square regression (Johansson et al., 2003), single pulse modeling (Zhao et al., 2001), k-means clustering (Tavazoie et al., 1999), QT-clustering (Heyer et al., 1999), singular value decomposition (Alter et al., 2000), and correspondence analysis (Fellenberg et al., 2001).

Transcription factors (TFs) play critical roles in controlling gene expressions (Adachi et al., 2000; Gissot et al., 2004; Kikuchi et al., 2005; Wu et al., 2006a; Wu et al., 2007; Lin et al., 2010; Chang et al., 2011). 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, including statistical methods (ANOVA analysis (Tsai et al., 2005) and Fisher's G test (Cheng and Li, 2008)), network component analysis (Yang et al., 2005), linear regression analysis (Cokus et al., 2006), rule-based modeling (Andersson et al., 2007), and dynamic system modeling (Wu and Li, 2008b). These existing computational approaches for identifying cell cycle TFs are mainly based on methods with a fixed selection criterion. That is, the same criterion is applied to each TF to determine whether it is a cell cycle TF or not. Since the characteristic of each TF may be quite different, it is not suitable to use a fixed selection criterion in identifying cell cycle TFs. Instead, variable selection criteria which depend on the characteristics of the TFs should be developed for identifying cell cycle TFs.

In this paper, we propose a method with variable selection criteria for identifying cell cycle TFs. Our method consists of two steps. The first step is to apply the relative R^2 method (Wang and Li, 2009; Hsieh and Wang, in press) to identify the regulatory targets of each TF in yeast. The second step is to use a hypothesis testing approach to determine whether a TF is a cell cycle TF or not. A TF is regarded as a

Abbreviations: TF, transcription factor; ChIP-chip, chromatin immunoprecipitation-chip; ANOVA, analysis of variance.

* Corresponding author. Tel.: +886 6 2757575x62426; fax: +886 6 2345482.
E-mail address: wessonwu@mail.ncku.edu.tw (W.-S. Wu).

0378-1119/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2011.06.001

cell cycle TF if a statistically significant portion of its regulatory targets is cell cycle-regulated genes.

2. Materials and methods

The data sources used in this study are introduced in Section 2.1. The two steps of the proposed method in identifying cell cycle TFs are illustrated in Sections 2.2 and 2.3.

2.1. Datasets

Two data sources were used in this study. First, the ChIP-chip data were from Harbison et al. (2004). They used genome-wide location analysis to determine the genomic occupancy of 203 TFs in rich media conditions. Second, the yeast cell cycle gene expression data were from Pramila et al. (2006). The alpha30 dataset 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 min and a total of 25 time points, which cover two cell cycles. That is, each gene has a 25-timepoint gene expression profile.

2.2. Identification of the regulatory targets of each TF in yeast by the relative R^2 method

The transcriptional regulatory mechanism of a target gene was modeled as a system with the expression profiles of several TFs as the inputs and the expression profile of the target gene as the output. The transcriptional regulation of the target gene is described by the following linear regression model

$$y_t = d_0 + d_1 z_{1,t} + d_2 z_{2,t} \ldots. + d_N z_{N,t} + \varepsilon_t \tag{1}$$

where y_t represents the target gene's expression profile at time point t, d_0 represents the target gene's basal expression level induced by RNA polymerase II, N denotes the number of TFs that bind to the promoter of the target gene (inferred from the ChIP-chip data), d_i indicates the regulatory ability of TFi, $z_{i,t}$ represents the expression profile of TFi 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 zero mean and unknown standard deviation σ.

Using the yeast cell cycle gene expression data from Pramila et al. (2006), the values of $\{y_t, z_{i,t}\}$ for $t \in \{1, 2, \cdots, 25\}, i \in \{1, 2, \cdots, N\}$ can be obtained. Then (1) at different time points can be rewritten as the following matrix form:

$$Y = Z\beta + e$$

where

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{25} \end{bmatrix}, Z = \begin{bmatrix} 1 & z_{1,1} & \cdots & z_{N,1} \\ 1 & z_{1,2} & \cdots & z_{N,2} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & z_{1,25} & \cdots & z_{N,25} \end{bmatrix}, \beta = \begin{bmatrix} d_0 \\ d_1 \\ \vdots \\ d_N \end{bmatrix}, e = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{25} \end{bmatrix}.$$

The parameter vector β can be estimated by the best linear unbiased estimator as follows (Bickel and Doksum, 2007)

$$\hat{\beta} = \left(Z^T Z\right)^{-1} Z^T Y = \left[\, \hat{d}_0 \quad \hat{d}_1 \quad \cdots \quad \hat{d}_N \,\right]^T.$$

Define $SS_{total} = \sum_{i=1}^{25} (y_i - \bar{y})^2$ and $SS_{reg} = \sum_{i=1}^{25} \left(\hat{y}_i - \bar{y}\right)^2$, where $\hat{y}_i = \left(Z\hat{\beta}\right)_i$ and $\bar{y} = \frac{1}{25}\sum_{i=1}^{25} y_i$. The R^2 value is defined as SS_{reg}/SS_{total}, which is used to measure how well the linear regression model fits the data. The value of R^2 lies between 0 and 1 and larger R^2 value means the model fits better.

Since d_i stands for the regulatory ability of TFi, a large absolute value of d_i means that TFi has a large regulatory effect on the target gene's expression. For each of the N TFs, say TFi, whether its regulatory ability d_i is statistically significantly different from zero is tested. The p-value for rejecting the null hypothesis H_0: $d_i = 0$ is computed as

$$P\left(|W| \geq \frac{|\hat{d}_i|}{\sqrt{Var\left(\hat{d}_i\right)}} \right)$$

where W denotes the standard normal variable (Bickel and Doksum, 2007). Then these N TFs are ranked from the one with the smallest p-value to the one with the largest p-value. The first TF in the ranked TF list is the most plausible regulator of the target gene and the last TF is the most unlikely regulator.

From the ranked list of the N TFs, those TFs that are unlikely to have any regulatory effect on the target gene's expression should be removed. A TF is removed if its p-value is larger than a cutoff p_0. Assuming that only the first k TFs in the ranked list have p-values less than p_0, these k TFs are then checked whether their expression profiles are able to fit the target gene's expression profile well in the linear regression model. The R^2 value of the linear regression model in terms of the N TFs, say g_N, is used as a baseline. Then these k TFs are regarded as the high-confidence TFs of the target gene if $g_k \geq s \cdot g_N$, where g_k is the R^2 value of the linear regression model based on these k TFs and s is a given constant. Note that g_k/g_N is called the relative R^2 value in Wang and Li's paper (2009). The criterion $g_k \geq s \cdot g_N$ is variable since $s \cdot g_N$ is different for each gene.

The same process is applied to each gene of the yeast genome. As a result, the high-confidence TFs of each of the 6000 genes in the yeast genome can be identified. Using the above results, the regulatory targets of each of the 203 TFs in yeast can also be inferred.

2.3. Identification of cell cycle TFs

Since the regulatory targets of each of the 203 TFs in yeast have been inferred, it is now possible to identify cell cycle TFs from these 203 TFs. Because the function of a cell cycle TF is to regulate the expression of the cell cycle-regulated genes, the regulatory targets of a cell cycle TF should be enriched with cell cycle-regulated genes. In this regard, a TF is considered as a cell cycle TF if a statistically significant portion of its regulatory targets is the cell cycle-regulated genes (identified by Spellman et al. (1998)). The hypergeometric distribution is used to test the statistical significance (Wu and Li, 2008a). The procedure for determining whether TFj is a cell cycle TF is as follows. Let F_j be the set of genes that are bound by TFj (inferred from the ChIP-chip data), G_j be the set of genes that are regulated by TFj selected by the relative R^2 method, V_j be the set of cell cycle-regulated genes (identified by Spellman et al. (1998)) that are also bound by TFj, and T_j be the set of cell cycle-regulated genes that are also regulated by TFj. Then the p-value for rejecting the null hypothesis (H_0: TFj is not a cell cycle TF) is calculated as

$$p = P\left(X \geq \left|T_j\right|\right) = \sum_{x \geq |T_j|} \frac{\binom{|V_j|}{x}\binom{|F_j| - |V_j|}{|G_j| - x}}{\binom{|F_j|}{|G_j|}}, \tag{2}$$

where X follows a hypergeometric distribution and $|G_j|$ means the number of genes in set G_j. A TFj 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 203 TFs in yeast.

Note that in (2), all terms (F_j, G_j, V_j, T_j) depend on j. Therefore, they are different for each TF. That is, the criterion for determining whether a TF is a cell cycle TF is a variable criterion.

3. Results

By integrating the ChIP-chip (Harbison et al., 2004) and yeast cell cycle gene expression data (Pramila et al., 2006), our method identified 15 cell cycle TFs. Among them, 12 are known cell cycle TFs listed in the MIPS database (Mewes et al., 2002) with solid experimental evidence, including Abf1, Hir3, Stb1, Yox1, and eight well-known major cell cycle TFs (Ace2, Fkh1, Fkh2, Mbp1, Mcm1, Swi4, Swi5, and Swi6).

The remaining three predicted cell cycle TFs (Hap4, Reb1 and Tye7) have not been reported in the literature with solid experimental evidence and are therefore novel cell cycle TFs. The biological relevance of our predictions is supported by four lines of indirect evidence. First, Hap4, Reb1 and Tye7 have been shown (Teixeira et al., 2006; Wu et al., 2006b) to have physical or genetic interactions with some known cell cycle TFs (see Fig. 1), suggesting that these three TFs may play a role in the yeast cell cycle. Second, Hap4, Reb1 and Tye7 have been shown (Teixeira et al., 2006; Wu et al., 2006b) to regulate some known cell cycle-regulated genes or have protein–protein interactions with some known cell cycle proteins (see Table 1), indicating that our prediction is biologically meaningful. Third, Hap4 and Reb1 were predicted to be cell cycle-regulated by previous computational studies (Pramila et al., 2006). Being cell cycle-regulated themselves, these TFs may play a role in the cell cycle process. Fourth, Hap4, Reb1 and Tye7 are also predicted as novel cell cycle TFs by previous computational studies (Tsai et al., 2005; Cheng and Li, 2008). 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.

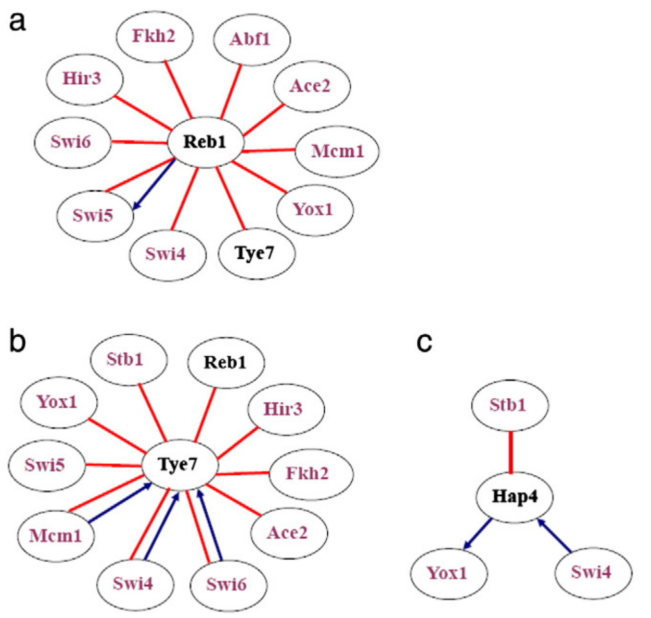

Fig. 1. Physical and genetic interactions between a novel cell cycle TF and the other identified cell cycle TFs. This figure shows the physical or genetic interactions between a novel cell cycle TF ((a) Reb1, (b) Tye7, and (c) Hap4) and the other identified cell cycle TFs. Each identified cell cycle TF is represented as an oval. The names of known cell cycle TFs (according to MIPS database) and newly predicted cell cycle TFs are colored purple and black, respectively. An undirected red line between two ovals indicates these two TFs have physical protein–protein interactions (Wu et al., 2006b). A directed blue line between two ovals indicate that these two TFs have genetic interactions indicated by ChIP-chip or/and mutant data (Teixeira et al., 2006). 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*.

Table 1
Known cell cycle-regulated genes and cell cycle proteins that have genetic or physical interactions with the three novel cell cycle TFs (Reb1, Tye7, and Hap4).

Known cell cycle-regulated genes which are regulated by Reb1 Teixeira et al. (2006)	*CDC5, CDC9, CDC21, CDC39, CDC50, CLB2, CLB3, SWI5*
Known cell cycle proteins which have protein–protein interaction with Reb1 Wu et al. (2006b)	Abf1, Ace2, Cdc28, Fkh2, Hcm1, Hir1, Hir2, Hir3, Mcm1, Mec1, Paf1, Swi4, Swi5, Swi6, Tos4, Tos8, Yox1
Known cell cycle-regulated genes which are regulated by Tye7 Teixeira et al. (2006)	*CDC19, HIR2*
Known cell cycle proteins which have protein–protein interaction with Tye7 Wu et al. (2006b)	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-regulated genes which are regulated by Hap4 Teixeira et al. (2006)	*CDC31, CDC36, CDC50, YOX1*
Known cell cycle proteins which have protein–protein interaction with Hap4 Wu et al. (2006b)	Bub1, Stb1

Note that Hap4 has been predicted by Tsai et al.'s method (2005) as one of the seventeen (1/17) novel cell cycle TFs and Reb1 and Tye7 have been predicted by Cheng and Li's method (2008) as two of the twenty nine (2/29) novel cell cycle TFs (see Supplementary Table 2 for details). Both studies predicted many novel cell cycle TFs but did not provide any evidence to validate Hap4, Reb1 and Tye7 as plausible cell cycle TFs. Therefore, Hap4, Reb1 and Tye7 may not be picked by biologists to do further study since there is no independent evidence to support their biological relevance to the cell cycle process and there are still many other candidates of novel cell cycle TFs that can be chosen for experimental testing. In contrast, our method only predicted three novel cell cycle TFs (Hap4, Reb1 and Tye7) and provided four lines of indirect evidence to validate our predictions. That is, our contribution is to provide high-confidence predictions of three novel cell cycle TFs (Hap4, Reb1 and Tye7), which are worthy of further experimental investigation by biologists.

4. Discussion

4.1. Evaluation of the usefulness of the two steps of the proposed method

The ChIP-chip data in Harbison et al. (2004) can only indicate the binding targets of a TF where the yeast cell is grown in the rich medium. Therefore, it cannot be known whether a TF bind DNA in a cell cycle dependent fashion from their ChIP-chip data. By using yeast cell cycle gene expression data, our method tries to extract the plausible regulatory targets of a TF related to the cell cycle process. Then a TF is regarded as a cell cycle TF if a statistically significant portion of its regulatory targets is cell cycle-regulated genes. In summary, with the aid of the information provided by cell cycle gene expression data, our method tries to identify cell cycle TFs from the 203 TFs which have ChIP-chip data. These 203 TFs contain many false positive cell cycle TFs and our method aims to eliminate them as many as possible.

There are two steps of the proposed method. The first step is to identify the regulatory targets of each TF in yeast by the relative R^2 method and the second step is to identify cell cycle TFs using a hypothesis testing approach. In order to evaluate the usefulness of these two steps, we redid the analysis as follows. First, cell cycle TFs were identified by the raw ChIP-chip data without using any step of the proposed method. Second, cell cycle TFs were identified by using only the first step of the proposed method. In these two situations, a TF is regarded as a cell cycle TF if its regulatory targets contain at least one cell cycle-regulated gene. Third, cell cycle TFs were identified by using both the two steps of the proposed method. In this situation, a TF is considered as a cell cycle TF if a statistically significant portion of

Table 2
Performance comparison of the methods applying none, only the first step, and both the two steps of the proposed method to retrieve the known cell cycle TFs annotated in the MIPS database. The Jaccard similarity score and F-measure value, both of which score the overlap between a method's prediction and the list of known cell cycle TFs, are used for performance comparison. The definition of the Jaccard similarity score is TP/(TP + FP + FN), where TP stands for true positives, FP for false positives, and FN for false negatives. The definition of the F-measure value is 2*precision*recall/(precision + recall) where precision = TP/(TP + FP) and recall = TP/(TP + FN). Note that the high Jaccard similarity score or high F-measure value indicates the high capability of a method in retrieving the known cell cycle TFs.

	TP	FP	FN	Jaccard similarity score	F-measure value
The method using both the two steps	12	3	24	0.308	0.471
The method using only the first step	27	75	9	0.243	0.391
The method without using any step	27	83	9	0.227	0.37

its regulatory targets is the cell cycle-regulated genes. As seen in Table 2, the performance of the method using only the first step is better than that without using any step of the proposed method, showing the usefulness of the first step. Similarly, the performance of the method using both the two steps is better than that using only the first step, showing the usefulness of the second step. The second step can greatly reduce the false positives with the expense of slightly increasing the false negatives. Note that the second step is based on the assumption “A TF is regarded as a cell cycle TF if a statistically significant portion of its regulatory targets is cell cycle-regulated genes”. The above analyses demonstrated that this assumption is useful even though it is not necessary true for all cell cycle TFs.

4.2. Selection of cutoffs

There are two cutoffs p_0 and s that we need to decide in the relative R^2 method. First, p_0 was used to remove those TFs (among all the TFs that bind to the target gene) that are unlikely to have any regulatory effect on the target gene's expression. Then s is used to check whether the left TFs are able to account for the dynamics of the target gene's expression. Since p_0 is used as a coarse filter to remove some TFs that are unlikely to have any regulatory effect on the target gene's expression, the selection of the value of p_0 should not be too stringent. Otherwise, some true TFs of the target gene may be accidentally removed. On the contrary, since s is used as a criterion to determine whether the TFs left can be regarded as the high-confidence TFs of the target gene, the selection of the value of s should be more stringent. Several combinations of p_0 and s were investigated and it can be seen in Table 3 that the combination $(p_0, s) = (0.72, 0.97)$ performs better than the other combinations. Therefore, (0.72, 0.97) are chosen as the default values of (p_0, s) in this study.

4.3. Performance comparison with five existing methods

Five previous methods, which used the same ChIP-chip dataset as we did, have been developed to identify the yeast cell cycle TFs. Tsai et al. (2005) identified 30 cell cycle TFs by applying a statistical method (ANOVA analysis) and Cheng and Li (2008) identified 40 cell cycle TFs by applying another statistical method (Fisher's G test). Cokus et al. (2006) identified 12 cell cycle TFs by applying linear regression analysis. Andersson et al. (2007) identified 15 cell cycle TFs by applying rule-based modeling. Wu and Li (2008b) identified 17 cell cycle TFs by using a time-lagged dynamic model of gene regulation.

Table 3
The Jaccard similarity scores of different combinations of (p_0, s).

$p_0\backslash s$	0.99	0.97	0.95
0.8	0.184	0.216	0.211
0.72	0.180	0.308	0.275
0.7	0.205	0.293	0.256
0.6	0.209	0.238	0.196
0.5	0.140	0.159	0.196
0.4	0.171	0.179	0.233
0.3	0.048	0.095	0.114

Table 4
Performance comparison of six cell cycle TF identification methods to retrieve the known cell cycle TFs annotated in the MIPS database.

	TP	FP	FN	Jaccard similarity score	F-measure value
Our method	12	3	24	0.308	0.471
Wu and Li's method	12	5	24	0.293	0.453
Tsai et al.'s method	13	17	23	0.245	0.394
Anderson et al.'s method	10	5	26	0.244	0.392
Cokus et al.'s method	9	3	27	0.231	0.375
Cheng and Li's method	13	29	23	0.200	0.333

Since these five approaches are different from ours, a performance comparison should be done. As suggested by de Lichtenberg et al. (2005), the ability of each of these six methods to retrieve the known cell cycle TFs according to the MIPS database (Mewes et al., 2002) was used as the performance index. Performance comparison was based on two metrics: the Jaccard similarity score (Shakhnovich et al., 2004) and F-measure value (van Rijsbergen, 1979), both of which score the overlaps between a method's result and the list of known cell cycle TFs (i.e., the true answers). Therefore, the high Jaccard similarity score or high F-measure value indicates high ability of a method to retrieve the known cell cycle TFs. As shown in Table 4, our method has the highest Jaccard similarity score and highest F-measure value among the six methods (see Supplementary Table 1 for more details). Therefore, our method outperforms the other five existing methods.

It should be note that some known cell cycle TFs have not been identified by any of the above six methods. Since the above six cell cycle TFs identification methods all relied on the ChIP-chip data, a cell cycle TF cannot be identified if it has no ChIP-chip data. For example, none of the six methods could successfully identify the known cell cycle TFs Ime1, Nnf2, Wtm2 and YBR267W because no binding targets of these four TFs could be found in the ChIP-chip data (Harbison et al., 2004).

4.4. The novelty of our method

Previous existing computational approaches for identifying the cell cycle TFs are mainly based on methods with a fixed selection criterion. That is, the same criterion is applied to each TF to determine whether it is a cell cycle TF or not. Since the characteristic of each TF may be quite different, it is not suitable to use a fixed selection criterion in identifying cell cycle TFs. To solve this problem, we proposed a method with variable selection criteria which depend on the characteristics of the TFs. Our method consists of two steps. The first step is to apply the relative R^2 method (Wang and Li, 2009) to identify the regulatory targets of each TF in yeast. The second step is to use a hypothesis testing approach to determine whether a TF is a cell cycle TF or not.

Wu and Li's method (2008b) has been shown to be better than other existing computational methods (Tsai et al., 2005; Cokus et al., 2006; Andersson et al., 2007; Cheng and Li, 2008) in identifying cell cycle TFs (see Table 4). Their method also consists of two steps as ours. In the first step, Wu and Li (2008b) used a regression model to identify the transcriptional regulators of each gene in the yeast genome. However, their regression model has not been evaluated by any model fitting criterion (e.g. R^2, AIC, BIC) from the statistical point of view. That is, their method was solely based on a linear regression model without associating with any model fitting criterion. We fixed this weakness by applying a more statistically rigorous method called the relative R^2 method (Wang and Li, 2009). The relative R^2 method is a linear regression approach associated with a model fitting criterion which can provide information about how well the linear regression model fits the expression profile of each gene in the yeast genome.

In the second step of Wu and Li's method, the *p*-value for rejecting the null hypothesis (H_0: TF*j* is not a cell cycle TF) is calculated as

$$p = P\left(X \geq \left|T_j\right|\right) = \sum_{x \geq |T_j|} \frac{\binom{|V|}{x}\binom{|F|-|V|}{\left|G_j\right|-x}}{\binom{|F|}{\left|G_j\right|}}, \qquad (3)$$

where G_j is the set of genes that are regulated by TF*j* selected by the regression method, T_j is the set of cell cycle-regulated genes that are also regulated by TF*j*, *F* is the set of all genes in the yeast genome and *V* be the set of cell cycle-regulated genes (identified by Spellman et al. (1998)). Therefore, *F* and *V* are the same for all TFs. Since the characteristic of each TF may be quite different, it is not suitable to use the same *F* and *V* for all TFs. Our method fixed this weakness by replacing *F* and *V*with F_j and V_j in (2). Both terms depend on *j*. Therefore, they are different for each TF. That is, our criterion for determining whether a TF is a cell cycle TF is a variable criterion. As shown in Table 4, our method outperforms Wu and Li's method, revealing that using variable selection criteria is very useful for improving the performance of identifying cell cycle TFs.

5. Conclusions

We developed a method with variable selection criteria 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 and 12 of which are known cell cycle TFs. The remaining three TFs (Hap4, Reb1 and Tye7) are novel cell cycle TFs. Our predictions are supported by previous computational studies, the protein-protein interaction data, ChIP-chip data or/and TF mutant data. Finally, we showed that our method outperformed five existing methods in identifying cell cycle TFs.

Supplementary materials related to this article can be found online at doi:10.1016/j.gene.2011.06.001.

Acknowledgment

We thank the reviewers for the valuable comments, which are very helpful for improving the quality of this paper. This study was supported by National Center for Theoretical Sciences, Taiwan and National Science Council, Taiwan, NSC99-2628-B-006-015-MY3, NSC-097-2118-M-009-003 and National Center for Theoretical Sciences (NCTS).

References

Adachi, N., Nomoto, M., Kohno, K., Koyama, H., 2000. Cell-cycle regulation of the DNA topoisomerase IIα promoter is mediated by proximal CCAAT boxes: possible involvement of acetylation. Gene 245, 49–57.

Alter, O., Brown, P.O., Botstein, D., 2000. Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. U.S.A. 97, 10101–10106.

Andersson, C.R., Hvidsten, T.R., Isaksson, A., Gustafsson, M.G., Komorowski, J., 2007. Revealing cell cycle control by combining model-based detection of periodic expression with novel cis-regulatory descriptors. BMC Syst. Biol. 1, 45.

Bähler, J., 2005. Cell-cycle control of gene expression in budding and fission yeast. Annu. Rev. Genet. 39, 69–94.

Bickel, P.J., Doksum, K.A., 2007. Mathematical statistics: basic ideas and selected topics, 2nd ed. Pearson Prentice Hall, New Jersey.

Chang, D.T.H., Huang, C.Y., Wu, C.Y., Wu, W.S., 2011. YPA: an integrated repository of promoter features in Saccharomyces cerevisiae. Nucleic Acids Res. 39 (1), D647–D652.

Cheng, C., Li, L.M., 2008. Systematic identification of cell cycle regulated transcription factors from microarray time series data. BMC Genomics 9, 116.

Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., et al., 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol. Cell 2 (1), 65–73.

Cokus, S., Rose, S., Haynor, D., Grønbech-Jensen, N., Pellegrini, M., 2006. Modelling the network of cell cycle transcription factors in the yeast saccharomyces cerevisiae. BMC Bioinformatics 7, 381.

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

Fellenberg, K., Hauser, N.C., Brors, B., Neutzner, A., Hoheisel, J.D., Vingron, M., 2001. Correspondence analysis applied to microarray data. Proc. Natl. Acad. Sci. U.S.A. 98, 10781–10786.

Gissot, M., Refour, P., Briquet, S., Boschet, C., Coupé, S., Mazier, D., et al., 2004. Transcriptome of 3D7 and its gametocyte-less derivative F12 plasmodium falciparum clones during erythrocytic development using a gene-specific micro-array assigned to gene regulation, cell cycle and transcription factors original research article. Gene 341, 267–277.

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., et al., 2004. Transcriptional regulatory code of a eukaryotic denome. Nature 431, 99–104.

Heyer, L.J., Kruglyak, S., Yooseph, S., 1999. Exploring expression data: identification and analysis of coexpressed genes. Genome Res. 9, 1106–1115.

Hsieh, W.J., Wang, H., in press. Human MicroRNA Target Identification by RRSM. J. Theor. Biol.

Johansson, D., Lindgren, P., Berglund, A., 2003. A multivariate approach applied to microarray data for identification of genes with cell-cycle coupled transcription. Bioinformatics 19 (4), 467–473.

Kikuchi, H., Takami, Y., Nakayama, T., 2005. GCN5: a supervisor in all-inclusive control of vertebrate cell cycle progression through transcription regulation of various cell cycle-related genes. Gene 347, 83–97.

Lin, Z., Wu, W.S., Liang, H., Yoo, Y., Li, W.H., 2010. The spatial distribution of cis regulatory elements in yeast promoters and its implications for transcription regulation. BMC Genomics 11, 581.

Mewes, H.W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., et al., 2002. MIPS: a database for genomes and protein sequences. Nucleic Acids Res. 30, 31–34.

Pramila, T., Wu, W., Miles, S., Noble, W.S., Breeden, L.L., 2006. 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. 20 (16), 2266–2278.

Rowicka, M., Kudlicki, A., Tu, B.P., Otwinowski, Z., 2007. High-resolution timing of cell cycle-regulated gene expression. Proc. Natl. Acad. Sci. U.S.A. 104 (43), 16892–16897.

Shakhnovich, B.E., Reddy, T.E., Galinsky, K., Mellor, J., Delisi, C., 2004. Comparisons of predicted genetic modules: identification of co- expressed genes through module gene flow. Genome Inform. 15 (1), 221–228.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., et al., 1998. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell 9, 3273–3297.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., Church, G.M., 1999. Systematic determination of genetic network architecture. Nat. Genet. 22, 281–285.

Teixeira, M.C., Monteiro, P., Jain, P., Tenreiro, S., Fernandes, A.R., Mira, N.P., et al., 2006. The YEASTRACT database: a tool for the analysis of transcription regulatory associations in saccharomyces cerevisiae. Nucleic Acids Res. 34, D446–D451.

Tsai, H.K., Lu, H.H., Li, W.H., 2005. Statistical methods for identifying yeast cell cycle transcription factors. Proc. Natl. Acad. Sci. U.S.A. 102 13532–12537.

van Rijsbergen, C.J., 1979. Information Retrieval, second ed. Butterworth-Heinemann, Massachusetts.

Wang, H., Li, W.H., 2009. Increasing microRNA target prediction confidence by the relative R-squared method. J. Theor. Biol. 259, 793–798.

Whitfield, M.L., Sherlock, G., Saldanha, A.J., Murray, J.I., Ball, C.A., Alexander, K.E., et al., 2002. Identification of gene periodically expressed in the human cell cycle and their expression in tumors. Mol. Biol. Cell 13 (6), 1977–2000.

Wu, W.S., Li, W.H., Chen, B.S., 2006a. Computational reconstruction of transcriptional regulatory modules of the yeast cell cycle. BMC Bioinformatics 7, 421.

Wu, W.S., Li, W.H., Chen, B.S., 2007. Identifying regulatory targets of cell cycle transcription factors using gene expression and ChIP-chip data. BMC Bioinformatics 8, 188.

Wu, W.S., Li, W.H., 2008a. Identifying gene regulatory modules of heat shock response in yeast. BMC Genomics 9, 439.

Wu, W.S., Li, W.H., 2008b. Systematic identification of yeast cell cycle transcription factors using multiple data sources. BMC Bioinformatics 9, 522.

Wu, X., Zhu, L., Guo, J., Fu, C., Zhou, H., Dong, D., et al., 2006b. SPIDer: saccharomyces protein–protein interaction database. BMC Bioinformatics 7, S16.

Yang, Y.L., Suen, J., Brynildsen, M.P., Galbraith, S.J., Liao, J.C., 2005. Inferring yeast cell cycle regulators and interactions using transcription factor activities. BMC Genomics 6 (1), 90.

Zhao, L.P., Prentice, R., Breeden, L., 2001. Statistical modeling of large microarray data sets to identify stimulus response profiles. Proc. Natl. Acad. Sci. U.S.A. 98, 5631–5636.