BMC Bioinformatics

Research article



# Identifying regulatory targets of cell cycle transcription factors using gene expression and ChIP-chip data

Wei-Sheng Wu*[1], Wen-Hsiung Li[2,3] and Bor-Sen Chen[1]

Address: [1]Lab of Control and Systems Biology, Department of Electrical Engineering, National Tsing Hua University, Hsinchu, 300, Taiwan, [2]Department of Evolution and Ecology, University of Chicago, 1101 East 57th Street, Chicago, IL, 60637, USA and [3]Genomics Research Center, Academia Sinica, Taipei, Taiwan

Email: Wei-Sheng Wu* - wessonwu@gmail.com; Wen-Hsiung Li - whli@uchicago.edu; Bor-Sen Chen - bschen@moti.ee.nthu.edu.tw

* Corresponding author

Published: 8 June 2007

*BMC Bioinformatics* 2007, **8**:188 doi:10.1186/1471-2105-8-188

Received: 21 October 2006
Accepted: 8 June 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/188



## Abstract

**Background:** ChIP-chip data, which indicate binding of transcription factors (TFs) to DNA regions *in vivo*, are widely used to reconstruct transcriptional regulatory networks. However, the binding of a TF to a gene does not necessarily imply regulation. Thus, it is important to develop methods to identify regulatory targets of TFs from ChIP-chip data.

**Results:** We developed a method, called Temporal Relationship Identification Algorithm (TRIA), which uses gene expression data to identify a TF's regulatory targets among its binding targets inferred from ChIP-chip data. We applied TRIA to yeast cell cycle microarray data and identified many plausible regulatory targets of cell cycle TFs. We validated our predictions by checking the enrichments for functional annotation and known cell cycle genes. Moreover, we showed that TRIA performs better than two published methods (MA-Network and MFA). It is known that co-regulated genes may not be co-expressed. TRIA has the ability to identify subsets of highly co-expressed genes among the regulatory targets of a TF. Different functional roles are found for different subsets, indicating the diverse functions a TF could have. Finally, for a control, we showed that TRIA also performs well for cell-cycle irrelevant TFs.

**Conclusion:** Finding the regulatory targets of TFs is important for understanding how cells change their transcription program to adapt to environmental stimuli. Our algorithm TRIA is helpful for achieving this purpose.

## Background

By organizing the genes in a genome into transcriptional regulatory modules (TRMs), a living cell can coordinate the activities of many genes and carry out complex functions. Therefore, identifying TRMs is useful for understanding cellular responses to internal and external signals. Advances in high-throughput tools such as DNA microarray [1,2] and chromatin immunoprecipitation-chip (ChIP-chip) [3,4] have made the computational reconstruction of TRMs of a eukaryotic cell possible.

Genome-wide gene expression analysis has been used to investigate TRMs controlling a variety of cellular processes in yeast [5-9]. Clustering and motif-discovering algorithms have been applied to gene expression data to find sets of co-regulated genes and have identified plausible binding motifs of their TFs [7,10,11]. Such approaches

have also been expanded to incorporate existing knowledge about the genes, such as cellular functions [12] or promoter sequence motifs [13]. Moreover, some researchers used model-based approaches such as random Boolean networks [14] and Bayesian networks [15,16] to infer regulatory network architectures. However, this approach provides only indirect evidence of genetic regulatory interactions and does not identify the relevant TFs.

On the other hand, the ChIP-chip technique was developed to identify physical interactions between TFs and DNA regions. Using ChIP-chip data, Simon *et al*. [17] investigated how the yeast cell-cycle gene-expression program is regulated by each of nine major transcriptional activators. Lee *et al*. [18] constructed a network of TF-gene interactions and Harbison *et al*. [19] constructed an initial map of yeast's transcriptional regulatory code. However, a weakness in the ChIP-chip technique is that the binding of a TF to a gene does not necessarily imply regulation. A TF may bind to a gene but has no regulatory effect on that gene's expression. Even if a TF does regulate a specific gene, the ChIP-chip data alone does not tell whether the regulation is activation or repression. Hence, additional information is required to solve this ambiguity inherent in ChIP-chip data.

To overcome this problem, several algorithms have been developed to combine gene expression and ChIP-chip data to infer the regulatory targets of a TF. For instance, NCA [20] and MA-Network [21] both use multivariate regression analysis and MFA [22] uses modified factor analysis of gene expression data to classify a TF's binding targets inferred from ChIP-chip data into regulatory and non-regulatory targets. In this paper, we use a different approach to explore the different biological possibilities for the same phenomenon. We develop a method, called Temporal Relationship Identification Algorithm (TRIA), which uses time-lagged correlation analysis between a TF and its binding targets to identify its regulatory targets. Our rationale is that a TF has a high time-lagged correlation with its regulatory targets, but has a low time-lagged correlation with its binding but non-regulatory targets. Time-lagged correlation analysis has the ability to infer causality and directional relationships between genes [23,24]. It has also been used to reconstruct the reaction network of central carbon metabolism [25] and the gene interaction networks of *Synechocystis sp* [26]. Therefore, time-lagged correlation analysis has the potential to be used to identify a TF's regulatory targets from its binding targets which may or may not be regulated by the TF.

## Results

### *Identification of the plausible regulatory targets of a TF*

Two previous papers [18,19] used a statistical error model to assign a *p*-value to the binding relationship of a TF-gene pair. They found that if $p$-value $\leq 0.001$, the binding relationship of a TF-gene pair is of high confidence and can usually be confirmed by gene-specific PCR. Therefore, we include a gene in the set $B^+$ if the TF-gene binding $p$-value in the ChIP-chip data is $\leq 0.001$, i.e. $B^+$ consists of genes that are significantly bound by a TF. Further, a gene in $B^+$ is assigned into $B^+R^+$ if it has a temporal relationship with the TF but into $B^+R^-$ otherwise. A TF-gene pair is said to have a temporal relationship if the gene's expression profile is significantly correlated with the TF's regulatory profile possibly with time lags (see Methods). Our hypothesis is that the genes in $B^+R^+$ are more likely to be the regulatory targets of a TF than are the genes in $B^+R^-$. TRIA is developed to classify $B^+$ into $B^+R^+$ and $B^+R^-$.

### *Only a subset of the binding targets are plausible regulatory targets of a TF*

We considered nine cell cycle TFs that have both sizes of $B^+R^+$ and $B^+R^- \geq 25$ (i.e. at least 25 genes in each group). The number of genes in each group ($B^+R^+$ and $B^+R^-$) is listed in Table 1. On average, 55% of significantly bound genes are identified as the plausible regulatory targets of a TF, similar to the result (58%) of [21], and 64% of the inferred regulatory targets have expression profiles that are positively correlated with the TF's regulatory profile possibly with time lags. Moreover, only 16% of the inferred regulatory targets and the TF are co-expressed (i.e. identified time lag = 0). That is, 84% of the inferred regulatory targets may not be found if we use the conventional correlation analysis that can only check whether a TF-gene pair are co-expressed or not (see Additional file 1 for details). The following analyses were performed to validate our method.

**Table 1: Classification of the binding targets of a TF into plausible and non-plausible regulatory ones. The numbers of genes in $B^+$, $B^+R^+$ and $B^+R^-$ are shown for each of the nine cell cycle TFs under study. $B^+R^+$ is further divided into two subsets depending on whether the gene's expression profile is positively ($TIC > 0$) or negatively ($TIC < 0$) correlated with the TF's regulatory profile, possibly with time lags (see Additional file 1 for details).**

| TF | $B^+$ | $B^+R^+$ ($TIC > 0$, $TIC < 0$) | $B^+R^-$ |
|---|---|---|---|
| Abf1 | 247 | 144 (85,59) | 103 |
| Ace2 | 81 | 44 (23,21) | 37 |
| Cin5 | 142 | 69 (35,34) | 73 |
| Fkh1 | 133 | 96 (62,34) | 37 |
| Fkh2 | 116 | 90 (60,30) | 26 |
| Rap1 | 147 | 82 (61,21) | 65 |
| Swi4 | 146 | 84 (66,18) | 62 |
| Swi5 | 106 | 42 (32,10) | 64 |
| Swi6 | 144 | 49 (25,24) | 95 |

### *Enrichment for specific functional categories*

$B^+R^+$ is shown to be more enriched than $B^+R^-$ for specific MIPS functional categories with adjusted $p$-value < 0.05 (after the Bonferroni correction for multiple tests) using the cumulative hypergeometric distribution (see Additional file 2 for details). In most cases (7/9), except for Rap1 and Swi5, the number of enriched MIPS functional categories in $B^+R^+$ is larger than that in $B^+R^-$ (see Figure 1). This result suggests that our criterion for distinguishing the plausible from non-plausible regulatory targets of a TF is reliable because co-regulated genes should have a greater probability to be involved in the same functional categories than non-co-regulated genes.

### *Enrichment for cell cycle genes*

We compute the proportions of genes of $B^+R^+$ and $B^+R^-$ that belong to the known cell cycle genes identified by Spellman *et al.* [7]. We then test whether the enrichment of the known cell cycle genes in $B^+R^+$ is statistically higher than that in $B^+R^-$. The cumulative hypergeometric distribution is used to assign a $p$-value for determining the statistical significance (see Appendix for details). In most cases (7/9), except for Abf1 and Ace2, the cell cycle genes are more enriched in $B^+R^+$ than in $B^+R^-$ (see Table 2). This result also suggests that our criterion for distinguishing the plausible from non-plausible regulatory targets of a cell cycle TF is reliable because regulatory targets of a cell cycle TF should be more enriched for the known cell cycle genes than should non-regulatory targets.

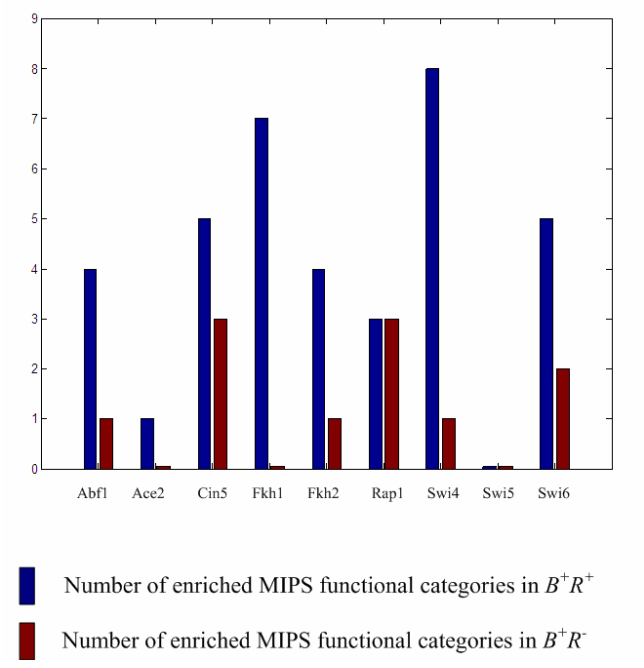


**Figure 1**
**Enrichment in functional annotation for the cell cycle TFs under study**. The numbers of significantly enriched MIPS functional categories in $B^+R^+$ (blue) and $B^+R^-$ (brown) for each of the nine cell cycle TFs under study are shown.

Taken together, the results mentioned above convincingly demonstrate that TRIA is a good method for identifying the plausible regulatory targets of a TF from its binding targets.

### *Identifying highly co-expressed genes among the plausible regulatory targets of a TF*

It is known that co-regulated genes may not be co-expressed [28]. Therefore, it is useful to identify highly co-expressed genes among co-regulated genes because these co-regulated and highly co-expressed genes should be more likely to be simultaneously co-activated or co-repressed by the same TF and involve in the same cellular process.

TRIA has the ability to identify subsets of highly co-expressed genes among the regulatory targets of a TF. First, we use TRIA to identify the plausible regulatory targets ($B^+R^+$) from the binding targets ($B^+$) of a TF. Then, we classify $B^+R^+$ into subsets $A_i$ and $R_i$, where $A_i$ ($R_i$) contains all genes whose expression profiles are positively (negatively) correlated with the TF's regulatory profile with a lag of $i$ time points. Finally, we test whether the expression coherence of $X_i$ is statistically higher than that of $B^+R^+$, where $X_i = A_i$ or $R_i$. The expression coherence of genes in a set $G$ (i.e. $EC(G)$) is defined as the fraction of gene pairs in $G$ with a correlation in expression level higher than a threshold $T$ [27]. The threshold $T$ was determined to be the 95$^{th}$ percentile correlation value of all pairwise correlations between 2000 randomly chosen genes in the yeast genome. Note that $0 \le EC(G) \le 1$. The cumulative hyper-

**Table 2: Enrichment of cell cycle genes. The proportions of genes that belong to the 793 cell cycle genes identified by Spellman *et al.* [7] are calculated for $B^+R^+$ and $B^+R^-$. We then test whether the enrichment of the known cell cycle genes in $B^+R^+$ is statistically higher than that in $B^+R^-$. The cumulative hypergeometric distribution is used to determine the statistical significance (see the Appendix for details). In most cases (7/9), except for Abf1 and Ace2, the known cell cycle genes are more enriched in $B^+R^+$ than in $B^+R^-$.**

| TF | $B^+R^+$ | $B^+R^-$ | $p$-value | $(n_a, m_a, n_b, m_b)$ |
|---|---|---|---|---|
| Abf1 | 19/144 | 6/103 | 0.0439 | (144,19,103,6) |
| Ace2 | 14/44 | 7/37 | 0.1433 | (44,14,37,7) |
| Cin5 | 24/69 | 11/73 | **0.0055** | (69,24,73,11) |
| Fkh1 | 41/96 | 3/37 | **5.9970e-005** | (96,41,37,3) |
| Fkh2 | 54/90 | 0/26 | **3.7043e-009** | (90,54,26,0) |
| Rap1 | 13/82 | 2/65 | **0.0092** | (82,13,65,2) |
| Swi4 | 60/84 | 15/62 | **1.2199e-008** | (84,60,62,15) |
| Swi5 | 22/42 | 14/64 | **0.0012** | (42,22,64,14) |
| Swi6 | 37/49 | 42/95 | **2.7593e-004** | (49,37,95,42) |

geometric distribution is used to assign a *p*-value for rejecting the null hypothesis $EC(X_i) = EC(B^+R^+)$, where $X_i = A_i$ or $R_i$ (see the Appendix for details).

Table 3 lists all subsets of $X_i$'s that contain highly co-expressed genes with *p*-value < 0.001. This result shows that in general several groups of highly co-expressed genes can be extracted from the co-regulated genes, consistent with the result of [28]. These co-regulated and highly co-expressed genes should be more likely to be simultaneously co-activated or co-repressed by the TF and can be used as candidates for further experimental studies. As shown in Table 3, different subsets may have different functional roles, indicating the diverse functions a TF might have.

### *Performance comparison with existing methods*

To identify the regulatory targets of a TF, Gao *et al*. [21] developed MA-Network that uses multivariate regression analysis of gene expression data and Yu *et al*. [22] developed a modified factor analysis (MFA) approach. We compare the identified regulatory targets of the TFs that are available in our study and at least one of the other two studies. On average, only 53% of our identified regulatory targets are also found by MA-Network and only 31% of our identified regulatory targets are also found by MFA. There is little overlap between the above three studies.

**Table 3: Identification of highly co-expressed genes among the regulatory targets of a TF. The expression coherence (*EC*) of $B^+R^+$, $A_i$ and $R_i$ are calculated. We then test whether the expression coherence of $X_i$ is statistically higher than that of $B^+R^-$, where $X_i = A_i$ or $R_i$. The cumulative hypergeometric distribution is used to assign a *p*-value for rejecting the null hypothesis $EC(X_i) = EC(B^+R^+)$. Only those $X_i's$ that have $p < 0.001$ (i.e., $-\log_{10} p > 3$) are shown (see the Appendix for details). In addition, we show the most enriched MIPS functional category for each $X_i$.**

| TF(*EC*($B^+R^+$)) | $X_i$ (*EC*($X_i$); $-\log_{10}$(*p*-value)) | | | | |
|---|---|---|---|---|---|
| Abf1(0.15) | $A_1$(0.64;Inf) PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT | $A_3$(0.51;Inf) TRANSCRIPTION | $R_6$(0.34;Inf) BIOGENESIS OF CELLULAR COMPONENTS | | |
| Ace2(0.07) | $A_3$(0.31;5.11) CELL CYCLE AND DNA PROCESSING | $A_4$(0.5;3.66) REGULATION OF METABOLISM AND PROTEIN FUNCTION | $R_3$(1;3.55) METABOLISM | | |
| Cin5(0.08) | $A_0$(0.73;9.14) ENERGY | $A_1$(0.43;6.29) CELLULAR TRANSPORT, TRANSPORT FACILITIES AND TRANSPORT ROUTES | $A_5$(0.61;11.63) METABOLISM | $R_0$(0.76;Inf) CELLULAR TRANSPORT, TRANSPORT FACILITIES AND TRANSPORT ROUTES | $R_2$(0.47;4.14) CELLULAR TRANSPORT, TRANSPORT FACILITIES AND TRANSPORT ROUTES |
| Fkh1(0.12) | $A_0$(0.65;11.29) CELL CYCLE AND DNA PROCESSING | $A_1$(0.49;11.03) CELL TYPE DIFFERENTIATION | $A_2$(0.27;4.18) CELL CYCLE AND DNA PROCESSING | | |
| Fkh2(0.16) | $A_0$(0.69;Inf) CELL CYCLE AND DNA PROCESSING | $A_1$(0.7;Inf) CELL CYCLE AND DNA PROCESSING | $A_2$(0.69;11.44) CELL TYPE DIFFERENTIATION | $A_3$(0.76;8.82) PROTEIN FATE | |
| Rap1(0.11) | $A_2$(0.58;Inf) PROTEIN SYNTHESIS | $A_4$(0.62;Inf) PROTEIN SYNTHESIS | $A_5$(1;9.46) UNCLASSIFIED PROTEINS | | |
| Swi4(0.2) | $A_0$(0.87;Inf) CELL CYCLE AND DNA PROCESSING | $A_1$(0.6;Inf) METABOLISM | $A_2$(0.79;Inf) CELL CYCLE AND DNA PROCESSING | $A_3$(0.71;6.19) CELL CYCLE AND DNA PROCESSING | |
| Swi5(0.17) | $A_0$(1;7.79) BIOGENESIS OF CELLULAR COMPONENTS | $A_2$(0.86;11.36) INTERACTION WITH THE ENVIRONMENT | $A_3$(0.64;7.78) CELL RESCUE, DEFENSE AND VIRULENCE | | |
| Swi6(0.23) | $A_0$(0.9;10.18) BIOGENESIS OF CELLULAR COMPONENTS | $A_6$(0.73;4.33) CELL CYCLE AND DNA PROCESSING | $A_7$(0.75;8.25) METABOLISM | $R_2$(0.61;4.76) CELL CYCLE AND DNA PROCESSING | |

This is not surprising biologically since the three methods study different biological possibilities for the same phenomenon. However, since the results of the three methods are not highly congruent, a performance comparison of these three methods should be done. Since a TF has to bind to its regulatory targets to regulate their expressions, enrichment of the high-confidence TF binding motifs among the identified regulatory targets of a TF can be used as a criterion for performance comparison. The high-confidence TF binding motifs were derived using six motif discovery methods, also including the requirement for conservation across at least three of the four related yeast species [19]. Let $S_1$ ($T_1$) be the set of regulatory targets of a TF that are identified by TRIA but not by MA-Network (MFA) and $S_2$ ($T_2$) be the set of regulatory targets of a TF that are identified by MA-Network (MFA) but not by TRIA. We tested over-representation of the high-confidence TF binding motifs in $S_1$ and $S_2$ ($T_1$ and $T_2$). The cumulative hypergeometric distribution is used to assign a *p*-value to the motif enrichment (see the Appendix for details). We found that in four of the five (4/5) cases the high-confidence TF binding motifs are enriched in $S_1$ with *p*-value < 0.001 but only two of the five (2/5) cases in $S_2$ are enriched (see Table 4). Similarly, we found that in six of the eight (6/8) cases the high-confidence TF binding motifs are enriched in $T_1$ with *p*-value < 0.001 but none of the eight (0/8) cases in $T_2$ is enriched (see Table 5). The results show that TRIA has a better ability to identify the regulatory targets of a TF than do MA-Network and MFA.

**Table 4: Performance comparison of TRIA with MA-Network. We tested over-representation of the high-confidence TF binding motif in $S_1$ and $S_2$, where $S_1$ is the set of regulatory targets of a TF that are identified by TRIA but not by MA-Network and $S_2$ is the set of regulatory targets of a TF that are identified by MA-Network but not by TRIA. The proportions of genes, whose promoter regions contain the high-confidence TF binding motif is calculated for $S_1$ and $S_2$. The cumulative hypergeometric distribution is used to determine the statistical significance of over-representation (see the Appendix for details). In four of the five (4/5) cases the high-confidence TF binding motifs are enriched in $S_1$ with *p*-value < 0.001 but only two of the five (2/5) cases in $S_2$.**

| TF | $S_1$ | *p*-value | $S_2$ | *p*-value |
|---|---|---|---|---|
| Abf1 | 46/62 | **0** | 28/56 | **3.0839e-011** |
| Ace2 | 2/28 | 0.0340 | 2/17 | 0.0132 |
| Fkh2 | 17/47 | **1.5357e-008** | 7/18 | **1.8019e-004** |
| Swi4 | 16/27 | **6.5301e-012** | 6/18 | 0.0021 |
| Swi5 | 9/25 | **2.4141e-004** | 7/30 | 0.0171 |

**Table 5: Performance comparison of TRIA with MFA. We tested over-representation of the high-confidence TF binding motif in $T_1$ and $T_2$, where $T_1$ is the set of regulatory targets of a TF that are identified by TRIA but not by MFA and $T_2$ is the set of regulatory targets of a TF that are identified by MFA but not by TRIA. The proportions of genes, whose promoter regions contain the high-confidence TF binding motif is calculated for $T_1$ and $T_2$. The cumulative hypergeometric distribution is used to determine the statistical significance of the over-representation (see the Appendix for details). In six of the eight (6/8) cases the high-confidence TF binding motifs are enriched in $T_1$ with *p*-value < 0.001 but none of the eight (0/8) cases in $T_2$.**

| TF | $T_1$ | *p*-value | $T_2$ | *p*-value |
|---|---|---|---|---|
| Abf1 | 75/105 | **4.0357e-012** | 10/106 | 0.9042 |
| Ace2 | 1/31 | 0.2782 | 3/35 | 0.0056 |
| Fkh1 | 30/64 | **3.1252e-007** | 5/109 | 1.0000 |
| Fkh2 | 20/49 | **6.6581e-011** | 10/100 | 0.2038 |
| Rap1 | 32/72 | **1.2579e-011** | 7/36 | 0.0052 |
| Swi4 | 28/56 | **5.3634e-012** | 2/36 | 0.7981 |
| Swi5 | 7/26 | 0.0076 | 4/32 | 0.3417 |
| Swi6 | 19/30 | **2.4500e-009** | 13/72 | 0.2932 |

## Discussion

Many researchers used ChIP-chip data to study regulatory networks of the yeast [17-19,29,30]. Most of them (except [29]) regarded that a gene is regulated by a TF if the gene is bound by the TF with a *p*-value ≤ 0.001 in the ChIP-chip data. However, a TF that binds to a gene may have no regulatory effect on that gene. Therefore, additional information is required to solve this uncertainty. TRIA was developed to overcome this problem and was applied to gene expression and ChIP-chip data to identify the plausible regulatory targets of nine cell cycle TFs. The effectiveness of TRIA was validated by statistically testing for the enrichment of functional groups and known cell cycle genes.

Since co-expressed genes are not necessarily co-regulated and vice versa [28], it is important to develop a method that can identify co-regulated genes that are not co-expressed. TRIA has the ability to do this task. Through identifying a TF's binding targets that have temporal relationships with the TF, we can find the TF's regulatory targets that may not be co-expressed. We can further identify subsets of highly co-expressed genes among the inferred regulatory targets according to the identified time lags and regulatory directions. These co-regulated and highly co-expressed genes should be more likely to be simultaneously co-activated or co-repressed by the TF and can be used as candidates for further experimental studies.

TRIA has been successfully used by two previous studies to investigate other biological problems. First, Tsai *et al*. [31] developed TFBSfinder, which utilizes several data sources (DNA sequences, phylogenetic information, microarray data and ChIP-chip data), to identify cell cycle TF binding sites in yeast. TRIA was used to select reliable target genes of a TF in the first step of their algorithm. The target gene selection is a important step that strongly enhances the performance of TFBSfinder [31]. Since the performance of

TFBSfinder is shown to be better than three well-known TF binding site identification algorithms (AlignACE, MDscan and MEME) [31], this confirmed that TRIA does have ability to identify the plausible regulatory targets of a TF. Second, Wu *et al*. [32] developed MOFA, which integrates gene expression and ChIP-chip data, to reconstruct transcriptional regulatory modules (TRMs) of the yeast cell cycle. TRIA was used as the first step of MOFA to refine the noisy raw ChIP-chip data and construct a binding score matrix. The quality of the binding score matrix strongly affects the performance of MOFA [32]. The TRMs identified by MOFA was validated by using existing experimental data, enrichment for genes in the same MIPS functional category, known DNA-binding motifs, etc. In addition, MOFA is capable of finding many novel TF-target gene relationships and can determine whether a TF is an activator or/and a repressor [32]. Since MOFA can reconstruct biologically relevant TRMs of the yeast cell cycle, this also attests to the usefulness of TRIA.

In this paper, TRIA is used to identify regulatory targets of cell cycle TFs. For a control, we show that TRIA can also perform well for cell-cycle irrelevant regulators. In this regard, we apply TRIA to identify regulatory targets of TFs that are activated by amino acid starved stress. The genome-wide gene expression and ChIP-chip data under amino acid starved growth condition are download from [8,19]. As shown in Figure 2, in most of the cases, $B^+R^+$ is more enriched than $B^+R^-$ for specific MIPS functional categories with adjusted $p$-value < 0.05 (after the Bonferroni correction for multiple tests) using the cumulative hypergeometric distribution. This result suggests that TRIA performed well for cell-cycle irrelevant TFs.

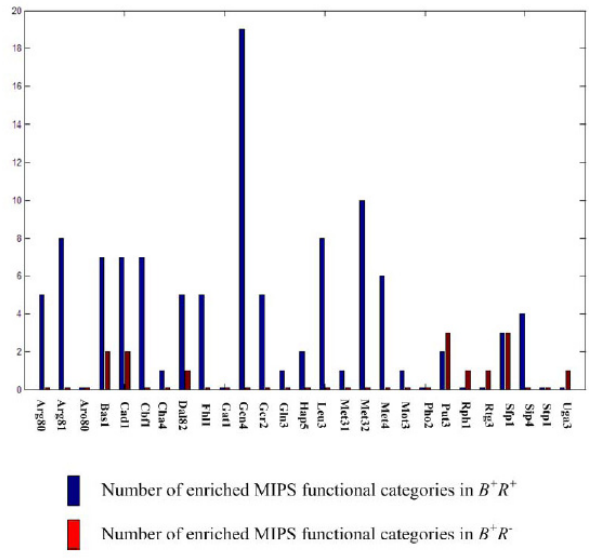


**Figure 2**
**Enrichment in functional annotation for the stress response TFs under study**. The numbers of significantly enriched MIPS functional categories in $B^+R^+$ (blue) and $B^+R^-$ (brown) for each of the 27 amino acid starved stress TFs under study are shown.

The development of TRIA was motivated by two biological observations. First, it is known that TF binding affects gene expression in a nonlinear fashion: below some level it has no effect, and above some level the effect may saturate. This type of behavior can be modeled using a sigmoid function. Therefore, we define a TF's regulatory profile as a sigmoid function of its expression profile as in previous studies [33-35]. Although this may not be true for TFs that are activated at the post-translational stage [20,36], it is not a serious problem for many cell cycle TFs whose expression levels significantly varies with times, indicating that they are under transcriptional control [24,33,34,37-39]. Second, the regulatory effect of a TF on its target genes may not be simultaneous but has a time lag [23,24,26,35,37,38,40-42]. This makes TRIA more general than previous studies [20-22,28] that regard a gene to be regulated by a TF only when the gene's expression profile are co-expressed with the transcription factor activity (TFA) profile. Actually, we found that TRIA performed better than two previous algorithms (MA-Network and MFA) [21,22]. This may result from the fact that TRIA is designed for cell cycle TFs and also considers a time-lagged correlation between a cell cycle TF and its regulatory targets.

In this study, we use time-lagged correlation analysis between a TF and its binding targets to identify its regulatory targets. However, in some cases, TFs may interact with each other and together regulate a group of target genes. This issue will be addressed in the future. We will try to define an overall regulatory profile of a TF complex and apply TRIA to identify target genes that are co-regulated by the same TF complex.

## Conclusion

An algorithm called TRIA is developed to identify the plausible regulatory targets of a TF from its binding targets. Since the binding of a TF to a gene does not necessarily imply regulation, TRIA is used to solve this ambiguity. We validated the effectiveness of TRIA by checking the enrichments for functional annotation and known cell cycle genes. Moreover, the performance of TRIA was shown to be better than two published methods (MA-Network and MFA). Moreover, TRIA has the ability to identify subsets of highly co-expressed genes among the regulatory targets of a TF. In addition, TRIA has been successfully applied to identify high-confidence cell cycle TF binding sites [31] and to reconstruct transcriptional regulatory modules of the yeast cell cycle [32]. Finally, for a control,

TRIA is shown to perform well for cell-cycle irrelevant TFs. In conclusion, TRIA can find biologically relevant results and should be useful for systems biology study.

## Methods

### *Data sets*

Three types of data are used in this study. First, the ChIP-chip data of the cell cycle TFs under the rich media are downloaded from [19]. Second, the gene expression data of the yeast cell cycle are downloaded from [7]. Although it is an old data set, it is still the best cell cycle data set that are available in the public domain. Genes that have only one missing point in their gene expression profiles are reconstructed by the spline algorithm [43], but genes that have more than one missing value in their gene expression profiles or have no ChIP-chip data are excluded. Third, the genome-wide distribution of the high-confidence TF binding motifs was downloaded from [19]. The high-confidence TF binding motifs were derived by using six motif discovery methods, with the requirement for conservation across at least three of four related yeast species [19].

### *Temporal Relationship Identification Algorithm (TRIA)*

Temporal Relationship Identification Algorithm (TRIA) is developed to identify TF-gene pairs that have a temporal relationship. A cell cycle TF and its binding target are said to have a positively (negatively) temporal relationship if the target gene's expression profile is significantly positively (negatively) correlated with the TF's regulatory profile possibly with a time lag. It is known that TF binding affects gene expression in a nonlinear fashion: below some level it has no effect, and above some level the effect may become saturated. This type of behavior can be modeled using a sigmoid function. Therefore, we define a TF's regulatory profile as a sigmoid function of its expression profile as in previous studies [33-35].

Let $\vec{x} = (x_1, ..., x_N)$ be the gene expression time profile of cell cycle TF $x$ and $\vec{y} = (y_1, ..., y_N)$ be the expression profile of gene $y$. The regulatory profile $RP(\vec{x}) = (f(x_1), ..., f(x_N))$ of TF $x$ is defined as a sigmoid function:

$$f(x_i) = \frac{1}{1 + e^{-(x_i - \bar{x})/s}} \quad i = 1, 2, \cdots, N$$

where $\bar{x}$ is the sample mean and $s$ is the sample standard deviation of $\vec{x}$. Compute the correlation between $\vec{y}$ and $RP(\vec{x})$ with a lag of $k$ time points [24,25]:

$$r(k) = \left( \sum_{i=1}^{N-k} (y_{i+k} - \bar{y})(f(x_i) - \bar{m}) \right) \Big/ \left( \sqrt{\sum_{i=1}^{N-k} (y_{i+k} - \bar{y})^2} \cdot \sqrt{\sum_{i=1}^{N-k} (f(x_i) - \bar{m})^2} \right), \quad k = 0, 1, ..., L$$

where $\bar{y} \triangleq \left( \sum_{i=1}^{N-k} y_{i+k} \right) \Big/ (N-k)$, $\bar{m} \triangleq \left( \sum_{i=1}^{N-k} f(x_i) \right) \Big/ (N-k)$ and $L$ is the maximal time lag of the TF's regulatory profile considered. In this study, we set $L = 8$ meaning that we compute the correlation between a gene and a TF with all possible time lags that are less than one cell cycle. The time lag may be interpreted as the time for a TF to have a regulatory effect on a gene.

Then we test the null hypothesis $H_0$: $r(k) = 0$ and the alternative hypothesis $H_1$: $r(k) \neq 0$ by the bootstrap method (see the Appendix) and get a $p$-value $p(k)$. The time-lagged correlation ($TlC$) of $\vec{y}$ and $RP(\vec{x})$ is defined as $r(j)$ that has the smallest $p$-value (i.e. $TlC(\vec{y}, RP(\vec{x})) = r(j)$ if $p(j) \leq p(k)\ \forall k \neq j$). Note that $-1 \leq TlC(\vec{y}, RP(\vec{x})) \leq 1$. Two possible temporal relationships between $\vec{y}$ and $RP(\vec{x})$ can be identified by TRIA: $\vec{y}$ and $RP(\vec{x})$ are (1) positively correlated with a lag of $j$ time points if $TlC(\vec{y}, RP(\vec{x})) = r(j) > 0$ & $p(j) \leq p_{Threshold}$ and (2) negatively correlated with a lag of $j$ time points if $TlC(\vec{y}, RP(\vec{x})) = r(j) < 0$ & $p(j) \leq p_{Threshold}$. The $p_{Threshold}$ is chosen to ensure that we have at most a 5% false discovery rate (FDR) [44]. We may consider that TF $x$, after a lag of $j$ time points, activates (represses) gene $y$ if $\vec{y}$ and $RP(\vec{x})$ are positively (negatively) correlated with a lag of $j$ time points.

## Appendix

### *Statistical test used in Table 2*

We want to test whether the enrichment of the known cell cycle genes (identified in [7]) in $B^+R^+$ is statistically higher than that in $B^+R^-$. Following Banerjee and Zhang [27], a model based on hypergeometric distribution is used.

We calculate:

$$P(m_a, m_b, n_a, n_b) = \frac{\binom{n_a}{m_a}\binom{n_b}{m_b}}{\binom{n_a + n_b}{m_a + m_b}} = \frac{\binom{n_a}{m_a}\binom{N - n_a}{M - m_a}}{\binom{N}{M}}$$

where $N = n_a + n_b$, $M = m_a + m_b$, $n_a(n_b)$ is the number of genes in $B^+R^+$ ($B^+R^-$), $m_a(m_b)$ is the number of the known cell cycle genes in $B^+R^+$ ($B^+R^-$), and

$$\binom{n_a}{m_a} \triangleq \frac{n_a!}{m_a!(n_a - m_a)!}$$

Then, we consider all possible combinations of $x_a$, $x_b$ such that $\sum_{i=\{a,b\}} x_i = \sum_{i=\{a,b\}} m_i = M$ and sum all probabilities calculated as above where $x_a \geq m_a$, which is taken as the *p*-value for rejecting the null hypothesis that enrichment of the known cell cycle genes in $B^+R^+$ is not statistically higher than that in $B^+R^-$.

$$p = P(x_a \geq m_a) = \sum_{x_a \geq m_a} \frac{\binom{n_a}{x_a}\binom{N-n_a}{M-x_a}}{\binom{N}{M}} = 1 - \sum_{x_a=0}^{m_a-1} \frac{\binom{n_a}{x_a}\binom{N-n_a}{M-x_a}}{\binom{N}{M}} \quad \text{(A1)}$$

### *Statistical test used in Table 3*

The expression coherence (*EC*) of sets $B^+R^+$, $A_i$ and $R_i$ are calculated, where $A_i$ ($R_i$) contains all genes in $B^+R^+$ whose expression profiles are positively (negatively) correlated with the TF's regulatory profile with a lag of *i* time points.

We want to test whether the expression coherence of $X_i$ is statistically higher than that of $B^+R^-$, where $X_i = A_i$ or $R_i$. The *p*-value for rejecting the null hypothesis $EC(X_i) = EC(B^+R^+)$ (the alternative hypothesis is $EC(X_i) > EC(B^+R^+)$) is defined as in Equation (*A*1). *N* is the number of gene pairs in $B^+R^+$, *M* is the number of gene pairs in $X_i$, where $X_i = A_i$ or $R_i$, $n_a$ is the number of gene pairs in $B^+R^+$ that have correlations higher than the threshold *T*, and $m_a$ is the number of gene pairs in $X_i$ that have correlations higher than the threshold *T*.

### *Statistical tests used in Tables 4 and 5*

The proportions of genes whose promoter regions contain the high-confidence TF binding motif are calculated for $S_1$ ($T_1$) and $S_2$ ($T_2$), where $S_1$ ($T_1$) is the set of regulatory targets of a TF that are identified by TRIA but not by MA-Network (MFA) and $S_2$($T_2$) is the set of regulatory targets of a TF that are identified by MA-Network (MFA) but not by TRIA. Only 5 TFs (Abf1, Ace2, Fkh2, Swi4, Swi5) are studied for both TRIA and MA-Network. Only 8 TFs (Abf1, Ace2, Fkh1, Fkh2, Rap1, Swi4, Swi5, Swi6) are studied for both TRIA and MFA.

The high-confidence TF binding motifs were derived by using six motif discovery methods, under the requirement for conservation across at least three of the four related yeast species [19]. The yeast genome has 6229 ORFs. Only 817 genes contain Abf1 binding sites, 65 genes contains Ace2 binding sites, 461 genes contain Fkh2 binding sites, 501 genes contain Swi4 binding sites, 575 genes contain Swi5 binding sites, 1181 genes contain Fkh1 binding sites, 379 genes contain Rap1 binding sites, and 946 genes contain Swi6 binding sites [19].

We tested over-representation of the high-confidence TF binding motif in $S_1$ ($T_1$) and $S_2$ ($T_2$). The cumulative hypergeometric distribution is used to determine the statistical significance. The *p*-value is defined as in Equation (*A*1), where $N = 6229$ is the number of genes in the yeast genome, *M* is the number of genes in *G*, where $G = S_1$ ($T_1$) or $S_2$($T_2$) (e.g. $M = 62$ for Abf1 if $G = S_1$ and $M = 56$ for Abf1 if $G = S_2$; $M = 72$ for Rap1 if $G = T_1$ and $M = 36$ for Rap1 if $G = T_2$), $n_a$ is the number of genes in the yeast genome that contain binding sites of a TF under study (e.g. $n_a = 817$ for Abf1; $n_a = 379$ for Rap1) and $m_a$ is the number of genes in *G* that contain binding sites of a TF under study (e.g. $m_a = 46$ for Abf1 if $G = S_1$ and $m_a = 28$ for Abf1 if $G = S_2$; $m_a = 32$ for Rap1 if $G = T_1$ and $m_a = 7$ for Rap1 if $G = T_2$).

### *The bootstrap method for testing the statistical significance of the difference between r(k) and 0*

We observed *N-k* pairs of observations, $Z = \{\mathbf{z}_i: i = 1, ..., N-k \text{ and } \mathbf{z}_i = (f(x_i), y_{i+k})\}$. The correlation coefficient from the sample is calculated and denoted as

$$r(k) = \left(\sum_{i=1}^{N-k} (y_{i+k} - \bar{y})(f(x_i) - \bar{m})\right) \Big/ \left(\sqrt{\sum_{i=1}^{N-k} (y_{i+k} - \bar{y})^2} \cdot \sqrt{\sum_{i=1}^{N-k} (f(x_i) - \bar{m})^2}\right), \quad k = 0,1,2,\cdots$$

where $\bar{y} \triangleq \left(\sum_{i=1}^{N-k} y_{i+k}\right) \Big/ (N-k)$, $\bar{m} \triangleq \left(\sum_{i=1}^{N-k} f(x_i)\right) \Big/ (N-k)$ and $-1 \leq r(k) \leq 1$. It is aimed to use these observations to test if $r(k)$ is different from 0 significantly. Suppose the null hypothesis is $H_0$: $r(k) = 0$ and the alternative hypothesis is $H_1$: $r(k) \neq 0$. We will apply the bootstrap method to perform this hypothesis testing based on the observations. Keeping the pair relationship of these *N-k* pairs to maintain the dependence between $(f(x_i), y_{i+k})$, $\mathbf{z}_i$ are sampled with replacement *N-k* times to form a bootstrap sample, $Z^* = \{\mathbf{z}_i^*: i = 1, ..., N-k \text{ and } \mathbf{z}_i^* \text{ belongs to } Z\}$. The correlation coefficient from the bootstrap sample $Z^*$ is computed and denoted as $r^*(k)$, $-1 \leq r^*(k) \leq 1$. Repeat the resampling procedure B times, we will observed $r_1^*(k)$, $r_2^*(k)$, ..., $r_B^*(k)$. These bootstrap correlation coefficients are sorted to be $-1 \leq r_{(1)}^*(k) \leq r_{(2)}^*(k) \leq ... \leq r_{(B)}^*(k) \leq 1$. Then, the (1-α) two-sided percentile interval is given by

$[\, r^*_{(B\times\alpha/2)}(k),\ r^*_{(B\times(1-\alpha/2))}(k)\,]$ in this case [45]. If this percentile interval does not contain 0, then the null hypothesis is rejected at the significance level of α. Otherwise, the data fail to reject the null hypothesis at the significance level of α. Since the *p*-value is the smallest value of α for which the null hypothesis will be rejected based on the observation, the *p*-value for this test is estimated by the following:

$$\hat{p}(k) = 2\times\min\{\hat{p}_+(k), 1-\hat{p}_+(k)\}, \quad \text{where } \hat{p}_+(k) = \sum_{i=1}^{B} I\{r_i^*(k) \geq 0\}/B,$$

where $\mathbf{I}\{\cdot\}$ is the indicator function whose value is one when the event is true and zero otherwise.

## Authors' contributions

WSW developed the algorithm, performed the simulation and wrote the manuscript. WHL and BSC gave the research topic, provided essential guidance and revised the manuscript. All authors read and approved the final manuscript.

## Additional material

### Additional file 1

*Supplementary Table 1*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-188-S1.xls]

### Additional file 2

*Supplementary Table 2*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-188-S2.pdf]

## Acknowledgements

This study was supported by the National Science Council and Academia Sinica, Taiwan, AS-95-TP-A05.

## References

1. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270:**467-470.
2. DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale.** *Science* 1997, **278:**680-686.
3. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA: **Genome-wide location and function of DNA binding proteins.** *Science* 2000, **290:**2306-2309.
4. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO: **Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF.** *Nature* 2001, **409:**533-538.
5. 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:**65-73.
6. Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I: **The transcriptional program of sporulation in budding yeast.** *Science* 1998, **282:**699-705.
7. 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.
8. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: **Genomic expression programs in the response of yeast cells to environmental changes.** *Mol Biol Cell* 2000, **11:**4241-4257.
9. Causton HC, Ren B, Koh SS, Harbison CT, Kanin E, Jennings EG, Lee T, True HL, Lander ES, Young RA: **Remodeling of yeast genome expression in response to environmental changes.** *Mol Biol Cell* 2001, **12:**323-337.
10. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95:**14863-14868.
11. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture.** *Nat Genet* 1999, **22:**281-285.
12. Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N: **Revealing modular organization in the yeast transcriptional network.** *Nat Genet* 2002, **31:**370-377.
13. Pilpel Y, Sudarsanam P, Church GM: **Identifying regulatory networks by combinatorial analysis of promoter elements.** *Nat Genet* 2001, **29:**153-159.
14. Liang S, Fuhrman S, Somogyi R: **REVEAL, a general reverse engineering algorithm for inference of genetic network architectures.** *Pac Symp Biocomput* 1998, **3:**18-29.
15. Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian networks to analyze expression data.** *J Comput Biol* 2000, **7:**601-620.
16. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.** *Nat Genet* 2003, **34:**166-176.
17. Simon I, Barnett J, Hannett N, Harbison CT, Rinaldi NJ, Volkert TL, Wyrick JJ, Zeitlinger J, Gifford DK, Jaakkola TS, Young RA: **Serial regulation of transcriptional regulators in the yeast cell cycle.** *Cell* 2001, **106:**697-708.
18. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA: **Transcriptional regulatory networks in *Saccharomyces cerevisiae*.** *Science* 2002, **298:**799-804.
19. 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.
20. Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP: **Network component analysis: reconstruction of regulatory signals in biological system.** *Proc Natl Acad Sci USA* 2003, **100:**15522-15527.
21. Gao F, Foat BC, Bussemaker HJ: **Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data.** *BMC Bioinformatics* 2004, **5:**31.
22. Yu T, Li KC: **Inference of transcriptional regulatory network by two-stage constrained space factor analysis.** *Bioinformatics* 2005, **21:**4033-4038.
23. Reis BY, Butte AJ, Kohane IS: **Approaching causality: discovering time-lag correlations in genetic expression data with static and dynamic relevance networks.** *RECOMB* 2000:p5.
24. Kato M, Tsunoda T, Takagi T: **Lag analysis of genetic networks in the cell cycle of budding yeast.** *Genome Inform* 2001, **12:**266-267.
25. Arkin A, Shen PD, Ross J: **A test case of correlation metric construction of a reaction pathway from measurements.** *Science* 1997, **277:**1275-1279.

26. Schmitt WA Jr, Raab RM, Stephanopoulos G: **Elucidation of gene interaction networks through time-lagged correlation analysis of transcriptional data.** *Genome Res* 2004, **14:**1654-1663.
27. Banerjee N, Zhang MQ: **Identifying cooperativity among transcription factors controlling the cell cycle in yeast.** *Nucleic Acids Res* 2003, **31:**7024-7031.
28. Zhou XJ, Kao MC, Huang H, Wong A, Nunez-Iglesias J, Primig M, Aparicio OM, Finch CE, Morgan TE, Wong WHZ: **Functional annotation and network reconstruction through cross-platform integration of microarray data.** *Nat Biotechnol* 2005, **23:**238-243.
29. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK: **Computational discovery of gene modules and regulatory networks.** *Nat Biotechnol* 2003, **21:**1337-1342.
30. Kato M, Hata N, Banerjee N, Futcher B, Zhang MQ: **Identifying combinatorial regulation of transcription factors and binding motifs.** *Genome Biology* 2004, **5:**R56.
31. Tsai HK, Hunag TW, Chou MY, Lu HS, Li WH: **Method for identifying transcription factor binding sites in yeast.** *Bioinformatics* 2006, **22:**1675-1681.
32. Wu WS, Li WH, Chen BS: **Computational reconstruction of transcriptional regulatory modules of the yeast cell cycle.** *BMC Bioinformatics* 2006, **7:**421.
33. Chen HC, Lee HC, Lin TY, Li WH, Chen BS: **Quantitative characterization of the transcriptional regulatory network in the yeast cell cycle.** *Bioinformatics* 2004, **20:**1914-1927.
34. Chen KC, Wang TY, Tseng HH, Huang CY, Kao CY: **A stochastic differential equation model for quantifying transcriptional regulatory network in *Saccharomyces cerevisiae*.** *Bioinformatics* 2005, **21:**2883-2890.
35. Chang WC, Li CW, Chen BS: **Quantitative inference of dynamic pathways via microarray data.** *BMC Bioinformatics* 2005, **6:**44.
36. Bussemaker HJ, Li H, Siggia ED: **Regulatory element detection using correlation with expression.** *Nat Genet* 2001, **27:**167-171.
37. Yu H, Luscombe NM, Qian J, Gerstein M: **Genomic analysis of gene expression relationships in transcriptional regulatory networks.** *Trends Genet* 2003, **19:**422-427.
38. Zhu Z, Pilpel Y, Church GM: **Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm.** *J Mol Biol* 2002, **318:**71-81.
39. Lin LH, Lee HC, Li WH, Chen BS: **Dynamic modeling and gene expression prediction for cis regulatory circuit by cross gene identification scheme.** *BMC Bioinformatics* 2005, **6:**258.
40. Liping J, Tan KL: **Identifying time-lagged gene clusters using gene expression data.** *Bioiformatics* 2005, **21(4):**509-516.
41. Qian J, Dolled-Filhart M, Lin J, Yu H, Gerstein M: **Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions.** *J Mol Biol* 2001, **314:**1053-1066.
42. Qian J, Lin J, Luscombe NM, Yu H, Gerstein M: **Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data.** *Bioinformatics* 2003, **19:**1917-1926.
43. Schumaker L: *Spline functions: basic theory* New York: Wiley; 1981.
44. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society, Series B* 1995, **57:**289-300.
45. Efron B, Tibshirani RJ: *An introduction to the Bootstrap* London: Chapman Hall; 1993.