Identifying Regulatory Targets of Cell Cycle Transcription Factors Using Gene Expression and ChIP-chip Data

Wei-Sheng Wu[1], Wen-Hsiung Li[2] and Bor-Sen Chen[1]
[1]*Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan*
[2]*Department of Evolution and Ecology, University of Chicago, 1101 East 57th Street, Chicago, IL, 60637, USA*
{wswu, bschen}@ee.nthu.edu.tw, whli@uchicago.edu

ABSTRACT

ChIP-chip data, which indicate binding of transcription factors (TFs) to DNA regions in vivo, are widely used to reconstruct transcriptional regulatory modules. 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. 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 expression coherence and 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).

1: INTRODUCTIONS

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. The advance in high-throughput chromatin immunoprecipitation-DNA chip (ChIP-chip) [1,2] has made the computational reconstruction of TRMs of a eukaryotic cell possible.

ChIP-chip technique was used to identify physical interactions between TFs and DNA regions. Using ChIP-chip data, Simon *et al.* [3] investigated how the yeast cell-cycle gene-expression program is regulated by each of nine major transcriptional activators. Lee *et al.* [4] constructed a network of TF-gene interactions and Harbison *et al.* [5] 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 [6,7] and ChIP-chip data to infer regulatory targets of a TF. For instance, NCA [8] and MA-Network [9] both use multivariate regression analysis and MFA [10] 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 [11,12]. It has also been used to reconstruct the reaction network of central carbon metabolism [13] and the gene interaction networks of *Synechocystis sp* [14]. 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.

2: METHODS

2.1: DATA SETS and ADDITIONAL FILES

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 [5]. Second, the gene expression data of the yeast cell cycle are downloaded from [15]. Third, the genome-wide distribution of the high-confidence TF binding motifs was downloaded from [5]. The high-confidence TF binding motifs were derived by using six motif discovery methods, also including the requirement for conservation across at least three of four related yeast species [5]. Moreover, due to the page limit, some details of the paper, which are included in the additional files, has to be found at http://oz.nthu.edu.tw/~d907907/TRIA/ICS2006/TRIA.htm.

2.2: TEMPORAL RELATIONSHIP IDENTIFICATION ALGORITHM (TRIA)

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 positively (negatively) correlated with the TF's regulatory profile, possibly with time lags. 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 just like previous studies [16-18]:

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

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 [12,13]:

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

where

$$\bar{y} \triangleq \left(\sum_{i=1}^{N-k} y_{i+k} \right) \Big/ (N-k), \quad \bar{m} \triangleq \left(\sum_{i=1}^{N-k} f(x_i) \right) \Big/ (N-k)$$

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 Additional file 1) 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 $p(j) \le p(k) \ \forall k \neq j$). Note that $-1 \le TlC(\vec{y}, RP(\vec{x})) \le 1$. 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 $TlC(\vec{y}, RP(\vec{x})) = r(j) > 0$ & $p(j) \le p_{Threshold}$ and (2) negatively correlated with a lag of j time points if $TlC(\vec{y}, RP(\vec{x})) = r(j) < 0$ & $p(j) \le p_{Threshold}$. The $p_{Threshold}$ is chosen to ensure that we have at most a 5% false discovery rate (FDR) [19]. 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.

2.3: IDENTIFICATION of PLAUSIBLE TF REGULATORY TARGETS

Two previous papers [4,5] used a statistical error model to assign a p-value to the binding relationship of a TF-gene pair. They found that if $p \le 0.001$, 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 ≤ 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. Our hypothesis is that the genes in B^+R^+ are more likely to be the TF regulatory targets than are the genes in B^+R^-. TRIA is developed to classify B^+ into B^+R^+ and B^+R^-.

3: RESULTS

3.1: ONLY a SUBSET of the TF BINDING TARGETS are PLAUSIBLE REGULATORY TARGETS

We considered nine cell cycle TFs that have both sizes of B^+R^+ and $B^+R^- \ge 25$ (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 plausible TF regulatory targets, similar to the result of [9], 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 2 for details). The following analyses were performed to validate our method.

3.2: EXPRESSION COHERENCE of PLAUSIBLE and NON-PLAUSIBLE TF REGULATORY TARGETS

We compute the expression coherences of B^+R^+ and B^+R^-. 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 [20]. T was determined to be the 95th percentile correlation value of all pairwise correlations between 2000 randomly chosen genes in the yeast genome. Note that $0 \le EC(G) \le 1$. We then test whether the expression coherence of B^+R^+ is statistically higher than that of B^+R^-. The cumulative hypergeometric distribution is used to assign a p-value for rejecting the null hypothesis $EC(B^+R^+) = EC(B^+R^-)$ (see Additional file 3 for details). Table 2 shows that in most cases (7/9), except for Ace2 and Rap1, the expression coherence of B^+R^+ is significantly higher than that of B^+R^- with $p<0.01$.

This result suggests that our criterion for distinguishing plausible from non-plausible TF regulatory targets is reliable because genes co-regulated by the same TF should be more strongly co-expressed than should non-co-regulated genes.

3.3: 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 4 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^- (Figure 1). This result suggests that our criterion for distinguishing plausible from non-plausible TF regulatory targets is reliable because co-regulated genes should have a greater probability to be involved in the same functional categories than should non-co-regulated genes.

3.4: ENRICHMENT for KNOWN 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.* [15]. 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 Additional file 3 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^- (Table 3). This result also suggests that our criterion for distinguishing 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. Taken together, the results mentioned above convincingly demonstrate that TRIA is a good method for identifying plausible regulatory targets of a TF from its binding targets.

3.6: IDENTIFYING HIGHLY CO-EXPRESSED GENES AMONG THE PLAUSIBLE TF REGULATORY TARGETS

It is known that co-regulated genes may not be co-expressed [21]. 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 similar cellular processes. TRIA has the ability to identify subsets of highly co-expressed genes among a TF's regulatory targets. First, we use TRIA to identify the plausible regulatory targets from the binding targets of a TF. Then, we classify the regulatory targets 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 cumulative hypergeometric distribution is used to assign a p-value for rejecting the null hypothesis $EC(X_i) = EC(B^+R^+)$ (see Additional file 3 for details). Table 4 lists all subsets of X_i's that contain highly co-expressed genes with p<0.01. 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 [21]. That is, co-expression does not imply co-regulation and vice versa.

3.7: PERFORMANCE COMPARISON with EXISTING METHODS

To identify TF regulatory targets, Gao *et al.* [9] developed MA-Network that uses multivariate regression analysis of gene expression data and Yu *et al.* [10] 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 TF regulatory targets are also found by MA-Network and only 31% of our identified TF regulatory targets are also found by MFA. There is little overlap between the above three studies. 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 TF regulatory targets 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 [5]. Let S_1 (or T_1) be the set of regulatory targets of a TF that are identified by TRIA but not by MA-Network (or MFA) and S_2 (or T_2) be the set of regulatory targets of a TF that are identified by MA-Network (or MFA) but not by TRIA. We tested over-representation of the high-confidence TF binding motifs in S_1 and S_2 (or T_1 and T_2). The cumulative hypergeometric distribution is used to assign a p-value to the motif enrichment (see Additional file 3 for details). We found that in four of the five (4/5) cases the high-confidence TF binding motifs is enriched in S_1 with p<0.001 but only two of the five (2/5) cases in S_2 (see Table 5). Similarly, we found that in six of the eight (6/8) cases the high-confidence TF binding motifs is

enriched in T_1 with p<0.001 but zero of the eight (0/8) cases in T_2 (see Table 6). Thus, TRIA has a better ability to identify plausible TF regulatory targets than do MA-Network and MFA.

4: DISCUSSION

The development of TRIA is 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 the regulatory profile of a TF as a sigmoid function of its expression profile just like previous studies [16-18]. Although this may not be true for TFs that are mainly regulated at the post-transcriptional level [8,22], 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 [12,16,17]. Second, the regulatory effect of a TF on its target genes may not be simultaneous but after some time lags [11,12,14,18]. This makes TRIA more general than previous studies [8-10,21] which regard a gene to be regulated by a TF only if 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) [9,10]. This may results from the fact that TRIA is specially designed for cell cycle TFs and also considers time-lagged correlation between a cell cycle TF and its regulatory targets.

Since co-expressed genes are not necessarily co-regulated and vice versa [21], 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 highly co-expressed. We can further identify subsets of highly co-expressed genes among the inferred TF 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.

5: CONSLUSIONS

In this study, an algorithm called TRIA is developed to identify 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 expression coherence and the enrichments for functional annotation and known cell cycle genes. Besides, the performance of TRIA was shown to be better than two published methods (MA-Network and MFA). Moreover, TRIA also has the ability to identify subsets of highly co-expressed genes among a TF's regulatory targets. In addition, in our two previous works, we have successfully applied TRIA to identify high-confidence TF binding sites [23] and to reconstruct transcriptional modules of the yeast cell cycle [24]. Taken together, we are confident that TRIA has the ability to find biologically relevant results and can be useful in systems biology study.

6: REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] Schena M, Shalon D, Davis RW, Brown PO, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science* 1995, 270:467–470.

[7] DeRisi JL, Iyer VR, Brown PO, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science* 1997, 278:680–686.

[8] 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.

[9] Gao F, Foat BC, Bussemaker HJ, "Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data," *BMC Bioinformatics* 2004, 5(1):31.

[10] Yu T, Li KC, "Inference of transcriptional regulatory network by two-stage constrained space factor analysis," *Bioinformatics* 2005, 21:4033–4038.

[11] Reis BY, Butte AJ, Kohane IS, "Approaching causality: discovering time-lag correlations in genetic expression data with static and dynamic relevance networks," *RECOMB2000*, p5.

[12] Kato M, Tsunoda T, Takagi T, "Lag analysis of genetic networks in the cell cycle of budding yeast," *Genome Inform* 2001, 12:266–267.

[13] Arkin A, Shen PD, Ross J, "A test case of correlation metric construction of a reaction pathway from measurements," *Science* 1997, 277:1275–1279.
[14] Schmitt WAJr, Raab RM, Stephanopoulos G, "Elucidation of gene interaction networks through time-lagged correlation analysis of transcriptional data," *Genome Res.* 2004, 14:1654–1663.
[15] 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.
[16] 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.
[17] 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.
[18] Chang WC, Li CW, Chen BS, "Quantitative inference of dynamic pathways via microarray data," *BMC Bioinformatics* 2005, 6(44):1-19.
[19] 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.
[20] Banerjee N, Zhang MQ, "Identifying cooperativity among transcription factors controlling the cell cycle in yeast," *Nucleic Acids Res* 2003, 31:7024–7031.
[21] 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.
[22] Bussemaker HJ, Li H, Siggia ED, "Regulatory element detection using correlation with expression," *Nat. Genet.* 2001, 27:167-171.
[23] Tsai HK, Hunag TW, Chou MY, Lu HS, Li WH, "Method for identifying transcription factor binding sites in yeast," *Bioinformatics* 2006, 22:1675–1681.
[24] Wu WS, Li WH, Chen BS, "Computational reconstruction of transcriptional regulatory modules of the yeast cell cycle," *BMC Bioinformatics*, revised.

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

Table 1 - Classification of TF binding targets 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 2 for details).

TF	$EC(B^+R^+)$	$EC(B^+R^-)$	p-value
Abf1	0.15	0.05	0
Ace2	0.07	0.05	0.1771
Cin5	0.08	0.04	3.7866e-010
Fkh1	0.12	0.04	0
Fkh2	0.16	0.03	1.9035e-012
Rap1	0.11	0.1	0.0742
Swi4	0.2	0.05	0
Swi5	0.17	0.06	2.1148e-012
Swi6	0.23	0.08	0

Table 2 - Expression coherences of B^+R^+ and B^+R^-. The expression coherences of B^+R^+ and B^+R^- are calculated for each of the nine cell cycle TFs under study. We then test whether the expression coherence of B^+R^+ is statistically higher than that of B^+R^-. The cumulative hypergeometric distribution is used to assign a p-value for rejecting the null hypothesis $EC(B^+R^+) = EC(B^+R^-)$ (see Additional file 3 for details).

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

Table 3 - Enrichment of cell cycle genes. The proportions of genes that belong to the 793 cell cycle genes identified by Spellman et al. [15] 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 Additional file 3 for details).

TF($EC(B^+R^+)$)	X_i ($EC(X_i)$; -$\log_{10}$(p-value))		
Abf1(0.15)	A_1(0.64;Inf)	A_2(0.33;2.98)	A_3(0.51;Inf)
	R_2(0.39;2.72)	R_6(0.34;Inf)	
Ace2(0.07)	A_3(0.31;5.11)	A_4(0.5;3.66)	R_3(1;3.55)
Cin5(0.08)	A_0(0.73;9.14)	A_1(0.43;6.29)	A_5(0.61;11.63)
	R_0(0.76;Inf)	R_1(0.4;2.24)	R_2(0.47;4.14)
Fkh1(0.12)	A_0(0.65;11.29)	A_1(0.49;11.03)	A_2(0.27;4.18)
	A_6(0.47;2.95)		
Fkh2(0.16)	A_0(0.69;Inf)	A_1(0.7;Inf)	A_2(0.69;11.44)
	A_3(0.76;8.82)		
Rap1(0.11)	A_2(0.58;Inf)	A_3(0.67;2.68)	A_4(0.62;Inf)
	A_5(1;9.46)		
Swi4(0.2)	A_0(0.87;Inf)	A_1(0.6;Inf)	A_2(0.79;Inf)
	A_3(0.71;6.19)		
Swi5(0.17)	A_0(1;7.79)	A_2(0.86;11.36)	A_3(0.64;7.78)
Swi6(0.23)	A_0(0.9;10.18)	A_6(0.73;4.33)	A_7(0.75;8.25)
	R_2(0.61;4.76)		

Table 4 - Identification of highly co-expressed genes among the TF regulatory targets. The *EC* scores 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 $p < 0.01$ (i.e., $-\log_{10} p > 2$) are shown (see Additional file 3 for details).

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 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 Additional file 3 for details).

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

Table 6 - 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 Additional file 3 for details).

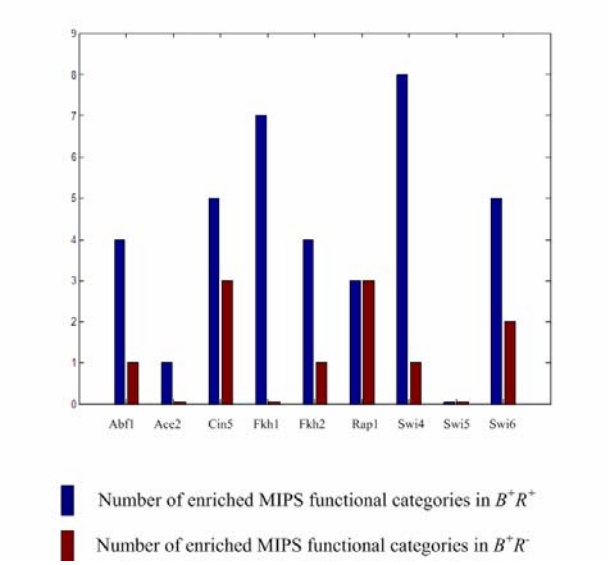

Figure 1 - Enrichment in functional annotation The numbers of significantly enriched MIPS functional categories in B^+R^+ (left, blue) and B^+R^- (right, brown) for each of the nine cell cycle TFs under study are shown.