Genome-wide analysis of gene expression relationships in transcriptional regulatory networks

Haiyuan Yu*, Nicholas M Luscombe*, Jiang Qian¶ and Mark Gerstein‡

Department of Molecular Biophysics and Biochemistry
Yale University
PO Box 208114, New Haven, CT 06520-8114, USA

* These authors contributed equally to this work

¶ Present address: Wilmer Institute, Johns Hopkins School of Medicine, Baltimore, MD, 21287

‡ To whom correspondence should be addressed. Tel: +1 203 432 6105; Fax: +1 360 838 7861; Email: Mark.Gerstein@yale.edu

Abstract

From merging a number of data sources, we created an extensive map of the transcriptional regulatory network in yeast, comprising 7419 interactions connecting 180 transcription factors (TFs) with their target genes. We integrated this network with gene-expression data, relating the expression profiles of TFs and target genes. We found that genes targeted by the same TF tend to be co-expressed, with the degree of co-expression increasing if genes share more than one TF. Moreover, shared targets of a TF tend to have similar cellular functions. In contrast, the expression relationships between the TFs and their targets are much more complicated, often exhibiting time-shifted or inverted behavior.

1. Introduction

An important question in molecular biology is how gene expression is regulated in response to changes in the environment. Previous studies have explored this by making genome-wide measurements of gene expression levels with DNA arrays[1-3] and by searching for transcription factor (TF) binding sites using genetic, biochemical, and large-scale ChIp-chip (chromatin immunoprecipitation and DNA chip) experiments[4-10]. Here, we integrate gene-expression and TF-binding data for *Saccharomyces cerevisiae* in order to determine the effect that regulatory networks have on the expression of targeted genes.

1.1 TF-target regulatory network

We compiled a yeast regulation dataset from merging the results of genetic, biochemical and ChIp-chip experiments[4,5,7,10]. It contains 7,419 TF-target pairs from 180 TFs and 3,474 target genes (Table 1). Regulatory networks can be simplified into six basic motifs (Fig. 1a)[9,10]. Here, we focus on the single input motif (SIM), multi-input motif (MIM) and feed-forward loop (FFL) as the data for the remaining motifs are too sparse.

1.2 Gene expression dataset

We obtained expression profiles of yeast genes through two complete cell cycles.[11] Between the expression profiles of pairs of genes, we used a local clustering method to calculate four types of temporal relationships as diagramed in Fig. 1b[12]: correlated, time-shifted, inverted, and inverted time-shifted. To find these relationships, expression levels must be assessed over a time-course, with many measurements, at small and uniform intervals. Most available datasets do not satisfy these conditions, being only suitable for simple correlation calculations (*ie* co-expression); thus, we can only conduct detailed analysis on the cell-cycle dataset. Nevertheless, similar overall results are observed in other microarray datasets.

1.3 Statistical formalism

We use several statistics to quantify the significance of our observations. The *p*-value is the probability that an observation (*eg* co-expression of target genes) would be made by chance, and is calculated using the cumulative binomial distribution:

$$\mathrm{P}(c \geq c_o) = \sum_{c=c_o}^{N} [\frac{N!}{N!(N-c)!}] p^c (1-p)^{N-c}$$

N is the total number of possible gene pairs in the data, c_o is the number of observed pairs with a specific relationship (*ie* from expression or function), and *p* is the probability of finding a gene pair with the same relationship randomly (picking from the entire genome).

The log odds ratio (LOD) is the enrichment a particular relationship in the presence of regulation with respect to random expectation for the occurrence of the relationship:

$$\mathrm{LOD} = \ln[\frac{P(relationship \mid regulation)}{P(relationship)}]$$

P(relationship | regulation) is the probability for gene pairs with certain regulatory relationship (*eg* TF=>target) to have a specific expression or functional relationship (*eg* correlated expression). *P(relationship)* is the probability for randomly selected gene pairs to have the same expression or functional relationship. When we report this together with p-values, we use the following notation {log p-value,LOD value}.

2. Relationships *between target genes*

2.1 Target genes are co-expressed

First, we investigate expression relationships between genes targeted by the same TFs. Overall, 3.3% of target gene pairs are co-expressed, which is four times greater than random expectation {-12,1.3} (Fig. 2a, bar-ALL). We detect few inverted or time-shifted relationships (§2.4).

The level of correlation is very dependent on the type of regulatory network motif (Fig. 2a). Genes targeted by individual TFs (SIM) are not strongly correlated: just 1.3% of target pairs are co-expressed though this is significantly higher than expected {-11,0.29}. Correlation is stronger for genes targeted by multiple, common TFs: 24.4% of MIM target pairs {-12,3.2} and 5.0% of FFL targets exhibit co-expression {-12,1.6}. Similar results are observed for other expression datasets[3,13-17] (Table 1).

The differences in enrichment (*ie* LOD values) indicate that expression is much more tightly regulated when multiple TFs are involved. However, with >100 yeast transcription factors yet to be investigated[18], unidentified TF-target relationships will probably alter the classification of SIM target genes to MIM or FFL networks in the future.

2.2 Target genes have similar functions

Previous studies showed that co-expressed genes tend to share similar functions[19,20]. By comparing the MIPS (level 2) functional classifications[21], we find that genes targeted by the same TFs are five times more likely to share functions than expected randomly {-12,1.6}(Fig. 2b). Comparing between regulatory motifs, we again see that target genes sharing more than one common TF tend to exhibit this effect to an even greater degree (SIM{-10,1.6}, MIM{-12,2.2}). Interestingly, FFL motifs display the smallest enrichment {-11,1.5}. We speculate that this is because they have specialized effects on

gene expression (see below) and so regulate a very precise subset of genes that do not necessarily share functions, but nonetheless require coordinated expression.

2.3 Co-expression is most likely for target genes with similar functions

We also examine the expression relationships for co-targeted genes that share functions (Fig. 2c). The degree of co-expression is extremely high if targets have the same function, but low if they do not. For example, 75% of MIM target genes are co-expressed if they share functions {-12,4.3} but only 3.6% if they do not {-6,1.3}. Thus, there must be a common set of TFs for genes of similar functions to be co-expressed. Furthermore, though TFs often target genes of various functions, there are regulatory subdivisions and co-expression does not usually extend across functional categories.

2.4 Effect of Regulatory-signal Type

We have limited experimental data describing type of regulatory signal (*ie* activation or repression) for 906 TF-target pairs. Overall, target genes display correlated expression relationships (§2.1). However, we observe more complex relationships once regulatory-signal type is considered (Fig. 2d). Unsurprisingly, co-activated genes have mostly correlated relationships {-12,2.3}. In contrast, co-repressed genes have a variety of relationships. The results indicate that genes activated by the same TFs co-express, but genes inhibited by the same repressors do not always co-express, though they shut down simultaneously.

3. Relationships *between TFs and target genes*

3.1 Complex expression relationships

Next we compare the expression profiles of TFs with their targets (Fig. 2e). Here the relationships are more complex than co-expression: SIMs exhibit time-shifted {-3,0.64} and inverted time-shifted relationships {-2,0.69}, whereas MIMs display inverted time-shifted relationships {-9,1.4}. This suggests that target genes have a delayed response to regulatory events.

FFL motifs present the most interesting and complex relationships. The leading TFs in the motif (denoted TF1) generally have negative relationships with the target genes -- i.e. inverted {-2,0.82} or inverted time-shifted {-10,2.0}. The intermediate TFs (TF2) exhibit all four types of relationships; The most common arrangement (55% of FFLs, supplementary table 2) is where the leading TF has a negative relationship with the target and the intermediate TF has a positive one (*ie* correlated or time-shifted). (Note, however, there are only 11 FFLs for which both TF1 and TF2 have significant expression relationships with the targets.)

3.2 Relation to Regulatory-signal Type

As in §2.4, we can measure the TF-target expression relationships when the type of regulatory signals is taken into account. Though the data is too sparse to make statistically sound conclusions, we try to make some observations. Unsurprisingly, activators are co-expressed with their targets {-2,0.63} (Fig. 2f), and comprise over 50% of TF-target pairs with significant expression relationships. We also find that repressors exhibit inverted {-2,1.1} and inverted time-shifted relationships {-2,1.2}. There are unexpected results too. Activators display significant inverted time-shifted relationships {-6,1.8} and repressors show (normal) time-shifted relationships. There are several reasons for this: A sizeable proportion of TFs (15%) act both as activators and repressors,

in some cases for the same target. Furthermore, the combined effect of multiple TFs in MIM and FFL motifs can have an unpredictable effect on target expression.

4. Examples of TF-target relationships

In Fig. 3 we examine specific regulatory networks.

4.1 SIM: ndd1 network

Ndd1, a cell cycle regulator during S and G_2/M transition[22,23], acts as the sole regulating TF for *MCM21,* kinetochore protein required for normal cell growth from late S to early M phase[24,25], and *STB5*, another transcription factor[26]. All three genes display cell cycle periodicity. *NDD1* peaks early in S and sustains high expression until G2. The targets are co-expressed and time-shifted with respect to *NDD1* by one time-point, peaking later in S.

4.2 MIM: forkhead network

Ndd1 is recruited to G_2/M-transition-specific promoters by Fkh1 and Fkh2, two forkhead transcription activators[22,23,27]. Collectively, these three TFs regulate Dbf2, a kinase needed for cell-cycle regulation[28], and HDR1 (function unknown). The expression profiles of the three TFs are only loosely correlated and peak at different points from early S to late G_2. The targets are time-shifted with respect to *FKH1* by two time-points and peak at the G_2/M transition. The local clustering scores show that their expression profiles are better correlated than in the preceding SIM example (Supplementary Table 3).

4.3 FFL: mbp1/swi4 network

In a feed-forward-loop, Mbp1 (a cell-cycle regulator controlling DNA replication and repair[6,29]) is the leading TF, Swi4 (a cell-cycle regulator controlling cell-wall and membrane synthesis[6,29]) is the intermediate TF, and *SPT21* (a TF involved in histone expression[30]) and *YML102C-A* (function unknown) are the target genes. The profiles of the intermediate TF and target genes are correlated and peak sharply in G_1. In contrast,

the leading TF displays an inverted relationship, which highlights its involvement as a target repressor. (Previous studies have shown Mbp1 acts as an activator for ~50% its targets during the G_1/S transition and as a repressor for ~10% of its targets later in the cycle[6,7,29].)

5. Conclusions

In summary, we find significant connections between the networks from TF-binding experiments and gene expression data. (i) Genes targeted by the same TF are generally co-expressed and the correlation in expression profiles is highest for genes targeted by multiple TFs. (ii) Genes targeted by the same TF tend to share cellular functions, and there are subdivisions within individual network motifs that separate the regulation of genes of distinct functions. (iii) The expression profiles of transcription factors and their target genes display more complex relationships than simple correlation, with the regulatory response of target genes often being delayed.

Acknowledgements

NML is sponsored by the Anna Fuller Fund and MG acknowledges support from the NSF DMS-0241160.

Reference

1 Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. [see comments.]. *Science* 270, 467-70

2 Chee, M. et al. (1996) Accessing genetic information with high-density DNA arrays. *Science* 274, 610-4

3 DeRisi, J.L., Iyer, V.R. and Brown, P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680-6

4 Wingender, E. et al. (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Research* 29, 281-3

5 Guelzim, N., Bottani, S., Bourgine, P. and Kepes, F. (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genetics* 31, 60-3

6 Iyer, V.R. et al. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 409, 533-8

7 Horak, C.E. et al. (2002) Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*. *Genes Dev.* 16, 3017-3033

8 Ren, B. et al. (2000) Genome-wide location and function of DNA binding proteins. *Science* 290, 2306-9

9 Shen-Orr, S.S., Milo, R., Mangan, S. and Alon, U. (2002) Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics* 31, 64-8

10 Lee, T.I. et al. (2002) Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*. *Science* 298, 799-804

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

12 Qian, J. et al. (2001) Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *Journal of Molecular Biology* 314, 1053-66

13 Gasch, A.P. et al. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell* 11, 4241-57

14 Gasch, A.P. et al. (2001) Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Molecular Biology of the Cell* 12, 2987-3003

15 Chu, S. et al. (1998) The transcriptional program of sporulation in budding yeast.[erratum appears in Science 1998 Nov 20;282(5393):1421]. *Science* 282, 699-705

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

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

18 Riechmann, J.L. et al. (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science* 290, 2105-10

19 Gerstein, M. and Jansen, R. (2000) The current excitement in bioinformatics-analysis of whole-genome expression data: how does it relate to protein structure and function? *Current Opinion in Structural Biology* 10, 574-84

20 Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95, 14863-14868

21 Mewes, H.W. et al. (1997) MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Research* 25, 28-30

22 Loy, C.J., Lydall, D. and Surana, U. (1999) NDD1, a high-dosage suppressor of cdc28-1N, is essential for expression of a subset of late-S-phase-specific genes in Saccharomyces cerevisiae. *Molecular & Cellular Biology* 19, 3312-27

23 Koranda, M., Schleiffer, A., Endler, L. and Ammerer, G. (2000) Forkhead-like transcription factors recruit Ndd1 to the chromatin of G2/M-specific promoters. *Nature* 406, 94-8

24 Ortiz, J., Stemmann, O., Rank, S. and Lechner, J. (1999) A putative protein complex consisting of Ctf19, Mcm21, and Okp1 represents a missing link in the budding yeast kinetochore. *Genes & Development* 13, 1140-55

25 Poddar, A., Roy, N. and Sinha, P. (1999) MCM21 and MCM22, two novel genes of the yeast Saccharomyces cerevisiae are required for chromosome transmission. *Molecular Microbiology* 31, 349-60

26 Kasten, M.M. and Stillman, D.J. (1997) Identification of the Saccharomyces cerevisiae genes STB1-STB5 encoding Sin3p binding proteins. *Molecular & General Genetics* 256, 376-86

27 Hollenhorst, P.C., Pietz, G. and Fox, C.A. (2001) Mechanisms controlling differential promoter-occupancy by the yeast forkhead proteins Fkh1p and Fkh2p: implications for regulating the cell cycle and differentiation. *Genes & Development* 15, 2445-56

28 Grandin, N., de Almeida, A. and Charbonneau, M. (1998) The Cdc14 phosphatase is functionally associated with the Dbf2 protein kinase in Saccharomyces cerevisiae. *Molecular & General Genetics* 258, 104-16

29 Koch, C. et al. (1993) A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase.[comment]. *Science* 261, 1551-7

30 Dollard, C. et al. (1994) SPT10 and SPT21 are required for transcription of particular histone genes in Saccharomyces cerevisiae. *Molecular & Cellular Biology* 14, 5223-8

Figure captions:

Figure 1. Schematic representations of transcription regulatory motifs and temporal gene expression relationships. (A) Depiction of the six basic regulatory motifs: ○ TF, □ target. (1) single input motif - target gene has one TF, (2) multi-input motif – target gene has multiple TFs, (3) feed-forward loop – leading TF (TF1) regulates an intermediate TF (TF2) and both regulate the target gene, (4) autoregulation – TF targets itself, (5) multi-component loop – two TFs regulate each other, and (6) regulator chain – set of TFs regulate each other in series. **(B)** Schematic of the four gene expression relationships: (1) correlated (*ie* co-expressed - genes have similar profiles), (2) time-shifted (genes have similar profiles, but one is delayed with respect to the other in the cell cycle), (3) inverted (genes have opposing profiles), and (4) inverted time-shifted. The local clustering method uses a dynamic programming algorithm to align the expression profiles of the genes in question. From the alignment, the method is able to determine which of the four types the relationship is and assign a clustering score measuring the significance of the relationship; for the Cho et al dataset, a score of 13 or above corresponds to a relationship significant to $p = 2.7\times10^{-3}$ (see supplementary materials).

Figure 2. Expression relationships between gene pairs. Log odds ratio (LOD) values above 0 signify observations that are more common than expected by chance, and vice versa (see supplementary materials). Parts A to D show relationships *between target genes (as indicated by the color coding)* for each of the different network motifs. (Note the category "All" includes all gene pairs co-regulated by at least one common transcription factor.**) (A)** LOD values of the likelihood that target gene pairs have correlated expression in different network motifs. **(B)** LOD values of the likelihood that target pairs share the same cellular function. **(C)** LOD values of the likelihood that target pairs with the same function have correlated expression. **(D)** LOD values of the likelihood that co-activated or co-repressed target pairs exhibit one of the four expression relationships. Parts E and F show Expression relationships *between TFs and target genes*. **(E)** LOD values of the likelihood that TFs and their target genes exhibit one of the four expression relationships in different network motifs. FFLs are divided into the TF-target relationship for the leading (TF1) and intermediate TFs (TF2). **(F)** LOD values of the likelihood that activator and repressor TF-target pairs exhibit one of the four expression relationships.

Figure 3. Expression profiles of example regulatory networks during the cell cycle. ⊖ TF, □ target. |→| indicates a time-shift relationship. The inset describes the TF and target genes involved in the example. **(A)** Single input motif, **(B)** multi-input motif, and **(C)** feed-forward loop.

Table 1. Summary of transcription regulatory network dataset.

	motifs†	SIM	MIM	FFL	ALL
	# TFs	119	118	97	188
	# targets	1754	986	511	3416
# TF-target pairs	Total	1754	2781	1523	7419
	Activation‡	37	50	19 - 33§	144
	Repression‡	12	34	23 - 10§	79
LOD values for co-expressed target pairs£	Stress response	0.44*	3.55*	0.59	0.88*
	Sporulation	0.03	0.25	0.08	-0.05
	Diauxic shift	0.11*	1.78*	0.30*	0.30*
	DNA damage	1.24*	4.87*	1.26*	2.14*
	Cell Cycle (Spellman et al.)	0.37*	2.09*	1.62*	0.52*
	" " (Cho et al)	0.29*	2.79*	1.35*	0.93*
	" " (Zhu et al)	0.22*	2.50*	0.91*	0.64*

* LOD values with P-value smaller than 1e-05 (see supplementary Table 1)

† The abbreviation for the motifs is the same as in the caption of Figure 1A. ALL, All the TF-target pairs. There are 3 smaller motifs: Auto, 22 targets, MCL, 31 targets, RC, 119 targets. The random expectation for the number of targets is 6130, the number of yeast genes. The random expectation for the number of gene pairs in yeast is 18785385 = 6130(6129)/2, which is obtained by counting all pairs between yeast genes.

‡ Positive expression relationships (correlated and time-shifted) are considered as activation signals, while negative relationships (inverted and inverted time-shifted) are considered as repression signals. Overall, 18 regulators activate some of their targets but repress others. Note this is distinct from the number of activator relations determined experimentally (as described in §2.4 and §3.2)

§ We show the number of relations for FFL:TF1 and FFL: TF2.

£ Log odds ratios for target gene pairs having correlated profiles in different expression datasets. The local clustering method cannot be applied, so expression correlation is measured using the Pearson correlation coefficient. Co-expressed gene pairs are those in the top 1% of largest correlation coefficients.

.

Supplementary Materials[§]

Determination of the expression relationships using local clustering method (excerpt from Jiang et al, JMB, 314:1053-1066)[*]

"We use a degenerate dynamical programming algorithm to find time-shifted and inverted correlations between expression profiles. The algorithm does not allow gaps between consecutive time points in the current version. However, there are some obvious extensions, which we explore later in the discussion section.

"Suppose there are n $(1,2,\ldots n)$ time-point measurements in the profile. First, the expression ratio is normalized in "Z-score" fashion, so that for each gene the average expression ratio is zero and standard deviation is 1. The normalized expression level at time point i for gene x is denoted as x_i. Consider a matrix of all possible similarities between the expression ratio for gene x and gene y. This matrix can also be called a 'score matrix'. In our algorithm, it is defined as $M(x_i,y_i) = x_i y_j$. For simplification, it will be referred as $M_{i,j}$ for comparison of any two genes.

"Then, two sum matrices **E** and **D** are calculated as $E_{i,j} = \max(E_{i-1,j-1}+M_{i,j}, 0)$ and $D_{i,j} = \max(D_{i-1,j-1}-M_{i,j}, 0)$. The initial conditions are $E_{0,j} = 0$ and $E_{i,0} = 0$, and the same initial conditions are also applied to the matrix of **D**. The central idea is to find a local segment that has the maximal aggregated score, i.e., the sum of $M_{i,j}$ in this segment. This can be accomplished by standard dynamic programming as in local sequence alignment [29] and results in an alignment of l aligned time points, where $l \leq n$.

"Finally, an overall maximal value S is found by comparing the maximums for matrices **E** and **D**. This is the match score S for the two expression profiles. If the maximum is off diagonal in its corresponding matrix, the two expression profiles have a time-shifted relationship. This involves an alignment over a smaller number of time points l than the total number n. A maximal value from matrix **D** indicates these two profiles have an inverted relationship.

"At the end of this procedure, one obtains a match score and a relationship, i.e., 'simultaneous,' 'time-delayed,' 'inverted,' or 'inverted time-delayed'. Obviously, for the gene pairs with a very low match score, even though they are also assigned a relationship, we can classify them as 'unmatched'.

"Figure 1E[†] is the corresponding matrix **E** for the expression profiles shown in Fig. 1B. The matrix **D** for these expression profiles is not shown here because the maximal value

---

[§] Please visit the supplementary website (http://bioinfo.mbb.yale.edu/regulation/TIG/) for further information.

[*] Please note that the "simultaneous" relationship discussed in the JMB paper is the "correlated" relationship discussed in this paper.

[†] Supplementary Figure 1 is the Figure 1 in the JMB paper.

is not in this matrix. The match score for these expression profiles, a score of *S*=19, is highlighted in the black cell. There is a time delay (time shift) in their relationship because the match score of 19 is not on the main diagonal of the matrix. Figure 1F is the corresponding matrix **D** for the profiles shown in Fig. 1C. The match score is *S*=20; and because the maximum value is from matrix **D** rather than **E** (not shown), these expression profiles are correlated in an inverted fashion. "

Supplementary Figure 1. "Three examples showing simultaneous (A), time-delayed (B), and inverted (C) relationships in the expression profiles. Note there are only 8 time points for each profile, while in the real yeast cell-cycle data there are 17 time points. Also, the expression ratio is not normalized, whereas in the real data each profile is normalized so that the averaged expression ratio is 0 and the standard deviation is 1. The thick segments of the expression profiles are the matched part. (D) The corresponding matrix **E** for the expression profile shown in (A). The corresponding matrix **D** is not shown because in this case the match score (the maximal score) is from **E** and not **D**. The numbers outside the border of the matrix are the expression ratio shown in (A). The black cell contains the overall match score *S* for these two expression profiles, and the light gray cells indicate the path of the optimal alignment between the expression profiles. The path starts from the match score and ends at the first encountered 0. (E) The corresponding matrix **E** for the expression profile shown in (B). Note the time-shifted relationship and how the length of the overall alignment can be shorter than 8 positions. (F) The corresponding matrix **D** for the expression profiles shown in (C). The matrix **E** is not shown because the best match score is not from this matrix in this case."

Calculation of the LOD values

Figure 2A

$$\mathrm{LOD} = \ln[\frac{P(co\text{-}exp \mid co\text{-}reg)}{P(co\text{-}exp)}]$$

where *P(co-exp | co-reg)* is the possibility for genes co-regulated by a certain motif to be co-expressed (i.e. correlated), which is calculated as the percentage of correlated pairs between all possible pairs of co-regulated genes. *P(co-exp)* is the possibility for gene pairs randomly chosen from the dataset to be co-expressed, which is calculated as the percentage of correlated pairs between all possible gene pairs in Cho's dataset.

Figure 2B

$$\mathrm{LOD} = \ln[\frac{P(same-function \mid co-reg)}{P(same-function)}]$$

where *P(same-function | co-reg)* is the possibility for gene pairs co-regulated by a certain motifs to have the same functions. *P(same-function)* is the possibility for gene pairs randomly chosen from the dataset to have the same functions.

Figure 2C

$$\mathrm{LOD} = \ln[\frac{P(co-exp \mid same-function, co-reg)}{P(co-exp)}]$$

where P(co-exp | same-function, co-reg) is the possibility for gene pairs that are co-regulated and have the same functions to be co-expressed.

Figure 2D

Log odd ratios for the co-activated gene pairs are calculated by the formula:

$$\mathrm{LOD} = \ln[\frac{P(Exp \mid co\text{-}activated)}{P(Exp)}]$$

Log odd ratios for the co-repressed gene pairs are calculated by the formula:

$$\mathrm{LOD} = \ln[\frac{P(Exp \mid co\text{-}repressed)}{P(Exp)}]$$

where *P(Exp | co-activated)* and *P(Exp | co-repressed)* are the possibilities of having certain expression relationship between co-activated and co-repressed gene pairs, respectively. *P(Exp)* is the possibility for gene pairs randomly chosen from the dataset to have the corresponding expression relationship.

Figure 2E

$$\mathrm{LOD} = \ln[\frac{P(Exp \mid TF - T)}{P(Exp)}]$$

where *P(Exp | TF-T)* is the possibility for the TF-target pairs (TF-T) to have certain expression relationship.

Figure 2F

Log odds ratios between the activators and their targets are calculated by the formula:

$$\mathrm{LOD} = \ln[\frac{P(Exp \mid A - T)}{P(Exp)}]$$

where *P(Exp | A-T)* is the possibility for the activator-target pairs (A-T) to have certain expression relationship.

Log odds ratios between the inhibitors and their targets are calculated by the formula:

$$\mathrm{LOD} = \ln[\frac{P(Exp \mid I - T)}{P(Exp)}]$$

where *P(Exp | I-T)* is the possibility for the inhibitor-target pairs (I-T) to have certain expression relationship.

Table 1

$$\mathrm{LOD} = \ln[\frac{P(co\text{-}exp \mid co\text{-}reg)}{P(co\text{-}exp)}]$$

where all the calculations are very similar to those in Figure 2A, except that the expression relationships between gene pairs are determined using Pearson correlation coefficient in different microarray datasets.

All the possibilities in the analysis are calculated in the same way as in Figure 2A.

Supplementary Table 1. P-values* for the LOD values in Table 1

Motif†	Stress response	Sporulation	Diauxic shift	DNA damage	Cell-cycle by Spellman et al	Cell-cycle by Cho et al	Cell-cycle by Zhu et al
SIM	2.50E-06	0.2958	4.88E-06	1.33E-11	2.28E-11	1.29E-11	2.28E-09
FFL	0.0097	0.2829	5.71E-07	5.81E-07	0	3.95E-13	0
MIM	0	0.1351	0	9.78E-13	3.73E-13	1.48E-12	3.22E-15
ALL	4.67E-11	0.9877	0	1.16E-10	5.96E-10	8.88E-10	0
Correlation coefficient Cut-off‡	0.70	0.95	0.90	0.80	0.70	0.70	0.70

* P-values are calculated by the formula given in text.

† The abbreviation for the motifs is the same as in the caption of Figure 1.

‡ Correlation coefficient cut-off is determined as the Pearson correlation coefficient, above which roughly top 1% gene pairs with the largest correlation coefficient are. The correlation coefficient cut-offs are equivalent to local clustering score of 13.

Supplementary Table 2. Number of FFLs with different regulatory relationships between the regulators and their targets determined from the expression data

Type of FFLs		# of FFLs
TF1-target	TF2-target	
P*	P	3
P	N	2
N	P	6
N	N	0

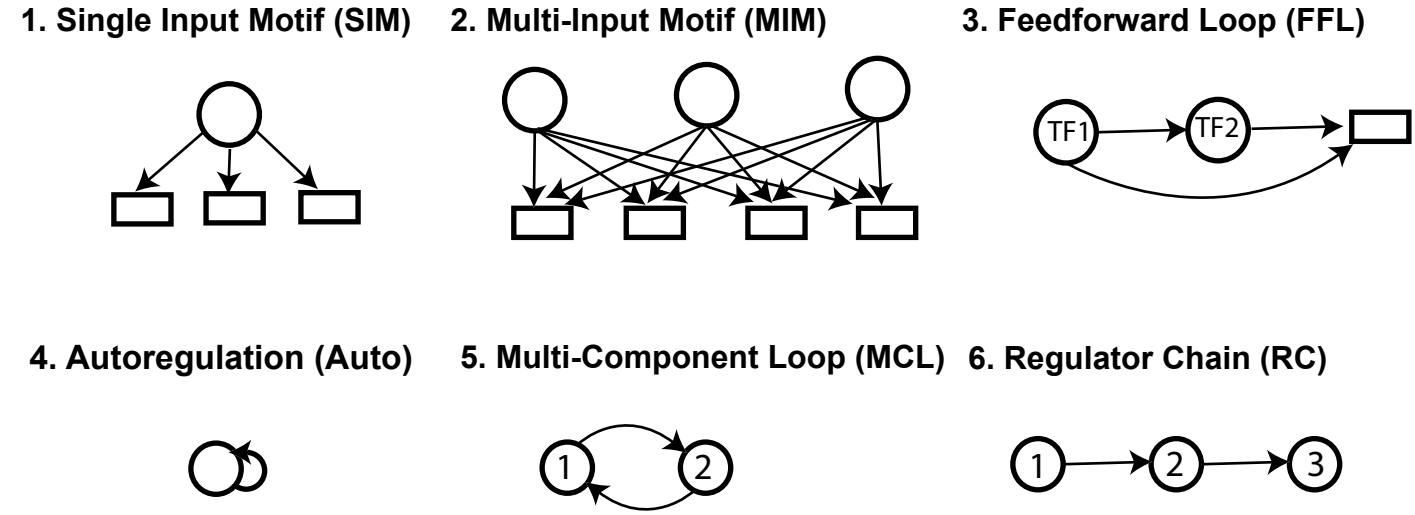
* P: positive relationships between the TFs and their targets; N: negative relationships between the TFs and their targets.

Supplementary Table 3. Relationships and scores between the genes in the examples determined by local clustering

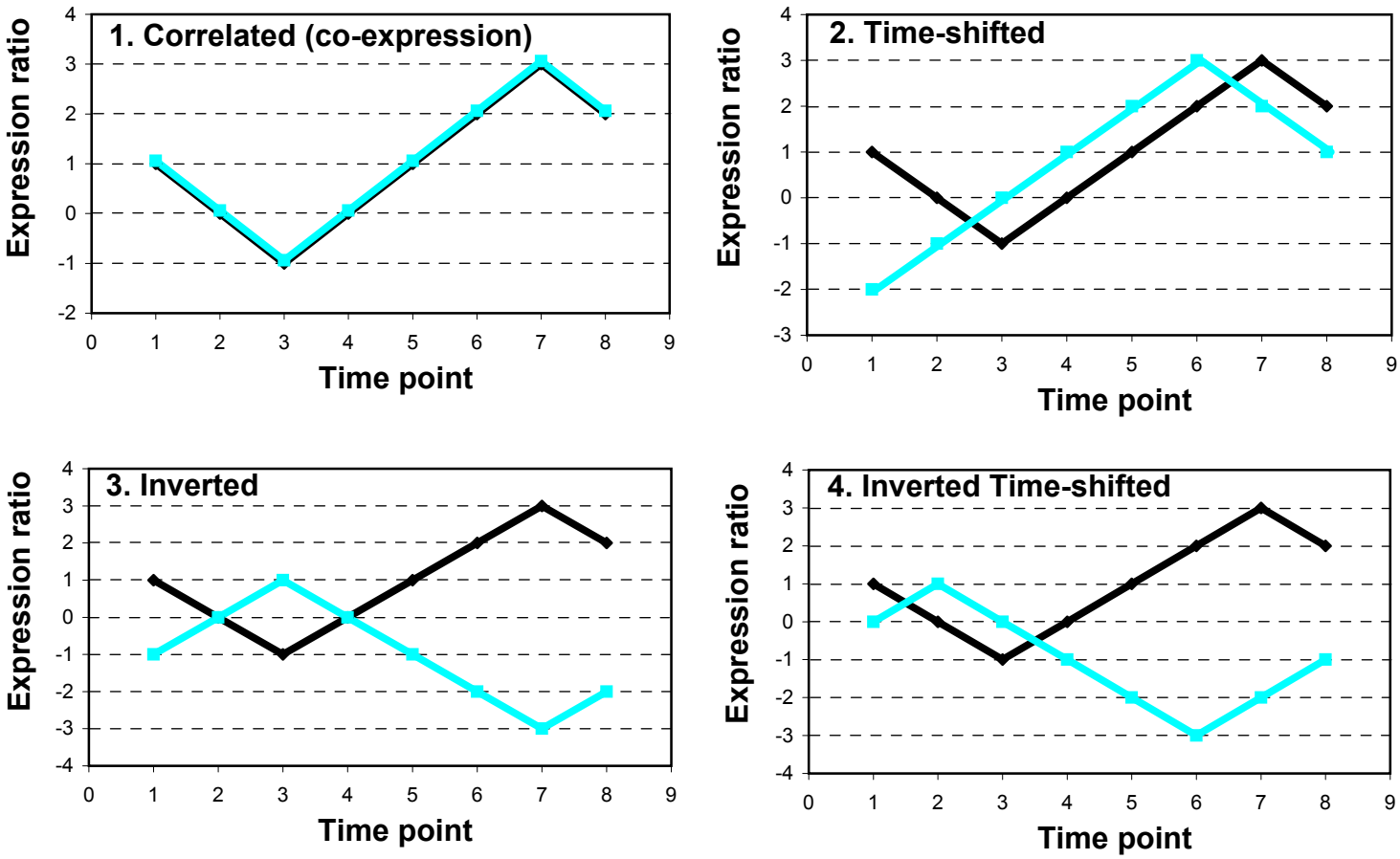
Motif	Gene 1	ORF name	Gene 2	ORF name	Relationship	Local clustering score	P-value
SIM	NDD1	YOR372C	MCM21	YHR178W	Time-shifted	13	2.7e-03
	NDD1	YOR372C	STB5	YDR318W	Time-shifted	13	2.7e-03
	MCM21	YHR178W	STB5	YDR318W	Correlated	13	2.7e-03
MIM	FKH1	YIL131C	FKH2	YNL068C	Time-shifted	12*	1.3e-02
	FKH1	YIL131C	NDD1	YOR372C	Time-shifted	12	1.3e-02
	FKH1	YIL131C	DBF2	YGR092W	Time-shifted	13	2.7e-03
	FKH1	YIL131C	HDR1	YBR138C	Time-shifted	13	2.7e-03
	FKH2	YNL068C	NDD1	YOR372C	Time-shifted	12	1.3e-02
	FKH2	YNL068C	DBF2	YGR092W	Time-shifted	13	2.7e-03
	FKH2	YNL068C	HDR1	YBR138C	Time-shifted	14	3.8e-04
	NDD1	YOR372C	DBF2	YGR092W	Time-shifted	13	2.7e-03
	NDD1	YOR372C	HDR1	YBR138C	Time-shifted	12	1.3e-02
	DBF2	YGR092W	HDR1	YBR138C	Correlated	15	2.9e-05
FFL	MBP1	YDL056W	SWI4	YER111C	Inverted	14	3.8e-04
	MBP1	YDL056W	SPT21	YMR179W	Inverted	12	1.3e-02
	MBP1	YDL056W	YML102C-A	YML102C-A	Inverted	13	2.7e-03
	SWI4	YER111C	SPT21	YMR179W	Correlated	14	3.8e-04
	SWI4	YER111C	YML102C-A	YML102C-A	Correlated	15	2.9e-05
	SPT21	YMR179W	YML102C-A	YML102C-A	Correlated	14	3.8e-04

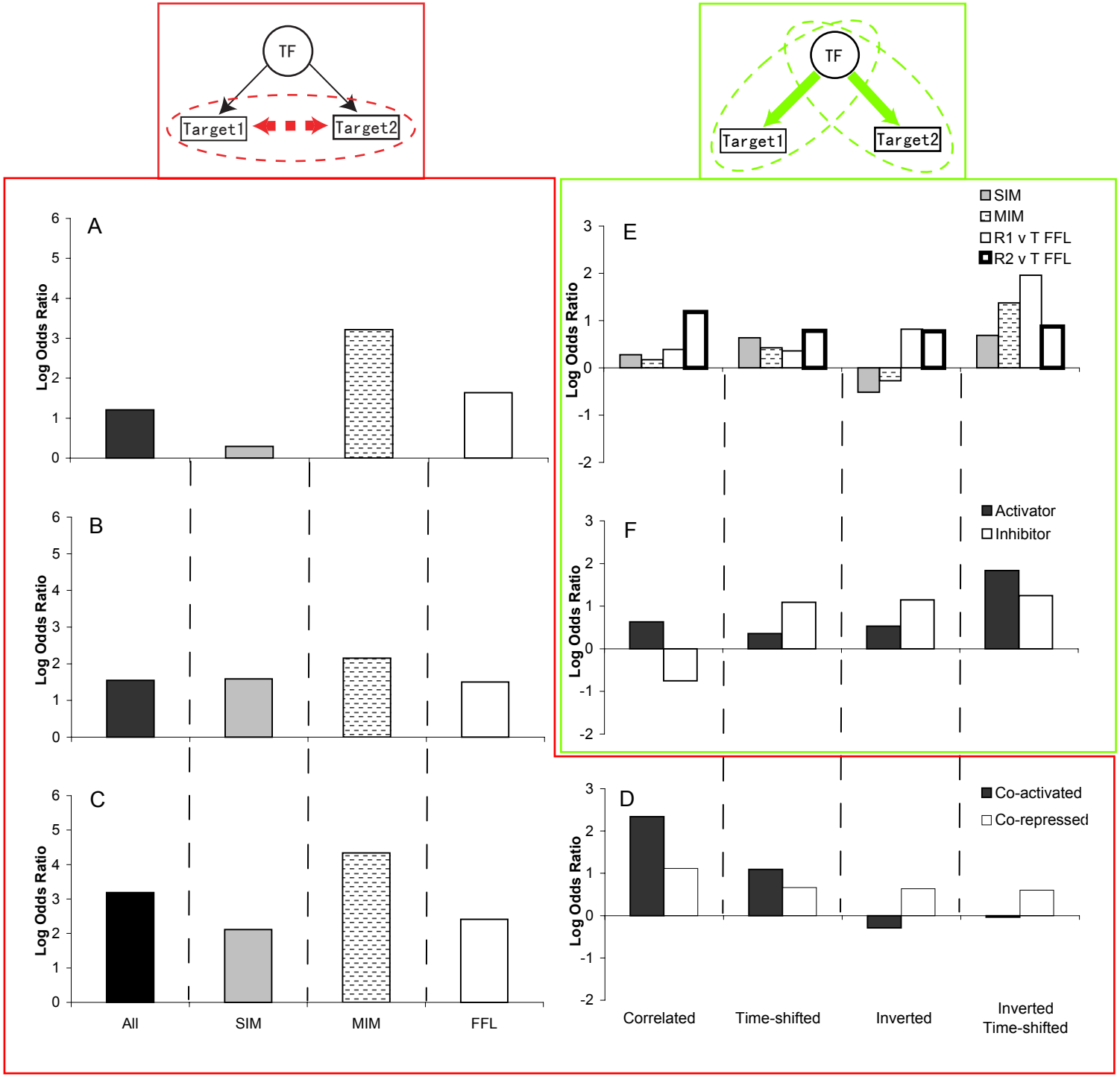
* Local clustering score of 12 is equivalent to correlation coefficient of about 0.6

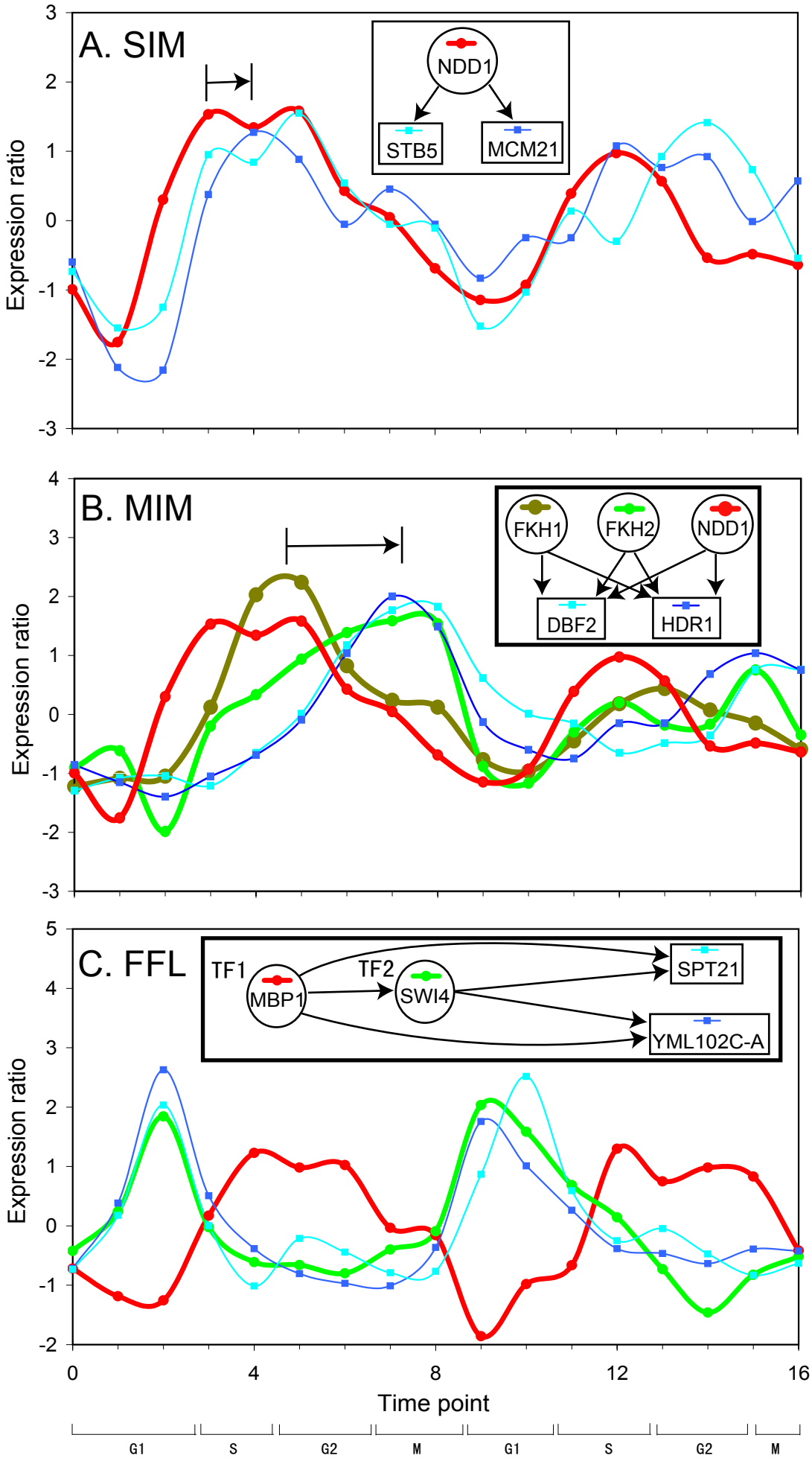
A. Regulatory motifs



B. Expression relationships




TF
Target1
Target2
TF
Target1
Target2
A
B
C
D
E
F
Log Odds Ratio
All
SIM
MIM
FFL
SIM
MIM
R1 v T FFL
R2 v T FFL
Activator
Inhibitor
Co-activated
Co-repressed
Correlated
Time-shifted
Inverted
Inverted Time-shifted


A. SIM
NDD1
STB5
MCM21
Expression ratio
B. MIM
FKH1
FKH2
NDD1
DBF2
HDR1
Expression ratio
C. FFL
TF1
MBP1
TF2
SWI4
SPT21
YML102C-A
Expression ratio
Time point
G1
S
G2
M
G1
S
G2
M