Genomic analysis of gene expression relationships in transcriptional regulatory networks

Haiyuan Yu[1]*, Nicholas M Luscombe[1]*, Jiang Qian[2] and Mark Gerstein[1]

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

[2]Wilmer Institute, Johns Hopkins School of Medicine, Baltimore, MD 21287, USA

Corresponding author: Mark Gerstein (Mark.Gerstein@yale.edu).

*These authors contributed equally to this work.

From merging several data sources, we created an extensive map of the transcriptional regulatory network in *Saccharomyces cerevisiae*, 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. By contrast, the expression relationships between the TFs and their targets are much more complicated, often exhibiting time-shifted or inverted behavior. Further information is available at http://bioinfo.mbb.yale.edu/regulation/TIG/

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* to determine the effect that regulatory networks have on the expression of targeted genes.

TF–target regulatory network

We compiled a yeast regulation dataset by merging the results of genetic, biochemical and ChIp-chip experiments [4,5,7,10]. It contains 7419 TF-target pairs from 180 TFs and 3474 target genes (Table 1). Regulatory networks can be simplified into six basic motifs [9,10] (Fig. 1a). 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.

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 [12] (Fig. 1b): 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 (i.e. co-expression); thus, we can only conduct detailed analysis on the cell-cycle dataset. Nevertheless, similar overall results are observed in other microarray datasets.

Statistical formalism

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

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

N is the total number of possible gene pairs in the data, c_o is the number of observed pairs with a specific relationship (i.e. 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:

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

$P(\text{relationship} \mid \text{regulation})$ is the probability of gene pairs with certain regulatory relationship (e.g. TF→target) having a specific expression or functional relationship (e.g. correlated expression). $P(\text{relationship})$ is the probability of randomly selected gene pairs having the same expression or functional relationship. When we report this together with *P*-values, we use the following notation {log *P*-value;LOD value}.

Relationships between target genes

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, column 'All'). We detect few inverted or time-shifted relationships (see section 'effect of regulatory-signal type').

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 although 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 (i.e. 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.

Target genes have similar functions

Previous studies showed that co-expressed genes tend to share similar functions [19,20]. By comparing the MIPS (Munich Information Center for Protein Sequences, 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.

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

We also examined 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, although TFs often target genes of various functions, there are regulatory subdivisions and co-expression does not usually extend across functional categories.

Effect of regulatory-signal type

We have limited experimental data describing type of regulatory signal (i.e. activation or repression) for 906 TF-target pairs [5]. Overall, target genes display correlated expression relationships (see section 'Target genes are co-expressed'). However, we observe more complex relationships once regulatory-signal type is considered (Fig. 2d). Unsurprisingly, co-activated genes mostly have correlated relationships {–12;2.3}. By 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, although they shut down simultaneously.

Relationships between TFs and target genes

Complex expression relationships

Next we compared 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; that is, inverted {–2;0.82} or inverted time-shifted {–10;2.0}. The intermediate TFs (TF2) exhibit all four types of relationship. The most common arrangement (55% of FFLs, Supplementary Table 2 at http://download.bmnqc.com/supp/tig/Ru230_Yu.pdf) is where the leading TF has a negative relationship with the target and the intermediate TF has a positive one (i.e. correlated or time-shifted). (Note, however, there are only 11 FFLs for which both TF1 and TF2 have significant expression relationships with the targets.)

Relation to regulatory-signal type

As in section 'Effect of regulatory-signal type', we can measure the TF–target expression relationships when the type of regulatory signals is taken into account. Although the data are 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.

Examples of TF-target relationships

In Fig. 3 we examine specific regulatory networks.

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*, encoding another transcription factor [26]. All three genes display cell cycle periodicity. *NDD1* expression 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.

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 at http://download.bmnqc.com/supp/tig/Ru230_Yu.pdf).

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. By 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].)

Conclusions

In summary, we find significant connections between the networks from TF-binding experiments and gene expression data. (1) Genes targeted by the same TF are generally co-expressed and the correlation in expression profiles is highest for genes targeted by multiple TFs. (2) 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. (3) 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. Note that our results are fairly robust with respect to specifics of the regulatiory network. As a check, we recalculated all our results using just the interactions in the Lee *et al.* dataset (106 regulators and 2416 genes) [10], and we got essentially the same results.

Data availability

The datasets used for this analysis are available at http://bioinfo.mbb.yale.edu/regulation/TIG/.

Acknowledgements

We thank Duncan Milburn for computational support. N.M.L. is sponsored by the Anna Fuller Fund and M.G. acknowledges support from the NSF DMS-0241160.

References

1 Schena, M. *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470

2 Chee, M. *et al.* (1996) Accessing genetic information with high-density DNA arrays. *Science* 274, 610–614

3 DeRisi, J.L. *et al.* (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686

4 Wingender, E. *et al.* (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.* 29, 281–283

5 Guelzim, N. *et al.* (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nat. Genet.* 31, 60–63

6 Iyer, V.R. *et al.* (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 409, 533–538

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, 2306–2309

9 Shen-Orr, S.S. *et al.* (2002) Network motifs in the transcriptional regulation network of *Escherichia coli. Nat. Genet.* 31, 64–68

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. *Mol. 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. *J. Mol. Biol.* 314, 1053–1066

13 Gasch, A.P. *et al.* (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* 11, 4241–4257

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

15 Chu, S. *et al.* (1998) The transcriptional program of sporulation in budding yeast. *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, 2105–2110

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? *Curr. Opin. Struct. Biol.* 10, 574–584

20 Eisen, M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863–14868

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

22 Loy, C.J. *et al.* (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. Mol. Cell. Biol.* 19, 3312–3327

23 Koranda, M. *et al.* (2000) Forkhead-like transcription factors recruit Ndd1 to the chromatin of G2/M-specific promoters. *Nature* 406, 94–98

24 Ortiz, J. *et al.* (1999) A putative protein complex consisting of Ctf19, Mcm21, and Okp1 represents a missing link in the budding yeast kinetochore. *Genes Dev.* 13, 1140–1155

25 Poddar, A. *et al.* (1999) MCM21 and MCM22, two novel genes of the yeast *Saccharomyces cerevisiae* are required for chromosome transmission. *Mol. Microbiol.* 31, 349–360
26 Kasten, M.M. and Stillman, D.J. (1997) Identification of the *Saccharomyces cerevisiae* genes STB1-STB5 encoding Sin3p binding proteins. *Mol. Gen. Genet.* 256, 376–386
27 Hollenhorst, P.C. *et al.* (2001) Mechanisms controlling differential promoter-occupancy by the yeast forkhead proteins Fkh1p and Fkh2p: implications for regulating the cell cycle and differentiation. *Genes Dev.* 15, 2445–2456
28 Grandin, N. *et al.* (1998) The Cdc14 phosphatase is functionally associated with the Dbf2 protein kinase in *Saccharomyces cerevisiae*. *Mol. Gen. Genet.* 258, 104–116
29 Koch, C. *et al.* (1993) A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase. *Science* 261, 1551–1557
30 Dollard, C. *et al.* (1994) SPT10 and SPT21 are required for transcription of particular histone genes in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* 14, 5223–5228

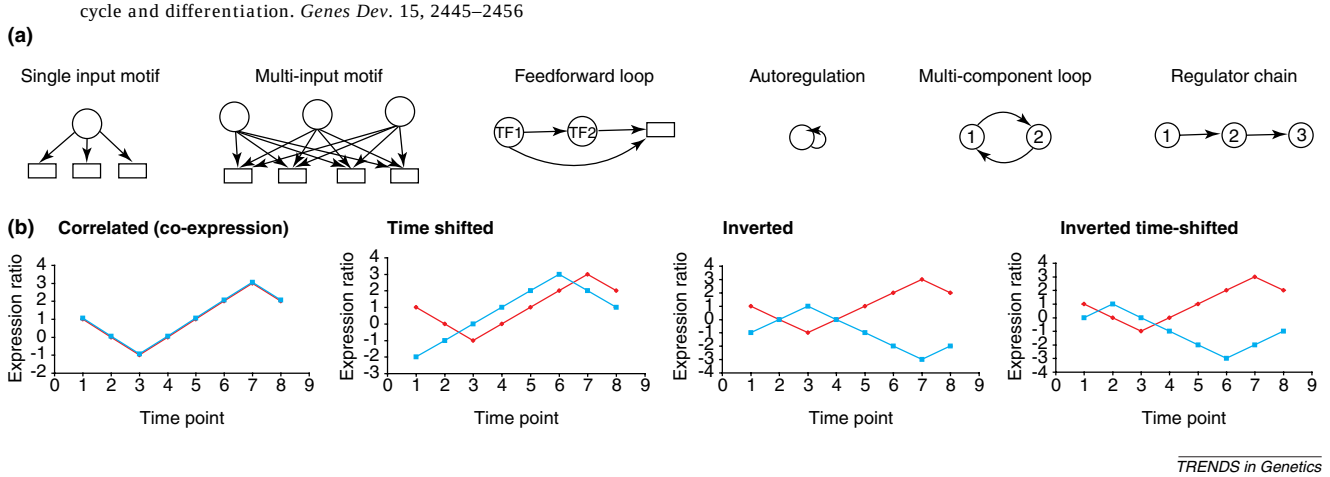

Fig. 1. Transcriptional regulatory motifs and temporal gene expression relationships. (a) The six basic regulatory motifs, where circles represent the transcription factors (TFs) and squares, targets. For the single input motif, the target gene has one TF; for the multi-input motif, target gene has multiple TFs. In the feed-forward loop, the leading TF (TF1) regulates an intermediate TF (TF2), and both regulate the target gene. In autoregulation, the TF targets itself, and in the multi-component loop, two TFs regulate each other. In a regulator chain, a set of TFs regulate each other in series. (b) The four gene expression relationships: correlated, where genes have similar profiles (co-expressed); time-shifted, where genes have similar profiles, but one is delayed with respect to the other in the cell cycle; inverted, where genes have opposing profiles; and 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 [11], a score of 13 or above corresponds to a relationship significant to $P = 2.7 \times 10^{-3}$ (see Supplementary data at http://download.bmnqc.com/supp/tig/Ru230_Yu.pdf).

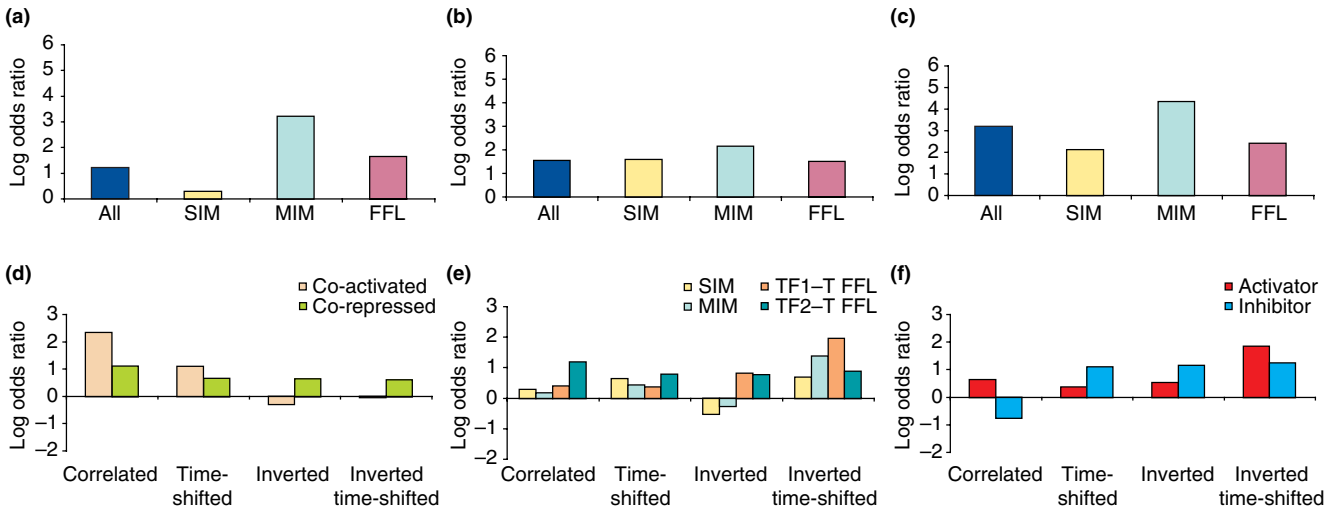

Fig. 2. Expression relationships between gene pairs. Log odds ratio (LOD) values above 0 signify observations that are more common than expected by chance, and vice versa (see Supplementary data at http://download.bmnqc.com/supp/tig/Ru230_Yu.pdf). (a–d) Relationships between target genes (as indicated 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. (e,f) 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. FFL, feed-forward loop; MIM, multi-input motif; SIM, single-input motif; TF1–T FFL, TF1–target pairs in FFLs; TF2–T FFLs, TF2–target pairs in FFLs.

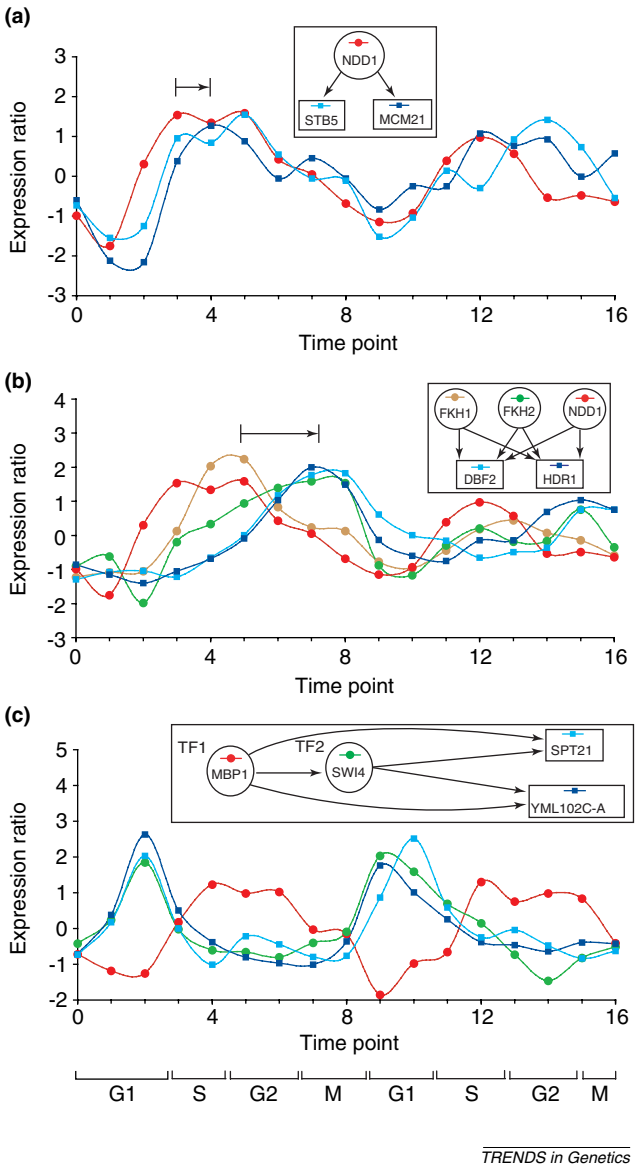

Fig. 3. Expression profiles of example regulatory networks during the cell cycle. Circles show transcription factors (TFs), squares show targets. The arrow indicates a time-shift relationship. The inset shows the relationships between the TF and target genes involved in the example. (a) Single input motif; (b) multi-input motif; (c) feed-forward loop.

Table 1. Summary of transcriptional regulatory network dataset

	Motifs[b]	SIM	MIM	FFL	All	Refs
	No. of TFs	119	118	97	188	
	No. targets targets	1754	986	511	3416	
No. of TF–target pairs	Total	1754	2781	1523	7419	
	Activation[c]	37	50	19–33[d]	144	
	Repression[c]	12	34	23–10[d]	79	
LOD values for co-expressed target pairs[e]	Stress response	0.44[a]	3.55[a]	0.59	0.88[a]	[13]
	Sporulation	0.03	0.25	0.08	-0.05	[15]
	Diauxic shift	0.11[a]	1.78[a]	0.30[a]	0.30[a]	[3]
	DNA damage	1.24[a]	4.87[a]	1.26[a]	2.14[a]	[14]
	Cell cycle	0.37[a]	2.09[a]	1.62[a]	0.52[a]	[17]
	Cell cycle	0.29[a]	2.79[a]	1.35[a]	0.93[a]	[11]
	Cell cycle	0.22[a]	2.50[a]	0.91[a]	0.64[a]	[16]

Abbreviations: All, all the transcription factor–-target pairs; FFL, feed-forward loop; LOD, log odds ratio; MIM, multi-input motif; SIM, single-input motif; TF, transcription factor.

[a]LOD values with *P*-value smaller than 1×10^{-5} (see Supplementary Table 1 at http://download.bmnqc.com/supp/tig/Ru230_Yu.pdf)

[b]There are three 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.

[c]Positive expression relationships (correlated and time-shifted) are considered as activation signals, whereas 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 sections 'Effect of regulatory-signal type' and 'Relations to regulatory-signal type)

[d]We show the number of relations for TF1–target pairs (N1) and TF2–target pairs (N2) in FFLs, using the following notation (N1–N2).

[e]LOD values 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.