Supplementary Materials[§]

Determination of the expression relationships using local clustering method (excerpt from Qian 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