PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2010, 37(4): 402~415
www.pibb.ac.cn

A Novel Method to Identify The Condition-specific Regulatory Sub-network That Controls The Yeast Cell Cycle Based on Gene Expression Model*

LIU Qi-Jun[1,2], WANG Zheng-Hua[1], LIU Wan-Lin[2], LI Dong[2], HE Fu-Chu[2,3]**, ZHU Yun-Ping[2]**
([1] *National Laboratory for Parallel & Distributed Processing, National University of Defense Technology, Changsha* 410073, China;
[2] *State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Radiation Medicine, Beijing* 102206, China;
[3] *Institutes of Biomedical Sciences, Fudan University, Shanghai* 200032, China)

Abstract The huge datasets produced from high-throughput microarray technology can elucidate unknown mechanisms of gene regulation in biological systems. Because biological processes are dynamic, it is relevant to focus on certain condition-specific gene regulatory sub-networks. The cell cycle is a basic cellular process, thus, identifying cell cycle specific regulatory sub-networks in yeast will provide a basis for understanding the cell cycle and may be important in other cellular conditions. With a gene expression differential equation model (GEDEM), dynamic cell cycle-related regulatory relationships were indentified from a static regulatory network. Compared to cell cycle-related regulatory interactions previously published, this method identified more true regulatory relationships and show higher performance than other methods. On larger datasets, the GEDEM identified regulatory sub-networks with high sensitivity and specificity. Further analysis on combinatorial regulation revealed that condition-specific regulatory sub-networks exhibited more significant correlations between transcription factors than previously implied in static network analyses, which infer that the condition-specific sub-networks are closer to reality than static network. Additionally, the GEDEM identified more potential co-regulatory transcriptional factors in the cell cycle.

Key words gene regulatory network, cell cycle, gene expression model, condition-specific sub-network, differential equation model
DOI: 10.3724/SP.J.1206.2009.00581

Gene regulation is the link between genetic material and proteins, thus, it is one of the key control mechanisms in biological systems. It is important to understand the details of control mechanisms in biological processes in order to grasp underlying biological principles and effectively guide drug design. It is well known that regulatory pathways are dynamic[1], and cell cycle regulation is a key biological process. Therefore, identifying dynamic regulatory sub-networks that control the cell cycle would be useful in understanding gene regulation in the cell cycle, and similar sub-networks may be involved in other biological processes.

There are several methods for identifying regulatory network modules (or sub-networks), including those in the cell cycle, though most methods are not explicitly designed to define condition-specific dynamic regulatory sub-networks. Some methods are based only on network topology, for example, the "origon" module [2] and RegulonDB [3]. Other methods combine network topology with gene expression information, for example, the GRAM (Genetic

*This work was supported by grants from The State Key Laboratory of Proteomics(SKLP-O200811), The National Natural Science Foundation of China (60603054, 30800200, 30621063), National Basic Research Program of China (2006CB910803, 2006CB910706, 2010CB912700), Hi-Tech Research and Development Program of China (2006AA02A312) and National S&T Major Project (2008ZX10002-016, 2009ZX09301-002).
**Corresponding author.
HE Fu-Chu. Tel: 86-10-68171208, Fax: 86-10-68214653
E-mail: hefc@nic.bmi.ac.cn
ZHU Yun-Ping. Tel: 86-10-80705225, Fax: 86-10-80705155
E-mail: zhuyunping@gmail.com
Received: December 15, 2009 Accepted: January 21, 2010

Regulatory Modules) algorithm[4]. There are also many methods for integrating numerous other types of information. One example is SAMBA (Statistical-Algorithmic Method for Bicluster Analysis)[5], which integrates information on growth profiles, binding location profiles, protein interactions, and complex interactions. Another example is MOFA (MOdule Finding Algorithm)[6], which is based on gene expression and ChIP-chip data[7–9]. Lee *et al*.[10] identified condition-specific regulatory modules that integrated separate units of gene expression profiles along with ChIP-chip and motif data in yeast.

However, most of the above methods did not focus on the time-course of changes in gene expression, the basic phenotype for regulatory relationships that we want to understand. Network component analysis (NCA) was presented as a method for uncovering hidden regulatory signals underlying gene expression[11]. NCA can be used on time-course data to infer transcriptional factor (TF) activity, which is useful for analyzing regulatory signal dynamics[11– 12]. Recently, Ernst *et al*. (2007)[13] identified bifurcation points from time series gene expression data to map dynamic regulation in response to stress. Apart from these condition-specific time-course methods, multiple datasets are useful for constructing pairwise regulatory relationships that are components of static regulatory networks[14–15].

Currently, many researchers are focused on studies related to the notion of dynamic or condition-specific regulatory networks[10, 16–17]. This notion was explicitly discussed by Luscombe *et al*. (2004)[10], who elucidated a dynamic network from their own method, the Trace-Back Algorithm (TBA). The TBA identifies dynamic sub-networks based on active target genes (TGs) that are differentially expressed, TF expression levels (either present or absent), and the relationships between TFs and target genes (TGs). In TBA, it is important to identify the status (present or absent) of TFs. This information can be obtained by comparing gene expression levels in the specific condition or phase with expression levels in the cell cycle. Kim *et al*. (2006)[16] improved the signature algorithm (SA, advanced by lhmels *et al*. in 2003[18]) by using bi-clustering to identify genes that were co-expressed in condition-specific gene transcriptional regulatory networks; this improved method is known as the united signature algorithm (USA). USA applies a condition score scheme based on gene expression, and then assigns a gene score, thereby identifying condition-specific genes that form sub-networks. Because this method regards the problem to be one of classification, rather than recognition, every regulatory relationship must be assigned to a condition. For a set of focus conditions, USA requires a corresponding set of static regulatory relationships. ChIP-chip data yields regulatory relationships according to the experimental designs, thus the USA method may produce many false positive results.

Gene expression datasets can be used to construct static regulatory networks[19]. Several mathematical models have been introduced that simulate regulatory networks[20], for example, directed and undirected graph boolean networks[21–23], Bayesian networks[24], ordinary and partial differential equation systems[25], stochastic equations[26], Markov chains[27–28], and some rule-based formalisms. Of these methods, the differential equation model is most suitable for uncovering detailed knowledge about gene regulatory mechanisms. Vu *et al*.(2007) used nonlinear differential equations to model gene expression in the cell cycle, but in that integrated model, they could only calculate the relationships of a single TF to a single TG.

To quantitatively characterize functional relationships at the molecular level, we improved the differential equation model of Vu *et al*. to represent the relationship between TGs and all their regulators (TFs). We divided the weight for each TF-TG into two parts, the modified weight and the indicator. The latter indicates whether this TF regulates the TG in the focus condition. Thus, compared to the other differential equation-based works[20, 25, 29], our method, the Gene Expression Differential Equation Model (GEDEM), aimed to identify sub-networks in a focus biological condition. Applying the GEDEM to a widely used microarray dataset[30–31] and a static regulatory network[32], we obtained several cell cycle related regulatory sub-networks. Compared to other current methods, the GEDEM exhibited superior performance. Based on a larger dataset, the GEDEM identified regulatory sub-networks with high sensitivity and specificity. Further analysis of combinatorial regulation revealed that condition-specific regulatory sub-networks exhibited more significant correlations between TFs than previously implied in static network analyses. In addition, more potential co-regulatory TFs were identified.

1 Methods

1.1 Differential equation model for gene expression

It is well known that a particular TG may have one or many regulators. In our model, TGs are regarded as single transcript processing units. The transcription regulatory effect is the combined input from its regulators (Figure 1).

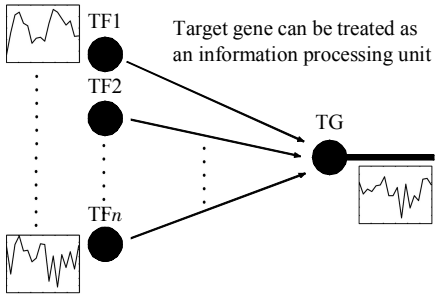

Fig. 1 A target gene can be treated as a transcript processing unit

The target gene processes the combined signals from its regulators. Its output is mRNA expression level.

The transcriptional process can be modeled by a simple first-order differential equation[33]:

$$y_i'(t)=R_i(t)-\lambda_i y_i(t)+\varepsilon_i(t)$$

Where $y_i(t)$ denotes the expression level of the i-th gene at time t; $R_i(t)$ and λ_i are the transcriptional rate and self-degradation rate, respectively, and $\varepsilon_i(t)$ represents noise that may be caused by uncertain data measurements or may be due to irregularities in model behavior. This equation means that changes in mRNA levels are due to the synthesis rate, which is controlled by the transcription and degradation rates. The transcription rate $R_i(t)$ is a complex function of the combined effects of all relevant TFs.

From another point of view, some groups have used a similar model in the following form[25, 29, 34]:

$$y_i'(t) = k_{1i} f(g_i)-k_{2i} y_i(t)+\varepsilon_i(t)$$

$$g_i = \sum_{j=1}^{n} \omega_{ij} x_j(t)+b_i$$

Where g_i is the combined effect of TFs on gene i, ω_{ij} is the weight that determines the influence of TF j on TG i, f is the activation function, k_{1i} denotes the reaction rate of gene i, k_{2i} represents the degradation constant for gene i, $x_j(t)$ describes the mRNA level of the TFs that regulate TG i, while b_i is a lower base level of gene i.

In these researches, they adopted two action functions mainly. One is the identity function $f(x)=x$[34] and the other is the sigmoid function $f(x)=(1+e^{-x})^{-1}$[25]. While fitting the model of gene expression datasets, the identity function has lower than the latter one[25]. Therefore, we choose the sigmoid function in our work.

In order to identify the regulatory sub-network, we introduced a novel variable v_{ij}:

$$v_{ij}=\begin{cases}1 & \text{TF } j \text{ regulates target gene } i \text{ under specific conditions}\\ 0 & \text{Otherwise}\end{cases} \quad (1)$$

The weight ω_{ij} can be expressed as $\omega_{ij}=u_{ij}v_{ij}$, where u_{ij} is the dynamic weight without considering the presence of regulatory relationships.

As we are all known, gene expression datasets are noisy, which is caused by systematic errors, the errors in measuring and sampling of the experiments. The datasets we used are not raw but standardized ones. The data preprocess has already included denoising step which would make background noise not affect gene expression levels. In other sides, the noises in the model often cause the large increasing of computation to resolve. Here, we omit the noises of gene expression levels.

Therefore, our dynamic gene expression differential equation model can be expressed as follows:

$$y_i'(t)=k_{1i}\frac{1}{1+e^{-(\sum_{j=1}^{n} u_{ij}v_{ij}x_j(t)+b_i)}}-k_{2i}y_i(t) \quad (2)$$

Where v_{ij} denotes the elements of the required condition-specific regulatory matrix.

1.2 The framework for implementation of the GEDEM

The GEDEM uses two types of datasets: the microarray datasets and the static regulatory network. Its implementation includes the following steps: preparing the dataset, selecting a TG, processing, and evaluation. The main framework and the details are described in Figure 2.

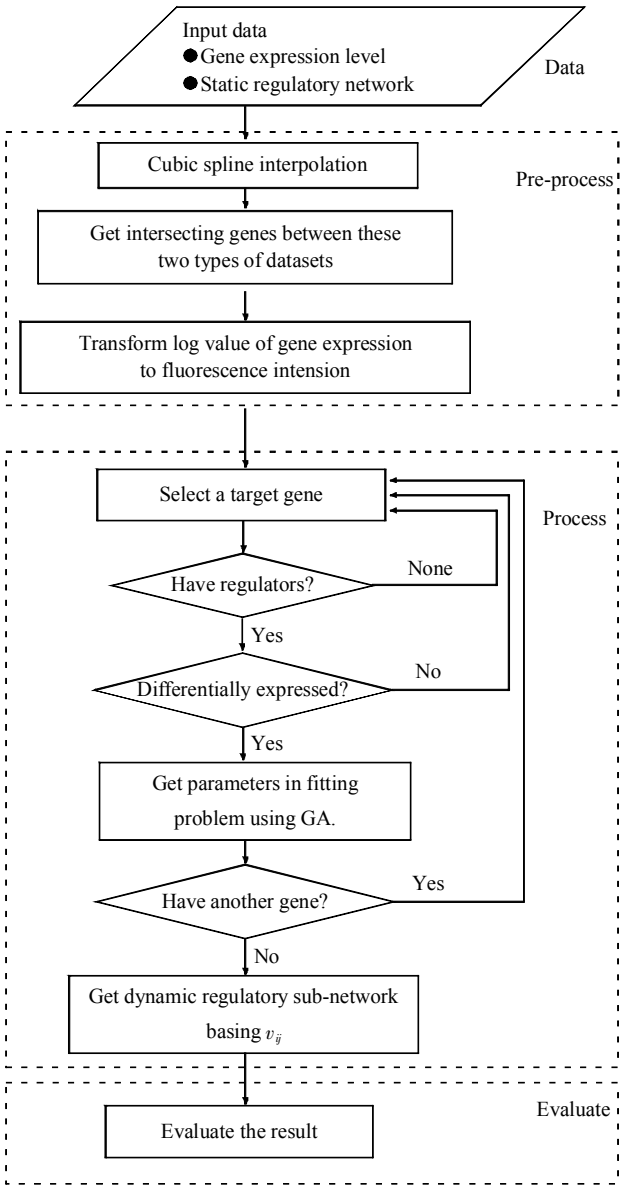

Fig. 2 The work flow for the identification of condition-specific regulatory sub-networks

1.3 The correlation between the expression levels of two genes

We calculated the correlation between the expression levels of two genes (including TFs) using the Pearson correlation coefficient, which gives the best estimate of a linear correlation:

$$corr(P, Q)=\frac{\sum p_i q_i - n\overline{pq}}{(n-1)s_p s_q}$$

$$=\frac{n\sum p_i q_i - \sum p_i \sum q_i}{\sqrt{n\sum p_i^2-(\sum p_i)^2}\ \sqrt{n\sum q_i^2-(\sum q_i)^2}}$$

Where P, Q are vectors whose elements are p_i, q_i (i=1, 2, ⋯, n).

1.4 The performance measurements for evaluating the recognization

Identifying the condition-specific sub-networks is a classification problem whose aim is to classify all regulatory associations into two groups: under and outside of condition-specific sub-network. To evaluate its performances, we can compare the classification result with the standard datasets (Figure 3).

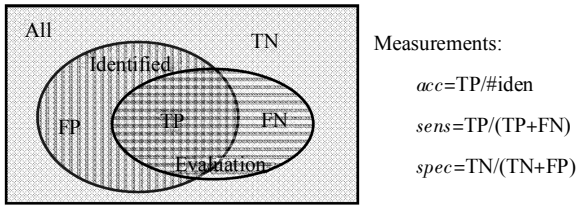

Fig. 3 The performance measurements for classification methods

TP means the true identified regulatory associations; FP is the false identified ones; FN denotes the true regulatory associations which did not identified and TN contains the negative ones not identified. : FP; : TP; : FN; : TN. TP +FP =Identified (#iden); TP+FN=Evaluation; TP+FP+FN+TN=All.

As defining in Figure 3, there are three popular performance measurements for this task, the accuracy (acc), the sensitivity ($sens$) and the specificity ($spec$).

2 Results

2.1 Preparing datasets for the GEDEM

We applied our gene expression model to the well-known yeast cell cycle dataset of Spellman *et al.* (1998)[30]. This dataset includes data from several types of experiments that used different synchronization methods for cell cultures, including alpha factor arrest, CDC15 mutant arrest, CDC28 mutant arrest [31], and elutriation. We used the static regulatory network of yeast from the YeastRACT database [35], which contains 169 TFs, 5 402 TGs, and 27 858 regulatory relationships (July 2007 version).

We collected evaluating datasets from references that provided evidence for the regulatory relationships in the YeastRACT database and the *Saccharomyces* Genome database (SGD) website. All sources were manually checked, and we found 529 regulatory relationships between 47 TFs and 276 TGs that occurred in the yeast cell cycle (Evaluation datasets are available for request).

In summary(Table 1), our dataset contained 5 350 genes that were in the static regulatory network and in gene expression experiments. Among these, 168 genes were TFs. Based on the two-fold rule (i.e., the maximum expression level should greater than or equal to two times the control expression level), we could expect that 2 350, 4 648, 3 845 and 2 985 genes would be differentially expressed in the four cell cycle experiments.

Table 1 The yeast cell cycle datasets before and after pre-processing

	Original datasets		After pre-processing	
Type	Dataset and reference	Size[1)]	Differential expression gene	Size of curated dataset[1)]
Microarray	Alpha factor [30]	6 178	2 350	(168, 5 350, 27 059)
	CDC15 mutation [30]	6 178	4 648	
	CDC28 mutation [31]	6 178	3 845	
	Elutriation [30]	6 178	2 985	
Static regulatory network	YeastRACT Database	(169, 5 571, 27 858)	---	
Evaluation dataset[2)]	Cell cycle genes [30]	800	---	(47, 276, 529)
	YeastRACT CGI and Manually checked	---	---	

[1)] The size of a dataset with full information is expressed with three numbers, including the number of TFs, TGs, and either the number of regulatory relationships for the static network, or the number of genes for the microarray data; but the size of a microarray dataset is expressed as the number of ORFs. [2)] The reported and manually checked cell cycle related regulatory relationships are available for request.

2.2 Resolution of the GEDEM with the genetic algorithm (GA)

The differential equation (2) can be solved with numerical methods, like the Runge-Kutta procedure. However, the expression levels of genes are noisy; this makes it difficult to discern the true level of gene expression. The noise caused by systematic bias can be reduced by subtracting one sample value from another in the same experiment. Thus, it is more suitable to use the differential value of gene expression $y_i'(t)$. To facilitate the solution of this equation, we regard it simply as a fitting problem rather than as a different equation.

For the time course of gene expression, the least square error is

$$E_i=\sqrt{\sum_{q=1}^{Q}(y_{iq}'-\hat{y}_{iq}')^2} \quad (3)$$

Where q is the discrete sample index, Q denotes the total number of samples, y_{iq}' is the true value of $\mathrm{d}y_i/\mathrm{d}t$, while $\hat{y}_{iq}'$ is the estimated value for gene i according to the right side of Equation (2). Regarding Equation (3) as an objective function of a constrained optimization problem, we can use GA, a rapid method for global optimization, to get the optimized resolution.

Genetic algorithms are implemented as a computer simulations in which a population of abstract representations (called chromosomes or the genotype of the genome) of candidate solutions (called individuals, creatures, or phenotypes) to an optimization problem evolves toward better improved solutions. Traditionally, solutions are represented in binary as strings of 0 and 1, but other encodings are also possible. The evolution usually starts from a population of randomly generated individuals and occurs in generations. In each generation, the fitness of every individual in the population is evaluated, multiple individuals are stochastically selected from the current population (based on their fitness), and modified (recombined and possibly randomly mutated) to form a new population. The new population is then used in the next iteration of the algorithm. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population. If the algorithm has terminated due to a maximum number of generations, a satisfactory solution may or may not have been reached (See

Wikipedia address for Genetic Algorithm, http://en.wikipedia.org/wiki/Genetic_algorithm).

We used the GA function in the Matlab® toolbox for resolving the optimization problem. The using syntax is: [x, fval, exitflag]=ga (fitnessfcn, nvars, A, b, Aeq, beq, LB, UB). The input and output parameters were descripted as following.

fitnessfcn: input parameter, fitness function;

nvars: input parameter, number of design variables;

A: input parameter, A matrix for inequality constraints: $Ax \leqslant b$;

B: input parameter, b vector for inequality constraints: $Ax \leqslant b$;

Aeq: input parameter, A matrix for equality constraints: Ax=b;

Beq: input parameter, b vector for equality constraints: Ax=b;

LB: input parameter, Lower bound on x;

UB: input parameter, Upper bound on x;

x: output parameter, optimized resolution;

fval: output parameter, the value of the fitness function at x;

exitflag: output parameter, An integer identifying the reason the algorithm terminated.

Given a TG j, the solution variables to resolve include u_{ij}, v_{ij}, k_1, k_2 and b, which formed x variables (u_{i1}, $u_{i2}, \cdots, u_{im}$, v_{i1}, $v_{i2}, \cdots, v_{im}$, k_{i1}, k_{i2}, b_i). Within the context of gene regulatory networks in biology, the constraints are as follows: the parameter u_{ij} represents free variables, the value set of v_{ij} is [0, 1], and k_1, k_2, and b must be greater than zero. Thus, the inequality constraints are:

$$\begin{pmatrix} \cdots & -1 & & \cdots & & & & \\ \cdots & & -1 & \cdots & & & & \\ \cdots & & & \ddots & & & & \\ \cdots & & & \cdots & -1 & & & \\ \cdots & 1 & & \cdots & & & & \\ \cdots & & 1 & \cdots & & & & \\ \cdots & & & \ddots & & & & \\ \cdots & & & \cdots & 1 & & & \\ \cdots & & & \cdots & & -1 & & \\ \cdots & & & \cdots & & & -1 & \\ \cdots & & & \cdots & & & & -1 \end{pmatrix} \begin{pmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{im} \\ v_{i1} \\ v_{i2} \\ \vdots \\ v_{im} \\ k_{i1} \\ k_{i2} \\ b_i \end{pmatrix} \leqslant \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

In fact, the inequality constraints are all very simple. We can use the boundary parameters to describe the inequality constraints. Thus, the LB is ($-\infty$, $-\infty$, $\cdots$, $-\infty$, 0, 0, $\cdots$, 0, 0, 0, 0), and the UB is (∞, ∞, $\cdots$, ∞, 1, 1, $\cdots$, 1, ∞, ∞, ∞).

Actually, GA is an iterative process. Different from other iterative processes implements, the "ga" function in MATLAB does not require initial values for unknown variables. In fact, the function generates initial parameter values randomly.

For most TGs, the fitting processes were convergent; the error distribution is shown in Figure 4. The correlation between the approached and the true

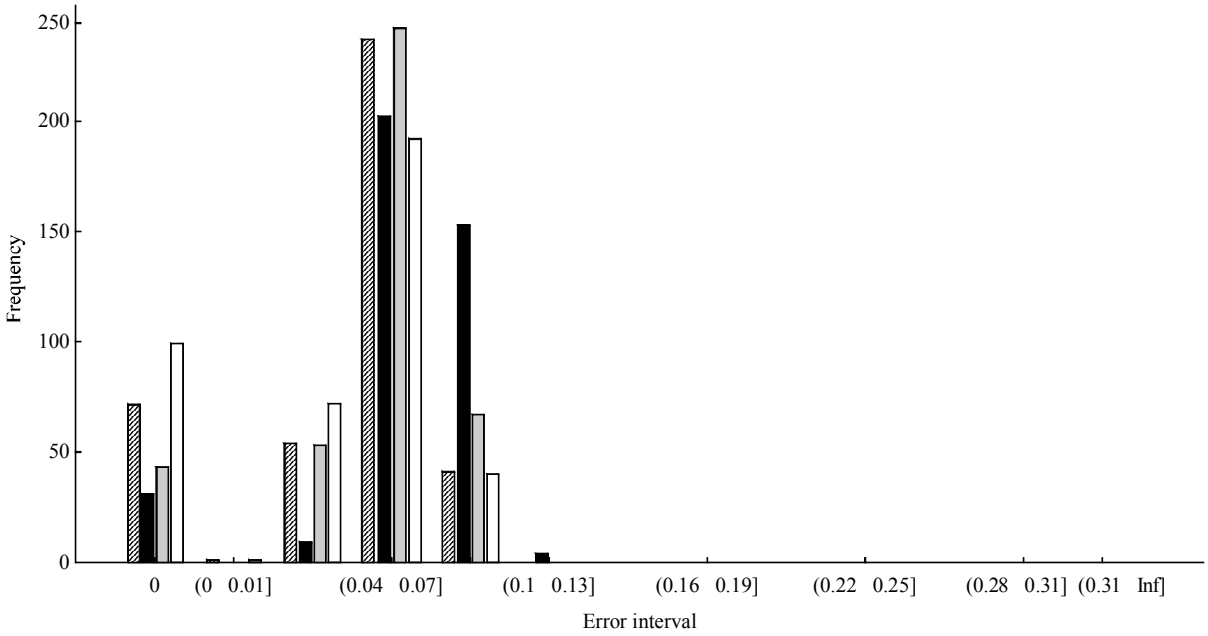

Fig. 4 The distribution of the errors in using the gene expression model of equation (3) to fit the data

The error value is the error in fitting the differences in normalized gene expression, which are relative errors. ▨: Alpha factor; ■: cdc15 mutation; ▩: cdc28 mutation; □: Elutriation.

rates of change in gene expression can also show the degree of fitting. We selected 35, 7, 42, and 33 TGs with greater than a 0.8 degree of fit from four microarray datasets (TGs are listed in Table 2 and the expression levels of the best-fitted genes are shown in Figure 5). The fitting results show that our method was able to use differential equations to describe the rate of change in gene expression for most TGs.

Table 2 The genes with a correlations larger than 0.8 between fitted and original observed changes in expression change ratio larger than 0.8 in the four experiments

Experiments	The lists of the genes	
Alpha factor	AGA1(*YNR044w*)(0.991 9)*	STE2(*YFL026w*)(0.842 3)
	SVS1(*YPL163c*)(0.951 8)*	SST2(*YLR452c*)(0.841)
	PRM5(*YIL117c*)(0.942 1)*	HHO1(*YPL127c*)(0.837 6)
	YLR463c(*YLR463c*)(0.939 8)*	YRF1-5(*YLR467w*)(0.830 4)
	KAR4(*YCL055w*)(0.939 4)*	STR3(*YGL184c*)(0.830 2)
	GPH1(*YPR160w*)(0.938 3)	PIR3(*YKL163w*)(0.828 4)
	SPI1(*YER150w*)(0.927 3)	YRF1-2(*YER190w*)(0.827 4)
	PHD1(*YKL043w*)(0.917 1)	TSL1(*YML100w*)(0.825)
	NIS1(*YNL078w*)(0.906 2)	CLN2(*YPL256c*)(0.823 8)
	CHS1(*YNL192w*)(0.882)	PUT3(*YKL015w*)(0.823 7)
	CIK1(*YMR198w*)(0.880 2)	SWI4(*YER111c*)(0.821 8)
	YBR071w(*YBR071w*)(0.874 4)	CLB1(*YGR108w*)(0.815 9)
	PST1(*YDR055w*)(0.869)	CLB2(*YPR119w*)(0.814 7)
	SWE1(*YJL187c*)(0.868 7)	FUS1(*YCL027w*)(0.808 5)
	HLR1(*YDR528w*)(0.864 5)	BUD4(*YJR092w*)(0.807 7)
	PIL1(*YGR086c*)(0.863 9)	YRF1-1(*YDR545w*)(0.805 5)
	YNL300w(*YNL300w*)(0.863 8)	HSP150(*YJL159w*)(0.800 7)
	PIR1(*YKL164c*)(0.852 7)	
CDC15 mutation	YNL046w(*YNL046w*)(0.978 8)*	KAR4(*YCL055w*)(0.908 3)*
	SCW10(*YMR305c*)(0.967 2)*	AGA1(*YNR044w*)(0.895 2)
	RNR1(*YER070w*)(0.963 1)*	YPS3(*YLR121c*)(0.860 3)
	DSF2(*YBR007c*)(0.9164)*	
CDC28 mutation	HXT10(*YFL011w*)(0.982 9)*	CLN1(*YMR199w*)(0.869 5)
	SWE1(*YJL187c*)(0.957 1)*	YJR054w(*YJR054w*)(0.862)
	YRO2(*YBR054w*)(0.954 5)*	PDS5(*YMR076c*)(0.858 4)
	SPI1(*YER150w*)(0.951 5)*	YPS3(*YLR121c*)(0.854 2)
	YOL014w(*YOL014w*)(0.943 1)*	RAD51(*YER095w*)(0.849 4)
	PCL1(*YNL289w*)(0.915 1)	CDC21(*YOR074c*)(0.848 7)
	HCM1(*YCR065w*)(0.913)	MCD4(*YKL165c*)(0.847)
	TOS2(*YGR221c*)(0.907 9)	LAP4(*YKL103c*)(0.841 4)
	YGR035c(*YGR035c*)(0.896 7)	DSE1(*YER124c*)(0.839 9)
	YHL026c(*YHL026c*)(0.896 4)	HST4(*YDR191w*)(0.838 7)
	CTF18(*YMR078c*)(0.892 9)	DUN1(*YDL101c*)(0.834 3)
	PRY1(*YJL079c*)(0.888 4)	MYO1(*YHR023w*)(0.833 6)
	AMN1(*YBR158w*)(0.887 2)	CLB1(*YGR108w*)(0.832 1)
	DBF2(*YGR092w*)(0.885 8)	CLB6(*YGR109c*)(0.816 5)
	MNN1(*YER001w*)(0.883 3)	CLB2(*YPR119w*)(0.815 7)
	MCM2(*YBL023c*)(0.882 5)	CDC6(*YJL194w*)(0.815)
	SUC2(*YIL162w*)(0.881 3)	WSC2(*YNL283c*)(0.811 2)
	NIS1(*YNL078w*)(0.872 2)	RNR1(*YER070w*)(0.81)
	CIN8(*YEL061c*)(0.871 2)	SPT21(*YMR179w*)(0.804 4)
	NDD1(*YOR372c*)(0.870 7)	YJL051w(*YJL051w*)(0.804 1)
	YIL141w(*YIL141w*)(0.870 7)	HSL1(*YKL101w*)(0.801 8)

Continued

Experiments	The lists of the genes	
Elutriation	SPI1(*YER150w*)(0.996)*	RNR3(*YIL066c*)(0.851)
	YHP1(*YDR451c*)(0.954 5)*	DSE2(*YHR143w*)(0.850 3)
	MET3(*YJR010w*)(0.913 5)*	NIS1(*YNL078w*)(0.850 3)
	PUT1(*YLR142w*)(0.911 4)*	YER189w(*YER189w*)(0.847 1)
	STP4(*YDL048c*)(0.907 6)*	ACE2(*YLR131c*)(0.846 5)
	PIR3(*YKL163w*)(0.904 5)	CTS1(*YLR286c*)(0.845 2)
	YPR013c(*YPR013c*)(0.898)	ICS2(*YBR157c*)(0.840 9)
	IQG1(*YPL242c*)(0.890 2)	ALD6(*YPL061w*)(0.840 8)
	HHO1(*YPL127c*)(0.888 5)	YHB1(*YGR234w*)(0.84)
	YRF1-5(*YLR467w*)(0.887 4)	RAD27(*YKL113c*)(0.836 3)
	PHO5(*YBR093c*)(0.880 9)	HXT2(*YMR011w*)(0.829 8)
	UGX2(*YDL169c*)(0.877 9)	YRF1-6(*YNL339c*)(0.829 1)
	SRL1(*YOR247w*)(0.866 3)	PST1(*YDR055w*)(0.829)
	PHO12(*YHR215w*)(0.864 3)	YPR045c(*YPR045c*)(0.824 1)
	MCH4(*YOL119c*)(0.864 2)	UGP1(*YKL035w*)(0.814 2)
	PHO11(*YAR071w*)(0.852 1)	PUT3(*YKL015w*)(0.805 6)
	PHO84(*YML123c*)(0.851 9)	

The data numbers in the brackets parentheses denote the correlation values of the TG. An asterisk at the end of the line means that the gene is plotted in Figure 5.

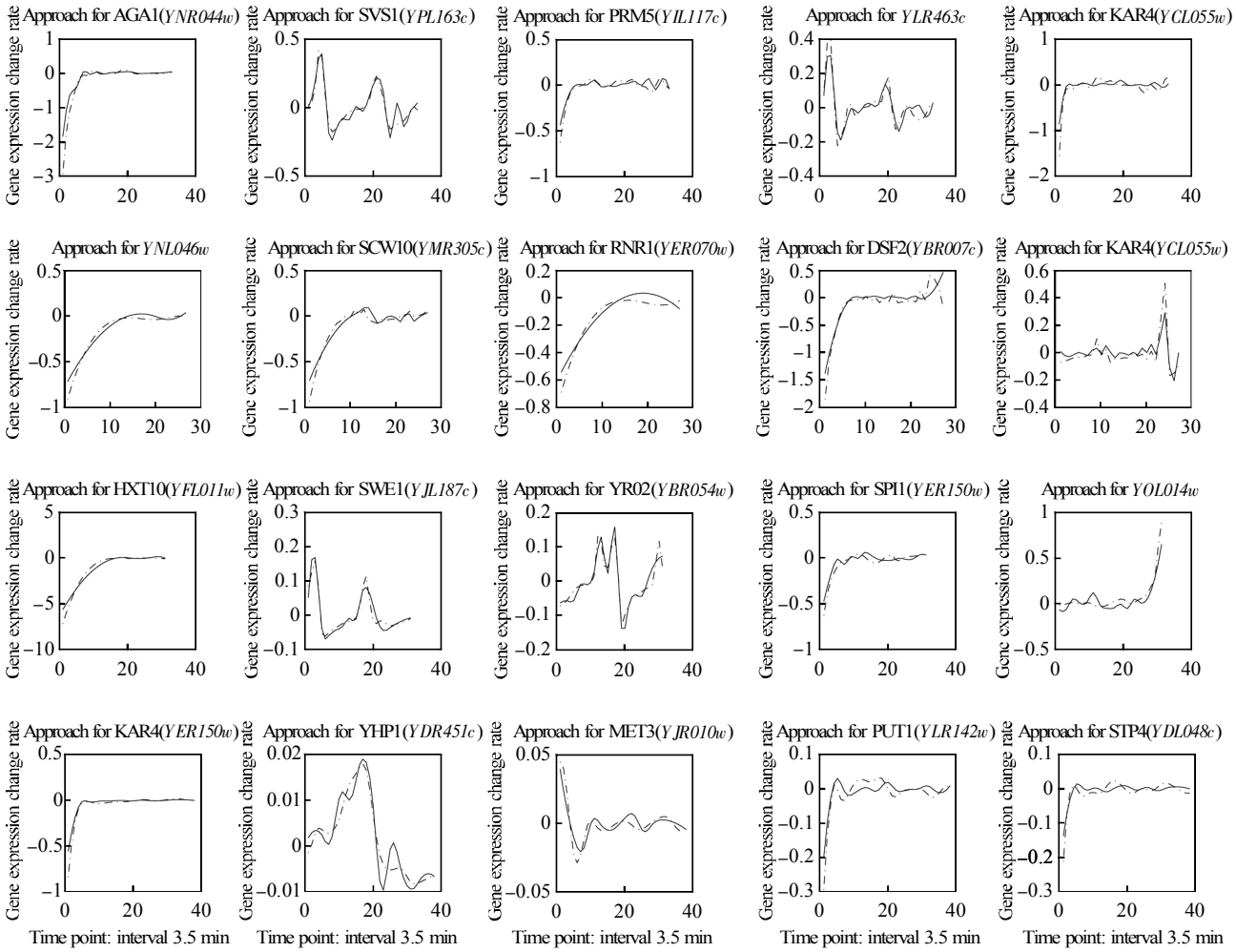

Fig. 5 The most highly correlated relationships between the true changes in gene expression and the calculated changes in gene expression for five TGs in each experiment

Each row represents one experiment: the alpha factor, CDC15 mutation, CDC28 mutation, and elutriation experiments are arranged from top to bottom, respectively. ····· : Fitted; — : Origin.

2.3 The regulatory sub-network in the yeast cell cycle

2.3.1 Comparison with the trace-back algorithm.

The TBA can be used to identify condition-specific dynamic regulatory sub-networks, as reported by Luscombe *et al*.[1] to compare our method with the TBA, the TBA and GEDEM were used to process the same dataset (Spellman′s gene expression data and the static regulatory network data from Luscombe *et al*).

The TBA traces the static regulatory network to identify whether a regulatory interaction occurred in the cell cycle process. Thus, the static regulatory network provides prior information to the TBA. Furthermore, additional prior information is provided by the previously identified cell cycle related genes[1]. The GEDEM also uses these two types of prior knowledge.

As shown in Table 3, the GEDEM recognized more condition-specific regulatory relationships than the TBA, though the actual number of true relationships was smaller in each of Spellman′s gene expression experiments. We compared the performance of the two methods with three commonly used measurements, and the performance values of the GEDEM were higher than those of the TBA.

Table 3 Performance comparison of the GEDEM with the TBA

Methods	Experiments	#iden.	#true	acc.%	sens.%	spec.%
GEDEM	Alpha factor	180	62	34.44	28.97	88.07
	CDC15 mutation	189	45	23.81	21.03	85.44
	CDC28 mutation	222	58	26.13	27.10	83.42
	Elutriation	133	29	21.80	13.55	89.48
	Total (union all 4)[1)]	541	134	24.77	62.62	58.85
TBA	Total (union all 4)	550	130	23.64	53.50	56.25

[1)] "Total (union all 4)" means that the dataset is the union set of all four experiment results. Because these four experiments are focus on the same condition-cell cycle, there are overlap between their results, which lead the number of identified regulatory associations (and true identified ones) in union set is not the sum of the size of four results and then affect the values of the performance measurements.

2.3.2 Condition-specific regulatory sub-network on our curated dataset.

Compared to the number of currently known regulatory interactions, the static regulatory network dataset of the TBA is quite small; only 215 regulatory interactions intersected with the cell cycle evaluation dataset. That is, the TBA dataset did not identify many TF-TG relationship pairs. Notably, our dataset contains 496 true cell cycle regulatory interactions.

Table 4 lists the results of the GEDEM analysis and TBA method on our larger dataset. We identified condition-specific regulatory relationships with higher

Table 4 The results from a GEDEM analysis and TBA method on our larger dataset

Methods	Experiments	#iden.	#true	acc.%	sens.%	spec.%
GEDEM	Alpha factor	472	83	17.58	30.63	90.66
	CDC15 mutation	588	68	11.56	25.09	87.52
	CDC28 mutation	589	84	14.26	31.00	87.88
	Elutriation	480	67	13.96	24.72	90.08
	Total (union all 4)[1)]	1757	227	12.92	83.76	63.27
TBA	Total (union all 4)	1716	169	8.97	62.36	62.85

[1)] "Total (union all 4)" is the same meaning with that in Table 3.

sensitivity and specificity (83.76% and 63.27%, respectively) than TBA, demonstrating better performance. However, the accuracy of identification on the larger dataset was worse than the accuracy on the smaller dataset. In fact, the accuracy was affected by the size of dataset; the larger the dataset, the harder it was to achieve good accuracy. The dataset for this analysis was four times larger than the former dataset.

2.3.3 Correlations between co-regulating TFs in a condition-specific regulatory sub-network and in a static regulatory network.

The correlations between co-regulating TFs found in our cell cycle specific sub-network were compared to those found in the static regulatory network(Figure 6 shows a comparison between all 4 experimental datasets and the static regulatory network). We found that the co-regulating TFs tended to be co-expressed. The mean correlations of co-regulating TFs in the cell cycle specific regulatory sub-network were closer to −1 or 1 than those of the static regulatory network. That is, each set of TFs identified in our sub-network tended to be either positively or negatively correlated with each other. In a biological context, the positively correlated TFs act as positive co-regulators, and negatively correlated TFs as negative co-regulators.

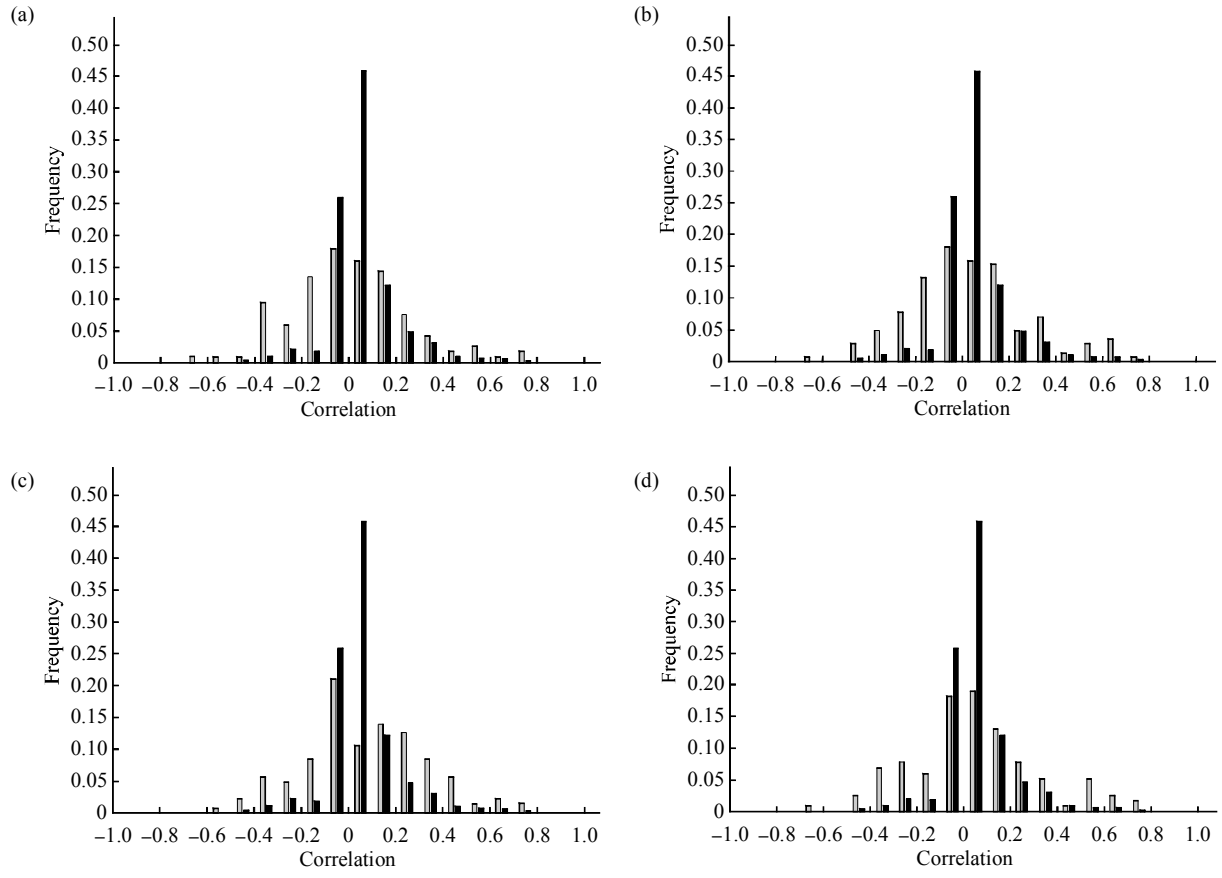

Fig. 6 The distributions of correlations between combinatorial TFs in the four experiments

The distribution found in the static regulatory network is compared to that found in our condition-specific regulatory sub-network. (a) Alpha. (b) cdc28. (c) cdc15. (d) Elutrition. □: Conditon-specific subnetwok; ■ : Static network.

We defined significant positive and negative correlations between TFs in the alpha experiment larger than 0.2 and smaller than −0.1, respectively (Table 5 ; there are similar results for other 3 experimental datasets; the result datasets for co-regulating TF sets identified are available for request).

Table 5 The co-regulating TF sets with significant mean correlations in the alpha experiment

Negatively correlated sets		Positively correlated sets	
Value	TF Set	Value	TF Set
–0.627 1	MAL33, MBP1	0.798 4	FKH1, FKH2
–0.543 1	HAP2, HAP5	0.774 4	CIN5, SOK2
–0.400 6	HAP4, STP1	0.658 5	ACE2, SWI5
–0.376 8	HAP4, MBP1	0.579 4	SWI5, YAP5
–0.370 8	HCM1, SWI5	0.557 1	FKH2, SWI5
–0.370 4	FKH1, HAP4	0.557	FKH2, YAP5
–0.359 4	ACE2, PDR3	0.496 2	FKH1, RAP1
–0.357 9	CAD1, SWI4	0.439 9	INO4, STP2
–0.350 4	MSN4, YAP1	0.394 7	GAT3, PDR1, YAP5
–0.342 6	ASH1, RPN4	0.378 2	ARR1, FKH1, FKH2
–0.334 3	DOT6, PHD1	0.354 5	FKH1, FKH2, GAT4
–0.332 2	MBP1, RFX1	0.314 8	ECM22, FKH1, INO4, RAP1
–0.319 7	SKN7, SWI4	0.310 6	MCM1, SOK2
–0.306 6	STE12, YOX1	0.295	ADR1, AFT2, NRG1, PDR1
–0.290 5	MET32, SKN7	0.284 5	GCR1, SWI5, SWI6
–0.249 9	MET31, RAP1	0.279 1	LEU3, STE12
–0.231 1	PHD1, STP2	0.278 5	ARR1, HAA1
–0.220 2	CIN5, REB1	0.260 3	ABF1, FKH1
–0.214 5	GCN4, MSN4	0.257 3	MGA2, SPT23
–0.205 9	MBP1, SWI4	0.245 7	ABF1, DAL82, STP2
–0.205 1	ROX1, SOK2	0.233 6	FKH2, YRR1
–0.192 1	SWI4, SWI6	0.208 7	ASH1, SWI4
–0.183 3	ASH1, STP2, SWI4		
–0.169 6	FKH2, STE12		
–0.166	HCM1, MBP1		
–0.157 1	SOK2, YHP1		
–0.136 8	INO2, MCM1, RAP1		
–0.130 5	MSN4, SOK2, SWI5		
–0.129 5	SMP1, SWI5		
–0.123 2	PHD1, RME1		
–0.125 4	HAP4, XBP1, YPR015c, YOX1		
–0.123 6	AFT1, AFT2, SOK2, ZAP1		
–0.113 2	STP2, TEC1		
–0.110 7	MET31, MSN2, RTG3, YAP6		
–0.108 1	HAP1, MOT3, STE12, XBP1		
–0.106 2	LEU3, SOK2, STB5		
–0.101 2	MCM1, SWI4, SWI5		

2.3.4 Regulatory complexes identified by the GEDEM.

In the cell cycle, there are several complexes that comprise a set of TFs that act as regulators. Some examples are SBF (SWI4 + SWI6), MBF (MBP1 + SWI6), FKH1/FKH2, and SWI5/ACE2. With the

GEDEM analysis, we identified SBF, FKH1/FKH2, and SWI5/ACE2, as shown in Table 5.

We notice that the correlation between SWI4 and SWI6 was negative. That maybe caused by several reasons. First, it is not clear that the elements of complex should be positive correlated with each other. Some elements take part in other processes, which maybe make their relationships more complex, for example, SWI4 and SWI6 both combined with other TFs. So, it is not available to identify the combined regulatory interactions only basing on the correlations of gene expression dataset. Second, SBF is expressed in the G1/S phase of the cell cycle [36], and the correlation was calculated from the gene expression dataset which contains 3 cycle cell cycle data [30]. Though the positive correlations may exist, they could be inundated by the other information. In addition, the gene expression dataset is noisy. The correlations between SWI4 and SWI6 in three cell cycle datasets are not coherent, the values are –0.175 8, –0.315 4 and 0.211 8, respectively.

In all, we identified 116 sets of TFs in the alpha factor experiment. Besides the above mentions, the other sets of co-regulating TFs may also tend to form complexes; this is currently an intense research topic.

3 Discussion

Here we describe the GEDEM, a novel method for recognizing dynamic regulatory sub-networks. Based on a TBA analysis of the same data, the GEDEM exhibits superior performance. The GEDEM explores the many-to-one control mechanisms of the regulatory network that occur in specific conditions. We can discern the weight of each TF that contributed to changes in expression of each TG; this can be used to investigate the mechanism of combinatorial regulation in future studies.

3.1 Combinatorial regulation is common in yeast

It is well known that combinatorial regulation is a common mechanism in yeast. In the evaluation dataset that contained the true regulatory network, we found 152 TGs with only one TF regulator, but 209 TGs had multiple TF regulators; thus, we confirmed that combinatorial regulation is common in the yeast cell cycle.

3.2 Capability and limits in extending GEDEM use

The GEDEM can be used to analyze many types of biological processes and conditions. Although there are many steady-state conditions in a cell, the GEDEM works best for conditions that give rise to multiple sequential changes. However, the GEDEM is limited by several factors. First, a basic assumption of the model is that the mRNA expression level of regulatory genes reflects the expression level of the corresponding proteins. Studies in many species, particularly mammalian, have shown that this relationship is not true for every protein. Additionally, there is often a time delay in regulation because the level of TF does not immediately lead to an expression change of target gene. This delay is ignored by the GEDEM.

3.3 Dataset requirements for the GEDEM

In equation (2), there is a “∑” symbol that describes the combined expression of multiple TFs. If the number of TFs for a TG is very large, the time course of their expression is linearly dependent; that is, some TFs can be represented by linear adjustments in other TFs. In that case, the model has several possible solutions. To avoid this scenario, the number of TFs for one TG must be less than the rank of the expression level matrix for these TFs. Fortunately, in the alpha factor experiment, the rank for the expression of all TFs is 18 in the static regulatory network. Thus, we can depict the distribution of in-degree TGs in the static regulatory network (Figure 7). Therefore, we conclude that the GEDEM is able to identify the true regulators of most TGs when there are sufficient samples in the microarray dataset.

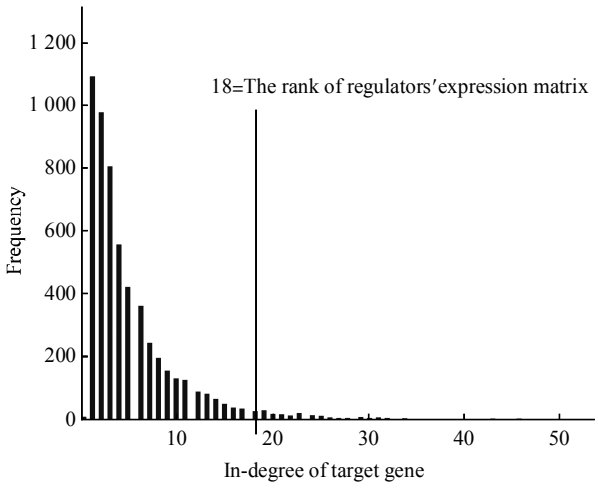

Fig. 7 The distribution of in-degree target genes for the static regulatory network

The line indicates the rank of the TF expression level matrix for the alpha factor experiment dataset. Most of the TGs have fewer than 18 regulators.

4 Conclusions

We have used a differential equation model to describe the time-course expression levels of TGs controlled by multiple TFs by treating each TG as an information processing unit. We call the model the GEDEM. In the GEDEM, there is an indicator variable that indicates whether a TF regulates the corresponding TG in the experimental condition. We solved the model with the GA, which showed convergence in the processing of most genes.

Based on analyses of the same datasets, the GEDEM performed better than the TBA. For larger datasets that we curated, the GEDEM exhibited high sensitivity and specificity. The correlations between co-regulating TFs in our condition-specific regulatory sub-network showed higher significance than those for the static regulatory network; thus, our condition-specific networks are closer to reality. Furthermore, identifying regulatory complexes in the cell cycle also indicated the ability of the GEDEM to discern condition-specific sub-networks from static networks.

References

[1] Luscombe N M, Babu M M, Yu H, *et al*. Genomic analysis of regulatory network dynamics reveals large topological changes. Nature, 2004, **431**(7006): 308−312

[2] Balazsi G, Barabasi A L, Oltvai Z N. Topological units of environmental signal processing in the transcriptional regulatory network of *Escherichia coli*. Proc of Nat Acad Sci, 2005, **102**(22): 7841−7846

[3] Resendis-Antonio O, Freyre-González J A, Menchaca-Méndez R, *et al*. Modular analysis of the transcriptional regulatory network of *E. coli*. TRENDS in Genetics, 2005, **21**(1): 16 − 20

[4] Bar-Joseph Z, Gerber G K, Lee T I, *et al*. Computational discovery of gene modules and regulatory networks. Nature Biotechnology, 2003, **21**(11): 1337−1342

[5] Tanay A, Sharan R, Kupiec M, *et al*. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. Proc of Nat Acad Sci, 2004, **101**(9): 2981−2986

[6] Wu W S, Li W H, Chen B S. Computational reconstruction of transcriptional regulatory modules of the yeast cell cycle. BMC Bioinformatics, 2006, **7**(1): 421−435

[7] Lee T I, Rinaldi N J, Robert F, *et al*. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science, 2002, **298**(5594): 799−804

[8] Milo R, Shen-Orr S, Itzkovitz S, *et al*. Network motifs: simple building blocks of complex networks. Science, 2002, **298**(5594): 824−827

[9] Harbison C T, Gordon D B, Lee T I, *et al*. Transcriptional regulatory code of a eukaryotic genome. Nature, 2004, **431**(7004): 99−104

[10] Lee H G, Lee H S, Jeon S H, *et al*. High-resolution analysis of condition-specific regulatory modules in *Saccharomyces cerevisiae*. Genome Biology, 2008, **9**(1): R2−R22

[11] Liao J C, Boscolo R, Yang Y L, *et al*. Network component analysis: Reconstruction of regulatory signals in biological systems. Proc Natl Acad Sci USA, 2003, **100**(26): 15522−15527

[12] Boorsma A, Lu X J, Zakrzewska A, *et al*. Inferring condition-specific modulation of transcription factor activity in yeast through regulon-based analysis of genomewide expression. PLoS ONE, 2008, **3**(9): e3112−e3121

[13] Ernst J, Vainas O, Harbison C T, *et al*. Reconstructing dynamic regulatory maps. Mol Syst Biol, 2007, **3**(3): 74−86

[14] Shi Y, Mitchell T, Bar-Joseph Z. Inferring pairwise regulatory relationships from multiple time series datasets. Bioinformatics, 2007, **23**(6): 755−763

[15] Kim J, Bates D G, Postlethwaite I, *et al*. Linear time-varying models can reveal non-linear interactions of biomolecular regulatory networks using multiple time-series data. Bioinformatics, 2008, **24**(10): 1286−1292

[16] Kim H, Hu W, Kluger Y. Unraveling condition specific gene transcriptional regulatory networks in *Saccharomyces cerevisiae*. BMC Bioinformatics, 2006, **7**(1): 165−181

[17] Li J J, Ren J, Liang Z, *et al*. Universal core modules govern the dynamics of cellular regulatory networks. arXiv preprint q-bioMN/0601012, 2006, http://arxiv.org/abs/q-bio/0601012

[18] Ihmels J, Friedlander G, Bergmann S, *et al*. Revealing modular organization in the yeast transcriptional network. Nat Genet, 2002, **31**(4): 370−377

[19] Friedman N, Linial M, Nachman I, *et al*. Using bayesian networks to analyze expression data. J Comput Biol, 2000, **7**(3−4): 601−620

[20] Klipp E. Systems Biology in Practice: Concepts, Implementation and Application. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA, 2005: 99−107

[21] Shmulevich I, Dougherty E R, Zhang W. Gene perturbation and intervention in probabilistic Boolean networks. Bioinformatics, 2002, **18**(10): 1319−1331

[22] Kauffman S, Peterson C, Samuelsson B, *et al*. Random boolean network models and the yeast transcriptional network. Proc Natl Acad Sci USA, 2003, **100**(25): 14796−14799

[23] Kim H, Lee J K, Park T. Boolean networks using the chi-square test for inferring large-scale gene regulatory networks. BMC Bioinformatics, 2007, **8**(2): 37−51

[24] Friedman N, Linial M, Nachman I, *et al*. Using bayesian networks to analyze expression data. J Comput Biol, 2000, **7**(3−4): 601−620

[25] Vu T T, Vohradsky J. Nonlinear differential equation model for quantification of transcriptional regulation applied to microarray data of *Saccharomyces cerevisiae*. Nucl Acids Res, 2007, **35**(1): 279−287

[26] Chen K C, Wang T Y, Tseng H H, *et al*. A stochastic differential equation model for quantifying transcriptional regulatory network in *Saccharomyces cerevisiae*. Bioinformatics, 2005, **21**(12): 2883−

2890

[27] Dewey T G, Galas D J. Dynamic models of gene expression and classification. Functional & Integrative Genomics, 2001, **1**(4): 269−278

[28] Ji X, Li Ling J, Sun Z. Mining gene expression data using a novel approach based on hidden Markov models. FEBS Letters, 2003, **542**(1−3): 125−131

[29] Vohradsky J. Neural network model of gene expression. The FASEB J, 2001, **15**(3): 846−854

[30] Spellman P, Sherlock G, Zhang M, *et al*. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Mol Biol Cell, 1998, **9**(12): 3273−3297

[31] Cho R J, Campbell M J, Winzeler E A, *et al*. A genome-wide transcriptional analysis of the mitotic cell cycle. Molecular Cell, 1998, **2**(1): 65−73

[32] Teixeira M C, Monteiro P, Jain P, *et al*. The YEASTRACT database: a tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae*. Nucl Acids Res, 2006, **34**(suppl_1): D446− D451

[33] Singer R H, Penman S. Messenger RNA in HeLa cells: kinetics of formation and decay. J Mol Biol, 1973, **78**(2): 321−334

[34] D′Haeseleer P, Wen X, Fuhrman S, *et al*. Linear modeling of mRNA expression levels during CNS development and injury. Pacific Symposium on Biocomputing, 1999, **4**(1): 41−52

[35] Teixeira M C, Monteiro P, Jain P, *et al*. The YEASTRACT database: a tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae*. Nucl Acids Res, 2006, **34**(suppl_1): D446−D451

[36] Simon I, Barnett J, Hannett N, *et al*. Serial regulation of transcriptional regulators in the yeast cell cycle. Cell, 2001, **106**(6): 697−708

一种基于基因表达模型识别酵母细胞周期条件特异调控子网的方法 *

刘齐军[1,2] 王正华[1] 刘万霖[2] 李 栋[2] 贺福初[2,3]** 朱云平[2]**

([1] 国防科学技术大学并行与分布处理国家重点实验室, 长沙 410073；
[2] 蛋白质组学国家重点实验室，北京蛋白质组研究中心，军事医学科学院放射与辐射医学研究所, 北京 102206；
[3] 复旦大学生物医学科学研究所，上海 200032)

摘要 由高通量微阵列技术产生的数据集可以用于解释生物系统基因调控的未知机制．生物过程是动态的，所以很有必要关注某些条件下特异的基因调控子网络．细胞周期是一个基本的细胞过程，识别酵母的细胞周期特异调控子网是理解细胞周期过程的基础，并且有助于揭示其他细胞条件的基因调控机理．使用一个基因表达微分方程模型(GEDEM)，从静态网络中识别了动态的细胞周期相关调控关系．与已经报道的细胞周期相关调控相互作用相比，该方法识别了更多的真实存在的条件特异调控关系，取得了比当前的方法更好的性能．在大数据集上，GEDEM 识别了具有高敏感性和特异性的调控子网．组合调控的深入分析显示，条件特异调控子网的转录因子之间的相关性呈现出比静态网络中转录因子相关性更强，这说明条件特异网络比静态网络更加接近真实情况．另外，GEDEM 方法还识别更多潜在的共调控转录因子．

关键词 基因调控网络，细胞周期，基因表达模型，条件特异子网，微分方程模型

学科分类号 Q756，Q811.4 **DOI:** 10.3724/SP.J.1206.2009.00581

* 蛋白质组学国家重点实验室开放课题基金(SKLP-O200811), 国家自然科学基金(60603054, 30800200, 30621063), 国家重点基础研究发展计划(973)项目 (2006CB910706, 2006CB910803, 2010CB912700), 国家高技术研究发展计划(863)项目(2006AA02A312)和国家科技重大专项项目(2008ZX10002-016, 2009ZX09301-002)资助.

** 通讯联系人.

贺福初. Tel: 010-68171208, Fax: 010-68214653, E-mail: hefc@nic.bmi.ac.cn

朱云平. Tel: 010-80705225, Fax: 010-80705155, E-mail: zhuyunping@gmail.com

收稿日期：2009-12-15，接受日期：2010-01-21