Published online 14 December 2006 *Nucleic Acids Research, 2007, Vol. 35, No. 1* ***279–287***
doi:10.1093/nar/gkl1001

Nonlinear differential equation model for quantification of transcriptional regulation applied to microarray data of *Saccharomyces cerevisiae*

Tra Thi Vu and Jiri Vohradsky*

Laboratory of Bioinformatics, Institute of Microbiology, ASCR, Videnska 1083, 14220 Prague, Czech Republic

Received June 19, 2006; Revised October 12, 2006; Accepted October 27, 2006

ABSTRACT

Microarray studies are capable of providing data for temporal gene expression patterns of thousands of genes simultaneously, comprising rich but cryptic information about transcriptional control. However available methods are still not adequate in extraction of useful information about transcriptional regulation from these data. This study presents a dynamic model of gene expression which allows for identification of transcriptional regulators using time series of gene expression. The algorithm was applied for identification of transcriptional regulators controlling 40 cell cycle regulated genes of *Saccharomyces cerevisiae*. The presented algorithm uses a dynamic model of time continuous gene expression with the assumption that the target gene expression profile results from the action of the upstream regulator. The goal is to apply the model to putative regulators to estimate the transcription pattern of a target gene using a least squares minimization procedure. The procedure iteratively tests all possible transcription factors and selects those that best approximate the target gene expression profile. Results were compared with independently published data and good agreement between the published and identified transcriptional regulators was found.

INTRODUCTION

Regulation of gene expression is one of the most important processes in the living cell, transmitting static information encoded in the DNA sequence into functional protein molecules, which consequently control most of the cellular processes. The regulation of gene expression depends on the recognition of specific promoter sequences by transcriptional regulatory proteins which allows binding of RNA polymerase and initiates transcription. The transcriptional programs are modified as cell progresses through development or through a reaction to changing environmental conditions.

Developments in microarray technology have permitted the recording of changes in gene expression over time during the cell cycle or other developmental processes. As the regulation of transcription is a dynamic process, analysis of the time series of changes in RNA amounts during the cell cycle can lead to the discovery of causal relations between genes and their regulators. Since, the gene expression data are the result of network interactions between regulators and target genes, it is reasonable to trace the interaction networks from microarray gene expression data. As mRNA levels are the result of the action of such networks, it should be possible to reverse engineer the network architecture from the microarray data.

Cell cycle control has been intensively studied in the budding yeast *Saccharomyces cerevisiae* and large transcriptomic databases of the alteration of RNA synthesis during the cell cycle have been created. Genome-wide microarray gene expression data relevant to the yeast cell cycle have been collected in a parallel manner (1,2). The data were analyzed using a variety of clustering methods (3,4) for the identification of cell cycle, or co-ordinately controlled genes. A singular value decomposition (5–7) was used to model the gene expression data. Instead of grouping genes, according to the similarity of their gene expression patterns, transcriptional regulatory networks were found by genome-wide location analysis identifying which transcriptional regulators bind to which promoters (8,9). Potential transcriptional regulatory networks were identified independently in the work of Lee *et al*. (10).

Several procedures for inference of transcriptional regulatory networks from experimental microarray data were published in recent years (11–13). The methods identified upstream regulatory genes by modeling the regulatory process using differential equation models of gene expression control. The goal was to fit a generalized linear model using a set of putative regulators to estimate the transcription pattern of a specific target gene. Alternatively, Woolf and Wang (14) used fuzzy logic for the prediction of transcriptional regulators.

*To whom correspondence should be addressed. Tel: +420 241062513; Fax: +420 241722257; Email: vohr@biomed.cas.cz

Nachman *et al*. (15) used a kinetic model in connection with dynamic Bayesian networks to infer the structure of transcriptional regulatory networks and the regulators from gene expression time series. One potentially useful approach, introduced by Bar-Joseph (16), combines genomic information with gene expression data analysis. This approach was extended recently by Wang *et al*. and Makita *et al*. who combined the analysis of gene expression data with promoter sequence analysis (17,18), or the sigma factor binding sequence motif (18). Such a combined approach has the advantage of the introduction of additional information independent of the gene expression data.

In this paper, we present an alternative method for the prediction of target gene regulators based on a nonlinear differential equation model of gene expression (19). In the beginning, a set of all potential regulators is selected. For a selected set of target genes, the procedure iteratively picks individual genes from the pool of possible regulators and applies the model to fit the gene expression profile of the target gene using the expression profile of the candidate regulatory gene. This procedure is repeated for all target genes and all possible regulators. Those regulators that are able to model the target gene expression profile correctly are considered to be the true regulators.

The procedure was applied to 40 target genes of *S.cerevisiae* transcriptomic data (2,11). A pool of 184 candidates for potential regulators was selected by combination from the previous reports of Lee (10), Chen (12), Chen (11), and the database of yeast transcriptional regulators YEASTRACT (http://www.yeastract.com). The data were also analyzed using more common linear model. The results of the presented algorithm were compared with the data computed using the linear model, a generalized linear model published recently by Chen (11), all the results were verified by comparison with independent experimental data collected in YEASTRACT. Results show that the method is capable of correctly identifying regulators and predicting their function as activators or repressors.

RESULTS

Dynamic model of transcriptional control

The presented model results from our previously published work (19–21) on the dynamic simulation of genetic networks. The model is derived by assuming the recursive action of regulators on the target over time. The model assumes that the regulatory effect on the expression of a particular gene can be expressed as a combinatorial action of its regulators. The target gene expression level z at time t + dt can be derived from the expression levels of regulators (y_j) at time t and the regulatory weights (w_j) of all genes controlling the target gene. Thus g, a regulatory effect for a particular gene, is

$$g \approx \sum_{j} w_j y_j - b, \tag{1}$$

for $j = 1, 2, \ldots, m$, where m is the number of regulators controlling the gene. The parameter b represents transcription initiation delay or an unspecific bias caused by regulatory effects associated with gene expression but independent of the particular regulator. Let the rate of expression of a target gene (dz/dt) be given by the regulatory effects of other genes ρ and the effect of degradation x. The degradation effect is modeled by the kinetic equation of a first order chemical reaction $x = k.z$. The function ρ represents the regulatory effect g (Equation 1) of regulators $j = 1, 2, \ldots, m$, transformed by a sigmoidal transfer function

$$\rho = \frac{1}{1 + \exp\left(-\sum_{j=1..m} w_j y_j + b\right)}, \tag{2}$$

for m regulators. The whole model for the control of target gene expression z has the form:

$$\frac{\mathrm{d}z}{\mathrm{d}t} = k_1 \frac{1}{1 + \exp\left(-\sum_{j=1..m} w_j y_j + b\right)} - k_2 z. \tag{3}$$

The constant k_2 represents the rate constant of degradation of the target gene product, and k_1 is its maximal rate of expression. Here, we consider the case of only one transcriptional factor. In such a case Equation 3 simplifies to

$$\frac{\mathrm{d}z}{\mathrm{d}t} = \frac{k_1}{1 + \exp\left(-wy + b\right)} - k_2 z, \tag{4}$$

where y is approximated with a polynomial of degree n

$$y \approx a_0 + a_1 t + a_2 t^2 + \ldots + a_n t^n. \tag{5}$$

Coefficients $\{a_0, \ldots, a_n\}$ are computed from the experimental gene expression profile using a least squares minimization procedure. The polynomial fit is used as an approximation of an underlying 'true' expression profile, which is obscured by experimental errors. It is assumed that the weight of the experimental error is the same for all points of the measurement. In principle other approximations can be used [the topic of modeling time series measurements in microarray experiments was discussed in the paper of Bar-Joseph (22)].

This simplified version of the model (Equation 4) was used for the identification of the target gene regulators throughout this paper. The degree n of the polynomial must be chosen so that it reflects the rate of changes in gene expression in the given experiment and must be chosen for each experiment individually. In the case of the yeast cell cycle analyzed here, the degree $n = 6$.

Using expression profiles $\mathbf{Z}$ $\{z(t_\tau)\}$ of the target and $\mathbf{Y}$ $\{y(t_\tau)\}$ of the regulator genes measured at time points $t = t_\tau$, $\tau = 1,2,\ldots,Q$, we search for all the gene profiles $\mathrm{Y} \in \{\mathbf{Y}_i, i = 1,2,\ldots,m\}$ (the pool of all m potential regulators) that minimize the mean square error function:

$$E = \frac{1}{Q} \sum_{\tau=1}^{Q} \left[z(t_\tau) - z^c(t_\tau)\right]^2, \tag{6}$$

where $\{z^c(t_\tau)\}$ denotes the reconstructed profile of $z(t) \in \mathbf{Z}$ at time points $t = t_\tau$, $\tau = 1,2,\ldots,Q$ for Q data points, computed using the model (Equation 4).

The problem now becomes an optimization problem, where the expression profiles of Z and Y are supplied to estimate parameters w, b, k_1, k_2 of the model (Equation 4) that minimize the error function (Equation 6).

For comparison we applied to the same data more commonly used linear model of the form

$$\frac{\mathrm{d}z}{\mathrm{d}t} = \mathrm{d}_0 + \mathrm{d}_1 y - \mathrm{d}_2 z, \qquad \mathbf{7}$$

and computed the parameters d_i (i = 1 . . 2) by minimizing error function 6.

Computational algorithm

The method aims to select a set of potential regulators of a particular target gene by estimating the expression profile of the target gene. The method searches for possible regulators from a pool of transcriptional regulators using least squares minimization and the model Equation 4, minimizing the error function, Equation 6. The problem of missing data points and experimental fluctuation in the gene expression profiles is bypassed by approximating the regulator gene profiles by a polynomial of degree n. The degree is chosen according to the number of data points in the profile and the level of fluctuations, individually for each experiment. Differential Equation 4 is solved numerically and the parameters w, b, k_1 and k_2 are optimized in a least squares minimization loop until the desired precision or a predefined number of iterations is reached.

The overall algorithm is described as follows:

(i) Fit regulator gene profiles with a polynomial of degree n (Equation 5).
(ii) Select a target gene.
(iii) Select a candidate regulatory gene from the pool of possible regulators.
(iv) Apply least squares minimization procedure to the target and regulator genes using Equation 4 with error function 6.
(v) Go to step 3 and repeat for all possible regulators.
(vi) Select regulators that best satisfy the selection criterion.
(vii) Go to step 2 and repeat for all target genes.

The procedure was repeated 100 times for each combination of regulator-target with randomly selected initial parameter values at the beginning of the optimization procedure. The parameter set giving the smallest value of the error function was then selected.

The optimization was performed using standard Levenberg–Marquardt procedure, Equation 4 was solved numerically using Runge–Kutta procedure (ode45 function in MATLAB). All computations were performed in the MATLAB environment.

For comparison of the results we identified potential regulators using the linear model (Equation 7) and the scheme given by points 1–7, with the linear model replacing Equation 4 in the step 4.

Dataset selection

To evaluate the performance of our model, we chose the eukaryotic cell cycle dataset published by Spellman *et al.* (2). This dataset records changes in gene expression measured as amounts of mRNA using microarrays at 18 timepoints over two cell cycle periods. The chip contained 6178 open reading frames. Using multivariate methods, Spellman identified 800 genes whose expression was associated with the cell cycle. Nevertheless, the real number of regulators controlling the cell cycle is much smaller. For the identification of yeast cell cycle regulators, we selected a pool of 184 potential regulator genes by combining data from previously published papers (10–12) and the YEASTRACT database. In order to enable comparison with previously published data, we chose 40 target genes, the same as those analyzed in the paper of Chen *et al.* (11).

Inference of regulators

The procedure was applied to 40 yeast cell cycle regulated target genes and 184 potential regulators. The data were in the form of a log base 2 of the ratio between the actual value of the mRNA amount divided by the value of a standard which was the same for all time points. Therefore, after exponentiation the whole time series was just scaled by the value of the standard. Before application of the algorithm the data were exponentiated to a power of 2. The least squares minimization procedure was applied to each target gene for all potential regulators.

It can be assumed that a least squares best fit of the polynomial of degree n to the target gene expression profile z^p,

$$z^p \approx c_0 + c_1 t + c_2 t^2 + \ldots + c_n t^n, \qquad \mathbf{8}$$

is an approximation of the unknown real profile, which is obscured by the experimental noise, inaccuracies and natural biochemical and physiological fluctuations [any other fitting function can be used. Choosing a different function will only change the approximation and consequently the selection of the threshold values (see below), but not the principle]. If repeated measurements of expression values are available, estimation of the overall error by means of a polynomial fit (Equation 8) can be replaced by a statistical model. We chose the polynomial representation as it has widely been accepted as an approximation of unknown functions, and as the data used here were available only as averages. Its deviation from the experimental data are given by

$$E_1 = \frac{1}{Q} \sum_{\tau=1}^{Q} \left[z(t_\tau) - z^p(t_\tau)\right]^2. \qquad \mathbf{9}$$

Finding the most probable regulator for the given target means finding a regulator profile, which models best the target profile using the model (Equation 4) and minimizing E (Equation 6). It can be assumed that the fit of the model to the target regulator profile is at least as good as the fit given by Equation 8, i.e. the deviation E (Equation 6) must be less than or equal to the deviation E_1 (Equation 9). Therefore we selected those regulators for which E was less than or equal to E_1. In addition, if we look at the plot of values of E for all potential regulators of a given target sorted in increasing order of E (Supplementary Figure 1), we can see that, in most cases, several first potential regulators have E noticeably smaller than the rest (an edge can be seen on the bar graph). This means that those regulators fit the target gene profile even better than the others (we call them 'best regulators'). A summary of 'correct identification' of the regulators for all targets is given in Table 1. Correct identification means that a gene identified as a regulator for a given target was also

Table 1. Summary of identification of regulators for 40 selected yeast cell cycle regulated genes

Id	Target		best	$E \leqslant E_1$	$E \leqslant 1.1 * E_1$	$E \leqslant 1.2 * E_1$	Min(m)	Min(m) lin	E	
			m	m	m	m			Nonlin	Lin
1	YER150W	SPI1	3	4	5	8	4	2	0.0253	0.8339
2	YOR323C	PRO2	1	8	75	182	7	35	0.0010	0.0236
3	YKL177W	NA	0	0	0	5	7	3	0.0006	0.0277
4	YMR288W	HSH155	2	10	11	26	10	12	0.0019	0.0588
5	YMR316W	DIA1	4	15	29	40	21	1	0.0052	1.0992
6	YPL223C	GRE1	0	0	0	1	5	6	0.0017	0.0373
7	YPR035W	GLN1	2	2	2	10	2	6	0.0021	0.2907
8	YER003C	PMI40	1	2	3	4	1	11	0.0017	0.2779
9	YJL155C	FBP26	2	16	157	180	10	4	0.0003	0.0892
10	YMR145C	NDE1	0	0	3	10	4	16	0.0010	0.1342
11	YBR089W	NA	2	4	4	5	4	13	0.0577	1.4703
12	YDR285W	ZIP1	2	6	45	76	4	1	0.0274	1.8964
13	YFR057W	NA	0	0	13	46	8	4	0.0039	0.1206
14	YAL018C	NA	5	18	68	148	5	22	0.0003	0.1219
15	YOR383C	FIT3	2	2	2	6	2	14	0.0219	1.4964
16	YOR319W	HSH49	12	18	31	44	12	32	0.0801	4.7275
17	YOR264W	DSE3	7	7	16	20	7	7	0.0097	1.1955
18	YOL116W	MSN1	4	6	32	84	4	4	0.0045	0.1843
19	YGR269W	NA	0	0	1	5	2	1	0.0108	0.0778
20	YKL001C	MET14	4	13	23	27	3	1	0.0019	0.1988
21	YDR146C	SWI5	0	0	0	1	4	12	0.0096	0.5309
22	YPL256C	CLN2	1	6	12	18	1	5	0.0253	1.2436
23	YJL187C	SWE1	1	2	3	6	1	4	0.0072	0.2139
24	YOR372C	NDD1	1	2	3	4	8	17	0.0062	0.1479
25	YLR274W	CDC46	2	7	5	6	7	7	0.0303	0.6388
26	YHR152W	SPO12	2	3	5	7	3	12	0.0012	0.3448
27	YCR065W	HCM1	2	6	6	8	6	16	0.0037	0.7056
28	YAL040C	CLN3	2	4	15	19	21	14	0.0105	0.7826
29	YDR224C	HTB1	1	3	3	3	3	2	0.0218	0.7135
30	YGL116W	CDC20	2	10	10	11	10	17	0.0050	0.5054
31	YPR119W	CLB2	4	7	9	13	8	21	0.0173	3.5841
32	YPL163C	SVS1	4	6	8	9	6	22	0.0360	7.7809
33	YLR210W	CLB4	0	0	0	0	15	3	0.0070	0.0858
34	YGR109C	CLB6	4	4	7	8	10	10	0.0922	5.9788
35	YBR010W	HHT1	0	0	1	1	7	5	0.0504	1.4994
36	YER111C	SWI4	2	21	24	27	1	1	0.0023	0.0000
37	YLR079W	SIC1	3	5	7	11	5	4	0.0384	0.5123
38	YER001W	MNN1	1	2	6	9	1	11	0.0193	3.5400
39	YDR225W	HTA1	1	4	4	3	4	9	0.0429	6.9192
40	YKL185W	ASH1	8	8	15	28	6	1	0.0173	0.0000
	% found		35	37.5	60	75	100	—	—	—

Columns 'm' indicate the number of regulators identified by the algorithm using five different criteria. 'best' means regulators with the smallest *E*. *E* is given by Equation 6, E_1 is given by Equation 9. 'min(m)' means the position of the first correctly found regulator in the list of regulators for the given target, sorted according to the value of *E*. % found—percentage of targets for which the regulators were correctly assigned. Correctly found regulators are defined as those which were also identified as regulators in the independent database of yeast regulators—YEASTRACT. Column 'Min(m) lin' represents Min(m) for the linear model. Column E represents *E* as defined by Equation 6 for the nonlinear model (nonlin, Equation 4) and for the linear model (lin, Equation 7).

annotated as the regulator in the independent YEASTRACT database. It is necessary to emphasize that the YEASTRACT database represents current knowledge, which is already quite comprehensive but still far from complete. Therefore, all comparisons with this database have limited information value. If the regulator was found in YEASTRACT than it was confirmed that the regulator was identified independently elsewhere. If the predicted regulator does not match with the YEASTRACT we can say only that according to the current state of knowledge the gene has not been identified as regulator in other studies.

Expression profiles of the target gene, the best fitting regulator and the reconstructed target gene profiles for 12 cell cycle regulated genes are shown in Figure 1. The remaining profiles are shown in Supplementary Figure 2. Table 1 shows that the regulators selected as 'best' were identified correctly as regulators for only 35% of the targets. However, the average false positive (FP) rate, defined as the ratio between regulators identified as FPs and the total number of potential regulators, was very low (average = 1.1%, data not shown). The FP rate defined here represents the number of predicted regulators which were not found in the YESTRACT database. When the criterion ($E \leqslant E_1$) for selection of the regulators was used, the probability of correct identification slightly increased (to 37.5% see Table 1) but the FP rate was doubled. Slight modification of the criterion to $E \leqslant 1.1 * E_1$ led to rapid increase of the probability of correct detection (see Table 1) but the FP rate increased as well. This trend continued when the criterion was softened to $E \leqslant 1.2 * E_1$. Close inspection of FP rates showed that most of the increases was caused by only a few targets, namely YOR323C, YJL155C, YDR285W and YAL018C, whose profiles could be modeled by almost any regulator. The result was a dramatic increase in the FP rate. All of these profiles exhibited very high

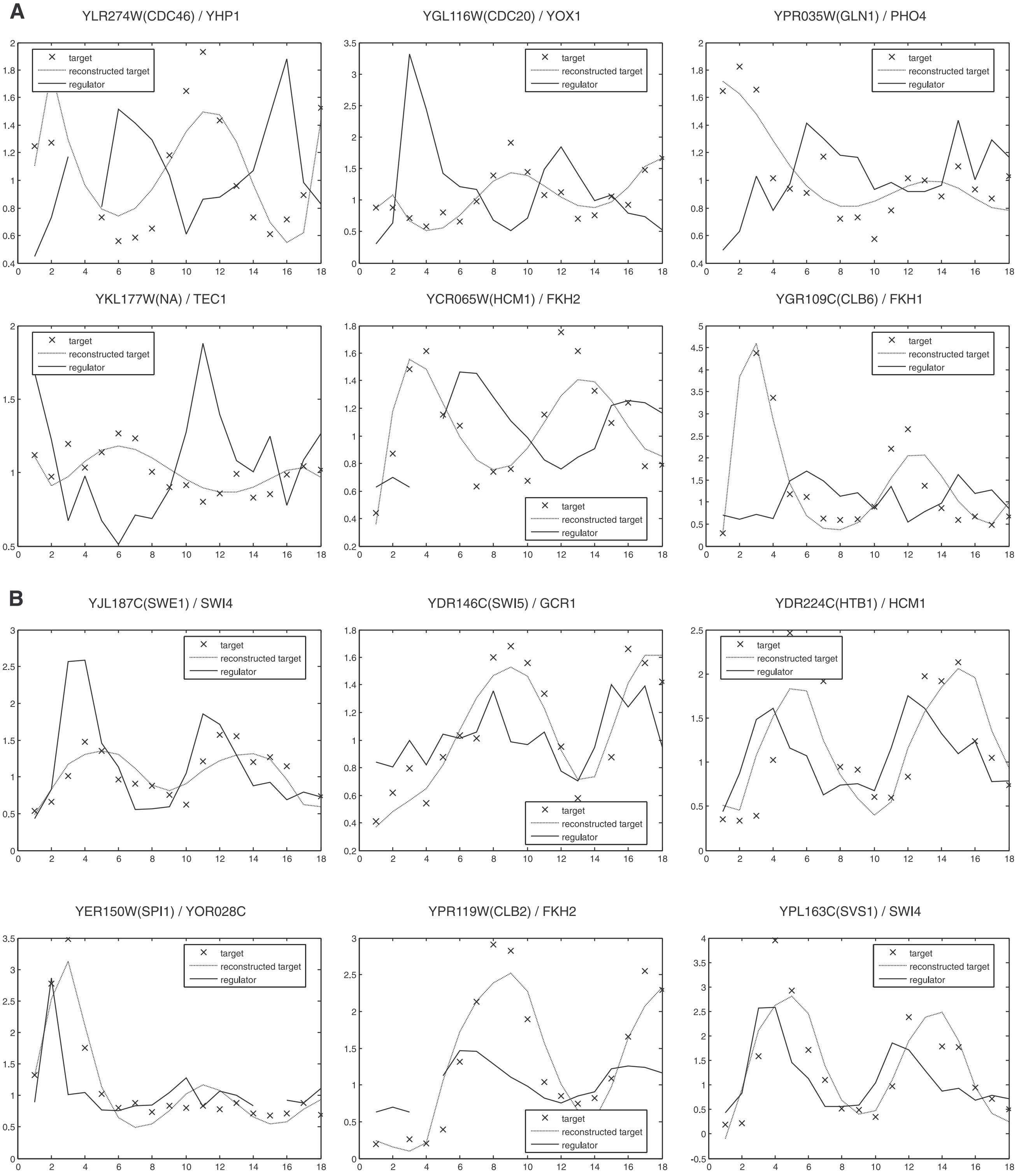

Figure 1. Expression profiles of 12 cell cycle regulated genes and their predicted regulators. (**A**) repressors, (**B**) activators. Horizontal axis—time points, vertical axis—expression relative to time point zero. Gene names in captions are arranged as target/regulator, symbols—target gene profile, dotted line—target gene profile fitted using the model, solid line—profile of the best fitting regulator (the lines are interrupted at the positions where the original data points were missing).

fluctuations (see their profiles in Supplementary Figure 2) and therefore had very high E_1. Moreover, the profiles were flat. Consequently, many regulator genes satisfied the criterion ($E \leqslant E_1$). The results for such target profiles are very difficult to evaluate and, in real situations, such genes should be excluded from the evaluation.

A complete list of the 40 target genes and the predicted regulators with $E \leqslant E_1$ is given in Supplementary Table 2 together with the values of model parameters (w, b, k_1, k_2, Equation 4) and the specificity Sp of prediction. The specificity of prediction is defined here as $\mathrm{Sp} = (\mathrm{N} - \mathrm{FP})/\mathrm{N}$, where N represents number of potential regulators and FP was defined above. This measure is used to indicate a relative improvement in number of experiments necessary for experimental verification of the results of the algorithm compared with the whole pool of the potential regulators.

When the regulators for each target were sorted according to increasing E and the position of the first correctly identified regulator was recorded, the column 'min(m)' of Table 1 was obtained. This column shows that most regulators for a given target were correctly found among the first 5 regulators in the sorted list. For 36 targets out of 40 (90%), a correct regulator was found among the first 10 regulators in the sorted list. If the first 21 regulators in the sorted list for each target were considered, all targets got at least one correctly assigned regulator. This observation greatly simplifies experimental verification. That means that if the 10 regulators with the smallest E are first selected, 90% of the targets get at least one correctly assigned regulator. If we recall that the pool of all potential regulators considered in this paper comprises 184 genes, the algorithm restricts the amount of putative regulators by almost 20-fold.

All identified regulators (the regulators with the smallest E) could also be identified as activators or repressors according to the sign of the weight w. Predictions made by the algorithm were compared with the data in the YEASTRACT database and 77.8% of the identified regulators were also correctly identified as activators or repressors (Supplementary Table 1). If for a specific target no regulator satisfied the above mentioned criteria, it was concluded that the given target did not have a regulator in the pool of regulators. The list of predicted regulators for all targets used in this paper, together with the values of parameters of Equation 4, is given in Supplementary Tables 1 and 2.

In comparison with Chen *et al.* (11), who analyzed a similar dataset, only one regulator was predicted by both methods. That means that both methods identified, in substantial part, different sets of regulators. As the results from Chen *et al.* were not compared with an independent data source, it cannot be evaluated whether their method gives more reliable results.

Some of the differences between the known and predicted functions could be caused by incomplete information in the YEASTRACT database. Other inaccuracies were caused by relatively high experimental noise which cannot be avoided. In the results of this paper we kept these profiles, but in the real world situation, such profiles should be excluded from evaluation. As a least squares minimization procedure was used to compute the target gene expression profile, there is always a risk that the procedure can get stuck in a local minimum of the parametric space and the optimal solution will not be reached. In such cases the real regulator is not found. This was partially avoided by running the least squares minimization procedure several times for different initial values of parameters and selecting the sets of parameters that gave the best solutions. We repeated the procedure 100 times for each pair target/regulator with randomly generated initial values and selected the parameter set which gave the best approximation of the target gene profile.

Comparison with linear model

In order to compare the results of our algorithm we processed the same data with more commonly used linear model, here given by Equation 7. Results are summarized in Table 1 and Figure 2. In both cases, the potential regulators were sorted according to increasing value of E (Equation 6) and the position of the first correctly identified regulator (as correct regulators were defined those which matched with the YEASTRACT database) in the sorted list for both models was recorded [columns Min(m) and Min(m) lin Table 1]. Figure 2A and B show histograms of distributions of these values. It can be seen that the histogram for the linear model, is broader and reaches higher values of Min(m) than the histogram for the nonlinear model.

If columns E (nolin,lin), representing the error function E (Equation 6) for the first correctly identified genes are compared, it is apparent that the goodness of fit for the nonlinear model (Equation 4) is one order of magnitude better than the linear fit. If we accept that the error of measurement in the microarray experiment is 10% of the value, and if it is required that the error of fit is within these 10%, an average threshold of E can be calculated. If 95% of all regulators has to lie within the 10% interval $E = 0.0268$. If the regulators with the value of E higher than this threshold are excluded, then for the nonlinear model 9 genes out of 40 are excluded, but for the linear model 37 out of 40 have to be excluded. It can be concluded that the nonlinear model presented here gives markedly better results than more common linear model.

The best fitting regulators also identified in YEASTRACT were compared with the predictions made by the nonlinear model and by the predictions made by Chen *et al.* (11) (Supplementary Table 1). No match between the predictions made by Chen *et al.* was found. Prediction made by the nonlinear model matched with the linear model in only 5 cases out of 40.

To summarize, our study provides not only a mathematical background for transcriptional regulation, but also makes correct predictions of regulators. The sign of the parameter w of the model function suggests whether the regulator acts as an activator or a repressor of the target gene. Time courses of expression of the target genes estimated by the procedure fit well with the observed ones (see Figures 1 and 2).

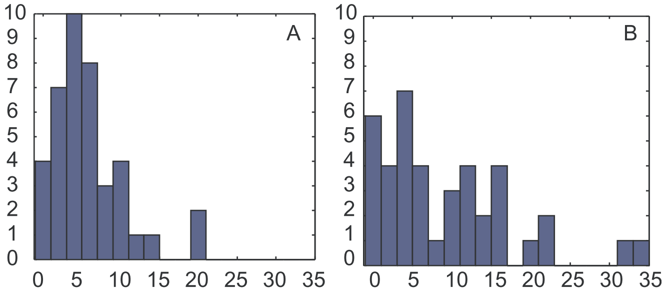

Figure 2. Histogram of distribution of the order of correctly identified regulators in the sorted list of potential regulators [columns Min(m) and Min(m) lin in Table 1], horizontal axis—the order in the sorted list. Regulators were sorted according to the error of approximation of the target gene expression profile (Equation 6). (**A**) Nonlinear model Equation 4, (**B**) linear model Equation 7.

The model can be extended to determine cooperative and competitive regulation by more than one regulator. Such extension increases the computational complexity in a combinatorial way and becomes computationally unfeasible. Currently, we are working on optimization of the computational procedure to simplify the algorithm and increase its speed.

DISCUSSION

Dynamic modeling of biological systems including genetic networks and cell regulatory networks has been around for a long time [for a review see (23)]. Some of the models successfully simulated simple real systems as the genetic network of lambda phage (20,24–26). Yeast cell cycle gene expression data were analyzed by various clustering methods (3,27–29) or by linear algebra (6,7,30). An alternative method of gene location analysis for identification of transcription factors was performed in the work of Lee and Iyer (10,31). A comprehensive list of yeast transcriptional regulators is maintained in the YEASTRACT database (http://www.yeastract.com). Recently, several papers on the mathematical modeling of transcriptional regulation in yeast using microarray time series (11–13,15), and a fuzzy logic approach for identification of triplets of target/activator/repressor (14) were published.

Here, we present a novel algorithm based on a more general concept of a simulated model of genetic networks (19). The model assumes that the kinetic profile of a target gene results from the activity of a particular regulator, which binds the gene upstream and initiates its transcription. From a modeling point of view it means that it is possible to generate the target gene expression profile from the gene expression profile of the regulator using the model and its parameters.

In order to find the correct set of parameters, the difference between the measured target gene expression profile and the profile computed from the expression profile of the regulator has to be minimized. Therefore the search for an appropriate regulator becomes a least squares minimization problem where all possible regulators are tested one by one to determine whether they can model the target gene expression profile. The regulators fitting the target profile best are then selected. In this paper, we show that the model is capable of correctly identifying regulators of yeast cell cycle-associated target genes and their functions (activator or repressor). In comparison with other previously published models which assume linear dependence between the time course of the regulator and the target gene expression profiles, our model assumes a nonlinear dependence given by Equation 2, which conforms better to the observed dynamic behavior of the transcriptional regulatory systems. Here we have focused on a direct one to one relation between the target and the regulator, but the model in principle can incorporate any number of regulators affecting expression of the target gene (see Equation 3).

For more complicated and indirect control processes comprising, e.g. a cascade of regulatory events, more complex interactions would have to be considered. The general model given by Equation 3, extended to all genes of the network, would still remain valid. The algorithm presented here models all combinations of the target/regulator and selects from the results those regulators, which give the best predictions. We then obtain complete information about the ability of all regulators to model each target gene profile. From this set the regulator, which models best the target is selected. Our approach is very comprehensive but on the other hand limits its extendibility to just a small number of regulators due to the high computational requirements. To systematically model all possible interactions controlling one target in a more complex regulatory network would require a large computational load. Fortunately, the algorithm can easily be parallelized and the comprehensiveness of the results might justify the use of parallel computation for full identification of such complex networks.

The presented algorithm was compared with a linear model and it was proven that the nonlinear model used here gives markedly better results than the linear model both in the sense correct identification of regulators and the goodness of fit of the computed target gene expression profile.

Previously published models [Chen *et al.* (11)], which iteratively incorporate more regulators in each step to improve the fit of the modeled profile to the experimentally measured target gene expression profile, suffer from principal inaccuracy; they improve the fit but do not follow the mechanism of gene expression. In reality, all members of the network perform at the same moment and the influence of each member is given by the value of the parameters associated with them. Therefore, hierarchical selection of regulators according to their capability of improving the fit of the model does not guarantee correct identification of the regulators.

Comparison of Chen *et al*'s predictions (11) with the predictions made by the linear and nonlinear models showed that all three methods gave different result set of genes. Best fit of the regulator profile is obtained by the Chen's algorithm but at the cost of prediction of non-documented regulators for the given target. Our nonlinear algorithm selected well genes documented in YEASTRACT database and also provided reasonable goodness of fit of the target gene expression profile. The linear model (Equation 7) performed worse giving lowest fit and lowest prediction ability (see Table 1 for comparison between linear and nonlinear model).

Probably, the only correct approach would be to create particular models for all known types of transcriptional regulation. Then all possible regulators should be tested for all models and an appropriate model and the genes which best approximate the target gene profile under the particular model would be identified. Nachman *et al.* (15) took this approach by suggesting a model of gene expression based on Michaelis–Menten kinetics and dynamic Bayesian networks and retrieving the best network structure from gene expression time series data. Similar to Chen *et al.* (11) they added new regulators during the network structure reconstruction to improve the match between the predicted and known behavior of the system. In contrast with Chen *et al.*, the search was not mechanistic but built a dedicated network structure that approached a probable model of transcriptional regulation.

Instead of attempting to model more complicated regulatory processes with the high risk of incorrect prediction, we focused here on the simplest case but with a reliable outcome.

A drawback of all published algorithms for inference of transcriptional regulatory networks, including this one, is that the candidate regulators are selected from the pool of potential regulators defined independently, usually by sequence similarity analysis, or by other genome annotation methods. If the regulator is not identified, it inevitably escapes identification by the modeling approach. The less characterized the genome of an organism is, the higher the probability of this type of error.

The procedure in principle can use not only the pool of potential regulators, but all genes immobilized on the chip. But by using this approach, the risk of identifying FPs increases. Also, other genes can have expression profiles satisfying the minimization criterion without having anything to do with regulation. Those FPs would eventually have to be sorted out anyway using some independent information. This is, in principle, the same as the a priori selection of potential regulators.

Also, when deciding that a target gene is not controlled by any of the regulators in the predefined set, the target could be controlled by some of the regulators in the set, but this effect could be masked by the inhibition of the activity of the regulator by some other regulator that is not included in the set. Such genes then would not be identified.

Transcriptional regulation is mediated by means of proteins whose active form quite often arises from post-translational modification, e.g. phosphorylation. Such changes cannot be recorded by microarrays, which measure concentration of mRNA. In the modeling using microarray data it is presumed that the regulator protein concentration profile copies the expression profile of mRNA. This assumption is not always valid and cannot be simply predicted just from the microarray data. From this point of view, the use of proteomic data instead of the microarray data in the modeling of transcriptional control seems to be more appropriate.

In general there is no best model. Only a group of models suitable for particular cases have been, or will be, designed. Each of them has particular properties more suitable for particular cases. Relevant results can be obtained by application of several approaches and by critical analysis of the results. One such model is presented in this communication. Results indicate that the model presented here captures the behavior of transcriptional regulation with good accuracy and the predicted regulators well match with documented function obtained independently. Therefore, we believe that our algorithm will be useful for interpretation of gene expression time series.

Although applied to microarray data of *S.cerevisiae*, the algorithm can be used for the data of any other organism, and can be analyzed using an experimental design similar to this case. In the future, the model will be developed so that it will be able to identify more complex transcriptional regulatory interactions.

CONCLUSIONS

The goal of the presented algorithm is not to recover all possible transcriptional regulatory interaction in a genetic network, but rather to focus on comprehensive analysis of the influence of all possible regulators of a given target gene and to recover basic transcriptional regulations with high confidence. Instead of trying to model a large genetic network with high probability of incorrect results we decomposed such a network into elementary interactions where a single regulator acts as an activator or repressor. The algorithm is capable of correct identification of such interactions with a higher accuracy than previously published algorithms. The algorithm not only selects the most probable regulator of a given gene but also provides information about the ability of all potential regulators to control the target gene, thus giving the possibility of investigating the role of other regulators and decreasing the probability of misidentification. The approach presented here is based on the correct physico-chemical background of the system and avoids building of the control network on the basis of improvement of the fit to the experimental data, which can lead to high rate of misidentifications which are difficult to discover. The complex transcriptional control network can in principle be decomposed into the elementary networks, such as the one presented here and the final network can be assembled form them. The hurdle in this approach is the combinatorial increase in the number of necessary computations that grows with the size of the network. For large scale networks, this can lead to an unrealistic number of computations. This suggests improvements in the speed of the algorithm and incorporation of independent information as genome-wide location data, DNA sequence information and targeted biochemical and molecular biological experiments, which can substantially reduce the number of combinations necessary to perform the optimization. The algorithm presented here allows for such extensions. We plan to investigate these possibilities on a model organism and extend the algorithm to the identification of more complex transcriptional regulatory networks.

ACKNOWLEDGEMENTS

The authors wish to thank Josef Panek for useful discussions. This work was supported by grants from the Czech Science Foundation No. 310/03/0293, 310/04/0804 and the EC Integrated Project ActinoGEN, LSHM-CT-2004-005224. Funding to pay the Open Access publication charges for this article was provided by ActinoGEN.

Conflict of interest statement. None declared.

REFERENCES

1. Cho,R.J., Campbell,M.J., Winzeler,E.A., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T.G., Gabrielian,A.E., Landsman,D., Lockhart,D.J. *et al*. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.
2. Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
3. 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.
4. Chu,S., DeRisi,J., Eisen,M., Mulholland,J., Botstein,D., Brown,P.O. and Herskowitz,I. (1998) The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699–705.

5. Holter,N.S., Maritan,A., Cieplak,M., Fedoroff,N.V. and Banavar,J.R. (2001) Dynamic modeling of gene expression data. *Proc. Natl Acad. Sci. USA*, **98**, 1693–1698.
6. Holter,N.S., Mitra,M., Maritan,A., Cieplak,M., Banavar,J.R. and Fedoroff,N.V. (2000) Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc. Natl Acad. Sci. USA*, **97**, 8409–8414.
7. Alter,O., Brown,P.O. and Botstein,D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA*, **97**, 10101–10106.
8. Iyer,R., Iverson,T.M., Accardi,A. and Miller,C. (2002) A biological role for prokaryotic ClC chloride channels. *Nature*, **419**, 715–718.
9. Simon,I., Barnett,J., Hannett,N., Harbison,C.T., Rinaldi,N.J., Volkert,T.L., Wyrick,J.J., Zeitlinger,J., Gifford,D.K., Jaakkola,T.S. *et al.* (2001) Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, **106**, 697–708.
10. Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I. *et al.* (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.
11. Chen,K.C., Wang,T.Y., Tseng,H.H., Huang,C.Y. and Kao,C.Y. (2005) A stochastic differential equation model for quantifying transcriptional regulatory network in *Saccharomyces cerevisiae*. *Bioinformatics*, **21**, 2883–2890.
12. Chen,H.C., Lee,H.C., Lin,T.Y., Li,W.H. and Chen,B.S. (2004) Quantitative characterization of the transcriptional regulatory network in the yeast cell cycle. *Bioinformatics*, **20**, 1914–1927.
13. Sasik,R., Iranfar,N., Hwa,T. and Loomis,W.F. (2002) Extracting transcriptional events from temporal gene expression patterns during *Dictyostelium* development. *Bioinformatics*, **18**, 61–66.
14. Woolf,P.J. and Wang,Y. (2000) A fuzzy logic approach to analyzing gene expression data. *Physiol. Genomics*, **3**, 9–15.
15. Nachman,I., Regev,A. and Friedman,N. (2004) Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*, **20**, I248–I256.
16. Bar-Joseph,Z., Gerber,G.K., Lee,T.I., Rinaldi,N.J., Yoo,J.Y., Robert,F., Gordon,D.B., Fraenkel,E., Jaakkola,T.S., Young,R.A. *et al.* (2003) Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.*, **21**, 1337–1342.
17. Wang,W., Cherry,J.M., Nochomovitz,Y., Jolly,E., Botstein,D. and Li,H. (2005) Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. *Proc. Natl Acad. Sci. USA*, **102**, 1998–2003.
18. Makita,Y., De Hoon,M.J., Ogasawara,N., Miyano,S. and Nakai,K. (2005) Bayesian joint prediction of associated transcription factors in *Bacillus subtilis*. *Pac. Symp. Biocomput.*, 507–518.
19. Vohradsky,J. (2001) Neural network model of gene expression. *FASEB J.*, **15**, 846–854.
20. Vohradsky,J. (2001) Neural model of the genetic network. *J. Biol. Chem.*, **276**, 36168–36173.
21. Vu,T.T. and Vohradsky,J. (2002) Genexp-a genetic network simulation environment. *Bioinformatics*, **18**, 1400–1401.
22. Bar-Joseph,Z. (2004) Analyzing time series gene expression data. *Bioinformatics*, **20**, 2493–2503.
23. de Jong,H. (2002) Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.*, **9**, 67–103.
24. Matsuno,H., Doi,A., Nagasaki,M. and Miyano,S. (2000) Hybrid Petri net representation of gene regulatory network. *Pac. Symp. Biocomp.*, 341–352.
25. McAdams,H.H. and Shapiro,L. (1995) Circuit simulation of genetic networks. *Science*, **269**, 650–656.
26. McAdams,H.H. and Arkin,A. (1998) Simulation of prokaryotic genetic circuits. *Annu. Rev. Biophys. Biomol. Struct.*, **27**, 199–224.
27. Dembele,D. and Kastner,P. (2003) Fuzzy C-means method for clustering microarray data. *Bioinformatics*, **19**, 973–980.
28. Gasch,A.P. and Eisen,M.B. (2002) Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol.*, **3**, research0059.1–research0059.22.
29. Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nature Genet.*, **22**, 281–285.
30. Alter,O., Brown,P.O. and Botstein,D. (2003) Generalized singular value decomposition for comparative analysis of genome-scale expression datasets of two different organisms. *Proc. Natl Acad. Sci. USA*, **100**, 3351–3356.
31. Iyer,V.R., Horak,C.E., Scafe,C.S., Botstein,D., Snyder,M. and Brown,P.O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533–538.