A Smooth Response Surface Algorithm for Constructing Gene Regulatory Network

Hongquan Xu[1,4], Peiru Wu[2], C.F. Jeff Wu[1,4], Carl Tidwell[1] and Yixin Wang[3]

[1]Department of Development Sciences Informatics, [2]Department of Discovery Research Informatics, and [3]Department of Molecular Sciences, Pfizer Global Research and Development, Ann Arbor, MI 48105

[4]Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1285

Abbreviated title: Smooth Response Surface Algorithm

Abstract

A smooth response surface algorithm is developed as an elaborate data mining technique for analyzing gene expression data and constructing gene regulatory network. A three-dimensional smooth response surface is generated to capture the biological relationship between the target and activator-repressor. This new technique is applied to functionally describe triplets of activators, repressors and targets, and their regulations in gene expression data. A diagnostic strategy is built into the algorithm to evaluate the scores of the triplets so that those with low scores are kept and a regulatory network is constructed based on this information and existing biological knowledge. The predictions based on the identified triplets in two yeast gene expression data sets agree with some experimental data in the literature. It provides a novel model with attractive mathematical and statistical features that make the algorithm valuable for mining expression or concentration information, assist in determining the function of uncharacterized proteins, and can lead to a better understanding of coherent pathways.

Keywords: activator-repressor-target model; data mining; diagnostic strategy; gene expression profiling.

Introduction

The rapid advance of genome-scale sequencing is a driving force in the development of methods to exploit this information. The knowledge of the coding sequences of virtually every gene in an organism, for instance, has enabled the development of technology to simultaneously monitor the expression of all the genes. Microarrays use either cDNA clones or PCR products and print them robotically onto a glass microscope slide surface. In contrast, Affymetrix GeneChip system designs oligonucleotides based on sequence information and synthesizes them in situ on the solid support using light-directed, solid-phase combinational chemistry. These arrays are hybridized under stringent conditions with a complex sample representing mRNAs expressed in the test cell or tissue. By doing so, the technology measures the expression level of thousands of genes simultaneously using the oligonucleotides bound to a silicon surface. The results from these expression profiling technologies are quantitative and highly parallel, thereby allowing us to take an accurate snapshot of the workings of the cell in a particular state.

Cells regulate the expression of their genes in response to environmental changes. Normally this regulation is beneficial to the cell, protecting it from starvation or injury. However, errors in this regulation can lead to serious diseases ranging from cancer to heart disease. The pharmaceutical industry is beginning to recognize that gene regulation can be useful for both assaying drugs and as a source for new molecular targets--assuming the regulatory network is well understood. As such, changes in gene expression patterns can be used to assay drug efficacy throughout the drug discovery

process. One assay that takes advantage of the existing level of sequence information and is complementary to sequence and genetic analysis is gene expression profiling. Expression profiling assays generate huge data that are not amenable to simple analysis. A great challenge in maximizing the use of these data is to develop algorithms to interpret and interconnect results for different genes under different conditions.

Most existing methods for analyzing gene expression data are classical and modern statistical clustering techniques, which group genes with similar expression patterns. It includes the methods and applications of hierarchical clustering (1, 4, 8, 11, 15), and self-organizing map (6, 16, 18). The clustering methods, which only distinguish between those genes that have the same and different expression profiles, cannot fully reveal the complex cell regulatory network. Recently, a fuzzy logic approach (19) is proposed to generate a connected network of genes using gene expression data. The fuzzy logic algorithm provides a way to transform precise numbers into qualitative descriptions, then analyze this qualitative data using heuristic rules, and finally transform a qualitative descriptor in the heuristic solution back into a precise number.

To improve and extend the fuzzy logic algorithm, a smooth response surface (SRS) algorithm is introduced and developed as a more elaborate data mining technique with attractive mathematical and statistical features for analyzing gene expression data. Response surface methodology focuses on the relationship between the response and the input factors in the study of a process or system and is widely used in manufacturing and high-tech industries. It can be used to optimize the response or to understand the

underlying mechanism (10, 20). For the present study, a smooth mathematical model is proposed to describe a biological model that governs the qualitative relationship between an activator gene, a repressor gene and a target gene. A quantitative statistical approach is developed that can efficiently extract information from gene expression data to bear on the activator-repressor-target model. The SRS algorithm uses a 3D response surface as a graphic representation of a high-dimensional decision matrix (*i.e.* $n \times n$ decision matrix as n tends to infinity), and leads to a direct process of plug-in quantitative expression data. In contrast, the fuzzy logic method uses some heuristic rules in a decision matrix, and consists of a stepwise process of fuzzification, decision-making and defuzzification. Advantages of the SRS algorithm over fuzzy logic approach include noise tolerance, computational efficiency, and simpler data processing (i.e., from stepwise process to direct plug-in).

The SRS algorithm was used to analyze two yeast expression data gathered from the Affymetrix GeneChip system and cDNA microarray. By using yeast gene expression data collected at different time points of the cell cycle, we were able to identify many regulatory elements and their target genes within the cell that work together to maintain and control certain cellular processes. Many cases are validated by available experimental results, including the signaling network controlling anaerobic and aerobic growth and cell proliferation. These results suggest that the SRS technique can indeed identify biologically relevant connections between sets of genes, which can in turn help describe the complex web of interactions that regulate gene expression.

Methods

Gene regulatory model. Transcriptional regulation has been extensively studied in both prokaryotic and eukaryotic organisms. In many cases, initiation of transcription is controlled by the promoter and its upstream regulatory elements. DNA binding proteins recognize these promoter sequences and activate or repress gene expression through their interactions with promoter and RNA polymerase (12). In developing the SRS algorithm, we employed the a*ctivator-repressor-target* (ART) model to search for triplets of genes *A*, *B* and *C* under which, the concentration of the target gene *C* should be high when the activator *A* is high and the repressor *B* is low. Conversely, when the concentration of the repressor *B* is high and the activator *A* is low, the concentration of the target *C* is low. These qualitative, or heuristic, rules are similar to the judgement calls made by expert systems in data interpretation and can be used to describe the expected behaviors in more complicated biological models in the future. Most importantly, it can be used by combining with more compact and explicit mathematical formulae to extend a decision matrix in the fuzzy logic algorithm to a high-dimensional decision matrix in the proposed SRS algorithm.

Gene expression data sets. Public domain GeneChip and cDNA microarray data describing the yeast cell cycle (2, 4) were chosen to validate the SRS algorithm.

Data processing flowchart. The data processing flowchart of the SRS algorithm is illustrated in Fig. 1. The input parameters and functions are shown in Table 1, which

describe mathematical and statistical constraints that are required for normalization and computation in the algorithm. The imputation and gene filtering steps are applied to remove noise from the data to ensure that the expression data is above the noise level and the observed signal change in the data is significant. In the SRS model, we define a three-dimensional response surface as a function of a pair of genes, which builds a relationship between the target and activator-repressor. First, a transformation step maps the data into the 3D response surface space, which describes the relationship for the triplet of activator-repressor-target. For each triplet of genes (*A, B, C)*, a lack-of-fit formula is used to filter the triplets from the initial screening. Then a diagnostic method is developed to further refine the selected triplets. A final score, based upon the strength of the triplet interrelationship, is defined in order to rank the triplets.

A US patent application is filed on the SRS algorithm and the series number is A0000407L1. Copies of the program are available upon request to the authors.

*Smooth response surface model***.** To fit the expression data into the response surface model, first, the raw data over the time points are transformed into the interval [0,1] such that the minimum and maximum values for each gene are 0 and 1 respectively. We define a three-dimensional smooth response surface given by $S(A,B)$, that is a piecewise linear-quadratic polynomial on $[0,1]\times[0,1]$. It describes a surface in 3D unit cell as shown in Fig. 2, and can be used to describe the biological relationship between the target and activator-repressor genes. The triplets that follow the activator-repressor-target relationship should lie closely to the response surface.

The 3D response surface may be interpreted as a graphic representation of a high dimensional decision matrix. After transformation, the normalized values A and B are broken into various classes, from "LOW" (if it is close to 0) to "HIGH" (if it is close to 1) or in the between (if it is around 0.5). The function $S(A,B)$ maps two normalized values A and B onto a 3D surface, in order to describe a surface response value C whose value also ranges from "LOW" to "HIGH" in association to the two given values A and B. It uses some heuristic rules to facilitate decision-making. As seen in Fig. 2, a triplet $(A,B,S(A,B))$ represents the biological relationship that follows the pattern of a target $S(A,B)$ controlled by an activator A and a repressor B as described in the activator-repressor-target model. The response surface captures the biological model with features such as compactness, simplicity and visualization.

For each triplet (A,B,C), the fitted value of a target gene C is given by $\hat{C} = S(A,B)$. If the activator-repressor-target relationship is strong, the residual, $\hat{C} - C$, should be small. The residual sum of squares measures the overall variation in C that is not explained in the response surface model. Then the lack-of-fit function $RT(A,B,C)$, *i.e.*, the ratio of the residual sum of squares and the total sum of squares, describes the proportion of variation in C that is not captured by the 3D response surface. A small value of lack-of-fit indicates that there is a strong activator-repressor-target relationship among A, B and C. To save storage and computation, only those triplets whose lack-of-fit values do not exceed a given constant RT are kept.

After the initial filtering, a diagnostic strategy $Diag(A,B,C)$ is applied to check the reliability of triplets. For each triplet (A, B, C), $Diag(A,B,C)$ measures robustness of the fitted model. It is observed that the intensity measurement of gene expression at one or two time points may deviate from the model and suggest that the measurement may be faulty and should be treated as an outlier. If such a value occurs at the *i*-th time point, then $RT_{(i)}(A,B,C)$, i.e., the lack-of-fit of (A, B, C) when the *i*-th time point (or the *i*-th column) is left out, will differ greatly from *RT(A,B,C).* $Diag(A,B,C)$ provides a summary measure over all time points for a given triplet. A larger *Diag* value would suggest that the information for the triplet is unreliable and should be removed for further consideration. This step leads to the criteria for selecting triplet candidates: $RT(A,B,C) \le RT$ and $Diag(A,B,C) \le Diag$, where *RT* and *Diag* are constants as specified by users.

A final score is defined to measure the strength of the triplet interrelationship. Score (A, B, C) is a function of the lack-of-fit value and the diagnostic measure, and focuses primarily on the *RT(A,B,C)* value and secondly on the *Diag(A,B,C)* values. Triplets with low values of *RT(A,B,C)* and *Diag(A,B,C)* will have low scores indicating a close relationship among A, B and C.

Results

GeneChip data on yeast cell cycle. The SRS algorithm is applied to a public oligonucleotide GeneChip data set that studied gene expression profiles during the yeast

cell cycle (2). The data were collected at 17 time points for 6457 genes on the Affymetrix Ye6000 chip. The fluorescence intensities are used for analysis. First, negative and small positive values, which are due to measurement error and thus not reliable, are imputed. The lowest 5% values are replaced by the 5^{th} percentile over all genes for each time point. The 5^{th} percentiles of fluorescence intensities vary from tens to twenties over the 17 time points. Next, genes are filtered such that the maximum fold-change (i.e., ratio of the maximum and minimum values over the 17 data points) is at least 3 and maximum intensity is at least 100. After filtering, 1,514 genes are retained and processed by the SRS algorithm to form a triplet candidate pool. There are 28,023 triplets, out of $1514 \times 1513 \times 1512 \approx 3.5 \times 10^9$ possible triplets, whose *RT(A,B,C)* values were found to be less than 0.1. Fig. 3 shows the best 9 fitting triplets from the initial screening. However, the low lack-of-fit scores, except for the first triplet, are caused by the extreme values at the 90-minute point. Therefore, most of these triplets are not reliable and should not be interpreted as having a potential biological interrelationship. These unreliable triplets have very large *Diag* values (>7.2) and are filtered out by the diagnostic procedure. Finally, 20,500 triplets with diagnostic measures less than 2 are selected and scored. Table 2 lists top 20 scoring triplets with known functions and Fig. 4 shows the first 9 triplets in graphics.

In order to evaluate the algorithm, the best scoring triplets were examined to see if they make biological sense. A complete table of all the triplets can be obtained from the authors. Fig. 5 shows a connected network of all triplets that have known functions. Within this network, many predicted activator-repressor-target relationships have been

confirmed by published experimental results (3). Looking at a few of common targets in this network, most associated regulatory genes either carry similar cellular functions or are involved in the same cellular process. For example, CDC9 encodes an ATP-dependent DNA ligase and is an essential gene for cell division and DNA recombination. Four of the identified regulators (i.e., SMC1, NIP29, BTT1 and NUM1) are functionally related. SMC1 acts as a positive regulator and is a chromosomal ATPase family member. Like CDC9, SMC1 is involved in chromosome structure and segregation. Another positive regulator is NIP29 that is a structural protein for microtubule nucleation and spindle body duplication. For the two negative regulators, BTT1 has repressor effects on the expression of several genes and NUM1 functions in nuclear migration and microtubule polymerization.

Another major node in this predicted network highlights HAP1. The transcription factor HAP1 has been shown to repress the nuclear encoding cytochrome gene CYC7 under anaerobic growth and activate CYC7 under aerobic growth. The prediction suggests that HAP1 repress CYC7, which in turn accurately predicts that the cells used in this data set were primarily grown under anaerobic conditions. Two other gene products, FAA1 and HES1 that are involved in cellular lipid metabolism and ergosterol biosynthesis have also been implicated in HAP1 regulation in the literature (14).

In addition, the SRS algorithm uncovered relationships for SPO13, CBF2 and YGP1. SPO13 acts as a transcriptional activator and controls meiotic chromosome segregation. CBF2 is a centromere-binding factor in a multisubunit kinetochore protein complex.

YGP1 codes for a glycoprotein synthesized in response to nutrient limitation. The common theme in these gene products is their functions in cell proliferation and cell division. As expected, many genes associated with each of them have been shown to carry similar functions.

The same data set was analyzed by fuzzy logic (19). We compared the analysis results of the fuzzy logic and the SRS methods. In the fuzzy logic approach, 1,898 genes (30%) that have at least 3 fold-change and maximum intensity 30 were retained. A simple network of six genes was constructed from about 470,000 triplets (0.007% of all possible triplets). The analysis took about 200 hours on an 8-processor SGI Origin 2000. In contrast, in the SRS approach, 1,451 genes (23%) that have at least 3 fold-change and maximum intensity 100 were retained for the SRS algorithm. A complex network was constructed from 20,500 triplets (0.0006% of all possible triplets). The analysis took about only 4 hours. The conclusion is that the SRS algorithm is more conservative, more reliable and significantly more efficient in computation.

Microarray data on yeast cell cycle. The SRS algorithm is also applied to a yeast cDNA microarray data set, which were collected at 15 time points of the yeast cell cycle by using a 2,467-gene microarray (4). The Cy5/Cy3 fluorescence ratios are used for analysis. First, the missing values are imputed and replaced by the average of the previous and following time points. The replacement is carried out one gene at a time. A gene is filtered out if it has two or more consecutive missing time points. After imputation, the data set is filtered using the criterion of minimum fold-change of 3. This

leads to 830 genes which are analyzed by the SRS algorithm. There are 2,393 triplets whose *RT(A, B, C)* values are less than 0.1 and *Diag(A,B)* measures are less than 2. These 2,393 triplets are scored for further investigation.

In order to evaluate and validate the results predicted by the SRS algorithm, the best scoring triplets that have known functions were examined first. Within the network illustrated in Fig. 6. many predicted activator-repressor-target relationships have been implicated in published results (17, 22). For example, RSR1 is a RAS GTPase involved in bud site selection. The predicted RSR regulators include CDC45, POL2, DPB2, SWI5 and RPM2 and these proteins also function in budding and cell proliferation. ASF1 causes depression of many silent chromosomal loci when overexpressed in a cell. Because of its broad functionality, a large number of triplets have been found with ASF1 in them. ASF1 represents a major node in the network of Fig. 6, suggesting that its associated proteins may have diverse cellular roles. Comparing to the GeneChip data, fewer triplets can be explained by existing yeast genetic knowledge. This may primarily be caused by the difference in experiment design. In addition, because we do not have access to the original intensity values of the experiment, no data filtering is performed based on expression level. Fold change ratios for some of the genes with low level of expression might not be reliable. Further experimentation is needed to explore these predicted relationships**.**

Discussion

The validation studies show that the SRS algorithm provides a powerful data mining tool for analyzing gene expression data. In general, the findings of the algorithm agree well with published experimental results. This should not come as a surprise, because the algorithm searches for relationships that fit our scientific understanding of how an activator, repressor, and target should interact. By using essentially the same criteria that an experimenter would use to describe the regulatory function of a protein, the SRS algorithm approximates the thought process an expert would use in interpreting or analyzing this data. However, by applying a computational algorithm to the analysis of the data, we have provided a process of data sorting in an unbiased manner, quickly and efficiently.

The mathematical and statistical features make the SRS algorithm powerful and valuable for mining gene expression information. One of the important features is the use of a diagnostic strategy. It ensures that the triplets with unreliable measurements at one time point is filtered out and the final selected triplets would not have the biased data points. For the yeast GeneChip data used in the validation study, there is one extreme time point (i.e. 90-minute) whose dynamic range of the intensities is quite different from other time points. One strategy is to remove the whole experiment from the data set (16). Removing the whole experiment can result in a severe loss of information as the extreme values only occur for some genes. For other genes in the experiment, the information they carry is still valuable and should be exploited in our algorithm. The leave-one-out diagnostic strategy has the ability to extract useful information from that time point and in the meanwhile to minimize error from the noise. To confirm this observation, we

apply the algorithm to a reduced data set with the 90-minute experiment removed and construct a new network. The new network is similar to Fig. 5. However, some parts are missing. The new network does not contain the pathways of CDC9 and YGP1, which are parts of Fig. 5.

Compared with the fuzzy logic approach, there are many advantages in analyzing gene expression data with the SRS algorithm. First, the SRS algorithm is less sensitive to noise because it employs a smooth response surface, while the fuzzy logic approach is more sensitive to noise because it makes discrete decisions (based on a discontinuous response surface). For example, given genes $A = 0.10$ and $B = 0.49$, if the noise causes gene $B' = 0.51$, then in the SRS algorithm the fitted values of the product genes are given by $\hat{C} = S(A, B) = 0.11$ and $\hat{C}' = S(A, B') = 0.098$, which results in the noise-induced-error $\hat{C} - \hat{C}' = 0.012$, which is about 10%. In contrast, in the fuzzy logic method, $\hat{C} = 0.305$ and $\hat{C}' = 0.1475$, and the noise-induced-error is given by $\hat{C} - \hat{C}' = 0.1575$, which is about 50%. This example shows the noise-induced-error in the fuzzy logic approach is more serious than that in the SRS algorithm. As a result, the fuzzy logic approach may miss important information inherent in the genes. The SRS algorithm predicts a larger and more complicated gene regulatory network than the fuzzy logic approach for the same GeneChip expression data. Secondly, the SRS algorithm is significantly more efficient in computation because of its mathematical simplicity and compactness. It has shown tremendous savings in computing time.

Gene expression profiling is a rapid high-throughput process that gives a large amount of information about the cell in a form that can be easily processed on a computer. By using statistical and data mining approaches to analyzing expression profile data, it is possible to confirm the function of a known gene. Moreover, because an exploratory algorithm like the SRS does not require biological information about the genes, genes with unknown functions can be included as easily as genes with known functions. Although in this study the algorithm was only used to search for triplets of activator, repressor, and target genes, the technique is general and can be applied to other relationships and more complicated systems. Examples include other classes of relationships such as co-activators and co-repressors (12) or more complicated systems that involve genes whose transcription is regulated in complex ways by any number of transcription factors. Similarly, although the validation of this algorithm was performed using GeneChip and microarray data in this paper, the algorithm should work equally well with other expression profiling techniques

References

1. **Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, and Levine AJ.** Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci USA* 96: 6745-6750, 1999.
2. **Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, and Davis RW.** A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* 2: 65-73, 1998.
3. **Deckert, J, Perini R, Balasubramanian B, and Zitomer RS.** Multiple elements and auto-repression regulate rox1, a repressor of hypoxic genes in *Saccharomyces cerevisiae*. *Genetics Soc Am* 139: 1149-1158, 1995.
4. **Eisen MB, Spellman PT, Brown PO, and Botstein D.** Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95: 14863-14868, 1998.
5. **Fytlovich S, Gervais M, Agrimonti C, and Guiard B.** Evidence for an interaction between the CYP1(HAP1) activator and a cellular factor during heme-dependent transcriptional regulation in the yeast *Saccharomyces cerevisiae*. *EMBO J* 12: 1209-1218, 1993.
6. **Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh M, Downing JR, Caligiuri MA, Bloomfield CD, and Lander**

ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286: 531-537, 1999.

7. **Hach A, Hon T, and Zhang L.** A new class of repression modules is critical for heme regulation of the yeast transcriptional activator HAP1. *Mol Cell Biol* 19: 4324-4333, 1999.
8. **Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JCF, Trent JM, Staudt LM, Hudson Jr. J, Boguski MS, Lashkari D, Shalon D, Botstein D, and Brown PO.** The transcriptional program in the response of human fibroblasts to serum. *Science* 283: 83-87, 1999.
9. **Lodi T, and Guiard B.** Complex transcriptional regulation of the *Saccharomyces cerevisiae* CYB2 gene encoding cytochrome b_2: CYP1(HAP1) activator binds to the CYB2 upstream activation site UAS1-B2. *Mol Cell Biol* 11: 3762-3772, 1991.
10. **Myers RH, and Montgomery DC.** *Response Surface Methodology: Process and Product in Optimization Using Designed Experiments.* New York: Wiley, 1995.
11. **Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, Ross DT, Pergamenschikov A, Williams CF, Zhu SX, Lee JCF, Lashkari D, Shalon D, Brown PO, and Botstein D.** Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc Natl Acad Sci USA* 96: 9212-9217, 1999.
12. **Ptashne M.** *A Genetic Switch.* Oxford: Cell Press, 1986.

13. **Prezant T, Pfeifer K, and Guarente L.** Organization of the regulatory region of the yeast CYC7 gene: multiple factors are involved in regulation. *Mol Cell Biol* 7: 3252-3259, 1987.

14. **Schneider JC, and Guarente L.** Regulation of the yeast CYT1 gene encoding cytochrome c_1 by HAP1 and HAP2/3/4. *Mol Cell Biol* 11: 4934-4942, 1991.

15. **Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, and Futcher B.** Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by Microarray Hybridization. *Mol Biol Cell* 9: 3273-3297, 1998.

16. **Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, and Golub TR.** Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA* 96: 2907-2912, 1999.

17. **Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, and Church GM.** Systematic determination of genetic network architecture. *Nat Gene* 22: 281-285, 1999.

18. **Toronen P, Kolehmainen M, Wong G, and Gastren E.** Analysis of gene expression data using self-organizing maps. *FEBS Letters* 451: 142-146, 1999.

19. **Woolf PJ, and Wang Y.** A fuzzy logic approach to analyzing gene expression data. *Physiol Genomics* 3: 9-15, 2000.

20. **Wu CFJ, and Hamada M.** *Experiments: Planning, Analysis and Parameter Design Optimization*. New York: Wiley, 2000.

21. **Zhang L, Hach A, and Wang C.** Molecular mechanism governing heme signaling in yeast: a higher-order complex mediates heme regulation of the transcriptional activator HAP1. *Mol Cell Biol* 18: 3819-3828, 1998.

22. **Zitomer RS, Limbach MP, Rodriguez-Torres AM, Balasubramanian B, Deckert J, and Snow PM.** Approaches to the study of rox1 repression of the hypoxic genes in the yeast *Saccharomyces cerevisiae.* *Methods Enzymol* 11: 279-288, 1997.

Figure Legends

Fig. 1. The data processing flowchart of the SRS algorithm. The imputation and gene filtering steps are applied to remove noise from the data. In the SRS model, a transformation is made to map the data into the 3D response surface space, which describes the relationship for the triplet of activator-repressor-target. For each triplet of genes, a lack-of-fit formula is defined to filter the triplets from the initial screening. Then a diagnostic method is developed to refine the selected triplets and a score, which reflects the strength of the triplet interrelationship, is defined to rank the refined triplets. Finally, a gene regulatory network is constructed based on the top scoring triplets.

Fig. 2. A 3D smooth response surface. It is a piecewise linear-quadratic polynomial on $[0,1]\times[0,1]$ and is used to model the biological relationship between the target and activator-repressor genes.

Fig. 3. The best fitting triplets without diagnostic. The graphs show the log transformed (base 2) concentration values of the triplets. The lack-of-fit values for these triplets are less than 0.017. Except for the first triplet, all others have large *Diag* values (>7.2). The low lack-of-fit is caused by the extreme values at the 90-minute experiment. These triplets are not reliable which are indicated by the large *Diag* values.

Fig. 4. The activator-repressor-target mechanism. The graphs show the log transformed (base 2) concentration values of the triplets. It captures the biological relationship among the triplets. When the concentration of the activator A (long-dash line) is high and that of

the repressor B (short-dash line) is low, the concentration of the target C (solid line) is high; when the concentration of A is low and that of B is high, the concentration of C is low.

Fig. 5. A predicted gene regulatory network for the GeneChip data. It is constructed with the top scoring triplets of known functions and existing biological knowledge. The result agrees well with the literature information.

Fig. 6. A predicted gene regulatory network for the microarray data. It is constructed with the top scoring triplets.

Table 1. List of input parameters and functions in the algorithm

Computational parameters:	
n	number of samples (or time points)
Gene	$A = (a_1,...,a_n)$
	$B = (b_1,...,b_n)$
	$C = (c_1,...,c_n)$
Gene-triplet	(A,B,C)
Computational functions:	
Response surface function	$S(a,b) = \begin{cases} 2a(1-b), & \text{if } 0 \le a \le 0.5 \text{ and } 0.5 \le b \le 1; \\ 1-2(1-a)b, & \text{if } 0.5 \le a \le 1 \text{ and } 0 \le b \le 0.5; \\ a-b+0.5, & \text{otherwise} \end{cases}$
Lack-of-fit function	$RT\ (A,B,C) = \dfrac{\sum_{i=1}^{n} (c_i - \hat{c}_i)^2}{\sum_{i=1}^{n} (c_i - \bar{c})^2}$
Diagnostic function	$Diag(A,B,C) = \dfrac{\left(\frac{1}{n} \sum_{i=1}^{n} \left[RT_{(i)}(A,B,C) - RT(A,B,C) \right]^2 \right)^{1/2}}{RT(A,B,C)}$
Score function	$Score(A,B,C) = RT(A,B,C)(1 + Diag(A,B,C))$

Table 2. Top scoring triplets with known functions from the yeast GeneChip data

Rank	A	B	C	RT(A,B,C)	Diag(A,B,C)	Score(A,B,C)
1179	TEC1	PDS1	YGP1	0.0553	0.48307	0.08201
1274	GAP1	MSB1	ARE2	0.06608	0.25991	0.08325
1339	PIR3	RNP1	CPS1	0.07207	0.16708	0.08411
1340	HPR5	GAP1	HAP1	0.07227	0.16418	0.08414
1380	PIR3	RNP1	FAA1	0.07604	0.1136	0.08468
1480	RAD27	MSK1	HES1	0.06075	0.41396	0.08589
1612	SPT21	CBF2	GPA1	0.07278	0.19939	0.0873
1645	HES1	TWT2	HAP1	0.06965	0.25932	0.08771
1920	TIP1	MCR1	SPO13	0.07994	0.13201	0.0905
1947	HPR5	PEP5	HAP1	0.08003	0.13536	0.09086
2117	TIP1	AGP1	SPO13	0.08185	0.12926	0.09243
2226	INH1	CBF2	YGP1	0.07592	0.22926	0.09333
2243	CYB2	CIK1	TIP1	0.07185	0.30092	0.09347
2277	RAD27	CPS1	HES1	0.06435	0.45645	0.09373
2503	CYB2	MCR1	SPO13	0.06806	0.40383	0.09555
3107	HES1	GPD2	HAP1	0.08957	0.12041	0.10036
3122	KAR3	FAA1	HAP1	0.08402	0.19595	0.10048
3186	INH1	PDS1	YGP1	0.08022	0.25682	0.10082
3941	IPL1	TSM1	CBF2	0.09555	0.09849	0.10497
4273	SPO16	ASE1	SMC3	0.09026	0.18177	0.10666

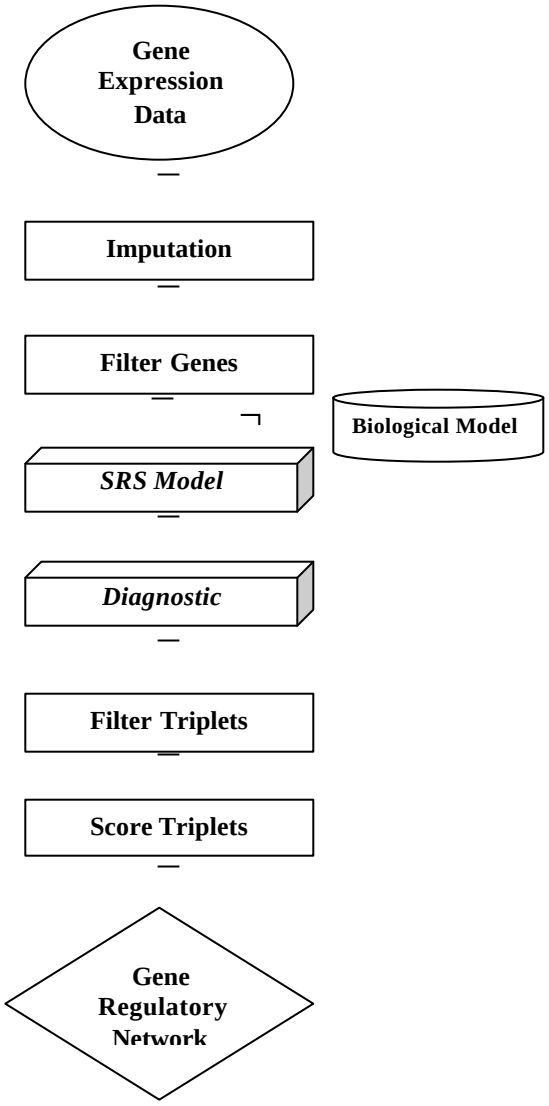
Gene Expression Data
Imputation
Filter Genes
Biological Model
SRS Model
Diagnostic
Filter Triplets
Score Triplets
Gene Regulatory Network

Fig. 1.

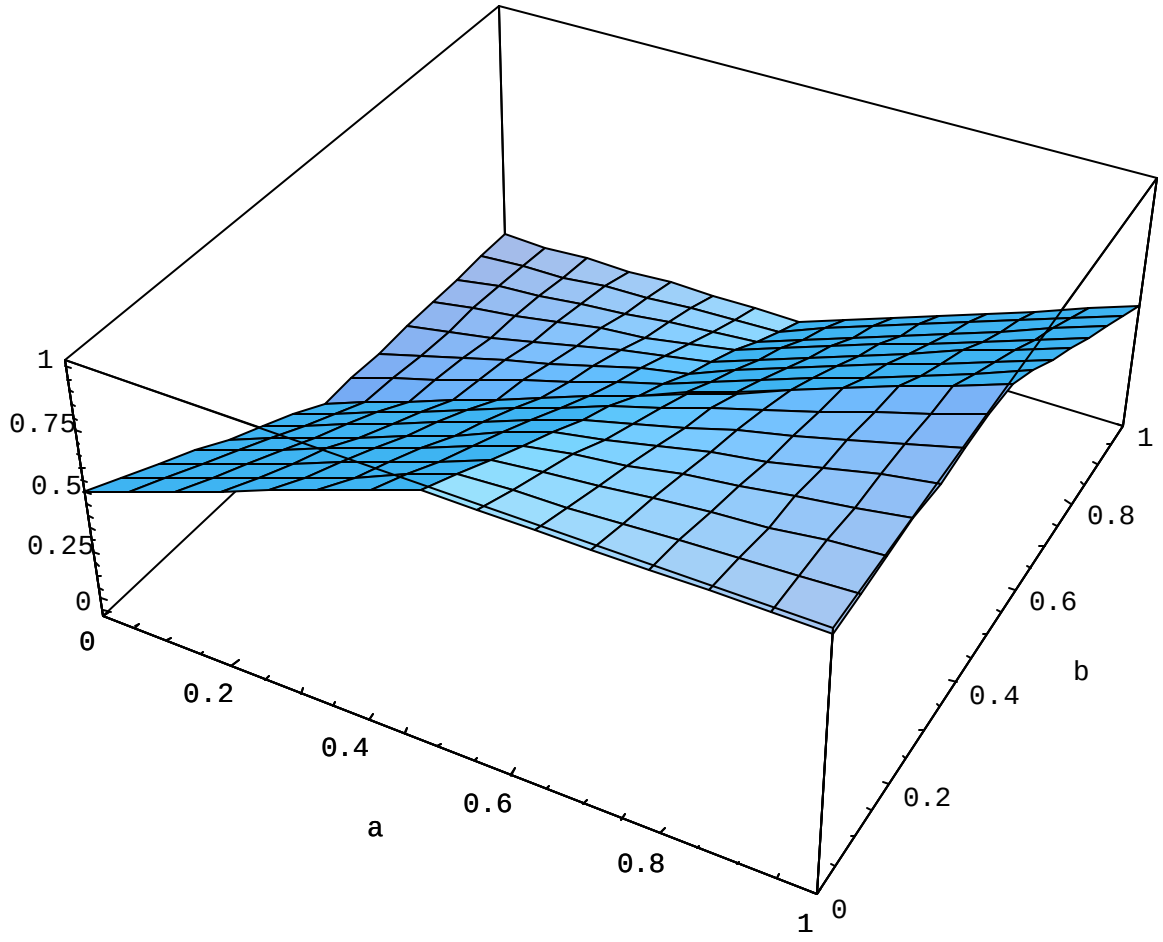

Fig. 2.

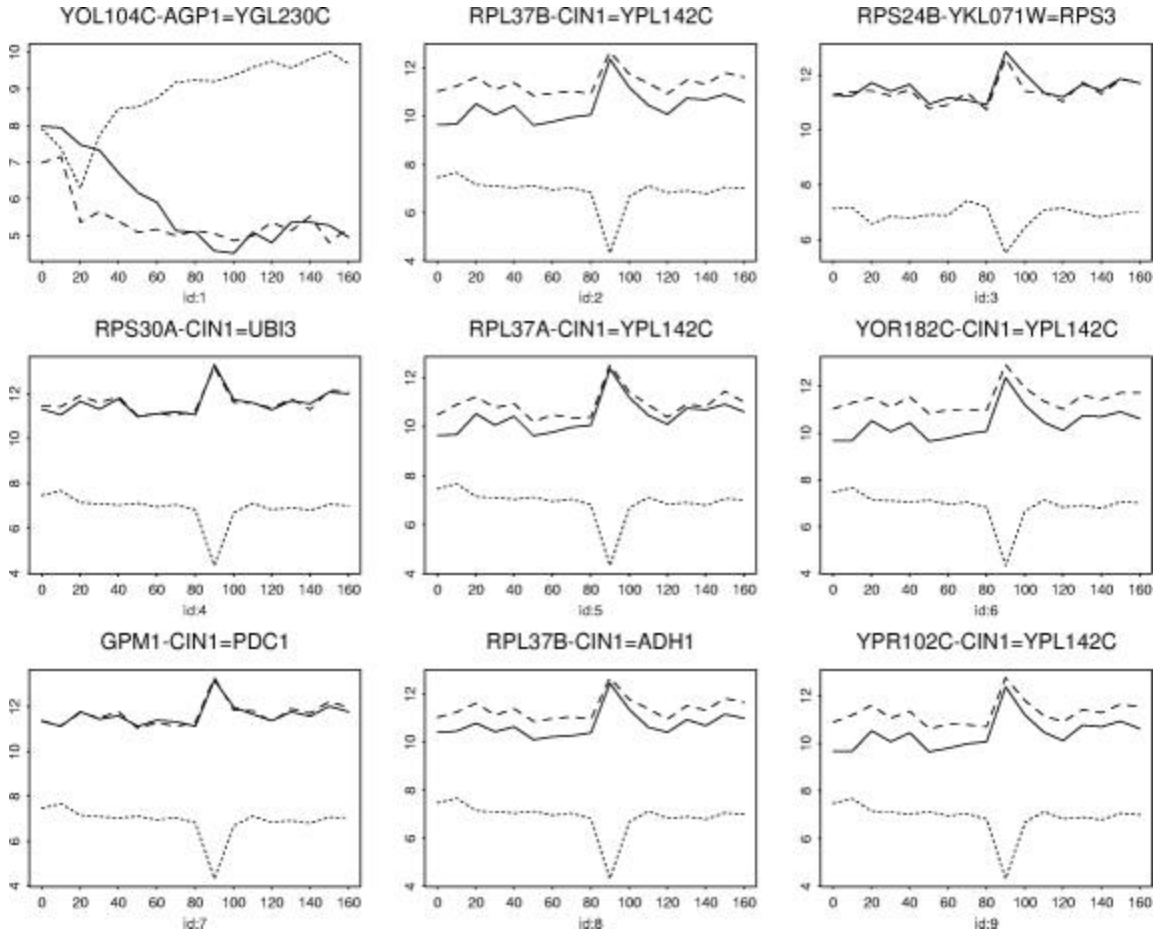
YOL104C-AGP1=YGL230C
RPL37B-CIN1=YPL142C
RPS24B-YKL071W=RPS3
RPS30A-CIN1=UBI3
RPL37A-CIN1=YPL142C
YOR182C-CIN1=YPL142C
GPM1-CIN1=PDC1
RPL37B-CIN1=ADH1
YPR102C-CIN1=YPL142C
id:1
id:2
id:3
id:4
id:5
id:6
id:7
id:8
id:9

Fig 3.

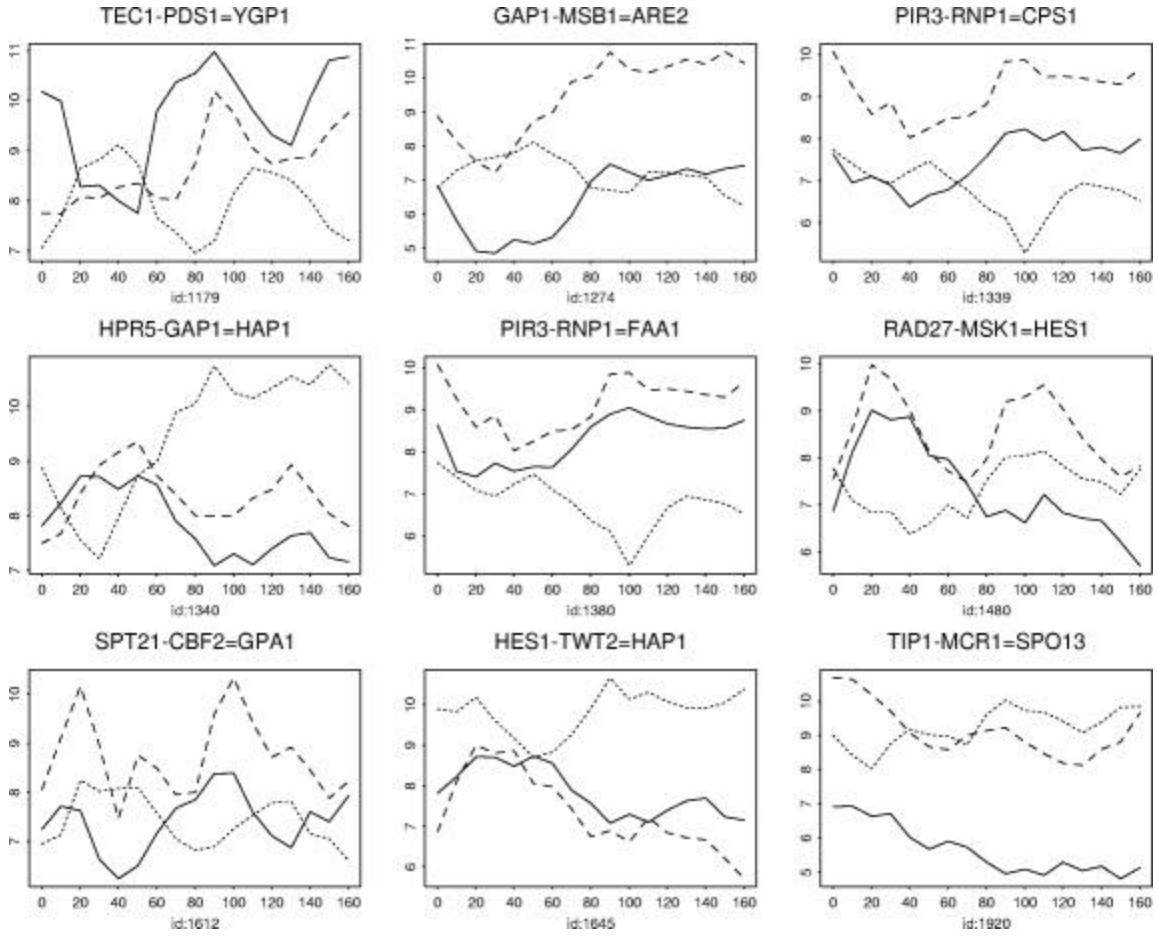
TEC1-PDS1=YGP1
id:1179
GAP1-MSB1=ARE2
id:1274
PIR3-RNP1=CPS1
id:1339
HPR5-GAP1=HAP1
id:1340
PIR3-RNP1=FAA1
id:1380
RAD27-MSK1=HES1
id:1480
SPT21-CBF2=GPA1
id:1612
HES1-TWT2=HAP1
id:1645
TIP1-MCR1=SPO13
id:1920

Fig. 4.

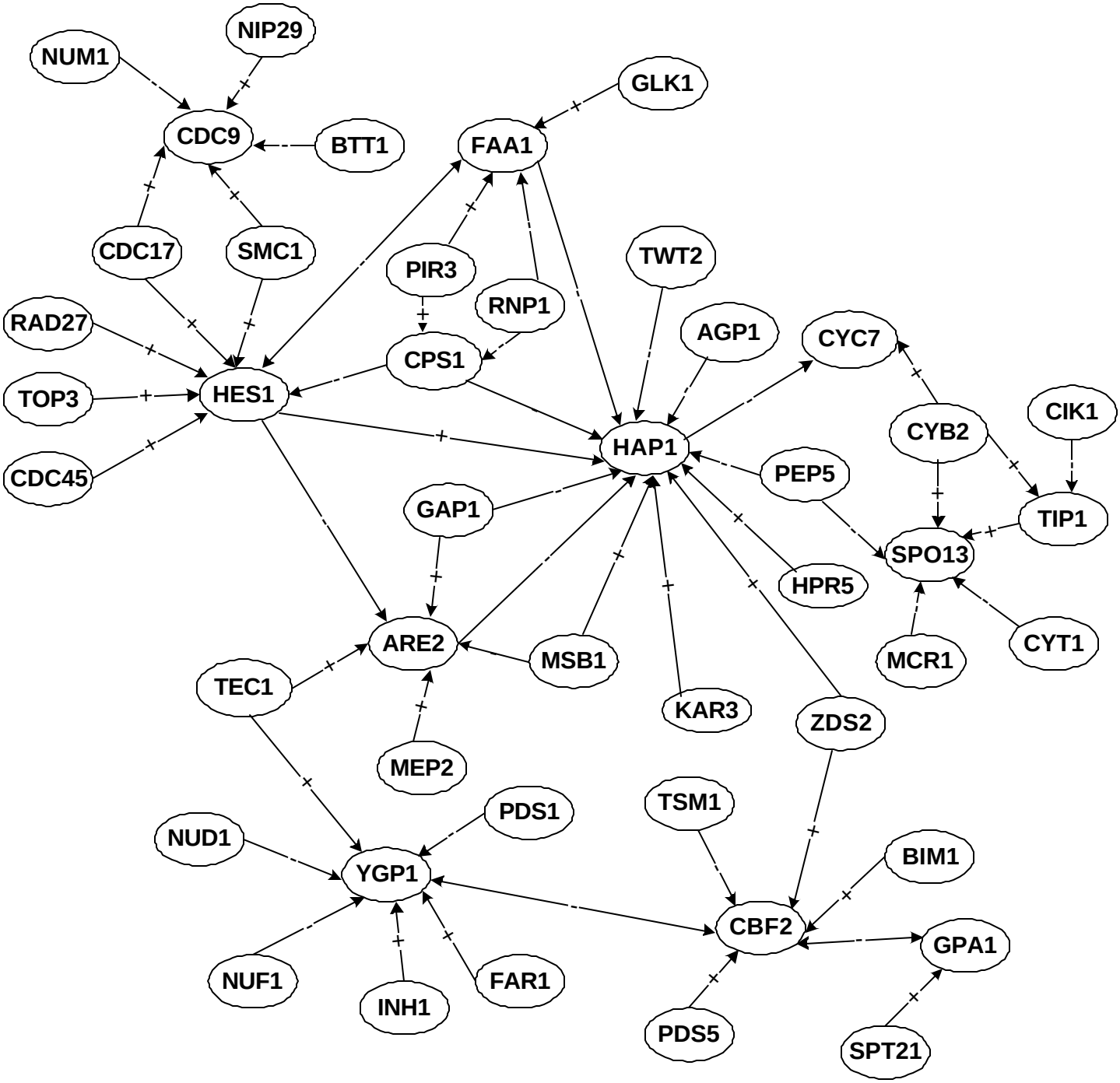
NIP29
NUM1
GLK1
CDC9
BTT1
FAA1
CDC17
SMC1
PIR3
TWT2
RNP1
RAD27
AGP1
CYC7
CPS1
TOP3
HES1
CYB2
CIK1
HAP1
CDC45
PEP5
GAP1
TIP1
SPO13
HPR5
ARE2
MSB1
MCR1
CYT1
TEC1
KAR3
ZDS2
MEP2
TSM1
PDS1
NUD1
BIM1
YGP1
CBF2
GPA1
NUF1
FAR1
INH1
PDS5
SPT21

Fig. 5.

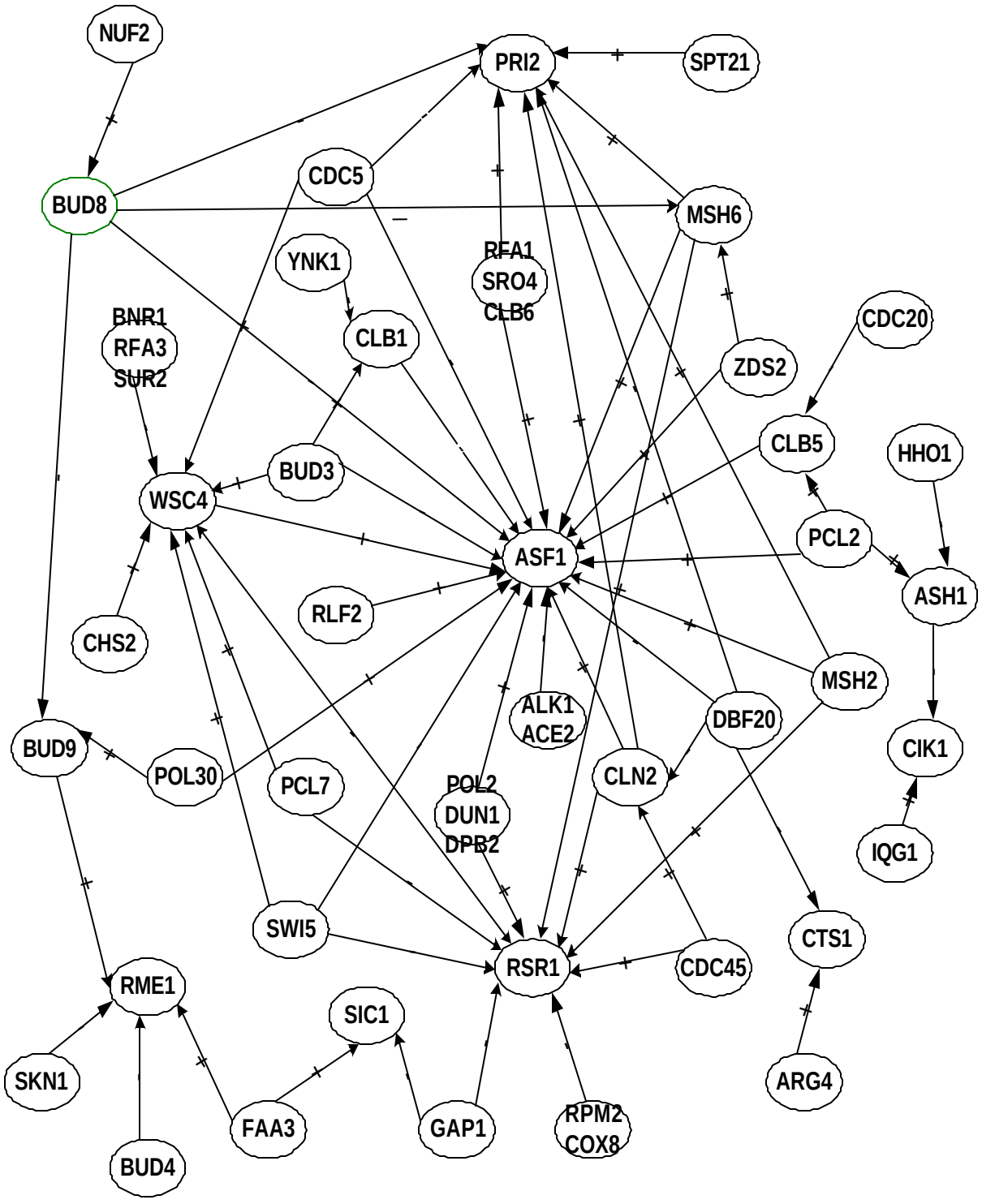
NUF2
PRI2
SPT21
CDC5
BUD8
MSH6
YNK1
RFA1
SRO4
CLB6
BNR1
RFA3
SUR2
CLB1
CDC20
ZDS2
CLB5
HHO1
BUD3
WSC4
ASF1
PCL2
ASH1
RLF2
CHS2
MSH2
ALK1
ACE2
DBF20
BUD9
POL30
PCL7
CLN2
CIK1
POL2
DUN1
DPB2
IQG1
SWI5
CTS1
RSR1
CDC45
RME1
SIC1
SKN1
ARG4
FAA3
GAP1
RPM2
COX8
BUD4

Fig. 6.