A smooth response surface algorithm for constructing a gene regulatory network

HONGQUAN XU,1,2 PEIRU WU,1 C. F. JEFF WU,3 CARL TIDWELL,1 AND YIXIN WANG2
1Department of Discovery Research Informatics, and 2Bioinformatics, Department of Molecular Sciences, Pfizer Global Research and Development, Ann Arbor 48105; and 3Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109
Received 24 July 2001; accepted in final form 6 August 2002

Xu, Hongquan, Peiru Wu, C. F. Jeff Wu, Carl Tidwell, and Yixin Wang. A smooth response surface algorithm for constructing a gene regulatory network. Physiol Genomics 11: 11–20, 2002; 10.1152/physiolgenomics.00060.2001.—A smooth response surface (SRS) algorithm is developed as an elaborate data mining technique for analyzing gene expression data and constructing a gene regulatory network. A three-dimensional SRS 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.

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

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 (8). 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 combinatorial chemistry. These arrays are hybridized under stringent conditions with a complex sample representing mRNAs expressed in the test cell or tissue (5). 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 (5, 16, 27).

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 (13, 23). 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 amounts of 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 classic and modern statistical clustering techniques, which group genes with similar expression patterns. It includes the methods and applications of hierarchical clustering (1, 10, 31) and self-organizing map (30, 33). 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 (35) 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 fea-
tures 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 (21, 36). 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 (ART) model. The SRS algorithm uses a three-dimensional (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 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 (24). In developing the SRS algorithm, we employed the 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 judgment 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 formulas 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 (5, 10) were chosen to validate the SRS algorithm.

**SRS algorithm.** 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 step is applied to replace missing or incorrect data. The filter-gene step is 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. Then 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.

In the SRS model, a 3D response surface as a function of a pair of genes is introduced to describe the biological relationship among the target and activator-repressor genes. Specifically, we use a piecewise linear-quadratic polynomial \(S(A,B)\)

\[
S(A,B) = \begin{cases} 
0 & A \leq a \land B \leq b \\
S_1(A,B) & A > a \lor B > b 
\end{cases}
\]

Fig. 1. The data processing flowchart of the smooth response surface (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 (ART). For each triplet of genes, a lack-of-fit formula is used to filter the triplets from the initial screening. Then a diagnostic strategy is applied to refine the selected triplets, and a score, which reflects the strength of the triplet interrelationship, is used to rank the refined triplets. Finally, a gene regulatory network is constructed based on the top scoring triplets.
as defined in Table 1, which generates a smooth response surface in the unit cube (Fig. 2). In the top left region of Fig. 2 (i.e., \(a < 0.5\) and \(b > 0.5\)), the activator \(A\) is small and the repressor \(B\) is large; therefore, the expression level of an ideal target should be small. In the bottom right region of Fig. 2 (i.e., \(a > 0.5\) and \(b < 0.5\)), the activator \(A\) is large and the repressor \(B\) is small; therefore, the expression level of an ideal target should be large. In the other two regions, the activator \(A\) and repressor \(B\) are close to each other; therefore, the expression level of a target should be close to medium. Ideally, the triplets that follow the ART relationship should lie closely to the response surface in Fig. 2. We observe that response functions other than \(S(A,B)\) given in Table 1 may be used to model the ART relationship. To describe different forms of regulatory functions such as “dominant” activators or repressors, different functions should be used. An example of such regulation in eukaryotic cells is the heterodimeric transcription factor AP1 that has Jun and Fos components. Fos cannot form homodimers, and Jun-Jun dimer binds the transcription factor AP1 that has Jun and Fos components. Ideally, the triplets that follow the ART relationship should lie closely to the response surface in Fig. 2. We observe that response functions other than \(S(A,B)\) given in Table 1 may be used to model the ART relationship. To describe different forms of regulatory functions such as “dominant” activators or repressors, different functions should be used. An example of such regulation in eukaryotic cells is the heterodimeric transcription factor AP1 that has Jun and Fos components. Fos cannot form homodimers, and Jun-Jun dimer binds the transcription factor AP1 that has Jun and Fos components.

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

<table>
<thead>
<tr>
<th>Computational parameters</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n)</td>
<td>number of samples (or time points)</td>
</tr>
<tr>
<td>Gene</td>
<td>(A = (a_1, \ldots, a_n))</td>
</tr>
<tr>
<td></td>
<td>(B = (b_1, \ldots, b_n))</td>
</tr>
<tr>
<td>Gene-triplet</td>
<td>(C = (c_1, \ldots, c_n))</td>
</tr>
<tr>
<td>((A,B,C))</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Computational functions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response surface function</td>
<td>(S(a,b) = \begin{cases} 2a(1-b), &amp; \text{if } 0 \leq a \leq 0.5 \text{ and } 0.5 \leq b \leq 1; \ 1-2(1-a)b, &amp; \text{if } 0.5 \leq a \leq 1 \text{ and } 0 \leq b \leq 0.5; \ a-b+0.5, &amp; \text{otherwise} \end{cases} )</td>
</tr>
<tr>
<td>Lack-of-fit function</td>
<td>(RT(A,B,C) = \frac{\sum_{i=1}^{n} (c_i - \hat{c}<em>i)^2}{\sum</em>{i=1}^{n} (c_i - \bar{c})^2} )</td>
</tr>
<tr>
<td>Diagnostic function</td>
<td>(\text{Diag}(A,B,C) = \frac{(\frac{1}{n} \sum_{i=1}^{n} [RT_{ij}(A,B,C) - RT(A,B,C)(1 - \text{Diag}(A,B,C))]^2)^{\frac{1}{2}}}{RT(A,B,C)} )</td>
</tr>
<tr>
<td>Score function</td>
<td>(\text{Score}(A,B,C) = RT(A,B,C)(1 + \text{Diag}(A,B,C)))</td>
</tr>
</tbody>
</table>

To find good triplets, the algorithm searches all possible triplet combinations. For each triplet \((A,B,C)\), the fitted value of a target gene \(C\) is given by \(\hat{C} = S(A,B)\). If the ART relationship is strong, then the residual, \(\hat{C} - C\), should be small. The residual sum of squares over all time points measures the overall variation in \(C\) that is not explained in

---

**Fig. 2.** A three-dimensional SRS. 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.
the response surface model. Then the lack-of-fit function $RT(A,B,C)$ defined in Table 1, which is the ratio of the residual sum of squares and the total sum of squares corrected for mean, 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 ART 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. The default value of $RT$ is 0.1.

After fitting the SRS model, a leave-one-out diagnostic strategy is applied to check the reliability of triplets. 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 triplet $(A,B,C)$ when the $i$-th time point (or the $i$-th column) is left out, will differ greatly from $RT(A,B,C)$. $Diag_i(A,B,C)$ defined in Table 1 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.

The above two steps lead to the criteria for selecting triplet candidates: $RT(A,B,C) \leq RT$ and $Diag_i(A,B,C) \leq Diag$, where $RT$ and $Diag$ are constants as specified by users. The final number of triplets identified relies on the choice of these parameters. A larger $RT$ and/or $Diag$ value would generate more triplets. In particular, $RT = 0.1$ and $Diag = 2$ are used in the validation study below.

A final score is defined to measure the strength of the triplet interrelationship. Score $(A,B,C)$ defined in Table 1 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 $Diag(A,B,C)$ value. 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 $C$. Finally, a gene regulatory network is constructed based on the scoring triplets.

A US patent application is filed on the SRS algorithm, and the series number is 60/297299. Copies of the program are available upon request, from the authors.

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 (5). The data were collected at 17 time points for 6,457 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. For each time point, the lowest 5% values are replaced by the 5th percentile over all genes at that particular time point. The 5th 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 $1,514 \times 1,513 \times 1,512 \approx 3.5 \times 10^8$ possible triplets, whose $RT(A,B,C)$ values were found to be less than 0.1. Figure 3 shows the best nine fitting triplets from the initial screening. However, the low lack-of-fit values, except for the first triplet, are caused by the extreme values at the 90-min 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.

To evaluate the algorithm, the best scoring triplets were examined to see whether they make biological sense. A complete table of all the triplets can be obtained from the authors. Figure 5 shows a connected network of all triplets that have known functions. 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, which 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 (15, 25, 29, 32, 34).

Another major node in this predicted network highlights HAP1. The transcription factor HAP1 has been shown to modulate the nuclear encoding cytochrome gene CYC1 and CYC7, the aerobic and hypoxic isoforms of cytochrome $c$, respectively (4, 7, 17, 20, 26). HAP1 induces CYC1 and CYC7 under aerobic conditions (4, 14, 20, 38). However, the hypoxic gene CYC7 can be made at very low oxygen concentrations (4, 12, 14, 17). Our prediction suggests that HAP1 represses CYC7, which might implicate that the cells used in this data set were primarily grown under anaerobic conditions. Further experiments are needed to test the prediction and reconcile this with the experimental results.

We also explored the biological significance of a newly identified relationship. The highest-ranking triplet in Table 2 contains three gene products: TEC1 as an activator, PDS1 as a repressor, and YGP1 as the target. TEC1 is a transcriptional factor and has been implicated in filamentation and agar-invasive growth (22). It may act as a positive modulator of YGP1, because YGP1 gene product is involved in cellular adaptations prior to stationary phase and response to nutrient depletion (9). PDS1, the predicted negative regulator is an inhibitor of cytokinesis, DNA replication, and anaphase (37). It is conceivable that overexpression of PDS1 would promote cell cycle progression and thus repress expression of the stationary phase
Fig. 3. The best fitting triplets without diagnostic. The x-axis is the time points in minutes, and the y-axis is log-transformed (base 2) concentration values of the triplets, where long-dashed line is activator A, short-dashed line is repressor B, and solid line is target C. 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-min time point. These triplets are not reliable which are indicated by the large Diag values.

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>ABC</th>
<th>RT(A,B,C)</th>
<th>Diag(A,B,C)</th>
<th>Score(A,B,C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1179</td>
<td>TEC1 PDS1 YGP1</td>
<td>0.0553</td>
<td>0.48307</td>
<td>0.08201</td>
</tr>
<tr>
<td>1274</td>
<td>GAP1 MSB1 ARE2</td>
<td>0.06608</td>
<td>0.25991</td>
<td>0.08325</td>
</tr>
<tr>
<td>1339</td>
<td>PI3 RNP1 CPS1</td>
<td>0.07207</td>
<td>0.16708</td>
<td>0.08411</td>
</tr>
<tr>
<td>1340</td>
<td>HPR5 GAP1 HAP1</td>
<td>0.07227</td>
<td>0.16418</td>
<td>0.08414</td>
</tr>
<tr>
<td>1380</td>
<td>PIR3 RNP1 FAA1</td>
<td>0.07604</td>
<td>0.1136</td>
<td>0.08468</td>
</tr>
<tr>
<td>1480</td>
<td>RAD27 MSK1 HES1</td>
<td>0.06075</td>
<td>0.41396</td>
<td>0.08589</td>
</tr>
<tr>
<td>1612</td>
<td>SPT21 CBF2 GPA1</td>
<td>0.07278</td>
<td>0.19939</td>
<td>0.0873</td>
</tr>
<tr>
<td>1645</td>
<td>HES1 TWT2 HAP1</td>
<td>0.06965</td>
<td>0.25932</td>
<td>0.08771</td>
</tr>
<tr>
<td>1920</td>
<td>TIP1 MCR1 SPO13</td>
<td>0.07994</td>
<td>0.13201</td>
<td>0.0905</td>
</tr>
<tr>
<td>1947</td>
<td>HPR5 PEP5 HAP1</td>
<td>0.08003</td>
<td>0.13536</td>
<td>0.09086</td>
</tr>
<tr>
<td>2117</td>
<td>TIP1 AGP1 SPO13</td>
<td>0.08185</td>
<td>0.12926</td>
<td>0.09243</td>
</tr>
<tr>
<td>2226</td>
<td>INH1 CBF2 YGP1</td>
<td>0.07582</td>
<td>0.22926</td>
<td>0.09333</td>
</tr>
<tr>
<td>2243</td>
<td>CYB2 CIK1 TIP1</td>
<td>0.07185</td>
<td>0.30692</td>
<td>0.09347</td>
</tr>
<tr>
<td>2277</td>
<td>RAD27 CPS1 HES1</td>
<td>0.06435</td>
<td>0.45645</td>
<td>0.09373</td>
</tr>
<tr>
<td>2503</td>
<td>CYB2 MCR1 SPO13</td>
<td>0.06806</td>
<td>0.40383</td>
<td>0.09555</td>
</tr>
<tr>
<td>3107</td>
<td>HES1 GPD2 HAP1</td>
<td>0.08957</td>
<td>0.12041</td>
<td>0.10036</td>
</tr>
<tr>
<td>3122</td>
<td>KAR3 FAA1 HAP1</td>
<td>0.08402</td>
<td>0.19595</td>
<td>0.10048</td>
</tr>
<tr>
<td>3186</td>
<td>INH1 PDS1 YGP1</td>
<td>0.08022</td>
<td>0.25682</td>
<td>0.10082</td>
</tr>
<tr>
<td>3941</td>
<td>IPL1 TSM1 CBP2</td>
<td>0.09555</td>
<td>0.08449</td>
<td>0.10497</td>
</tr>
<tr>
<td>4273</td>
<td>SPO16 ASE1 SMC3</td>
<td>0.09026</td>
<td>0.18177</td>
<td>0.10666</td>
</tr>
</tbody>
</table>

Rank is the index of triplets sorted by score.
gene YGP1 to a basal level. Further experimental data is required to test such hypotheses.

The same data set was analyzed by fuzzy logic (35). 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 threefold change and maximum intensity $/H_{11350}$ were retained. A simple network of 6 genes was constructed from about 470,000 triplets ($0.007\%$ of all possible triplets). The analysis took about 200 h on an 8-processor SGI Origin 2000. In contrast, in the SRS approach, 1,451 genes (23%) that have at least threefold change and maximum intensity $/H_{11350}$ 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 4 h. Only a small fraction of triplets were identified by both approaches, although they started with a large number of common genes. In conclusion, the SRS algorithm is more conservative because less genes are used and less triplets are generated, more reliable because a leave-one-out diagnostic strategy is applied to remove unreliable triplets, and more efficient in computation because a compact mathematical formula replaces the complicated stepwise process of fuzzification, decision making, and defuzzification.

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 (10). 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. 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 that 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.

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, several predicted ART relationships suggested the gene products be involved in

Fig. 4. The ART mechanism. The $x$-axis is the time points in minutes, and the $y$-axis is log-transformed (base 2) concentration values of the triplets. The graphs show the putative biological relationship among the triplets. When the concentration of the activator $A$ (long-dashed line) is high and that of the repressor $B$ (short-dashed 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.
related cellular processes. For example, RSR1 is a RAS GTPase involved in bud site selection (6). The predicted RSR regulators include CDC45, POL2, DPB2, SWI5, and RPM2, and these proteins also function in budding and cell proliferation (3, 11, 19, 28, 39). ASF1 causes derepression of many silent chromosomal loci when overexpressed in a cell (18). 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.

when compared 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 are 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 min) whose dynamic range of the intensities is quite different from other time points. One strategy is to remove the entire time point from the data set (30). Removing the entire time point 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 test this observation, we apply the algorithm to a data set without the 90-min experiment and construct a new network. The new network is similar to Fig. 5. However, some parts of the network are missing. The network did 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, whereas 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. Second, the SRS algorithm is easier to implement because of its mathematical simplicity and compactness. The computational complexity of the SRS algorithm is $O(g^2n)$, where $g$ is the number of genes retained after filtering, and $n$ is the number of samples. The memory require-
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 generic and can be applied to other relationships and more complicated systems. Examples include other classes of relationships such as coactivators and corepressors 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


