**Abstract**—A fuzzy logic approach, called FuzzyTRN, to infer transcriptional regulatory networks (TRN) in *Saccharomyces cerevisiae* is proposed. FuzzyTRN predicts potential regulators and their target genes using sequences analysis on transcription factor binding sites (TFBS) of transcriptional factors (TF) and promoter region of target genes. Those potential regulators and target genes are used to form vertices in the TRN. Furthermore, multiple sets of microarray gene expression data (MGED) are used by FuzzyTRN to predict links in the TRN. FuzzyTRN predicts transcriptional interactions by recognizing expression patterns of genes. In this study, a number of confirmed genetic interactions are utilized to train FuzzyTRN. 112 indirect genetic interactions that were confirmed by quantitative real-time polymerase chain reaction (qRT-PCR) experiments, and 259 and 86 direct genetic interactions that were collected by TRANSFAC database and literature surveying, were used as training set in this work. A simulation that encompasses 170 TFs and 40 target genes has been conducted and checked against YEASTRACT database to evaluate the performance of the proposed algorithm.

---

**I. INTRODUCTION**

The importance of studying genetic interactions is that it can reveal gene functions, the nature of mutations, gene function compensations and redundancies, and protein interactions (Tong et al., 2004) [1]. With the emergence of modern biotechnologies, abundant amount of data provided by advanced DNA sequencing and microarray technologies are available for scientists to disclose the unknown mechanisms of various biological processes in living cells. However, the high-throughput wet-lab methodologies in the post-genomic era led us to a problem of exponential accumulation of biological data. Computational power and storage that are required to analyze and manage these data also exponential increased. In recently years, the developments in genomics and semiconductor have provided researchers with a number of whole-genome sequences and high-performance computing systems. Hence, analyzing and managing the biological data in whole-genome scale have become feasible.

With the abundant and different types of information produced by microarray technology, how to properly utilize available data in the analysis became an essential issue. Various approaches have been proposed to predict genetic interactions or transcriptional interactions. According to the type of data they used, many of them are based on sequence data, localization data, expression data, structural data, or orthologs information across different species (Shoemaker and Panchenko, 2007) [2].

For sequence based approaches, some of them aimed at groups of genes, and predict gene-gene interactions within the groups (Li et al., 2006) [3]. However, these approaches are not specifically designed to reveal transcriptional interactions. Mining of conserved motif sequences among transcriptional factors (TFs) is a good approach to find a common target gene (Fang et al., 2005; Neduva et al., 2005) [4, 5]. Nevertheless, the resulting motif sequences are still insufficient to be utilized in prediction of interactions since it is still not clear about how to use conserved motif sequences as a key feature in interaction prediction. Support vector machines have been applied using either amino acid triplet occurrences (Shen et al., 2007) [6] or sequence signature products (Martin et al., 2005) [7]; but they did not provide a direct way to obtain information about which parts of the sequence are important for interactions.

Since abundant gene expression data produced by microarray technology in recently years, prediction of gene regulatory networks becomes a hot topic in gene expression data analysis, which aimed to reveal the underlying network of genetic interactions from one or a set of microarray gene expression data (Hartemink, 2005; Basso et al., 2005). [8, 9] Most of them may be classified into three classes: graphical models, discrete variable models and continuous variable models. Graphical models (Schäfer and Strimmer, 2005) [10] depict genetic interactions through directed graphs to characterize the types of interactions. However, specificities of most of the graphical models are low in most cases. Discrete variable models discretize gene expressions into a few states. The dynamics of gene expressions may be perceived as transitions of finite states. Typical discrete variables models proposed are Boolean networks,
probabilistic Boolean networks and discrete Bayesian networks (Friedman et al., 2000) [11]. Continuous variable models characterize the expression of a gene or its change by a linear or nonlinear continuous function of the expression of other genes. The genetic interactions are frequently modeled by a first-order or a second-order differential (or difference) equation. Continuous variable models consist of two major types: continuous Bayesian networks (Hughes et al., 2000; Husmeier, 2003) [12–13], state-space models (Perrin et al., 2003; Beal et al., 2005) [14, 15], deterministic differential systems (Kimura et al., 2005; Chen and Aihara, 2002; de Jong, 2002) [16–18], and optimization techniques (Wang et al., 2006) [19]. Although each class of models has been shown to be informative for understanding gene-gene interactions, most of them simplified the regulatory mechanism in living cells into models which is unnecessary true, it leads the estimation to biased results due to model mis-specification. Some of them tried to solve this problem by cooperating with Bayesian theory (Perrin et al., 2003; Beal et al., 2005) [14–15], but there is no direct information about the latent factors. Another major drawback of some of these models is that they are usually computational overhead, which limited the analysis in a small group of genes, and hard to be applied to large scale genome analysis.

Here, we present a fuzzy logic based approach, called FuzzyTRN, which unifies analysis of sequence data, expression data, and experimental confirmed results to infer transcriptional regulatory networks (TRN). FuzzyTRN consists of a fuzzy logic classifier (TFBS-FIS) using sequence data to predict potential transcription factor binding sites (TFBS). The primary reason to use sequence information is that it is an ultimate way to understand transcriptional interactions (Levitt, 2007) [20], so the specificity of FuzzyTRN can be increased. Furthermore, FuzzyTRN also contains a neuro-fuzzy classifier using expression data (LINK-ANFIS). It is used to predict the types of the potential transcriptional interactions. By cooperating with experimental confirmed results that are produced in wet-lab, LINK-ANFIS of the proposed FuzzyTRN can learn the gene expression patterns from known interactions, and then predict similar interactions among genes of interest with higher accuracy.

The rest of this paper is organized as follows. Section II introduces the proposed FuzzyTRN. In Section III, detailed information about training data obtained by quantitative real-time polymerase chain reaction (qRT-PCR) experiments, TRANSFAC database (Matys et al., 2003) [21] and literature surveying are given. Also, FuzzyTRN was applied to integrated real microarray gene expression data (MGED) in yeast (Spellman et al., 1998) [22] to infer a TRN that consists of 170 TFs and 40 target genes. The prediction results were checked against known transcriptional interactions in YEASTRACT database (Teixeira et al., 2006) [23]. The performance of FuzzyTRN was compared with two latest approaches (Schäfer and Strimmer, 2005; Vu and Vohradsky, 2007) [10, 24] in contrast. Conclusions and discussions are provided at the final part of this paper.

II. THE PROPOSED ALGORITHM

A. Overview

The proposed FuzzyTRN consists of three major portions: identification of consensus sequence motif, TFBS-FIS for identification of potential regulator, and LINK-ANFIS for inference of transcriptional interaction. The overall procedure of FuzzyTRN is drawn in Fig. 1, and detailed information of each process is described in the following subsections.

B. Identification of Consensus Sequence Motif

As the first step in FuzzyTRN, we collect candidate sequence motifs of all known TFs from YEASTRACT database. Next, we used ClustalW (Thompson et al., 1994) [25] multiple sequence alignment algorithm to identify regions of similarity, and the template formed by the aligned sequences may be a consensus sequence motif that can be recognized by a given TF. An example of aligned candidate sequence motifs for ABF1 is given in Fig. 2.

In most of the sequence analysis studies, conventional nucleotide symbols {A, T, C, and G} are commonly used to accurately represent sequences of nucleotides. However, it may not be a good way to model the biological systems that are known to be constantly variable. In this work, we brought uncertainty symbols into FuzzyTRN to represent the patterns of consensus sequence motifs that allow various ambiguities.

---

[Fig. 1. Conceptual block diagram of the proposed FuzzyTRN.]

[Fig. 2. Aligned candidate sequence motifs of ABF1 using ClustalW.]
A list of symbols defined by IUPAC (International Union of Pure and Applied Chemistry) [26] one letter codes, as summarized in Table I, is used to describe the frequencies of nucleotides in the candidate sequence motifs. Let us suppose that we have \( n \) candidate sequence motifs of length \( k \) for a given TF:

\[
\begin{bmatrix}
  b_1^{(1)} & b_1^{(2)} & \cdots & b_1^{(k)} \\
  b_2^{(1)} & b_2^{(2)} & \cdots & b_2^{(k)} \\
  \vdots & \vdots & \ddots & \vdots \\
  b_n^{(1)} & b_n^{(2)} & \cdots & b_n^{(k)}
\end{bmatrix},
\]

where \( b_i^{(j)} \in \{ \text{symbols of nucleotides} \} \), \( 1 \leq i \leq n \), and \( 1 \leq j \leq k \). Let \( N(i) \) be a vector of the frequencies that

\[
N(j) = \left[ N(A,j) \ N(C,j) \ N(G,j) \ N(T,j) \right],
\]

where \( N(A,j) \), \( N(C,j) \), \( N(G,j) \), and \( N(T,j) \) are the summed frequencies of base \( A \), \( C \), \( G \), and \( T \) at position \( j \), respectively. This vector takes values from Table I. Let \( S(j) \) be the number of valid bases at position \( j \), where ‘-’ is not counted as a valid base.

Thus, suppose that the average probability of observing any base \( \{ A, C, G, T \} \) in whole-genome is assumed to likely equal to \( 1/4 \), using a maximum log-likelihood approach and the method of Lagrange multipliers, we can estimate the position-specific score:

\[
P(b, j) = \log_2 \left\{ \exp \left[ \frac{N(b,j) - \sum_{b_i \in \{ A, C, G, T \}} N(b_i,j)}{\exp \left[ \frac{1}{4} \right]} \right] \right\},
\]

where \( b \) is one of the bases \( \{ A, C, G, T \} \). In order to determine whether the frequency of a nucleotide and its position are highly correlated as measured, or are highly correlated due to lack of data, we can calculate a position-specific weighted score:

\[
R(b, j) = P(b, j) \frac{S(j)}{n}.
\]

The position-specific weighted score \( R(b, j) \) is maximal when both of the following conditions are satisfied: 1) the nucleotide \( b \) dominates the numbers of nucleotides at position \( j \) of candidate sequence motifs, and 2) all nucleotides at position \( j \) of candidate sequence motifs are valid. Hence, we calculate values of \( R(b, j) \) throughout all position \( j \) of candidate sequence motifs, a position-specific weighted matrix \( M \) for a given TF can be obtained:

\[
M = \begin{bmatrix}
  R(A,1) & R(A,2) & \cdots & R(A,k) \\
  R(C,1) & R(C,2) & \cdots & R(C,k) \\
  R(G,1) & R(G,2) & \cdots & R(G,k) \\
  R(T,1) & R(T,2) & \cdots & R(T,k)
\end{bmatrix}.
\]

For instance, the aligned candidate sequence motifs of ABF1 (as demonstrated in Fig. 2) can be transformed into a position-specific weighted matrix \( M \) as shown in Fig. 2. The saturation levels of colors in Fig. 3 represent the magnitudes of \( R(b, j) \) in \( M \). Then, \( M \) can be used as a template to identify potential TBFS in the promoter region of a given target gene.

C. Identification of Transcription Factor Binding Sites: (TFBS-FIS) Stage 1: Similarity Scoring

TFBS exist in the promoter region of a gene, and they can be recognized by proteins known as TF. Once TFs bind to the TFBS in the promoter sequence, the recruited RNA polymerase will start to synthesize the mRNA from the coding region of the gene. This is the first step to activate a gene. To fully activate the functions of a gene requires lots of complex post-processes, which are not in the scope of this study. The promoter regions discussed in this study is the 1000 bp fragments located in the 5'-upstream region of genes. The length 1000 bp was determined based on the knowledge that most of known TFBS are located within 1000 bp upstream region of genes in the higher eukaryotic cells.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Frequency (Probabilistic)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>12 (100%)</td>
<td>Adenine</td>
</tr>
<tr>
<td>C</td>
<td>12 (100%)</td>
<td>Cytosine</td>
</tr>
<tr>
<td>G</td>
<td>12 (100%)</td>
<td>Guanine</td>
</tr>
<tr>
<td>T</td>
<td>12 (100%)</td>
<td>Thymine</td>
</tr>
<tr>
<td>R</td>
<td>6 (50%)</td>
<td>Purine</td>
</tr>
<tr>
<td>Y</td>
<td>6 (50%)</td>
<td>Pyrimidine</td>
</tr>
<tr>
<td>K</td>
<td>6 (50%)</td>
<td>Keto group</td>
</tr>
<tr>
<td>M</td>
<td>6 (50%)</td>
<td>Amino group</td>
</tr>
<tr>
<td>S</td>
<td>6 (50%)</td>
<td>Strong hydrogen binding</td>
</tr>
<tr>
<td>W</td>
<td>6 (50%)</td>
<td>Weak hydrogen binding</td>
</tr>
<tr>
<td>B</td>
<td>6 (33%)</td>
<td>Not adenine</td>
</tr>
<tr>
<td>D</td>
<td>4 (33%)</td>
<td>Not cytosine</td>
</tr>
<tr>
<td>H</td>
<td>4 (33%)</td>
<td>Not guanine</td>
</tr>
<tr>
<td>V</td>
<td>4 (33%)</td>
<td>Not thymine</td>
</tr>
<tr>
<td>N</td>
<td>3 (25%)</td>
<td>Any nucleotide</td>
</tr>
<tr>
<td>−</td>
<td>0 (0%)</td>
<td>No nucleotide exists</td>
</tr>
</tbody>
</table>

![Fig. 3. Visualization of the position-specific weighted matrix M that was obtained by the candidate sequence motifs of ABF1.](image)
A fuzzy logic classifier, called TFBS-FIS, is designed for the purpose of identification of potential TFBS in the promoter region of a given target gene. Before we jump into the TFBS-FIS, we introduce a feature vector that contains three characteristic values: 1) maximum likelihood, 2) proportion of matched nucleotides, and 3) p-value of the previous proportion. The feature vector is designed to serve as input of the TFBS-FIS.

For a given target gene, let us suppose that we have a fragment of its upstream sequence length $t$:

$$u^{(1)} \ldots u^{(2)} u^{(3)},$$

where $u^{(j)} \in \{A, C, G, T\}$, and $1 \leq j \leq t$. Then, we use the position-specific weighted matrix $M$ of a TF to find the region of maximum likelihood in the upstream sequence of a target gene by calculating:

$$L(i: k | M) = \frac{\sum_{j=1}^{k} R(u^{(i+j-1)}, j)}{\sum_{j=1}^{k} \max_{b \in \{A, C, G, T\}} R(b, j)},$$

where $-t \leq i \leq (-k-1)$. This equation is meant to measure the similarity between the consensus sequence motif of a TF and the sequence fragment from $i$ bp to $(i + k - 1)$ bp in the upstream region of a target gene. For all $i$, the maximum likelihood in the feature vector is $max L(i: k | M)$.

The proportion of matched nucleotides is the second characteristic value in the feature vector. The sequence fragment that produces the maximum likelihood is used in this step. First, we calculate the number of nucleotides in the sequence fragment that have non-negative position-specific weighted score $R(b, j)$ by

$$I_{NN}(i_{max}, j | M) = \begin{cases} 1 & \text{if } R(u^{(i_{max} + j-1)}, j) \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $i_{max} = \arg \max_{-t \leq i \leq (-k-1)} L(i: k | M)$.

The notation $I_{NN}(i_{max}, j | M)$ denotes whether the nucleotide $u^{(i_{max} + j-1)}$ is corresponded to a non-negative $R(u^{(i_{max} + j-1)}, j)$ in a given $M$ or not. Since the number of valid nucleotides at position $j$ of the aligned candidate sequence motifs represents the reliabilities of values at $j$th column in $M$, the proportion of matched nucleotides is the weighted sum of $I_{NN}(i_{max}, j | M)$ throughout all $j$:

$$P_M(i_{max} | M) = \frac{\sum_{j=1}^{k} S(j) \cdot I_{NN}(i_{max}, j | M)}{\sum_{j=1}^{k} S(j)}.$$  

Nevertheless, short consensus sequence motif usually yields large values in $max L(i: k | M)$ and $P_M(i_{max} | M)$ due to the possibility to get a random match is high. More specifically, all bound regions are assumed to equally likely have the sequence motifs, and the motifs are assumed to reside at any position of bound regions with equal probability. For that reason, we need to test the hypothesis:

$$H_0: \bar{P}_M(i | M) = P_M(i_{max} | M),$$  

$$H_1: \bar{P}_M(i | M) \neq P_M(i_{max} | M),$$

where $\bar{P}_M(i | M)$ is the average proportion of matched nucleotides, and $-t \leq i \leq (-k-1)$. The standard score of $P_M(i_{max} | M)$ is calculated by

$$Z(i_{max} | M) = \frac{\bar{P}_M(i | M) - P_M(i_{max} | M)}{\sigma_{P_M(i_{max} | M)}/\sqrt{k-1}},$$

where $\sigma_{P_M(i_{max} | M)}$ is the standard deviation of $P_M(i | M)$. We can obtain the p-value of $Z(i_{max} | M)$ by computing the cumulative distribution function for normal distribution by

$$P(Z(i_{max} | M) = \text{cdf}(Z(i_{max} | M), N(0,1)).$$

where $N(0,1)$ is normal distribution with mean of zero and variance of one. Thus, for any possible gene pair of TF and target gene, a feature vector that consists of three characteristic values is now available and applied in the second stage of TFBS-FIS.

D. Identification of Transcription Factor Binding Sites: (TFBS-FIS) Stage 2: Fuzzy Logic Classifier

In the lack of precise mathematical model to determine the degree of confidence for a obtained feature vector, fuzzy inference system (FIS), which was introduced in Zadeh (1965) [27], is a good tool to solve the problem. Fuzzy logic inference is a process of mapping a given crisp quantitative input to an output using fuzzy logic. The process of fuzzy inference involves: membership functions, fuzzifier, fuzzy rules base, fuzzy inference engine, and defuzzifier. There are two types of FIS, Mamdani-type (Mamdani, 1974) [28], and Sugeno-type (Takagi, 1974) [29]. In this work, we use Mamdani-type FIS, called TFBS-FIS, to identify TFBS in the upstream sequence of a given target gene.

The membership function is the mathematical function which defines the degree of membership of a crisp input in a linguistic fuzzy set. Since we have three characteristics values in a feature vector, an individual set of membership functions is required for each one of them.

The membership functions associated to the maximum likelihood $max L(i: k | M)$ are built on the Gaussian distribution curves with means 1 and 0, and standard deviations 0.15 and 0.3, respectively.
In the same formulation, we can define the membership functions for mapping the proportion of matched nucleotides \( P_M(i_{\text{max}} \mid M) \) into fuzzy sets by:

\[
\begin{align*}
MF_{lh}(x) &= e^{\frac{x-1}{2(0.15)}} \quad \text{for high-likelihood,} \\
MF_{ll}(x) &= e^{\frac{x-1}{2(0.3)}} \quad \text{for low-likelihood.}
\end{align*}
\]  

(14)

The membership functions of the fuzzy sets in (14) and (15) are depicted together in Fig. 4. In order to increase the specificity of the TFBS-FIS, the degrees of spreading of \( MF_{lh} \) and \( MF_{lh} \) are larger than that of \( MF_{ll} \) and \( MF_{ll} \), so the FIS can distinguish the given set of \( L(i : k \mid M) \) and \( P_M(i_{\text{max}} \mid M) \) into correct fuzzy sets.

The membership function for fuzzifying \( P_v(i_{\text{max}} \mid M) \) into degree of confidence is formulated by a sigmoid function:

\[
MF_{pv}(x) = \frac{1}{1 + e^{500(x-0.01)}} \quad \text{for significance.} 
\]  

(16)

where sign parameter is -500 and shift parameter is 0.01. The membership function \( MF_{pv} \) is depicted in Fig. 5, which helps us to conduct the test in (11) at level of significance 99% using fuzzy logic.

The output of the TFBS-FIS determines the degree of confidence that the upstream sequence of a target gene contains TFBS that can be recognized by a given TF. Hence, the membership functions for the fuzzy sets associated to the output are built on Gaussian distribution curves with means 0 and 1, and standard deviations 0.3 as follow:

\[
\begin{align*}
MF_{c}(x) &= e^{\frac{x-1}{2(0.3)}} \quad \text{for TFBS,} \\
MF_{n}(x) &= e^{\frac{x-1}{2(0.3)}} \quad \text{for Non-TFBS.}
\end{align*}
\]  

(17)

The membership functions \( MF_{c} \) and \( MF_{n} \) are drawn in Fig. 6. We can see that the larger output represents the higher confidence to make sure that a target gene contains TFBS that can be recognized by a given TF.

A set of fuzzy inference rules is required to perform the fuzzy logical operations. The fuzzy inference rules are implemented by “and” and “or” operations, which are the minimum and maximum functions, respectively. The Mamdani method was chosen as the defuzzification procedures, which means that the fuzzy sets to the input data were joined through the add function, and the output of the system was then computed as the centroid of the output membership functions. The established inference rules are defined as:

- High-likelihood and high-match and significant \( \rightarrow \) TFBS
- Low-likelihood or low-match or (not) significant \( \rightarrow \) Non-TFBS

Hence, a complete TFBS-FIS has been established, and the architecture of the TFBS-FIS is as shown in Fig. 7. We can apply any possible pair of TF and target gene to the TFBS-FIS to identify TFBS in the upstream sequence of the target gene. If the inference output is larger than 0.9, microarray gene expression profiles of the TF and the target gene are then applied to an LINK-ANFIS to infer the type of the transcriptional interaction between the TF and the target gene.
E. Inference of Transcriptional Interactions using ANFIS (LINK-ANFIS) Stage 1: Feature Extraction

The feature extraction portion of this study was inspired by observing the patterns in the plots of the gene expression levels of paired genes. In this study, two types of interactions, called activator-target (AT) interactions and repressor-target (RT) interactions (Lesage et al., 2004; Kafri et al., 2005; Wong and Roth, 2005 [30–32]), are investigated. In qRT-PCR experiments, a RT interaction is defined as a phenomenon that after a TF’s loss occurs, the expression quantity for its compensatory gene increases. Besides RT, some cases show that following a TF’s absence, the expression quantity of its compensatory gene decreased, and we call this phenomenon AT. Two types of patterns, complementary pattern (CP) and similar pattern (SP), were found to have association with RT and AT interactions, respectively. Specifically, when the expression of an enhancer gene increases (decreases), the expression of its related target gene decreases (increases), and this type of gene’s expression pattern is called as CP. On the other hand, if the expression of an enhancer gene increases (decreases) followed by the expression of its target gene decreases (increases), this pattern is called as SP. A test was conducted to support that CP and SP are associated to RT and AT interactions, respectively. For detailed information, please refer to (Chuang et al., 2008) [33].

For gene expression curves of SP, their slopes and curvatures tend to have same signs during the experimental period. On the other hand, for gene expression curves of CP, their slopes and curvatures tend to have different signs for the whole experimental time course. To capture characteristics of SP and CP of expression curves, we calculate the products of their first and second derivatives with respect to time as follows:

\[ E_{i,j}^{(D_1)} = \frac{\partial G_{i}(t)}{\partial t} \cdot \frac{\partial G_{j}(t)}{\partial t} , \]
\[ E_{i,j}^{(D_2)} = \frac{\partial^2 G_{i}(t)}{\partial t^2} \cdot \frac{\partial^2 G_{j}(t)}{\partial t^2} , \]

where \( t' = t + 1 \), \( G_{i}(t') \) is the lag-1 gene expression level of target gene \( G_{i} \) at time point \( t' \), and \( G_{j}(t) \) is the expression level of regulating gene \( j \) at time point \( t \). Similar patterns of AT pairs will result in positive values of \( E_{i,j}^{(D_1)} \) and \( E_{i,j}^{(D_2)} \). On the contrary, \( E_{i,j}^{(D_1)} \) and \( E_{i,j}^{(D_2)} \) of RT pairs will be negative or small positive values. Please note that there is a time lag in the expression of a target gene behind its partner gene. The reason is that from time course of gene expression data, most of co-expressed genes do not actually regulate each other (Ji and Tan, 2005) [34]. Therefore, observing gene expression data taken with time lags of one or two data point might be useful to analyze the causal relationships in which the expression behavior of one gene leads to a delayed pattern of expression of another (Reis et al., 2000) [35].

Time course gene expression data are discrete in time, so the partial derivatives in Equations (2) and (3) are discretized into

\[ \frac{\partial G_{i}(t)}{\partial t} = \frac{G_{i}(t+1) - G_{i}(t)}{\Delta t} \]  

and

\[ \frac{\partial^2 G_{i}(t)}{\partial t^2} = \frac{G_{i}(t+2) - 2G_{i}(t+1) + G_{i}(t)}{(\Delta t)^2} - \frac{G_{i}(t+1) - G_{i}(t)}{\Delta t} , \]

where \( t + 1 \) denotes the \((t + 1)th\) time point and \( \Delta(t + 1) \) is the time interval between time points \( t + 1 \) and \( t + 2 \). When all time intervals are equal, then

\[ \frac{\partial^2 G_{i}(t)}{\partial t^2} = \frac{G_{i}(t+2) - 2G_{i}(t+1) + G_{i}(t)}{(\Delta t)^2} \]

Furthermore, correlation coefficient between \( G_{i}(t') \) and \( G_{i} \) is also served and one of the features in the MGED of paired genes. Based on the theory of signal processing, a set of nonlinear signals can be treated as a combination of linear ones if they are divided into smaller fragments. Therefore, we used a shifting window to extract fragments of MGED of paired genes, and measure their correlation coefficient. The correlation \( E_{i,j}^{(corr)} \) between the \( i \)th target gene and the \( j \)th regulating gene can be formulated as below:
\[ E_{i,j}^{\text{Corr}} = \gamma \left( G_i(t': t' + \kappa), G_j(t: t + \kappa) \right) \] (23)

where \( \gamma \) is the function to compute the correlation coefficient, and \( \kappa \) is the size of the shifting window. If the paired genes express in SP, \( E_{i,j}^{\text{Corr}} \) will results in positive value. On the other hand, \( E_{i,j}^{\text{Corr}} \) will be a negative value if the expression for paired genes belongs to CP category.

Finally, three primary features of expression curves for paired genes, \( E_{i,j}^{\text{D1}}, E_{i,j}^{\text{D2}}, \) and \( E_{i,j}^{\text{Corr}} \), are fed into an ANFIS, called LINK-ANFIS, to infer genetic interactions.

F. Inference of Transcriptional Interactions using ANFIS (LINK-ANFIS) Stage 2: Multilayer ANFIS

One of our previous works, called GeneCFE-ANFIS (Chuang et al., 2008) [33], based on ANFIS were proposed to infer genetic interactions. In this work, we extend this algorithm by integrating multiple sets of MGED measured from biological microarray experiments under different experimental conditions. Therefore, we can make use of any possible strength in all available MGED to infer transcriptional interactions. The architecture of the LINK-ANFIS is shown in Fig. 8. We use the LINK-ANFIS to predict two types of interactions between TF and its target gene, which are AT interaction, and RT interaction.

For instance, if we are looking for the type of interaction gene \( G_i \) acts on gene \( G_j \), the processed gene expression data of \( G_i \) and \( G_j \) are used to extract \( E_{i,j}^{\text{D1}}, E_{i,j}^{\text{D2}}, \) and \( E_{i,j}^{\text{Corr}} \), and these features are then separately input to three front-end ANFIS. Each front-end ANFIS has one input, and 5 fuzzy membership functions (functions for strong-AT, weak-AT, no interactions, weak-RT and strong-RT) are utilized to fuzzify the crisp values (\( E_{i,j}^{\text{D1}}, E_{i,j}^{\text{D2}}, \) or \( E_{i,j}^{\text{Corr}} \)) into quantifiers in fuzzy space. The fuzzy rule-base of each ANFIS is formed by 25 rules (5 \( \times \) 5), and preliminary inference of front-end ANFIS is obtained by summing up all responses produced by each rule with corresponding weighting factors. The weighting factors can be determined by training the ANFIS using the known genetic interactions that have been confirmed by qRT-PCR experiments or literatures.

The preliminary inferences yielded by front-end ANFIS are then fed to a back-end ANFIS to infer type of interaction that \( G_i \) acts on \( G_j \). The back-end ANFIS has three inputs, and each input has 5 membership functions, which are the same setting as those front-end ANFIS have. However, since to the back-end ANFIS has three input variables, there are 75 rules (5 \( \times \) 5 \( \times \) 3) in the fuzzy rule-base of the back-end ANFIS. The fuzzy inference engine and the premise parameters (properties of membership functions, definition of rule bases) can be tuned by the training data set, and the thorough review of ANFIS training algorithm, we refer to Jang (1993) [36].

In order to distinguish significance results from all predictions, a set of cutoff values [\( \text{TH}_{RT}, \text{TH}_{AT} \)] can be determined by the output values of all training data. Suppose that the output values of all training data is \( O = [o_1, o_2, \ldots, o_{trn}] \), where \( o_i \) is the output value of \( i\text{th} \) training data, and \( trn \) is the number of training data. The cutoff values \( [\text{TH}_{RT}, \text{TH}_{AT}] = [-\sigma_O, \sigma_O] \), where \( \sigma_O \) is the zero mean standard deviation of all values in \( O \). For instance, for lag-1 gene pairs, we predict the interaction of \( G_i \) and \( G_j \) is AT interaction or RT interaction; if the output of the ANFIS computed with lag-1 is more extreme than the associated cutoff, e.g. larger than \( \sigma_O \) or smaller then -\( \sigma_O \), and no interaction if the output value falls in the interval of the two cutoffs.

Finally, the LINK-ANFIS that is trained by training data can be used to predict potential transcriptional interactions that are not-yet-confirmed by biological experiments.
III. EXPERIMENTAL RESULTS

A. Nucleotide Sequence Data

In this study, FuzzyTRN was applied to identify transcriptional regulators that control 40 cell-cycle regulated genes of *Saccharomyces cerevisiae*. These genes served as target genes in the simulations, and 1000 bp upstream sequences of them are retrieved from the database of yeast transcriptional regulators YEASTRACT (Teixeira et al., 2006) [23]. A pool of 170 TF candidates for potential regulators is collected. The candidate sequence motifs of these TFs are also collected from YEASTRACT.

B. Microarray Gene Expression Data

In this study, the proposed FuzzyTRN is applied to the cDNA MGED in Spellman et al. (1998) [22] to infer TC and TD pairs. Four data sets are integrated and utilized to train and test the proposed algorithm. For each experiment, experimental and control groups were mRNAs extracted from synchronized yeast cultured by alpha factor, cdc15, cdc28 and elutriation mutants, respectively. There are 18, 24, 17 and 14 sampling points in each of these experiments with no replicates. A full description of experimental protocol and complete datasets are available at http://cellecycle-www.stanford.edu.

In order to make use of all possible data, MGED of alpha, cdc15, cdc28, and elu are integrated by appending one after another, sequentially. Therefore, we can get an integrated MGED with 73 time points. More time points in MGED can help us to prevent the LINK-ANFIS from overfitting problems.

C. Training Data Sets for the LINK-ANFIS

The goal of the LINK-ANFIS in the FuzzyTRN is to predict transcriptional interactions by recognition of expression patterns of paired genes. Since LINK-ANFIS is a classifier with supervised learning, the proposed FuzzyTRN requires training before it becomes fully functional.

The training data encompasses 457 paired genes and their corresponding interaction types that were confirmed by experiments. Among all of the training data, 112 gene pairs were confirmed by quantitative real-time polymerase chain reaction (qRT-PCR) experiments, 259 were obtained from TRANSFAC database (Matys et al., 2003) [21], and 86 were collected by a series of surveys of 36 previously published literatures [37–72]. By using the training data and the integrated MGED, the LINK-ANFIS in the FuzzyTRN yielded 90.2% training accuracy after 500 iterations of training, and the cutoff values $[TH_{RT}, TH_{AT}] = [-0.8085, 0.8085]$ are determined to perform further predictions.

D. Inference of Regulators and Interactions

The proposed FuzzyTRN is applied to 40 yeast cell-cycle regulated target genes and 170 candidate regulators. First, FuzzyTRN calculates the position-specific weighted matrices of 170 candidate regulators, and then use them to identify consensus sequence motif in the upstream sequence of the 40 preselected target genes. 849 potential links are identified, and been applied to the LINK-ANFIS to predict the types of the transcriptional interactions (AT or RT). 442 out of 849 potential links are predicted to have significant transcriptional interactions. In order to compare the results of our algorithm, a Gaussian graphical model (GGM) (Schäfer and Strimmer, 2005) [10] and a nonlinear differential equations model (NDEM) (Vu and Vohradsy, 2007) [24] are also applied to predict links among 40 target genes and 170 candidate regulators. Results yielded by FuzzyTRN, GGM, and NDEM have been checked against experimental confirmed links annotated in YEASTRACT, and summarized in Table II. We can see that FuzzyTRN yielded more positive links than other approaches. Due to limited space, detailed information about genes involved in the simulation is available upon request from the authors.

IV. CONCLUSION

The goal of the proposed FuzzyTRN is to predict transcriptional regulatory interactions in a genetic network. We focus on comprehensive analysis of sequence data and microarray gene expression data with experimental confirmed links, and to recover unknown transcription regulations with strong confidences. By comparing with previously published algorithms, FuzzyTRN can correctly identify interactions between transcription factors and target genes with a higher accuracy. By cooperating with sequence data, FuzzyTRN is able to reduce possible false-positive predictions, which were usually high in algorithms that used solely microarray gene expression data. FuzzyTRN is also capable of handling large scale networks since the computational time required to train the LINK-ANFIS in FuzzyTRN is lesser than 3 minutes using a Pentium-M 1.86 GHz computer; and the rest portions of the FuzzyTRN require only linear computational time. FuzzyTRN yielded more promising results comparing with previously published algorithms that might be a useful tool to recover more potential transcriptional regulatory interactions that are not yet confirmed by biological experiments.

ACKNOWLEDGMENT

The authors thank Dr. Ting-Fang Wang at the Institute of Biological Chemistry, Academia Sinica for providing us with the qRT-PCR results, Mr. Yu-Bin Wang for constructive discussions, and Dr. Shang-Kai Tai, Mr. Chia-Chang Wang for collecting transcriptional interactions. This research was supported by NSC grants NSC 96-2628-E-002-252-MY3 and NSC 96-2218-E-002-015.

REFERENCES


[2] B.A. Shoemaker and A.R. Panchenko, “Deciphering Protein-Protein Interactions. Part II. Computational Methods to Predict Protein and
### TABLE II

**LIST OF TARGET GENES AND NUMBER OF PREDICTED REGULATORS WITH EVIDENCES FOUND IN YEASTRACT DATABASE**

<table>
<thead>
<tr>
<th>#</th>
<th>Target</th>
<th>Number of Correctly Predicted TFs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FusyTRN</td>
</tr>
<tr>
<td>1</td>
<td>ASH1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>CDC20</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>CDC46</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>CLB2</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>CLB4</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>CLB6</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>CLN2</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>CLN3</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>DIA1</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>DSE3</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>FBP26</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>FIT3</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>GLN1</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>GRE1</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>HCM1</td>
<td>3</td>
</tr>
<tr>
<td>16</td>
<td>HHT1</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>HSH155</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>HSH49</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>HTA1</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>HTB1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#</th>
<th>Target</th>
<th>Number of Correctly Predicted TFs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FusyTRN</td>
</tr>
<tr>
<td>21</td>
<td>MET14</td>
<td>3</td>
</tr>
<tr>
<td>22</td>
<td>MNN1</td>
<td>3</td>
</tr>
<tr>
<td>23</td>
<td>MSN1</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>NDD1</td>
<td>0</td>
</tr>
<tr>
<td>25</td>
<td>NDE1</td>
<td>0</td>
</tr>
<tr>
<td>26</td>
<td>PMI40</td>
<td>0</td>
</tr>
<tr>
<td>27</td>
<td>PR02</td>
<td>1</td>
</tr>
<tr>
<td>28</td>
<td>SIC1</td>
<td>0</td>
</tr>
<tr>
<td>29</td>
<td>SPI1</td>
<td>5</td>
</tr>
<tr>
<td>30</td>
<td>SPO12</td>
<td>2</td>
</tr>
<tr>
<td>31</td>
<td>SVS1</td>
<td>1</td>
</tr>
<tr>
<td>32</td>
<td>SWE1</td>
<td>2</td>
</tr>
<tr>
<td>33</td>
<td>SW14</td>
<td>1</td>
</tr>
<tr>
<td>34</td>
<td>SW15</td>
<td>4</td>
</tr>
<tr>
<td>35</td>
<td>YAL018C</td>
<td>0</td>
</tr>
<tr>
<td>36</td>
<td>YBR089W</td>
<td>0</td>
</tr>
<tr>
<td>37</td>
<td>YFR057W</td>
<td>0</td>
</tr>
<tr>
<td>38</td>
<td>YGR269W</td>
<td>0</td>
</tr>
<tr>
<td>39</td>
<td>YKL177W</td>
<td>1</td>
</tr>
<tr>
<td>40</td>
<td>ZIP1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Total Number of Links**

| | 49 | 8 | 34 |


