Inferring Gene Regulatory Networks From Expression Data by Discovering Fuzzy Dependency Relationships

Patrick C. H. Ma and Keith C. C. Chan

***Abstract*—For one to infer the structures of a gene regulatory network (GRN), it is important to identify, for each gene in the GRN, which other genes can affect its expression and how they can affect it. For this purpose, many algorithms have been developed to generate hypotheses about the presence or absence of interactions between genes. These algorithms, however, cannot be used to determine if a gene activates or inhibits another. To obtain such information to better infer GRN structures, we propose a fuzzy data mining technique here. By transforming quantitative expression values into linguistic terms, it defines a measure of fuzzy dependency among genes. Using such a measure, the technique is able to discover interesting fuzzy dependency relationships in noisy, high dimensional time series expression data so that it can not only determine if a gene is dependent on another but also if a gene is supposed to be activated or inhibited. In addition, the technique can also predict how a gene in an unseen sample (i.e., expression data that are not in the original database) would be affected by other genes in it and this makes statistical verification of the reliability of the discovered gene interactions easier. For evaluation, the proposed technique has been tested using real expression data and experimental results show that the use of fuzzy-logic based technique in gene expression data analysis can be quite effective.**

***Index Terms*—Bioinformatics, data mining, fuzzy logic, gene regulatory networks (GRNs).**

I. Introduction

LARGE-SCALE monitoring of gene expression such as DNA microarrays [1]–[4] is considered to be one of the most promising techniques for making the discovery of gene regulatory networks (GRNs) [5] feasible. A GRN [6] is a complex biological system in which a regulator binds to a target gene and acts as a complex input-output system for performing various cellular processes. Since the expression of the gene, which encodes the regulator, is also regulated by the functional products of some other genes, this forms many complicated regulatory interactions that constitute the structures of underlying GRNs.

Even though all steps in gene expression, from the DNA-RNA transcription step to post-translational modification of a protein, may be regulated, the majority of the known regulators that control gene expression work at the level of transcription. Better understanding of gene regulatory interactions at the transcriptional level may therefore lead to better understanding of how cellular processes are carried out to accommodate changes in different external environment (i.e., temperature, pH values, pressure, etc.) [7]. Unfortunately, since living cells contain thousands of genes with each interacting with the regulators encoded by one or more other genes, the task of inferring the structures of GRNs is very difficult. Besides the complicated gene regulatory mechanisms, there are two important challenges that face bioinformatics researchers [8]. The first challenge comes from the presence of noise inherent in the data. Sources of noise in gene expression data include experimental, measurement, reporting and other data processing errors. The second challenge comes from the sparseness of samples relative to the great number of genes.

In an attempt to discover GRNs from gene expression data, various algorithms have been developed. They include Boolean networks [9]–[12], Bayesian networks [13]–[15], differential equations [16]–[18], data mining [19]–[24], and fuzzy logic based algorithms [25]–[27]. These algorithms were developed only to generate hypotheses about the presence or absence of interactions between genes and since they cannot be used to predict how a gene could be affected by other genes, statistical verification of the hypothesis is difficult. To do so, laboratory tests often need to be carried out. In addition to this, existing algorithms do not normally make good use of the temporal patterns inherent in time series gene expression data and they also do not take into consideration the directionality of regulation.

To better infer GRN structures, we propose a fuzzy data mining technique here. By transforming quantitative expression values into linguistic terms, the proposed technique is able to uncover hidden fuzzy dependency relationships among genes using the proposed fuzzy interestingness measure. It can handle very noisy, high-dimensional time series gene expression data and can represent discovered fuzzy dependency relationships explicitly as "if a gene is highly expressed, its dependent gene (target gene) is then lowly expressed," etc. These discovered relationships can not only make hidden regularities more easily interpretable, it can also determine if a gene is supposed to be activated or inhibited and can be used to predict how a gene would be affected by other genes from an unseen sample (i.e., expression data that are not in the original database). The proposed technique has been tested with real expression data. Experimental results show that it can be relatively effective. The fuzzy dependency relationships discovered can reveal biologically meaningful gene regulatory relationships that could be used to infer underlying GRN structures.

Manuscript received November 17, 2005; revised May 2, 2006, August 10, 2006, and October 28, 2006. This work was supported in part by the Hong Kong Polytechnic University under Grant RG1E.

The authors are with the Department of Computing, the Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China (E-mail: cschma@comp.polyu.edu.hk, cskcchan@comp.polyu.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TFUZZ.2007.894969

The rest of the paper is organized as follows. In Section II, we provide an overview of existing algorithms for discovering GRNs from gene expression data. In Section III, the proposed technique is described in details. The effectiveness of this technique has been evaluated and compared through various experiments with real expression data. The experimental setup, together with the results, is discussed in Section IV. Finally, in Section V, we give a summary of this paper.

II. Literature Review

Boolean networks [9]–[12] have been used to infer underlying GRN structures. In a Boolean network, the state of a gene is represented by a Boolean variable (ON or OFF) and interactions between genes are represented by Boolean functions. Boolean networks require that a number of assumptions be made to simplify analysis. For example, the activation of a single gene is represented as a Boolean switch that can either be on or off. Unfortunately, the validity of these assumptions has been questioned by many researchers, especially those in the biological community. To these researchers, there is a perceived lack of connection between simulation results and empirically testable hypotheses [28].

Instead of Boolean networks, Bayesian networks [13]–[15] can also be used for GRN inferences. Bayesian network is a probabilistic model that describes the multivariate probability distribution of a set of genes whose interdependencies are known. A Bayesian network allows the conditional dependencies and independencies to be displayed by means of a directed acyclic graph. To construct Bayesian networks, one needs to discover the interdependencies among genes. To do so, one can use Bayesian statistics [29] to find the network structures and the corresponding model parameters that best describe the probability distribution for which a dataset is drawn. The goodness-of-fit of a network with respect to a dataset can be assessed by assigning a score based on a statistically motivated scoring function such as the Bayesian score [30]. However, this approach to the learning of network structures is a NP-hard problem, especially for high-dimensional data such as gene expression data. Another problem that needs to be tackled when using the Bayesian network approaches for gene expression data analysis is concerned with the effect of small sample sizes. This, together with high-dimensional data, can make the estimation of the many parameters required for a Bayesian network difficult, if not impossible [31].

GRNs can be discovered based on the biochemically inspired models, such as reaction kinetics [32]. This approach has the advantage that the models constructed can be most directly related to biological processes. Unfortunately, the types of reactions and their parameters are often unknown and at present, the data collected are not sufficient for regulatory networks to be understood at this level of details. As a result, some researchers used approximations to reaction-kinetic formulations to derive systems of coupled differential equations [16]–[18] to describe the time course of gene expression levels. However, the primary disadvantages with these approaches are that they are computationally intensive and they require large data samples for the estimation of numerous parameters.

Other than the above, data mining approaches [19]–[24] for such tasks as clustering and classification are also very popular when it comes to GRN-discovery. In [19] and [21], for example, different clustering algorithms have been used to group co-expressed genes into clusters according to how similar they are to each other. Co-expressed genes are genes that have similar transcriptional response to the external environment. Genes, if found to belong to be same cluster, may be co-expressed genes and this can be an indication that they might be co-regulated by the same regulatory mechanism. The key step in data clustering is the use of a measure of similarity between two genes. For such purpose, a number of measures can be used. For example, one can consider using Kendall or Spearman correlation measures that can capture non-linear correlation [8]. However, perhaps for its simplicity, the Pearson correlation coefficient, despite it being a linear estimator, is the most popular similarity measures used for gene expression data analysis [8]. Since such a measure can only determine if two genes have a significant linear relationship with each other, the regulatory relationships, such as which gene affects which other genes, cannot be discovered. For classification-based approaches, decision tree-based classifier [20], [22] is often used to infer the structures of GRNs. This classifier normally makes use of a greedy procedure to select genes (attributes) that yield, for example, the maximum information gain, to recursively partition a training data set. Based on such a procedure, a decision tree can be formed. The leaf nodes of a decision tree are used to represent classes of target genes. The advantages of such an approach are that it can identify gene interactions in an explicit manner, and also the discovered tree can be used to predict how a target gene would be affected by other genes from an unseen sample. However, due to the fact that the pruning methods which a decision-tree-based classifier adopts are based on hill-climbing approaches, important information can be overlooked [33].

It is noteworthy that, to handle continuous values, many existing approaches have to perform discretization. The crisp discretization procedure they normally rely on does not take into account values at the interval boundaries. These values may end up be assigned to different intervals even though they are very similar. This actually adds noise to the data and may result in important patterns being overlooked. In most cases, the resulting intervals cannot be meaningfully interpreted and are hard to understand. In light of the prowess of fuzzy logic [34]–[37] in dealing with the uncertainties arising from noisy and inexact data which are quite commonplace in expression data and also the patterns discovered are easily interpretable by human users, some fuzzy logic-based approaches have been proposed [25]–[27] to infer the structures of GRNs from gene expression data.

In [25], a fuzzy logic-based approach has been proposed to search gene expression data for regulatory triplets consisting of an activator, repressor and target gene. Gene expression levels are first converted into three different states (low, medium and high) to varying degrees based on a set of predefined membership functions. Genes are then paired into as activator-repressor pairs to determine the expression value of the target gene based on a set of fuzzy rules. These regulatory triplets are then ranked based on a residual score between the predicted and actual expression values and a variance score of the activator and repressor gene pair. The triplets with low residual score and low

variance score are then the most likely to exhibit the regulatory relationships. However, because of the nonlinear scalability of this approach, it only allows for one activator and one inhibitor for each gene, the rules generated therefore provide limited biologically meaningful insights and experimentally testable hypotheses. In [26], a data clustering method, which was used as a preprocessing step for discovering potential regulatory triplets in order to reduce computational complexity, has been proposed. As reported in [26], although the performance of this improved method is better than the previous one [25] in terms of computation time and robustness to noise, the pre-determination of a suitable number of clusters to form and the percentage of cluster combinations to keep remain difficult. If these parameters are not properly set, there is a chance that interesting gene interactions may be overlooked. The selection of appropriate parameters when using this approach may add an additional level of difficulty to the gene interaction discovery process as it requires many trials and errors. In addition to the above, a linear fuzzy logic-based approach with exhaustive search technique has also been proposed [27]. Even though some gene interactions discovered by this approach are consistent with the existing biological literature on GRNs, the number of possible inputs for each gene is limited. Also, this approach was only tested on a small set of genes (12 genes), its applicability to larger GRNs consisting of several hundreds of genes, is not guaranteed [27].

For some of the problems that face existing algorithms to be better tackled, we therefore propose a fuzzy data mining technique to infer GRNs.

III. The Proposed Fuzzy Data Mining Technique

A. Linguistic Variables and Terms

To describe the proposed technique, let us assume that we are given a set of gene expression data, G, consisting of M time series collected from experiments with M genes. Each of these M time series consists, in turn, of N data points collected under N different experimental conditions, $E_1, \ldots, E_{t-1}, E_t, E_{t+1}, \ldots, E_N$, carried out, one after the other, at N different time instances. The data set, G, can therefore be represented as $G = \{G_1, \ldots, G_j, \ldots, G_M\}$, where each G_j, $j = 1, \ldots, M$, represents a time series gene expression profile and G_j takes on the expression value, e_{tj}, under the experimental condition, E_t.

As discussed above, in order to minimize the impact of noisy data in the GRN inference process, we propose to represent this quantitative gene expression data in linguistic variables and terms using the concepts of fuzzy set. To do so, we let $L = \{L_1, \ldots, L_j, \ldots, L_M\}$ be a set of linguistic variables such that $L_j \epsilon L$ corresponds to $G_j \epsilon G$. For each quantitative attribute, G_j, we denote the domain of the attribute as $\mathrm{dom}(G_j) = [l_j, u_j] \subseteq \mathcal{R}$, where l_j and u_j represent the lower and upper bounds, respectively. In other words, the linguistic variable, L_j, that represents G_j, has to take on the linguistic terms defined in $\mathrm{dom}(G_j)$. The set of these linguistic terms is denoted as $T(L_j) = \{l_{jk} \mid k = 1, \ldots, s_j\}$, where l_{jk} is a linguistic term characterized by a fuzzy set, F_{jk} with membership function, $\mu_{F_{jk}}$, defined on $\mathrm{dom}(G_j)$ so that $\mu F_{jk} : \mathrm{dom} G_j \rightarrow [0, 1]$. Given the above notations, we represent the value of the linguistic variable, L_j, in E_t as $l_{jk}(e_{tj})$ and the corresponding degree of membership as $\mu_{F_{jk}}(e_{tj})$.

B. Discovery of Fuzzy Dependency Relationships Among Genes

Step 1) Mining for first-order fuzzy dependency relationships: This step is to determine, for each G_j (a target gene), which other genes it is dependent on. For gene regulatory relationships, if the expression of G_j is dependent on the expression of G_p, where $G_p \epsilon G$ and $p \neq j$, one would expect to observe a certain amount of time delay between two corresponding observations made on their expression values. For this reason, we detect if there exists interesting dependency relationships between $l_{jk}(e_{tj})$ of G_j in E_t and $l_{pq}(e_{(t-1)p})$ of G_p in E_{t-1}, where $q = 1, \ldots, s_p$. Since our interpretation of the linguistic terms is based on fuzzy set theory, the dependency relationships that are expressed in these terms are referred to as fuzzy dependency relationships.

Let $l_{pq}(e_{(t-1)p}) \rightarrow l_{jk}(e_{tj})$ be the fuzzy dependency relationship between the linguistic terms $l_{pq}(e_{(t-1)p})$ and $l_{jk}(e_{tj})$. Since this dependency relationship involves only one linguistic term in its antecedent, it is referred here as first-order fuzzy dependency relationship. Then, the observed total degree, $o(l_{pq}(e_{(t-1)p}) \rightarrow l_{jk}(e_{tj}))$, of the occurrences of this relationship is defined as follows:

$$o(l_{pq}(e_{(t-1)p}) \rightarrow l_{jk}(e_{tj})) = \sum_{t=2}^{N} \min(\mu_{F_{pq}}(e_{(t-1)p}), \mu_{F_{jk}}(e_{tj})). \quad (1)$$

To decide whether the fuzzy dependency relationship, $l_{pq}(e_{(t-1)p}) \rightarrow l_{jk}(e_{tj})$, is interesting, we determine whether the difference between the observed total degree, $o(l_{pq}(e_{(t-1)p}) \rightarrow l_{jk}(e_{tj}))$, and the expected total degree, $e(l_{pq}(e_{(t-1)p}) \rightarrow l_{jk}(e_{tj}))$, is statistically significant. To determine if this is the case, we can use the standardized residual [38] to scale the difference as shown in (2)

$$z_{pj} = \frac{o(l_{pq}(e_{(t-1)p}) \rightarrow l_{jk}(e_{tj})) - e(l_{pq}(e_{(t-1)p}) \rightarrow l_{jk}(e_{tj}))}{\sqrt{e(l_{pq}(e_{(t-1)p}) \rightarrow l_{jk}(e_{tj}))}} \quad (2)$$

where

$$e(l_{pq}(e_{(t-1)p}) \rightarrow l_{jk}(e_{tj})) = \frac{\sum_{t=2}^{N} \mu_{F_{pq}}(e_{(t-1)p})}{\sum_{t=2}^{N} \sum_{q=1}^{s_p} \mu_{F_{pq}}(e_{(t-1)p})} \times \frac{\sum_{t=2}^{N} \mu_{F_{jk}}(e_{tj})}{\sum_{t=2}^{N} \sum_{k=1}^{s_j} \mu_{F_{jk}}(e_{tj})} \times \sum_{t=2}^{N} \sum_{q=1}^{s_p} \sum_{k=1}^{s_j} \min(\mu_{F_{pq}}(e_{(t-1)p}), \mu_{F_{jk}}(e_{tj})). \quad (3)$$

As this statistic approximates the standard normal distribution only when the asymptotic variance of z_{pj} is close to one, it is, in practice, adjusted by its variance for a more precise analysis. The new test

statistic, which is called the adjusted residual, can be expressed as follows:

$$d_{pj} = \frac{z_{pj}}{\sqrt{v_{pj}}} \tag{4}$$

where v_{pj} is the maximum likelihood estimate of its asymptotic variance [38] and is defined as:

$$v_{pj} = \left(1 - \frac{\sum_{t=2}^{N} \mu_{F_{pq}}(e_{(t-1)p})}{\sum_{t=2}^{N}\sum_{q=1}^{s_p} \mu_{F_{pq}}(e_{(t-1)p})}\right) \times \left(1 - \frac{\sum_{t=2}^{N} \mu_{F_{jk}}(e_{tj})}{\sum_{t=2}^{N}\sum_{k=1}^{s_j} \mu_{F_{jk}}(e_{tj})}\right). \tag{5}$$

This statistic d_{pj} has an approximate standard normal distribution [39]–[42] and the fuzzy dependency relationship $l_{pq}(e_{(t-1)p}) \rightarrow l_{jk}(e_{tj})$ is interesting when the test statistic is statistically significant. In other words, if $d_{pj} > 1.96$ ((4)), we can conclude, with a confidence level of 95 percent, that $l_{jk}(e_{tj})$ of G_j in E_t is dependent on $l_{pq}(e_{(t-1)p})$ of G_p in E_{t-1}.

Step 2) Mining for high-order fuzzy dependency relationships: Since the expression of G_j can be dependent on more than one gene, i.e., a target gene may have multiple regulators, $l_{jk}(e_{tj})$ can be characterized by more than one linguistic term. For this to be discovered, let σ be a subset of integers so that $\sigma = \{p_1, \ldots, p_h\}$ where $\sigma \subseteq \{1, \ldots, M\}$ and $|\sigma| = h \geq 1$. We also let G_σ is a subset of G so that $G_\sigma = \{G_p \mid p \in \sigma\}$. Given any G_σ, it is associated with a set of linguistic terms, $T(L_\sigma) = \{l_{\sigma q} \mid q = 1, \ldots, s_\sigma = \prod_{p\in\sigma} s_p\}$ where $l_{\sigma q}$ is represented by a fuzzy set, $F_{\sigma q}$, so that $F_{\sigma q} = F_{p_1 q_1} \cap \cdots \cap F_{p_k q_k} \cap \cdots \cap F_{p_h q_h}, p_k \in \sigma, q_k \in s_{p_k}$. The degree to which E_{t-1} is characterized by the linguistic term $l_{\sigma q}(e_{(t-1)\sigma})$ is defined as follows:

$$\mu_{F_{\sigma q}}(e_{(t-1)\sigma}) = \min(\mu_{F_{p_1 q_1}}(e_{(t-1)p_1}), \ldots, \mu_{F_{p_h q_h}}(e_{(t-1)p_h})). \tag{6}$$

Based on the first-order dependency relationships discovered in Step 1), the proposed technique determines if there are interesting second-order dependency relationships in which the expression of G_j in E_t is dependent on the expression of two other genes in E_{t-1}. To determine if this is the case, it makes use of a heuristic in which the dependency relationship between $l_{\sigma q}(e_{(t-1)\sigma})$, where $G_\sigma = G_{p_1} \cup G_{p_2}$, and $l_{jk}(e_{tj})$ is tested to determine whether it is interesting only if the dependency relationship between $l_{p_1 q}(e_{(t-1)p_1})$ and $l_{jk}(e_{tj})$, and the dependency relationship between $l_{p_2 q}(e_{(t-1)p_2})$ and $l_{jk}(e_{tj})$ are previously found to be interesting. Similarly, the proposed technique determines if there are interesting third-order dependency relationships when all combinations of second-order dependency relationships are interesting. In general, the proposed technique tests if an nth-order dependency relationships are interesting when all its $(n-1)$th order dependency relationships are found to be interesting. By searching for high-order dependency relationships using such a technique, the proposed technique can effectively prevent an exhaustive search for all possible combinations of the linguistic terms.

C. Assignment of Weights to the Discovered Fuzzy Dependency Relationships

Since the fuzzy dependency relationship is not completely deterministic, the uncertainty associated with $l_{\sigma q}(e_{(t-1)\sigma}) \rightarrow l_{jk}(e_{tj})$ can be modeled with the confidence measure defined as $\Pr(l_{jk}(e_{tj}) \mid l_{\sigma q}(e_{(t-1)\sigma}))$. For the purpose of making use of $l_{\sigma q}(e_{(t-1)\sigma})$ to predict if L_j should take on $l_{jk}(e_{tj})$ in a future time point, we use a weight of evidence measure [43], $W(l_{\sigma q}(e_{(t-1)\sigma}) \rightarrow l_{jk}(e_{tj}))$ which is defined, in terms of the mutual information, $I(l_{jk}(e_{tj}) : l_{\sigma q}(e_{(t-1)\sigma}))$, as follows:

$$\begin{aligned} W(l_{\sigma q}(e_{(t-1)\sigma}) \rightarrow & l_{jk}(e_{tj})) \\ &= W\left(l_{jk}(e_{tj})/l_{jk'}(e_{tj}) \mid l_{\sigma q}(e_{(t-1)\sigma})\right) \\ &= I(l_{jk}(e_{tj}) : l_{\sigma q}(e_{(t-1)\sigma})) \\ &\quad - I(l_{jk'}(e_{tj}) : l_{\sigma q}(e_{(t-1)\sigma})) \end{aligned} \tag{7}$$

where

$$\begin{aligned} & I(l_{jk}(e_{tj}) : l_{\sigma q}(e_{(t-1)\sigma})) \\ &= \log \frac{\Pr(l_{jk}(e_{tj}) \mid l_{\sigma q}(e_{(t-1)\sigma}))}{\Pr(l_{jk}(e_{tj}))} \\ &= \log \frac{\sum_{t=2}^{N} \min(\mu_{F_{\sigma q}}(e_{(t-1)\sigma}), \mu_{F_{jk}}(e_{tj}))}{\sum_{t=2}^{N}\sum_{q=1}^{s_\sigma} \min(\mu_{F_{\sigma q}}(e_{(t-1)\sigma}), \mu_{F_{jk}}(e_{tj}))} \end{aligned} \tag{8}$$

and

$$\begin{aligned} & I(l_{jk'}(e_{tj}) : l_{\sigma q}(e_{(t-1)\sigma})) \\ &= \log \frac{\Pr(l_{jk'}(e_{tj}) \mid l_{\sigma q}(e_{(t-1)\sigma}))}{\Pr(l_{jk'}(e_{tj}))} \\ &= \log \frac{\sum_{t=2}^{N}\sum_{k'=1,k'\neq k}^{s_j} \min(\mu_{F_{\sigma q}}(e_{(t-1)\sigma}), \mu_{F_{jk'}}(e_{tj}))}{\sum_{t=2}^{N}\sum_{k'0=1,k'\neq k}^{s_j}\sum_{q=1}^{s_\sigma} \min(\mu_{F_{\sigma q}}(e_{(t-1)\sigma}), \mu_{F_{jk'}}(e_{tj}))}. \end{aligned} \tag{9}$$

The term $\Pr(l_{jk}(e_{tj}) \mid l_{\sigma q}(e_{(t-1)\sigma}))$ can be considered as being the probability $l_{jk}(e_{tj})$ is observed after $l_{\sigma q}(e_{(t-1)\sigma})$ is observed in the previous time point. $\Pr(l_{jk'}(e_{tj}) \mid l_{\sigma q}(e_{(t-1)\sigma}))$ can be considered as being the probability $l_{jk'}(e_{tj})$, where $k \neq k'$, is observed after $l_{\sigma q}(e_{(t-1)\sigma})$ is observed in the previous time point. $W(l_{\sigma q}(e_{(t-1)\sigma}) \rightarrow l_{jk}(e_{tj}))$ measures the amount of positive or negative evidence that is provided by $l_{\sigma q}(e_{(t-1)\sigma})$ supporting or refuting $l_{jk}(e_{tj})$ being observed in the next time

point. Since this measure is probabilistic, it can work effectively even when the data being dealt with contains incomplete, missing, or erroneous values.

D. Prediction Based on the Discovered Fuzzy Dependency Relationships

Given a set of time series expression data collected from a set of M' genes from an unseen sample (i.e., gene expression data that are not in the original database). This set of M' genes can be represented by $G' = \{G_{(1)}, \ldots, G_{(j)}, \ldots, G_{(M')}\}$, where $G' \subseteq G$ and $M' \leq M$. To predict the value of $L_{(j)}$ in E_t, the discovered fuzzy dependency relationships can be searched to see which other genes $G_{(j)}$ is dependent on. If the dependency relationship, $l_{(\sigma)q}(e_{(t-1)(\sigma)}) \rightarrow l_{(j)k}(e_{t(j)})$, which indicates that $l_{(j)k}(e_{t(j)})$ is dependent on $l_{(\sigma)q}(e_{(t-1)(\sigma)})$, is previously discovered in the original database, then we can conclude that there is some evidence supporting the value of $L_{(j)}$ in E_t is $l_{(j)k}(e_{t(j)})$ if value of $L_{(\sigma)}$ in E_{t-1} is $l_{(\sigma)q}(e_{(t-1)(\sigma)})$. Then, the weight of evidence, $W'(l_{(\sigma)q}(e_{(t-1)(\sigma)}) \rightarrow l_{(j)k}(e_{t(j)}))$, supports this assignment in the unseen sample can be defined as follows:

$$W'(l_{(\sigma)q}(e_{(t-1)(\sigma)}) \rightarrow l_{(j)k}(e_{t(j)})) = W(l_{\sigma q}(e_{(t-1)\sigma}) \rightarrow l_{jk}(e_{tj})) \times \mu_{F_{(\sigma)q}}(e_{(t-1)(\sigma)}). \quad (10)$$

Let β be the set that contains all the fuzzy dependency relationships (previously discovered in the original database) supporting the value of $L_{(j)}$ in E_t is $l_{(j)k}(e_{t(j)})$. To calculate the total weight of evidence measure, the following steps are performed.

1) Each of the highest order, nth, dependency relationships is searched first.
2) If it is found to be matched with the expression profile of the unseen sample, then it is put into a set, β'.
3) If matched, this relationship and its corresponding subsets (its lower order dependency relationships) are then removed from β. Otherwise, only this relationship is removed from β.
4) If there is another nth order relationship in β, repeat step 1) for this. Otherwise, repeat step 1 for other lower-order, $(n-1)$th, relationships.
5) The searching process is terminated until β is empty.

Then we can combine the evidences provided by the fuzzy dependency relationships in β' that support the value of $L_{(j)}$ in E_t is $l_{(j)k}(e_{t(j)})$ by computing a total weight of evidence measure as follows:

$$TW'(l_{(\sigma)q}(e_{(t-1)(\sigma)}) \rightarrow l_{(j)k}(e_{t(j)})) = \sum_{\sigma=1}^{|\beta'|} W'(l_{(\sigma)q}(e_{(t-1)(\sigma)}) \rightarrow l_{(j)k}(e_{t(j)})). \quad (11)$$

Therefore, the value of $L_{(j)}$ in E_t is assigned to $l_{(j)k}(e_{t(j)})$ if

$$TW'(l_{(\sigma)q}(e_{(t-1)(\sigma)}) \rightarrow l_{(j)k}(e_{t(j)})) = \mathrm{MAX}_{k=1}^{s_{(j)}} TW'(l_{(\sigma)q}(e_{(t-1)(\sigma)}) \rightarrow l_{(j)k}(e_{t(j)})).$$

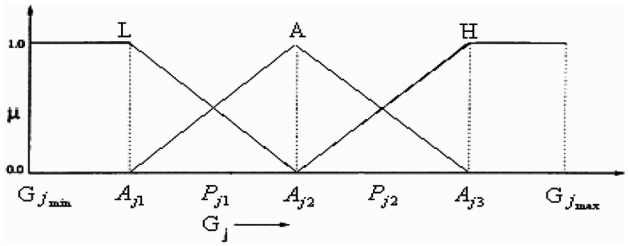

Fig. 1. Membership function.

IV. Experimental Results

The proposed technique has been tested with real expression data. In this section, we give the details of the experiments and discuss the results.

A. Experimental Data

For our experiments, we used a set of time series gene expression data that contains a series of gene expression measurements of the transcript (mRNA) levels of *S. cerevisiae* genes [44]. In this dataset, the biological samples were synchronized by three different methods: α factor arrest, arrest of a cdc15, and cdc28 temperature-sensitive mutant. Using periodicity and correlation algorithms, a total of about 800 genes that meet an objective minimum criterion for cell cycle regulation were identified [44]. Since, gene expression can be described in a finite number of different states such as "highly expressed" and "highly repressed," "upregulated," and "downregulated," "expressed," and "not expressed," or other different number of states [20], etc. For our application here, we define three different states, "highly expressed (H)," "averagely expressed (A)," and "lowly expressed (L)," in terms of three fuzzy sets as shown in Fig. 1.

For each quantitative attribute, G_j, where $j = 1, \ldots, M$, let $G_{j_{\max}}$ and $G_{j_{\min}}$ denote the maximum and minimum values that G_j take on under the N experimental conditions. Let the values of G_j be sorted in ascending order. Let P_{j1} be the value of G_j that exceeds one-third of the measurements and is less than the remaining two-thirds and P_{j2} be the value of G_j that exceeds two-thirds of the measurements and is less than the remaining one-third [45]. In order to determine P_{j1} and P_{j2}, the measurements was divided into a number of small class intervals, n_c, of equal width, δ, (i.e., $n_c = 10$), and counted the corresponding class frequencies, f_i, where $i = 1, 2, \ldots, n_c$. The position of the kth partition value, where $k = 1, 2$ for three partitions, is calculated as follows:

$$P_{jk} = \mathrm{low}_i + \frac{R_k - cf_{i-1}}{f_i} \times \delta \quad (12)$$

where low_i is the lower limit of the ith class interval, $R_k = N \times k/N_F$ is the rank of the kth partition value, N is the total number of records, N_F is the total number of fuzzy sets, and cf_{i-1} is the cumulative frequency of the immediately preceding class interval such that $cf_{i-1} < R_k < cf_i$. Then

$$A_{j1} = \frac{G_{j\min} + P_{j1}}{2}$$
$$A_{j2} = \frac{P_{j1} + P_{j2}}{2}$$
$$A_{j3} = \frac{P_{j2} + G_{j\max}}{2}.$$

$$\mu_L(e_{tj}) = \begin{cases} 1, & if \quad e_{tj} \le A_{j1} \\ \dfrac{A_{j2} - e_{tj}}{A_{j2} - A_{j1}}, & if \quad A_{j1} < e_{tj} < A_{j2} \\ 0, & otherwise \end{cases}$$

$$\mu_A(e_{tj}) = \begin{cases} 0, & if \quad e_{tj} \le A_{j1} \\ \dfrac{e_{tj} - A_{j1}}{A_{j2} - A_{j1}}, & if \quad A_{j1} < e_{tj} < A_{j2} \\ 1, & if \quad e_{tj} = A_{j2} \\ \dfrac{A_{j3} - e_{tj}}{A_{j3} - A_{j2}}, & if \quad A_{j2} < e_{tj} < A_{j3} \\ 0, & otherwise \end{cases}$$

$$\mu_H(e_{tj}) = \begin{cases} 0, & if \quad e_{tj} \le A_{j2} \\ \dfrac{e_{tj} - A_{j2}}{A_{j3} - A_{j2}}, & if \quad A_{j2} < e_{tj} < A_{j3} \\ 1, & otherwise \end{cases}$$

Fig. 2. Degree of membership.

Given the above representation, the degree of membership of a gene expression value, e_{tj}, of G_j in E_t to each fuzzy set can be computed as shown in Fig. 2.

B. Result

For our experiments, we chose the cdc15 data set for training. The data set has 24 experimental conditions and is relatively larger in size when compared to the others. The proposed technique was evaluated to determine its prediction accuracy in three different ways as suggested in [22]: i) using a ten-fold cross validation [46] strategy on the cdc15 data set for both training and testing, ii) using the alpha data set, which has 18 conditions, for testing, and iii) using the cdc28 data set, which has 17 conditions, for testing. During the training process, interesting fuzzy dependency relationships in the 800 cell cycle regulated genes were discovered. Based on the discovered dependency relationships, the states of each target gene (each of the 800 genes) in the testing sets were predicted. The predicted states were then compared with the original states of the genes and the average prediction accuracy was calculated.

For performance evaluation, we compared the prediction accuracy of the proposed technique with two well-known classification methods that have been applied to gene expression data analysis. The first one is a decision tree-based approach [22], and the second one is a support vector machine (SVM)-based approach [47]. To predict the states of the target gene with these approaches, the expression values of the target gene were divided into two different states as suggested in [22]: "expressed more than average (H)," and "expressed less than average (L)." The expression value, e_{tj}, of the target gene can then be mapped to H if $e_{tj} > \overline{e}_j$, and L if $e_{tj} \le \overline{e}_j$, where $\overline{e}_j$ is the average expression value of G_j under all the experimental conditions. According to [22], the decision tree-based approach of C4.5 (see Section II) was used and the discretization approach proposed in [48], [49] was used to discretize the expression values of all the genes, except the target gene, before C4.5 was used. For performance comparison against the SVM-based approach [47], we chose to use linear SVM as it is more commonly used for gene expression data analysis [50], [51]. Unlike C4.5, the original expression values of all the genes, except the target gene, were used to train the SVM without discretization. In addition C4.5 and SVM, we also compared the proposed technique with one of the popular fuzzy logic-based classification methods [52], [53] called FID [53]. FID combines symbolic decision trees with approximate reasoning offered by fuzzy representation. It extends non-fuzzy logic based classifiers such as C4.5 by using splitting criteria based on fuzzy restrictions and using different inference procedures to exploit fuzzy sets. For FID, the expression values of all the genes were transformed into the linguistic terms with different degree of memberships using the membership functions shown in Fig. 1.

In the experiments, we have tried to use different subsets of genes when trying to build classifiers with C4.5, SVM and FID. Specifically, we have tried to use five top-ranked genes, and then tried 10, 15, . . ., until we included all genes [8], [54], [55]. On average, we found that the highest prediction accuracy could be obtained when the top 50 ranked genes were used. Also, for feature ranking, the coefficients of the weight vector of a linear SVM were used for SVM [51] and the t-value [8] was used for both decision-tree based approaches. In Table I, the comparisons of the average prediction accuracy of different approaches are shown (for the cdc15 data set—standard deviation: proposed-4%, C4.5-4.9%, SVM-5.1%, and FID-4.6%). The statistics as shown in the table indicate that the proposed technique has higher prediction accuracy than the others for all training and testing sets. The FID approach is second to the proposed but better than the other two non-fuzzy approaches. This seems to indicate that the use of fuzzy modeling of gene expression data does indeed improve performance.

For the 800 cell cycle regulated genes, the number of possible gene-pairs that can be formed is over 600,000. Some of these gene-pairs will have important biological significance. One evaluation criteria for the effectiveness of these different approaches to GRN-inference is, therefore, for us to determine which of the discovered fuzzy dependency relationships in the cdc15 data set are already known to have regulatory relationships as reported in biomedical literature [56]–[59]. For this purpose, we compared the results of the proposed technique with those obtained by 5 different popular approaches: C4.5 [22], FID [53], fuzzy logic-based approaches [25] and [26], and non-network-based approach [24]. For both C4.5 and FID, the classification rules discovered for prediction accuracy evaluation were used. To prevent any genes that have known gene regulatory relationships from being eliminated from the feature selection process, we examined all the results obtained before and after feature extraction was applied. For the two fuzzy logic-based approaches (as discussed in Section II), the triplets with $r^2 < 0.015$ and variance ≤ 1.5 were selected as suggested in [25]. For the fuzzy approach proposed in [26], SOM was used as a pre-processing step and the number of clusters (12) was set to equal to half of the number of time points in the dataset (24 time points) as suggested in [26]. To avoid loss of biologically meaningful regulatory relationships, 100% of the cluster combinations were kept. For the non-network-based approach, we compared the proposed technique with the edge-detection approach [24]. This approach detects for gene

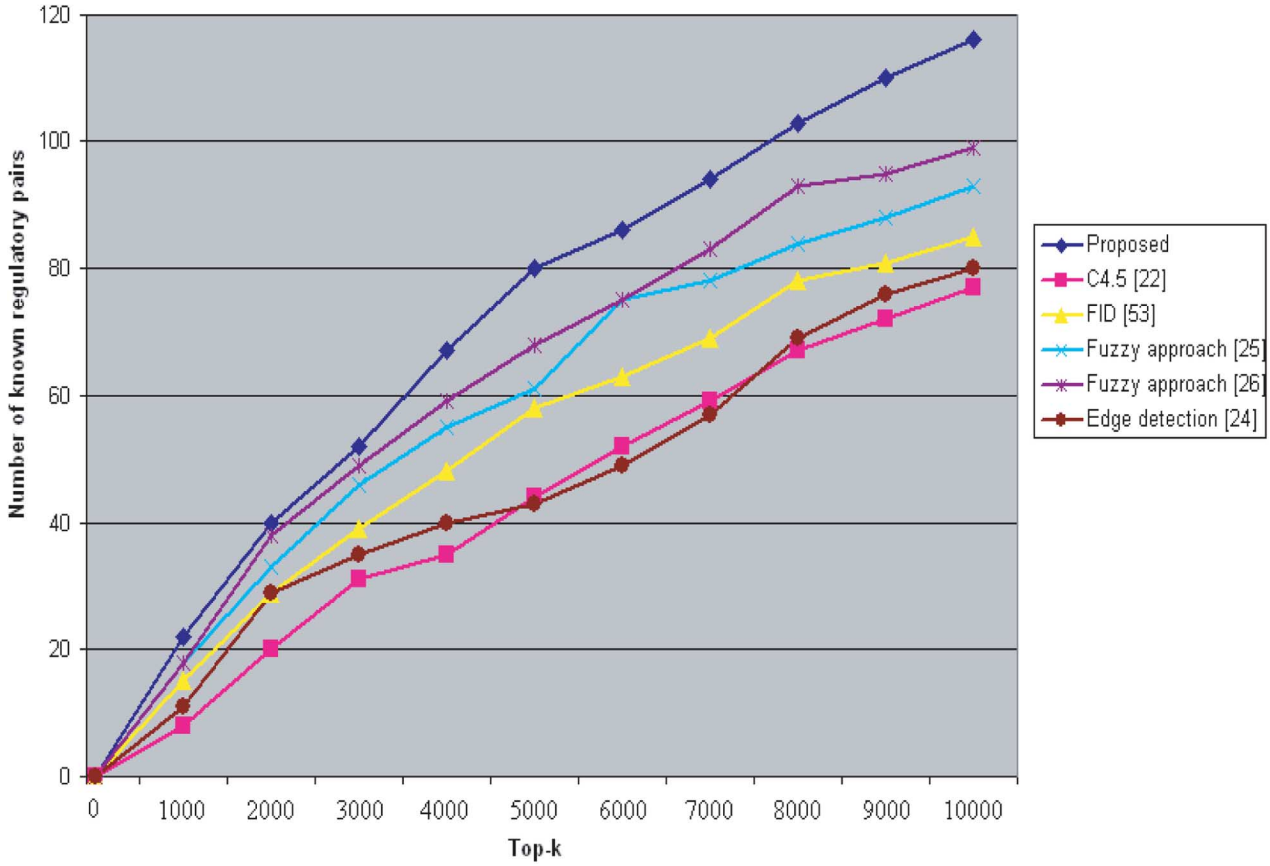

Fig. 3. Comparison of the number of known regulatory pairs found by different approaches (among 800 cell cycle regulated genes).

TABLE I
COMPARISON OF THE AVERAGE PREDICTION ACCURACY (800 CELL CYCLE REGULATED GENES)

	Proposed	C4.5	SVM	FID
cdc15	74%	55%	61%	69%
alpha	76%	60%	65%	67%
cdc28	69%	52%	58%	61%
average	73%	56%	61%	66%

pairs that have a high degree of similarity in their respective gene expression profiles. The similarity measure used is based on a correlation coefficient function. Since the approaches proposed in [24]–[26] were not developed to be used to predict gene expression in unseen samples, they cannot be compared and hence, they are not shown in Table I. Also, since SVM is a black-box approach, the discovered gene interactions cannot be explicitly revealed for possible biological interpretation.

In Fig. 3, we show the numbers of known regulatory relationships found from the top k-ranking interactions discovered by all these different approaches (k is allowed to vary from 0 to 10,000). As shown, the proposed technique can discover more known gene regulatory relationships than other approaches. In fact, all fuzzy-logic based approaches seem to have performed better than the non-fuzzy approaches. To further evaluate the proposed technique, the fuzzy dependency relationships discovered are presented in details here. A set of 16 genes, whose functional products are well-known and important in cell cycle regulation, was selected from the 800 cell cycle regulated genes to be presented here. They are listed in Table II.

TABLE II
SUMMARY OF 16 TARGET GENES SELECTED

Standard Name	Systematic Name	Function
CLN1	YMR199W	cyclin-dependent protein kinase regulator activity
CDC20	YGL116W	enzyme activator activity
CLB1	YGR108W	cyclin-dependent protein kinase regulator activity
CLB4	YLR210W	cyclin-dependent protein kinase regulator activity
HTA2	YBL003C	histone H2A
SWI4	YER111C	transcription factor activity
CLB5	YPR120C	cyclin-dependent protein kinase regulator activity
HTB2	YBL002W	histone H2B
CLN2	YPL256C	cyclin-dependent protein kinase regulator activity
CLB6	YGR109C	cyclin-dependent protein kinase regulator activity
HTA1	YDR225W	histone H2A
CLB2	YPR119W	cyclin-dependent protein kinase regulator activity
SIC1	YLR079W	kinase inhibitor activity
SWI5	YDR146C	transcriptional activator activity
HTB1	YDR224C	histone H2B
CLN3	YAL040C	cyclin-dependent protein kinase regulator activity

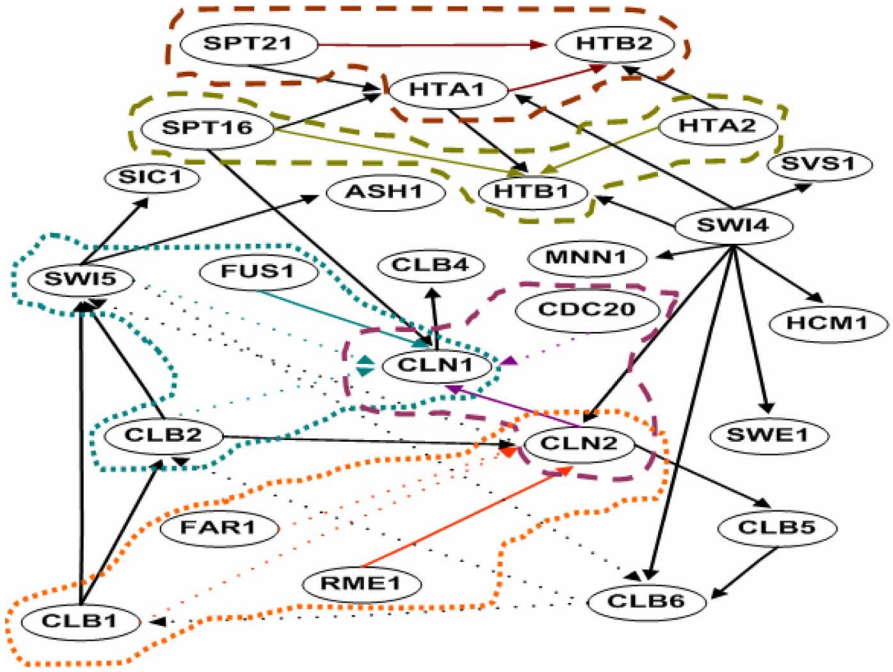

Fig. 4. Gene interaction diagram constructed by some biologically meaningful fuzzy dependency relationships discovered by the proposed technique. Solid lines correspond to activation relationships and broken lines correspond to inhibition relationships. All the discovered high-order fuzzy dependency relationships are highlighted in the figure. For example, the second-order fuzzy dependency relationships discovered are: i) $\mathrm{SPT21} = \mathrm{H}$ and $\mathrm{HTA1} = H- > \mathrm{HTB2} = \mathrm{H}$, ii) $\mathrm{SPT16} = \mathrm{H}$ and $\mathrm{HTA2} = \mathrm{H}- > \mathrm{HTB1} = \mathrm{H}$, iii) $\mathrm{CLN2} = \mathrm{H}$ and $\mathrm{CDC20} = \mathrm{L}- > \mathrm{CLN1} = \mathrm{H}$. The third-order fuzzy dependency relationships discovered are: i) $\mathrm{SWI5} = \mathrm{L}$ and $\mathrm{FUS1} = \mathrm{H}$ and $\mathrm{CLB2} = \mathrm{L}- > \mathrm{CLN1} = \mathrm{H}$, ii) $\mathrm{CLB1} = \mathrm{L}$ and $\mathrm{FAR1} = \mathrm{L}$ and $\mathrm{RME1} = \mathrm{H}- > \mathrm{CLN2} = \mathrm{H}$.

TABLE III
COMPARISON OF THE AVERAGE PREDICTION ACCURACY (16 TARGET GENES)

	Proposed	C4.5	SVM	FID
cdc15	87%	77%	79%	83%
alpha	89%	71%	76%	81%
cdc28	86%	70%	74%	79%
average	87%	73%	76%	81%

TABLE IV
COMPARISON OF THE AVERAGE PREDICTION ACCURACY (CDC15 DATASET)

	Proposed	C4.5	SVM	FID
CLN1	96%	80%	83%	86%
CDC20	80%	73%	80%	80%
CLB1	93%	90%	90%	93%
CLB4	83%	73%	80%	80%
HTA2	90%	83%	76%	90%
SWI4	80%	66%	70%	73%
CLB5	80%	76%	76%	76%
HTB2	86%	80%	80%	83%
CLN2	93%	83%	86%	90%
CLB6	90%	83%	90%	86%
HTA1	86%	73%	70%	73%
CLB2	96%	90%	86%	96%
SIC1	83%	76%	66%	76%
SWI5	90%	76%	80%	86%
HTB1	80%	76%	76%	80%
CLN3	80%	63%	70%	76%

In Table III, a summary of the average prediction accuracy (16 target genes) of the different approaches is given (for the cdc15 data set—standard deviation: proposed-1.9%, C4.5-3.2%, SVM-2.9%, and FID-2.2%). Tables IV–VI show the average prediction accuracy (each target gene) of the different approaches in each of the three data sets respectively.

TABLE V
COMPARISON OF THE AVERAGE PREDICTION ACCURACY (ALPHA DATASET)

	Proposed	C4.5	SVM	FID
CLN1	94%	83%	83%	94%
CDC20	83%	55%	66%	72%
CLB1	94%	83%	77%	94%
CLB4	83%	61%	72%	77%
HTA2	94%	83%	83%	94%
SWI4	77%	44%	72%	61%
CLB5	88%	83%	88%	88%
HTB2	94%	77%	83%	88%
CLN2	100%	77%	77%	83%
CLB6	88%	83%	77%	83%
HTA1	88%	77%	83%	83%
CLB2	94%	77%	83%	88%
SIC1	77%	55%	61%	66%
SWI5	88%	72%	72%	77%
HTB1	100%	77%	83%	94%
CLN3	77%	50%	55%	61%

The results show that the proposed technique has the highest prediction accuracy. It is followed by FID. In general, the fuzzy approaches performed better as expected as the use of fuzzy modeling makes these approaches better able to handle noise. During the training process, interesting fuzzy dependency relationships that can be used to construct characteristic descriptions of the states of each target gene (16 genes) were discovered. Based on these findings, for example, we can then construct the gene interaction diagram [6] as showed in Fig. 4. This diagram might provide important clues for the reconstruction of underlying GRNs.

It should be noted that there are two major types of gene regulatory relationships at the level of transcription [6], [7]. They

TABLE VI
COMPARISON OF THE AVERAGE PREDICTION ACCURACY (CDC28 DATASET)

	Proposed	C4.5	SVM	FID
CLN1	88%	76%	82%	88%
CDC20	76%	58%	64%	64%
CLB1	94%	82%	82%	82%
CLB4	76%	52%	58%	70%
HTA2	82%	76%	76%	82%
SWI4	76%	52%	64%	70%
CLB5	88%	82%	76%	82%
HTB2	94%	70%	76%	88%
CLN2	94%	82%	82%	94%
CLB6	88%	82%	76%	82%
HTA1	94%	70%	76%	88%
CLB2	94%	88%	82%	88%
SIC1	82%	64%	70%	70%
SWI5	82%	76%	76%	82%
HTB1	94%	64%	76%	88%
CLN3	76%	52%	64%	64%

are activation and inhibition. Activation and inhibition can take place through the regulator (the protein product of G_p) directly binding to G_j (the target gene), or by binding to other regulators, thereby, controlling G_j indirectly. In the activation process, if one is hypothesizing that G_p activates G_j, one would expect to see in the data that, if the state of G_p is high, it is to be followed by the state of G_j being also high and if the state of G_p is low, it is to be followed by the state of G_j being low. The expectation would be reversed for inhibition. Hence, based on the fuzzy dependency relationships discovered among the various states of different genes, we may determine whether or not one gene activates and inhibits another. The following is an example of a discovered third-order fuzzy dependency relationship. $\mathrm{SWI5} = \mathrm{L}$ and $\mathrm{FUS1} = \mathrm{H}$ and $\mathrm{CLB2} = \mathrm{L}- > \mathrm{CLN1} = \mathrm{H}$. The above biologically meaningful fuzzy dependency relationship discovered by the proposed technique indicates that if the states of SWI5 and CLB2 are low (L) and the state of FUS1 is high (H), then the state of CLN1 is high (H) in the next time instance. This also implies that FUS1 activates CLN1 and both SWI5 and CLB2 inhibit CLN1. Searching the gene interactions discovered by other approaches, we found that the regulatory relationship between $\mathrm{FUS1} = \mathrm{H}$ and $\mathrm{CLN1} = \mathrm{H}$ cannot be discovered by both C4.5 and FID. And also, the regulatory relationship between $\mathrm{SWI5} = \mathrm{L}$ and $\mathrm{CLN1} = \mathrm{H}$ cannot be discovered by the edge detection approach. Another example of a third-order fuzzy dependency relationship discovered from the data is given as follow: $\mathrm{CLB1} = \mathrm{L}$ and $\mathrm{FAR1} = \mathrm{L}$ and $\mathrm{RME1} = \mathrm{H}- > \mathrm{CLN2} = \mathrm{H}$. We found that the regulatory relationships among $\mathrm{FAR1} = \mathrm{L}$, $\mathrm{RME1} = \mathrm{H}$ and $\mathrm{CLN2} = \mathrm{H}$, cannot be discovered by both C4.5 and FID. SBF is a transcription factor binding on the promoters of many genes induced at the G1/S transition. Since SBF contains Swi4 as the DNA-binding protein, therefore, the periodic expression of SWI4 indicates that Swi4 plays an essential role in regulating the expression of genes peaked at start (at G1 phase) [56]–[59]. The following first-order fuzzy dependency relationships were discovered by the proposed technique can reveal such interactions. $\mathrm{SWI4} = \mathrm{H}- > \mathrm{SVS1} = \mathrm{H}$, $\mathrm{SWI4} = \mathrm{H}- > \mathrm{HCM1} = \mathrm{H}$, and $\mathrm{SWI4} = \mathrm{H}- > \mathrm{SWE1} = \mathrm{H}$. Since SVS1, HCM1 and SWE1 contain an SBF binding site in their promoter regions and their expression peaked at G1 phase, therefore, this coincides with the biological evidence that their expression were regulated by SWI4 [60], [61]. However, we found that the above regulatory relationships cannot be discovered by the two fuzzy logic-based approaches and also the edge detection approach. Comparing the gene interactions discovered by the proposed technique with those discovered by other approaches as above, we found that some of the known regulatory relationships can only be discovered by the proposed technique. As discussed in Section II, since the pruning methods which the decision-tree-based classifiers adopt are based on hill-climbing approaches, some biologically meaningful gene interactions may therefore be overlooked. For the edge detection approach, as it focuses on detecting gene-pairs that have a high degree of similarity in their respective gene expression profiles, those weaker but important local similarities between the expression profiles could be overlooked. Also, the two fuzzy logic-based approaches allow for only one activator and one repressor for each gene to be found, the above high-order fuzzy dependency relationships that involve multiple regulators for each target gene cannot be found by them.

V. CONCLUSION

In this paper, we have presented an effective fuzzy data mining technique for the discovery of GRNs from time series gene expression data. By transforming the quantitative expression values into linguistic terms, the proposed technique can discover fuzzy dependency relationships in high-dimensional and very noisy data without the need for additional feature selection procedures. Based on the discovered fuzzy dependency relationships, the user can not only determine those genes affecting a target gene but also can identify whether or not the target gene is supposed to be activated or inhibited. In addition, the discovered fuzzy relationships can also be used to predict gene expression from other unseen samples. Experimental results on real expression data show that the use of fuzzy modeling can be effective and the fuzzy dependency relationships discovered can reveal biologically meaningful gene regulatory relationships that could be used to infer underlying structures of GRNs. Similar to the existing approaches, the proposed technique starts from searching for all possible pairs of genes to discover the fuzzy dependency relationships. In order to cope with very large gene expression datasets, for future research, we intend to exploit the inherently parallel nature of the proposed technique when discovering the dependency relationship. With such improvement, the proposed technique will be able to better tackle highly complex data sets and problems.

REFERENCES

[1] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science*, vol. 270, no. 5235, pp. 467–470, 1995.
[2] D. Shalon, S. J. Smith, and P. O. Brown, "A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization," *Genome Res.*, vol. 6, no. 7, pp. 639–645, 1996.
[3] G. Ramsay, "DNA chips—States-of-the-art," *Nature Biotechnol.*, vol. 16, no. 1, pp. 40–44, 1998.
[4] D. J. Lockhart and E. A. Winzeler, "Genomics, gene expression and DNA arrays," *Nature*, vol. 405, no. 6788, pp. 827–836, 2000.
[5] N. L. v. Berkum and F. C. Holstege, "DNA microarrays: Raising the profile," *Curr. Opin. Biotechnol.*, vol. 12, no. 1, pp. 48–52, 2001.

[6] J. M. Bower and H. Bolouri, *Computation Modeling of Genetic and Biochemical Networks. Cambridge*. Cambridge, MA: MIT Press, 2001.

[7] J. D. Watson, *Molecular Biology of the Gene*. San Francisco, CA: Pearson/Benjamin Cummings, 2004.

[8] D. Stekel, *Microarray Bioinformatics*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[9] T. Akutsu, S. Miyano, and S. Kuhara, "Identification of genetic networks from a small number of gene expression patterns under the boolean network model," in *Proc. Pacific Symp. Biocomput.*, 1999, pp. 17–28.

[10] T. Akutsu, S. Miyano, and S. Kuhara, "Inferring qualitative relations in genetic networks and metabolic pathways," *Bioinf.*, vol. 16, no. 8, pp. 727–734, 2000.

[11] L. Raeymaekers, "Dynamics of boolean networks controlled by biologically meaningful functions," *J. Theor. Biol.*, vol. 218, pp. 331–341, 2002.

[12] L. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang, "Probabilistic boolean networks: A rule-based uncertainty model for gene regulatory networks," *Bioinf.*, vol. 18, no. 2, pp. 261–274, 2002.

[13] D. Husmeier, "Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks," *Bioinf.*, vol. 19, no. 17, pp. 2271–2282, 2003.

[14] B. E. Perrin, L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, and F. Buc, "Gene networks inference using dynamic Bayesian networks," *Bioinf.*, vol. 19, pp. ii138–ii148, 2003.

[15] M. Zou and S. D. Conzen, "A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data," *Bioinf.*, vol. 21, no. 1, pp. 71–79, 2005.

[16] J. C. Leloup and A. Goldbeter, "Toward a detailed computational model for the mammalian circadian clock," *Proc. Nat. Acad. Sci.*, vol. 100, pp. 7051–7056, 2003.

[17] K. C. Chen, T. Y. Wang, H. H. Tseng, C. Y. Huang, and C. Y. Kao, "A stochastic differential equation model for quantifying transcriptional regulatory network in saccharomyces cerevisiae," *Bioinf.*, vol. 21, no. 12, pp. 2883–2890, 2005.

[18] J. J. Tyson, "Models of cell cycle control in eukaryotes," *J. Biotechnol.*, vol. 71, pp. 239–244, 1999.

[19] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Nat. Acad. Sci.*, vol. 95, no. 25, pp. 14863–14868, 1998.

[20] Y. P. P. Chen, *Bioinformatics Technologies*. New York: Springer-Verlag, 2005.

[21] A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering gene expression patterns," *J. Comput. Biol.*, vol. 6, no. 3–4, pp. 281–297, 1999.

[22] L. A. Soinov, M. A. Krestyaninova, and A. Brazma, "Towards reconstruction of gene networks from expression data by supervised learning," *Genome Biol.*, vol. 4, no. 1, R6, 2003.

[23] D. Zhu, A. O. Hero, Z. S. Qin, and A. Swaroop, "High throughput screening of co-expressed gene pairs with controlled false discovery rate (FDR) and minimum acceptable strength (MAS)," *J. Comput. Biol.*, vol. 12, no. 7, pp. 1029–1045, 2005.

[24] V. Filkov, S. Skiena, and J. Zhi, "Analysis techniques for microarray time-series data," in *Proc. Int. Conf. Res. Comput. Molecular Biol. (RECOMB)*, 2001, pp. 124–131.

[25] P. J. Woolf and Y. Wang, "A fuzzy logic approach to analyzing gene expression data," *Physiol. Genomics*, vol. 3, no. 1, pp. 9–15, 2000.

[26] H. Ressom, R. Reynolds, and R. S. Varghese, "Increasing the efficiency of fuzzy logic-based gene expression data analysis," *Physiol. Genom.*, vol. 13, no. 2, pp. 107–117, 2003.

[27] B. A. Sokhansanj, J. P. Fitch, J. N. Quong, and A. A. Quong, "Linear fuzzy gene network models obtained from microarray data by exhaustive search," *BMC Bioinf.*, vol. 5, pp. 108–120, 2004.

[28] D. Endy and R. Brent, "Modeling cellular behaviour," *Nature*, vol. 409, pp. 391–395, 2001.

[29] D. M. Chickering, Learning Bayesian Networks From Data UCLA Cognitive Systems Laboratory, Los Angeles, CA, 1996, Tech. Rep. R-245.

[30] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning Bayesian networks: The combination of knowledge and statistical data," *Mach. Learn.*, vol. 20, pp. 197–243, 1995.

[31] M. Dejori and M. Stetter, "Bayesian inference of genetic networks from gene expression data: Convergence and reliability," in *Proc. Int. Conf. Artif. Intell.*, 2003, pp. 321–327.

[32] P. W. Atkins, *Physical Chemistry*, 6th ed. New York: Freeman, 1998.

[33] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA: Morgan Kaufmann, 1993.

[34] L. A. Zadeh, "Fuzzy logic and approximate reasoning," *Synthese*, vol. 30, pp. 407–428, 1975.

[35] L. A. Zadeh, "A theory of approximate reasoning," *Mach. Intell.*, vol. 9, pp. 149–194, 1979.

[36] L. A. Zadeh, "The role of fuzzy logic in the management of uncertainty in expert systems," *Fuzzy Sets Syst.*, vol. 11, pp. 199–227, 1983.

[37] J. Yen, "Fuzzy logic-A modern perspective," *IEEE Trans. Knowl. Data Eng.*, vol. 11, no. 1, pp. 153–165, Jan. 1999.

[38] S. J. Haberman, "The analysis of residuals in cross-classified tables," *Biometrics*, vol. 29, pp. 205–220, 1973.

[39] K. C. C. Chan and A. K. C. Wong, "A statistical technique for extracting classificatory knowledge from databases," in *Knowledge Discovery in Databases*, G. Piatesky-Shapiro and W. Frawley, Eds. Cambridge, MA: MIT Press, 1991, pp. 107–123.

[40] Y. Wong and A. K. C Chan, "From association to classification: Inference using weight of evidence," *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 3, pp. 764–767, Jun. 2003.

[41] W. H. Au, K. C. C. Chan, and X. Yao, "A novel evolutionary data mining algorithm with applications to churn modeling," *IEEE Trans. Evol. Comput.*, vol. 7, no. 6, pp. 532–545, Jun. 2003.

[42] K. C. C. Chan, A. K. C. Wong, and D. K. Y. Chiu, "Learning sequential patterns for probabilistic inductive prediction," *IEEE Trans. Syst. Man Cybern.*, vol. 24, no. 10, pp. 1532–1547, Oct. 1994.

[43] D. B. Osteyee and I. J. Good, *Information, Weight of Evidence, the Singularity between Probability Measures and Signal Detection*. Berlin, Germany: Springer-Verlag, 1974.

[44] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Lyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization," *Mol. Biol. Cell.*, vol. 9, no. 12, pp. 3273–3297, 1998.

[45] S. Mitra and T. Acharya, *Data Mining: Multimedia, Soft Computing, and Bioinformatics*. Hoboken, NJ: Wiley, 2003.

[46] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. San Francisco, CA: Morgan Kaufmann, 2001.

[47] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[48] U. Fayyad and K. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, 1993, pp. 1022–1029.

[49] R. Kohavi, "Wrappers for Performance Enhancement and Oblivious Decision Graphs," Ph.D. dissertation, Stanford University, Computer Science Department, Stanford, CA, 1995.

[50] B. Scholkopf, K. Tsuda, and J. P. Vert, *Kernel Methods in Computational Biology*. Cambridge, MA: MIT Press, 2004.

[51] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, pp. 389–422, 2002.

[52] S. A. Vinterbo, E. Y. Kim, and L. O. Machado, "Small, fuzzy and interpretable gene expression based classifiers," *Bioinf.*, vol. 21, no. 9, pp. 1964–1970, 2005.

[53] C. Z. Janikow, "Fuzzy decision trees: Issues and methods," *IEEE Trans. Syst. Man Cybern. B, Cybern.*, vol. 28, no. 1, pp. 1–14, Jan. 1998.

[54] L. Li, C. R. Weinberg, T. A. Darden, and L. G. Pedersen, "Gene selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method," *Bioinf.*, vol. 17, no. 12, pp. 1131–1142, 2001.

[55] K. Duan and J. C. Rajapakse, "SVM-RFE peak selection for cancer classification with mass spectrometry data," in *Proc. 3rd Asia-Pacific Bioinf. Conf.*, 2005, pp. 191–200.

[56] T. I. Lee *et al.*, "Transcriptional regulatory networks in saccharomyces cerevisiae," *Science*, vol. 298, pp. 799–804, 2002.

[57] E. Wingender *et al.*, "The transfac system on gene expression regulation," *Nucleic Acids Res.*, vol. 29, pp. 281–283, 2001.

[58] M. Kanehisa *et al.*, "From genomics to chemical genomics: New developments in KEGG," *Nucleic Acids Res.*, vol. 34, pp. D354–357, 2006.

[59] C. A. Ball *et al.*, "Saccharomyces genome database provides tools to survey gene expression and functional analysis data," *Nucleic Acids Res.*, vol. 29, no. 1, pp. 80–81, 2001.

[60] V. R. Iyer, C. E. Horak, C. S. Scafe, D. Botstein, M. Snyder, and P. O. Brown, "Genome binding sites of the yeast cell-cycle transcription factors SBF and MBF," *Nature*, vol. 409, pp. 533–538, 2001.

[61] I. Simon, J. Barnett, N. Hannett, C. T. Harbison, N. J. Rinaldi, T. L. Volkert, J. J. Wyrick, J. Zeitlinger, D. K. Gifford, T. S. Jaakkola, and R. A. Young, "Serial regulation of transcriptional regulators in the yeast cell cycle," *Cell*, vol. 106, pp. 697–708, 2001.

Patrick C. H. Ma received the B.A. and Ph.D. degrees in computer science from the Hong Kong Polytechnic University, Hong Kong, China, in 2001 and 2006, respectively.

Currently, he is a Project Assistant in the Department of Computing, the Hong Kong Polytechnic University, Hong Kong, China. His research interests are in bioinformatics, data mining, and computational intelligence.

Keith C. C. Chan received the B.Math. degree in computer science and statistics, and the M.A.Sc. and Ph.D. degrees in systems design engineering from the University of Waterloo, Waterloo, ON, Canada, in 1983, 1985, and 1989, respectively.

He was with IBM Canada Laboratory, Toronto, ON, Canadam where he was involved in the development of software engineering tools. In 1993, he joined the Department of Electrical and Computer Engineering, Ryerson University, Toronto, Ontario, Canada, as an Associate Professor. In 1994, he joined the Department of Computing, the Hong Kong Polytechnic University, Hong Kong, China, where he is now Professor and Head. He is also a Guest Professor of the Graduate School and an Adjunct Professor of the Institute of Software, the Chinese Academy of Sciences, Beijing, China. He has served as a Consultant to government agencies and various companies in Hong Kong, China, Singapore, Malaysia, and Canada. His research interests are in data mining, computational intelligence, bioinformatics, and software engineering.