INFERENCE OF GENE REGULATORY NETWORKS FROM MICROARRAY DATA: A FUZZY LOGIC APPROACH

PATRICK C.H. MA AND KEITH C.C. CHAN[†]

Department of Computing, The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong SAR, China

Recent developments in large-scale monitoring of gene expression such as DNA microarrays have made the reconstruction of gene regulatory networks (GRNs) feasible. Before one can infer the structures of these networks, it is important to identify, for each gene in the network, which genes can affect its expression and how they affect it. Most of the existing approaches are useful exploratory tools in the sense that they allow the user to generate biological hypotheses about transcriptional regulations of genes that can then be tested in the laboratory. However, the patterns discovered by these approaches are not adequate for making accurate prediction on gene expression patterns in new or held-out experiments. Therefore, it is difficult to compare performance of different approaches or decide which approach is likely to generate plausible hypothesis. For this reason, we need an approach that not only can provide interpretable insight into the structures of GRNs but also can provide accurate prediction. In this paper, we present a novel fuzzy logic-based approach for this problem. The desired characteristics of the proposed algorithm are as follows: (i) it is able to directly mine the high-dimensional expression data without the need for additional feature selection procedures, (ii) it is able to distinguish between relevant and irrelevant expression data in predicting the expression patterns of predicted genes, (iii) based on the proposed objective interestingness measure, no user-specified thresholds are needed in advance, (iv) it can make explicit hidden patterns discovered for possible biological interpretation, (v) the discovered patterns can be used to predict gene expression patterns in other unseen tissue samples, and (vi) with fuzzy logic, it is robust to noise in the expression data as it hides the boundaries of the adjacent intervals of the quantitative attributes. Experimental results on real expression data show that it can be very effective and the discovered patterns reveal biologically meaningful regulatory relationships of genes that could help the user reconstructing the underlying structures of GRNs.

1 Introduction

Large-scale monitoring gene expression such as DNA microarrays [1,2] is considered to be one of the most promising techniques for reconstructing the gene regulatory networks (GRNs). A GRN is typically a complex biological system in which proteins and genes bind to each other and act as an input-output system for controlling various cellular processes. Since, living cells contain thousands of genes, each of which codes for one or more proteins. Many of these proteins in turn regulate the expression of some other genes through complex regulatory pathways to accommodate changes in different external environments or carry out the essential developmental programs. The key to understanding living processes is therefore to uncover the structures of these regulatory networks that underlie the regulations of cells.

[†] E-mail: {cschma,cskcchan}@comp.polyu.edu.hk

Previous attempts have been reported to inferring the underlying structures of GRNs such as the biochemically driven approaches [3,4], the Boolean network approaches [5], the Bayesian network approaches [6] and the data mining approaches [7-9]. However, these approaches have several limitations need to be overcome in order to effectively deal with the problem. For example, for the biochemically driven approaches, most of the biochemical reactions under participation of proteins do not follow linear reaction kinetics and also gene expression data seems not sufficient to globally understand regulatory networks at this level of detail [3,4]. For the Boolean network approaches, the validity of the pre-defined assumptions [5] and the values of the Boolean approach in general, have been questioned by a number of researchers, particularly in the biological community, where there is a perceived lack of connection between simulation results and empirically testable hypotheses [10]. For the Bayesian network approaches, the task of learning model parameters is NP-hard especially for high-dimensional data. Moreover, many parameters need to be estimated accurately and this requires a large amount of samples that may not always be readily available [6]. For the data mining approaches, clustering of gene expression data [7] only measures whether genes share a significant linear relationship with each other. The regulatory relationships such as which gene affects which other genes cannot be discovered. On the other hand, the crisp discretization procedures of the classification algorithms [8,9] such as C4.5 [11] do not take into account that values at the borderline between value categories may be very similar. This makes the classifiers less resilient to noise and some useful patterns exist at this borderline can be overlook.

Besides the above limitations, the patterns discovered by most of the existing approaches are not adequate for making accurate prediction on gene expression patterns in new or held-out experiments. Hence, it is difficult to compare performance of them or decide which approach is likely to generate plausible hypothesis. Therefore, we need an approach that not only can provide interpretable insight into the structures of GRNs but also can provide accurate prediction. For this reason, we propose a novel fuzzy logic-based approach in this paper. The rest of the paper is organized as follows. In Section 2, the proposed algorithm is described in details. The effectiveness of the proposed algorithm has been evaluated and compared through various experiments with real expression data. The experimental set-up, together with the results, is discussed in Section 3. Lastly, in Section 4, we give a summary of the paper.

2 The proposed algorithm

Fuzzy logic and fuzzy sets allow the modeling of language-related uncertainties by providing a symbolic framework for knowledge comprehensibility [12,13]. Fuzzy representation is becoming increasingly popular in dealing with problems of uncertainty, noise and inexact data. Recently, fuzzy logic has successfully been used for clustering gene expression data. For example, the fuzzy *k*-means algorithms [14,15] have been applied to discover the clusters of co-expressed genes so that genes have similar biological functions can be revealed. However, for the inference of GRNs, only limited studies have been proposed [16]. Due to the fact that there is a need to have an effective fuzzy logic-based algorithm, here, we propose such an algorithm and discuss the details in this section.

2.1. *Linguistic variables and linguistic terms representation*

Given a set of data D, each record r (experimental condition), is characterized by a set of attributes (genes), $A=\{A_1,\ldots,A_i,\ldots,A_n\}$. For any record, $r \in D$, $r[A_i]$ denotes the value in r for attribute A_i. Let $L=\{L_1,\ldots,L_i,\ldots,L_n\}$ be a set of linguistic variables such that $L_i \in L$ represents $A_i \in A$. For any quantitative attribute, A_i, let $dom(A_i)=[l_i,u_i]\subseteq \Re$ denote the domain of the attribute, where l_i and u_i represent the lower and upper bounds of A_i respectively. Moreover, A_i is represented by a linguistic variable, L_i, whose value is a linguistic term in $T(L_i)=\{l_{ij} \mid j=1,2,\ldots,s_i\}$ where l_{ij} is a linguistic term characterized by a fuzzy set, F_{ij}, that is defined on $dom(A_i)$ and whose membership function is $\mu_{F_{ij}}$. The degree of membership of the value in r with respect to F_{ij} is given by $\mu_{F_{ij}}(r[A_i])$. The degree to which r is characterized by l_{ij}, $\lambda_{l_{ij}}(r)$, is therefore defined as follows:

$$\lambda_{l_{ij}}(r)=\mu_{F_{ij}}(r[A_i]). \tag{1}$$

If $\lambda_{l_{ij}}(r)=1$, the attribute A_i of r is completely characterized by the linguistic term l_{ij}. If $\lambda_{l_{ij}}(r)=0$, the attribute A_i of r is not characterized by the linguistic term l_{ij}. If $0<\lambda_{l_{ij}}(r)<1$, the attribute A_i of r is partially characterized by the linguistic term l_{ij}. In the case where $r[A_i]$ is unknown, $\lambda_{l_{ij}}(r)=0.5$, which indicates that there is no information available concerning whether the attribute A_i of r is or is not characterized by the linguistic term l_{ij}.

2.2. *Discovering the interesting patterns*

Let $o(l_{ij})$ be the observed degree to which the records in the given dataset are characterized by the linguistic term l_{ij}. It is defined as follows:

$$o(l_{ij})=\sum_{r\in D}\lambda_{l_{ij}}(r)\cdot \tag{2}$$

Also, let $l_{ij} \Leftrightarrow l_{pq}$ be the association between the linguistic terms l_{ij} and l_{pq}. Then, the observed degree to which the records are characterized by this association, $o(l_{ij} \Leftrightarrow l_{pq})$, is defined as follows:

$$o(l_{ij} \Leftrightarrow l_{pq})=\sum_{r\in D}\min(\lambda_{l_{ij}}(r),\lambda_{l_{pq}}(r))\cdot \tag{3}$$

To decide whether an association, $l_{ij} \Leftrightarrow l_{pq}$, is interesting, it is objectively evaluated using the proposed objective interestingness measure, $d(l_{ij} \Leftrightarrow l_{pq})$. This measure reflects the differences in the observed $o(l_{ij} \Leftrightarrow l_{pq})$ and expected $e(l_{ij} \Leftrightarrow l_{pq})$ degrees to which the records are characterized by these linguistic terms. The objective interestingness measure, $d(l_{ij} \Leftrightarrow l_{pq})$, is defined as follows:

$$d(l_{ij} \Leftrightarrow l_{pq}) = \frac{z(l_{ij} \Leftrightarrow l_{pq})}{\sqrt{v(l_{ij} \Leftrightarrow l_{pq})}}, \tag{4}$$

where

$$z(l_{ij} \Leftrightarrow l_{pq}) = \frac{o(l_{ij} \Leftrightarrow l_{pq}) - e(l_{ij} \Leftrightarrow l_{pq})}{\sqrt{e(l_{ij} \Leftrightarrow l_{pq})}}, \tag{5}$$

$$e(l_{ij} \Leftrightarrow l_{pq}) = \frac{o(l_{ij}) \times o(l_{pq})}{\sum_{r \in D} \sum_{q=1}^{s_p} \sum_{j=1}^{s_i} \min(\lambda_{l_{ij}}(r), \lambda_{l_{pq}}(r))}, \tag{6}$$

$$v(l_{ij} \Leftrightarrow l_{pq}) = (1 - \frac{o(l_{ij})}{\sum_{r \in D} \sum_{j=1}^{s_i} \lambda_{l_{ij}}(r)})(1 - \frac{o(l_{pq})}{\sum_{r \in D} \sum_{q=1}^{s_p} \lambda_{l_{pq}}(r)}). \tag{7}$$

If $d(l_{ij} \Leftrightarrow l_{pq}) > 1.96$ (i.e., the 95th percentile of the normal distribution) [17-19], we can conclude that the association $l_{ij} \Leftrightarrow l_{pq}$ is interesting. It means that it is more likely for a record to be characterized by both l_{ij} and l_{pq}.

2.3. *Prediction based on the discovered patterns*

Given that $l_{ij} \Leftrightarrow l_{pq}$ is interesting, the patterns can be constructed as follows:

$$w(l_{ij} \Leftrightarrow l_{pq}) = \log \frac{\Pr(l_{ij} \Leftrightarrow l_{pq})}{\Pr(l_{ij} \Leftrightarrow \neg l_{pq})}, \tag{8}$$

where

$$\Pr(l_{ij} \Leftrightarrow l_{pq}) = \frac{\sum_{r \in D} \min(\lambda_{l_{ij}}(r), \lambda_{l_{pq}}(r))}{\sum_{r \in D} \sum_{j=1}^{s_i} \min(\lambda_{l_{ij}}(r), \lambda_{l_{pq}}(r))},$$

$$\Pr(l_{ij} \Leftrightarrow \neg l_{pq}) = \frac{\sum_{r \in D} \sum_{u=1, u \neq q}^{s_p} \min(\lambda_{l_{ij}}(r), \lambda_{l_{pu}}(r))}{\sum_{r \in D} \sum_{u=1, u \neq q}^{s_p} \sum_{j=1}^{s_i} \min(\lambda_{l_{ij}}(r), \lambda_{l_{pu}}(r))}. \tag{9}$$

The term $\Pr(l_{ij} \Leftrightarrow l_{pq})$ can be considered as being the probability that a record is characterized by l_{ij} and l_{pq}, and the term $\Pr(l_{ij} \Leftrightarrow \neg l_{pq})$ can be considered as being the probability that a record is characterized by l_{ij} and l_{pu}, where $u \neq q$. Then, the term $w(l_{ij} \Leftrightarrow l_{pq})$ is a confidence measure that represents the uncertainty associated with $l_{ij} \Leftrightarrow l_{pq}$. It can be interpreted as being a measure of the difference in the information gain when a record that is characterized by l_{ij} is also characterized by l_{pq} as opposed to being characterized by other linguistic terms l_{pu}, where $u \neq q$.

Given a testing record r and it is characterized by n attribute values, $r[A_1],\ldots,r[A_p],\ldots,r[A_n]$, where $r[A_p]$ is the value that is to be predicted. Let l_p be a linguistic term with a domain of $T(L_p)$. The value of $r[A_p]$ is determined according to l_p. To predict $r[A_p]$, the discovered patterns are searched. If an attribute value, say $r[A_i]$, $A_i \neq A_p$, of r is characterized by the linguistic term in the antecedent of a pattern that implies l_{pq}, then it can be considered as providing some confidence that the value of l_p should be assigned to l_{pq}. By repeating this procedure, that is, by matching each attribute value of r against the discovered patterns, the value of l_p can be determined by computing the total confidence measure. Since each attribute of r may or may not provide a contribution to the total confidence measure and those that may support the assignment of different values. Therefore, the different contributions to the total confidence measure are measured quantitatively and then combined for comparison in order to find the most suitable value of l_p. For any attribute value, $r[A_i]$, $A_i \neq A_p$, of r, it is characterized by a linguistic term, l_{ij}, to a degree of compatibility, $\lambda_{l_{ij}}(r)$. Given the patterns those imply the assignment of l_{pq}, then, the confidence provided by $r[A_i]$ for such as assignment is as follows:

$$W_{l_{pq}r[A_i]} = w(l_{ij} \Leftrightarrow l_{pq}) \times \lambda_{l_{ij}}(r). \qquad (10)$$

Suppose that among the $n-1$ attributes excluding A_p, only some combinations of them, $r[A_1],\ldots,r[A_i],\ldots,r[A_\beta]$, are found to match one or more patterns. Then, the total confidence measure of assigning the value of l_p to l_{pq} is given as follows:

$$TW_q = \sum_{i=1}^{\beta} W_{l_{pq}} r[A_i]. \qquad (11)$$

Based on the above total confidence measure, in the case if $TW_q > TW_c$, where $q \neq c$. Then, l_p is assigned to l_{pq}.

3 Experimental results

3.1. *Experimental data*

For experimentation with real data, we used a set of gene expression data that contains a series of gene expression measurements of the transcript (mRNA) levels of S. cerevisiae genes [7,20]. In this dataset, the 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 [7]. The expression data we used is available at [21]. Since gene expression can be described in a finite number of different states/patterns [22]. We therefore represented it in terms of three fuzzy sets: low (L), medium (M) and high (H). For any quantitative attribute A_i, the degree of membership of a record, $r[A_i]$, can be computed as follows [23] (in Fig. 1):

$$\mu_{low}(r[A_i]) = \begin{cases} 1, & if \quad r[A_i] < Av_{i1} \\ \dfrac{P_{i2} - r[A_i]}{P_{i2} - Av_{i1}}, & if \quad Av_{i1} \le r[A_i] < P_{i2} \\ 0, & otherwise \end{cases} \tag{12}$$

$$\mu_{medium}(r[A_i]) = \begin{cases} 0, & if \quad r[A_i] < P_{i1} \\ \dfrac{r[A_i] - P_{i1}}{Av_{i2} - P_{i1}}, & if \quad P_{i1} \le r[A_i] < Av_{i2} \\ \dfrac{P_{i2} - r[A_i]}{P_{i2} - Av_{i2}}, & if \quad Av_{i2} \le r[A_i] < P_{i2} \\ 0, & otherwise \end{cases} \tag{13}$$

$$\mu_{high}(r[A_i]) = \begin{cases} 0, & if \quad r[A_i] < P_{i1} \\ \dfrac{r[A_i] - P_{i1}}{Av_{i3} - P_{i1}}, & if \quad P_{i1} \le r[A_i] < Av_{i3} \\ 1, & otherwise \end{cases} \tag{14}$$

where A_i is sorted in the ascending order of its values, P_{i1} is the value of A_i that exceeds one-third of the measurements and is less than the remaining two-thirds and P_{i2} is the value of A_i that exceeds two-thirds of the measurements and is less than the remaining one-third. And also, $A_{i_{\max}}$ and $A_{i_{\min}}$ denote the maximum and minimum values encountered along attribute A_i, and $Av_{i1} = \frac{A_{i\min} + P_{i1}}{2}$, $Av_{i2} = \frac{P_{i1} + P_{i2}}{2}$ and $Av_{i3} = \frac{P_{i2} + A_{i\max}}{2}$ [23].

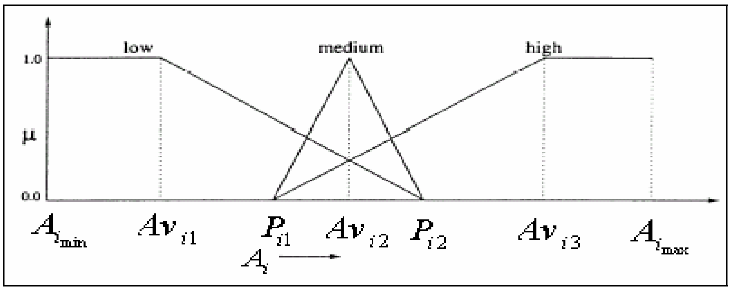


Figure 1. Membership function.

3.2. *Method of evaluating the results*

In this analysis, we chose the cdc15 experiment as the training set. Another two datasets: alpha and cdc28 experiments were used as the testing sets. For experimentation, we randomly selected 6 genes (CLN1, HTA1, HTB1, CLB1, CLN2, and CLB6) to evaluate the effectiveness of the proposed algorithm. Using the proposed algorithm, the patterns of these genes in the independent testing sets are predicted. Then, the predicted patterns are compared with the original patterns of these genes and the percentage of accurate prediction can therefore be determined.

3.3. *Results*

To evaluate the performance of the proposed method, we also compared it to the popular decision-tree based algorithm called C4.5 [11] as discussed in Section 1. Moreover, since one of the desirable features of the proposed algorithm is its feature selection capability, it is able to distinguish between relevant and irrelevant expression data. Therefore, for fair performance comparisons, we performed additional experiments to compare it to C4.5 with feature selection approach. There are many feature selection methods have been proposed for gene expression data such as filter and wrapper methods [24,25]. In this analysis, we adopted *t*-statistics measure [25]. Based on the *t*-statistics measure, the new subset of genes with largest *t*-values was obtained. The selection method of genes with largest *t*-values is as follows: (i) sorted the genes in descending order based on their *t*-values, (ii) initially, 5% (empirically set) of genes were selected from top of the rank list, (iii) the classification performance based on this subset of genes was measured by C4.5 (10-fold cross validation), (iv) added another 5% of genes from the rank list into this subset, (v) repeat steps (iii) and (iv) until the classification performance converged, (vi) the final subset of genes was selected.

In Tables 1 and 2, the comparisons of average prediction accuracy are showed. According to these tables, we found that the performance of C4.5 can be improved with the feature selection procedure. In addition, we also compared another well-known decision-tree based algorithm called FID [26] and trained the algorithm only using the significant features identified by C4.5 during the feature selection process as discussed above. FID is a fuzzy logic-based classifier that combines symbolic decision trees with approximate reasoning offered by fuzzy representation. It extends C4.5 by using splitting criteria based on fuzzy restrictions and using different inference procedures to exploit fuzzy sets. The experimental results of FID are also showed in Tables 1 and 2. According to these results, we found that the performance of the proposed algorithm is not only better than other popular algorithms and also the average prediction accuracy in each testing set is high. This indicates that the proposed algorithm is very effective in predicting gene expression patterns in the unseen samples.

Table 1. Result comparison (alpha dataset).

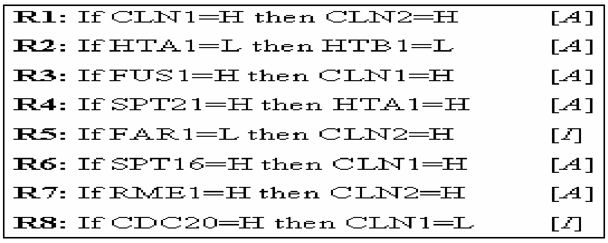
Gene	Proposed	C4.5	C4.5 + Feature selection	FID + Feature selection
CLN1	0.94	0.67	0.83	0.94
HTA1	0.89	0.61	0.78	0.83
HTB1	1	0.67	0.78	0.94
CLB1	0.94	0.67	0.83	0.94
CLN2	1	0.67	0.78	0.83
CLB6	0.89	0.72	0.83	0.83
Avg.	0.94	0.67	0.81	0.89

Table 2. Result comparison (cdc28 dataset).

Gene	Proposed	C4.5	C4.5 + Feature selection	FID + Feature selection
CLN1	0.88	0.65	0.76	0.88
HTA1	0.94	0.58	0.71	0.88
HTB1	0.94	0.53	0.65	0.88
CLB1	0.94	0.71	0.82	0.82
CLN2	0.94	0.71	0.82	0.94
CLB6	0.88	0.65	0.82	0.76
Avg.	0.92	0.64	0.76	0.86

3.4. *Biological interpretation*

In order to evaluate the biological significance of the discovered patterns, we tried to verify that any known regulatory relationships of genes could be revealed from them. In Fig. 2, it shows some of the discovered patterns (with high confidence measures, Section 2.3) represented in rules that reveal known regulatory relationships [27]. Based on the discovered relationships, we can then construct the gene interaction diagrams [28] as showed in Fig. 3 that might provide important clues in reconstructing the structures of the underlying GRNs. One of the appealing advantages of network reconstruction using the proposed algorithm is that the user can easily improve the classifier by adding new samples or experimental conditions and reproduce the architecture of a network consistent with the data. Since such iterative improvements can be part of an interactive process. Therefore, the proposed algorithm can be considered as a basis for an interactive expert system for gene network reconstruction.



R1: If CLN1=H then CLN2=H [A]
R2: If HTA1=L then HTB1=L [A]
R3: If FUS1=H then CLN1=H [A]
R4: If SPT21=H then HTA1=H [A]
R5: If FAR1=L then CLN2=H [I]
R6: If SPT16=H then CLN1=H [A]
R7: If RME1=H then CLN2=H [A]
R8: If CDC20=H then CLN1=L [I]

Figure 2. Patterns discovered (*A* - known activation relationships and *I* - known inhibition relationships).

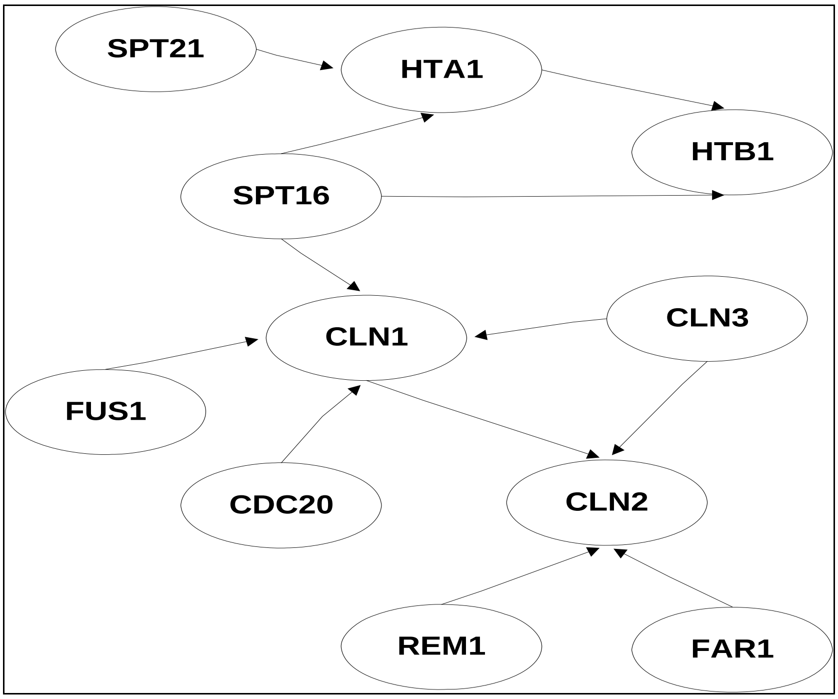


Figure 3. Gene interaction diagram discovered (12 known regulatory relationships involved). Solid lines correspond to activation relationships and broken lines correspond to inhibition relationships.

4 Conclusions

In this paper, we have presented a novel fuzzy logic-based approach for the inference of GRNs. The proposed algorithm is able to distinguish between relevant and irrelevant expression data in predicting the expression patterns of predicted genes without the need for additional feature selection procedures. And also, it is able to explicitly reveal the discovered patterns for possible biological interpretation. With the proposed objective interestingness measure, no user-specified thresholds are needed in advance. Experimental results on real expression data show that the proposed algorithm can be very effective and the discovered patterns reveal biologically meaningful regulatory relationships of genes that could help the user reconstructing the underlying GRNs.

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*, 270(5235):467—470, 1995.
2. D.J. Lockhart and E.A. Winzeler. Genomics, gene expression and DNA arrays. *Nature*, 405(6788):827—836, 2000.
3. J.C. Leloup and A. Goldbeter. Toward a detailed computational model for the mammalian circadian clock. *Proc. of the National Academy of Science, USA*, 100:7051—7056, 2003.
4. 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. *Bioinformatics*, Advance Access published online on March, 2005.
5. T. Akutsu, S. Miyano and S. Kuhara. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. *Pacific Sym. on Biocomputing*, 17—28, 1999.
6. B.E. Perrin, L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet and F. Buc. Gene networks inference using dynamic bayesian networks. *Bioinformatics*, 19:138—148, 2003.
7. 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.*, 9(12):3273—3297, 1998.
8. L. Wong. *The Practical Bioinformatician.* World Scientific, 2004.
9. M. Middendorf, A. Kundaje, C. Wiggins, Y. Freund and C. Leslie. Predicting genetic regulatory response using classification. *Bioinformatics*, 20:232—240, 2004.
10. D. Endy and R. Brent. Modeling cellular behaviour. *Nature*, 409:391—395, 2001.
11. J.R. Quinlan. *C4.5: Programs for Machine Learning.* San Fran., CA: Morgan Kaufmann, 1993.
12. L.A. Zadeh. Fuzzy sets. *Inf. Contr.*, 8:338—353, 1965.
13. L.A. Zadeh. Fuzzy logic and approximate reasoning. *Synthese*, 30:407—428, 1975.
14. A.P. Gasch and M.B. Eisen. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol.*, 3(11): RESEARCH0059.1—0059.22, 2002.
15. C. Arima, T. Hanai and M. Okamoto. Gene expression analysis using fuzzy k-means clustering. *Genome Informatics*, 14:334—335, 2003.
16. P.J. Woolf and Y. Wang. A fuzzy logic approach to analyzing gene expression data. *Physiol Genomics*, 3:9—15, 2000.
17. K.C.C. Chan and A.K.C. Wong. A statistical technique for extracting classificatory knowledge from databases. *Knowledge Discovery in Databases*, G. Piatesky-Shapiro and W.J. Frawley, Eds. Menlo Park, CA:/Cambridge, MA: AAAI/MIT Press, 107—123, 1991.
18. P.C.H. Ma, K.C.C. Chan and D.K.Y. Chiu. Clustering and re-clustering for pattern discovery in gene expression data. *Journal of Bioinformatics and Computational Biology*, 3(2):281—301, 2005.
19. Y. Wang and A.K.C. Wong. From association to classification: Inference using weight of evidence. *IEEE Trans. Knowledge and Data Engineering*, 15(3):764—767, 2003.
20. R.J. Cho, M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockhart and R.W. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell.*, 2(1):65—73, 1998.
21. http://genome-www.stanford.edu/cellcycle
22. C. Creighton and S. Hansah. Mining gene expression databases for association rules. *Bioinformatics*, 19(1):79—86, 2003.
23. S. Mitra, K.M. Konwar and S.K. Pal. Fuzzy decision tree, linguistic rules and fuzzy knowledge-based network: generation and evaluation. *IEEE Trans. on Systems, Man and Cybernetics - Part C: Applications and Reviews*, 32:328—339, 2002.
24. M. Xiong, X. Fang and J. Zhao. Biomarker identification by feature wrappers. *Genome Res.*, 11:1878—1887, 2001.
25. Y. Su, T.M. Murali et. al. RankGene: Identification of diagnostic genes based on expression data. *Bioinformatics,* 19(12):1578—1579, 2003. Available: http://genomics10.bu.edu/yangsu/rankgene/.
26. C.Z. Janikow. Fuzzy decision trees: issues and methods. *IEEE Trans. on Systems, Man and Cybernetics - Part B: Cybernetics*, 28(1):1—14, 1998.
27. V. Filkov, S. Skiena and J. Zhi. Analysis techniques for microarray time-series data. *In Proceedings of RECOMB*, 124—131, 2001.
28. J.M. Bower and H. Bolouri. *Computation Modeling of Genetic and Biochemical Networks.* Cambridge, Mass.: MIT Press, 2001.