

# Modeling Gene Expression Networks using Fuzzy Logic

Pan Du, Jian Gong, Eve Syrkin Wurtele, and Julie A. Dickerson, *Member, IEEE*

**Abstract**—Gene regulatory networks model regulation in living organisms. Fuzzy logic can effectively model gene regulation and interaction to accurately reflect the underlying biology. A new multi-scale fuzzy clustering method allows genes to interact between regulatory pathways and across different conditions at different levels of detail. Fuzzy cluster centers can be used to quickly discover causal relationships between groups of coregulated genes. Fuzzy measures weight expert knowledge and help quantify uncertainty about the functions of genes using annotations and the gene ontology database to confirm some of the interactions. The method is illustrated using gene expression data from an experiment on carbohydrate metabolism in the model plant, *Arabidopsis thaliana*. Key gene regulatory relationships were evaluated using information from the Gene Ontology database. A new regulatory relationship concerning trehalose regulation of carbohydrate metabolism was also discovered in the extracted network.

**Index Terms**—fuzzy logic, microarray analysis, gene expression networks, fuzzy clustering.

## I. INTRODUCTION

THE behavior of biological systems is inherently fuzzy. Genes influence one another and are active at different level to different degrees. Many organisms have had their genomes completely sequenced, making it possible to begin to identify all the genes and their function in the organism. The major challenge in the post-genome era is to understand how interactions among molecules in a cell determine its form and function. This points to the need to develop methodologies to identify and analyze the complex biological networks that regulate metabolism. Metabolic networks form the basis for the net accumulation of biomolecules in living organisms. Regulatory networks modulate the action of these metabolic networks, leading to physiological and morphological changes. Even though new high-throughput transcriptomic,

Manuscript received May 31, 2004. This work was supported in part grants from the National Science Foundation in the Arabidopsis 2010 (DBI-0209809) and Information Technology Research (IBN-0219366) Programs. Seed funding was also provided by the Iowa State University Plant Sciences Institute and the Roy J. Carver Foundation. All of the authors are affiliated with the Virtual Reality Applications Center, Iowa State University.

Pan Du, Jian Gong and Julie A. Dickerson are with the Electrical and Computer Engineering Department, Iowa State University, Ames, IA 50011-3060 USA (e-mail: julied@iastate.edu).

Eve Syrkin Wurtele, is with the Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011 US (e-mail: mash@iastate.edu).

proteomic, and metabolomic analysis technologies give biologists vast amounts of valuable data, techniques that model uncertainty are needed to cope with the many genes of uncertain function and to understand complex interactions.

Gene expression (or transcriptomic) data in the form of high-throughput microarray experiments measures the amount of RNA associated with each of thousands of genes in parallel. The expression of each gene, as reflected by level of accumulation of the corresponding RNA, is not just turned on and off like a light switch. Clustering analysis has been used to hypothesize gene function under the assumption that genes that show similar expression patterns must be coregulated or part of the same regulatory pathway. Fuzzy clustering methods allow genes to belong to multiple clusters and participate in multiple pathways, thus reflecting the known biological reality of cellular metabolism. Fuzzy systems also aid in incorporating known information about some genes into the network.

Gene expression networks show how genes regulate metabolism. Previous work used different machine learning methods to construct hypothetical networks. These methods produced high numbers of false positive connections due to inadequate sampling of the biological process in time and the on/off assumption described previously. In order to get biological meaningful results, information must be combined from a variety of sources to construct networks. Such fuzzy expert knowledge includes databases of genes and their products, as well as information about the interactions that occur between them. This work models the interactions between genes in gene regulatory pathways using fuzzy weights.

## II. BACKGROUND

### A. Transcriptomics Data

Gene expression describes the transcription of the information contained within the DNA, the repository of genetic information, into messenger RNA (mRNA) molecules. mRNA molecules are then translated (Here “translate” means that messenger RNA directs the amino acid sequence of a growing polypeptide during protein synthesis) into the proteins that perform most of the critical functions of cells. The analysis of the types and quantities of mRNAs produced by a cell (transcriptomics) indicates which genes are transcribed under specific conditions. Gene expression is a

highly complex and tightly regulated process that allows a cell to respond dynamically both to environmental stimuli and to its own changing needs. This mechanism controls which genes are expressed in a cell and acts as a “volume control” that increases or decreases the level of expression of particular genes as necessary [1]. Fuzzy metrics can express both concepts simultaneously. The challenge currently facing biological researchers is to discover the functions of the genes and how they interact.

DNA microarray technology exploits the ability of a given mRNA molecule to bind specifically to, or hybridize to, the DNA template from which it originated. Microarrays allow scientists to measure, in a single experiment, the expression levels of thousands of genes within a cell. The amount of mRNA bound to the spots on the microarray is precisely measured, generating a profile of RNAs accumulated in the cell. This work uses microarray data from the Affymetrix Arabidopsis ATH1 genome array, that analyzes 22K genes at a time [2].

Researchers use microarrays to detect expression patterns—the extent to which each particular gene(s) is being expressed more or less under a set of specific circumstances. These gene expression patterns can give insights into the gene functions and the underlining gene regulatory networks.

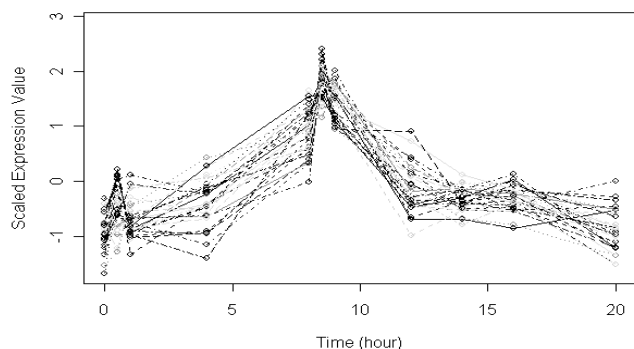


Figure 1. Coregulated gene expression patterns behave similarly across a range of conditions. In this example, the index is hours into a short growth day. The expression values are normalized to a mean of zero and a standard deviation of one. A cluster window size of  $\sigma = 0.1$  was used.

### B. Finding Patterns in Microarray Data

In related biological processes, many genes are highly coregulated (i.e., their gene expression patterns are similar). Figure 1 shows an example of highly coregulated gene expression profiles in the diurnal biological process of the model plant *Arabidopsis* (a member of the mustard family, widely used as a model organism in plant biology)[3, 4]. Clustering is widely used to find these coregulated genes[5-8]. Many popular cluster algorithms are hard clustering algorithms, e.g., hierarchical clustering or k-means clustering. In these algorithms, a gene can only belong to one cluster. In actuality, a single gene may be involved in different biological processes. Furthermore, gene expression patterns may be similar only under a subset of conditions. Hard clustering algorithms cannot extract the gene relations described above.

Fuzzy K-means uses membership values to measure the relationship between a gene and its clusters [9, 10]. As a result, a gene can belong to several clusters to a degree.

Clustering, by itself, does not delineate the causal relationship between genes. RNA profiles are very noisy and may be unequally sampled in time. Using cluster centers, instead of individual gene expression profiles, smoothes by averaging individual gene profiles within the cluster. This is equivalent to a low-pass filter. Thus, clusters of highly coregulated genes can be modeled as a single entity when inferring the gene regulatory relations. A gene transcription response usually can occur in from tens of minutes to several hours, so time delay correlation can help determine the causal relationship.

### C. Gene Regulatory Networks

Regulatory networks reflect causal interactions among biomolecules in living systems. Gene regulatory networks can be defined as regulatory networks that consider transcriptomics data. Several types of models have been proposed for representing regulatory networks in biological systems, including Boolean networks [11, 12], linear weighting networks [13], differential equations [14], and Bayesian Networks [15-17]. Circuit simulations and differential equations require detailed information that is not yet known about the regulatory mechanisms between entities. Boolean networks analyze binary state transition matrices to look for patterns in gene expression. Each part of the network is either on or off depending on whether a signal exceeds a pre-determined threshold. Generalized Logical Networks [18-20] allow the variables in Boolean networks to have more than two values and use generalized Boolean functions to define the relationship. Probabilistic Boolean Networks combine several promising predictors or Boolean functions together, so that each makes a contribution to the prediction of a target gene. A probabilistic model randomly selects one of these promising predictors. Linear weighting networks have the advantage of simplicity since they use simple weight matrices to additively combine the contributions of different regulatory elements. Bayesian networks model probabilistic transitions between network states. Bayesian networks assume that there are no cycles in a network. However cycles are the major mechanism to ensure stability or homeostasis. Dynamic Bayesian Networks combine the features of Hidden Markov Models to incorporate feedback [15-17].

This work models interactions (also referred to as edges or links) in the network as fuzzy functions that depend on the detail known about the network. Fuzzy cognitive maps are fuzzy digraphs that model causal flow between concepts [21] or, in this case, biomolecular entities, including RNAs, metabolites, and proteins [22, 23]. Entities stand for causal fuzzy sets where events occur to some degree. The entities are linked by interactions that show the degree to which these entities depend on each other. Interactions stand for causal flow. The sign of an interaction (+ or -) shows causal coregulation between entities. The fuzzy structure allows the

entities levels to be expressed as continuous values. This modeling has demonstrated regulation in the Arabidopsis network, in the case of gibberellin conversion from an inactive form to an active form [23]. Fuzzy cognitive maps (FCMs) have the potential to deal with the lack of quantitative information on how different variables interact. The FCModeler tool uses fuzzy methods for modeling networks and interprets the results using fuzzy cognitive maps. The FCModeler tool is intended to capture the intuitions of biologists, help test hypotheses, and provide a modeling framework for assessing the large amounts of data captured by RNA microarrays and other high-throughput experiments [24].

For regulatory network modeling, there are a number of significant problems. All of these models are based on information about the quantities of one or more classes of entities. However, these values alone cannot give a complete picture of how the metabolism of living things works [25]. The number of measurements for each object is very limited due to experimental constraints. This is true especially for the complex models and large-scale networks. This makes it difficult to get enough data to use classical machine learning approaches.

Another difficulty is that different models and algorithms often produce different results. It is important to interpret the resulting network model from a biological viewpoint. The Gene Ontology (GO: <http://www.geneontology.org>) provides a way to do this [26, 27]. GO is a shared, controlled vocabulary that is being developed to cover all organisms. GO has three categories: molecular function (MF), biological process (BP), and cellular component (CC). The existence of GO is not only providing us a controlled vocabulary, but paved another way to gene function prediction, clustering interpretation, and evaluation [28]. This work uses an additive fuzzy system to assess the evidence for gene function in a cluster and for the interactions in gene regulatory networks.

### III. ANALYSIS METHODS

The analysis and creation of gene regulatory networks involves first clustering the data at different levels, then searching for weighted time correlations between the cluster center time profiles. The link validity and strength is then evaluated using a fuzzy metric based on evidence strength and co-occurrence of similar gene functions within a cluster.

#### A. Multi-scale Fuzzy K-Means Clustering

The Fuzzy K-means algorithm minimizes the objective function [9, 10]:

$$J(F, V) = \sum_{i=1}^N \sum_{j=1}^K m_{ij}^2 d_{ij}^2 \quad (1)$$

$F = \{X_i, i = 1, \dots, N\}$  are the  $N$  data samples;  $V = \{V_j, j = 1, \dots, K\}$  represent the  $K$  cluster centers.  $m_{ij}$  is the membership of  $X_i$  in cluster  $j$ , and  $d_{ij}$  is the Euclidean

distance between  $X_i$  and  $V_j$ . One commonly used fuzzy membership function is adapted as:

$$m_{ij} = \frac{1/d_{ij}^2}{\sum_{k=1}^K 1/d_{ik}^2} W(d_{ij}) \quad (2)$$

where  $W(d)$  is the window function centered at  $V_j$  and can take any form. Adding a window function  $W(d)$  to the membership function limits the size of clusters. This work uses truncated Gaussian windows with values outside the range of  $3\sigma$  set to zero:

$$W(d_{ij}) = \begin{cases} e^{-(d_{ij})^2 / (2\sigma^2)} & d_{ij} < 3\sigma \\ 0 & \text{elsewhere} \end{cases} \quad (3)$$

The window function  $W(d)$  insures that genes with distances larger than  $3\sigma$  will have no effect on the cluster centers.

#### 1) Multi-scale Algorithm

The multi-scale algorithm is similar to the ISODATA algorithm with cluster splitting and merging [29, 30]. There are four parameters:  $K$  (initial cluster number),  $\sigma$  (scale of the window  $W(d)$ ),  $T_{split}$  (split threshold),  $T_{combine}$  (combine threshold). Whenever the genes are further away from the cluster center than  $T_{split}$ , the cluster is split and faraway genes form new clusters. Also, if two cluster centers are separated by less than  $T_{combine}$ , then the clusters are combined. Usually  $T_{combine} \leq \sigma$  and  $2\sigma \leq T_{split} \leq 3\sigma$ . The algorithm is given in Table I.  $\varepsilon_1$  and  $\varepsilon_2$  are small numbers to determine whether the clustering converged. The advantage of this algorithm is that it dynamically adjusts the number of clusters based on the splitting and merging heuristics.

TABLE I. MULTI-SCALE FUZZY K-MEANS ALGORITHM

1	Initialize parameters: $K$ , $\sigma$ , $T_{split}$ and $T_{combine}$
2	Iterate using Fuzzy K-means until convergence to threshold $\varepsilon_1$
3	Split process: do split if there are elements farther away from cluster center than $T_{split}$ .
4	Iterate using Fuzzy K-means until convergence to threshold $\varepsilon_1$
5	Combine Process: combine the clusters whose distance between cluster centers is less than $T_{combine}$ . If the cluster after combining has elements far away from cluster center (distance larger than $3\sigma$ ), stop combining.
6	Iterate steps 1-5 until converging to a given threshold $\varepsilon_2$ .

#### 2) Effects of window size

Changing the window size can affect the level of detail captured in the clusters. If  $\sigma \ll 1$ , then clusters are individual elements. As  $\sigma$  increases, the window gets larger. The result is a hierarchical tree that shows how the clusters interact at different levels of detail. This work uses three level of multi-scale fuzzy K-mean clustering ( $\sigma = 0.1, 0.2$  and  $0.3$ ). The

initial number of clusters is  $K = N$ , the total number of data points,  $T_{combine} = \sigma$ , and  $T_{split} = 3\sigma$ . Clustering results with different window sizes provide different levels of information. At  $\sigma = 0.1$ , the cluster sizes are very small. These clusters represent very highly correlated profiles (correlation coefficients between gene profiles within one  $\sigma$  window size are larger than 0.9) or just the individual gene profiles because many clusters only contain a single element. At  $\sigma = 0.2$ , smaller clusters are combined with nearby clusters. Highly correlated profiles are detected. The  $\sigma = 0.3$  level is the coarsest level.

### B. Construction of gene regulatory networks

Clustering provides sets of genes with similar RNA profiles. The next step is finding the relationships among these coregulated genes. If gene  $A$  and gene  $B$  have similar expression profiles, there are several possible relationships: 1.  $A$  and  $B$  are coregulated by other genes; 2.  $A$  regulates  $B$  or vice versa; 3. There is no causal relationship, just coincidence. Here, the regulation may be indirect, i.e., interaction through intermediates. These cases cannot be differentiated solely by clustering. Cubic spline interpolation generates equally sampled profiles as in [31].

The gene regulatory model can be simplified as a linear model [32]:

$$x_A(t + \tau_A) = \sum_B w_{BA} x_B + b_A \quad (4)$$

where  $x_A$  is the expression level of gene  $A$  at time  $t$ ,  $\tau_A$  is the gene regulation time delay of gene  $A$ ,  $w_{BA}$  is the weight indicating the inference of gene  $B$  to  $A$ ,  $b_A$  is a bias indicating the default expression level of gene  $A$  without regulation.

Standardizing gene expression profiles to 0 mean and 1 standard deviation removes  $b_A$  from equation (4). The goal is to find out if genes  $A$  and  $B$  have a regulatory relationship, the weight  $w_{BA} = [0, 1]$  (0 means no regulatory relation, 1 means strongly regulated). The time correlation between genes  $A$  and  $B$  can be expressed in discrete form as:

$$R_{AB}(\tau) = \sum_n x_A(n) x_B(n - \tau) \quad (5)$$

where  $x_A$  and  $x_B$  are the standardized (zero mean, standard deviation of unity) expression profiles of genes  $A$  and  $B$ .  $\tau$  is the time shift. For a periodic time profile, we can use circular time correlation, i.e., the time points at the end of the time series will be rewound to the beginning of series after time shifting. For multiple data sets, the time correlation results of each data set are combined as:

$$R_{AB}^C(\tau) = \sum_k w_k R_{AB}^k(\tau) \quad (6)$$

where  $R_{AB}^C(\tau)$  is the combined time correlation result,  $R_{AB}^k(\tau)$  is the time correlation result of the  $k^{\text{th}}$  data set,  $w_k$  is the weight of  $k^{\text{th}}$  data set that depends on the experiment reliability and the length of the expression profile.

The value  $\max |R_{AB}^C(\tau)|$  can be used to estimate the time delay  $\tau'$  between expression profiles of genes  $A$  and  $B$ . Given a correlation threshold  $T_R$ , if  $\max |R_{AB}^C(\tau)| > T_R$ , there is significant regulation between genes or clusters. By defining the clusters as nodes and significant links as edges, we can get the gene regulation network of these clusters. We can define four types of regulation:

$R_{AB}^C(\tau') > 0, \tau' \neq 0$ , positive regulation between genes  $A$  and  $B$ ;

$R_{AB}^C(\tau') < 0, \tau' \neq 0$ , negative regulation between genes  $A$  and  $B$ ;

$R_{AB}^C(\tau') > 0, \tau' = 0$ , genes  $A$  and  $B$  are positively coregulated;

$R_{AB}^C(\tau') < 0, \tau' = 0$ , genes  $A$  and  $B$  are negatively coregulated.

The sign of  $\tau'$  determines the direction of regulation.  $\tau' > 0$  means gene  $B$  regulates gene  $A$  with time delay  $\tau'$ ;  $\tau' < 0$  means gene  $A$  regulates gene  $B$  with time delay  $\tau'$ .

### C. Network evaluation using fuzzy metrics

The available gene ontology (GO) annotation information can estimate a fuzzy measure for the types or functions of genes in a cluster. The GO terms in each cluster are weighted according to the strength of the supporting evidence information and the distance to cluster center. An additive fuzzy system is used to combine this information [33]. Every GO annotation indicates the type of supporting evidence. This evidence is used to set up a bank of fuzzy rules for each annotated data point. Different fuzzy membership values are given to each evidence code. For example, evidence inferred by direct assays (IDA) or from a traceable author statement (TAS) in a refereed journal has a value of one. The least reliable evidence is electronic annotation which is known to have high rates of false positives.

Each gene in a cluster is weighted by the Gaussian window function in equation (3). This term weights the certainty of the gene's GO annotation using product weighting. Each gene and its associated GO term are combined to find the possibility distribution for each single GO term that occurs in the GO annotations in one cluster. One gene may be annotated by several GO terms, and each GO term has one evidence code. Each GO term may occur  $K$  times in one cluster, but with a different evidence code and in different genes. For the  $n^{\text{th}}$  unique GO term in the  $j^{\text{th}}$  cluster, the fuzzy weight is the sum of the weights for each occurrence of the term:

$$W_{GO}(j, n) = \sum_{i=1}^K w_{GO, j}(i, n) \quad (7)$$

where  $w_{GO, j}(i, n) = w_{evi}(i, n) \cdot W(d_{ij})$ ,  $w_{evi}$  is given in table II, and  $W(d_{ij})$  is the same as equation (3).

TABLE II: EVIDENCE CODES AND THEIR WEIGHTS  
([HTTP://WWW.GEONTOLOGY.ORG/GO.EVIDENCE.HTML](http://www.geneontology.org/GO.evidence.html))

EVIDENCE CODE	MEANING OF THE EVIDENCE CODE	EVIDENCE WEIGHT, $W_{EVI}$
IDA	Inferred from direct assay	1.0
TAS	Traceable author statement	1.0
IMP	Inferred from mutant phenotype	0.9
IGI	Inferred from genetic interaction	0.9
IPI	Inferred from physical interaction	0.9
IEP	Inferred from expression pattern	0.8
ISS	Inferred from structural similarity	0.8
NAS	Non-traceable author statement	0.7
IEA	Inferred from electronic annotation	0.6
	Other	0.5

This provides a method of pooling uncertain information about gene function for a cluster of genes. This gives an additive fuzzy system that assesses the credibility of any GO terms associated to a cluster [33]. The results can be left as a weighted fuzzy set or be defuzzified by selecting the most likely annotation. For each cluster, the weight is normalized by the maximum weight and the amount of unknown genes. This is the weighted percentage of each GO term  $p_{weight}$ :

$$p_{weight}(j, n) = \frac{W_{GO}(j, n)}{W_{root}(j) - W_{unknown}(j)} * 100\% \quad (8)$$

where  $W_{GO}(j, n)$  represents the weight of the  $n^{\text{th}}$  GO term in the  $j^{\text{th}}$  cluster.  $W_{unknown}(j)$  is the weight of GO term in cluster  $j$ : xxx unknown, e.g., GO: 0005554 (molecular\_function unknown).  $W_{root}(j)$  is the weight of root in cluster  $j$ . GO terms are related using directed acyclic graphs. The root of the graph is the most general term. Terms further from the root provide more specific detail about the gene function and are more useful for a researcher. The weight of each node is computed by summing up the weights of its children (summing the weights of each of the  $N$  GO terms in a cluster):

$$W_{root}(j) = \sum_{n=1}^N W_{GO}(j, n) \quad (9)$$

The higher weighted nodes further from the root are the most interesting since those nodes refer to specific biological processes.

#### IV. FUZZY CLUSTERING RESULTS

The tested data set compared *Arabidopsis thaliana* plants, wild-type (WT) and transgenic plants containing antisense *ACLA-1* behind the constitutive CaMV 35S promoter (referred to as a*ACLA-1*). The microarray type was an Affymetrix GeneChip. The data consisted of two replicates; each with eleven time points (0, 0.5, 1, 4, 8, 8.5, 9, 12, 14, 16, 20 hours), and changing from light (from 0 to 8 hours) to dark

(from 8 to 20 hours) [3, 4]. Only *ACLA-1* seedlings exhibiting features characteristic of the antisense phenotype were used. Total RNA was extracted from leaves and used for microarray analyses.

The Affymetrix microarray data were normalized with the Robust Multichip Average (RMA) method [34]. The replicates of each gene expression profile are standardized to zero mean, one standard deviation. The data was filtered by comparing the expression values between the WT and *ACLA1* gene mutated at 1, 8.5 and 12 hours. Differentially expressed genes having fold changes larger than 2 times at any of the time points 1, 8.5 and 12 hours were kept. 484 genes remained after filtering. The gene expression patterns used for clustering are the time point measurements for the wild-type plant.

The data was combined so that each data point consists of a gene evaluated at a series of time points. Three-level multi-scale fuzzy k-means clustering was used, with window sizes of  $\sigma = 0.1, 0.2$ , and  $0.3$ . The initial number of clusters,  $K$ , was the number of genes. There were 236 clusters at  $\sigma = 0.1$ ; 28 clusters at  $\sigma = 0.2$ ; and 5 clusters at the  $\sigma = 0.3$  level.

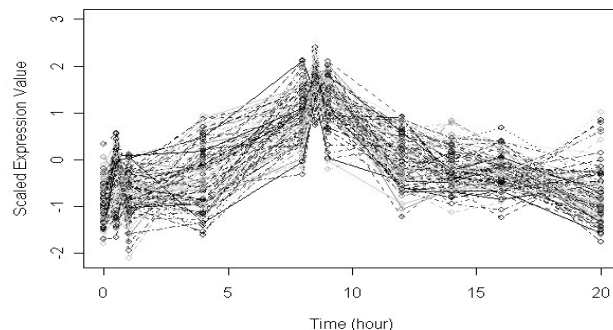


Figure 2. Coregulated gene expression patterns behave similarly across a range of conditions. In this example, the index is hours into a short growth day. The expression values are normalized to a mean of zero and a standard deviation of one. A cluster window size of  $\sigma = 0.2$  was used.

Figures 1 and 2 show typical cluster patterns for at  $\sigma = 0.1$  and  $0.2$  window sizes respectively. The cluster in figure 1 is much more tightly coregulated than figure 2 with less variation. Figure 3 shows the cluster center profiles of 5 cluster centers at the  $\sigma = 0.3$  level. At this coarse level, information such as whether the gene expression level increases or decreases in the day or night is given. Figure 3 shows that clusters 2 and 3 decrease in the day and increase at night, while cluster 1, 4 and 5 are opposite. At  $\sigma = 0.2$ , the regulatory relationships can be studied at a more detailed level. There are 28 clusters at this level. Figure 4 shows their relationship with the  $\sigma = 0.3$  clustering. Several clusters from  $\sigma = 0.2$  belong to more than one cluster at  $\sigma = 0.3$ . This is due to genes in these sub-clusters being involved in multiple related biological processes. In figure 3, clusters 2 and 3 represent the genes active at night, and clusters 1, 4, and 5 are active in the day. Figure 4 shows that genes active at night are more tightly coregulated than those active in the day.

Biologically, this indicates the ACLA1 related genes are mainly active in the day and their expressions are diversified. At  $\sigma = 0.1$ , clusters were further subdivided into 236 clusters. Many of these clusters only included 1 or 2 genes. Given the noise in microarray experiments and the small number of genes in each cluster, we did not further study at this level.

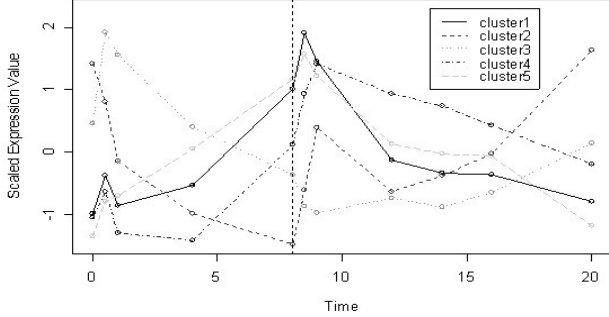


Figure 3. Cluster center profiles for the window size  $\sigma = 0.3$  level.

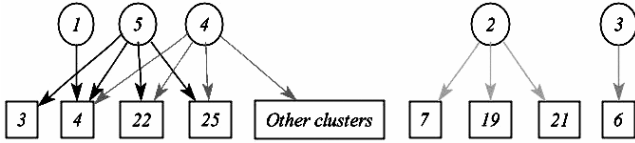


Figure 4. Relationship between the clusters from the  $\sigma = 0.2$  case (cluster numbers in rectangles) and the clusters in the  $\sigma = 0.3$  case (cluster numbers in circle).

## V. INFERRING AND MODELING GENE REGULATORY NETWORKS

### A. Construct the genetic network using time correlation

The genetic networks among the clusters of highly coregulated genes can be constructed based on their cluster center profiles. Since the data used were unequally sampled with 0.5h as minimum interval, we interpolated the gene expression profiles as equally sampled 41 time points with 0.5h intervals using cubic spline interpolation. The time correlation of each replicate  $R_{ij}^k(\tau)$ ,  $k=1, 2$  was computed using equation (5), then combined using equation (6) as  $R_{ij}^C(\tau)$  with weight  $w_k = 0.5, k=1, 2$ .  $\tau$  was limited to the range of [-4h, 4h] because the light period only lasted 8 hours in this data set. The genetic networks were constructed with a correlation threshold of  $T_R = 0.65$ . The strength of correlation was mapped into three categories: [0.65, 0.75), [0.75, 0.85), and [0.85, 1]. Three types of line thickness from thin to thick represent the strength of the correlation. Dark dashed lines represent positive coregulation; gray dashed lines represent negative coregulation; solid lines with bar head represent negative regulation; solid lines with arrowheads represent positive regulation.

Figure 5 shows the constructed gene regulatory networks

based on the cluster center profiles shown in figure 3. The networks indicate clusters 1 and 5 are highly coregulated (0 time delay), clusters 1 and 5 positively regulate cluster 4 with time delays of 2.5h and 3h, and both negatively regulated cluster 3 with a time delay of 1.5h; cluster 4 is negatively regulated by cluster 3 with delay 1h, the correlation between cluster 2 and cluster 4, and cluster 1 and 3 is not strong. All of these relations are correspond to the cluster center profiles. This means the algorithm correctly resolved the relationships between cluster centers.

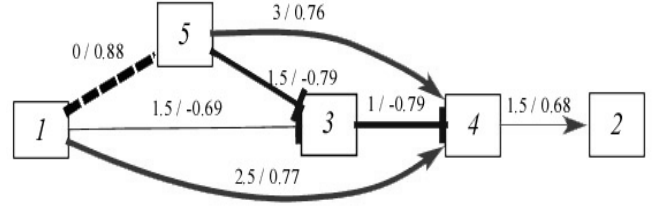


Figure 5. Gene regulatory networks inferred from the case with  $\sigma = 0.3$ . The numbers on each link show the time delay for the interaction on top and the correlation coefficient of the interaction on the bottom.

Figure 6 shows the constructed regulatory networks of the 28 cluster centers at  $\sigma = 0.2$  level. The graph notations are the same as in figure 5. The graph shows that there is one highly connected group of clusters. The other clusters at the upper right corner are less connected. The relations between clusters may become complex with a large number of edges. Simplification of the networks is necessary when there are many highly connected clusters.

Figure 6 shows possible duplicate relationships. This can be analyzed using the path search function in FCModeler. In figure 6, from cluster 15 to 19, there are two paths: one is directly from cluster 15  $\rightarrow$  19 with time delay 1h and correlation coefficient,  $\rho = -0.85$ ; another path is cluster 15  $\rightarrow$  7 with time delay 0.5 h and correlation coefficient,  $\rho = -0.89$ , and then from 7  $\rightarrow$  19 with time delay 0.5h and  $\rho = 0.81$ . The total time delays of both paths are the same. So it is very possible one of the paths is redundant. Figure 7 shows part of the simplified graph of figure 6.

### B. Cluster and Network Evaluation using Weighted GO Terms

Cluster evaluation makes use of the available GO information to find out what kind of functions or processes a cluster involves. In figure 6, the graphs in the upper right corner are less connected. The Gene Ontology shows most of these clusters are not annotated. This means these clusters have no biological evidence of direct relation with the highly connected group. It also shows how the multi-scale fuzzy algorithm successfully separates those unrelated genes.

Figure 7 shows that cluster 3 and 4 are highly coregulated (correlation coefficient between cluster centers is 0.91). The cluster is split because the combined cluster 3 and 4 has a cluster radius larger than  $3\sigma$ . Table III shows the fuzzy weights for the GO terms in each cluster. The BP (Biological



compounds, and signal transducer activity.

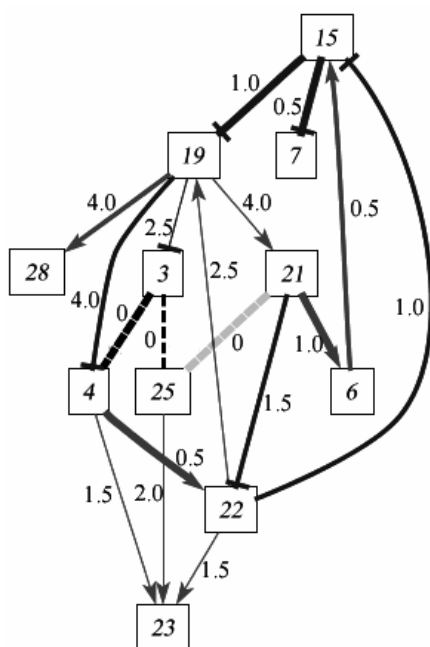


Figure 7. Simplified regulatory networks with redundant edges removed for the window size  $\sigma = 0.2$  level. The number on each link represents the estimated time delay.

Most of the clusters have the following molecular functions: binding, catalytic activity, and transcription regulator activity. Clusters 3 and 4 are the most similar clusters in the sense of molecular function. The largest weight is on DNA binding, and they both include: purine nucleotide binding, oxygen binding, and carbohydrate binding. Also, both clusters contain active genes that attend transferase activity (transferring phosphorus-containing groups), hydrolase activity (acting on glycosyl bonds), and oxidoreductase activity. The only difference is that cluster 4 contains genes acting in transporter activity.

## VI. CONCLUSIONS AND FUTURE WORK

Fuzzy logic can be applied to all aspects of gene regulatory network analysis from clustering to assessing network credibility. Multi-scale fuzzy k-means clustering provides the cluster information in different scales and captures interactions in terms of gene function and across regulatory pathways. It makes the results more reliable. The regulatory network construction algorithm uses the cluster centers efficiently to evaluate the time delay information. The algorithm also allows feedback in the networks, which most qualitative regulatory network algorithms cannot provide at present. Visualizing the cluster relationships helps show biological interactions. GO and pathway evaluations indicate the algorithm is promising and demonstrate that it yields detailed biological hypotheses of the regulatory connections with known metabolic networks. Future work will focus on

integrating the regulatory network model with existing metabolic networks to simulate cellular processes.

## ACKNOWLEDGMENT

The network visualization was performed using the facilities at the Virtual Reality Application Center at Iowa State University. The authors would like to thank Dr. Carol Foster, and Ling Li for kindly making their microarray expression data available for this work.

## REFERENCES

- [1] National Center for Biotechnology Information (NCBI), "Microarrays: Chipping Away At The Mysteries Of Science And Medicine," vol. 2004: NCBI, 2004.
- [2] Affymetrix Inc., "Statistical Algorithms Reference Guide," Affymetrix, Inc., Santa Clara, CA 701110 Rev 1, 2001.
- [3] B. F. Fatland, J. Ke, M. Anderson, W. Mentzen, L. W. Cui, C. Allred, J. L. Johnston, B. J. Nikolau, and E. S. Wurtele, "Molecular Characterization of a Novel Heteromeric ATP-Citrate Lyase that Generates Cytosolic Acetyl-CoA in Arabidopsis," *Plant Physiology*, vol. 130, pp. 740-756, 2002.
- [4] C. M. Foster, L. Ling, A. M. Myers, M. G. James, B. J. Nikolau, and E. S. Wurtele, "Expression of genes in the starch metabolic network of Arabidopsis during starch synthesis and degradation," *In preparation.*, 2004.
- [5] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings National Academy of Science*, vol. 95, pp. 14863-14868, 1998.
- [6] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, 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," *Molecular Biology of the Cell*, vol. 9, pp. 3273-3297, 1998.
- [7] V. R. Iyer, M. B. Eisen, D. T. Ross, G. Schuler, T. Moore, J. C. F. Lee, J. M. Trent, L. M. Staudt, J. Hudson Jr., M. S. Boguski, D. Lashkari, D. Shalon, D. Botstein, and P. O. Brown, "The Transcriptional Program in the Response of Human Fibroblasts to Serum," *Science*, vol. 283, pp. 83-87, 1999.
- [8] T. Hastie, R. Tibshirani, M. B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W. C. Chan, D. Botstein, and P. O. Brown, "Gene shaving" as a method for identifying distinct sets of genes with similar expression patterns," *Genome Biology*, vol. 1, pp. research0003.1-0003.21, 2000.
- [9] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press, 1981.
- [10] A. P. Gasch and M. B. Eisen, "Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering," *Genome Biol*, vol. 3, pp. RESEARCH0059, 2002.
- [11] S. Liang, S. Fuhrman, and R. Somogyi, "REVEAL, A general reverse engineering algorithm for inference of genetic network architectures," presented at Pacific Symposium on Biocomputing 3, Hawaii, 1998.
- [12] T. Akutsu, S. Miyano, and S. Kuhara, "Identification of Genetic Networks from a Small Number of Gene Expression Patterns Under the Boolean Network Model," presented at Pacific Symposium on Biocomputing 4, Hawaii, 1999.
- [13] D. C. Weaver, C. T. Workman, and G. D. Stormo, "Modeling Regulatory Networks with Weight Matrices," presented at Pacific Symposium on Biocomputing 4, Hawaii, 1999.
- [14] T. Akutsu, S. Miyano, and S. Kuhara, "Algorithms for Inferring Qualitative Models of Biological Networks," presented at Pacific Symposium on Biocomputing 5, Hawaii, 2000.
- [15] K. Murphy, Mian, S., "Modelling Gene Expression Data using Dynamic Bayesian Networks," Computer Science Division, University of California, Berkeley 1999.

- [16] K. P. Murphy, "Dynamic Bayesian Networks: Representation, Inference and Learning," in *Computer Science: UNIVERSITY OF CALIFORNIA, BERKELEY*, 2002, pp. 255.
- [17] B. E. Perrin, L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, and F. D'Alche-Buc, "Gene networks inference using dynamic Bayesian networks," *Bioinformatics*, vol. 19 Suppl 2, pp. III138-III148, 2003.
- [18] R. Thomas, D. Thieffry, and M. Kaufman, "Dynamical behaviour of biological regulatory networks--I. Biological role of feedback loops and practical use of the concept of the loop-characteristic state," *Bull Math Biol*, vol. 57, pp. 247-76, 1995.
- [19] L. Mendoza and E. R. Alvarez-Buylla, "Dynamics of the genetic regulatory network for Arabidopsis thaliana flower morphogenesis," *J Theor Biol*, vol. 193, pp. 307-19, 1998.
- [20] L. Mendoza, D. Thieffry, and E. R. Alvarez-Buylla, "Genetic control of flower morphogenesis in Arabidopsis thaliana: a logical analysis," *Bioinformatics*, vol. 15, pp. 593-606, 1999.
- [21] J. A. Dickerson and B. Kosko, "Virtual Worlds as Fuzzy Cognitive Maps," *Presence*, vol. 3, pp. 173-189, 1994.
- [22] Z. Cox, A. Fulmer, and J. A. Dickerson, "Interactive Graphs for Exploring Metabolic Pathways," presented at ISMB, 2002, Edmonton, CA, 2002.
- [23] J. A. Dickerson, Z. Cox, E. S. Wurtele, and A. W. Fulmer, "Creating Metabolic and Regulatory Network Models using Fuzzy Cognitive Maps," presented at North American Fuzzy Information Processing Conference (NAFIPS), Vancouver, B.C., 2001.
- [24] J. A. Dickerson, D. Berleant, Z. Cox, W. Qi, and E. Wurtele, "Creating Metabolic Network Models using Text Mining and Expert Knowledge," presented at Atlantic Symposium on Molecular Biology and Genome Information Systems and Technology (CBGIST 2001), Durham, North Carolina, 2001.
- [25] V. Hatzimanikatis and K. H. Lee, "Dynamical Analysis of Gene Networks Requires Both mRNA and Protein Expression Information," *Metabolic Engineering*, vol. 1, pp. 275-281, 1999.
- [26] J. Blake and M. Harris, "The Gene Ontology Project: Structured vocabularies for molecular biology and their application to genome and expression analysis," in *Current Protocols in Bioinformatics*, D. B. D. A.D. Baxevanis, R. Page, G. Stormo and L. Stein, Ed. New York: Wiley and Sons, Inc., 2003.
- [27] M. Ashburner and S. Lewis, "On ontologies for biologists: the Gene Ontology - uncoupling the web," *In Silico Biology Novartis Found Symp*, vol. 247, pp. 66-80; discussion 80-3, 84-90, 244-52, 2002.
- [28] F. Al-Shahrour, R. Diaz-Uriarte, and J. Dopazo, "FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes," *Bioinformatics*, vol. 20, pp. 578-80, 2004.
- [29] G. H. Ball, "Data analysis in the social sciences: what about the details," *AFIPS Proc. Cong. Fall Joint Comp.*, vol. 27, pp. 533-559, 1965.
- [30] G. H. Ball and D. J. Hall, "ISODATA, a novel method of data analysis and pattern classification," Stanford Research Institute, Technical Report 1965.
- [31] P. D'Haeseleer, "Reconstructing Gene Networks from Large Scale Gene Expression Data," in *Computer Science*. Albuquerque, NM: The University of New Mexico, 2000, pp. 207.
- [32] P. D'Haeseleer, S. Liang, and R. Somogyi, "Gene expression analysis and modeling," *Pac Symp Biocomput*, 1999.
- [33] B. Kosko, *Neural Networks and Fuzzy Systems*. Englewood Cliffs: Prentice Hall, 1992.
- [34] L. Gautier, L. Cope, B. Bolstad, and R. Irizarry, "affy--analysis of Affymetrix GeneChip data at the probe level," *Bioinformatics*, vol. 20, pp. 307-315, 2004.
- [35] P. J. Eastmond and I. A. Graham, "Trehalose metabolism: a regulatory role for trehalose-6-phosphate?," *Curr Opin Plant Biol*, vol. 6, pp. 231-5, 2003.



**Pan Du** received the B.S. and M.S. degrees in Electrical Engineering from National University of Defense Technology, Changsha, China, in 1995 and 1998, respectively.

He is currently a co-major Ph.D. student in Electrical Engineering major and Bioinformatics and Computational Biology major at Department of Electrical and Computer Engineering, Iowa State University. His research interests include systems biology, genetic network modeling and inference, microarray data analysis, signal processing and pattern recognition.



**Jian Gong** received the B.S. degree in the Department of Instrumentation and Detection from Xidian University, Xi'an, China, 1998. From 1998 to 1999, she was studying Signal Processing in graduate college, Xidian University, Xi'an, China.

She is currently a masters student at Department of Electrical and Computer Engineering, Iowa State University. Her research interests include communication and signal processing, pattern

recognition, and bioinformatics.



**Eve Syrkin Wurtele** received the B.S. degree in Biology from U.C. Santa Cruz, CA, USA, 1971, and the Ph.D. degree in Biology from U.C. Los Angeles, CA, USA, 1980. She was a Postdoctoral Fellow at Department of Biochemistry, U.C. Davis from 1980 to 1983, Senior Research Scientist at Cell Biology Division, NPI, Inc. from 1983 to 1988. In 1988, she joined the department of Botany and Food Technology as an Affiliate Assistant Professor, Iowa State University. She became an Assistant Professor and Associate

Professor at Department of Botany, Iowa State University in 1990 and 1995, respectively. Since 1998, she has been the Professor at Department of Development & Cell Biology, Iowa State University, Ames, IA, USA.

Dr. Wurtele was the organizer of International Symposium of Metabolic Networking in Plants, April, 1999 (Supported in part by NSF), Elementary School presentations and laboratory experiments, and design of Virtual Plant Cell for teaching. She received the Herman Frasch Foundation Award, American Chemical Society in 1997. She is also the panel member of National Science Foundation, study section participant of National Institute of Health, and the Co-Organizer of Third International Congress on Plant Metabolomics, Iowa State University, June, 2004 (Supported in part by NSF and DOE).



**Julie Dickerson** received her B.S. degree from the University of California, San Diego and her M.S. and Ph.D. degrees from the University of Southern California. She is currently an Associate Professor of Electrical and Computer Engineering at Iowa State University.

Dr. Dickerson designed radar systems for Hughes Aircraft Company and Martin Marietta while getting her Ph.D. Her current research activities are intelligent systems, bioinformatics, pattern recognition, and data visualization. She is a Carver Fellow in the Virtual Reality Applications

Center and a member of the Baker Center for Bioinformatics in the Plant Sciences Institute at Iowa State.

**TABLE III. CLUSTER ANNOTATION OF BIOLOGICAL PROCESS GO**  
 $(W_{root}, W_{GO}(j, n)$  AND  $p_{weight}(j, n)$  AS DEFINED IN EQUATION (9))

CLUSTER INDEX ( $W_{root}$ )	MAJOR GO TERM	$W_{GO}(J,N)$	$P_{WEIGHT}(J,N)$
Cluster 3 (24.81)	Response to water derivation	4.11	16.6
	Regulation of transcription, DNA-dependent	3.16	12.7
	Carboxylic acid metabolism	2.82	11.4
	Protein amino acid phosphorylation	2.63	10.6
Cluster 4 (36.03)	Protein amino acid phosphorylation	8.34	<b>23.1</b>
	Carboxylic acid metabolism	3.58	9.9
	Response to abiotic stimulus	3.35	9.3
	Regulation of transcription, DNA-dependent	2.44	6.8
Cluster 6 (8.48)	Regulation of transcription, DNA-dependent	1.99	<b>23.5</b>
	myo-inositol biosynthesis	0.95	11.2
	Abscisic acid mediated signaling	0.83	9.8
	Protein amino acid phosphorylation	0.57	6.7
Cluster 7 (13.58)	Carbohydrate metabolism	3.02	<b>22.2</b>
	Cell surface receptor linked signal transduction	1.71	12.6
	Nucleobase, nucleoside, nucleotide, and nucleic acid metabolism	1.62	11.9
	Protein amino acid phosphorylation	1.59	11.7
Cluster 15 (2.52)	Regulation of transcription, DNA-dependent	1.32	<b>52.4</b>
	Electron transport	0.7	27.8
Cluster 19 (3.32)	Cell-cell signaling	0.78	23.5
	Response to auxin stimulus	0.68	20.5
	Protein folding	0.65	19.6
	N-terminal protein myristoylation	0.61	18.4
Cluster 21 (9.71)	Carbohydrate metabolism	2.93	<b>29.1</b>
	Response to gibberellic acid stimulus	1.86	19.2
	Photosynthesis, dark reaction	0.91	9.4
Cluster 22 (23.76)	Protein amino acid phosphorylation	6.74	<b>28.4</b>
	Macromolecule biosynthesis	3.38	14.2
	Regulation of transcription DNA-dependent	2.50	10.5
	Signal transduction	2.30	9.7
Cluster 23 (4.61)	Response to endogenous stimulus	2.79	<b>60.5</b>
	Response to biotic stimulus	1.83	39.7
Cluster 25 (39.16)	Carboxylic acid metabolism	8.19	<b>20.9</b>
	Response to pest/pathogen/parasite	5.66	14.5
	Lipid biosynthesis	3.55	9.1
	Transport	3.52	9.0
Cluster 28 (0.95)	Carbohydrate metabolism	0.95	<b>100</b>

**TABLE IV: SUMMARY OF MOLECULAR FUNCTION FOR EACH CLUSTER**

MF LEVEL 2	CLUSTER INDEX											
	3	4	6	7	15	19	21	22	23	25	27	28
Carbohydrate binding	1.3	3.6	0	3.4	0	0	0	3.8	0	0	0	0
Nucleic acid binding	23.7	10.4	41.6	4.4	55.6	0	0	16	0	3.5	0	0
Nucleotide binding	6.14	7.5	0	12.4	0	0	17.6	0	0	4.8	0	0
Protein binding	0	0	0	0	0	0	10.4	0	0	2.8	0	0
Oxygen binding	5.5	4	0	0	15.4	0	0	0	0	9.3	0	0
Lipid binding	0	1.6	0	0	0	0	0	2.7	0	3.4	0	0
Metal ion binding	2.9	1.4	8	0	0	0	0	0	0	0	0	0
Kinase activity	27.2	15.7	6.4	0	0	0	0	0	0	12	0	0
Transferase activity	23.3	19.6	0	28.2	11	0	0	59.3	0	8.6	0	0
Hydrolase activity	15.8	12.8	16.5	27	0	71.9	32.1	8	0	8.1	0	100
Oxidoreductase activity	8.2	8.3	16.8	6	0	0	23.8	11	0	14.1	0	0
Signal transducer activity	0	0	0	7.8	0	28.5	0	0	0	0	0	0
Isomerase activity	0	0	10.7	0	0	0	0	0	0	0	0	0
Transcription regulator activity	14.2	5.7	25.6	0	27.9	0	0	8.9	0	3.5	0	0
Transporter activity	0	1.5	0	0	17.8	0	23.8	0	0	5.4	0	0