GFBA: A Biclustering Algorithm for Discovering Value-Coherent Biclusters *

Xubo Fei[1], Shiyong Lu[1], Horia F. Pop[2], and Lily R. Liang[3]

[1] Dept. of Computer Science, Wayne State University, USA
{xubo,shiyong}@wayne.edu
[2] Dept. of Computer Science, Babes-Bolyai University,Romania
hfpop@cs.ubbcluj.ro
[3] Dept. of Computer Science and IT, University of the District of Columbia, USA
lliang@udc.edu

Abstract. Clustering has been one of the most popular approaches used in gene expression data analysis. A clustering method is typically used to partition genes according to their similarity of expression under different conditions. However, it is often the case that some genes behave similarly only on a subset of conditions and their behavior is uncorrelated over the rest of the conditions. As traditional clustering methods will fail to identify such gene groups, the biclustering paradigm is introduced recently to overcome this limitation. In contrast to traditional clustering, a biclustering method produces biclusters, each of which identifies a set of genes and a set of conditions under which these genes behave similarly. The boundary of a bicluster is usually fuzzy in practice as genes and conditions can belong to multiple biclusters at the same time but with different membership degrees. However, to the best of our knowledge, a method that can discover fuzzy value-coherent biclusters is still missing. In this paper, (i) we propose a new fuzzy bicluster model for value-coherent biclusters; (ii) based on this model, we define an objective function whose minimum will characterize good fuzzy value-coherent biclusters; and (iii) we propose a genetic algorithm based method, Genetic Fuzzy Biclustering Algorithm (GFBA), to identify fuzzy value-coherent biclusters. Our experiments show that GFBA is very efficient in converging to the global optimum.

1 Introduction

Clustering has been one of the most popular approaches used in gene expression data analysis. It is used to group genes according to their expression under multiple conditions or to group conditions based on the expression of a number of genes. When a clustering method is used for grouping genes, it typically partitions genes according to their similarity of expression under all conditions. However, it is often the case that some genes behave similarly only on a subset of conditions and their behavior is uncorrelated over the rest of the conditions. Therefore, traditional clustering methods will fail to identify such gene groups.

* This work was partially supported by the Agricultural Experiment Station at the University of the District of Columbia (Project No.: DC-0LIANG; Accession No.: 0203877).

Consider the gene expression data matrix shown in Table 1. If we consider all conditions, genes 1, 2, and 4 do not seem to behave similarly since their expression values are uncorrelated under condition 2 - while genes 1 and 2 have an increased expression value from condition 1 to condition 2, the expression of gene 4 drops from condition 1 to condition 2. However, these genes behave similarly under conditions 1, 3, and 4 since all their expression values increase from condition 1 to condition 3 and increase again under condition 4. A traditional clustering method will fail to recognize such a cluster since the method requires the three genes to behave similarly under all conditions which is not the case.

Table 1. A sample gene expression data matrix and a hidden bicluster.

	cond. 1	cond. 2	cond. 3	cond. 4
gene1	**0.0**	5.0	**1.0**	**2.0**
gene2	**1.0**	20.0	**2.0**	**3.0**
gene3	10.0	10.0	20.0	6.0
gene4	**2.0**	0.0	**3.0**	**4.0**

To overcome the limitation of traditional clustering, the biclustering paradigm [1] was introduced recently and several biclustering methods have been developed under this paradigm; see [2] for a recent survey of existing bicluster models and methods. In contrast to traditional clustering, a biclustering method produces biclusters, each of which identifies a set of genes and a set of conditions under which these genes behave similarly. For example, an appropriate biclustering method will recognize highlighted hidden bicluster from Table 1.

However, in practice, the boundary of a bicluster is usually fuzzy for three reasons: (i) the microarray dataset might be noisy and incomplete, (ii) the similarity measurement between genes is continuous and there is no clear cutoff value for group membership, and (iii) a gene might behave similarly to gene A under a set of conditions and behave similarly to another gene B under another set of conditions. Therefore, there is a great need for a fuzzy biclustering method, which produces biclusters in which genes and conditions can belong to a cluster partially and to multiple biclusters at the same time with different membership degrees.

The main contributions of this paper are:

1. We propose a new fuzzy bicluster model for value-coherent biclusters.
2. Based on this model, we define an objective function whose minimum will characterize good fuzzy value-coherent biclusters and facilitate the tradeoff between the number of genes, the number of conditions, and the quality of returned biclusters.
3. We propose a genetic algorithm based method, Genetic Fuzzy Biclustering Algorithm (GFBA), to identify fuzzy value-coherent biclusters. Our experiments show that GFBA is efficient in its converging to the global optimum and produces biclusters with higher quality than existing methods.

Organization. The rest of this paper is organized as follows. In section 2, We introduce the related work on biclustering. In section 3, we formally define a new fuzzy bicluster model and objective function. In section 4, we propose GFBA for discovering biclusters. Experimental results and comparison are analyzed in section 5 and section 6 concludes the paper.

2 Related Work

A recent survey by Madeira and Oliveira [2] identifies four major classes of biclusters: 1) biclusters with constant values, 2) biclusters with constant values on rows or columns, 3) biclusters with coherent values, and 4) biclusters with coherent evolutions.

The first class of biclusters are submatrixes where all values are equal. However, in real datasets, constant biclusters are usually masked by noise. Therefore, a merit function is needed to quantify the quality of constant biclusters. Hartigan [3] defines the variance of a bicluster as such a merit function and proposes a partition-based algorithm to discover constant biclsuters. Given a user specified parameter *K*, the algorithm partitions the original matrix into K biclusters and then calculates the variance for each bicluster. Based on the above variance, Tibshirani et al. [4] propose a permutation-based method to induce the optimal number of biclsuters, K. In addition to using the variance defined by Hartigan [3], Cho et al. [5] also use the total squared residue to quantify the homogeneity of a bicluster. Therefore, their framework can find not only constant biclusters, but also value-coherent biclusters (class 3).

The second class of biclusters are submatrixes with constant values on rows or columns. Getz et al. [6] propose to apply a normalization procedure first to the input matrix before their coupled two-way clustering method is performed. The normalization procedure transforms a row-constant or column-constant bicluster into a constant bicluster (on both rows and columns). To discover row-constant biclusters hidden in noisy data, Califano et al. [7] propose an approach to discover biclusters that for each row, the difference between two extreme values is within some user specified value δ. The same approach can be applied to the discovery of column-constant biclusters. Sheng et al. [8] propose a Bayesian based approach and use either the row-column orientation or column-row orientation to discover column-constant or row-constant biclusters. Finally, Segal et al. [9] introduce a probabilistic model based approach to discover column-constant biclusters, which are more general than the previous approaches.

The third class of biclusters are submatrixes with coherent values on both rows and columns. Cheng and Church [1] define the mean squared residue score to quantify the incoherence of a bicluster and propose a greedy algorithm to discover biclusters with scores lower than some threshold. Yang et al. [10] generalize the definition of δ-bicluster to deal with null values and use the FLexible Overlapped biClustering ($FLOC$) algorithm to discover a set of biclusters simultaneously. The Coupled Two-Way Clustering ($CTWC$) proposed by Getz el al. [6] and the Interrelated Two-Way Clustering ($ITWC$) proposed by Tang et al. [11] are two iterative algorithms based on a combination of the results of one-way clustering on both dimensions. Lazzeroni and Oven [12] introduce the plaid model where the value of an element in the data matrix is viewed as a sum of layers which take into account the interactions between biclusters. Bleuler et al. [13] propose a EA (evolutionary algorithm) framework that embeds a greedy strategy. Chakraborty et al. [14] use genetic algorithm to eliminate the threshold of the maximum allowable dissimilarity in a bicluster.

The fourth class of biclusters are submatrixes that have coherent evolutions across the rows and/or columns of the data matrix regardless of their exact values. Ben-Dor et al. [15] define a bicluster as an order-preserving submatrix (OPSM) in which the sequence of values in every row is strictly increasing. Liu and Wang [16] generalize the OPSM model by allowing grouping so that columns with insignificant value differences will be considered to have the same ranking and assigned to the same group.

The fuzzy sets theory, developed by Zadeh [17], allows an object to partially belong to one cluster and can belong to more than one clusters at the same time with different membership degrees. As opposed to traditional clustering techniques, fuzzy clustering does not require exclusive membership of data items to a particular class which is advantageous, in that it accommodates uncertainty and imprecision. The most famous fuzzy clustering technique is fuzzy C-means (FCM) algorithm [18] and many other analytic fuzzy clustering approaches are derived from it. A noise clustering (NC) algorithm has been proposed to overcome sensitivity of the FCM algorithms to noisy data [19]. R. Krishnapuram et al.[20] introduce an approach called Possibilistic C Means (PCM) algorithm to overcome the relative membership problem of the FCM.

3 Our Proposed Fuzzy Value-Coherent Bicluster Model and Its Objective Function

In this section, we first give an overview of the value coherent bicluster model proposed by Cheng and Church [1], and then based on this model, we develop our fuzzy value-coherent bicluster model. Through out the whole paper, we will use the notations summarized in Table 2. These notations will be defined later.

Table 2. Notations used in this paper.

a_{ij}	the entry in row i and column j
a_{iJ}	the mean of row i of a bicluster
a_{Ij}	the mean of column j of a bicluster
a_{IJ}	the mean of a whole bicluster A_{IJ}
R_{ij}	the residue of a_{ij} in a bicluster
Q_{iJ}	the row sum squared residue of row i
Q_{Ij}	the column sum squared residue of column j
$f_I(i)$	the fuzzy membership value for row i
$f_J(j)$	the fuzzy membership value for column j
$H(I, J)$	the (fuzzy) mean squared residue for bicluster A_{IJ}

3.1 The Value-Coherent Bicluster Model

Consider a gene expression dataset represented by a M by N matrix A, in which rows represent genes, columns represent conditions, and a_{ij} represents the expression of gene

i under condition j. Let $G = \{g_1, \cdots, g_M\}$ and $C = \{c_1, \cdots, c_n\}$ be the sets of genes and conditions in A, respectively. Let $I \subseteq G$ and $J \subseteq C$, we use A_{IJ} to denote the submatrix formed by extracting from A all rows and columns in I and J. A (crisp) *bicluster* is a submatrix of A.

Cheng and Church [1] define a value-coherent bicluster model based on an additive model in which each entry a_{ij} is obtained by the sum of the background effect μ, row i's effect α_i, and column j's effect β_j:

$$a_{ij} = \mu + \alpha_i + \beta_j \tag{1}$$

These effects are defined as $\mu = a_{IJ}$, $\alpha_i = a_{iJ} - a_{IJ}$, and $\beta_j = a_{Ij} - a_{IJ}$, where a_{iJ}, a_{Ij}, and a_{IJ} are the means of row i, column j, and the whole bicluster A_{IJ}, respectively, which are defined as follows: $a_{iJ} = \frac{\sum_{j \in J} a_{ij}}{|J|}$, $a_{Ij} = \frac{\sum_{i \in I} a_{ij}}{|I|}$, $a_{IJ} = \frac{\sum_{i \in I} \sum_{j \in J} a_{ij}}{|I| \cdot |J|}$. | I | and | J | denote the cardinalities of sets I and J, respectively. As a result, each a_{ij} satisfies the following equation.

$$a_{ij} = a_{iJ} + a_{Ij} - a_{IJ} \tag{2}$$

A bicluster A_{IJ} is *fully value-coherent* if all entries in A_{IJ} satisfy Equation (2).

However, given an arbitrary submatrix A_{IJ}, it might not be a fully value-coherent bicluster. To quantify the value coherence of a bicluster, the notion of *residue* is introduced to calculate the difference between the observed value of a_{ij} and the expected value of a_{ij} if A_{IJ} was a fully value-coherent bicluster.

$$R_{ij} = a_{ij} - a_{iJ} - a_{Ij} + a_{IJ} \tag{3}$$

Finally, the incoherence of the whole bicluster A_{ij} is defined as the mean squared residue of the bicluster as follows.

$$H(I, J) = \frac{\sum_{i \in I} \sum_{j \in J} R_{ij}^2}{|I| \cdot |J|} \tag{4}$$

Problem Statement. Under this model, a formal statement of *the value-coherent biclustering problem* is as follows: given a gene expression data matrix A and a user provided value δ, return all biclusters A_{IJ} with $H(I, J) \leq \delta$.

3.2 Our Proposed Fuzzy Value-Coherent Bicluster Model

In this section, we extend the above value-coherent bicluster model to a fuzzy model. In contrast to a crisp bicluster, which either contains a gene or a condition completely or does not contain it at all, a fuzzy bicluster can contain a gene or a condition *partially*.

More formally, given a gene expression dataset represented by a M by N matrix A with gene set G and condition set C. Let I be a fuzzy set defined over G with a fuzzy membership function f_I where $0 \leq f_I(i) \leq 1$ for $1 \leq i \leq M$. Similarly, let J be a fuzzy set defined over G with a fuzzy membership function f_J where $0 \leq f_J(j) \leq 1$ for $1 \leq j \leq N$. We use A_{IJ} to denote a fuzzy bicluster that is formed by associating each gene i with a membership value $f_I(i)$ and each condition j with a membership value $f_J(j)$ to reflect the degrees that they belong to the fuzzy bicluster. Therefore, a crisp bicluster a is special case of a fuzzy bicluster. The cardinalities of fuzzy sets I and J are represented as $|I|$ and $|J|$ such that $|I| = \sum_{i=1}^{M} f_I(i)$ and $|J| = \sum_{j=1}^{N} f_J(j)$.

Let $a_{iJ} = \frac{\sum_{j=1}^{N} f_J(j)^m \cdot a_{ij}}{\sum_{j=1}^{N} f_J(j)^m}$, $a_{Ij} = \frac{\sum_{i=1}^{M} f_I(i)^m \cdot a_{ij}}{\sum_{i=1}^{M} f_I(i)^m}$, and $a_{IJ} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} f_I(i)^m \cdot f_J(j)^m \cdot a_{ij}}{\sum_{i=1}^{M} \sum_{j=1}^{N} f_I(i)^m \cdot f_J(j)^m}$ be the means of row i, column j, and the whole bicluster A_{IJ}, respectively, m is a user defined value called *fuzziness parameter*, which is used to adjust the power of $f_I(i)$ or $f_J(j)$. The larger the value of m is, the greater the power of $f_I(i)$. We define the residue of an entry a_{ij} as:

$$R_{ij} = a_{ij} - a_{iJ} - a_{Ij} + a_{IJ} \tag{5}$$

We then define the incoherence of the whole fuzzy bicluster A_{IJ} as the mean squared residue of the bicluster.

$$H(I,J) = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} f_I(i)^m \cdot f_J(j)^m \cdot R_{ij}^2}{|I| \cdot |J|} \tag{6}$$

The reader can verify that if we restrict each fuzzy membership value to be 0 or 1 then Equation (6) is reduced to Equation (4). Therefore, our model is a fuzzy generalization of the basic value-coherent bicluster model.

In addition, we define the row sum squared residue $Q_{iJ} = \sum_{j=1}^{N} f_J(j)^m \cdot R_{ij}^2$ and the column sum squared residue $Q_{Ij} = \sum_{i=1}^{M} f_I(i)^m \cdot R_{ij}^2$ to indicate the contributions of gene i and condition j to the total incoherence, respectively:

3.3 Our Proposed Objective Function

Suppose a user is interested in returning the fuzzy bicluster A_{IJ} with $|I| = \xi_I$ and $|J| = \xi_J$ where ξ_I and ξ_J are two user provided fixed values such that $H(I,J)$ is minimized. The following theorem states one necessary condition for the minimization of $H(I,J)$. This condition can be used in an iterative procedure to update I and J's membership functions to achieve the minimization of $H(I,J)$.

Theorem 1. *Given fixed values $m \in (1,\infty)$, $\xi_I > 0$, and $\xi_J > 0$, $H(I,J)$ with constraints $|I| = \xi_I$ and $|J| = \xi_J$ is globally minimal only if :*

$$f_I(i) = \xi_I \cdot \frac{\frac{1}{Q_{iJ}^{1/(m-1)}}}{\sum_{i=1}^{I} \frac{1}{Q_{iJ}^{1/(m-1)}}} \tag{7}$$

$$f_J(j) = \xi_J \cdot \frac{\frac{1}{Q_{Ij}^{1/(m-1)}}}{\sum_{j=1}^{J} \frac{1}{Q_{Ij}^{1/(m-1)}}} \tag{8}$$

Proof. see [21].

The condition stated in the above theorem can be used to discover biclusters with a given size. To discover biclusters with arbitrary sizes, if one uses the mean squared residue as the objective function, our experiments show that most discovered biclusters will have small sizes (volumes), this is very undesirable since larger biclusters should be more interesting to a biologist as they tend to be more significant. This phenomenon can easily be explained by the definition of the mean squared residue – the larger a bicluster is, the closer its mean squared residue is to the mean squared residue of the whole matrix. Thus smaller biclusters tend to have extreme mean square residues, while larger biclusters tend to have mean square residues that are in the middle.

To solve this problem, inspired by PCM [20], we propose the following objective function.

$$\begin{aligned} H_m(I,J) &= H(I,J) \\ &+ \frac{\Sigma_{i=1}^{M} \eta \cdot H(I,J) \cdot (1 - f_I(i))^m}{M - |I|} \\ &+ \frac{\Sigma_{j=1}^{N} \xi \cdot H(I,J) \cdot (1 - f_J(j))^m}{N - |J|} \end{aligned} \tag{9}$$

where the first term is used to control the quality of a bicluster by minimizing its incoherence, the second and third terms are used to promote a bicluster with more genes and more conditions. η and ξ are parameters provided to satisfy different requirements on the incoherence and the sizes of the biclusters. If biologists need biclusters with more genes, they can use greater η value; if they need biclusters with more genes they can increase ξ values. On the other side if they require higher quality they can use smaller values. $H(I,J)$ is the mean squared residue score which is used to adjust the weight of compensation.

Theorem 2. *Given fixed value $m \in (1,\infty)$, $H_m(I,J)$ is globally minimal only if:*

$$f_I(i) = \frac{1}{1 + (\frac{Q_{iJ} \cdot (M-|I|)}{|I||J| \cdot \eta \cdot H(I,J)})^{1/(m-1)}} \tag{10}$$

$$f_J(j) = \frac{1}{1 + (\frac{Q_{Ij} \cdot (N-|J|)}{|I||J| \cdot \xi \cdot H(I,J)})^{1/(m-1)}} \tag{11}$$

Proof. see [21].

It is obvious from (10) and (11) that $f_I(i)$ and $f_J(j)$ lie in the desired range. The genes and conditions that produce large residues will be reassigned with low membership degrees while those co-expressed genes and conditions can get high membership degrees as we expect.

4 Our Proposed GFBA Algorithm

4.1 An Overview of GFBA

Equations (10) and (11) provide an efficient and stable optimization method to minimize the objective function in (9). Unfortunately, it is dependent on initial conditions and might end in a local optimum. In order to ensure the algorithm to converge to a global optimal solution, we propose a genetic algorithm (GA) based fuzzy biclustering algorithm, called GFBA whose pseudocode is outlined in Figure 1, in which g is the number of generations, p is the number of biclusters in each population, mp is the probability of mutation, r is the fraction of the population to be replaced by crossover in each population, cp is the fraction of the population to be replaced by crossover in each population, z is the number of biclusters in each population, and m is the fuzziness parameter. It is different from the normal genetic algorithm in that we have an additional step called *BiclusterOptimization* to speed up the convergence process. Psudocode of BiclusterOptimization is shown in Figure 2.

```
Algorithm GFBA
Input: $g$,$p$,$mp$,$r$,$cp$,$z$,$m$;
Output: Biclusters
Begin
    Initialization()
    For $i$ =0 to $Z - 1$
        Selection()
        Crossover()
        Mutation()
        BiclusterOptimization()
    End For
End Algorithm
```

Fig. 1. Pseudocode of GFBA.

```
Algorithm BiclusterOptimization
Input: $< I, J >$;
Output: $< I', J' >$;
Begin
    $I^1 = I, J^1 = J$
    While $||I^{k+1} - I^k|| + ||J^{k+1} - J^k|| < \varepsilon$
        Calculate $H(I, J)$ using (6)
        Update $I^{k+1}$ using (10)
        Update $J^{k+1}$ using (11)
    End While
$I' = I^{k+1}, J' = J^{k+1}$
End Algorithm
```

Fig. 2. Pseudocode of BiclusterOptimization.

4.2 Solution Encoding and Fitness Function Definition

As in GA, GFBA maintains a population of coded solutions. A natural way of coding a bicluster is to consider two chromosomes of length M and N representing the genes and conditions. In this case, each allele corresponds to the fuzzy membership degree of a gene or condition. Thus in GFBA, each solution is encoded with a pair of vectors $S_z = (I, J)$ where $I = \{I_1, \cdots, I_M\}(0 < I_i < 1)$ and $J = \{J_1, \cdots, J_N\}(0 < J_j < 1)$. Then we define the fitness function as:$F(S_z) = \frac{1}{H_m(S_z)}$ in which$H_m(S_z)$ can be calculated by (9). We choose the inverse of $H_m(S_z)$ as the fitness value because our goal is to minimize the objective function shown in Equation (9) and thus those biclusters with lower values will be given higher probabilities to survive.

4.3 Operators

Initialization: The genetic algorithm begins with an initial population. In our experiment, the initial population is randomly generated, that is, all the membership degrees are initialized with random numbers between 0 and 1.

Selection: We use Roulette Wheel Selection (RWS), the most commonly used form of GA selection, for the selection operator. When using RWS, a certain number of biclusters $(1 - cp) \cdot z$ of the next generation are selected probabilistically, where the probability of selecting a bicluster S_z is given by

$$Pr(S_z) = \frac{Fitness(S_z)}{\Sigma_{z=1}^{Z} Fitness(S_z)} \tag{12}$$

With RWS, each solution will have a probability to survive by being assigned with a positive fitness value. A solution with a smaller H_m has a greater fitness value and hence has a higher probability to survive. On the other side, weaker solutions also have a chance to survive the selection process. This is an advantage, as though a solution may be weak, it may still contain some useful components.

Crossover and Mutation: Then $cp{\cdot}z/2$ pairs of parents are chosen probabilistically from the current population and the crossover operator will produce two new offsprings for each pair of parents using one point crossover technique on genes and conditions separately. Now the new generation contains the desired number of members and the mutation will increase or decrease the membership degree of each gene and conditions with a small probability of mutation mp.

Bicluster Optimization: Although the ordinary GA algorithm with above operators may converge and discover biclusters, it will take a lot of time since the initial assignments are random and subsequent evolution process are blind and probabilistic. To solve this problem, We try to leverage the advantage of efficiency and robustness of FCM and proposed BiclusterOptimization function Fig.2. BiclusterOptimization function is a simply Picard iteration through necessary condition $||I^{k+1} - I^k|| + ||J^{k+1} - J^k|| < \varepsilon$. In each generation we use it to update the membership degrees of the genes and conditions resulting from a new generation.

4.4 Interpretation of Fuzzy Biclustering Results

In contrast to traditional biclustering algorithms, our GFBA produces a set of fuzzy biclusters, each of which includes a subset of genes and conditions along with their membership values. For example, given the matrix in Table 3, our algorithm will return the following fuzzy bicluster that is presented by two vectors: Gene(0.99, 0.99, 0.99, 0.99, 0.13) and Con(0.99, 0,99, 0.99, 0.99, 0.81). Each vector indicates the degree of possibility for each gene/condition belonging to the bicluster.

Table 3. A sample gene expression data matrix.

	cond.1	cond.2	cond.3	cond.4	cond.5
gene1	2.0	3.0	6.0	1.0	10.0
gene2	3.0	4.0	7.0	2.0	11.0
gene3	5.0	6.0	9.0	4.0	12.0
gene4	6.0	7.0	10.0	5.0	13.0
gene5	40.0	40.0	40.0	40.0	40.0

Fuzzy biclusters provide richer information than regular biclusters as fuzzy biclusters associate with each gene/condition a membership value. However, humans are more familiar with the analysis and reasoning of regular biclusters that are not fuzzy, to accommodate this, our GFBA algorithm provides an optional step to select the most powerful genes and conditions from a fuzzy bicluster. More specifically, given a fuzzy bicluster A_{IJ} and a user specified threshold δ, we can interpret A_{IJ} as a regular bicluster $A_{I'J'}$ as follows.

$$I' = \{i | f_I(i) >= \alpha\} \tag{13}$$

$$J' = \{j | f_J(j) >= \alpha\} \tag{14}$$

Following the above example, if we choose α = 0.8, then the interpretation of the resulting fuzzy bicluster is a regular bicluster that contains genes 1-4 and conditions 1-5 since the membership value of gene 5 is smaller than 0.8 but the membership value for condition 5 is more than 0.8.

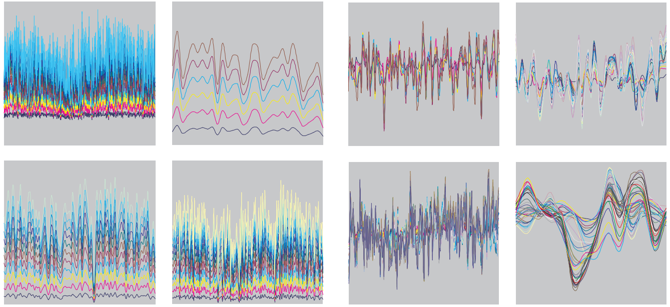
5 Experimental Results

We conducted experiments using GFBA on the same gene expression data sets as used by Y. Cheng and G.M. Church [1]. The yeast Saccharomyces cerevisiae cell cycle expression dataset contains 2,884 genes and 17 conditions. These genes were selected according to Tavazoie et al. (1999). The gene expression values were transformed by scaling and logarithm $x- > 100log(10^5x)$. So the values were mapped into the range 0 and 600 and missing values were represented by -1 in the yeast dataset.

The human lymphoma dataset contains 4026 genes and 96 conditions. The human data was downloaded from the website for supplementary information for the article by Alizadeh et al. (2000). The expression levels were reported as log ratios and after a scaling by a factor of 100, the data values are in the range between -750 and 650, with 47,639 missing values. The matrices introduced above were obtained from `http://arep.med.harvard.edu/biclustering`.

Fuzziness parameter m is important which determines the degree of fuzziness. In general, the larger m is, the “fuzzier” are the membership assignments. Doulaye Dembele. and Philippe Kastner [22] applied traditional FCM on microarray data using yeast and human dataset and found 2 is not a good choice for microarray data. In our experiment we found 1.6 is an appropriate fuzziness parameter for fuzzy biclustering on microarray data. Parameters η and ξ in Eq.22 do not necessary needed as input since we can randomly choose different values for different solutions thus we can find biclusters with different sizes. Parameter ε is used in the procedure BiclusterOptimization as a termination criterion between 0 and 1. In our experiment we choose 0.2.

There are also some parameters used for genetic framework: cp the fraction of the population to be replaced by crossover in each population is 0.7; z the number of biclusters in each population is 300; mp probability of mutation is 0.01; g the number of generations is optional, the higher the better. In our experiment we choose 40. Figure 3 shows Four sample biclusters discovered from yeast expression and human lymphoma expression dataset. The numbers of genes and conditions in each are reported in the format of (residue value, number of genes, number of conditions) as follows: for yeast expression dataset (194.5, 687, 10), (81.9, 143, 6), (95.9, 105, 11), (135.8, 232, 12); for



(a) Yeast. (b) Human lymphoma.

Fig. 3. Sample biclusters discovered by our GFBA algorithm.

human lymphoma expression dataset (423.4, 120, 6), (539.9, 51, 16), (778.4, 240, 25), (541.8, 14, 39).

Our fuzzy bicluter model is a natural extension of δ-Bicluster [1] and our objective function is a complementarity of fuzzy mean squared residue. It is difficult to compare the quality of different algorithms. Cheng and Church's algorithm is very efficient in finding biclusters with small residues while our algorithm show advantages in fuzziness and flexible and are more capable in discovering large biclusters with relatively small residue. Our objective function can well quantify the quality of biclusters but we can not use it to compare with other algorithm since we are the only ones to use it. Here we use a bicluster quality measure $\frac{1}{n} \cdot \sum_{i=1}^{n} \frac{Residue_i}{Volume_i}$ proposed by A.Chakraborty and H. Maka [14] by calculating the average residue/volume ratio of biclusters. Here n is the number of biclusters returned by an algorithm. Although this metric is better than the mean squared residue, for the control of quality, it is not desirable. This is because a small change of the threshold might affect the number of returned biclusters dramatically, either too many or too few will be returned. However it can be a fair measurement for comparison. On Yeast Dataset, the mean bicluster quality value of our algorithm is 0.02878 while Cheng and Church's algorithm is 1.39978 and Chakraborty's Genetic algorithm 0.05132. On Lymphoma Dataset, our algorithm is 0.109 while Cheng and Church's algorithm is 0.8156 and Chakraborty's Genetic algorithm 0.1247. More detailed results are available at [21].

6 Conclusions and Future Work

In this paper, we proposed an innovative fuzzy bicluster model, in which biclusters can be represented by the degree of possibilities that the genes and conditions belong to these biclusters. Based on this model, we defined an objective function whose minimum will characterize good fuzzy value-coherent biclusters and proposed a genetic algorithm based optimization method. Our experiments showed that our method is very efficient in converging to the global optimum. Future work includes the investigation of methods for discovering other classes of fuzzy biclusters and the applications of these methods to gene expression data analysis.

References

1. Cheng, Y., Church, G.M.: Biclustering of expression data. In: the 8th International Conference on Intelligent Systems for Molecular Biology. (2000) 93–103
2. Madeira, S.C., Oliveira, A.L.: Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Transactions on Computational Biology and Bioinformatics **1** (2004) 24–45
3. Hartigan, J.: Direct clustering of a data matrix. Journal of American Statistical Association **67**(337) (1972) 123–129
4. Tibshirani, R., Hastie, T., Eisen, M., Ross, D., Botstein, D., Brown, P.: Clustering methods for the analysis of DNA microarray data. Technical report, Dept. of Health Research and Policy, Dept. of Genetics, and Dept. of Biochemistry, Stanford Univ. (1999)
5. Cho, H., Dhillon, I., Guan, Y., Sra, S.: Minimum sum-squared residue coclustering of gene expression data. In: Fourth SIAM Intl Conf. Data Mining. (2004)
6. Getz, G., Levine, E., Domany, E.: Coupled two-way clustering analysis of gene microarray data. In: the Natural Academy of Sciences USA. (2000) 12079–12084
7. Califano, A., Stolovitzky, G., Tu, Y.: Analysis of gene expression microarays for phenotype classification. In: Intl Conf. Computacional Molecular Biology. (2000) 75–85
8. Sheng, Q., Moreau, Y., Moor, B.D.: Biclustering microarray data by gibbs sampling. Bioinformatics **19** (2003) ii196–ii205
9. Segal, E., Taskar, B., Gasch, A., Friedman, N., Koller, D.: Rich probabilistic models for gene expression. Bioinformatics **17** (2001) S243–S252
10. Yang, J., Wang, W., Wang, H., Yu, P.: Enhanced biclustering on expression data. In: 3rd IEEE Conference on Bioinformatics and Bioengineering. (2003) 321–327
11. Tang, C., Zhang, L., Ramanathan, M.: Interrelated two way clustering: an unsupervised approach for gene expression data analysis. In: Proc. of the 2nd IEEE International Symposium on Bioinformatics and Bioengineering. (2001) 41–48
12. Lazzeroni, L., Owen, A.: Plaid models for gene expression data. Technical report, Stanford Univ. (2000)
13. Bleuler, S., Prelic, A., Zitzler, E.: An EA framework for biclustering of gene expression data. In: Congress on Evolutionary Computation CEC2004. Volume 1. (2004) 166–173
14. Chakraborty, A., Maka, H.: Biclustering of gene expression data using genetic algorithm. In: the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology. (2005) 1–8
15. Ben-Dor, A., Chor, B., Karp, R., Yakhini, Z.: Discovering local structure in gene expression data: The order- preserving submatrix problem. In: Proc. of the Sixth Int Conf. Computational Biology. (2002) 49–57
16. Liu, J., Wang, W.: OP-Cluster: Clustering by tendency in high dimensional space. In: Third IEEE Intl Conf. Data Mining. (2003) 187–194
17. Zadeh, L.: Fuzzy sets. Information and Control **8** (1965) 338–353
18. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, New York (1981)
19. Dave, R.N.: Characterization and detection of noise in clustering. Pattern Recognition Letters **12**(11) (1991) 657–664
20. Krishnapuram, R., Keller, J.M.: A possibilistic approach to clustering. IEEE Transactions on Fuzzy Systems **1**(2) (1993) 98–110
21. Fei, X., Lu, S., Pop, H.F., Liang, L.: GFBA: A genetic fuzzy biclustering algorithm for discovering value-coherent biclusters, TR-DB-102006-FLPL", institution =. Technical report
22. Dembele, D., Kastner, P.: Fuzzy c-means method for clustering microarray data. **19**(8) (2003) 973–980