Evolutionary Metaheuristic for Biclustering based on Linear Correlations among Genes.

Juan A. Nepomuceno
Department of Computer Science
University of Sevilla, Spain
janepo@us.es

Alicia Troncoso Jesús S. Aguilar-Ruiz
Area of Computer Science
Pablo de Olavide University of Sevilla, Spain
{ali,aguilar}@upo.es

ABSTRACT

A new measure to evaluate the quality of a bicluster is proposed in this paper. This measure is based on correlations among genes. Moreover, a new evolutionary metaheuristic based on Scatter Search, which uses this measure as the fitness function, is presented to obtain biclusters that contain groups de highly-correlated genes. Later, an analysis of the correlation matrix of these biclusters is made to select these groups of genes that define new biclusters with shifting and scaling patterns. Experimental results from human B-cell lymphoma are presented.

Keywords

Biclustering, Gene Expression Data, Evolutionary Computation

1. INTRODUCTION

The human genome sequence was completely decoded with the *Human Genome Project* at the beginning of the present century. Every cell belonging to our body contains the same information but not all genes are expressed for every cell. This codification process of genes is known as *gene expression* and it is one of the most important research topics in Biology. DNA microarrays technology enable us to measure the gene expression level of thousand of genes under multiple sets of experimental conditions [3]. Data mining techniques are needed to analyze the huge volume of all this biological information [12]. Traditional clustering methods are not good enough in this field because they study similarities among genes along the complete set of conditions of the microarray dataset. The goal of biclustering techniques is to identify genes with the same pattern only under a specific group of conditions.

Several surveys about biclustering techniques have been published as for example [13] or more recently [5]. Many biclustering algorithms are optimization techniques which combine a search procedure with a merit function to evaluate biclusters. Evolutionary computation techniques are used in [8, 15] or Simulated Annealing in [4]. The fitness function used in these cases has been the mean squared residue (MSR) which was defined in [7] and measures the degree of coherence of biclusters. However, some kind of very important patterns from a biological point of view can not be obtained using the MSR. It is proven in [1] that MSR is not an appropriate measure to find scaling patterns when the variance of gene values is high in the bicluster. Thus, new measures have been proposed as fitness function in several optimization methods [16, 17].

Recently, the study of the nature of different patterns in data has motivated the development of new methods that work from a geometrical point of view [10, 11, 19, 9]. Most of previous approximations can only detect biclusters with a limited set of patterns. However, many of the main patterns to infer gene regulatory mechanism share the geometry of linear manifolds. Therefore, biclustering can be seen as the problem of searching for hyperplanes in high–dimensional data space.

The study of linear manifolds composed of genes has been the main motivation in this work. Thus, a new measure based on the linear dependence among genes is proposed to evaluate the quality of biclusters. An evolutionary metaheuristic, with this new measure as fitness function, is used to find biclusters. This optimization procedure is based on a Scatter Search scheme and uses several typical aspects of Genetic Algorithms as the way of generating the offspring. Biclusters obtained by the proposed method enclose groups of highly–correlated genes that are discovered by analyzing the correlation matrix. These groups of genes provide new biclusters with shifting and scaling patterns.

This paper is organized as follows. In section 2, it is shown basic concepts about linear correlations and how linear dependence among genes can be evaluated. Section 3 presents the proposed measure. Section 4 presents the methodology in order to find high–quality biclusters. Some experimental results from a real dataset are reported in Section 5. Finally, Section 6 outlines the main conclusions of the paper and future works.

2. CORRELATIONS AMONG GENES

Two discrete random variables X and Y, with values $\{x_1, ..., x_n\}$ and $\{y_1, ..., y_n\}$ respectively, are linearly dependent if one of them can be written as a linear combination of the other one. That is,

$$y_i = \alpha x_i + \beta \text{ with } i = 1, ..., n \text{ and } \alpha, \beta \in \mathbb{R} \qquad (1)$$

The linear correlation coefficient [18] is defined by:

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{n\sigma_X\sigma_Y} \qquad (2)$$

where $cov(X,Y)$ is the covariance of the variables X and Y, $\overline{x}$ and $\overline{y}$ are the mean of the values of the variables X and Y and σ_X and σ_Y are the standard deviations of X and Y, respectively.

SAC'10 March 22-26, 2010, Sierre, Switzerland.

The correlation coefficient, $\rho(X, Y)$, measures the grade of linear dependence between the variables X and Y. The values for the correlation coefficient vary between -1 and 1. If $\rho(X, Y) = 0$, the variables X and Y are linearly independent, and if $\rho(X, Y) = \pm 1$ the variables are linearly dependent. When the correlation value is equal to -1, the variables X and Y are dependent with negative correlation, that is, when the values of the variable X increase the values of the variable Y decrease linearly.

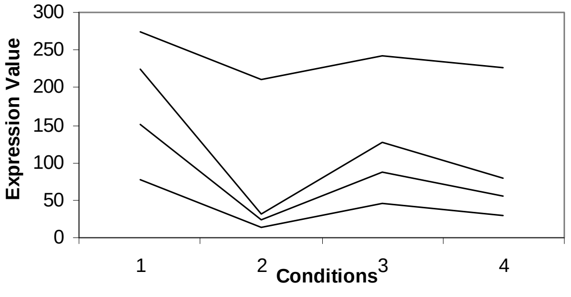
After preprocessing and normalization steps, a microarray can be seen as a real matrix M composed of N genes and L conditions. The element (i, j) of the matrix indicates the level of expression of gene i under the condition j. A bicluster B is a submatrix of the matrix M composed of $n \leq N$ rows or genes and $l \leq L$ columns or conditions. Submatrices associated to genes with shifting and scaling patterns are preferred. A discussion about what kind of submatrices are associated to different patterns can be read in [13].

A bicluster can be seen as a group of variables (or genes) whose values are the expression values of genes under the specific conditions. Biclusters with genes expressed in the same way, that is, genes with the same behavior, are the most interesting from a biological point of view. A group of genes has a *shifting pattern* when the expression values vary in the addition of a fixed value for all the genes. Sometimes, a group of genes presents the same behavior regarding the regulation but not with the same intensity. A group of genes has a *scaling pattern* when the expression values vary in the multiplication of a fixed value for all the genes. The behavior of genes is the same as they only vary on the slope when the expression value of the gene for each condition increases or decreases. Thus, two genes show a shifting or scaling pattern or both at the same time when both follow the equations (3) or (4) or (5) respectively:

$$g_Y = g_X + \beta \tag{3}$$
$$g_Y = \alpha g_X \tag{4}$$
$$g_Y = \alpha g_X + \beta \tag{5}$$

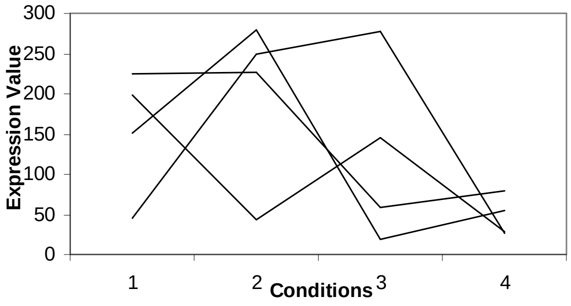
where α and β are real numbers.

Consequently, two genes with shifting or/and scaling patterns are linearly dependent. Thus, the correlation coefficient is set to 1 or -1 if such linear relationship is positive or negative, respectively. Therefore, a good measure, for finding theses patterns in data, will be a measure which detects linear dependence among genes from a microarray. A measure of quality of biclusters based on the average of the linear correlations of each pair of genes is proposed in this work.



$$\begin{bmatrix}151&23&87&55\\77&13&45&29\\224&32&128&80\\275&211&243&227\end{bmatrix}\rightarrow\frac{1}{6}\sum_{i=1}^{4}\sum_{j=i+1}^{4}|\rho_{g_ig_j}|=1$$

Figure 1: Bicluster with highly–correlated genes



$$\begin{bmatrix}151&279&18&55\\199&43&146&29\\224&227&59&80\\45&250&278&27\end{bmatrix}\rightarrow\frac{1}{6}\sum_{i=1}^{4}\sum_{j=i+1}^{4}|\rho_{g_ig_j}|=0.26$$

Figure 2: Bicluster with lowly–correlated genes.

In Figure 1 a bicluster with 4 genes and 4 conditions, which contains a scaling pattern, is illustrated. All genes have the same tendency but with different intensity. Every row of the matrix represents a gene and the average correlation is equal to 1 which means that the bicluster is composed of highly–correlated genes. Thus, two genes represent two variables linearly dependent and the correlation coefficient for each pair of genes is equal to 1.

Figure 2 presents a bicluster without patterns, that is, every gen has a different behavior from the remaining ones in the bicluster. In this case, the average correlation among genes is 0.26 and, therefore, this bicluster is composed of lowly–correlated genes.

3. BICLUSTERS EVALUATION

Given a bicluster B composed of N genes, $B = [g_1, \ldots, g_N]$, the average correlation of B, $\rho(B)$, is defined as follows,

$$\rho(B) = \frac{1}{\binom{N}{2}}\sum_{i=1}^{N}\sum_{j=i+1}^{N}|\rho_{g_ig_j}| \tag{6}$$

where, $\rho_{g_ig_j}$ is the correlation coefficient between the gene i and the gene j.

Firstly, it must be noticed that $\rho_{g_ig_j} = \rho_{g_jg_i}$, therefore only $\binom{N}{2} = \frac{N(N-1)}{2}$ elements have been considered in the sum because each pair of genes must be taken into account only one time. Secondly, $0 \leq \rho^2 \leq 1$ and thus $0 \leq \rho(B) \leq 1$ too.

Generally, biclusters with highly–correlated genes and high volume are preferred. Therefore, the fitness function is defined by:

$$f(B) = (1 - \rho(B)) + M_1\left(\frac{1}{nG}\right) + M_2\left(\frac{1}{nC}\right) \tag{7}$$

where nG and nC are the number of genes and conditions of the bicluster B, respectively, and M_1 and M_2 are penalization factors to control the volume of the bicluster B. Best biclusters are the ones with the lowest value for the fitness function. Thus, it has been considered $(1-\rho(B))$ to evaluate biclusters with highly–correlated genes as good biclusters.

Basically, the average of the linear correlations of each pair of genes has been considered as the measure of quality of biclusters and, therefore, with this fitness function the high–quality biclusters will be groups of correlated genes.

4. DESCRIPTION OF THE ALGORITHM

The proposed algorithm is an evolutionary metaheuristic for biclustering that uses as fitness function the one defined from (7). This search procedure finds biclusters composed of a huge number of genes with high average correlations.

The evolutionary metaheuristic is based on a general scheme of Scatter Search and some features of Genetic Algorithms such as the way of generating offspring. Scatter Search [14] is an optimization metaheuristic based on the evolution of a population. On the opposite to other evolutionary heuristics, Scatter Search emphasizes systematic processes against random procedures. The main ideas are the diversification of solutions in order to avoid local minima and the intensification in order to find high-quality solutions. A reference set of solutions is built to introduce this two ideas in the search. Reference set is composed of the best solutions according to the fitness function and the most scattered ones according to a certain distance. The evolution of this set along the iterative process provides the optimum solution.

The combination method is the mechanism to create new biclusters in Scatter Search. All pairs of biclusters belonging to the reference set are combined generating new biclusters. In this work the well-known uniform crossover operator used in Genetic Algorithms is the proposed combination method. The reference set is updated with the best biclusters, according to the fitness function, from the joining of the reference set and the new biclusters generated by the combination method. This process is repeated iteratively until the reference set does not change. After getting the stability of reference set in the updating process, this set is rebuilt to introduce diversity in the search process. The pseudocode of the proposed evolutionary method is presented in Algorithm 1.

Biclusters are encoded by binary strings of length $N + L$ where N and L are the number of genes and conditions from microarray M, respectively [8].

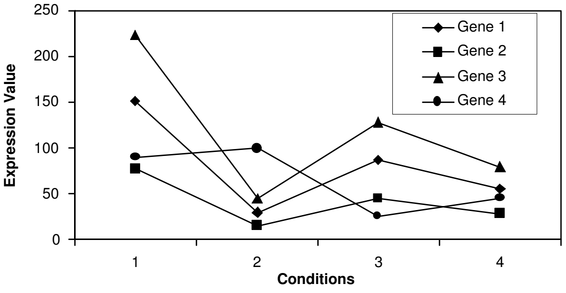


Figure 3: Bicluster with high correlation.

The fitness function is based on the average of the correlations of pair of genes. Due to the nature of the average, the value of the fitness function can be high for a bicluster and this bicluster can contain several non–correlated genes with the remaining ones of the bicluster as it is shown in Figure 3. Figure 3 presents three genes with a high correlation and one gene which is not correlated with the other three. Due to this fact, new biclusters have to be built by analyzing the correlation matrix of biclusters obtained by the Algorithm 1. Therefore, all groups of genes with a correlation coefficient greater or equal than a given minimum correlation are selected and these groups define new biclusters that present interesting patterns from a biological point of view as shifting and scaling patterns.

Algorithm 1 EVOLUTIONARY SEARCH PROCEDURE

```
INPUT Microarray $M$, penalization factors $M_1$ and $M_2$, number
  of biclusters $numBi$ to be found, maximum number of iterations
  $numIter$, size of the population and size $S$ of the reference set.
OUTPUT The set $Results$ containing $numBi$ biclusters.
begin
  $num \leftarrow 0$, $Results \leftarrow \emptyset$
  while ($num < numBi$) do
    Initialize population $P$
    //Building Reference Set
    $R_1 \leftarrow S/2$ best biclusters from $P$ (according to the fitness function)
    $R_2 \leftarrow S/2$ most scattered biclusters, regarding $R_1$, from $P \setminus R_1$ (according to a distance).
    $RefSet \leftarrow (R_1 \cup R_2)$
    $P \leftarrow P \setminus RefSet$
    //Initialization
    stable $\leftarrow$ FALSE, $i \leftarrow 0$
    while ($i < numIter$) do
      while (NOT stable) do
        $A \leftarrow RefSet$
        $B \leftarrow CombinationMethod(RefSet)$
        $RefSet \leftarrow S$ best biclusters from $RefSet \cup B$
        if ($A = RefSet$) then
          $stable \leftarrow TRUE$
        end if
      end while
      //Rebuilding Reference Set
      $R_1 \leftarrow S/2$ best biclusters from $RefSet$
      $R_2 \leftarrow S/2$ most scattered biclusters from $P \setminus R_1$
      $RefSet \leftarrow (R_1 \cup R_2)$
      $P \leftarrow P \setminus RefSet$
      $i \leftarrow i + 1$
    end while
    //Storage in Results
    $Results \leftarrow$ the best from $RefSet$
    $num \leftarrow num + 1$
  end while
end
```

5. EXPERIMENTAL RESULTS

A well known dataset [7] has been used to show the performance of the proposed method, the *human B-cells lymphoma* expression data, which has 4026 genes and 96 conditions [2]. Original data were processed replacing missing values with random values. The main parameters of the proposed evolutionary method are as follows: 200 for the size of the initial population, 10 for the reference set, 20 for the maximum number of iterations and 10 for the number of biclusters to be found. After an experimental study to test the influence of the penalization factors on the number of genes and conditions of biclusters to be obtained, the penalization factors $M_1 = 10$ and $M_2 = 10$ have been chosen.

Table 1 presents information about ten biclusters found by the proposed evolutionary search procedure from the lymphoma dataset. For each bicluster an identifier of the bicluster, the number of genes and conditions, the value of the MSR and the value of the average correlation among the genes are presented. It can be observed that the average correlation of these biclusters can be considered high taking into account their remarkable volume. Most of papers published in the literature present algorithms based on the MSR measure and a bicluster is considered a high–quality bicluster for the lymphoma dataset if the value of its MSR is less than 1200 [7, 8]. Thus, it could be concluded that the biclusters shown in this table can not be considered good biclusters since the values of the MSR exceed by far 1200. However, these biclusters contain groups de highly–correlated genes that define new biclusters with shifting and scaling patterns at the same time.

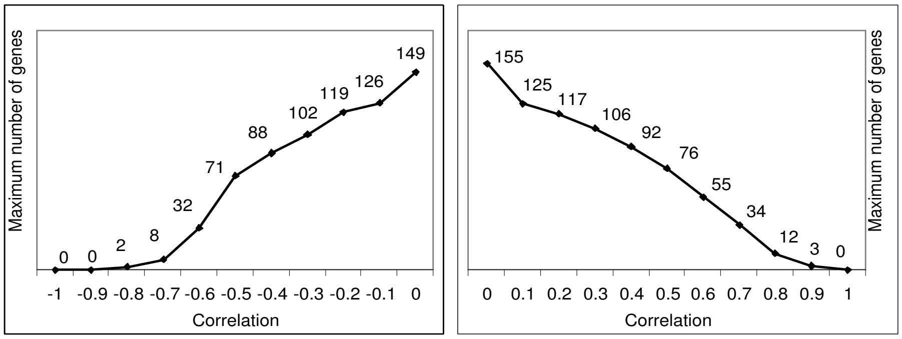


Figure 4: Volume of biclusters for different average correlations.

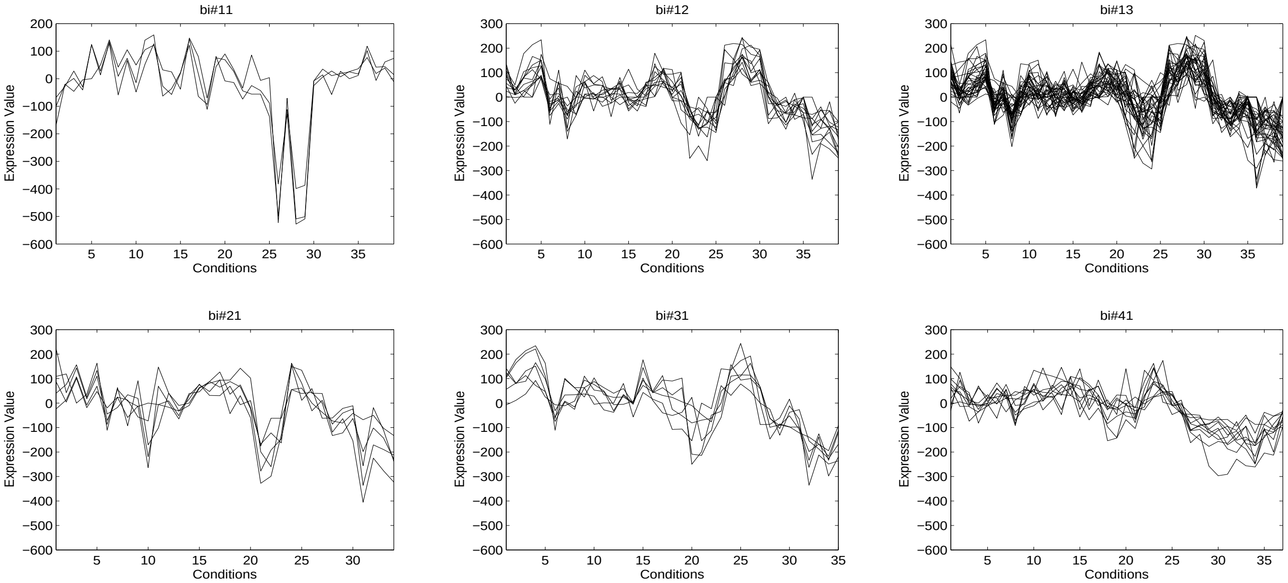


Figure 6: Biclusters found for the lymphoma dataset.

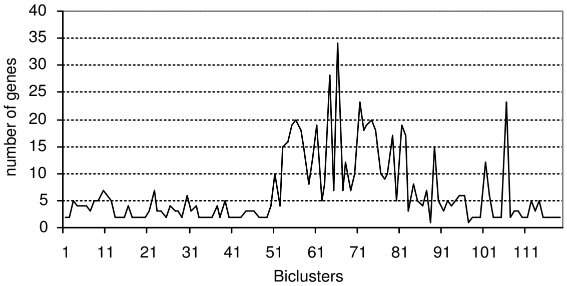


Figure 5: Number of genes of biclusters with average correlation of 0.7.

Id bi.	Genes	Conditions	MSR	Correlation
bi #1	235	39	7243.0	0.31
bi #2	249	36	6018.5	0.34
bi #3	171	34	5363.9	0.36
bi #4	193	35	7177.5	0.32
bi #5	279	37	5882.8	0.30
bi #6	302	43	6612.4	0.23
bi #7	204	32	6408.4	0.30
bi #8	249	37	7083.7	0.26
bi #9	308	37	5587.7	0.22
bi #10	288	32	6290.2	0.25

Table 1: Biclusters found by Algorithm 1 for the Lymphoma dataset.

Figure 4 shows the highest volumes of the biclusters, contained in the bicluster $bi\#1$, found for different values of correlation coefficients. For example, if the correlation coefficient is equal to 0.7, a bicluster composed of 34 genes can be obtained from the bicluster $bi\#1$ and the remaining biclusters found with that coefficient have a number of genes lower than 34. It can be noted that the higher value for correlation, the smaller the volume of biclusters is. As it is expected, patterns can be identified easier in biclusters with correlations near to 1.

Figure 5 shows the number of genes of biclusters, with an average correlation of 0.7, which have been found by analyzing the correlation matrix of the bicluster bi#1. It can be observed that a bicluster with 34 genes is obtained as it was expected. There are

Id bi.	Genes	MSR	Correlation	Variance
bi #11	3	1122.5	0.9	21371.9
bi #12	12	2136.5	0.8	8697.5
bi #13	34	3007.9	0.7	8349.6
bi #21	5	2712.1	0.8	13091.3
bi #31	6	2368.7	0.8	12503.8
bi #41	10	1811.65	0.8	7202.1

Table 2: Biclusters obtained from biclusters of Algorithm 1.

some biclusters with only two correlated genes but biclusters with more of 10 genes are found too.

Figure 6 shows six biclusters found by the proposed methodology for the lymphoma dataset. The biclusters $bi\#11$, $bi\#12$ and $bi\#13$ are biclusters with 3, 12 and 34 genes and with average correlations of 0.9, 0.8 and 0.7, respectively (see Figure 4). These biclusters have been obtained from the bicluster $bi\#1$. The biclusters $bi\#21$, $bi\#31$ and $bi\#41$ have been obtained by analyzing the groups of genes with average correlation of 0.8 from $bi\#2$, $bi\#3$ and $bi\#4$, respectively. It can be appreciated that the genes of the biclusters $bi\#21$, $bi\#31$ and $bi\#41$ show similar trends (shifting and scaling) along the different experimental conditions.

The information about these six biclusters is reported in Table 2. The number of genes, the MSR, the average correlation among genes and the variance of gene values for each bicluster are presented. The number of conditions is the same than that of the original biclusters (Table 1). These biclusters, except $bi\#11$, have values of MSR greater than 1200, however they can be considered good biclusters because they present shifting and scaling patterns as can be seen in Figure 6. It can be observed that the variance of gene values is high. This is a clear example of biclusters that can not be found by algorithms based on the MSR measure due to these algorithms might not find scaling patterns when the variance of gene values is high [1].

6. CONCLUSIONS

An evolutionary method to obtain biclusters has been presented in this paper. This algorithm is based on Scatter Search and uses a new measure based on correlations among genes. First, the algorithm searches for biclusters with groups of highly–correlated genes. Later, new biclusters with shifting and scaling patterns are built by analyzing the correlation matrix. Experiments and discussions from human B-cell lymphoma dataset have been reported. Interesting results, which indicate that the proposed measure improves the well-known MSR, have been provided.

Future works will be focussed on the comparison with other biclustering techniques using Gene Ontology Database [6, 9]. The standard deviation will be included in the fitness function to avoid the situation shown in Figure 3. Other aspects will be analyzed, as for example whether or not negative correlations must to be considered as quality patterns in data and the overlapping among genes.

7. ACKNOWLEDGMENTS

The financial support given by the Spanish Ministry of Science and Technology, project TIN2007-68084-C02 and by the Junta de Andalucía, project P07-TIC-02611 is acknowledged.

8. REFERENCES

[1] J. Aguilar-Ruiz. Shifting and scaling patterns from gene expression data. *Bioinformatics*, 21(20):3840–3845, 2005.

[2] A. Alizadeh and et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

[3] P. Brown and D. Botstein. Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, 21:33–37, 1999.

[4] K. Bryan. Biclustering of Expression Data Using Simulated Annealing. In *Proceedings of the 18th IEEE International Symposium on Computer-Based Medical Systems*, pages 383–388, 2005.

[5] S. Busygin, O. Prokopyev, and P. Pardalos. Biclustering in data mining. *Computers and Operations Research*, 35(9):2964–2987, 2008.

[6] C. Cano, L. Adarve, J. López, and A. Blanco. Possibilistic approach for biclustering microarray data. *Computers in Biology and Medicine*, 37(10):1426–1436, 2007.

[7] Y. Cheng and G. Church. Biclustering of Expression Data. *In Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 93–103, 2000.

[8] F. Divina and J. Aguilar-Ruiz. Biclustering of Expression Data with Evolutionary Computation. *IEEE Transactions on Knowledge and Data Engineering.*, 18(5):590–602, 2006.

[9] X. Gan, A. Liew, and H. Yan. Discovering biclusters in gene expression data based on high-dimensional linear geometries. *BMC Bioinformatics*, 9:209, 2008.

[10] R. Harpaz and R. Haralick. Exploiting the Geometry of Gene Expression Patterns for Unsupervised Learning. In *Proceedings of the 18th International Conference on Pattern Recognition (ICPR 2006)*, pages 670–674, 2006.

[11] R. Harpaz and R. Haralick. Mining Subspace Correlations. In *In Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining*, pages 335–342, 2007.

[12] P. Larranaga and et al. Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1):86–112, 2006.

[13] S. Madeira and A. Oliveira. Biclustering Algorithms for Biological Data Analysis: A Survey. *IEEE Transactions on Computational Biology and Bioinformatics*, 1(1):24–45, 2004.

[14] R. Marti and M. Laguna. *Scatter Search. Methodology and Implementation in C.* Kluwer Academic Publishers, 2003.

[15] S. Mitra and H. Banka. Multi-objective evolutionary biclustering of gene expression data. *Pattern Recognition*, 39(12):2464–2477, 2006.

[16] J. Nepomuceno, A. Troncoso, J. Aguilar-Ruiz, and J. Garcıa-Gutierrez. Biclusters Evaluation Based on Shifting and Scaling Patterns. *Lecture Notes in Computer Science*, 4881:840–849, 2007.

[17] B. Pontes, F. Divina, R. Giraldez, and J. Aguilar-Ruiz. Virtual Error: A New Measure for Evolutionary Biclustering. *Lecture Notes in Computer Science*, 4447:217–226, 2007.

[18] V. K. Rohatgi. *An introduction to probability theory and mathematical statistics*. New Delhi, 1988.

[19] H. Zhao, A. Liew, X. Xie, and H. Yan. A new geometric biclustering algorithm based on the Hough transform for analysis of large-scale microarray data. *Journal of Theoretical Biology*, 251(2):264–274, 2008.