A Local Search in Scatter Search for Improving Biclusters

Juan A. Nepomuceno
Dpt. Lenguajes y Sistemas Informáticos
University of Seville, Spain
janepo@us.es

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

***Abstract*—Scatter Search is a population-based metaheuristic that emphasizes systematic processes against random procedures. A local search procedure is added to a Scatter Search for Biclustering in order to improve the quality of biclusters. This local search constitutes the existing Improvement Method in most of Scatter Search schemes which intensifies the optimization process, and, consequently, improves the quality of biclusters according to a fitness function. The fitness function is based on linear correlations among genes and, therefore, biclusters with shifting and scaling patterns are obtained. Thus, the improvement of a bicluster consists in removing every pair of genes of such bicluster that has a correlation lower than a given threshold which is automatically chosen by the algorithm. Experimental results from a Yeast microarray data set with different stress conditions have been reported and compared to another algorithm based on Scatter Search recently published in the literature. Experiments show a remarkable performance of the Biclustering algorithm with the proposed local search.**

***Keywords*-Gene Expression Data, Biclustering, Scatter Search.**

I. Introduction

DNA microarrays technology enables us to measure the gene expression level of thousand of genes simultaneously. This technology provides a huge volume of biological information for understanding how biological functions are codified by genes. After several preprocessing steps, the expression matrix of a microarray is a numerical matrix where rows represent genes and columns represent experimental conditions. Each value of the matrix is the expression level of a gene under a specific condition.

Biclustering is an Unsupervised Data Mining technique that searches for local patterns in the gene expression data matrix. The goal of Biclustering is to discover groups of genes with the same behavior under a specific group of conditions. A large number of existing algorithms for Biclustering can be found in different surveys [1], [2], [3]. In the context of microarray analysis, Biclustering was firstly considered by Cheng and Church algorithm [4]. This algorithm is a greedy iterative search method which consists in building a bicluster by adding or removing rows or columns iteratively, and improving its quality through a measure called the *Mean Squared Residue* (MSR). In the last years, biclustering algorithms based on metaheuristics have been published such as evolutionary approaches [5], multiobjective evolutionary approaches [6] or Memetics Algorithms [7]. All these algorithms used the MSR as a part of their fitness function. However, several important patterns from a biological point of view can not be obtained using the MSR. It is shown in [8] that MSR is not an appropriate measure to find scaling patterns when the gene variance values are high in the bicluster.

Recently, some biclustering algorithms are based on the search of linear correlations among genes and, therefore, it is assumed that correlated genes imply co-expressed genes. Moreover, the correlation can capture some classical patterns studied in the bibliography such as shifting, scaling or geometric patterns [8]. A more general pattern, called correlated bicluster, is defined in [9] and a technique based on singular value decomposition is proposed to identify this new kind of patterns. A metaheuristic based on a Scatter Search scheme that uses linear correlations among genes as fitness function is presented in [10]. This algorithm searches for linear correlations among genes but does not capture negative correlations which represent important biological patterns. In fact, negative correlations such as opposite expression patterns or complementary patterns are very important to understand how molecular pathways work [11].

In this work, a local search procedure is included in a Scatter Search scheme that work with positive and negative correlations for Biclustering. This local search constitutes the Improvement Method of the proposed Scatter Search scheme and it intensifies the optimization process. This paper is organized as follows. The proposed Scatter Search algorithm with the above mentioned local search is presented in Section 2. Experimental results from Yeast microarray data set with different stress conditions is reported in Section 3. Moreover, in this section a comparison with a recently published algorithm is presented. Finally, Section 4 outlines the main conclusions of the paper and future works.

II. Biclustering with Scatter Search

Scatter Search [12] is a population-based metaheuristic where a set of individuals that represent possible solutions evolves in order to find optimal solutions of the problem. Scatter Search uses strategies to diversify and intensify the search considering the evolution of a special set of solutions

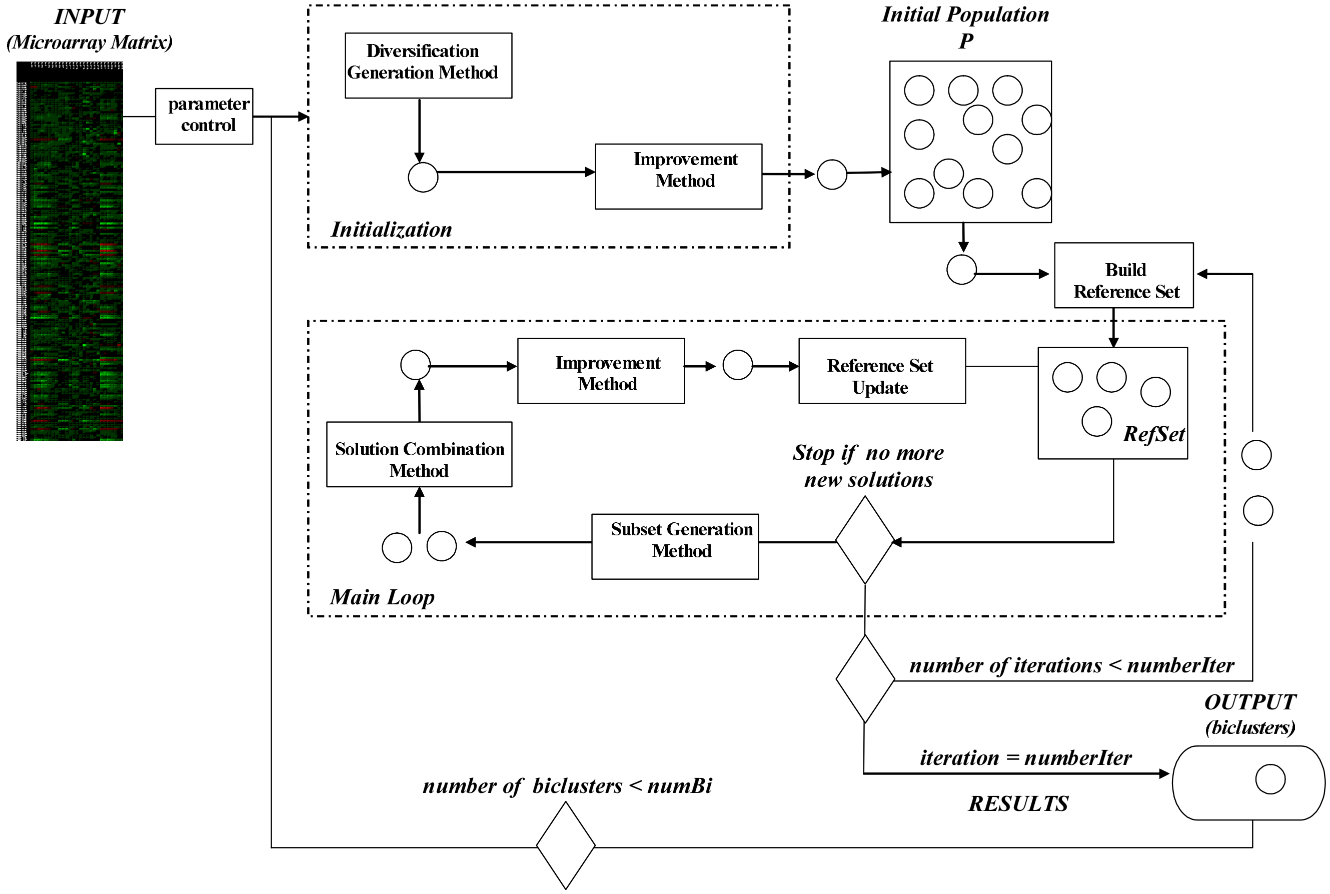

Figure 1. Scatter Search Scheme for Biclustering.

called *Reference Set*. This metaheuristic emphasizes systematic processes against well-known random procedures used in Genetic Algorithms.

The scheme of the proposed Scatter Search for Biclustering is shown in Figure 1. The input data are the matrix of gene expression of the microarray and two parameters M_1 and M_2 in order to control the volume of the bicluster to be obtained. Basically, the algorithm works as follows. After the Initial Population has been built, the Reference Set is initially built with the best solutions from the population, according to the value of their fitness function, and the most scattered ones from the population regarding the previous best solutions. This set evolves by considering new solutions generated by the *Subset Generation Method* and the *Solution Combination Method*. These new solutions are improved by the *Improvement Method* and finally, the *Reference Set Update Method* updates the set according to quality criteria that depend on the fitness function. The process is repeated until the set is stable, that is, it does not change. Later, the Reference Set is rebuilt by considering the best solutions from the new solutions and the most scatter ones from the initial population. After a given number of iterations, the final solution is the best one in the Reference Set. The intensification of the search is due to a local search procedure called Improvement Method where the solutions are improved by exploiting the knowledge of the problem.

Solutions are codified as binary strings which represent whether a gene or a condition of the microarray matrix belongs or not to the bicluster. The Initial Population is built with the solutions generated by the *Diversification Generation Method* where solutions are generated as scatter each other as possible. Every solution is improved before being inserted in the population. Once the Reference Set is stable, it has to be rebuilt with the best solutions from the Reference Set, according to the fitness function, and the most distant solutions, regarding the previously chosen best ones, from the initial population. The first time the Reference Set is built, the best solutions are chosen from the initial population and the remaining times from the Reference Set of the previous iteration. The initial population has to be updated in the evolutionary process by removing solutions which have already been considered in the building of the Reference Set. New solutions are inserted in the search process by the Subset Generation Method and by the Solution Combination

Method. By means of both methods, all pairs of biclusters in the Reference Set are combined using an uniform crossover operator. The Improvement Method is a local search that improves the value of the fitness function for each bicluster.

A. Fitness Function based on linear correlations

Given a bicluster B composed by N genes, the average absolute correlation $\rho_{|\cdot|}(B)$, is defined as:

$$\rho_{|\cdot|}(B) = \frac{1}{\binom{N}{2}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} |\rho_{ij}| \qquad (1)$$

where ρ_{ij} is the correlation coefficient between the gene i and the gene j. Due to $\rho_{ij} = \rho_{ji}$ only $\binom{N}{2}$ elements have been considered. The absolute value for correlation among genes has been considered in order to find negative correlations among genes.

Note that $-1 \leq \rho_{ij} \leq 1$, if the value is close to 0 the gene i and the gene j show different behavior. However, if ρ_{ij} is close to 1 both genes have the same behavior. Values from 0 to 1 indicate that two genes are *positively correlated* and they show the same tendency, that is, if one gene increases its value, the other one increases its value too. Values from -1 to 0 indicate that both genes are *negatively correlated* and they have complementary tendency, that is, if one gene increases the other one decreases with the same intensity and vice versa.

The fitness function used to evaluate the quality of biclusters is defined by:

$$f(B) = (1 - \rho_{|\cdot|}(B)) + \sigma_\rho + M_1 \left(\frac{1}{nG}\right) + M_2 \left(\frac{1}{nC}\right) \quad (2)$$

where nG and nC are the number of genes and conditions of the bicluster B, respectively, M_1 and M_2 are penalty factors to control the volume of the bicluster B, and σ_ρ is the standard deviation of the values ρ_{ij}. The standard deviation is included in order to avoid that the value of the average correlation can be high for a bicluster and this bicluster can contain several non-correlated genes with the remaining ones of the bicluster.

B. Local search procedure

The Improvement Method is a local search that improves the value of the fitness function for a solution in order to intensify the search process. In Scatter Search schemes, the improvement method is usually defined depending on the problem under study. The optimization process of the considered fitness function (Equation 2) maximizes the average absolute correlation and the volume of each bicluster. When the function fitness is minimized by the proposed Scatter Search the $\rho_{|\cdot|}$ tends to 1.

The Improvement Method is a local search that improves the value of $\rho_{|\cdot|}$ because of the proximity of its value to 1. Hence, the improvement of the fitness function is only related with the correlations of biclusters. The Improvement Method consists in removing every pair of genes in a bicluster that has a correlation value lower than a threshold.

Figure 2 presents an example where a bicluster composed by four genes is improved. The value of its $\rho_{|\cdot|}$ is equal to 0.68 but after removing $gen4$ the value becomes in 0.91. Genes 1, 2 and 3 are highly correlated ($\rho_{12} = 0.80$, $\rho_{13} = 0.99$ and $\rho_{23} = 0.81$). However, the $gen4$ is not so correlated with the remaining genes ($\rho_{14} = 0.42$, $\rho_{24} = 0.00$ and $\rho_{34} = 0.42$). Therefore, if 0.5 is, for example, chosen as threshold for removing genes, $gen4$ is deleted from the bicluster. Only a ρ_{ij} value lower than the threshold is enough to remove the gene from the bicluster. Note that the $gen1$ and $gen3$ are highly positively correlated, $gen2$ is highly negatively correlated with $gen1$ and $gen3$ as it has the opposite tendency to them and finally, $gen4$ shows a different behavior regarding the other three genes.

This Improvement Method needs a parameter as threshold to remove genes. For each execution of the Scatter Search algorithm, this parameter is automatically generated by an empirical study with random biclusters. One hundred biclusters are generated and the value from 0.1 to 0.9 that improves the greater number of biclusters is chosen. Therefore, the best value for the parameter is chosen depending on the data set.

III. Experiments

The data set used for experiments has been *yeast Saccharomyces Cerevisiae Stress conditions* which is a microarray composed by several time series that represent the Yeast response to different stress conditions. This data set was provided in [13] and it was used in [14] where can be downloaded as supplementary data. It is composed by 2993 genes and 173 conditions. The proposed algorithm has been executed for several values of the penalization parameters, in particular, $M_1 = 1$, $M_2 = 1$ and $M_1 = 10$, $M_2 = 10$. These parameters control the number of genes and conditions of the biclusters to be obtained, respectively.

The results obtained by the Scatter Search approach for Biclustering published in [10] have been used in order to compare to that of the proposed algorithm. This algorithm, SScorr, uses as fitness function the average correlation (without absolute values) and an Improvement Method that removes the negatively correlated genes.

Table I shows several measures for the 100 biclusters obtained by the proposed method and the SScorr algorithm. Both algorithms provide one bicluster for each run, therefore, 100 executions have to be made. It is shown the average of the number of genes, the average of the number of conditions, the average volume (number of genes x number of conditions), the average absolute correlation, the average correlation, the average of the number of negatively correlated genes and the value of the automatically adjusted threshold used in the proposed local search. This last value is

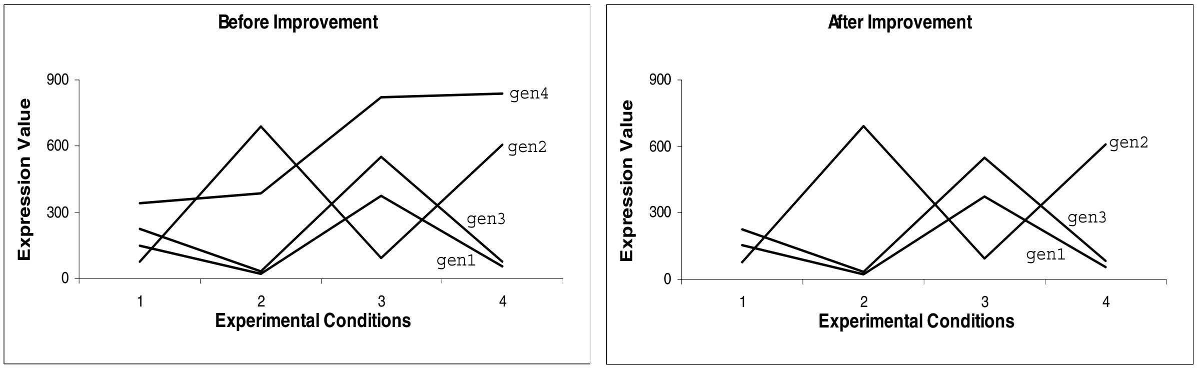

Figure 2. Local Search for Biclustering based on linear correlations among genes.

	Number of genes	Number of conditions	Volume	Average absolute correlation	Average correlation	Correlated negatively	Threshold	MSR	Gene variance
Proposed method $(M_1 = 1, M_2 = 1)$	15.47	18.56	289.19	**0.94**	0.66	18	0.6	0.55	1.53
Proposed method $(M_1 = 10, M_2 = 10)$	72.9	29.45	2106.51	**0.81**	0.35	1082.14	0.2	0.85	1.24
SScorr $(M_1 = 1, M_2 = 1)$	16.36	14.8	237.62	0.89	**0.89**	0	–	0.31	1.50
SScorr $(M_1 = 10, M_2 = 10)$	46.69	27.19	1269.44	0.72	**0.72**	0	–	0.37	1.02

Table I
RESULTS FROM EXPERIMENTS: 100 BICLUSTERS.

not an average, it is the same value for all executions of the algorithm. Although the MSR and the gene variance are not included in the fitness function of the proposed algorithm, the average value of the MSR and the average of the gene variance are specified in the two last columns as they are commonly used in the literature to determinate the quality of biclusters.

It can be easily observed the influence of M_1 and M_2 values over the size of biclusters found by both algorithms. The higher values for penalization values, the higher volume is.

The average absolute correlation values are in bold for the proposed method because this measure is included in its fitness function which has been minimized. However, the average correlation is shown in the Table but this measure has not been considered in the optimization process. For SScorr algorithm the situation is the opposite, the average correlation is in bold because it is the measure which is used in the fitness function. Both measures are presented in order to show the remarkable difference among the results obtained when using them in the fitness function.

It can be noted that high values for the average absolute correlation are obtained by the proposed algorithm and how this situation does not imply high values for the average correlation. For example, the average absolute correlation and the average correlation are 0.81 and 0.35, respectively for the proposed method with the penalizations $M_1 = 10$ and $M_2 = 10$. This is due to the pairs of negatively correlated genes decrease the value of average correlation but increase the value of the average absolute correlation. Both measures are the same for SScorr algorithm as its Improvement Method removes pairs of negatively correlated genes, and, therefore, all correlation coefficients among genes are greater than 0.

As expected, the number of negatively correlated genes is equal to 0 for SScorr algorithm. However, the proposed method finds biclusters with a huge number of negative correlations. All these biclusters are not able to be provided with the SScorr algorithm, and, however they show relevant patterns from a biological point of view..

It can be appreciated that the minimum allowed correlation used in the proposed local search to improve the biclusters is 0.6 and 0.2 using the proposed algorithm for $M_1 = 1, M_2 = 1$ and $M_1 = 10, M_2 = 10$, respectively. Note that if M_1 and M_2 are high then the terms related to the volume in the fitness function (Equation 2) have a weight greater than the remaining terms and, therefore, the threshold for the minimum allowed correlation must be not so restrictive to improve the fitness function values.

Finally, the biclusters obtained by the SScorr algorithm are better, regarding the MSR, than that of the proposed method. However, the first ones do not contain any pair of negatively correlated genes, which can be very relevant from a biological point of view. This fact is coherent with some

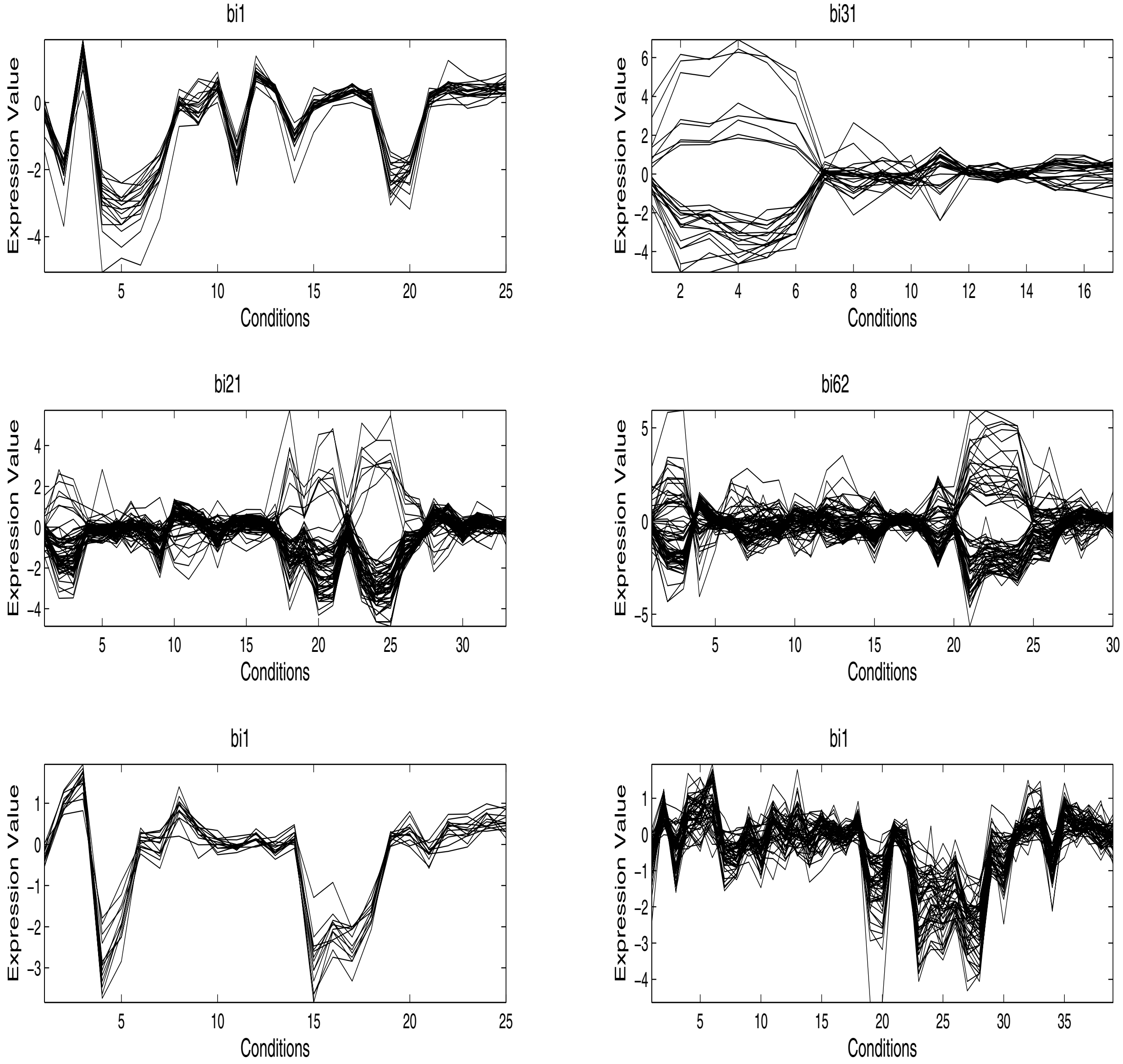

Figure 3. Geometrical representation of several biclusters from experiments.

reviews made to the MSR measure in [8].

Figure 3 presents the gene expression values along with the experimental conditions for 6 biclusters selected from the 100 biclusters. The $bi1$ and $bi31$ biclusters and $bi21$ and $bi62$ biclusters are obtained by the proposed algorithm for penalization factors $M_1 = 1, M_2 = 1$ and $M_1 = 10, M_2 = 10$, respectively. The $bi1$ bicluster contains a low number of genes due to the penalization parameters are low ($M_1 = 1, M_2 = 1$). It can be noted that all genes are positively correlated in this bicluster. Therefore, it could be found by the SScorr algorithm as the average absolute correlation is equal to the average correlation, concretely, 0.97. However, the $bi31$ bicluster is a clear example of a bicluster that can be found by the proposed method but not by the SScorr algorithm as its Improvement Method increases the average correlation of a bicluster by avoiding biclusters with this kind of patterns. In fact, the average absolute correlation is 0.94 and the average correlation is 0.09. It can be seen that there is a group of genes with opposite tendency but analogous intensity under conditions from 1 to 7, that is, when this group of genes increase the remaining genes decrease and vice versa. A similar scenario shows the $bi21$ and $bi62$ biclusters on the middle of the Figure. These biclusters have high values for the average absolute correlation (0.86 and 0.78, respectively), and low values for the average correlation (0.45 and 0.12, respectively). It can be easily observed that both biclusters contain negatively correlated genes. By the other hand, the two $b1$ biclusters, on the bottom of the Figure, are obtained by the SScorr algorithm for $M_1 = 1, M_2 = 1$ (on the left) and $M_1 = 10, M_2 = 10$ (on the right), respectively. Both biclusters have highly positively correlated genes but they have not any pair of negatively correlated genes. Therefore, although the SScorr algorithm obtains good biclusters as they contain genes with scaling and shifting patterns, the local search applied to improve the biclusters is not adequate because it deletes negative correlations which are important from a biological point of view [11].

IV. Conclusions

A local search procedure has been added to a Scatter Search for Biclustering in order to improve the quality of biclusters. Biclusters with certain patterns such as shifting and scaling patterns are considered high-quality biclusters. Therefore, the fitness functions is based on linear correlations such that as negative correlations as positive correlations are taken into consideration. The proposed local search has been used as the existing Improvement Method in Scatter Search schemes and it removes every pair of genes in a bicluster that has a correlation value lower than a given threshold which is automatically chosen by the algorithm for each data set. Experimental results from a Yeast microarray data set with different stress conditions have been reported showing that biclusters with negatively correlated groups of genes have been found. Moreover, a comparison with an recently published algorithm based on Scatter Search has been made.

Future works will focus on a biological study of biclusters found in order to establish a comparison with other algorithms.

Acknowledgment

The financial support given by the Ministry of Science and Innovation, projects TIN2007-68084-C02, PCI2006-A7-0575, and by the Junta de Andalucía, project P07-TIC-02611 is acknowledged.

References

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

[2] A. Tanay, R. Sharan, and R. Shamir. Biclustering Algorithms: A Survey. *Handbook of Computational Molecular Biology*, 9:26–1, 2005.

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

[4] Y. Cheng and G.M. Church. Biclustering of Expression Data. *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, 8:93–103, 2000.

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

[6] H. Banka and S. Mitra. Evolutionary Biclustering of Gene Expressions. *Ubiquity*, 7(42):1–12, 2006.

[7] Cristian Andrés Gallo, Jessica Andrea Carballido, and Ignacio Ponzoni. Microarray biclustering: A novel memetic approach based on the pisa platform. In *EvoBIO 2009: Proceedings of the 7th European Conference on Evolutionary Computation, Machine Learning and Data Mining*, pages 44–55, 2009.

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

[9] Hong Yan Wen-Hui Yang, Dao-Qing Dai. Finding correlated biclusters from gene expression data. *IEEE Transactions on Knowledge and Data Engineering*, IEEE computer Society Digital Library. IEEE Computer Society.(4):568–584, 2011.

[10] Juan A. Nepomuceno, Alicia Troncoso, and Jesus Aguilar-Ruiz. Biclustering of gene expression data by correlation-based scatter search. *BioData Mining*, 4(1):3, 2011.

[11] Tao Zeng and Jinyan Li. Maximization of negative correlations in time-course gene expression data for enhancing understanding of molecular pathways. *Nucl. Acids Res.*, page gkp822, 2009.

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

[13] Audrey P. Gasch, Paul T. Spellman, Camilla M. Kao, Orna Carmel-Harel, Michael B. Eisen, Gisela Storz, David Botstein, and Patrick O. Brown. Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. *Mol. Biol. Cell*, 11(12):4241–4257, 2000.

[14] A. Prelic, S. Bleuler, P. Zimmermann, A. Wille, P. Buhlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129, 2006.