Evolutionary Search of Biclusters by Minimal Intrafluctuation

Raúl Giráldez, Federico Divina, Beatriz Pontes and Jesús S. Aguilar–Ruiz

Abstract— **Biclustering techniques aim at extracting significant subsets of genes and conditions from microarray gene expression data. This kind of algorithms is mainly based on two key aspects: the way in which they deal with gene similarity across the experimental conditions, that determines the quality of biclusters; and the heuristic or search strategy used for exploring the search space. A measure that is often adopted for establishing the quality of biclusters is the *mean squared residue*. This measure has been successfully used in many approaches. However, it has been recently proven that the mean squared residue fails to recognize some kind of biclusters as quality biclusters, mainly due to the difficulty of detecting scaling patterns in data. In this work, we propose a novel measure for trying to overcome this drawback. This measure is based on the area between two curves. Such curves are built from the maximum and minimum standardized expression values exhibited for each experimental condition. In order to test the proposed measure, we have incorporated it into a multiobjective evolutionary algorithm. Experimental results confirm the effectiveness of our approach. The combination of the measure we propose with the mean squared residue yields results that would not have been obtained if only the mean squared residue had been used.**

I. Introduction

Microarray technologies offer the possibility of studying the activity of thousand of genes simultaneously [1], [2], and allow us to analyze biological phenomena such as development or evolution, determining the functionality of new genes, or even detect how a large number of genes interact one each other. Thus, from a collection of microscopic DNA spots attached to a solid surface, gene expression values can be represented as bidimensional numerical matrices, where each column corresponds to a gene, and each row corresponds to a sample (experimental condition). Each entry of the matrix denotes, therefore, the expression level of a gene under a certain condition.

Clustering techniques can be used to group genes that behave similarly under a set of conditions [3]. A set of genes is said to behave in a similar way when all the genes exhibit coherent rises and falls of their expression values across all the conditions. However, relevant genes are not necessarily related to every condition. In other words, some genes may be relevant only for a subset of conditions. From this point of view, clustering cannot be addressed with respect to genes or conditions independently, but also in the two dimensions simultaneously. This approach is called biclustering and aims at grouping genes presenting similar trends under a subset of conditions. That is, genes in a bicluster would be correlated with regard to a specific subgroup of samples. Biclustering was first introduced by Hartigan [4], and was first applied to biological data by Cheng and Church [5]. Since then, a number of methods for biclustering of biological and biomedical data have been proposed [6], [7]. Biclustering techniques can be characterized by two main features: the search strategy used for finding relevant subsets of genes and samples, and the specific measure used for evaluating the quality of the biclusters.

With regard to the first feature, it is important to mention those methods based on a greedy search strategy, such as those proposed by Cheng and Church [5], by Ben–Dor et al. [8], and most recently a polynomial time algorithm presented by Liu and Wang [9] (however, with several constraints). Other search methods include exhaustive search [10], [11], or stochastic techniques [12], [13], [14].

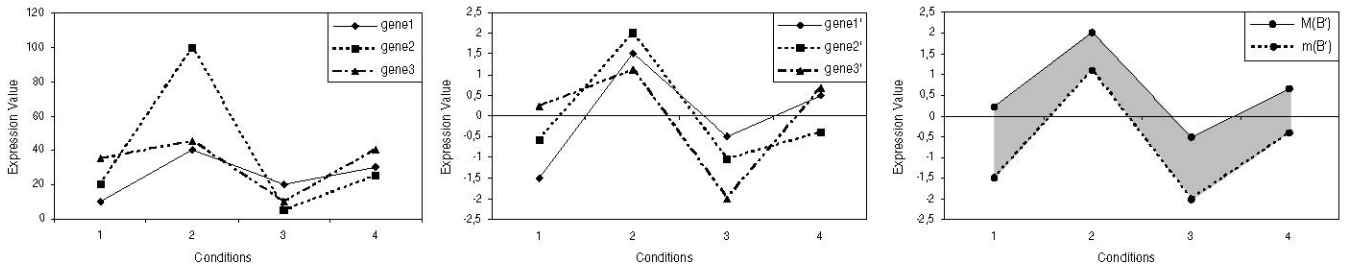
One of the most widely used measures for assessing the quality of biclusters is the *mean squared residue* (MSR) [5]. MSR evaluates the numerical coherence of gene expression values across the selected experimental conditions. In order to do so, the arithmetic means of the values in each row, column, and the full matrix are computed, and the numerical differences among the data are quantified. MSR has proven to be an effective quality measure. Nevertheless, it may fail to recognize some kind of quality biclusters, as it has been demonstrated in [15]. In this work, the author unambiguously settles the definition of the two main elements inherent to biclusters: shifting and scaling patterns, which induce linear correlations among genes. In the case of a perfect shifting pattern, the genes of the bicluster present a parallel behavior [15], i.e., the expression values of a gene can be obtained by adding a certain offset to the value of another gene. A bicluster presents a scaling pattern when the expression level values of genes are proportional to the value of another gene, i.e., the genes can assume the same trend but not necessarily parallel. Thus, genes in a bicluster might present either one of these patterns or both of them simultaneously. With regard to the effectiveness of MSR for evaluating biclusters, it has been proven that this measure is effective for recognizing shifting patterns, but it may fail to identify scaling trends within a bicluster, due to the variance of the scaling factor (for a detailed discussion, see [15]).

This fact represents the main motivation for this work, where we propose a new measure for evaluating biclusters. This new measure is named MSA (*Maximal Standard Area*), due to the way in which the coherence among genes is calculated.

In order to test the performance of our proposal, we incorporated MSA in a multi–objective Evolutionary Algorithm (MOEA). The main reason why we have chosen to use a MOEA, is that when searching for biclusters in microarray

Raúl Giráldez, Federico Divina and Jesús S. Aguilar–Ruiz are with the School of Engineering, Pablo de Olavide University, Seville, Spain (email: {aguilar,giraldez,fdivina}@upo.es). Beatriz Pontes is with the Department of Computer Science, University of Seville, Spain (email: bepontes@lsi.us.es).

1-4244-1210-2/07/$25.00 ©2007 IEEE.



(a) Bicluster B (b) Standardized Bicluster B' (c) M(B'), m(B') and MSA

Fig. 1. Example to illustrate the Maximal Standard Area (MSA). In (a) a bicluster example; in (b) its standardization; in (c) the MSA.

data, several objectives, e.g., the volume and MSA of the biclusters, are to be optimized at the same time. Often, these objectives are in conflict with each other. It follows that the problem of finding biclusters can be straightforwardly seen as a multi–objective optimization problem.

The remainder of the paper is organized as follows: Section II describes the quality measure which is proposed. A description of the MOEA used in this paper, and of how MSA has been included into the algorithm, is given in Section III, while some experimental results are discussed in Section IV. Finally, Section V summarizes the main conclusions.

II. Maximal Standard Area

As already stated, the MSR cannot recognize some kind of biclusters as quality biclusters. In order to overcome this drawback, we propose a novel measure for assessing the quality of biclusters on microarray, called *Maximal Standard Area* (MSA). The idea we base our proposal on is the area of the region between the maximum and minimum values of expression levels that genes assume under the conditions contained in the bicluster. Thus, what is measured is the area depicted by the maximal fluctuation of expression levels for each experimental condition. For each condition, the minimum and maximum values of expression level for all the genes contained in the bicluster are taken. These pairs of values define a band across all the conditions in the bicluster, and the area of this band is therefore the measure MSA. In the following of this section we will provide a detailed description of the measure we are proposing.

An important observation that can be extracted from an analysis of previously found biclusters (e.g., [5], [12], [14]), is that the range of expression values assumed by genes can vary substantially depending on the specific microarray taken as input. Therefore, in order to make an appropriate comparison between each gene and the pattern, it would be desirable to define a mechanism for scaling the expression levels to a common range. This mechanism would also be responsible for softening every gene behaviour, since the most important aspect is to characterize its tendency rather than its numerical values.

In the following, let $\mathcal{B}$ be a bicluster containing J genes and I conditions, and let b_{ij} denote the elements of $\mathcal{B}$, for $1 \leq i \leq I$ and $1 \leq j \leq J$.

Definition 1 (Standardization) We define the standardization of $\mathcal{B}$ as the bicluster $\mathcal{B}'$, whose element b'_{ij} are obtained as follows:

$$b'_{ij} = \frac{b_{ij} - \mu_{g_j}}{\sigma_{g_j}}, 1 \leq i \leq I, 1 \leq j \leq J$$

where σ_{g_j} is the standard deviation of all the expression values of gene j and μ_{g_j} is the mean of column j.

By means of the standardization, two distinct tasks are carried out. The first one is to shift all the genes to a similar range of values (near 0 in this case). The second one is to homogenize the expression values for each gene, modifying in this way their values under all the conditions, and smoothing their graphical representation, due to the correction of the global scaling factor in the denominator (notice that the *global scaling factor* is not the same as the *local scaling factor*, as it is described in [15]) .

Definition 2 (Bounds of bicluster $\mathcal{B}$) We define the upper bound of a bicluster $\mathcal{B}$ for condition i as

$$M_i(\mathcal{B}) = max_j \; b_{ij}, \forall j$$

and similarly the lower bound of bicluster $\mathcal{B}$ for condition i as

$$m_i(\mathcal{B}) = min_j \; b_{ij}, \forall j$$

We can now define the proposed measure:

Definition 3 (MSA) We define the Maximal Standard Area, MSA($\mathcal{B}'$), as the area delimited by the bounds for each condition as follows:

$$\text{MSA}(\mathcal{B}') = \sum_{i=1}^{I-1} \left| \frac{M_i(\mathcal{B}') - m_i(\mathcal{B}') + M_{i+1}(\mathcal{B}') - m_{i+1}(\mathcal{B}')}{2} \right|$$

where $\mathcal{B}'$ is the standardized bicluster.

As an example, Figure 1(a) shows a bicluster $\mathcal{B}$ containing three genes and four conditions. In Figure 1(b) the resulting standardized bicluster $\mathcal{B}'$ is displayed. It can be noticed that the standardized genes assume closer values than the original ones. In particular, in Figure 1(a), the value assumed by the second gene under the second condition is about 2.5 times greater than the value assumed by the other two genes under

```
Output: List of Biclusters L
    load Expression Matrix EM
    list L = {}
    bicluster b
    repeat
        b = MOEB (EM)
        if b is not null
            add b to L
            adjust weights of EM
        else
            end_cond met
        if max_iter is reached
            end_cond met
    until end_cond is met
    return L
```

(a) **SMOB** for Sequential Covering

```
Input: Expression Matrix EM
Output: Best individual in population
    initialize Population
    evaluate Population
    repeat
        select parents
        recombine pairs of parents
        mutate the resulting offspring
        evaluate new individuals
        select individuals for next generation
    until max_iter is reached
    best_ind = best individual in Population
    return best_ind
```

(b) Procedure **MOEB**

Fig. 2. A general scheme of both the sequential covering algorithm and of the procedure MOEB.

the same condition. As a result of the standardization, this difference is much less stressed, while the general tendency of the three genes is maintained. Finally, Figure 1(c) shows $M(\mathcal{B}')$ and $m(\mathcal{B}')$ for each condition. MSA($\mathcal{B}'$) is illustrated by the grey region.

If the genes of a bicluster $\mathcal{B}$ have a perfectly coherent trend then MSA($\mathcal{B}$) is equal to zero. On the contrary, MSA will assume higher values when the genes are less correlated with each other, due to the fact $M(\mathcal{B}')$ and $m(\mathcal{B}')$ are more distant from each other. It follows that we assume that biclusters characterized by a low MSA are interesting for further biological studies.

III. The Algorithm

In order to assess the validity of MSA, we introduce a multi–objective evolutionary algorithm (MOEA), called SMOB (for Sequential Multi–Objective Biclustering). SMOB is outlined in Figure 2. The algorithm is similar to SEBI [14] and adopts a sequential covering strategy. Unlike SEBI, where a single–objective EA was used, SMOB invokes several times a multi–objective evolutionary algorithm. Each time the MOEA is called, a bicluster is returned. Biclusters returned are stored in a list, until the evolutionary algorithm is called a maximum number of times. In [14] a threshold δ on MSR was used in order to reject biclusters. In SMOB we do not use any threshold. This is because several objectives are optimized at the same time, and thus biclusters cannot be rejected based on a bad result of a single objective.

Four objectives are to be optimized simultaneously: MSA, MSR, volume and gene variance. By optimizing MSA and MSR, we aim at overcoming the drawbacks of the MSR regarding the large size of some biclusters, which are usually flat, with very low variance of expression levels. Still, the advantages of such measure with respect to the discover of shifting patterns are taken into account. In short, the MSR and MSA will be minimized, while the volume and the gene variance will be maximized.

We can individuate two main reasons that justify the use of a MOEA for finding biclusters. First, the problem of finding biclusters in an expression matrix can be straightforwardly seen as a multi–objective problem. Indeed, we are interested in finding biclusters with high volume, low mean square residue, low area and relatively high expression level variance. Thus, there are at least four objectives to optimize and these objectives are in conflict with each other. For example, a bicluster consisting of just one element has mean squared residue equal to zero, or, again, a constant bicluster have gene variance equal to zero, but also mean squared residue equal to zero. Second, by using a MOEA it will not be necessary to combine all the objectives into a single fitness function, which might become complicated, especially when both maximization and minimization are involved. Finding a way to combine the objectives in a single function can be problematic, and may require more parameters to set [16].

The encoding of biclusters is the one proposed in [14], where bit strings are used. A bit is associated to each gene and each condition of the expression matrix. If a bit is set to one, it means that the relative gene/condition belongs to the bicluster, otherwise it does not.

Individuals are initialized in the following way. First the number of genes J and of conditions I contained in the biclusters are randomly determined. Then, J bits corresponding to genes and I bits corresponding to conditions are randomly selected. The selected bits are set to one, which means that the relative gene/condition is contained in the bicluster encoded by the individual. We perform this initialization instead of a pure random initialization of bit–strings, because in that way the initial biclusters would contain all about the same number of genes and conditions.

The fitness $f(x)$ of an individual x is calculated on the basis of the Pareto dominance, namely $f(x)$ is based on the number of individuals n that x dominates. x is said to dominate another individual y if x is not worst than y on all objectives, and x is better than y on a least one objective.

Moreover, in order to promote diversity in the population, two distance measures are used: one is calculated on the objective set and the other one on the decision set. The former is implemented by calculating the distance from the nearest neighbor in terms of objectives, i.e., MSA, MSR, volume and expression level variance. The latter is the normalized average number of individuals covering the same elements of the expression matrix included in the bicluster

being evaluated. The inverses of these two distances are added to the fitness.

The fitness of an individual x is then given by $f(x) = \frac{1}{n} + \frac{1}{dist_{obj}} + P(x)$, where n is computed as described above, $dist_{obj}$ is the distance considering the objectives. $P(x)$ is used in order to avoiding overlapping among biclusters, and is described in the following. Notice that the fitness has to be minimized.

Individuals are selected with a tournament mechanism, with a tournament size of four. Three crossover operators are used with different probabilities: one–point crossover, two–point crossover and uniform crossover. The application of the uniform crossover is the one having the highest probability. Uniform crossover is preferred to the other two crossovers because one–point and two–point crossover would prohibit certain combinations of bits to be crossed over together [12].

Three mutation operators are used: a classical mutation operator, one that can add a row and one that can add a column. We consider columns and rows separately, because typically there are many more columns than rows, thus considering them together, would give more probability of mutation to columns than to rows.

Elitism is applied by letting the non–dominated individuals survive to the next generation.

In order to avoid overlapping among biclusters, after each call of MOEB, we assign a weight to each element e_{ij} of the expression matrix. This weight w_{ij} is equal to the number of biclusters stored in the *Results* list that contain e_{ij}.

When a bicluster x is evaluated inside MOEB, a penalty $P(x) = 1 - \frac{V_x - \sum_{i,j \in x} w_{e_{ij}}}{V_x}$ is added to the fitness of x, where V_x is the volume of x. In this way, if a bicluster has low volume and it covers elements of the expression matrix that are already contained in many biclusters already found, $P(x)$ will be high. On the other hand, if the bicluster has a high volume and it overlaps with few biclusters, the penalty will be lower. If the bicluster x does not overlap with any bicluster then $P(x)$ is zero.

IV. Experiments

In order to show the validity of our approach, we perform experiments on two–well known datasets: the yeast *Saccharomyces cerevisiae* cell cycle expression dataset [17], and the human B–cell expression data [18]. The former dataset is a microarray which contains 2884 genes and 17 conditions, while the latter consists of 4026 genes and 96 conditions.

With regard to the parameters of SMOB, we used a population size of 200 individuals and a number of generations of 100. The crossover probability is set to 0.85 and the mutation probability to 0.2. The number of biclusters was set to 100, that is, SMOB generated one hundred biclusters for each dataset.

We want to highlight the fact that all the biclusters found on both datasets showed interesting behaviors, especially regarding the shifting patterns located in them. In the following we show six biclusters for each dataset. These biclusters were selected in order to show the different types of biclusters that were found by the algorithm.

TABLE I
Information relative to the six biclusters shown in Figure 3.

Bicluster	MSA	MSR	Volume	Gene Var.
yeast-39	30.9	504.0	154	1590
yeast-53	14.2	176.1	110	1023
yeast-78	23.8	524.2	120	1752
yeast-95	26.4	445.3	240	2402
yeast-96	23.8	338.8	143	1028
yeast-97	18.8	226.7	132	1007

A. Yeast Dataset

Figure 3 shows the six selected biclusters out of the one hundred found on the yeast dataset. We can notice from a visual inspection of the six biclusters that the genes present a similar behavior under a subset of conditions. As we mentioned before, this holds also for all the biclusters found. Moreover, most of the obtained biclusters show both shifting and scaling patterns.

Table I gives the numerical results for such bicluster. The first column of the table shows the name of the biclusters and the next four columns give the obtained values for the objectives considered: the MSA, the mean squared residue, the volume and the gene expression level variance, respectively.

In general, most of the biclusters obtained on this dataset, presented high values of MSR. Nevertheless many of such biclusters were interesting, presenting good level of coherence among genes. For instance, taking into account that the δ value used by Cheng and Church [5] in the experiment with the yeast dataset was 300, four out of the six selected biclusters would have been rejected. Among the biclusters that would not have been accepted, bicluster *yeast-95* is the clearest example of the drawback of MSR discussed in the Introduction. This is an important issue, and it is due to the dependency of the mean squared residue on the scaling factor, as it was demonstrated in [15]. Therefore, our approach is able to discover these kind of biclusters as well.

The other biclusters shown in Figure 3 are also interesting. Bicluster *yeast-97* shows a clear scaling pattern. Two genes present a parallel behavior as the others with different level of expression values. Another example of noticeable bicluster is *yeast-96*. This bicluster also shows a clear scaling pattern. In fact, under $5^{th}, 6^{th}$ and 7^{th} conditions the expression values assumed by the genes decrease and increase again with different slopes. This bicluster has also a value of MSR higher than the δ used in [5].

It is also worth noticing that MSR and MSA are not necessarily correlated. This can be observed from biclusters *yeast-96* and *yeast-78*. These two biclusters present the same level of MSA while the MSR of *yeast-78* is higher than the MSR of *yeast-96*.

B. Human Dataset

Six out of one hundred biclusters obtained on the human dataset are shown in Figure 4. Information about the shown biclusters is given in table II.

As it can be noticed these are very interesting biclusters. The genes contained in them present an extremely similar behavior. In particular the four genes contained in bicluster

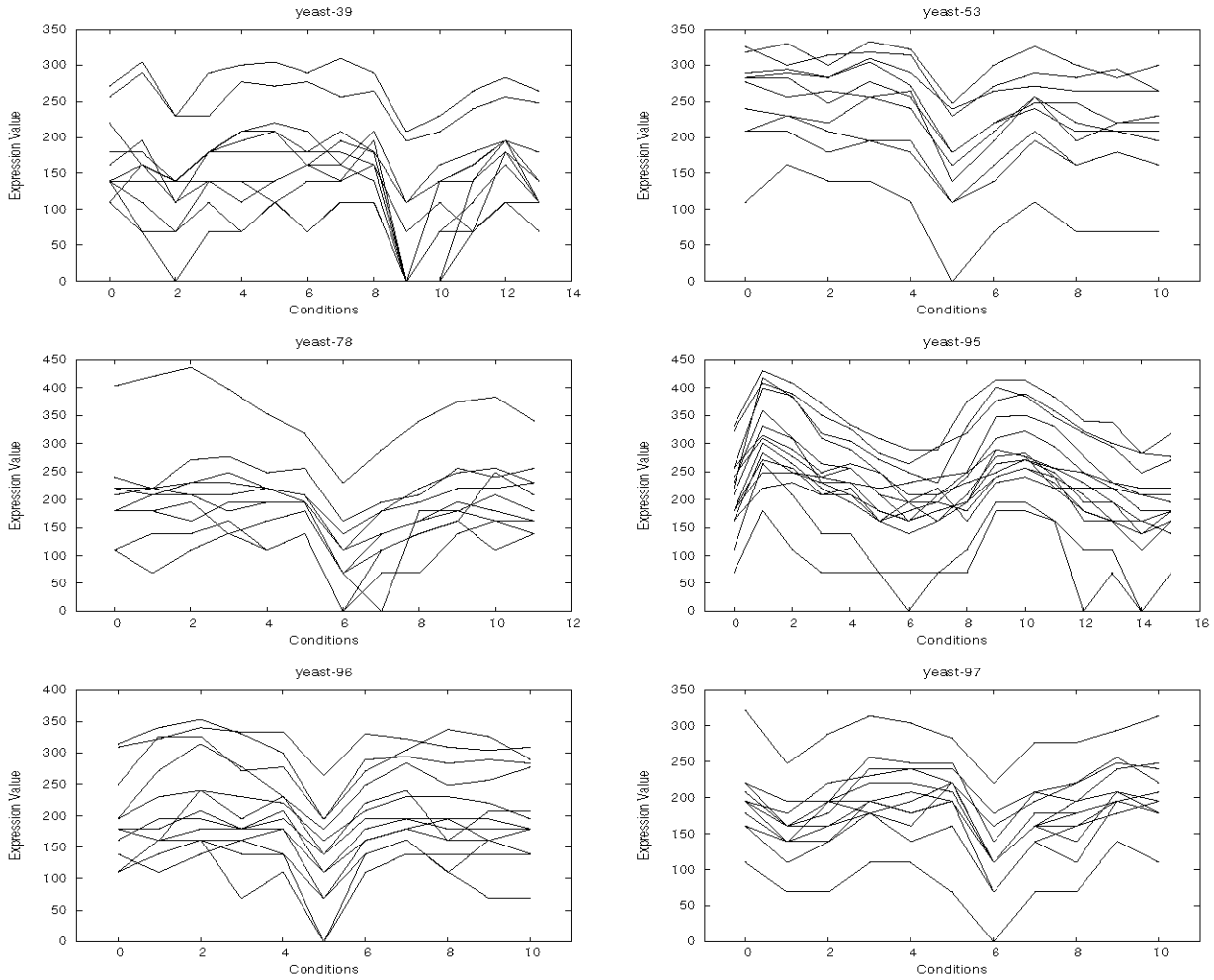


Fig. 3. Six biclusters found on the Yeast dataset.

human-45 behave practically in the same way. This fact is reflected by the very low MSR and MSA characterizing this bicluster. Moreover, of the six selected bicluster, *human-45* is the only one having a level of MSR smaller than the δ (1200) used in [5]. As in the yeast dataset, also most of the biclusters found by SMOB on the human dataset present values of MSR higher than δ.

As an example of such bicluster, we can consider biclusters *human-34*, which presents a very high level of MSR. Nevertheless, this bicluster is graphically interesting. Another instance of this fact is bicluster *human-14*, whose genes behave very similarly. Finally, another evidence that MSR and MSA are not correlated is provided by biclusters *human-15* and *human-34*.

V. Conclusions

In this paper we have presented a novel measure for assessing the quality of bicluster on microarray gene expression data, called MSA. Our main motivation is represented by the fact that even if MSR has been successfully used by many algorithms for finding biclusters on microarray data, it fails to recognize some kind of biclusters as quality biclusters.

In order to test the validity of our proposal, we used MSA as an objective to be optimized in a MOEA, together with the MSR, the volume and the gene variance. In this way we wanted to exploit the good characteristics of MSR and try to overcome its drawback by using MSA.

TABLE II
Information relative to the six biclusters shown in Figure 4.

Bicluster	MSA	MSR	Volume	Gene Var.
human-11	50.4	3470.2	434	20877
human-14	33.2	2077.2	238	16571
human-15	40.3	2054.0	420	7794
human-34	38.3	5523.0	315	19919
human-45	8.1	83.3	124	7388
human-100	26.9	2916.6	162	11548

Experimental results obtained on two datasets shows that MSA is effectiveness for overcoming the drawback of MSR. In fact, by using MSA as an objective, the algorithm could find biclusters that are interesting, but that are characterized by high values of MSR. Discovering such biclusters would be difficult for a heuristic based exclusively on MSR.

Another conclusion that we can draw regards the effectiveness of the MOEA used in this paper. A multiobjective approach allows to optimize simultaneously different objectives that are in conflict with each other, without the need of combining them into a single fitness function.

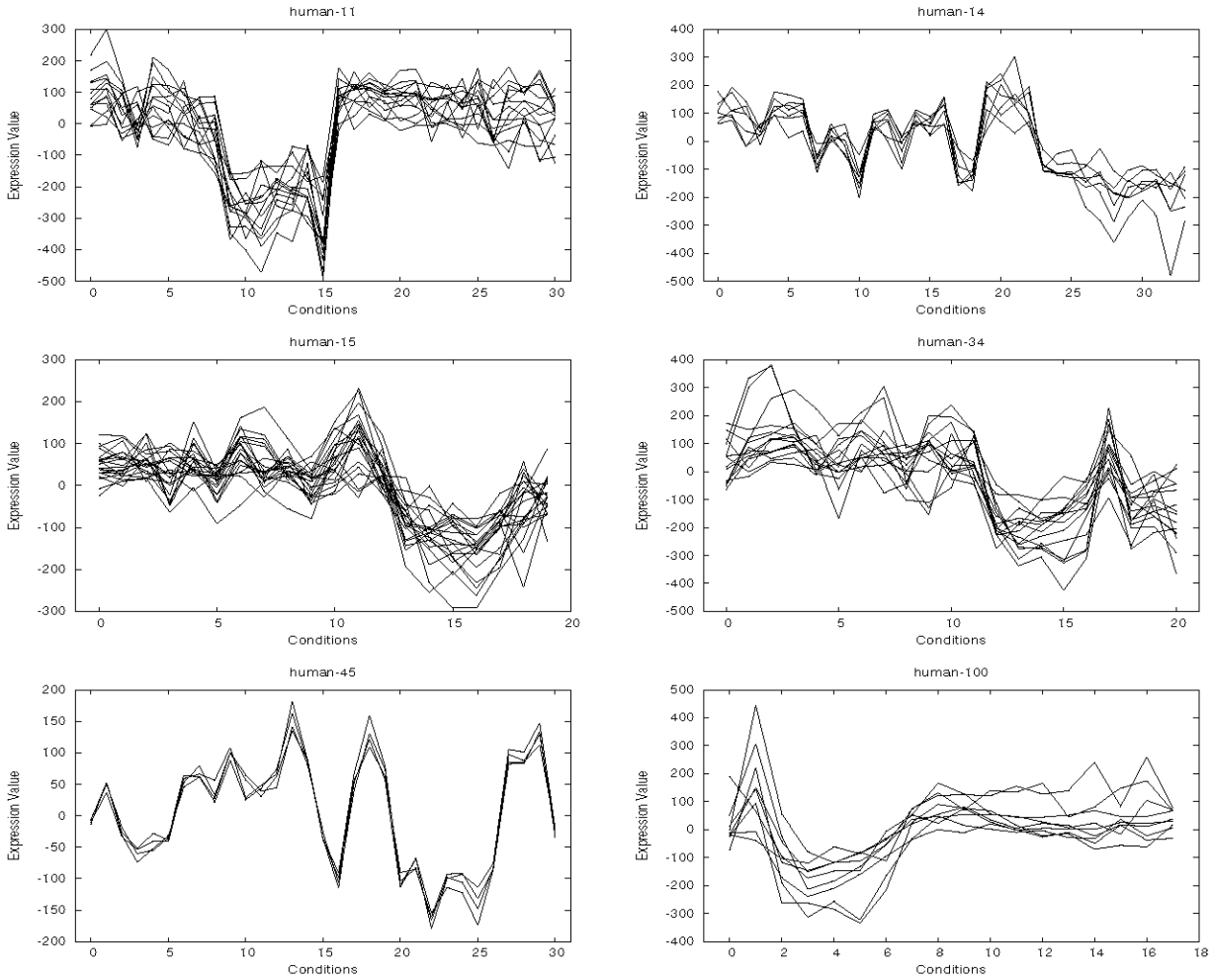


Fig. 4. Six biclusters found on the Human dataset.

ACKNOWLEDGMENTS

The research was partially supported by the Spanish Research Agency CICYT under grant TIN2004–00159 and Junta de Andalucía (III Research Plan).

REFERENCES

[1] P. Baldi, *DNA Microarrays and Gene Expression : From Experiments to Data Analysis and Modeling*. Cambridge University Press, 2002.

[2] C. Tilstone, "Dna microarrays: Vital statistics," *Nature*, vol. 424, pp. 610–612, 2003.

[3] A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering gene expression patterns," *Journal of Computational Biology*, vol. 6(3-4), pp. 281–297, 1999.

[4] J. A. Hartigan, "Direct clustering of a data matrix," *Journal of the American Statistical Association*, vol. 67, no. 337, pp. 123–129, 1972.

[5] Y. Cheng and G. M. Church, "Biclustering of expression data," in *Proceedings of the 8th International Conference on Itelligent Systems for Molecular Biology (ISMB'00)*, 2000, pp. 93–103.

[6] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: A survey," *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 1, pp. 24–25, 2004.

[7] A. Prelić, S. Bleuler, P. Zimmermann, and et al., "A systematic comparison and evaluation of biclustering methods for gene expression data," *Bioinformatics*, vol. 22, pp. 1122–1129, 2006.

[8] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini, "Discovering local structure in gene expression data: The order–preserving submatrix problem," *Proceedings of the 6th International Conference on Computacional Biology*, pp. 49–57, 2002.

[9] X. Liu and L. Wang, "Computing the maximum similarity bi-clusters of gene expression data," *Bioinformatics*, vol. 23, pp. 50–56, 2007.

[10] A. Tanay, R. Sharan, and R. Shamir, "Discovering statistically significant biclusters in gene expression data," *Bioinformatics*, vol. 18, pp. 136–144, 2002.

[11] H. Wang, W. Wang, J. Yang., and P. S. Yu, "Clustering by pattern similarity in large data sets," in *ACM SIGMOD International Conference on Management of Data*, Madison, WI, 2002, p. 394–405.

[12] S. Bleuler, A. Prelić, and E. Zitzler, "An EA framework for biclustering of gene expression data," in *Congress on Evolutionary Computation (CEC-2004)*. Piscataway, NJ: IEEE, 2004, pp. 166–173.

[13] K. Bryan, P. Cunningham, and N. Bolshakova, "Biclustering of expression data using simulated annealing," in *18th IEEE Symposium on Computer-Based Medical Systems*, Dublin, Ireland, 2005, pp. 383–388.

[14] F. Divina and J. S. Aguilar-Ruiz, "Biclustering of expression data with evolutionary computation," *IEEE Transactions on Knowledge & Data Engineering*, vol. 18, no. 5, pp. 590–602, 2006.

[15] J. S. Aguilar-Ruiz, "Shifting and scaling patterns from gene expression data," *Bioinformatics*, vol. 21, pp. 3840–3845, 2005.

[16] D. Corne, K. Deb, and P. J. Fleming, "The good of the many outweighs the good of the one: evolutionary multi-objective optimization," *IEEE Connections Newsletter*, vol. 1, no. 1, pp. 9–13, February 2003.

[17] R. Cho, M. Campbell, E. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. Wolfsberg, A. Gabrielian, D. Landsman, D. Lockhart, and R. Davis, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Cell*, vol. 2, pp. 65–73, 1998.

[18] A. A. Alizadeh, M. B. Eisen, R. E. Davis, and et al., "Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, p. 503–511, 2000.