BiHEA: A Hybrid Evolutionary Approach for Microarray Biclustering

Cristian Andrés Gallo[1], Jessica Andrea Carballido[1], and Ignacio Ponzoni[1,2]

[1] Laboratorio de Investigación y Desarrollo en Computación Científica (LIDeCC),
Departamento de Ciencias e Ingeniería de la Computación,
Universidad Nacional del Sur, Av. Alem 1253, 8000, Bahía Blanca, Argentina
[2] Planta Piloto de Ingeniería Química (PLAPIQUI) - UNS – CONICET
Complejo CRIBABB, Co. La Carrindanga km.7, CC 717, Bahía Blanca, Argentina
{cag,jac,ip}@cs.uns.edu.ar

Abstract. In this paper a new hybrid approach that integrates an evolutionary algorithm with local search for microarray biclustering is presented. The novelty of this proposal is constituted by the incorporation of two mechanisms: the first one avoids loss of good solutions through generations and overcomes the high degree of overlap in the final population; and the other one preserves an adequate level of genotypic diversity. The performance of the memetic strategy was compared with the results of several salient biclustering algorithms over synthetic data with different overlap degrees and noise levels. In this regard, our proposal achieves results that outperform the ones obtained by the referential methods. Finally, a study on real data was performed in order to demonstrate the biological relevance of the results of our approach.

Keywords: gene expression data, biclustering, evolutionary algorithms.

1 Introduction

The task of grouping genes that present a related behavior constitutes a growing investigation area into the research field of gene expression data analysis. The classification is performed according to the genes' expression levels in the Gene Expression Data Matrix (GEDM). The success in this task helps in inferring the biological role of genes. The study of these complex interactions constitutes a challenging research field since it has a great impact in various critical areas. In this context, the microarray technology arose as a fundamental tool to provide information about the behavior of thousands of genes. The information provided by a microarray experiment corresponds to the relative abundance of the mRNA of genes under a given condition. The abundance of the mRNA is a metric that can be associated to the expression level of the gene. This information can be arranged into a matrix, namely GEDM, where rows and columns correspond to genes and experiments respectively.

In most cases, during the process of detecting gene clusters, all of the genes are not relevant for all the experimental conditions, but groups of them are often co-regulated and co-expressed only under some specific conditions. This observation has led the attention to the design of biclustering methods that simultaneously group genes and

K.S. Guimarães, A. Panchenko, T.M. Przytycka (Eds.): BSB 2009, LNBI 5676, pp. 36–47, 2009.

samples [1]. In this regard, a suitable bicluster consists in a group of rows and columns of the GEDM that satisfies some similarity score [2] in union with other criteria.

In this context, a new multi-objective evolutionary approach for microarray biclustering is presented, which mixes an aggregative evolutionary algorithm with features that enhance its natural capabilities. To the best of our knowledge, this methodology introduces two novel features that were never addressed, or partially dealt-with, by other evolutionary techniques designed for this problem instance. The first contribution consists in the design of a recovery process that extracts the best solutions through the generations. The other new characteristic is the incorporation of an elitism procedure that controls the diversity in the genotypic space. The paper is organized as follows: in the next section some concepts about microarray biclustering are defined; then, a brief review on relevant existing methods used to tackle this problem is presented; in Section 4 our proposal is introduced; then, in Section 5, the experiments and the results are put forward; finally some conclusions are discussed.

2 Microarray Biclustering

As abovementioned, expression data can be viewed as a matrix **E** that contains expression values, where rows correspond to genes and columns to the samples taken at different experiments. A matrix element e_{ij} contains the measured expression value for the corresponding gene *i* and sample *j*. In this context, a bicluster is defined as a pair (*G*, *C*) where $G \subseteq \{1,\ldots, \mathrm{m}\}$ is a subset of genes (rows) and $C \subseteq \{1,\ldots, \mathrm{n}\}$ is a subset of conditions [2]. In general, the main goal is to find the largest bicluster that does not exceed certain homogeneity constrain. It is also important to consider that the variance of each row in the bicluster should be relatively high, in order to capture genes exhibiting fluctuating coherent trends under some set of conditions. The size *g*(*G,C*) is the number of cells in the bicluster. The homogeneity *h*(*G,C*) is given by the mean squared residue score, while the variance *k*(*G,C*) is the row variance [2]. Therefore, our optimization problem can be defined as follows:
maximize

$$g(G,C) = |G||C| \; . \tag{1}$$

$$k(G,C) = \frac{\sum_{g\in G,c\in C}\left(e_{gc} - e_{gC}\right)^2}{|G|\cdot|C|} \; . \tag{2}$$

subject to

$$h(G,C) \le \delta \; . \tag{3}$$

with $(G,C)\in X$, $X = 2^{\{1,\ldots,m\}} \times 2^{\{1,\ldots,n\}}$ being the set of all biclusters, where

$$h(G,C) = \frac{1}{|G|\cdot|C|}\sum_{g\in G,c\in C}\left(e_{gc} - e_{gC} - e_{Gc} + e_{GC}\right)^2 \; . \tag{4}$$

is the mean squared residue score,

$$e_{gC} = \frac{1}{|C|}\sum_{c\in C} e_{gc}, \quad e_{Gc} = \frac{1}{|G|}\sum_{g\in G} e_{gc} \; . \tag{5,6}$$

are the mean column and row expression values of (*G*,*C*) and

$$e_{GC} = \frac{1}{|G| \cdot |C|} \sum_{g \in G, c \in C} e_{gc} \ . \qquad (7)$$

is the mean expression value over all the cells that are contained in the bicluster (*G*,*C*). The user-defined threshold δ>0 represents the maximum allowable dissimilarity within the cells of a bicluster. In other words, the residue quantifies the difference between the actual value of an element e_{gc} and its expected value as predicted for the corresponding row mean, column mean, and bicluster mean. A bicluster with a mean square residue lower than a given value δ is called a δ-bicluster. The problem of finding the largest square δ-bicluster is NP-hard [2] and, in particular, Evolutionary Algorithms (EAs) are well-suited for dealing these problems [3, 4, 5].

3 GEDM: Main Biclustering Methods

Cheng and Church's Approach (CC): Cheng and Church [2] were the first to apply the concept of biclustering on gene expression data. Given a data matrix **E** and a maximum acceptable mean squared residue score (*h*(*G*,*C*)), the goal is to find subsets of rows and subsets of columns with a score no larger than δ. In order to achieve this goal, Cheng and Church proposed several greedy row/column removal/ addition algorithms that are then combined in an overall approach. The multiple node deletion method removes all rows and columns with row/column residue superior to $\delta . \alpha$ in every iteration, where α is a parameter introduced for the local search procedure. The single node deletion method iteratively removes the row or column that grants the maximum decrease of *h*(*G*,*C*). Finally, the node addition method adds rows and columns that do not increase the actual score of the bicluster. In order to find a given number of biclusters, the approach is iteratively executed on the remained rows and columns that are not present in the previous obtained biclusters.

Iterative Signature Algorithm (ISA): The most important conceptual novelty of this approach [6] is the focus on the desired property of the individual co-regulated bicluster that is going to be extracted from the expression data matrix. According to the definition of the authors, such a transcription bicluster consists of all genes that are similar when compared over the conditions, and all conditions that are similar when compared over the genes. This property is referred as self-consistency. In this regard, they proposed to identify modules by iteratively refining random input gene sets, using the signature algorithm previously introduced by the same authors. Thus, self-consistent transcription modules emerge as fixed-points of this algorithm.

BiMax: The main idea behind the Bimax algorithm [7] consists in the use of a divide and conquer strategy in order to partition **E** into three submatrices, one of which contains only 0-cells and therefore can be ignored in the following. The procedure is then recursively applied to the remaining two submatrices **U** and **V**; the recursion ends if the current matrix represents a bicluster, i.e. contains only 1s. If **U** and **V** do not share any rows and columns of **E**, the two matrices can be processed independently from each other. Yet, if **U** and **V** have a set of rows in common, special care is necessary to only generate those biclusters in **V** that share at least one common column with **CV**. A drawback of this approach is that only works on binary data matrices. Thus, the results strongly depend on the accuracy of the discretization step.

Order Preserving Submatrix Algorithm (OPSM): Ben-Dor et al. [8] defined a bicluster as an order-preserving submatrix (OPSM). According to their definition, a bicluster is a group of rows whose values induce a linear order across a subset of the columns. The work focuses on the relative order of the columns in the bicluster rather than on the uniformity of the actual values in the data matrix. More specifically, they want to identify large OPSMs. A submatrix is order-preserving if there is a permutation of its columns under which the sequence of values in every row is strictly increasing. In this way, Ben-Dor et al. aim at finding a complete model with highest statistically significant support. In the case of expression data, such a submatrix is determined by a subset of genes and a subset of conditions, such that, within the set of conditions, the expression levels of all genes have the same linear ordering. As such, Ben-Dor et al. addressed the identification and statistical assessment of co-expressed patterns for large sets of genes, and considered that, generally, data contains more than one such pattern.

Evolutionary Approaches: The first reported approach that tackled microarray biclustering by means of an EA was proposed by Bleuler *et al.* [5]. In this work, the use of a single-objective EA, an EA combined with a LS strategy [2] and the LS strategy alone [2] are analyzed. In the case of the EA, one novelty consists in a form of diversity maintenance on the phenotype space that can be applied during the selection procedure. For the case of the EA hybridized with a LS strategy, whether the new individual yielded by the LS procedure should replace the original individual or not is considered. As regards the LS as a stand alone strategy, they propose a new non-deterministic version, where the decision on the course of execution is made according to some probability. In the work of Mitra and Banka [3], the first approach that implements a Multi-Objective EA (MOEA) based on Pareto dominancy is presented. The authors base their work on the NSGA-II, and look for biclusters with maximum size and homogeneity. A LS strategy is applied to all of the individuals at the beginning of every generational loop. Finally, Gallo *et al.* [9] presents the SPEA2^{LS}, another MOEA combined with a LS [2] strategy. In this case, the authors base the algorithm on the SPEA2 [10], and seek biclusters with maximum rows, columns, homogeneity and row variance. A novel individual representation to consider the inverted rows of the data matrix is introduced. Also, a mechanism for re-orienting the search in terms of row variance and size is provided. The LS strategy is applied to all of the individuals in the resultant population of each generation.

4 BiHEA: Biclustering via a Hybrid Evolutionary Algorithm

The aim of our study is to use an evolutionary process to generate near optimal biclusters with coherent values following an additive model, according to the classification given by [1]. Thus, the EA is used to globally explore the search space *X*. However, it was observed that, in the absence of local search, stand-alone single-objective or MOEAs could not generate satisfactory solutions [3, 5, 9]. In that context, a LS technique based on Chung and Church's procedure is applied after each generation, thus orienting the exploration and speeding up the convergence of the EA by refining the chromosomes. Furthermore, two additional mechanisms were incorporated in the evolutionary process in order to avoid the loss of good solutions: an elitism procedure that maintains the best biclusters as well as the diversity in the genotypic space through the generations, and a recovery process that extracts the best solutions of each generation and then copies these individuals into an archive. This archive is actually the set of biclusters returned by the algorithm. Although these two mechanisms appear to be similar to each other, there are several differences between them. The elitism procedure selects the *b* best biclusters that do not overlap in a certain threshold, passing them to the next generation. These solutions can be part of the selection process of further generations allowing production of new solutions based on these by means of the recombination operator. However, due to imperfections on the selection process and of the fitness function, some good solutions can be misplaced through generations. To deal with this issue, we have incorporated an archive, which keeps the best generated biclusters through the entire evolutionary process. It is important to remark that this "meta" population is not part of the selection process, i.e., the evolution of the population after each generation is monitored by the recovery process without interfering in the evolutionary process.

Main Algorithm

As aforementioned, the main loop is a basic evolutionary process that incorporates the LS, the elitism and the recovery procedure. Algorithm 1 illustrates these steps.

Algorithm 1 (Main loop)

```
Input:   pop_size    (population size)
         max_gen     (max number of generations)
         mut_prob    (probability of mutation)
         $\delta$           (threshold for homogeneity)
         $\alpha$           (parameter for the local search)
         $\theta$           (overlap degree of the recovery process)
         GEDM        (gene expression data matrix)
Output:  arch        (a set of biclusters)

Step 1:  Initialization. Load the data matrix GEDM. Generate a random population $P_0$ of size pop_size.
         Generate an empty population arch.
Step 2:  Main loop. If max_gen is reached, go to Step 9.
Step 3:  Selection. Perform binary tournament selection over $P_t$ to fill the pool of parents $Q_t$ of size
         pop_size.
Step 4:  Elitism procedure. Select at most the best pop_size/2 individuals of $P_t$ that do not overlap each
         other in at most the 50% of cells. Copy the individuals to $P_{t+1}$.
Step 5:  Offspring. Generate the remained (at least pop_size-pop_size/2) individuals of $P_{t+1}$ applying
         recombination over two random parents of $Q_t$. Apply uniform mutation to those individuals.
Step 6:  Local Search. Apply the local search optimization to the individuals of $P_{t+1}$ with mean squared
         residue above $\delta$.
Step 7:  Recovery procedure. For each individual $I \in P_{t+1}$ with mean squared residue bellow $\delta$, try to
         add I to arch in the following way: find the individual $J \in$ arch who shares at least the $\theta$% of
         cells and then replace J with I only if I is larger than J. If no J where found, add I to arch in an
         empty slot only if the size of arch is bellow to pop_size. Otherwise discard I.
Step 8:  End of the loop. Go to Step 2.
Step 9:  Result. Return arch.
```

At this point, the differences between the elitism and the recovery procedure should be clear. The threshold for the overlap level in the elitism procedure, as well as the proportion of elitism, was empirically determined after several runs of the algorithms over different datasets. It is important to note that, as a consequence of a careful design of the recovery procedure, and by means of choosing an adequate value for the θ parameter, the resulting set of biclusters is slightly overlapped in comparison to the high overlapping degree present in the other EAs for biclustering [3, 5, 9].

Individual's Representation

Each individual represents one bicluster, which is encoded by a fixed size binary string built by appending a bit string for genes with another one for conditions. The individual constitutes a solution for the problem of optimal bicluster generation. If a string position is set to 1 the relative row or column belongs to the encoded bicluster, otherwise it does not. Figure 1 shows an example of such encoding for a random individual.

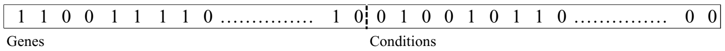

Fig. 1. An encoded individual representing a bicluster

Genetic Operators

After some preliminary tests we decided to apply independent bit mutation to both strings with mutation rates that allow the expected number of bits to be flipped to be the same for both strings. A two point crossover is preferred to one point crossover because the latter would prohibit certain combinations of bits to be crossed over together, especially in cases where the differences in size of rows and columns are notable. In this context, one random point is selected on the rows and the other random point is select over the columns, thus performing the recombination over both search spaces. Then, when both children are obtained combining each one of the two parents' parts, the individual selected to be the descendant is the best in terms of the fitness function.

Fitness Function

As regards the objectives to be optimized, we observed that it was necessary to generate maximal sets of genes and conditions while maintaining the "homogeneity" of the bicluster with a relatively high row variance, as it was established in the equations 1-3. These bicluster features, conflicting to each other, are well suited for multi-objective modeling. An aggregative fitness function that incorporates these features is presented in equation 8. In view of the fact that the local search procedure guarantees the residue constraint [2], the main reason for having a special consideration of the individuals with residue above δ is in the first generation, where the individuals in the population are randomly created. In this context, only those solutions with small residue are preferred. It is also important to consider that an individual can violate the residue constraint during the creation of offspring solutions. Therefore, as the crossover operator returns the best of both children's, individuals with small residue are again preferred in this case, as it can be seen in the fitness function formulation (eq. 8).

$$
fitness\ (G,C) = \begin{cases} h(G,C) & if \quad h(G,C) > \delta \\ 1 - \frac{|G||C|}{mn} + \frac{h(G,C)}{\delta} + \frac{1}{k(G,C)} & if \quad h(G,C) <= \delta \wedge k(G,C) > 1 \\ 1 - \frac{|G||C|}{mn} + \frac{h(G,C)}{\delta} + 1 & if \quad otherwise \end{cases} \tag{8}
$$

However, when the individuals meet the homogeneity constraint, the LS is not applied. Thus, the improvement of the solutions only depends on the evolutionary process, and then the consideration of biclusters' features such as size, mean squared residue and variance become important. The practical advantage on the consideration of the variance of a bicluster is to avoid constant biclusters [1], since they can be trivially obtained [2]. Note that the fitness function is minimized.

Local Search

The LS procedure that hybridizes the EA was already described. As aforementioned, the greedy approach is based on Chung and Church's work [2], with a small change that avoids the consideration of inverted rows, as applied in [5]. The algorithm starts from a given bicluster (G, C). The genes or conditions having mean squared residue above (or below) a certain threshold are selectively eliminated (or added) according to the description given in the previous sections.

5 Experimental Framework and Results

Two different goals were established for our study. First we need to analyze the quality of the results of BiHEA in the extraction of biclusters with coherent values that follow an additive model. For this analysis, the new approach was tested over synthetic data matrices with different degrees of overlap and noise and then, the results were compared with several of the most important methods for biclustering. Although performing over synthetic data can give an accurate view of the quality of the method, since the optimal biclusters are known beforehand, any artificial scenario inevitably is biased regarding the underlying model and only reflects certain aspects of biological reality. To this end, and in a second experimental phase, we will analyze the biological relevance of the results of BiHEA over a real life data matrix.

Performance Assessment

In order to assess the performance of the biclustering approach over synthetic data, the general bicluster match score is introduced, which is based on the gene match score proposed by [7]. Let M_1, M_2 be two sets of biclusters. The bicluster match score of M_1 with respect to M_2 is given by the equation 9, which reflects the average of the maximum match scores for all biclusters in M_1 with respect to the biclusters in M_2.

$$S^*(M_1,M_2)=\frac{1}{|M_1|}\sum_{(G_1,C_1)\in M_1}\max_{(G_2,C_2)\in M_2}\frac{|(G_1,C_1)\cap(G_2,C_2)|}{|(G_1,C_1)\cup(G_2,C_2)|}. \tag{9}$$

In this case, instead of considering only the genes of the biclusters of each set [7], the conditions will also be taken into account, i.e., the amount of cells of each bicluster will be assessed. Thus, this measure is more accurate than the metric presented in [7]. Now, let M_{opt} be the set of implanted biclusters and M the output of a biclustering method. The average bicluster precision is defined as $S^*(M, M_{opt})$ and reflects to what extent the generated biclusters represent true biclusters. In contrast, the average bicluster coverage, given by $S^*(M_{opt}, M)$, quantifies how well each of the true biclusters is recovered by the biclustering algorithm under consideration. Both scores take the maximum value of 1 if $M_{opt}= M$.

As regard the real data, since the optimal biclusters are unknown, the above metric can not be applied. However, prior biological knowledge in the form of natural language descriptions of functions and processes to which the genes are related has become widely available. Similar to the idea pursued in [7, 11, 12], whether the groups of genes delivered by BiHEA show significant enrichment with respect to a specific Gene Ontology (GO) annotation will be investigated. Then, a novel measure was designed in order to assess the molecular function and biological process enrichment of the results of a biclustering method. Let *M* be a set of biclusters, *GO* a specific GO annotation and α a statistic significant level. The overall enrichment indicator of *M* with respect of *GO* on a statistically significant level of α is given by:

$$E^{*}(M,GO,\alpha)=\frac{1}{|M|}\sum_{(G,C)\in M}\frac{Maxenrichment(G,GO,\alpha)}{|G|}\sum_{(G,C)\in M}\frac{|G|}{n}. \qquad (10)$$

where *Maxenrichment*(*G, GO, α*) is the maximum gene amount of *G* with a common molecular function/biological process under *GO* with a statistically significant α level. The metric of equation 10 measures the average of the maximum gene proportion statistically significant of molecular function/biological process enrichment of a set of biclusters *M* on a specific GO annotation, pondered with the average genes of *M*. It is an indicator of the quality of the results of a clustering/biclustering method on real data, and can be used to compare several methods, being the highest values the best.

First Experimental Phase: Synthetic Data

Data preparation

The artificial model used to generate synthetic gene expression data is similar to the approaches proposed by [7, 13]. In this regard, the biclusters represent transcription modules, where these modules are defined by a set G of genes regulated by a set of common transcription factors and a set C of conditions in which these transcription factors are active. Varying the amount of genes and conditions that two modules have in common, it is possible to vary the overlap degree in the implanted biclusters. To this end, we define the overlap degree *d*, as an *indicator* of the maximum amount of cells that two transcription modules can share. The amount of shared cells is actually d^2.

This model enables the investigation of the capability of a method to recover known groupings, while at the same time, further aspects like noise and regulatory complexity can be systematically studied [7]. The datasets are kept small, n = 100 and m = 100. This, however, does not restrict the generality of the results. In the case of *d* = 0, 10 non-overlapped biclusters (size = 10 rows *times* 10 columns) were implanted. For every *d*>0, the size of the artificial biclusters was increased in *d* rows and *d* columns, except for the rightmost bicluster, for which its size remains unchanged. For *d*>1, 18 additional biclusters appear in the data matrices, as a consequence of the overlap of the implanted transcription modules. These extra biclusters are also included in our study since they are equally suitable for being extracted by a biclustering method, although the overlap degree is higher than for the artificial transcription factors. The figure 2 depicts the previous scenario. The values of each bicluster were determined as follows: for the first row, random real numbers between 0 and 300 from a uniform distribution were incorporated. Then, for each one of the remainder rows, a unique random value between 0 and 300 is obtained and added to each element of the first row. The result is a bicluster with coherent values under an additive model [1], with a mean squared residue equal to 0 and a row variance greater than 0. The remainder slots in the matrix were filled with random real values between 0 and 600.

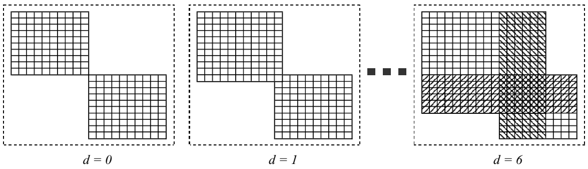

Fig. 2. Overlapping levels of artificial biclusters according to *d*. In *d=6,* the diagonal lines represent the extra biclusters generated by the overlapping of the implanted biclusters.

Synthetic datasets built following the aforementioned procedure are useful to analyze the behavior of a biclustering method in increasing regulatory complexity levels. However, these datasets represent an ideal scenario without noise, i.e., far away from realistic data. To deal with this issue, and in view of the fact that real scenarios have a great regulatory complexity, the behavior of this proposal with d=6 and with increasing noise levels will be also investigated. For the noisy model, the expression values of the biclusters are changed adding a random value from a uniform distribution between $-k$ and k, with k=0, 5, 10, 15, 20 and 25 to each cell.

Results

For referential purposes, several important biclustering algorithms were run: BiMax, CC, OPSM, ISA, and SPEA2^{LS}. For the first four implementations, the BicAT [14] tool was used. All the parameters for these methods were set after several runs, in order to obtain the best results of each strategy. For BiHEA, the parameters' setting is the following: population = 200; generations = 100; δ = 300; α = 1.2; mutation probability = 0.3; and θ = 70. Since the number of generated biclusters strongly varies among the considered methods, a filtering procedure, similar to the recovery process of our approach, has been applied to the output of the algorithms to provide a common basis for a fair comparison. The filtering procedure extracts, for each of the resulting set of biclusters, at most q of the largest biclusters that overlap in at most the θ=70% of cells. For d<2, q is set to 10, and for the rest, q is set to 28. As regards the results, in figures 3a and 3b, the average precision and the average coverage obtained by the different biclustering methods are shown, for the scenarios with increasing overlapping degrees. Similarly, in figures 3c and 3d, the results for the scenarios with increasing noise levels are illustrated.

As it can be observed, BiHEA outperforms the referential methods in all the scenarios, in terms of both the precision and the coverage of biclusters. As the overlapping degree increases, figures 3a and 3b show that the results obtained by our method improve, reaching an almost perfect score with d=6. This can be explained in terms of the theory of basic schemes in genetic algorithms, since in higher degrees of overlap, useful schemes shared between the optimal biclusters are larger in size. This feature facilitates the construction of solutions that meet the homogeneity constraint by means of the crossover operator. The last observation should be true for most EAs. Nonetheless, the imperfections on the selection process and fitness functions can derive on a misuse of this advantage, as it happens with SPEA2^{LS}. This clearly shows the need of the recovery process introduced on the BiHEA.

In regard to the effects of noise, the results are the expected ones. As the levels of noise augment, the degradation of a perfect bicluster increases the residue and

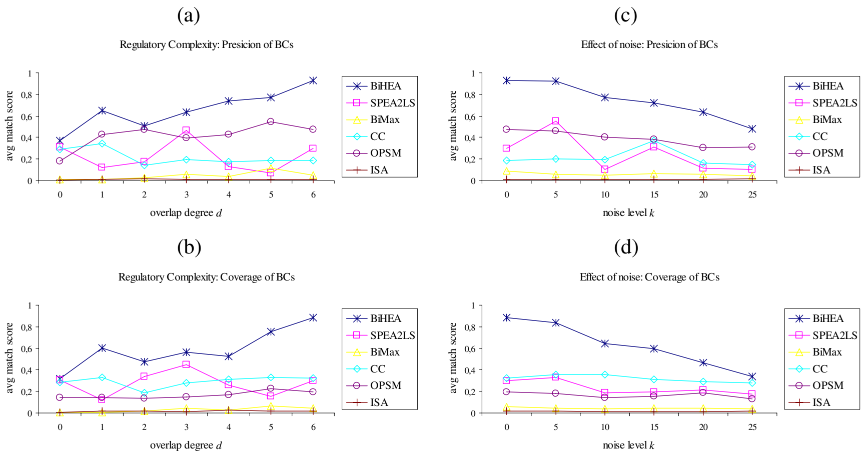

Fig. 3. Results for the artificial scenarios. Figures 3a and 3b show the average precision and the average coverage respectively in overlapped scenarios. Figures 3c and 3d show the average precision and the average coverage respectively in noisy scenarios.

possibly, the homogeneity constraint can no longer be satisfied for the entire bicluster. For the reference methods, OPSM, CC and SPEA2^{LS} show similar results, OPSM being the more precise one although the coverage appears to be worse than the results achieved by CC and SPEA2^{LS}. However, these methods appear to be less susceptible to the noise than BiHEA. On the other hand, both BiMax and ISA can not obtain significant biclusters, which contrasts with the conclusions published in [7] where both methods achieve almost perfect scores. Nevertheless, we argue that this may be a consequence of the way in which synthetic data are constructed, since in the case of BiMax, the discretization method is unable to obtain an appropriate binary representation of the synthetic data matrices. On the other hand, the notion of similarity of rows and columns in the ISA algorithm might be different from the one used here. However, the synthetic data used in this work was designed in the aforementioned manner since, according to our knowledge it represents general and relevant GEDMs, which allow a fair comparison in the evaluation of the algorithms.

Second Experimental Phase: Real Data

In this subsection, the results of BiHEA on a real GEDM will be briefly analyzed. This study will be focused on a colon cancer data [15] that consists in a GEDM of 62 colon tissue samples, 22 of which are normal and 40 are tumor tissues. This analysis will be focused on the 2000 genes with the highest minimal intensity [15]. For the experimentation, an ontological analysis of the 10 first resulting biclusters found by BiHEA, CC, ISA, OPSM and SPEA2^{LS} will be performed. The BiMax algorithm is not included since an adequate parameter setup for the discretization step could not be found. The parameters for the proposed approach remain almost the same, except for the following: $\delta = 150$; $\alpha = 2.0$. All the ontological classification was performed with the ontology tool Onto-Express [12], applying a hyper geometric distribution and referencing the calculations by the 2000 genes analyzed.

As regard the results, the figure 4 depicts the values achieved by the previous methods in terms of the overall enrichment indicator (cf. eq. 10 with $\alpha = 0.05$) for molecular function and biological process enrichment. It is clear that BiHEA is the method that obtains the better results, since the quality of the outcomes outperforms the results of the referential algorithms in terms of the overall enrichment indicator. Only OPSM remains close, whereas the other approaches obtain significantly worse results. These results are consistent with the ones obtained on the synthetic datasets, thus showing the correctness of the artificial model selected.

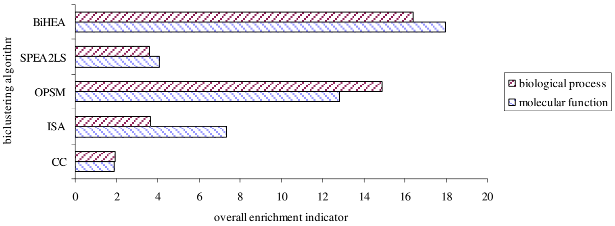

Fig. 4. Overall enrichment indicator of BiHEA, SPEA2^{LS}, OPSM, ISA and CC for molecular function and biological process enrichment

6 Conclusions

In this paper, we have introduced a new memetic evolutionary approach for microarray biclustering. The original EA was hybridized with a LS procedure for finer tuning, and also two novel features were introduced: the first one was designed in order to avoid the loss of good solutions through generations, while keeping a low degree of overlap between the final biclusters, and the other one was conceived so as to maintain a satisfactory level of diversity in the genotypic space.

In a first experimental phase on synthetic datasets, the results obtained with our method outperform the outcomes of several biclustering approaches of the literature, especially in the case of coherent biclusters with high overlap degrees. Nonetheless, this can not be considered as a drawback because, in general, the regulatory complexity of an organism is far from the model of non-overlapped biclusters. Furthermore, an analysis on a real dataset was performed and, in terms of the proposed measure, the quality of the outcomes of BiHEA is clearly better than the results of the reference methods. In fact, this shows the correctness of the model designed to build the biclusters, i.e., coherent biclusters following an additive model. Although this is consistent with the results obtained in the synthetic datasets, an extensive analysis on several real datasets in needed to confirm these results.

Finally, the framework for the comparison of biclustering algorithms was refined by means of the introduction of two new measures: the bicluster match score S^* and the overall enrichment indicator E^*. The first one is useful to test on synthetic data since the optimal biclusters are known beforehand. The last one can be used to assess the performance of several methods in real data in terms of a specific GO annotation. Both metrics are indispensable in any quality assessment of biclustering algorithms since they provide a fair framework in which the methods can be compared.

Acknowledgements. Authors acknowledge CONICET, the ANPCyT for Grant N°11-1652 and SeCyT (UNS) for Grant PGI 24/ZN15.

References

1. Madeira, S., Oliveira, A.L.: Biclustering Algorithms for Biological Data Analysis: A Survey. IEEE-ACM Trans. Comput. Biol. Bioinform. 1, 24–45 (2004)
2. Cheng, Y., Church, G.M.: Biclustering of Expression Data. In: Proceedings of the 8th Inter-national Conf. on Intelligent Systems for Molecular Biology, pp. 93–103 (2000)
3. Mitra, S., Banka, H.: Multi-objective evolutionary biclustering of gene expression data. Pattern Recognit. 39, 2464–2477 (2006)
4. Divina, F., Aguilar-Ruiz, J.S.: Biclustering of Expression Data with Evolutionary Computation. IEEE Trans. Knowl. Data Eng. 18(5), 590–602 (2006)
5. Bleuler, S., Prelic, A., Zitzler, E.: An EA framework for biclustering of gene expression data. In: Proceeding of Congress on Evolutionary Computation, pp. 166–173 (2004)
6. Ihmels, J., Bergmann, S., Barkai, N.: Defining transcription modules using large-scale gene expression data. Bioinformatics 20(13), 1993–2003 (2004)
7. Zimmermann, P., Wille, A., Buhlmann, P., Gruissem, W., Hennig, L., Thiele, L., Zitzler, E., Prelic, A., Bleuler, S.: A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics 22(9), 1122–1129 (2006)
8. Ben-Dor, A., Chor, B., Karp, R., Yakhini, Z.: Discovering Local Structure in Gene Expression Data: The Order-Preserving Submatrix Problem. In: Proc. Sixth Int'l Conf. Computational Biology (RECOMB 2002), pp. 49–57 (2002)
9. Gallo, C., Carballido, J.A., Ponzoni, I.: Microarray Biclustering: A Novel Memetic Approach Based on the PISA Platform. LNCS, vol. 5483, pp. 44–55. Springer, Heidelberg (2009)
10. Zitzler, E., Laumanns, M., Thiele, L.: SPEA2: Improving the strength pareto evolutionary algorithm for multiobjective optimization. In: Giannakoglou, Tsahalis, Periaux, Papailiou, Fogarty (eds.) Evolutionary Methods for Design, Optimisations and Control, pp. 19–26 (2002)
11. Tanay, A., et al.: Discovering statistically significant biclusters in gene expression data. Bioinformatics 18(suppl. 1), S136–S144 (2002)
12. Draghici, S., Khatri, P., Bhavsar, P., Shah, A., Krawetz, S., Tainsky, M.: Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design, and Onto-Translate. Nuc. Acids Res. 31(13), 3775–3781 (2003)
13. Ihmels, J., et al.: Revealing modular organization in the yeast transcriptional network. Nat. Genet. 31, 370–377 (2002)
14. Barkow, S., Bleuler, S., Prelic, A., Zimmermann, P., Zitzler, E.: BicAT: a biclustering analysis toolbox. Bioinformatics 22(10), 1282–1283 (2006)
15. Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., Levine, A.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. 96, 6745–6750 (1999)