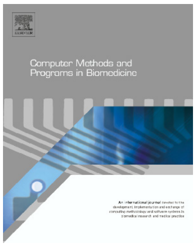


journal homepage: www.intl.elsevierhealth.com/journals/cmpb

A new measure for gene expression biclustering based on non-parametric correlation

Jose L. Flores[a,*], *Iñaki Inza*[a,1], *Pedro Larrañaga*[b,2], *Borja Calvo*[a,1]

[a] *Intelligent Systems Group, Department of Computer Sciences and Artificial Intelligence, University of the Basque Country, P.O. Box 649, 20080 Donostia – San Sebastian, Spain*

[b] *Computational Intelligence Group, Department of Artificial Intelligence, Technical University of Madrid, Campus de Montegancedo, s/n, Boadilla del Monte, 28660 Madrid, Spain*

ARTICLE INFO

Article history:
Received 16 September 2012
Received in revised form
14 June 2013
Accepted 26 July 2013

Keywords:
Biclustering
Biomedicine
Artificial intelligence
Machine learning

ABSTRACT

Background: One of the emerging techniques for performing the analysis of the DNA microarray data known as biclustering is the search of subsets of genes and conditions which are coherently expressed. These subgroups provide clues about the main biological processes. Until now, different approaches to this problem have been proposed. Most of them use the mean squared residue as quality measure but relevant and interesting patterns can not be detected such as shifting, or scaling patterns. Furthermore, recent papers show that there exist new coherence patterns involved in different kinds of cancer and tumors such as inverse relationships between genes which can not be captured.

Results: The proposed measure is called Spearman's biclustering measure (SBM) which performs an estimation of the quality of a bicluster based on the non-linear correlation among genes and conditions simultaneously. The search of biclusters is performed by using a evolutionary technique called estimation of distribution algorithms which uses the SBM measure as fitness function. This approach has been examined from different points of view by using artificial and real microarrays. The assessment process has involved the use of quality indexes, a set of bicluster patterns of reference including new patterns and a set of statistical tests. It has been also examined the performance using real microarrays and comparing to different algorithmic approaches such as Bimax, CC, OPSM, Plaid and xMotifs.

Conclusions: SBM shows several advantages such as the ability to recognize more complex coherence patterns such as shifting, scaling and inversion and the capability to selectively marginalize genes and conditions depending on the statistical significance.

1. Introduction

The quantitative and qualitative analysis of the DNA has become one of the most important areas in the biomedical research. To accomplish this analysis, the microarray technology is used [1]. This technology allows to measure simultaneously the level of expression of thousands of genes under a set of conditions. The measurement of these levels of expression leads to the storing of huge volumes of data which are organized in matrices. The analysis of these matrices is a key issue to better understand the biological interactions

[*] *Corresponding author*. Tel.: +34 943015245; fax: +34 943015590.
E-mail addresses: joseluis.flores@ehu.es (J.L. Flores), ccpincai@si.ehu.es (I. Inza), pedro.larranaga@fi.upm.es (P. Larrañaga), borja.calvo@ehu.es (B. Calvo).

[1] Tel.: +34 943015245; fax: +34 943015590.

[2] Tel.: +34 913367399; fax: +34 913367412.

0169-2607/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cmpb.2013.07.025

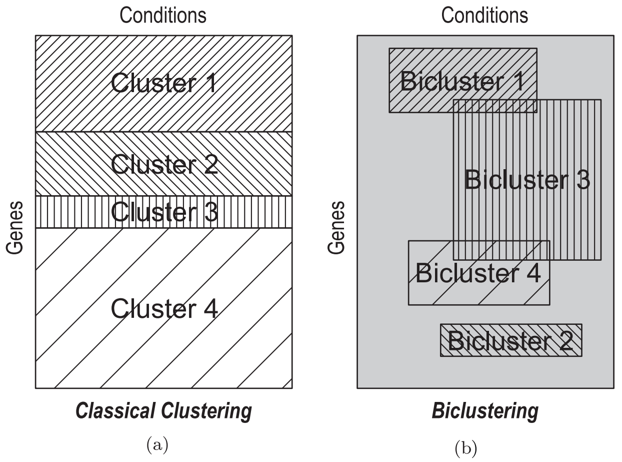


Fig. 1 – Clustering approaches: (a) classical clustering on genes and (b) biclustering on genes and conditions.

between genes. The complexity of this analysis arises from the huge quantity of variables, the noise which can affect these measures, and the low number of conditions.

Clustering is one of the techniques that has been broadly applied to analyze DNA microarrays [2,88,92]. Clustering allows to group genes that are expressed in a coherent way in different sets (see Fig. 1a).

However, the application of clustering shows that the methods have some limitations in this setting. Many expression level patterns are common to a subgroup of genes in a specific subset of conditions. In fact, the general understanding of biological processes indicates subgroups of genes to be co-expressed only under specific subsets of conditions, but they behave in a different way under other conditions (see Fig. 1b). The discovering of co-expressed subgroups of genes for some of the conditions may be the key to understand the biological interactions that are not apparent otherwise. This discovery has motivated the development of specific clustering algorithms known as biclustering algorithms.

The biclustering problem is a NP-hard problem that was originally considered by [3], subsequently by [4,5,91]. The complexity of this problem in the context of DNA microarray data analysis has motivated the creation of different algorithmic approaches.

Cheng and Church (CC) algorithm [6] is a greedy iterative search method based on a quality measure called Mean Squared Residue. The algorithm uses this quality measure in order to add or to remove rows or columns iteratively, thus, improving the quality. The MSR measures how adequate each expression value is with regard to the rest of values of the bicluster (see [6] for more details). Subsequently, this method was improved with the FLOC algorithm which incorporates the capability to deal with missing values and to obtain several biclusters simultaneously following a hierarchical clustering strategy in each dimension and combining the results. On the other hand, SAMBA takes a different approach performing an exhaustive biclusters enumeration by means of bipartite graph based model and subsequent greedy approach adding or removing nodes in order to find the maximum weight subgraphs. Spectral biclustering uses linear algebra in order to identify biclusters patterns. The Plaid Model takes a different approach which represents the input microarray matrix as a sum of layers where each layer represents a bicluster. The BiMax algorithm discretizes the microarray data before applying a recursive process until to obtain a matrix with a single value. On the other hand there is a group of biclusters algorithms based on metaheuristics such as simulated annealing, particle swarm optimization, memetic algorithms, etc. All these approaches use the MSR as a bicluster quality evaluation function in order to detect co-expressed behaviour patterns.

In this sense, the MSR is nowadays the most commonly measure used to perform the search of biclusters. The joint use of this measure with evolutionary approaches to perform the search in the candidate biclusters space has provided successful results correctly identifying the majority of coherence patterns [7–13].

However, this measure presents several problems. First, it does not evaluate correctly some coherence evolution patterns such as scaling patterns [14]. Furthermore, as we demonstrate in this paper there are other patterns such as inversion patterns which show some difficulties in being correctly identified by MSR. Secondly, it is not possible to determinate whether the collection of genes and conditions represents a reliable grouping in statistical terms.

In order to overcome these problems, we propose an alternative measure called Spearman's biclustering measure (SBM). It is based on the Spearman's non-parametric correlation coefficient as an evaluation function of fitness of candidate biclusters. Although several correlation-based measures have been proposed in the literature, all of them have been oriented to detect only direct relationships between genes. The proposed measure is able to detect direct and inverse relationships between genes and conditions. The inclusion of inverse relationships is based on recent papers related with the discovering of inverse relationships in different types of cancer and tumors [15–17]. It has been also considered the detection of relationships between conditions to increase the capability of detection of patterns. Metrics proposed in the literature the measures only consider the correlation between genes. The proposed measure is able to identify complex patterns that can not be identified by benchmarking MSR. Furthermore, based on its statistical properties, it is possible to determine the reliability of a collection of genes and conditions in statistical terms. The absence of this property in MSR measure does not allow to qualify the biclusters in statistical terms. The problem arises in the evaluation of the quality of the results provided of the measure with regard to classical MSR measure. It is also necessary to take into account that the evaluation of SBM has to be performed analyzing its capabilities with the actual taxonomy of coherence patterns [18] with regard to MSR as the biclustering reference measure. In order to compare both measures fairly, it would be desirable to use a methodology capable of identifying the weaknesses and strengths of any measure on every type of coherence pattern.

In order to answer these questions, we propose an additional set of contributions. Firstly, a set of global quality indexes is proposed to evaluate both the global relative performance and the global quality of the measure to discover genes and conditions. Secondly, a set of local quality indexes is proposed in order to individually assess the local quality of the measure with regard to the selected genes and conditions. Next, a fair comparison procedure based on statistical tests is proposed to evaluate the results using previous quality indexes. Finally, the search of biclusters based on different measures is performed using the same search algorithm: estimation of distribution algorithms (EDAs). A recent alternative to the use of genetic algorithms which have provided successful results in different fields [19–21].

Once this analysis has been performed, a biological analysis is performed from different points of view. Firstly, the algorithm is assessed individually with real microarrays following a quality process which allows to filter low quality genes and/or conditions in the bicluster before proceeding with the biological assessment. Finally, the algorithm is compared with biclustering methods such as BiMax, OPSM, PLAID, CC and Xmotifs. Following the methodology in [22] the performance of all algorithms is evaluated biologically with the percentage of biclusters enriched by any Gene Ontology Consortium (GO) category at different levels of significance.

This article is organized as follows: Section 2 presents the state of the art in biclustering algorithms. Section 2 shows the definitions of the coherence measures. Section 4 describes the search algorithm used for this problem, the estimation of distribution algorithms. Section 5 describes the proposed experimentation framework and finally in Section 6 we present our conclusions.

2. State of the art

In order to properly understand the notation used throughout the paper, a set of basic definitions will be provided.

2.1. *Basic definitions*

A DNA microarray dataset is represented as a matrix denoted by A_{RC}, where the set of rows is $R = \{1, \ldots, N\}$ and the set of columns is $C = \{1, \ldots, M\}$, each element a_{ij} of this matrix represents the level of expression of the gene i under condition j. Following this notation, given two subsets $I \subseteq R$ and $J \subseteq C$ of genes and conditions respectively, then B_{IJ} denotes the corresponding submatrix with $|I|$ genes and $|J|$ conditions and where b_{ij} denotes the level of expression of the gene i under the condition j of the submatrix.

A bicluster is therefore a submatrix B_{IJ} where the involved subset of genes (I) exhibits a similar behaviour pattern through the subset of conditions (J) [23,18,24].

2.2. *Taxonomy of biclustering algorithms*

In this paper an updated taxonomy of biclustering algorithms will be shown (see Tables 3 and 4 for an enumeration of the methods). The aim of this update is to include new dimensions to assess as accurately as possible the capabilities and specific characteristics of new biclustering algorithms developed since the key taxonomy of Madeira in [18]. Two new dimensions to characterize biclustering techniques have been incorporated: the nature of the data and the validation procedure. Finally, a new set of more complex coherence patterns is considered in the modeled coherence patterns set based on the presence of these behaviours in different types of cancer.

The axes to do a taxonomy of biclustering approaches are as follows:

- The nature of the data. In order to develop a biclustering algorithm it is necessary to decide whether the method will be able to extract the biclusters directly from continuous-value expression data or it is needed to perform a previous discretization task. In this sense current biclustering algorithms can be divided in different categories:
  1. Binary discretized data. This category groups all those algorithms that require a process of discretization of the gene expression data matrix to a binary data matrix [25,22,26]. It has the advantage of requiring much less computational resources to detect coherence patterns. Its drawback is the limited set of coherence of patterns detectable as pointed out by Ahmad in [27], the dependence of the discretization methodology [28] and therefore the impossibility to detect new continuous complex patterns.
  2. Integer discretized data. In this category the discretization process is still required to extract biclusters. The difference lies in the number of level of expressions obtained after the discretization process. Although the discretization can suppose a loss of information, it can be an advantage in terms of computational resources. This kind of approach provides faster algorithms and the loss of coherence pattern is not so considerable compared to the binary ones. However, it is easy to check that some shifting/scaling patterns are not detectable. Two subtypes are present in the biclustering literature depending on the number of levels of expressions. In the first subtype the gene expression data matrix is only discretized to three levels of expressions as it is commonly performed in DNS microarray studies: over-expressed, under-expressed, and baseline [29]. In the second subtype the gene expression data matrix is discretized to multiple levels of expression (more than three) [30–32].
  3. Original real numbers data matrix. In this category biclustering algorithms are able to extract directly biclusters from the original, continuous-value, microarray data. This kind of algorithms requires more complex processes and more computational resources. However, they have the advantage of detecting richer and more complex coherence patterns than previous scenarios. In this sense, the majority of the developed biclustering algorithms are able to deal with real data.
- Validation function. A second axe to divide bicluster learning algorithms is based on the quality requirements of the biclusters found and the algorithm's capability to recover known bicluster structures before to proceed with the biological analysis. Some biclustering approaches propose a set of validation procedures prior to assess the results from the biological perspective. These procedures aim to guarantee that it is useful to apply the proposed biclustering algorithm to real microarray data in order to perform a

posterior biological assessment. Two main kinds of validation approaches can be found in the literature:

1. Method/Model validation. The aim of this step is to evaluate the accuracy of the results of the method based on the computation of a set of measures on a predefined set of artificial microarray datasets. Different validation metrics can be found in the literature:
   - Precision/Recall metric. Most of the biclustering algorithms that use measures based on the precision/recall metrics such as false positive discovery rate [26], true positive discovery rate [33], etc.
   - Upper bounds on probability models. In this case upper bounds on the probability are computed based on random datasets in order to know the statistical significance of the probability of the obtained model [34].
2. Bicluster quality assessment. In this case the aim is to evaluate the coherence degree of a detected single bicluster. The different types of coefficients are the following:
   - Clustering coefficients. This kind of validation is based on the use of existing clustering coefficients that are adapted to the biclustering problem [35].
   - Information theory. This approach is used to measure the quantity of information that the bicluster found provide based on the whole microarray data information [36].
   - Specific coefficients. These are ad-hoc coefficients developed to measure the quality from different points of view such as the statistical significance of the biclusters obtained [29], or ad-hoc coefficients [37,32].

- Coherence patterns. An interesting criteria to evaluate a biclustering algorithm concerns the identification of the type of data pattern the algorithm is able to find. We identified several major classes of patterns already introduced in the taxonomy of Madeira [18]; a new pattern has been added:
  1. *Biclusters with constant values* [4]. The values of all elements of the bicluster are equal. Although this approach produces good results, in ideal conditions it is necessary to consider the noise that the real microarray data incorporates. This means that the values of the bicluster can be computed as:

$$b_{ij} = \mu + \eta_{ij} \tag{1}$$

  where μ represents the average value of the bicluster and η_{ij} represents the noise added.
  2. *Biclusters with constant values on rows or columns* [38,31]. This kind of pattern exhibits coherent variations in the rows or in the columns. That is, the value of each element b_{ij} of the bicluster can be broken down as the contribution of the average value of the bicluster (μ), the contribution of the value of the row (α_i) or the contribution of the value of the column (β_j). This set of contributions can be expressed in two possible ways: the additive model or the multiplicative model. In the additive model the contribution is expressed as the sum of the values of all involved elements: the average value of the bicluster and the value of the row, or the average value of the bicluster and the value of the column. On the other hand the multiplicative model is expressed as the product of the values of the corresponding elements (see Table 1).

Table 1 – Biclusters with constant values on rows or columns.

	Additive model	Multiplicative model
Rows	$b_{ij} = \mu + \alpha_i + \eta_{ij}$	$b_{ij} = \mu \times \alpha_i + \eta_{ij}$
Columns	$b_{ij} = \mu + \beta_j + \eta_{ij}$	$b_{ij} = \mu \times \beta_j + \eta_{ij}$

Table 2 – Biclusters with coherent values in rows or columns.

Additive model	Multiplicative model
$b_{ij} = \mu + \alpha_i + \beta_j + \eta_{ij}$	$b_{ij} = \mu \times \alpha_i \times \beta_j + \eta_{ij}$

  3. *Biclusters with coherent values* [6,39,40]. This kind of pattern exhibits coherent variations on the rows or on the columns. That is, the value of each element b_{ij} of the bicluster can be broken down as the contribution of the average value of the bicluster (μ), the contribution of the value of the row (α_i) and the contribution of the value of the column (β_j). This set of contributions can also be expressed in two possible ways: the additive model or the multiplicative model (see Table 2).
  4. *Biclusters with coherent evolutions* [41,29]. This kind of pattern represents the most difficult coherence pattern to be discovered. This pattern addresses the problem of discovering subsets of rows and subsets of columns with coherent behaviours, regardless of the exact numeric values in the data matrix. This collective coherent evolution can be observed in the entire bicluster, on both rows and columns of the bicluster, on the rows of the bicluster, or on the columns of the bicluster.

  Although there exist many kinds of coherent evolution patterns, two main cases are distinguished in the literature:
     - *Shifting pattern* [14]. A group of genes (conditions) follows a shifting pattern if their values under a specific condition (gen) can be expressed as the addition of some constant dependent on the condition (α_j) plus a value dependent on the specific gene (c_i). This can be expressed by means of a matrix as follows:

$$B_{IJ} = \begin{pmatrix} b_{11} & \ldots & b_{1J} \\ \ldots & \ldots & \ldots \\ b_{I1} & \ldots & b_{IJ} \end{pmatrix} = \begin{pmatrix} c_1 + \alpha_1 & \ldots & c_1 + \alpha_J \\ \ldots & \ldots & \ldots \\ c_I + \alpha_1 & \ldots & c_I + \alpha_J \end{pmatrix} \tag{2}$$

     In order to clarify this kind of pattern, Fig. 2a shows an example of a matrix that represents a bicluster with three genes and five conditions, where the genes follow a shifting pattern. Fig. 2b shows the graphical representation of the bicluster showing that the shifting pattern is a collection of copies of a vertically shifted pattern. Finally, Fig. 2c shows a graphical representation of the effects of the shifting patterns in the conditions.
     - *Scaling pattern* [14]. A group of genes (conditions) follows a scaling pattern if their values under a specific condition (gen) can be expressed as the product of

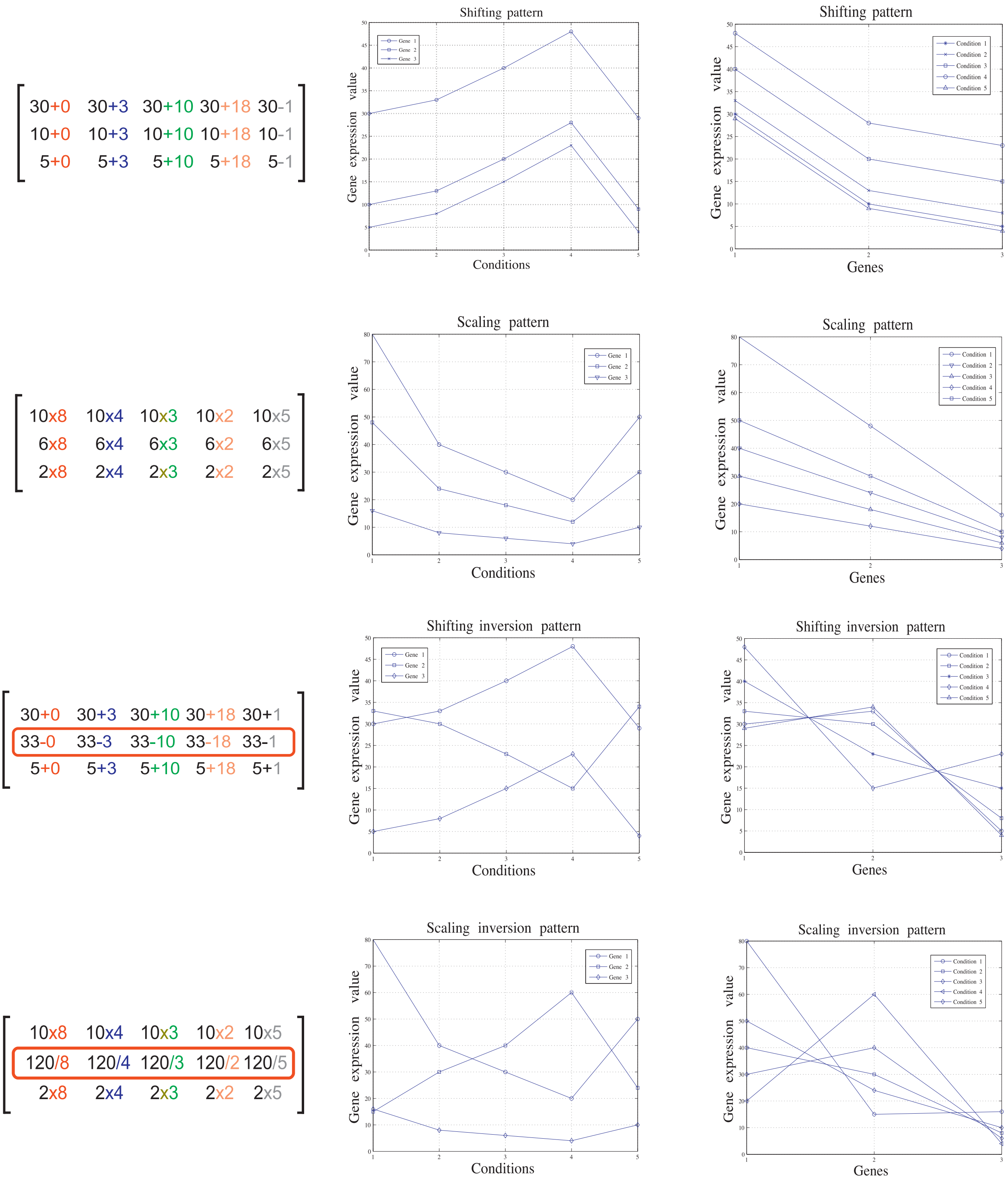


Fig. 2 – Coherence patterns: shifting, scaling and two types of inversion (shifting and scaling): (a) shifting pattern, (b) shifting pattern: genes, (c) shifting pattern: conditions, (d) scaling pattern, (e) scaling pattern: genes, (f) scaling pattern: conditions, (g) inverted shifting pattern, (h) inverted shifting pattern: genes, (i) inverted shifting pattern: conditions, (j) inverted scaling pattern, (k) inverted scaling pattern: genes, (l) Inverted scaling pattern: conditions.

some constant dependent on the condition (β_j) and a value dependent on the specific gene (c_i). This can be expressed by means of a matrix as follows:

$$B_{IJ} = \begin{pmatrix} b_{11} & \ldots & b_{1J} \\ \ldots & \ldots & \ldots \\ b_{I1} & \ldots & b_{IJ} \end{pmatrix} = \begin{pmatrix} c_1 \times \beta_1 & \ldots & c_1 \times \beta_J \\ \ldots & \ldots & \ldots \\ c_I \times \beta_1 & \ldots & c_I \times \beta_J \end{pmatrix} \tag{3}$$

We illustrate an example of this pattern in Fig. 2d. In this figure we show a matrix that represents a bicluster with three genes and five conditions where the genes follow a scaling pattern. As in the previous case Fig. 2e shows the corresponding graphical representation. Finally, Fig. 2f shows, the effects of the pattern in the conditions.

- *Shifting and inverted shifting pattern*. Recently, new interesting coherence patterns have been proposed. Some papers reflect inverse relationships associated with biological functions [42]. Furthermore, inverse relationships have been found in different kinds of cancer and tumors [15–17]. With the aim of enriching this set of coherence patterns, we have decided to include two new sets of inversely correlated coherence patterns in this taxonomy of biclustering algorithms.
  The first one is the shifting and inverted shifting pattern. This new category is defined as a subset of genes (conditions) that follows a shifting inversion pattern if their values under a specific condition (gene) can be expressed as the addition or subtraction of a constant dependent on the condition (α_j) from a value dependent on the specific gene (c_i). This can be also expressed by means of a matrix as follows:

$$B_{IJ} = \begin{pmatrix} \multicolumn{3}{c}{\overbrace{\hphantom{b_{11} \quad \ldots \quad b_{1J}}}^{\textit{Conditions}}} \\ b_{11} & \ldots & b_{1J} \\ \ldots & \ldots & \ldots \\ b_{i1} & \ldots & b_{iJ} \\ \ldots & \ldots & \ldots \\ b_{j1} & \ldots & b_{jJ} \\ \ldots & \ldots & \ldots \\ b_{I1} & \ldots & b_{IJ} \end{pmatrix} = \begin{pmatrix} \multicolumn{3}{c}{\overbrace{\hphantom{c_1 + \alpha_1 \quad \ldots \quad c_1 + \alpha_J}}^{\textit{Conditions}}} \\ c_1 + \alpha_1 & \ldots & c_1 + \alpha_J \\ \ldots & \ldots & \ldots \\ c_i - \alpha_1 & \ldots & c_i - \alpha_J \\ \ldots & \ldots & \ldots \\ c_j - \alpha_1 & \ldots & c_j - \alpha_J \\ \ldots & \ldots & \ldots \\ c_I + \alpha_1 & \ldots & c_I + \alpha_J \end{pmatrix} \tag{4}$$

In this matrix we have incorporated two inverted shifting patterns associated to the genes i and j. However the rest of the genes follow a classical shifting pattern. Fig. 2g shows an example of a matrix that represents a bicluster with three genes and five conditions where the second gene follows an inverted shifting pattern, whereas the first and the third gene follow a classical shifting pattern. Fig. 2h shows the graphical representation of the bicluster showing the shifting and the inverted shifting patterns as a collection of copies of the same vertically shifted pattern and vertically inverted pattern. Finally, Fig. 2i shows the effects of this pattern in the conditions but, contrary to previous patterns, the effect on the conditions is not recognizable.

- *Scaling and inverted scaling pattern*. This is the second category introduced and it is defined as a subset of genes within a group of genes (conditions) follows a shifting inversion pattern if their values under a specific condition (gene) can be expressed as the product or division of a value dependent on the specific gene (c_i) by some constant dependent on the condition (β_j). This can be expressed by means of a matrix:

$$B_{IJ} = \begin{pmatrix} \multicolumn{3}{c}{\overbrace{\hphantom{b_{11} \quad \ldots \quad b_{1J}}}^{\textit{Conditions}}} \\ b_{11} & \ldots & b_{1J} \\ \ldots & \ldots & \ldots \\ b_{i1} & \ldots & b_{iJ} \\ \ldots & \ldots & \ldots \\ b_{j1} & \ldots & b_{jJ} \\ \ldots & \ldots & \ldots \\ b_{I1} & \ldots & b_{IJ} \end{pmatrix} = \begin{pmatrix} \multicolumn{3}{c}{\overbrace{\hphantom{c_1 \times \alpha_1 \quad \ldots \quad c_1 \times \alpha_J}}^{\textit{Conditions}}} \\ c_1 \times \alpha_1 & \ldots & c_1 \times \alpha_J \\ \ldots & \ldots & \ldots \\ c_i / \alpha_1 & \ldots & c_i / \alpha_J \\ \ldots & \ldots & \ldots \\ c_j / \alpha_1 & \ldots & c_j / \alpha_J \\ \ldots & \ldots & \ldots \\ c_I \times \alpha_1 & \ldots & c_I \times \alpha_J \end{pmatrix} \tag{5}$$

  In this matrix we have also incorporated two inverted scaling patterns associated to the gene i and to the gene j. However, the rest of the genes follow a classical scaling pattern.
  Fig. 2j shows an example of a matrix that represents a bicluster with three genes and five conditions where the second gene follows a inverted scaling pattern, whereas the first and the third gene follow a classical scaling pattern. Fig. 2k, as in the previous case, shows the graphical representation of this kind of patterns. Finally, as in the previous pattern, Fig. 2l shows the effects of the pattern with the same result.

As a summary, a list of published biclustering techniques is shown in Tables 3 and 4, enumerating their capabilities based

Table 3 – Enumeration of biclustering methods and its classification.

Method	Pattern	Structure	Discovery	Data	Validation
	1 2 3 4 5	1 2 3 4 5 6 7 8 9	1 2 3	1 2 3 4	1 2
	1. Constant values 2. Constant values: row/col 3. Coherent values 4. Coherent evolutions I 5. Coherent evolutions II Inverted	1. Single 2. Exclusive row/col 3. Checkerboard 4. Exclusive row 5. Exclusive column 6. Non-overlapping: tree 7. Non-overlapping: non-exclusive 8. Overlapping: hierarchical 9. Arbitraly positioned overlapping	1. One at time 2. One set at time 3. Simultaneously	1. Binary 2. Integer I 3. Integer II 4. Real	1. Bicluster 2. Method
BlockClustering [4]	■□□□□	■■■■■■□□□	■■□	□□■■	■■
δ-Biclusters [6]	■■■□□	■■■■■■■■■	■□□	□□□■	■□
FLOC [97,98]	■■■□□	■■■■■■■■■	■■■	□□□■	■□
pClusters [99]	■■■□□	■■■■■■■□□	□□□	□□□■	□□
Plaid Models [39]	■■■□□	■■■■■■■■■	■□□	□□■■	□□
PRMs [100]	■■■□□	■■■■■■■■■	■■■	□□■■	■□
CTWC [38]	■■□□□	■■■■■■■■■	■■□	□□□■	■■
ITWC [40]	■■■□□	■■■■■□□□□	■■□	□□■■	■□
DCC [75]	■□□□□	■■■□□□□□□	■■■	□□■■	■□
δ-patterns [101]	■■□□□	■■■■■■■■■	■■■	□□□■	□■
Spectral [68]	■■■□□	■■■□□□□□□	■■■	□□■■	□□
Gibbs [31]	■■□□□	■■■■■□□□□	■■□	□□■□	■■
OPSMs [34]	■■■■□	■■■■■■■■■	■□□	□□■■	■■
SAMBA [29]	■■■■□	■■■■■■■■■	■■■	□■□□	■■
xMOTIFs [73]	■■■■□	■■■■■■■■■	■■■	□□□■	■□
OP-Clusters [102]	■■■■□	■■■■■■■■■	■■■	□□■■	□□
Scatter/Correlation [103]	■■■■■	■■■■■■■■■	■■■	□□■■	□□
FABIA [36]	■■■■□	■■■■■■■■■	■□□	□□■■	■■
KM-GS Hybrid [104]	■■■■□	■■■■■■■■■	■□□	□□■■	□□

on the exposed taxonomy. Its first column shows the name of the method. The second column describes the coherence patterns that the method is able to detect: by means of a row of five squares, each of them is filled if the method is able to detect the specific coherence pattern or empty otherwise (patterns are ordered by ascending complexity). The third column describes the bicluster structure as Madeira [18] defined in the original taxonomy and it is represented by a row of nine squares: each of them representing the capability of detection of artificially predesigned biclusters. The fourth column represents the discovery methodology carried out by the method. That is, depending on the algorithm there are methodologies which are able to discover one bicluster, a subset at a time or all biclusters simultaneously. So, those methodologies which are able to find several biclusters at the same time in the majority of cases are obviously able to find only one if these are correctly parameterized. Methodologies which are able to find all biclusters simultaneously are also able to find a subset of

Table 4 – Enumeration of biclustering methods and its classification (continuation).

BILS [105]	■■■■□	■■■■■■■■■	■□□	□□■■	□□
BCCA [106]	■■■■■	■■■■■■■■■	■□□	□□■■	□□
BiMine [37]	■■■■■	■■■■■■■■■	■□□	□□■■	■□
QUBIC [30]	■■■■□	■■■■■■■■■	■■□	□□■□	■□
VOTE [32]	■■■■□	■■■■■■■■■	■□□	□□■□	■□
BiCBin [26]	■□□□□	■■■■■■■■■	■□□	■□□□	■□
Non-Over [107]	■■■■■	■■■■■■■■■	■■□	□□■□	□□
MOEA-B [13]	■■■■□	■□□□□□□□□	■□□	□□■■	□■
EC [79]	■■■■□	■■■■■■■■■	■■□	□□■■	□□
Plaid-Enh [108]	■■■□□	■■■■■■■■■	■■■	□□■■	□□
Bic GF [109]	■□□□□	□□□□□□□□□	■■□	■□□□	□□
Mutual Information [33]	■■■■■	■■■■■■■■■	■■□	□□■■	□□
Diam-Clustering [110]	■■■■□	■■■■■■■■■	■■■	■■■■	□□
Geometric Gan [111]	■■■■□	■■■■■■■■■	■■■	■■■■	□□
Geometric Zhao [112]	■■■■□	■■■■■■■■■	■■■	■■■■	□□
Geometric Wang [113]	■■■■□	■■■■■■■■■	■■■	■■■■	□□

these. The fifth column represents the type of input needed by the algorithm, and finally the last column represents its validation capability: the presence/absence of an specific method validation procedure and the presence/absence of a procedure for bicluster assessment.

3. Actual and proposed coherence measures

3.1. *Classical mean squared residue (MSR)*

The mean squared residue (MSR) is a measure introduced by Cheng and Church [6] which has been widely used in different biclustering approaches and it is considered as the *benchmark* measure in biclustering literature [89]. The computation of this measure is defined as follows:

$$MSR(B_{IJ}) = \frac{1}{|I||J|}\sum_{i\in I, j\in J}(b_{ij} - b_{iJ} - b_{Ij} + b_{IJ})^2 \qquad (6)$$

where

$$b_{iJ} = \frac{1}{|J|}\sum_{j\in J} b_{ij} \qquad (7)$$

$$b_{Ij} = \frac{1}{|I|}\sum_{i\in I} b_{ij} \qquad (8)$$

$$b_{IJ} = \frac{1}{|I||J|}\sum_{i\in I, j\in J} b_{ij} \qquad (9)$$

b_{iJ} is the average of the elements of the column J, i.e. the average of the gene expression level through different conditions; b_{Ij} is the average of the elements of the row I, i.e. the average of the gene expression value of a condition through different genes; b_{IJ} is the average of all elements of the bicluster.

In spite of the success of this measure it has also been demonstrated that it shows some problems. The first problem was pointed out by Aguilar [14] and is related with the identification of the scaled versions of the same pattern. MSR gets different values for scaled versions of the same pattern. That is, it is not able to recognize the pattern. Furthermore, in this work we will show experimentally that it is not able to detect different inverted versions of the same pattern.

The second problem is related with the absence of a statistical base. The authors originally describe the measure as a submatrix A_{IJ} with a specific score. There is no reference with regard to any statistical base in the whole paper. Therefore, this measure does not include any approach to determine the statistical significance which can be used to assess results. The biclusters obtained with a methodology based on MSR can not be evaluated in statistical terms and therefore it is not possible to filter specific genes and/or conditions within the bicluster which do not meet basic statistical quality requirements. Furthermore, the original paper [6] describes the addition/deletion process of nodes by using the value of the score exclusively.

3.2. *Spearman's biclustering measure*

The main contribution of this paper is the proposal of a new measure capable of detecting normal and modified (shifted, scaled or inverted) versions of the same pattern between genes or conditions independent of the value.

A measure based on the concept of correlation is demonstrated as a suitable tool to quantify the distance between two genes through a set conditions [43]. The correlation is not only able to quantify the similar trends but also the direction of the trend. Although these capabilities are common to all correlation coefficients, not all of them are equally suitable to be used with DNA microarray data. As well as having datasets without outliers, some correlation coefficients can require normality conditions to provide reliable values of correlation.

Taking into account these potential problems, Spearman's correlation coefficient has some known interesting characteristics such as:

- **Invariability to shifting and scaling**, that is, a shifting/scaling pattern affects to specific values but it does not change the ranks of the data (the relative order of the data), therefore the correlation value does not change. We will illustrate this invariability by means of two examples where exists a shifting and scaling pattern (see Fig. 19). In both cases, the shifting or the scaling do not change ranking values and therefore the difference between the ranks "D" is zero and the correlation value is one. The invariability to shifting and scaling is formally proved in Appendix A.
- **Capability to detect nonlinear monothonic relationships**: since the dependence on the specific values of the data is relaxed and the relative order of the data is maintained, more complex patterns than shifting and scaling can be identified (e.g. inversion patterns).
- **It does not perform a Gaussian assumption of the data** [44,45].
- **Robustness against outliers** which are common in DNA microarrays datasets [46,47]. In robust statistics there are several basic tools to describe and measure robustness such as the breakdown point, the influence function and the sensitivity curve. In our proposed measure based on Spearman's correlation coefficient influence functions have been used to analyze the robustness of the measure concluding that our proposed correlation coefficient has a bounded smooth influence functions with a high efficiency [48].

The computation of this coefficient is accomplished in two steps. Firstly, the relative order of n measurements of X and Y written as x_i and y_i ($i = 1, \ldots, n$) is computed, i.e. the ranks. Finally, once the data is converted into ranks the coefficient is computed as follows:

$$r_{X,Y} = \frac{\sum_{i=1}^{n} x'_i y'_i - \frac{\left(\sum_{i=1}^{n} x'_i\right)\left(\sum_{i=1}^{n} y'_i\right)}{n}}{\left[\sum_{i=1}^{n} x'^2_i - \frac{\left(\sum_{i=1}^{n} x'_i\right)^2}{n}\right]^{1/2}\left[\sum_{i=1}^{n} y'^2_i - \frac{\left(\sum_{i=1}^{n} y'_i\right)^2}{n}\right]^{1/2}} \qquad (10)$$

The proposed Spearman's biclustering measure (SBM) makes use of this coefficient to detect trends in genes and in conditions within a bicluster. It is defined as follows:

$$\mathrm{SBM}(B_{IJ}) = \alpha(B_{IJ}) r^{\bar{G}}_{B_{IJ}} \beta(B_{IJ}) r^{\bar{C}}_{B_{IJ}} \tag{11}$$

The first set, $r^{\bar{G}}_{B_{IJ}}$, denotes the summarized expression of the trends observed in the genes of the considered bicluster. It is computed as follows:

$$r^{\bar{G}}_{B_{IJ}} = \frac{2}{|I|(|I|-1)} \sum_{i=1}^{|I|} \sum_{i'=i+1}^{|I|} |r^{G}_{ii'}| \tag{12}$$

where $|r^{G}_{ij}|$ is the absolute value of the Spearman's nonparametric correlation coefficient between the evolution of the gene *i* and the evolution of the gene *j* under all involved conditions of the bicluster. The computation of the absolute value is oriented to provide independence from the direction of the trend.

The second term, $r^{\bar{C}}_{B_{IJ}}$, denotes the summarized expression of the trends observed in the conditions and it is computed as:

$$r^{\bar{C}}_{B_{IJ}} = \frac{2}{|J|(|J|-1)} \sum_{j=1}^{|J|} \sum_{j'=j+1}^{|J|} |r^{C}_{jj'}| \tag{13}$$

where $|r^{C}_{ij}|$ represents the absolute value of the Spearman's nonparametric correlation coefficient between the evolution of the condition *i* and the evolution of the condition *j* under all involved genes of the bicluster.

Finally, the terms $\alpha(B_{IJ})$ and $\beta(B_{IJ})$ are the *reliability coefficients* and they weight the influence of the trends observed in the genes and the conditions respectively. The weighting is necessary because the computation of the correlation coefficients is performed with different sample sizes. In the case of computing the coefficient associated with the conditions, the coefficient is very reliable because it is computed with hundreds or thousands of samples. In the case of genes, the situation changes radically, the computation is performed with dozens of samples, that is one or two orders of lower magnitude. This situation involves a different influence in the computation of the SBM. In order to compensate this difference, we introduce the terms α and β. The term β is the coefficient associated to the conditions and therefore its reliability is very high (it is computed from hundreds of genes) and it is set to one. On the other hand, α is the coefficient associated to the conditions and therefore its reliability is very low and it is computed with few samples (subsets of conditions). This value is computed taking into account the number of samples as follows:

$$\alpha(B_{IJ}) = \begin{cases} 1 & \text{if } |J| > \mathit{Threshold}; \\ \dfrac{|J|}{M} & \text{if } |J| \leq \mathit{Threshold}. \end{cases} \tag{14}$$

where $|J|$ represents the number of the conditions of the bicluster and *M* the total number of conditions of the microarray. The selection of the threshold is based on the experimentation and allows to decide when it is necessary or not to take into account the weighting of the conditions. We have fixed this value to nine, it is minimum value that provides a coherent evolution in the results.

This expression represents a natural way of weighting the trends of both the conditions and the genes. An increase of the SBM measure value necessary implies a joint increase of the coherence in the genes and in the conditions.

4. The search of biclusters by means of randomized optimization heuristics

The whole set of all possible groupings of genes and conditions shapes the space of candidate biclusters. In this space each point represents a candidate bicluster that can be evaluated by a measure or merit function. The aim is to find the best candidate biclusters within this space. To accomplish the task of searching in the space of possible biclusters, several approaches have been used including simulated annealing [9], genetic algorithms [7,8,11,80,82] and greedy search techniques [49].

In this work we propose the use of a new search strategy based on a new bicluster quality measure combined with recent efficient evolutionary computation algorithms: estimation of distribution algorithms [50].

4.1. *Estimation of distribution algorithms*

Estimation of distribution algorithms (EDAs) is a novel class of evolutionary optimization methodology that was developed in the last decade as a natural alternative to genetic algorithms. The principal advantages of EDAs over genetic algorithms are the absence of multiple parameters to be tuned (e.g. crossover and mutation probabilities) and the expressiveness and transparency of the probabilistic model that guides the search process. In addition, EDAs have been proven to be better suited to some applications than GAs, while achieving competitive and robust results in the majority of tackled problems including in a variety set of bioinformatics problems [51].

4.1.1. *Introduction*

Estimation of distribution algorithms [50,52–55] are evolutionary algorithms that work with a population of candidate solutions (in our case candidate biclusters). Fig. 3 illustrates the general pseudocode for any EDA approach. Initially, a random sample of candidate biclusters is generated. These candidate biclusters are evaluated using an objective function. Based on this evaluation, the best points are selected. Then, a probabilistic model of the selected solutions is learned, and a new set of points is sampled from the model. The process is iterated until the optimum has been found or another termination criterion is fulfilled.

4.1.2. *A basic taxonomy of EDAs*

Since several EDAs have been proposed with a variety of models and learning algorithms, the selection of the best type of EDA to deal with a given optimization problem is not always straightforward. A criterion is to trade off the complexity of the probabilistic model with respect to the computational cost of storing and learning the selected model. Both issues are also related to the problem dimensionality (i.e. number

```
Step 1 $D_0 \longleftarrow$ Generate and evaluate $R$ individuals (the initial population) at random
Repeat for $l$=1,2,... until the stopping criterion is met
        Step 2 $D_{l-1}^{Se} \longleftarrow$Select $P \leq R$ individuals from $D_{l-1}$ according
               to the selection method
        Step 3 $p_l(\boldsymbol{x}) = p(\boldsymbol{x}|D_{l-1}^{Se})$ Estimate the probability distribution
               of an individual being among the selected individuals
        Step 4 $D_l \longleftarrow$ Sample and evaluate $R$ individuals (the new population)
               from $p_l(\boldsymbol{x})$
End repeat
```

Fig. 3 – EDA pseudocode.

of variables) and to the type of representation (e.g. discrete, continuous, mixed).

Researchers should be aware that simple models generally have minimal storage requirements, and are easy to learn. On the other hand, more complex models, which are able to represent more complex relationships, may require sophisticated data structures and costly learning processes. The impact that the choice between simple and more complex models has in the search efficiency will depend on the addressed optimization problem. Another criterion that should be taken into consideration is whether there is any previous knowledge about the problem structure, and which kind of probabilistic model is best suited to represent this knowledge. The following classification of EDAs is intended to help the bioinformatic researcher to find a suitable EDA type to be applied.

EDAs can be broadly divided according to the complexity of the probabilistic model used to capture the interdependencies between the variables: univariate, bivariate or multivariate approaches. Univariate EDAs, such the univariate marginal distribution algorithm (*UMDA*) [56,92], assume that all variables are independent and factorize the joint probability of the selected points as a product of univariate marginal probabilities. Consequently, these algorithms are the simplest EDAs with the fastest CPU execution times.

Bivariate models can represent low order dependencies between the variables. mutual information maximization for input clustering (*MIMIC*) [57], the bivariate marginal distribution algorithm BMDA [58], dependency tree-based EDAs [59] and the tree-based estimation of distribution algorithm (Tree-EDA) [60] are examples of this subclass.

Multivariate EDAs factorize the joint probability distribution using statistics of order greater than two. As the number of dependencies among the variables is higher than in the above categories, the complexity of the probabilistic structure, as well as the computational effort required to find the structure that best suits the selected points, is greater. Some algorithms that belong to this group are *EBNA* [61], and *BOA* [62].

For detailed information about the characteristics and different algorithms that constitute the family of EDAs, see [50,63].

4.1.3. *Univariate marginal distribution algorithm*

The biclustering problem requires the handling of thousands of variables which implies that in general the CPU time and memory required to handle complex models can be huge. Biclustering requires performing a searching task in a huge dimensionality space: it is typical to have thousands of genes and dozens of conditions. This task can be performed efficiently by using an EDA which explicitly represents each of the elements involved in the problem, in this case the genes and the conditions. The decision of selecting complex models where relationships between genes and conditions are explicitly represented can have several disadvantages such as:

- The first one is related to the complexity of the model to represent a subset of biclusters. If each point of the initial population must represent a codified set of biclusters then it involves the use of more complex probabilistic model. Therefore, the use of these kind of models involves more computational resources in different parts of the algorithm such as:
  - The size of the population. This required for correctly estimating the parameters of the model increases. Taking into account that there are thousands of variables this could represent a significant increase of time and computational resources whenever we decide to maintain the quality of the search.
  - The sampling of the model. Based on a more complex model the time spent to sample increases and this time which it is now moderated it could be prohibitive
  - The learning of the probabilist model. The existence of a more complex model involves more computations and a significant increase of population in order to correctly estimate the parameters.
  - Computation of the measure. With a bigger population the increase of the computation time is very significant. We must take into account that our experiments were performed with microarrays with few thousands of genes if we try more complex models that is with bigger microarrays the computational resources involved could be prohibitive.

- Secondly, we must take into account that our measure evaluates the quality of the relationships between genes and conditions not only between genes. This extra computational load increase the resources needed in order to evaluate a candidate bicluster.
- Finally, our proposed algorithm performs a search in a huge space, this space is exponential in the number of variables to be considered. Any kind of increase on the number of variables has a great impact on the difficulty of finding good promising candidate biclusters.

Therefore, the modelization of the interrelations between the genes and conditions is discarded due to the prohibitive costs cited previously. Based on this premise, EDAs such as UMDA [51] can be an ideal candidate to handle biclustering problems.

The model chosen is known as univariate marginal distribution algorithm for discrete domains: $UMDA_d$. This method considers a model where there are no interrelations between the variables, and the probability distribution can be learnt as:

$$p_l(\mathbf{x}) = \prod_{i=1}^{R} p_l(x_i) \tag{15}$$

This distribution represents the product of R independent probability distributions associated to the genes and the conditions of the DNA microarray. $p_l(x_i)$ is a Bernoulli probability distribution which takes two values: 1 if the gene or the condition is selected and 0 if the gene or the condition is not selected in the considered bicluster.

Based on this model, the estimation of the parameters of the model is performed based on marginal frequencies of the selected subset of biclusters:

$$p_l(x_i) = \frac{\sum_{j=1}^{R} \delta_j(X_i = x_i|D_{l-1}^{Se})}{R} \tag{16}$$

where

$$\delta_j(X_i = x_i|D_{l-1}^{Se}) = \begin{cases} 1 & \text{if in the } j\text{th case of } D_{l-1}^{Se}, X_i = x_1 \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

Finally, a new population is obtained sampling the built joint probability distribution from selected candidate solutions.

4.2. Description of the proposed algorithm for bicluster search

In order to better understand the proposed algorithm we will first proceed with the description of the representation of a candidate bicluster.

A candidate bicluster ($\mathbf{x}$) is represented by a binary array of $N+M$ elements:

$$\mathbf{x} = (\overbrace{x_1, \ldots, x_N}^{Genes}, \overbrace{x_{N+1}, \ldots, x_{N+M}}^{Conditions}). \tag{18}$$

The first N elements, ($x_1, \ldots, x_N$), represent each gene of the microarray. The next M elements, ($x_{N+1}, \ldots, x_{N+M}$), represent the conditions. A value of 1 in the ith position (where $i \in \{1, \ldots, N\}$) indicates that the ith gene has been selected for inclusion in the considered bicluster whereas a value of 0 indicates that the gene has not been selected. On the other hand, a value of 1 in the $(N+j)$th position (where $j \in \{1, \ldots, M\}$) indicates that the $(N+j)$th condition has been selected for inclusion in the bicluster and a value of 0 indicates that the condition has been excluded.

For example, let's consider a DNA microarray with 8 genes and 4 conditions. An example of a bicluster B_{IJ} where $I = \{3, 4, 5\}$ are the selected genes and $J = \{2, 3\}$ are the selected conditions, is codified as: $\mathbf{x} = (\overbrace{0, 0, 1, 1, 0, 0, 0}^{Genes} \ \overbrace{0, 1, 1, 0}^{Conditions})$.

The proposed methodology is a two-step process. The first step is a standard sequential extraction search process [6,13,64] and the second step is a statistical assessment process. The sequential extraction search process is composed of two coordinated algorithms: an outer algorithm and an inner algorithm. The role of the outer algorithm is to obtain a set of biclusters by requesting them from the inner algorithm and to control the diversity of this set, that is, to guarantee that there are not two biclusters with an overlapping greater than 50%.

The role of the inner algorithm is to find the best bicluster following a typical evolutionary sequence of steps, it is therefore a bicluster finder.

Initially, a collection of candidate biclusters is randomly generated. The diversity of the bicluster collection is controlled by means of the Hamming distance, measuring the overlapping rate with regard to the list of biclusters provided by the outer algorithm. A generated bicluster is rejected if the overlapping rate is greater than 50%. Otherwise it is accepted to be included in the population of biclusters.

Secondly, the whole collection of candidate biclusters is evaluated using the SBM measure (see Eq. eq:HS1). Based on this evaluation, half of the most fitted candidate biclusters are selected. Then, a probabilistic model of the selected biclusters is built, and a new collection of biclusters is sampled from the model within the exposed EDA evolutionary process. The process is iterated until a previously fixed number of iterations is reached. The best bicluster obtained in this process is provided to the outer algorithm and it is included in the final list of biclusters.

The inner algorithm is a bicluster finder which can be again invoked by the outer algorithm, that is, the controller algorithm, until a fixed number of biclusters is achieved. The algorithm concludes with a set of biclusters with different kinds of bicluster patterns only constrained by the predefined overlapping rate.

The list of biclusters obtained by the outer algorithm is then assessed by the next process: the statistical assessment process. This is a quality filtering process where every pair of genes and condition is statistically tested and whose result is a subset of biclusters considered to be statistically significant. Throughout this process some genes and/or conditions can be removed, even a whole bicluster can be removed from the original set.

5. Experimentation: analysis of the proposed SBM measure

In this work, we have considered a two-step experimentation process: an individual analysis of the SBM measure and a comparative analysis. The first part of the analysis was aimed to show the suitability and applicability of the proposed approach in the problem of biclustering and before performing any kind of comparative analysis. It has been therefore, a necessary requisite to continue with the next step of the experimentation.

In this step, we consider the combination of previous validation approaches (see Section 2.2) based on the use of artificial and real DNA microarray datasets to evaluate different aspects of the proposed algorithm. These aspects consist of analyzing the relative performance of different quality indexes such as the identification of patterns, the stability of the results [64–66], and the statistical significance of the quality of the final biclusters. Another aspects involved in this procedure are the required steps to perform a biological assessment of the results based on the common practices found in the biclustering literature. To accomplish this task the procedure of experimentation has been divided into two phases.

5.1. *First phase: artificial experimentation*

The first step entails the evaluation and comparison of MSR and SBM by means of a set of quality indexes and the statistical assessment of the quality of the biclusters found by the SBM.

In order to accomplish the comparison, a reference set of artificial DNA microarray has been created. Each artificial microarray includes a bicluster with a different specific pattern (see Section 5.1.1). Based on this set of artificial DNA microarrays, a search of the biclusters is performed by means of the same *UMDA*$_d$ procedure using as fitness function the SBM measure and MSR measure as fitness functions with the same stopping condition. The results of the search have been used to compute the quality indexes for both measures. The comparison of both measures has been carried out by means of a set of statistical tests using the results provided by the quality indexes.

Finally, based on the statistical properties of the SBM, we have performed an assessment of the quality of the final discovered bicluster.

In the following subsections each of the elements and each of the steps involved in this phase of experimentation are explained.

5.1.1. *Definition of the artificial DNA microarrays datasets*

The process of generating the set of different reference microarray has been accomplished in several steps. Firstly, we have used a microarray data simulation software known as SIMAGE [44]. This simulation software uses a model that allows to perform simulations of microarray data (see Tables 5 and 6) based on the following set of parameters: gene expression, missing data, scanning device bias, nonlinearity effect, background surface variation, random error, left-tail

Table 5 – SIMAGE model parameters.

SIMAGE parameters (Part I)		
Array number of grid rows	n_{row}	4
Array number of grid columns	n_{col}	4
Number of spots in a grid row	n^2_{spot}	20
Number of spots in a grid column	n^2_{spot}	20
Number of spot pins	n_{pin}	4
Number of technical replicates	n_{rep}	2
Number of genes		200
Number of slides	n_{slide}	20
Perform dye swap		Yes
Gene expression filter		Yes
Reset gene filter for each slide		no
Mean signal	μ	11.492
Change in log2ratio due to upregulation	μ_+	0.832
Change in log2ratio due to downregulation	μ_-	-0.605
Variance of gene expression	σ^2_G	1.775
% of upregulated genes	π_+	0
% of downregulated genes	π_-	0
Correlation between channels	ρ	0.981
Dye filter		Yes
Reset dye filter for each slide		Yes
Channel (dye) variation	$\sigma_{channel}$	0.51
Gene x Dye	X_{gk}	0
Error filter		Yes
Reset error filter for each slide		Yes
Random noise standard deviation	$\sigma_{epsilon}$	0.219
Tail behaviour in the MA plot	δ	0.09
Non-linearity filter		Yes
Reset non-linearity filter for each slide		Yes
Non-linearity parameter curvature	α_l	0.025
Non-linearity parameter tilt	α_2	0.777
Non-linearity from scanner filter		Yes
Reset non-linearity scanner filter for each slide		Yes
Scanning device bias (0 = clipped, 1 = fully non-linear)	w	0.295
Spotpin deviation filter		Yes
Reset spotpin filter for each slide		no
Spotpin variation	σ_{pin}	0.36
Background filter		Yes
Reset background filter for each slide		Yes
Number of background densities	n_{bg}	5
Mean standard deviation per background density	σ_{bg}	0.3
Maximum of the background signal (%) relative to the non-background signals	b	100

Table 6 – SIMAGE model parameters.

SIMAGE parameters (Part II)		
Standard deviation	σ_ϵ	0.1
Background gradient filter		Yes
Reset gradient filter for each slide		Yes
Maximum slope of the linear tilt	s	700
Missing values filter		Yes
Reset missing spots filter for each slide		Yes
Number of hairs	n_h	10
Maximum length of hair	l_h	20
Number of discs	n_d	6
Average radius disc	l_d	10
Number of missing spots	n_s	0

Table 7 – Artificial patterns.

Pattern	Name	Figure
# 1	Bicluster with constant values	4a
# 2	Bicluster with constant values in rows	4b
# 3	Bicluster with constant values in columns	4c
# 4	Bicluster with coherent values	4d
# 5	Bicluster with coherent evolutions	4e
# 6	Bicluster with inverted coherent evolutions	4f

behaviour, gene × dye interaction, channel error, spot pin error and a replication error.

From the data provided by this software, we obtained a reference microarray dataset. Based on this reference microarray dataset, we generated six different microarray datasets. Within each newly created microarray we insert only one different kind of coherence pattern following the taxonomy of coherence patterns explained in Section 2.2.

Six artificial DNA microarrays datasets incorporating a single coherence pattern inside each microarray (see Table 7 and Fig. 4) have been created: a bicluster with constant values, a bicluster with constant values in rows, a bicluster with constants values in columns, a bicluster with coherent values, a bicluster with coherent evolution (with some scaled versions of the same pattern), and a bicluster with inverted coherent evolutions. In all cases the microarray has 100 genes and 20 conditions, that is, a scaled version of the real microarrays.

5.1.2. *Definition of the biclustering global quality indexes*

In order to carry out a comprehensive comparison among various biclustering measures, two sets of quality indexes are used. Some of these have been proposed in different ways in recent works [67–69].

The first set provides information about the following features: stability and global discovery capability of biclusters, represented by a set of global quality indexes. Stability is a relatively neglected issue of the methods used in high-dimensional optimization tasks. Stability, defined originally [64–66] as the sensitivity of a method to variations in the training set, has been recently applied to biclustering methods [67]. Our application to biclustering has been performed in a different way with the aim of assessing the sensitivity of the method to the variations in the DNA microarray datasets and to the choice of the parameters of the model. This concept has been adapted to our evaluation framework as:

- *Stability*. Measurement of the variability of the results obtained in different search processes of biclusters. It is computed as the average value of the differences between the best candidate biclusters found codified as binary arrays obtained in different search processes. It is formally expressed as:

$$\mathit{Stability} = \frac{2}{r(r-1)}\sum_{i=1}^{r}\sum_{j=i+1}^{r} d(\mathbf{x}^{i}, \mathbf{x}^{j}) \tag{19}$$

where r represents the number of search processes performed, $\mathbf{x}^i$ and $\mathbf{x}^j$ represent the best candidate biclusters codified as binary arrays in the ith and jth search processes respectively, and $d(\mathbf{x}^i, \mathbf{x}^j)$ represents the Manhattan's relative distance between two candidate biclusters:

$$d(\mathbf{x}^{i}, \mathbf{x}^{j}) = \frac{\sum_{p=1}^{N+M} |x_p^i - x_p^j|}{N+M} \tag{20}$$

The computation of this metric allows to quantify the stability by means of a value between 0 and 1. The stability is considered high when the obtained values are close to 0 and low when the values are close to 1.

The problems related with the absence of a valid descriptor of the discovery capability will be shown by means of a simple example. The aim of this example is to show graphically the candidate bicluster discovered by using MSR and SBM. In order to fairly compare the results, both measures use the same search algorithm, the same stopping condition and obviously the same DNA microarray dataset with an artificial bicluster inserted. Therefore we will perform two search processes, the first one by using SBM and the second one by using MSR. At the end of the search process we will show two kinds of results for each measure, the first one is the graphical expression of the discovered bicluster within the microarray and the second one is the evolution of the value of the measure. The results of these independent search processes carried out by using MSR and SBM will be shown in several figures. Fig. 5a shows the evolution of the MSR value along the generations whereas Fig. 5b shows the final coherence pattern discovered. On the other hand, Fig. 5c shows the evolution of the On the other hand, Fig. 5c shows the evolution of the SBM value along the generations whereas Fig. 5d shows the final coherence pattern.

Several conclusions can be extracted from these figures: the first one is that the shape of the discovered bicluster by MSR and SBM are quite different. MSR discovers a bicluster where all conditions are selected but not all genes. Whereas SBM discover partially the original bicluster with the corresponding conditions. The second one is that the value of the biclustering measure is not a good descriptor of the quality of the discovered bicluster. It is necessary to introduce some kind of combination of characteristics to describe the properties of the discovery capability of a biclustering method.

In order to incorporate this information into the global properties, some contributions [67,69] propose a combination of characteristics to describe the properties of the global discovery capability of a biclustering method.

In the case of the patterns of simulated biclusters, these characteristics are based on the fact that it is possible to know the detection errors since we know which gene-condition combinations belong to the true biclusters. Therefore it is possible to know not only the amount but also the kind of errors produced in the identification of the elements. Using this information it is possible to classify the different kinds of errors within a confusion matrix specifically defined for biclustering problems: the columns represent the identifications of the elements performed by the biclustering method and the rows represent the real elements that form the bicluster. This type of table summarizes the correct and incorrect identifications of the gen-condition specific combinations. We call it *biclustering confusion matrix* (see Fig. 6).

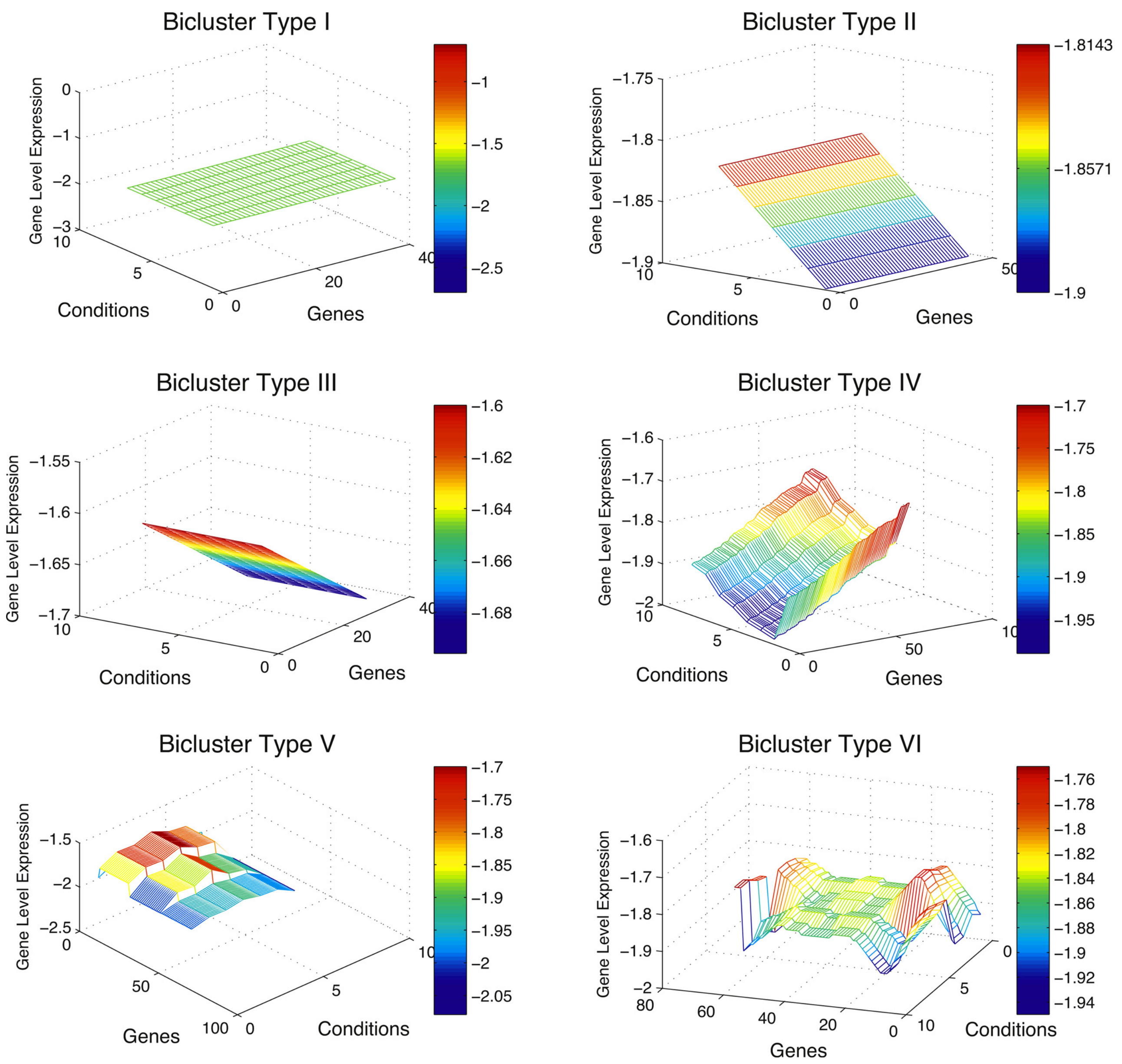


Fig. 4 – Considered patterns in the experiments with artificial data: (a) bicluster with constant values, (b) bicluster with constant values in conditions, (c) bicluster with constant values in genes, (d) bicluster with coherent values, (e) bicluster with coherent evolutions, and (f) bicluster with inverted coherent evolutions.

Based on this matrix it is possible to measure each kind of error by means of the following quality indexes:

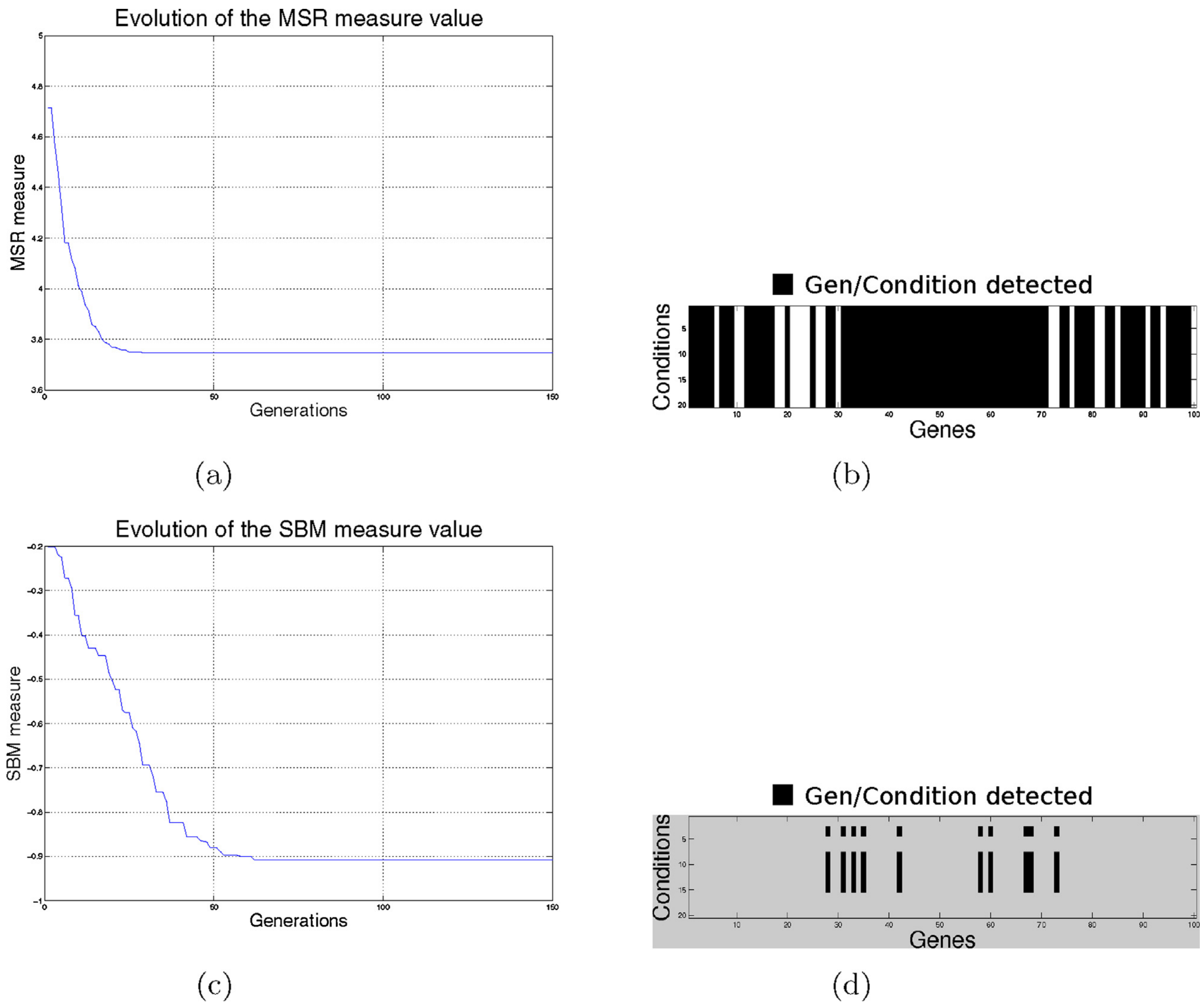
- *Bicluster global accuracy (BGA)*. This quality index determinates the percentage of elements of the artificial DNA microarray correctly identified. The identification includes those gen-condition combinations belonging to the true bicluster and those belonging to rest of the microarray. This metric is computed as:

$$\text{BGA} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Positives} + \text{Negatives}} \tag{21}$$

The values of this index lie between 0 and 1, representing a high accuracy those values near to 1. On the other hand values near to 0 represent a low accuracy.

- *Bicluster global specificity (BGP)*. This quality index measures the false positive rate. That is, the rate of identification of those gen-condition combinations belonging to the rest of the microarray incorrectly identified as belonging to the true bicluster. It is therefore the worst error committed if we take into account the costs associated with the study of the identified genes. This value is computed as:

$$\text{BGP} = \frac{\text{True Negatives}}{\text{False Positives} + \text{True Negatives}} \tag{22}$$



(a) (b)

(c) (d)

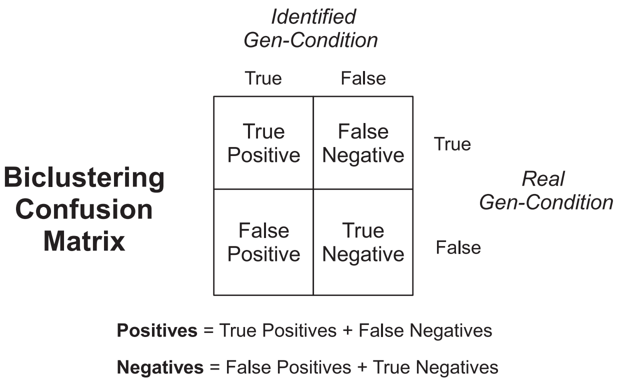
Fig. 5 – Capability of detection in a single execution.

The BGP values lie between 0 and 1. As in the previous index, values close to 1 point out that the BGP is low whereas values near to 0 represent a high false positive rate.

- *Bicluster global sensitivity (BGE)*. This quality index computes the rate of those gen-condition combinations belonging to the true bicluster identified as belonging to the rest of the microarray. This value is computed as:

$$\mathrm{BGE} = \frac{\text{True Positive}}{\text{True Positives} + \text{False Negatives}} \tag{23}$$

The values of this metric are between 0 and 1, representing a low false negative rate values close to 1. Values near to 0 point out a high false negative rate.



	True	False
True	True Positive	False Negative
False	False Positive	True Negative

Positives = True Positives + False Negatives

Negatives = False Positives + True Negatives

Fig. 6 – Biclustering confusion matrix.

In order to clarify these metrics, an example is used. Let's consider an artificial DNA microarray dataset with 8 genes and 5 conditions (see Fig. 7). Let's suppose that we have an artificial true bicluster (represented by lined cells) formed by the subset of genes $I = \{2, 3\}$ and by the subset conditions $J = \{2, 3, 4\}$.

On the other hand, let's suppose that we have a biclustering algorithm which identifies a bicluster formed by the subset of genes $I' = \{3, 4\}$ and the subset of conditions $J' = \{2, 3\}$. It is represented by bold cells. Using this information we build the biclustering confusion matrix. Next, the values of the quality indexes are computed (see Fig. 7).

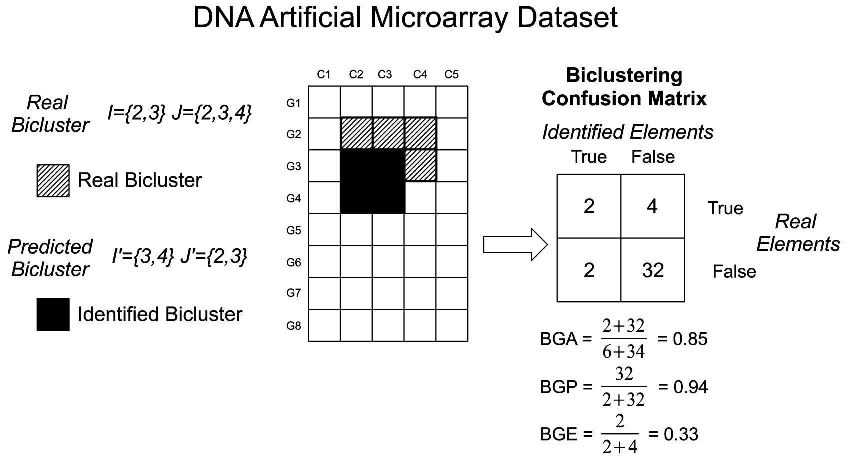


Fig. 7 – Procedure of constructing the biclustering confusion matrix and its associated BGA, BGP and BGE.

5.1.3. *Definition of the biclustering specific quality indexes*

Previous quality indexes provide a description of the global behaviour of a measure. However they do not explain the influence of the genes and the conditions in the global discovery capability. That is, it is not possible to know either the specific rate of genes correctly detected individually or the rate of conditions correctly detected individually. In order to fill this gap, we propose the creation of a second set of quality indexes dissociating the genes and the conditions. That is, we propose to create a set of quality indexes to be able to classify the number and the kind of errors separately for the genes and for the conditions. We call the table that summarizes the detection errors associated to the genes as *genes biclustering confusion matrix* and the table associated to the conditions as *conditions biclustering confusion matrix* (see Fig. 8).

Based on these matrices it is possible to define the following set of specific quality indexes:

- Bicluster Local Gen/Condition Accuracy (BLA_G/BLA_C). This determinates the percentage of genes/conditions of the artificial DNA microarray correctly identified. As in the case of global quality indexes, this identification includes those genes/conditions belonging to the bicluster and those belonging to the rest of the microarray. It is computed as:

$$\mathrm{BLA}_{G/C} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Positives} + \text{Negatives}} \tag{24}$$

  where G/C represents genes or conditions.
- Bicluster Local Gen/Condition Specificity (BLP_G/BLP_C). This measures the false positive rate of genes/conditions. This value is computed as:

$$\mathrm{BLP}_{G/C} = \frac{\text{True Negatives}}{\text{False Positives} + \text{True Negatives}} \tag{25}$$

  where G/C also represents genes or conditions.
- Bicluster Local Gen/Condition Sensitivity (BLE_G/BLE_C). This computes the false negative rate of genes/conditions. This value is computed as:

$$\mathrm{BLE}_{G/C} = \frac{\text{True Positive}}{\text{True Positives} + \text{False Negatives}} \tag{26}$$

  where G/C represents genes or conditions.

In order to clarify these new quality indexes, the previous example (see Fig. 7) has been used to compute this new set of quality indexes (see Fig. 9).

5.1.4. *Statistical assessment of SBM and MSR by means of biclustering quality indexes*

Firstly, a search of the biclusters contained in the set of designed artificial DNA microarray datasets by means of the exposed $UMDA_d$ procedure has been performed. The search is repeated 20 times for both measures: SBM and MSR obtaining 20 different biclusters for each measure.

The average and associated standard values of biclustering global quality indexes and biclustering specific quality indexes are computed for 20 found biclusters for MSR and SBM. Finally, all information is grouped into four tables (Tables 8–11).

Tables 8 and 9 shows the results corresponding to global quality metrics. The first column shows the artificial microarray used (see Table 7). The next 4 columns show the results of the global quality indexes.

Tables 10 and 11 show the results corresponding to the specific quality metrics computed by SBM and MSR respectively. Columns are organized in the same way as in Table 5, showing the values of the specific quality metrics associated with the genes in the first 3 columns and the values associated with the conditions in the last 3 columns. The best average values for each bicluster quality index are marked in bold in all tables.

The results show that SBM presents a good performance with regard to the quality indexes with the exception of the quality indexes associated to the false negative rate. In these cases, the SBM shows a lower performance than MSR, that is SBM has a more conservative policy than MSR in the selection process.

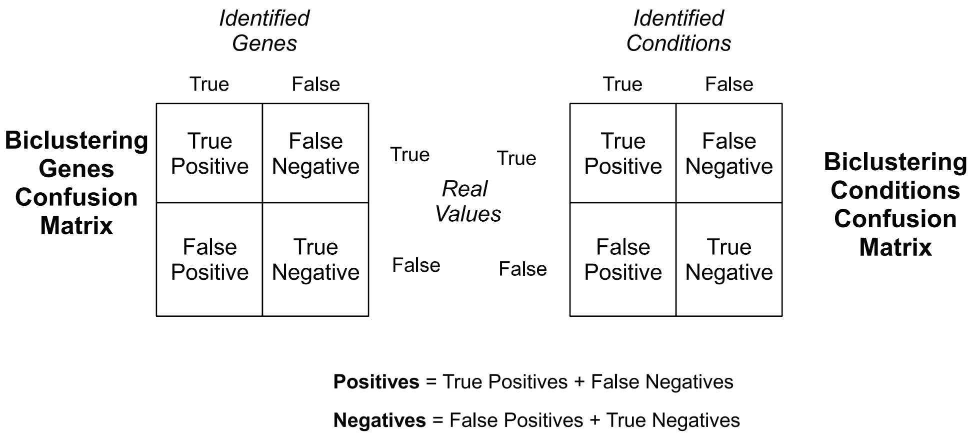


Fig. 8 – Biclustering genes and conditions confusion matrices.

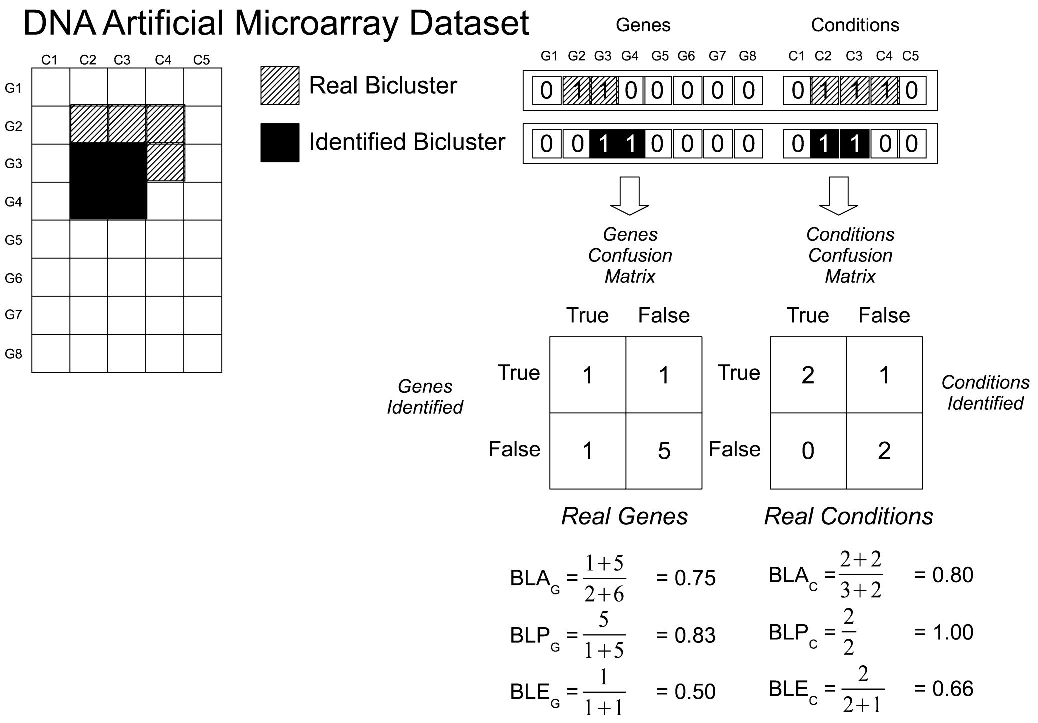


Fig. 9 – Procedure of evaluation of artificial data.

Based on these tables, we have carried out a statistically hypothesis-testing comparison among MSR and SBM biclustering measures by means of the Wilcoxon paired-signed rank test as suggested by Demsar [70].

SBM obtains a significant difference in performance for the accuracy of the false positive rate in the global quality indexes. However, it obtains a significative worse difference in the false negative rate. MSR obtains significative difference in the

Table 8 – Summary of the results obtained by SBM for the biclustering global quality indexes.

Pattern	SBM			
	Stability	BGA	BGP	BGE
# 1	15.97 ± 0.00	88.33 ± 0.40	98.16 ± 0.32	90.16 ± 0.65
# 2	7.21 ± 0.00	85.38 ± 0.32	98.36 ± 0.24	87.01 ± 0.57
# 3	15.19 ± 0.00	90.28 ± 0.56	97.97 ± 0.51	92.31 ± 0.80
# 4	10.56 ± 0.00	78.73 ± 0.65	98.69 ± 0.22	80.04 ± 0.87
# 5	3.97 ± 0.00	79.79 ± 0.55	96.82 ± 0.57	82.97 ± 0.21
# 6	6.17 ± 0.00	71.43 ± 0.49	94.16 ± 0.64	77.27 ± 0.29
Average	9.85 ± 4.93	82.07 ± 7.03	97.35 ± 1.69	84.79 ± 5.88

Table 9 – Summary of the results obtained by MSR for the biclustering global quality indexes.

Pattern	MSR			
	Stability	BGA	BGP	BGE
# 1	0.00 ± 0.00	35.00 ± 0.00	35.00 ± 0.00	100.00 ± 0.00
# 2	0.00 ± 0.00	31.80 ± 0.00	32.20 ± 0.00	100.00 ± 0.00
# 3	0.00 ± 0.00	38.40 ± 0.00	39.40 ± 0.00	100.00 ± 0.00
# 4	0.00 ± 0.00	33.20 ± 0.00	33.20 ± 0.00	100.00 ± 0.00
# 5	0.00 ± 0.00	30.00 ± 0.00	30.00 ± 0.00	100.00 ± 0.00
# 6	0.00 ± 0.00	34.80 ± 0.00	34.80 ± 0.00	100.00 ± 0.00
Average	0.00 ± 0.00	33.76 ± 00.00	33.98 ± 00.00	100.00 ± 00.00

Table 10 – Summary of the results obtained by SBM for the biclustering specific quality indexes.

Pattern	Genes			Conditions		
	BLA_G	BLP_G	BLE_G	BLA_C	BLP_C	BLE_C
# 1	88.08 ± 1.86	99.90 ± 0.31	71.90 ± 1.86	85.00 ± 0.00	85.00 ± 0.00	100.00 ± 0.00
# 2	89.56 ± 2.28	100.00 ± 0.00	60.46 ± 10.52	85.00 ± 0.00	85.00 ± 0.00	100.00 ± 0.00
# 3	88.92 ± 2.28	99.20 ± 0.70	78.05 ± 2.28	85.00 ± 0.00	85.00 ± 0.00	100.00 ± 0.00
# 4	86.87 ± 2.17	100.00 ± 0.00	50.10 ± 2.17	90.00 ± 0.00	90.00 ± 0.00	100.00 ± 0.00
# 5	88.65 ± 0.59	97.05 ± 1.10	51.35 ± 0.59	85.00 ± 0.00	85.00 ± 0.00	100.00 ± 0.00
# 6	92.10 ± 0.97	92.60 ± 0.64	45.90 ± 0.97	60.75 ± 4.38	74.50 ± 2.24	86.25 ± 2.75
Average	89.02 ± 1.75	98.09 ± 2.93	58.50 ± 12.95	81.14 ± 10.50	83.95 ± 5.10	97.56 ± 5.61

stability quality index. The only remarkable fact is the accuracy of the inversion pattern in SBM and MSR. In this case, the high rate of false positive allows to obtain higher accuracies than SBM. SBM is more conservative with regard to the selection of genes. This explains the high rate of false negative.

5.1.5. *Statistical assessment of the bicluster pattern coherence degree of a discovered bicluster*

The assessment of the bicluster pattern coherence degree is a problem which has yet to be settled [68]. Our aim is to incorporate an assessment method of the bicluster pattern coherence degree in statistical terms, which is one of the contributions of our work. This is feasible due to the statistical nature of the Spearman's non-parametric correlation coefficient involved in the computation of the proposed SBM measure to discriminate genes and conditions. This bicluster pattern coherence degree must also be verified in the genes and the conditions independently in order to be able to remove genes and conditions which do not reach the required level of coherence degree.

Therefore, the computation of the bicluster pattern coherence degree only requires to recover previously computed Spearman's nonparametric correlation coefficients to obtain the value of the SBM. These coefficients are arranged into two matrices. As explained previously, the first matrix is the genes correlation matrix which reflects the quality of the relationships between genes based on the correlation value computed. The second one is the conditions correlation matrix which reflects the quality of the relationships between conditions based on the same computation.

Both matrices reflect the coherence degree of the bicluster associated with the selected genes and with the conditions.

Based on these matrices, we perform the statistical tests associated to the Spearman's coefficient using the traditional Bonferroni method to adjust significance thresholds for multiple testing. In our case we fixed $\alpha = 0.05/n$ being n the number of correlation coefficients associated to the different paired genes/conditions combinations. This value depends on the number of genes and conditions.

The result of applying this method to the set of artificial bicluster coherence patterns confirms that the selection of the whole set of genes and conditions performed by the SBM fulfills the minimum statistical requirements of quality.

In order to clarify this process, let us take an example with a bicluster with 4 genes and 5 conditions (see Fig. 20). Based on this bicluster we compute the correlation coefficients between each pair of genes and conditions. After this

Table 11 – Summary of the results obtained by MSR for the biclustering specific quality indexes.

Pattern	Genes			Conditions		
	BLA_G	BLP_G	BLE_G	BLA_C	BLP_C	BLE_C
# 1	61.00 ± 0.00	61.00 ± 0.00	100.00 ± 0.00	35.00 ± 0.00	35.00 ± 0.00	100.00 ± 0.00
# 2	63.00 ± 0.00	63.00 ± 0.00	100.00 ± 0.00	35.00 ± 0.00	35.00 ± 0.00	100.00 ± 6.00
# 3	56.00 ± 0.00	66.00 ± 0.00	100.00 ± 0.00	40.00 ± 0.00	40.00 ± 0.00	100.00 ± 0.00
# 4	71.00 ± 0.00	61.00 ± 0.00	100.00 ± 0.00	40.00 ± 0.00	40.00 ± 0.00	100.00 ± 0.00
# 5	69.00 ± 0.00	69.000 ± 1.00	100.00 ± 5.00	35.00 ± 0.00	35.00 ± 0.00	100.00 ± 7.00
# 6	72.00 ± 0.00	72.00 ± 0.00	100.00 ± 6.00	40.00 ± 0.00	40.00 ± 0.00	100.00 ± 0.00
Average	65.07 ± 5.78	65.21 ± 4.11	100.00 ± 0.00	37.42 ± 2.50	37.42.00 ± 2.50	100.00 ± 0.00

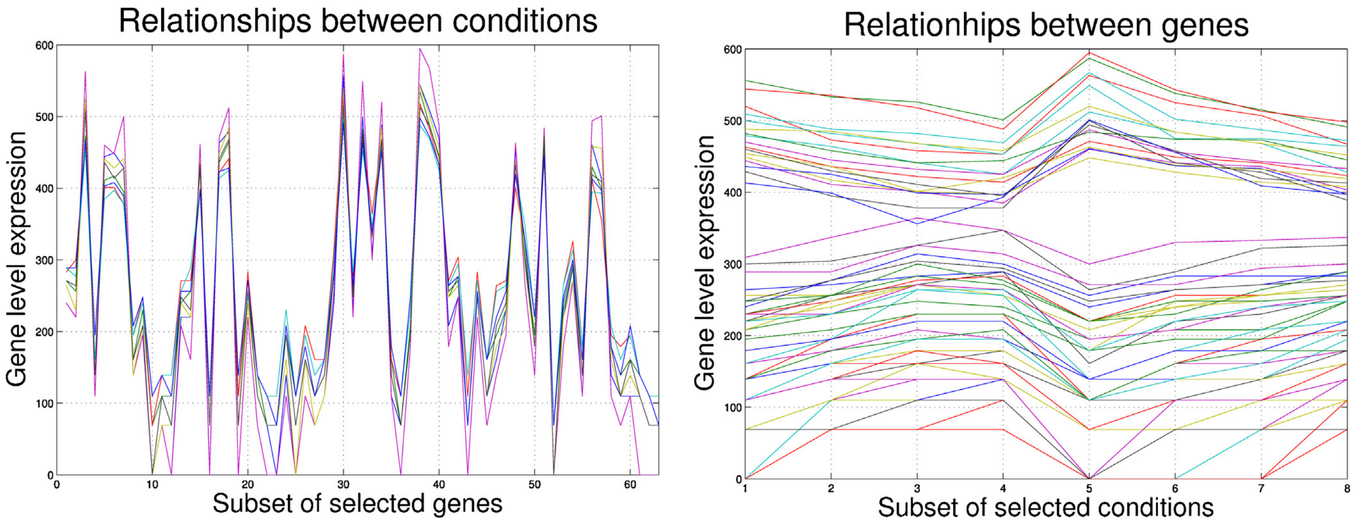


Fig. 10 – First real microarray: coherence evolution patterns detected.

Table 12 – Significant GO terms (function, process, component) of the selected bicluster.

Molecular function	Biological process	Cellular component
Structural constituent of ribosome	Translation	Cytosol
17 Genes	19 Genes	20 Genes
GO-e 8.84	GO-e 2.91	GO-e 4.48
Structural molecule activity	Cellular macromolecule biosynthetic	Cytosolic ribosome
19 Genes	29 Genes	17 Genes
GO-e 8.59	GO-e 2.36	GO-e 11.17
	Macromolecule biosynthetic process	Cytosolic part
	29 Genes	17 Genes
	GO-e 2.35	GO-e 9.98
		Ribosome
		19 Genes
		GO-e 8.21
		Ribosomal subunit
		17 Genes
		GO-e 9.01
		Cytosolic large ribosomal subunit
		10 Genes
		GO-e 6.14
		Cytosolic small ribosomal subunit
		7 Genes
		GO-e 4.48
		Large ribosomal subunit
		10 Genes
		GO-e 4.60
		Small ribosomal subunit
		7 Genes
		GO-e 2.84
		Ribonucleoprotein complex
		20 Genes
		GO-e 3.96
		Macromolecular complex
		Genes 34
		GO-e 3.48
		Intracellular non-membrane-bounded organelle
		Genes 25
		GO-e 3.29
		Non-membrane bounded organelle
		Genes 25
		GO-e 3.29

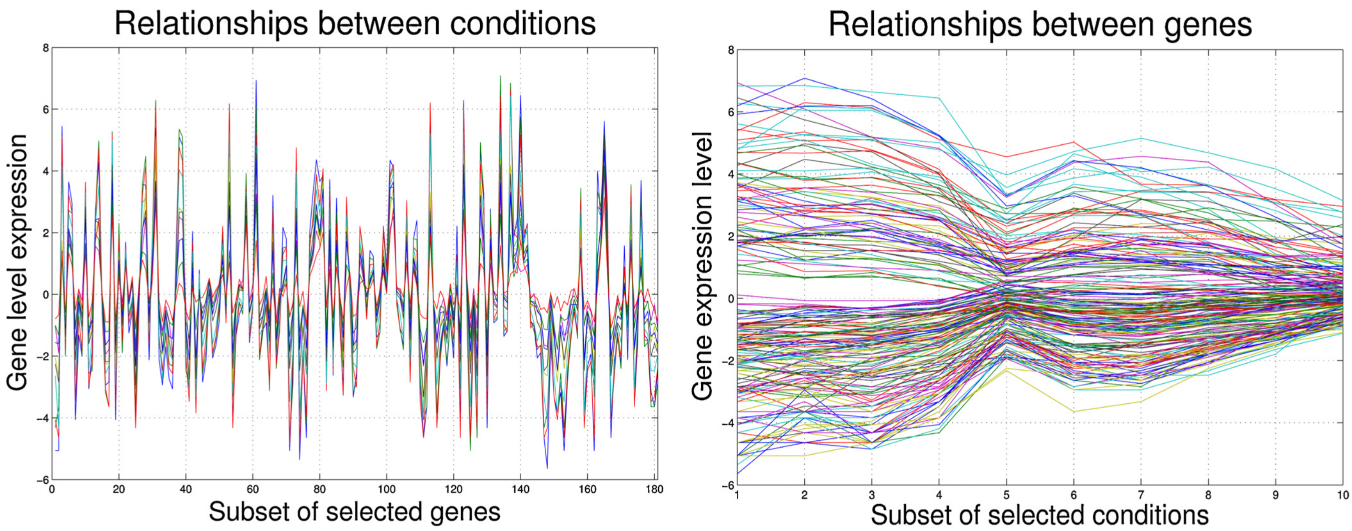


Fig. 11 – Second real microarray: coherence evolution patterns detected.

process we obtain two matrices, the first one (Fig. 20) is the gene correlation matrix which contains all correlation coefficients between genes, the second one (Fig. 20) is the condition correlation matrix which contains all correlation coefficients between conditions.

Based on these matrices we compute the statistical significance of each Spearman's correlation correlation of the correlation matrix: the result is composed of two matrices which contain the p-values associated with the genes and the conditions. The filtering process is based on these values, that is, when a gene or condition does not show statistical significant correlation for all genes/conditions.

In this example, based on $\alpha = 0.10$, gene 4 and condition 5 are filtered because all its p-values do not reach the minimum statistical significance.

In summary, our proposed bicluster pattern coherence degree assessment method provides additional statistical guarantees about the subset of genes prior to any kind of

Table 13 – Significant GO terms (function, process, component) of the selected bicluster.

Molecular function	Biological process	Cellular component
Catalytic activity	Cellular process	Cell part
76 Genes	163 Genes	176 Genes
GO-e 2.11	GO-e 12.39	GO-e 16.33
	Metabolic process	Cell
	123 Genes	177 Genes
	GO-e 3.49	GO-e 16.31
	Cellular metabolic process	Intracellular
	116 Genes	162 Genes
	GO-e 2.86	GO-e 8.16
	Biosynthetic process	Intracellular part
	76 Genes	161 Genes
	GO-e 2.21	GO-e 7.85
	Celullar bionsynthetic process	Intracelullar organelle
	74 Genes	134 Genes
	GO-e 2.01	GO-e 4.60
		Organelle
		134 Genes
		GO-e 4.59
		Membrane-bounded organelle
		121 Genes
		GO-e 2.81
		Intracelullar membrane-bounded organelle
		121 Genes
		GO-e 2.81
		Cytoplasm
		120 Genes
		GO-e 2.01

biology assessment. This statistical assessment can be viewed as an additional quality control in the process of biclustering analysis. This type of analysis is not possible with the classical approach of the MSR measure.

5.2. *Second experimentation phase: real gene expression datasets*

The aim of this second phase of the experimentation process is to assess the proposed measure on two real DNA microarrays datasets based on the well-known Yeast organism. The first dataset is a benchmark and widely used microarray in the biclustering literature [6]. The second is a microarray used to study the response of Yeast to environmental changes [71].

This assessment is based on the common practice found in the biclustering literature. It has been applied to both microarray datasets and it is formed by the following sequence of phases:

- Description. Characteristics of the microarray dataset, that is, the number of genes and conditions.
- Search and selection of biclusters. In this step we perform a search process of biclusters by means of the exposed $UMDA_d$ evolutionary method, using the proposed SBM measure as fitness function. This process is repeated 20 times in order to check the stability of the experiments and their replicability. The size of the population in each search process is fixed to $2.5(N+M)$ individuals, a minimal size in order to check whether the algorithm is able to provide significative results in the worst conditions. Each run of the algorithm is configured to output a single bicluster, resulting in a final set of 20 different biclusters.

  Next, we proceed with the assessment of the stability degree (see Eq. (19)) of the obtained biclusters. In order to consider the resulting subset of genes as acceptable, it is important that the computed values are near zero.
- Selection of genes. This is a new step which aims to collect a subset of genes which are common to all biclusters.
- Statistical assessment of the quality of a bicluster. This is a step which aims to filter low quality biclusters and therefore to detect spurious bicluster coherence models. To this aim, we apply the procedure described in Section 5.1.5, that is, to perform multiple statistical testing to the bicluster formed by all selected common genes and conditions to all biclusters.

  The application of statistical tests to detect spurious bicluster coherence models has been accomplished on each combination of pairs of genes and on each combination of pairs of conditions. So, the results of these tests are used to discard or accept a gene or condition. A gene or condition is discarded when all the results of the tests associated to the same element are negative.

  Therefore, the results of the tests determinate whether the subset of genes is statistically correlated among them and therefore it is possible to continue with the biological assessment. Otherwise, the subset of genes is rejected. The same is applied to selected conditions.
- Biological assessment. Any artificial scenario is inevitably biased regarding the underlying model and only reflects certain aspects of the biological reality [13]. Therefore, the biological relevance of the obtained biclusters with the selected genes is assessed by means of their functional enrichment. This enrichment is based on the information of the Gen Ontology (GO) annotations [72]. In this database genes are assigned to three structured, controlled vocabularies (ontologies) that describe the gene products in terms of associated biological processes, components and molecular functions in a species-independent manner. In our case, we have analyzed the results based on the biological processes, and component and molecular functions.

  The degree of enrichment has been measured using a cumulative hypergeometric distribution, which involves the probability of observing the number of genes from a particular GO category (i.e., function, process, component) within each bicluster [13]. The p-values are calculated for each functional category, which shows the statistical degree with which they match the different GO categories. Note that a smaller p-value is indicative of a better match. The Bonferroni correction is adopted for multiple comparison hypothesis. The enrichment is finally computed based on the corrected p-values as follows:

  $$Enrichment = -log_{10}(p - value) \qquad (27)$$

  Based on this widely used approach, we have selected the Saccharomyces Genome Database [72] to perform the analysis of functional enrichment. In order to perform the analysis, the common practice found in the literature has been adopted and therefore we have selected information from three databases: the molecular function, the biological process, and the cellular component.

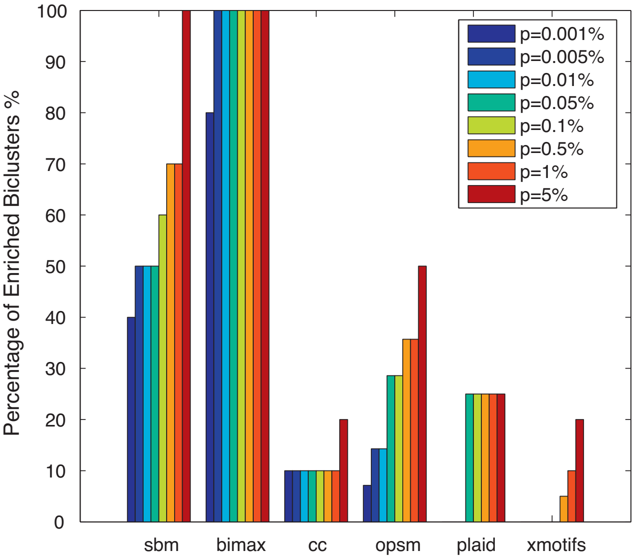


Fig. 12 – Comparison of different biclustering algorithms from Yeast data set.

5.2.1. *First microarray dataset: benchmark yeast DNA microarray*

Description. The first microarray is formed by 2884 genes under 17 conditions and it has been used in many works to assess the performance of biclustering methods [6].

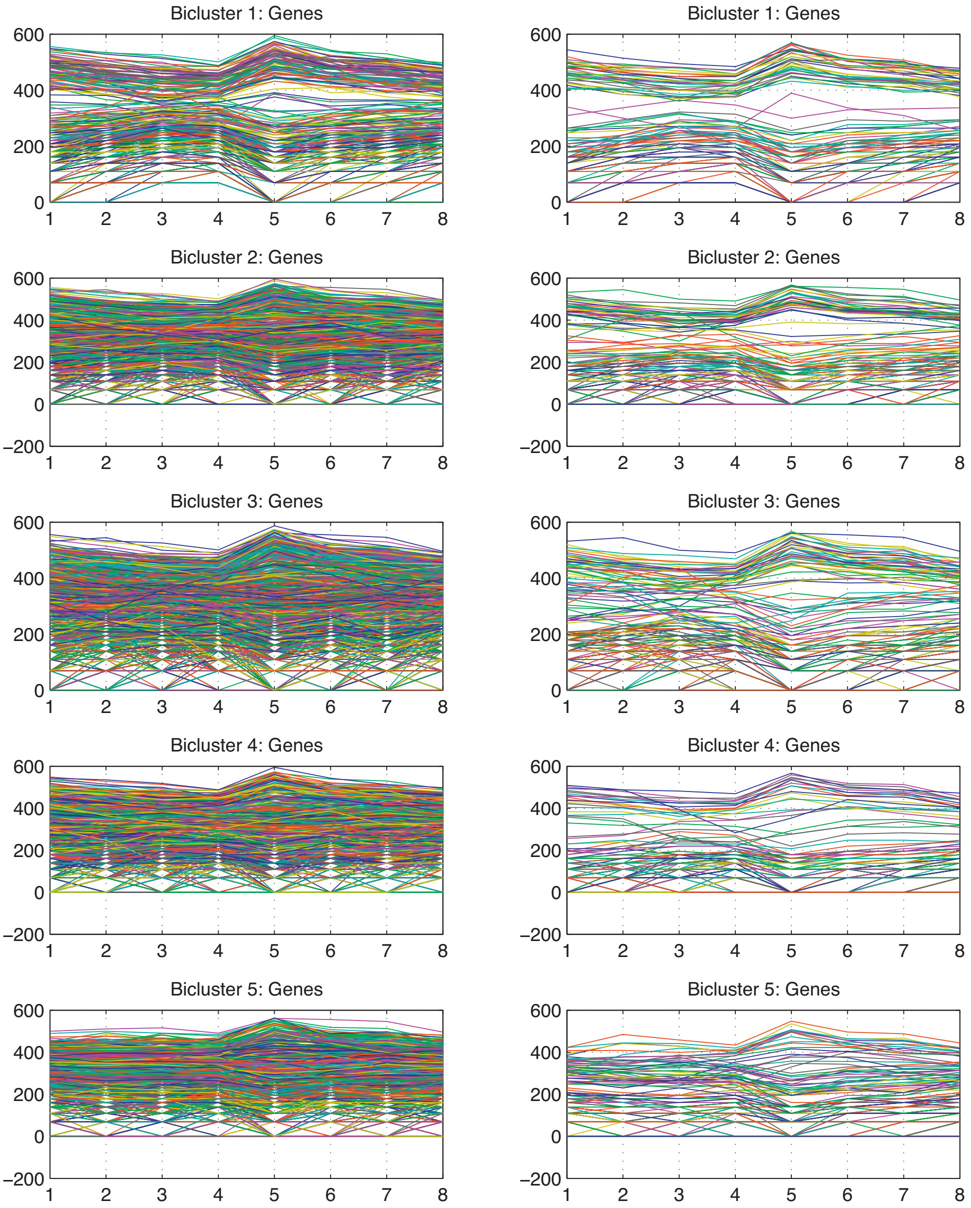


Fig. 13 – Biclustering with statistical filtering applied to the genes. Part I.

Search and selection of biclusters. By means of the application of the proposed procedure, the result of this step is a bicluster of 133 genes. Next, we have computed the variability of the bicluster found along different executions of the algorithms providing a result of 0.032. That is, only 3.2% of the genes and conditions differ along 20 different search processes. The 96.8% of the selected genes and conditions are common in all found 20 biclusters. This result is satisfactory considering the high number of genes and conditions.

Selection of genes. Starting from the group of 133 genes detected we proceed with the selection of those genes which are well documented in the Saccharomyces Genome Database website [72], showing that each stable gene with regard to the biological function, the biological process and the biological component. Therefore, all the genes were accepted to proceed with the next step.

Statistical assessment of the quality of a bicluster. The result of this filtering step based on the application of the exposed statistical tests to each combination of pairs of genes and conditions is that 69 genes were rejected. The final bicluster is a subset of 64 genes which shows a reliable coherence in statistical terms (see Fig. 10). The algorithm has been able to detect with high reliability different kinds of coherence patterns such as shifting, scaling and even inverse coherence patterns (see Fig. 10). The reliability of the inverse coherence patterns is specially important in some diseases such as some kinds of cancer and tumors [15–17].

It is also remarkable that the algorithm finds almost perfect coherence patterns in conditions (see Fig. 10) taking into account the 64 genes.

The final bicluster shows an almost perfect reliability in the coherence of the conditions and a rich set of reliable coherence

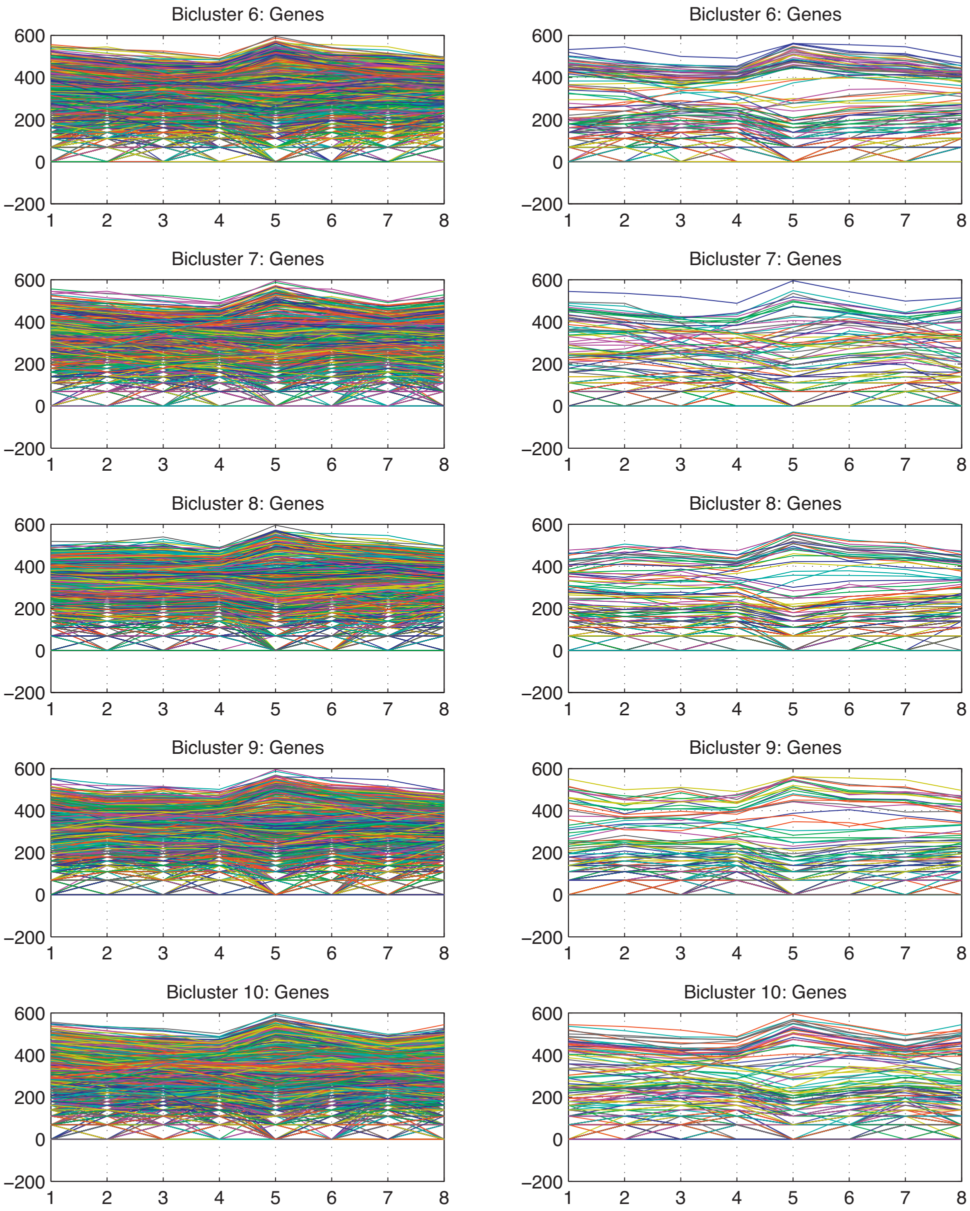


Fig. 14 – Biclustering with statistical filtering applied to the genes. Part II.

patterns in genes. That is, all genes has passed the statistical tests with solvency.

It is therefore possible to proceed with the biological assessment.

Biological assessment. The result of the analysis is shown in Table 12. The columns show the significant GO terms (function, process, component) used to describe the selected collection of genes. In each cell of the table the rows show the specific biological description, the number of involved genes, and the computed biological enrichment degree.

Firstly, the results of the molecular function analysis show two large groups of genes with a high degree of enrichment. These groups are associated with the structural constituent of ribosome (GO-e 8.84) and structural molecule activity (GO-e 8.59).

Secondly, the results of the biological process show that there are clear processes identified by several groups of genes: translation process (GO-e 2.91), cellular macromolecule biosynthetic (GO-e 2.36) and the macromolecule biosynthetic process (GO-e 2.35).

Finally, the results of the cellular component show a large group of genes involved in the following processes: Cytosol (GO-e 4.48), Cytosolic Ribosome (GO-e 11.17), Cytosolic part (GO-e 9.98), Ribosome (GO-e 8.21), Ribosomal subunit (GO-e 9.01), Cytosolic large ribosomal subunit (GO-e 6.14), Cytosolic small ribosomal subunit (GO-e 4.48), Large ribosomal subunit (GO-e 4.60), Small large ribosomal subunit (GO-e 2.84), Ribonucleoprotein complex (GO-e 3.96), Macromolecular complex (GO-e 3.48), Intracellular non-membrane bounded organelle (GO-e 3.29), Intracellular non-membrane-bounded organelle

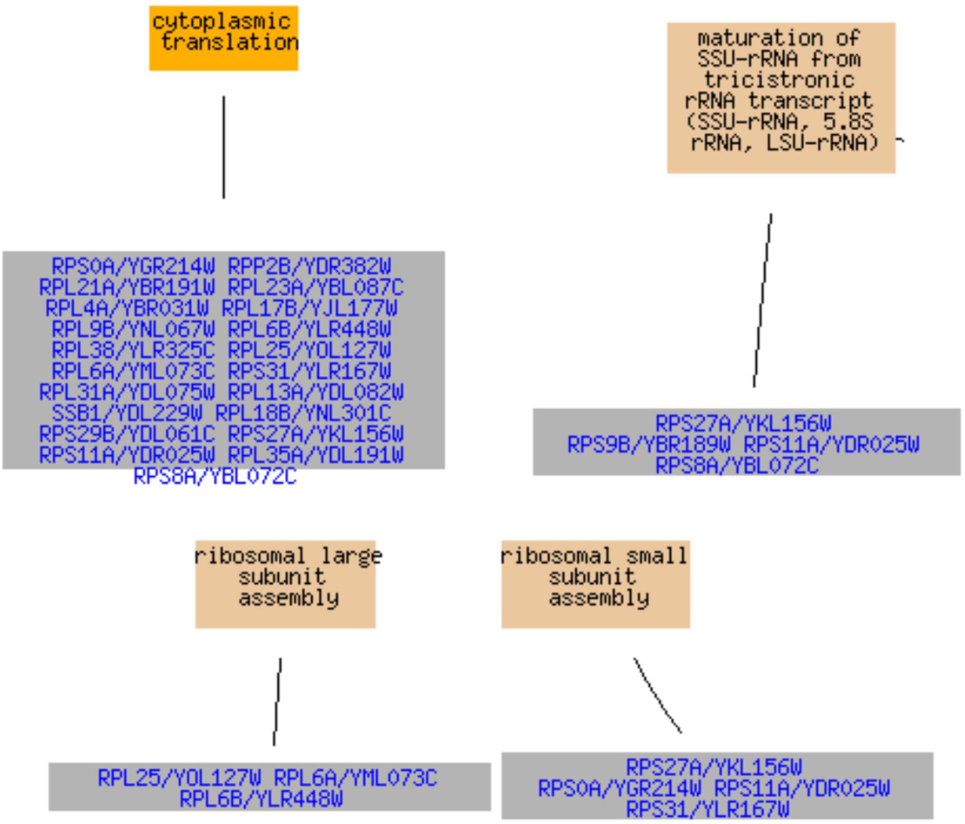

cytoplasmic translation
maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
RPS0A/YGR214W RPP2B/YDR382W
RPL21A/YBR191W RPL23A/YBL087C
RPL4A/YBR031W RPL17B/YJL177W
RPL9B/YNL067W RPL6B/YLR448W
RPL38/YLR325C RPL25/YOL127W
RPL6A/YML073C RPS31/YLR167W
RPL31A/YDL075W RPL13A/YDL082W
SSB1/YDL229W RPL18B/YNL301C
RPS29B/YDL061C RPS27A/YKL156W
RPS11A/YDR025W RPL35A/YDL191W
RPS8A/YBL072C
RPS27A/YKL156W
RPS9B/YBR189W RPS11A/YDR025W
RPS8A/YBL072C
ribosomal large subunit assembly
ribosomal small subunit assembly
RPL25/YOL127W RPL6A/YML073C
RPL6B/YLR448W
RPS27A/YKL156W
RPS0A/YGR214W RPS11A/YDR025W
RPS31/YLR167W

Fig. 15 – Detailed biological functions of the first bicluster.

(GO-e 3.29) and the Non-membrane bounded organelle (GO-e 3.29).

The results show the consistence of the identified groups of genes with regard to their involvement in the different biological parts. However, it is necessary to remark the high degrees of the GO-enrichment term obtained by the groups of genes identified in the cellular component and the molecular function.

These results show that SBM provides competitive results with regard to biological enrichment. We have obtained a large subset of genes compared with other approaches based on MSR [13].

5.2.2. *Second microarray dataset: response of Yeast cells to environmental changes DNA microarray*

Description. This second microarray is the result of the study of genomic expression programs in the response of Yeast cells to environmental changes [71]. This study has allowed to measure changes in transcript levels over time for almost every yeast gene, as cells responded to temperature shocks,

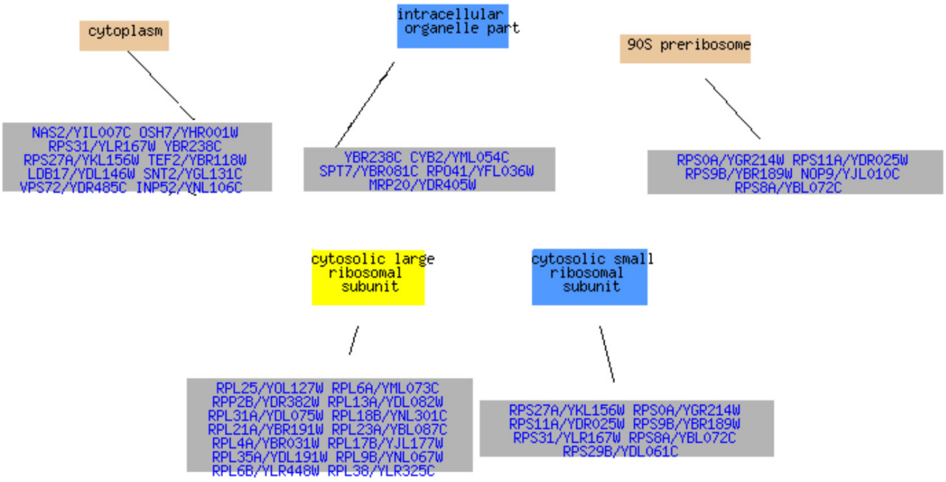

cytoplasm
intracellular organelle part
90S preribosome
NAS2/YIL007C OSH7/YHR001W
RPS31/YLR167W YBR238C
RPS27A/YKL156W TEF2/YBR118W
LDB17/YDL146W SNT2/YGL131C
VPS72/YDR485C INP52/YNL106C
YBR238C CYB2/YML054C
SPT7/YBR081C RPO41/YFL036W
MRP20/YDR405W
RPS0A/YGR214W RPS11A/YDR025W
RPS9B/YBR189W NOP9/YJL010C
RPS8A/YBL072C
cytosolic large ribosomal subunit
cytosolic small ribosomal subunit
RPL25/YOL127W RPL6A/YML073C
RPP2B/YDR382W RPL13A/YDL082W
RPL31A/YDL075W RPL18B/YNL301C
RPL21A/YBR191W RPL23A/YBL087C
RPL4A/YBR031W RPL17B/YJL177W
RPL35A/YDL191W RPL9B/YNL067W
RPL6B/YLR448W RPL38/YLR325C
RPS27A/YKL156W RPS0A/YGR214W
RPS11A/YDR025W RPS9B/YBR189W
RPS31/YLR167W RPS8A/YBL072C
RPS29B/YDL061C

Fig. 16 – Detailed biological components of the first bicluster.

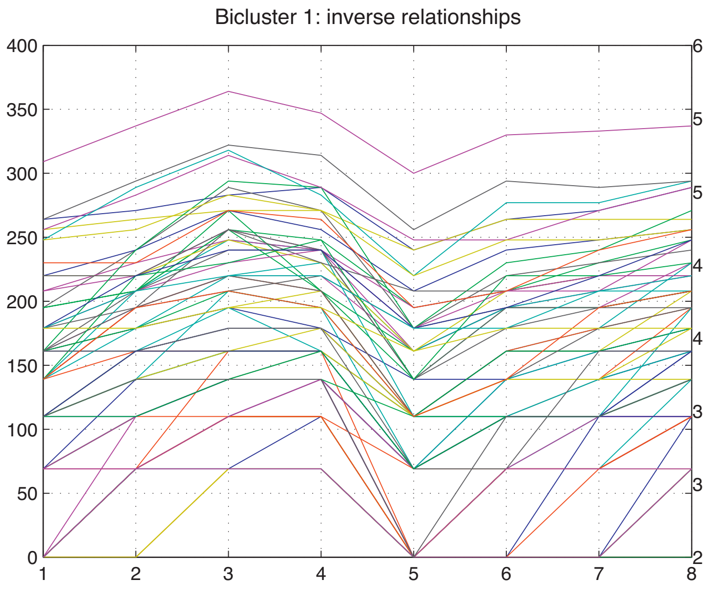


Fig. 17 – Bicluster 1: direct relationships.

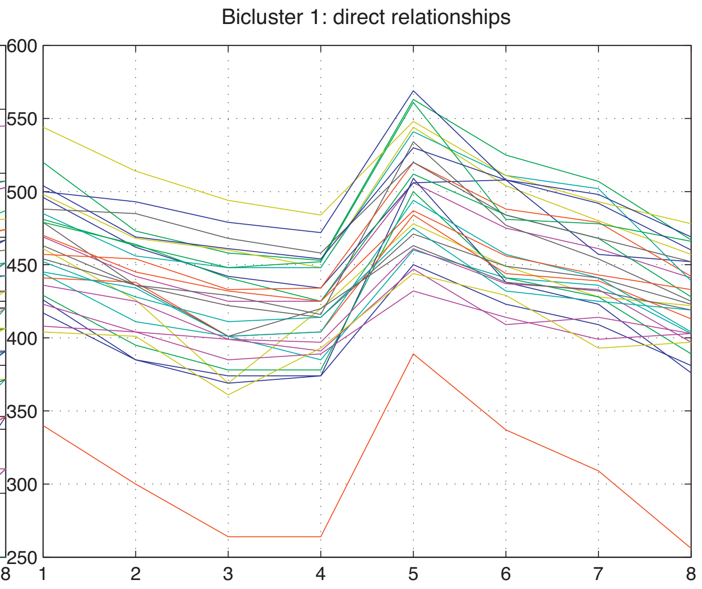


Fig. 18 – Bicluster 1: inverse relationships.

hydrogen peroxide, the superoxide-generating drug menadione, the sulfhydryl-oxidizing agent diamide, the disulfide-reducing agent dithiothreitol, hyper- and hypo-osmotic shock, amino acid starvation, nitrogen source depletion, and progression into stationary phase.

In the initial microarray, undocumented genes were removed and a subset of all conditions was selected: temperature shocks evolution (5 conditions), the sulfhydryl-oxidizing agent diamide (8 conditions) and the amino acid starvation (5 conditions). This is carried out to check whether the proposed algorithm is able to detect specific subsets of genes and conditions. The result is a microarray formed of 2601 genes and 18 conditions.

Search and selection of biclusters. The result of this step is a bicluster composed of 178 genes. Next, we have computed the variability of the bicluster found along different executions of the algorithms providing a result of 0.002. That is 0.2% of the genes and/or conditions are different genes along 20 repetitions from different starting points. This value shows a low variability and therefore a good result in terms of stability of the algorithm.

Selection of genes. We have performed the selection of the common genes to all biclusters found terminating with a collection of 178 genes. From this collection we proceed with the filtering of the undocumented genes, while maintaining the documented ones. The result is that all genes were documented and therefore it was not necessary to discard any gene and it is possible to proceed with the next step.

Statistical assessment of the quality of a bicluster. In this second microarray, the result of this filtering step shows that only one gene (see the graphical results of the coherence patterns in Fig. 11). As in the previous case, the approach has been able to detect different kinds of coherence patterns with high reliability in statistical terms, including the most complex coherence patterns, the inverse coherence patterns.

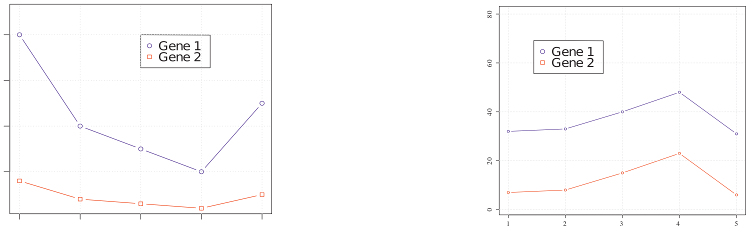
Biological assessment. The result of the analysis is shown in Table 13. The columns show the significant GO terms (function, process, component) used to describe the collection of genes. In each cell, the rows show the specific biological description, the number of involved genes, and the computed biological enrichment degree.

The results show that there is a subset of genes involved in catalytic activity, which is expected under the experimental conditions applied. So, in order to verify these results, it is necessary to consider the results of the original paper [71]. It shows that there is a specific subset involved in all experimental conditions. That is, this subset is involved in the *Environmental Stress Response (ESR)*. The result is that only 1.6% of the genes detected are in the ESR, while the rest of the genes are specific to the heat shock and to the sulfhydryl-oxidizing agent diamide and not to the starvation condition. This is coherent with the results presented in the original paper.

6. Experimentation: comparative analysis

The performance of the proposed approach has been compared with a set of state of the art biclustering methods such as BiMax [22], CC [6], OPSM [34], Plaid [39] and xMotifs [73] for the Yeast microarray data set. Following the methodology exposed in [22] the performance of all algorithms is evaluated with regard to the percentage of biclusters enriched by any Gene Ontology Consortium (GO) category at different levels of significance. GO [71] is used to know whether a group of genes belonging to a bicluster shows significant enrichment with regard to a specific GO term. Although there are different tools to analyze GO enrichment, a recently published tool AGO [74] has been used to study the percentage of significant biclusters obtained. The enrichment of each group of genes with regard to a specific GO term is established by the returned *p*-value. A bicluster is said to be *overrepresented* in a functional category if its *p*-value is small or below the preset threshold. The percentage of overrepresented biclusters in one or

$$\begin{array}{l} \text{Gene 1} \\ \text{Gene 2} \end{array} \begin{bmatrix} 10x8 & 10x4 & 10x3 & 10x2 & 10x5 \\ 2x8 & 2x4 & 2x3 & 2x2 & 2x5 \end{bmatrix} \qquad \begin{array}{l} \text{Gene 1} \\ \text{Gene 2} \end{array} \begin{bmatrix} 30+0 & 30+3 & 30+10 & 30+18 & 30-1 \\ 5+0 & 5+3 & 5+10 & 5+18 & 5-1 \end{bmatrix}$$



Gene 1	Gene 2	Gene 1 - Ranks	Gene 2 - Ranks	D
10x2	2x2	1	1	0
10x3	2x3	2	2	0
10x4	2x4	3	3	0
10x5	2x5	4	4	0
10x8	2x8	5	5	0

The ranks remains equal.

Gene 1	Gene 2	Gene 1 - Ranks	Gene 2 - Ranks	D
30-1	5-1	1	1	0
30+0	5+0	2	2	0
30+3	5+3	3	3	0
30+10	5+10	4	4	0
30+18	5+18	5	5	0

The ranks remains equal.

Fig. 19 – Invariability to shifting and scaling.

more GO annotations is used to compare different algorithmic approaches.

In order to perform this comparison with respect to previously cited biclustering algorithms it is necessary to parameterize them to return the same number of biclusters, whenever it is possible. In the case of our proposed SBM algorithm, the overlapping rate is fixed to 50%. The value of this rate could depend on the user's specific previous knowledge of the problem.

The biclusters obtained by the set of algorithms are processed by the AGO tool. Fig. 12 represents the percentage of enriched biclusters for each method in which one or several GO terms are overrepresented for different levels of significance (p-values from $p = 0.001\%$ to $p = 5\%$). The figure shows that for all levels of significance BiMax obtains the best results. This algorithm only searches binary coherence patterns in the space of binary data. Therefore, it is important to consider the level of significance and the capabilities of pattern detection. Complex new patterns are based on real data and they are hard to detect whereas the simple patterns can be expressed with binary data and they are therefore easy to find and detect. In the rest of cases, the SBM approach obtains the best results in all levels of significance. In the worst case, the percentage of enriched bicluster is above the 40% whereas the rest of approaches is very low or inexistent (see Fig. 12).

We have also followed our previously exposed methodology with the set of biclusters provided by the SBM algorithm. This has resulted in a reduction near to 85% of the previously selected genes in all candidate biclusters (see Figs. 13 and 14). These figures show in the first column all biclusters found and in the second column the same set of biclusters after applying a statistical filtering to reduce the set of genes. The figures show graphically a clear reduction in the number of genes in all biclusters. This result confirms that the statistical assessment of the results is a necessary step before proceeding to the biological analysis, showing that the majority of the genes do not fulfill the statistical requirements. As we have advised previously, the bicluster set that an algorithm can find does not necessarily imply that they are good candidates, this needs to be confirmed by means of a set of statistical tests. In general terms, biclustering algorithms do not incorporate biological information when they look for the best bicluster. They are based on mathematical quality measures and the search is performed under these conditions, so these biclusters can be considered as mathematical biclusters. This means that algorithms can find good biclusters but in mathematical terms.

In our experiments we assessed the concept of mathematical bicluster. The result of the experiments is the detection of several biologically related groups of genes involved in the process of translation. Therefore, not only the genes in a bicluster are strongly correlated, but also the main functions of these genes in different biclusters are also correlated (see Fig. 15). The whole set of filtered genes shows the relationship among the different biological functions related with the process of translation in the cell (see Figs. 15 and 16) where different elements are involved such as the small and large subunit, the preribosome and the cytoplasmic translation. The algorithm detects not only the activity of genes within a biological function but the associated biological functions in the same process. This can be useful when the knowledge of the organism is unknown because it can discover biological functions.

On the other hand, Figs. 13, 14 and 16 show different kind of patterns such as scaled patterns, shifting patterns and the proposed pattern the inverse patterns.

For instance, Bicluster 1 (see Fig. 13) can be decomposed in two subsets: the direct relationships subset and the subset covering inverse relationships (see Figs. 17 and 18). Scaling and shifting patterns in both subsets, with independence of the kind of relationship, can be seen. Our experiments have confirmed the capability of detection for these two kinds of

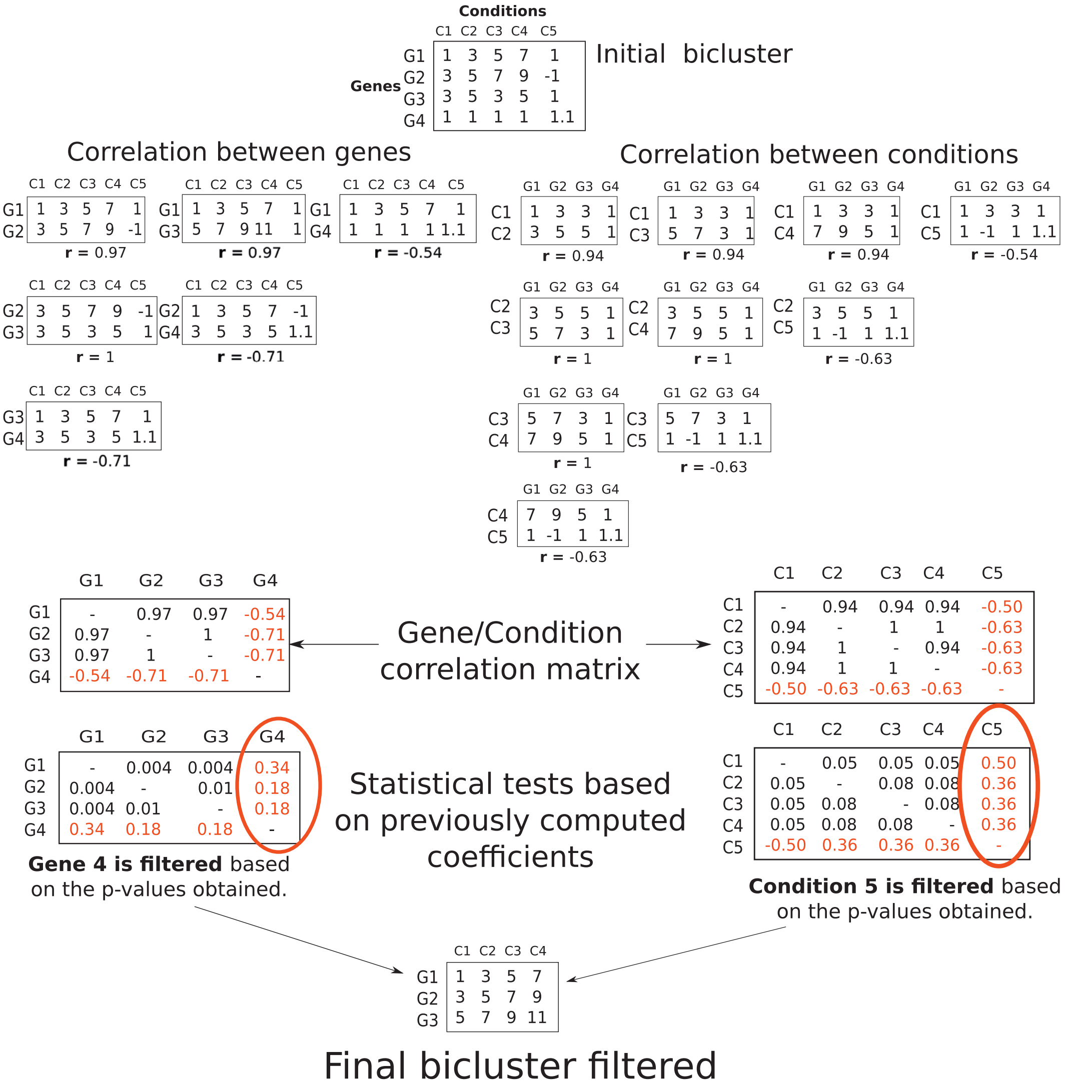

Conditions
Genes
Initial biclusterb
Correlation between genes
Correlation between conditions
r = 0.97
r = 0.97
r = -0.54
r = 1
r = -0.71
r = -0.71
r = 0.94
r = 0.94
r = 0.94
r = -0.54
r = 1
r = 1
r = -0.63
r = 1
r = -0.63
r = -0.63
Gene/Condition correlation matrix
Statistical tests based on previously computed coefficients
Gene 4 is filtered based on the p-values obtained.
Condition 5 is filtered based on the p-values obtained.
Final bicluster filtered

Fig. 20 – Gene/Condition filtering process.

patterns. This confirms that direct and inverse patterns are present in the same biological components for almost all biological biclusters.

The absence of the capability of detection for this kind of patterns can lead, in many cases, to the omission of key genes in the discovery of biological functions. Real cases where the inverse patterns are relevant can be found in different types of cancer and tumors as previously cited [15–17]. The detection of this kind of relationships can, therefore, show more complete biological information in the study of a disease.

In this situation where almost all biclusters contain relevant direct and inverse relationships, MSR has been unable to consider them as good biclusters because the computed value of MSR metric for all biclusters was greater than 300 [6]. This value is considered the quality threshold and therefore the results would been rejectable but not for the proposed SBM algorithm.

Finally, during the course of the analysis two types of mathematical biclusters were found. The first set reveals relationships with a known biological process. However, the second set of biclusters was not linked to any specific biological function, that is the biological functions of this set are classified as “unknown”. This means that we have found new unknown and undocumented biological functions. Considering the results we re-analyzed the relationships mathematically using a different toolbox but the result was the same. The result of the analysis with a different toolbox confirms that the correlation between genes and conditions is considered as “perfect correlation” because the values obtained were $\rho = 1$ or $\rho = -1$ for each accepted gene. This means that the detected relationships between genes by the algorithm were correct and therefore they suggest the initial discovering of new undocumented biological relationships. Subsequently, the results confirm the validity of the measure and the search mechanism to find quality biclusters.

7. Conclusions

In the present work we have shown a competitive alternative to the widely used MSR metric for biclustering is presented: the Spearman’s Biclustering Measure (SBM). In contrast to

MSR, this measure has several advantages: a lower rate of false positives, and the possibility to detect inverse patterns. The drawbacks detected is the high false negatives rate.

From the point of view of the biological assessment, the results confirm that SBM is able to detect groups of genes that perform related biological functions, as well as inverted biological functions. It therefore allows to detect a large variety of biological functions than classical MSR.

We also present a new set of quality indexes for the analysis and comparison of biclustering measures. This new set if indexes is used in combination with a predefined group of reference artificial DNA microarray datasets that fulfills the classical coherence patterns characteristics. This allows to quantify a biclustering measure in a fair and honest way with different quality indexes such as the accuracy, the false positive rate, the false negative rate and the stability degree.

We hope that this new measure and the set of proposed biclustering quality indexes actively helps in the data analysis and knowledge discovery process in gene expression experiments.

Conflict of interest

No competing financial interests exists.

Acknowledgments

This work has been partially supported by the Etortek, Saiotek and Research Groups 2007–2012 (IT-242-07) programs (Basque Government), TIN2010-14931 and Consolider Ingenio 2010 - CSD2007-00018 projects (Spanish Ministry of Education and Science) and COMBIOMED network in computational biomedicine (Carlos III Health Institute) and INF07/39.

Appendix A.

The correlation based on ranks is a tool which is invariable to scaling and shifting problems. Following, we will describe the necessary steps to demonstrate this.

First, the computation of the Spearman's correlation between the samples corresponding to two variables is based on the differences between the ranks of the samples. Therefore, in order to demonstrate the invariability between two sets of samples where one of them is a modified version of the other, it is only necessary to demonstrate that the associated ranks of the modified version do not change. So, to show that the ranks do not change, it is only necessary to demonstrate that the relative order of the samples does not change.

In all cases we will suppose that there is a set of samples x_1, . . ., x_n and its modified version $y_1, \ldots, y_n$ associated respectively to X and Y variables. The relative order is expressed by means of the following inequations $x_1^* \leq \ldots \leq x_n^*$, where x_i^* represents the ith ordered from the smallest sample to highest of the original set.

Theorem 1 ((Invariability to the shifting)). *Let's suppose a shifting pattern with regard to an original set, the shifting pattern can be expressed as $y_1 = x_1 + \alpha, \ldots, y_n = x_n + \alpha$, being $x_1, \ldots, x_n$ the original pattern and $\alpha \in R$. If $x_1^* \leq \ldots \leq x_n^*$ then $x_1^* + \alpha \leq \ldots \leq x_n^* + \alpha$.*
Proof. This is trivial due to the fact that the addition or the subtraction of a real number α does not change an inequation.

Theorem 2 ((Invariability to the scaling)). *Let's suppose a shifting pattern with regard to an original set, the shifting pattern can be expressed as $y_1 = x_1\alpha, \ldots, y_n = x_n\alpha$, being $x_1, \ldots, x_n$ the original pattern and $\alpha \in R^+$. If $x_1^* \leq \ldots \leq x_n^*$ then $x_1^*\alpha \leq \ldots \leq x_n^*\alpha$.*
Proof. This is trivial due to the fact that the product of a real positive number α does not change an inequation.
In the case of a real negative number $\alpha \in R^-$ the whole inequation changes, the ranks are completely inverted, and therefore the correlation computes a negative number but equal in absolute value to the previous correlation coefficient. The SBM always computes the absolute value of the correlation coefficient, therefore in these cases there is no negative effect.

REFERENCES

[1] J. Quackenbush, Computational analysis of microarray data, Nat. Rev. Genet. 6 (2001) 418–427.
[2] A. Ben-Dor, R. Shamir, Z. Yakhini, Clustering gene expression patterns, J. Comput. Biol. 6 (3/4) (1999) 281–297.
[3] J.N. Morgan, J.A. Sonquist, Problems in the analysis of survey data, and a proposal, J. Am. Stat. Assoc. 58 (302) (1963) 415–434.
[4] J.A. Hartigan, Direct clustering of a data matrix, J. Am. Stat. Assoc. 67 (337) (1972) 123–129.
[5] B. Mirkin, Mathematical Classification and Clustering (Nonconvex Optimization and Its Applications), Kluwer, Hingham, MA, USA, 1996.
[6] Y. Cheng, G.M. Church, Biclustering of expression data, in: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, AAAI Press, San Diego, USA, 2000, pp. 93–103.
[7] J. Aguilar, F. Divina, GA-based approach to discover meaningful biclusters, in: Proceedings of the 2005 Genetic and Evol. Comput. Conference, ACM Press, Washington, DC, USA, 2005, pp. 473–474.
[8] S. Bleuler, A. Prelić, E. Zitzler, An EA framework for biclustering of gene expression data, in: Proceedings of the 2004 Congress on Evol. Comput., vol. 1, IEEE, 2004, pp. 166–173.
[9] K. Bryan, P. Cunningham, N. Bolshakova, Biclustering of expression data using simulated annealing, in: Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems, IEEE Computer Society, 2005, pp. 383–388.
[10] C. Cano, L. Adarve, J. López, A. Blanco, Possibilistic approach for biclustering microarray data, Comput. Biol. Med. 37 (2007) 1426–1436.
[11] H.-S. Chiu, T.-W. Huang, C.-Y. Kao, Biclustering gene expression data by using iterative genetic algorithm, in: Proceedings of the 9th Annual International Conference on Research in Computational Molecular Biology, 2005.
[12] H.-S. Chiu, T.-W. Huang, C.-Y. Kao, Identifying biclusters by genetic algorithm, in: Proceedings of the 13th International Conference on Intelligent Systems for Molecular Biology, 2005.
[13] S. Mitra, H. Banka, Multi-objective evolutionary biclustering of gene expression data, Pattern Recogn. 39 (12) (2006) 2464–2477.
[14] J. Aguilar, Shifting and scaling patterns from gene expression data, BMC Bioinform. 21 (20) (2005) 3840–3845.

[15] G. Fontanini, D. Bigini, A. Mussi, M. Lucchi, C. Angeletti, Bcl-2 protein: a prognostic factor inversely correlated to p53 in non-small-cell lung cancer, Br. J. Cancer (1995) 1003–1007.

[16] N. Kopelman, D. Lancet, I. Yanai, Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms, Nat. Genet. (2005) 588–589.

[17] V. Subbarayan, X. Xu, J. Kim, P. Yang, A. Hoque, A. Sabichi, N. Llansa, G. Mendoza, C. Logothetis, R. Newman, S. Lippman, D. Menter, Inverse relationship between 15-liposygenase-2 and PPAR-γ gene expression in normal epithelia compared with tumor epithelia, Neoplasia (2005) 280–293.

[18] S.C. Madeira, A.L. Oliveira, Biclustering algorithms for biological data analysis: a survey, IEEE/ACM Lect. Notes Comput. Sci. 1 (2004) 24–45.

[19] E. Bengoetxea, P. Larrañaga, I. Bloch, A. Perchant, C. Boeres, Inexact graph matching by means of estimation of distribution algorithms, Pattern Recogn. 35 (12) (2002) 2867–2880.

[20] I. Inza, P. Larrañaga, R. Etxeberria, B. Sierra, Feature subset selection by Bayesian network-based optimization, Artif. Intell. 123 (1) (2000) 157–184.

[21] P. Larrañaga, R. Etxeberria, J. Lozano, J. Peña, Combinatorial optimization by learning and simulation of Bayesian networks, in: Proceedings of the 16th Annual Conference on Uncertainty in Artif. Intell. (UAI-00), Morgan Kaufmann, San Francisco, USA, 2000, pp. 343–352.

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

[23] J. Ahn, Y. Yoon, S. Park, Noise-robust algorithm for identifying functionally associated biclusters from gene expression data, Inform. Sci. 181 (February 2011) 435–449.

[24] C.-J. Wu, S. Kasif, GEMS: a web server for biclustering analysis of expression data, Nucleic Acids Res. 33 (Web-Server-Issue) (2005) 596–599.

[25] M. Koyutürk, W. Szpankowski, A. Grama, Biclustering gene-feature matrices for statistically significant dense patterns, in: Proceedings of the 3rd International Congress IEEE Computational Systems Bioinformatics, IEEE Computer Society, 2004, pp. 480–484.

[26] M. van Uitert, W. Meuleman, L.F.A. Wessels, Biclustering sparse binary genomic data, J. Comput. Biol. 15 (10) (2008) 1329–1345.

[27] W. Ahmad, A. Khokhar, cHawk: an efficient biclustering algorithm based on bipartite graph crossing minimization, in: Proceedings of the 2nd International Workshop on Data Mining and Bioinformatics, 2007.

[28] C. Gallo, J. Carballido, I. Ponzoni, BiHEA: a hybrid evolutionary approach for microarray biclustering, in: K. Guimarães, A. Panchenko, T. Przytycka (Eds.), Advances in Bioinformatics and Computational Biology, vol. 5676 of Lect. Notes Comput. Sci., Springer, Berlin, Heidelberg, 2009, pp. 36–47, http://dx.doi.org/10.1007/978-3-642-03223-3-4.

[29] A. Tanay, R. Sharan, R. Shamir, Discovering statistically significant biclusters in gene expression data, BMC Bioinform. 18 (2002) S136–S144.

[30] G. Li, M. Qin, H. Tang, A.H. Paterson, Y. Xu, QUBIC: a qualitative biclustering algorithm for analyses of gene expression data, Nucleic Acids Res. 37 (15) (2009) e101.

[31] Q. Sheng, Y. Moreau, B. Moor, Biclustering microarray data by Gibbs sampling, BMC Bioinform. 19 (Suppl. 2) (2003), ii196–ii205.

[32] J. Xiao, L. Wang, X. Liu, T. Jiang, An efficient voting algorithm for finding additive biclusters with random background, J. Comput. Biol. 15 (10) (2008) 1275–1293.

[33] X. Zhou, X. Wang, E.R. Dougherty, D. Russ, E. Suh, Gene clustering based on clusterwide mutual information, J. Comput. Biol. 11 (1) (2004) 147–161.

[34] A. Ben-Dor, B. Chor, R. Karp, Z. Yakhini, Discovering local structure in gene expression data: the order-preserving submatrix problem, in: Proceedings of the 6th Annual International Conference on Computational Biology, ACM Press, Washington, DC, USA, 2002, pp. 49–57.

[35] D. Hanisch, A. Zien, R. Zimmer, T. Lengauer, Co-clustering of biological networks and gene expression data, Bioinformatics 18 (Suppl. 1) (2002) 145–154.

[36] S. Hochreiter, U. Bodenhofer, M. Heusel, A. Mayr, A. Mitterecker, A. Kasim, T. Khamiakova, S.V. Sanden, D. Lin, W. Talloen, L. Bijnens, H.W.H. Göhlmann, Z. Shkedy, D.-A. Clevert, FABIA: factor analysis for bicluster acquisition, BMC Bioinform. 26 (12) (2010) 1520–1527.

[37] W. Ayadi, M. Elloumi, J.-K. Hao, A biclustering algorithm based on a bicluster enumeration tree: application to DNA microarray data, BioData Mining 2 (1) (2009) 9.

[38] G. Getz, H. Gal, I. Kela, D.A. Notterman, E. Domany, Coupled two-way clustering analysis of breast cancer and colon cancer gene expression data, BMC Bioinform. 19 (2003) 1079–1089.

[39] L. Lazzeroni, A. Owen, Plaid models for gene expression data, Tech. Rep., Department of Statistics, Stanford University, 2000 March.

[40] C. Tang, L. Zhang, A. Zhang, M. Ramanathan, Interrelated two-way clustering: an unsupervised approach for gene expression data analysis, in: Proceedings of the IEEE 2nd International Symposium on Bioinformatics and Bioengineering Conference, 2001, pp. 41–48.

[41] J. Liu, J. Yang, W. Wang, Biclustering in gene expression data by tendency, in: Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference, IEEE Computer Society, 2004, pp. 182–193.

[42] A. Laegrid, T. Hvidsten, H. Midelfart, J. Komorowski, Predicting gene ontology biological process from temporal gene expression patterns, Genome Res. (2003) 965–979.

[43] P. D'haeseleer, X. Wen, S. Fuhrman, R. Somogyi, Mining the gene expression matrix: inferring gene relationships from large scale gene expression data, in: Proceedings of the Second International Workshop on Information Processing in Cell and Tissues, Plenum Press, New York, NY, USA, 1998, pp. 203–212.

[44] C. Albers, R. Jansen, J. Kok, O. Kuipers, S. van Hijum, Simage: simulation of DNA-microarray gene expression data, BMC Bioinform. 7 (1) (2006) 205.

[45] M. Nykter, T. Aho, M. Ahdesmäki, P. Ruusuvuori, A. Lehmussola, O. Yli-Harja, Simulation of microarray data with realistic characteristics, BMC Bioinform. 7 (1) (2006) 3–49.

[46] A.L. Drobyshev, C. Machka, M. Horsch, M. Seltmann, V. Liebscher, M. Hrabé de Angelis, J. Beckers, Specificity assessment from fractionation experiments (SAFE): a novel method to evaluate microarray probe specificity based on hybridisation stringencies, Nucleic Acids Res. 31 (2) (2003) E1–E11.

[47] Y.H. Yang, S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngai, T.P. Speed, Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, Nucleic Acids Res. 30 (4) (2002) e15.

[48] C. Croux, C. Dehon, Influence functions of the Spearman and Kendall correlation measures, Stat. Methods Appl. 19 (4) (2010) 497–515.

[49] F. Angiulli, E. Cesario, C. Pizzuti, Random walk biclustering for microarray data, Inform. Sci. 178 (6) (2008) 1479–1497.

[50] P. Larrañaga, J. Lozano, Estimation of Distribution Algorithms. A New Tool for Evol. Comput., Kluwer Academic Publisher, New York, 2002.

[51] R. Armañanzas, I. Inza, R. Santana, Y. Saeys, J. Flores, J. Lozano, Y. Peer, R. Blanco, V. Robles, C. Bielza, P. Larrañaga, A review of estimation of distribution algorithms in bioinformatics, BioData Mining 1 (1) (2008) 6.

[52] P. Bosman, D.D. Thierens, Linkage information processing in distribution estimation algorithms, in: Proceedings of the Genetic and Evol. Comput. Conference, Morgan Kaufmann Publishers, San Francisco, 1999, pp. 60–67.

[53] J. Lozano, P. Larrañaga, I. Inza, E. Bengoetxea, Towards a New Evolutionary Computation: Advances on Estimation of Distribution Algorithms, Springer-Verlag, New York, 2006.

[54] H. Mühlenbein, G. Paass, From recombination of genes to the estimation of distributions I. Binary parameters, in: Proceedings of the 4th International Conference on Parallel Problem Solving, vol. 1141 of Lect. Notes Comput. Sci., Berlin, Germany, 1996, pp. 178–187.

[55] M. Pelikan, Hierarchical Bayesian Optimization Algorithm: Toward a New Generation of Evolutionary Algorithms, Springer, Berlin, 2005.

[56] H. Mühlenbein, The equation for the response to selection and its use for prediction, Evol. Comput. 5 (3) (1998) 303–346.

[57] J.S.D. Bonet, C. Isbell, P. Viola, MIMIC: finding optima by estimating probability densities, in: M. Jordan, M. Mozer, M. Perrone (Eds.), Advances in Neural Information Processing, vol. 9 (Denver 1996), MIT Press, Cambridge, 1996.

[58] M. Pelikan, H. MÃŒhlenbein, The bivariate marginal distribution algorithm, Adv. Soft Comput. A: Eng. Des. Manuf. (1999) 521–535.

[59] S. Baluja, S. Davies, Using optimal dependency-trees for combinatorial optimization: learning the structure of the search space, in: Proceedings of the 14th International Conference on Machine Learning, Morgan Kaufmann, Nashville, Tennessee, USA, 1997, pp. 30–38.

[60] R. Santana, E.P. de LeÃn, A. Ochoa, The edge incident model, in: Proceedings of the 2nd Symposium on Artif. Intell., 1999, pp. 352–359.

[61] R. Etxeberria, P. Larrañaga, Global optimization using Bayesian networks, in: Proceedings of the 2nd Symposium on Artificial Intelligence Adaptive Systems, vol. 9, (La Habana), 1999, pp. 332–339.

[62] M. Pelikan, D.E. Goldberg, Proceedings of the 6th International Conference on Parallel Problem Solving from Nature, vol. 1917 of Lecture Notes In Computer Science, Springer-Verlag, London, UK, 2000, pp. 385–394.

[63] M. Pelikan, D.E. Goldberg, F.G. Lobo, A survey of optimization by building and using probabilistic models, Comput. Optim. Appl. 21 (January) (2002) 5–20.

[64] A. Kalousis, J. Prados, M. Hilario, Stability of feature selection algorithms: a study on high-dimensional spaces, Knowl. Inf. Syst. 12 (1) (2007) 95–116.

[65] P. Krizek, J. Kittler, V. Hlavac, Improving stability of feature selection methods, in: Proceedings of the 12th International Conference on Computer Analysis of Images and Patterns, vol. 4673, Springer, Vienna, Austria, 2007, pp. 926–936.

[66] L.I. Kuncheva, A stability index for feature selection, in: Proceedings of the 25th IASTED International Multi-Conference on Artif. Intell. and Applications, ACTA Press, Innsbruck, Austria, 2007, pp. 390–395.

[67] J. Gu, J. Liu, Bayesian biclustering of gene expression data, BMC Genom. 9 (Suppl 1) (2008) S4.

[68] Y. Kluger, R. Basri, J. Chang, M. Gerstein, Spectral biclustering of microarray data: coclustering genes and conditions, Genome Res. 13 (2003) 703–716.

[69] H. Turner, T. Bailey, W. Krzanowski, Improved biclustering of microarray data demonstrated through systematic performance tests, Comput. Stat. Data Anal. 48 (2) (2005) 235–254.

[70] J. Demsar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.

[71] A.P. Gasch, P.T. Spellman, C.M. Kao, O. Carmel-Harel, M.B. Eisen, G. Storz, D. Botstein, P.O. Brown, P.A. Silver, Genomic expression programs in the response of yeast cells to environmental changes, Mol. Biol. Cell. 11 (2000) 4241–4257.

[72] http://www.yeastgenome.org/ http://www.yeastgenome.org/ http://www.yeastgenome.org/ http://www.yeastgenome.org/, “Sccharomyces genome database.”.

[73] T.M. Murali, S. Kasif, Extracting conserved gene expression motifs from gene expression data, in: Pacific Symposium on Biocomputing, 2003, pp. 77–88.

[74] Y.M.K.F.M. Al-Akwaa, An automatic gene ontology software tool for bicluster and cluster comparisons, in: Proceedings of the Sixth Annual IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB’09, IEEE Press, Piscataway, NJ, USA, 2009, pp. 163–167.

[75] S. Busygin, G. Jacobsen, E. Kramer, Double conjugated clustering applied to leukemia microarray data, in: Proceedings of the 2nd International Conference on Data Mining, Workshop on Clustering High Dimensional Data, 2002.

[79] F. Divina, J. Aguilar, Biclustering of expression data with evol. comput., IEEE Trans. Knowl. Data Eng. 18 (5) (2006) 590–602.

[80] B. Pontes, R. Giráldez, J.S. Aguilar, Shifting patterns discovery in microarrays with evolutionary algorithms, in: Proceedings of the 10th International Conference Knowledge-based Intelligent Information and Engineering Systems, vol. 4252 of Lect. Notes Comput. Sc, Springer, Bournemouth, UK, 2006, pp. 1264–1271.

[82] C. Cano, A. Blanco, F. Garcia, F. López, Evolutionary algorithms for finding interpretable patterns in gene expression data, in: Proceedings of the 2006 IADIS International Journal on Computer Science and Information System, vol. V I,2, 2006, pp. 88–99.

[88] D. Jiang, C. Tang, A. Zhang, Cluster analysis for gene expression data: a survey, IEEE Trans. Knowl. Data Eng. 16 (11) (2004) 1370–1386.

[89] H. Cho, I.S. Dhillon, Y. Guan, S. Sra, Minimum sum-squared residue co-clustering of gene expression data, in: Proceedings of the 4th SIAM International Conference on Data Mining, SIAM, 2004, pp. 114–125.

[91] R. Peeters, The maximum edge biclique problem is NP-complete, Discrete Appl. Math. 131 (2003) 651–654.

[92] C. Gonzalez, J. Lozano, P. Larrañaga, Mathematical modeling of UMDA algorithm with tournament selection, Int. J. Approx. Reason 31 (2002) 313–340.

[97] J. Yang, W. Wang, H. Wang, P. Yu, δ-Clusters: capturing subspace correlation in a large data set, in: Proceedings of 18th International Conference on Data Engineering, 2002, pp. 517–518.

[98] J. Yang, H. Wang, W. Wang, P. Yu, Enhanced biclustering on expression data, in: Proceedings of the Third IEEE Symposium on BioInformatics and Bioengineering, 2003, pp. 1–7.

[99] H. Wang, W. Wang, J. Yang, P. Yu, Clustering by pattern similarity in large data sets, in: Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data, ACM, 2002, pp. 394–405.

[100] E. Segal, B. Taskar, A. Gasch, N. Friedman, D. Koller, Rich probabilistic models for gene expression, Bioinformatics 17 (Supple. 1) (2001) S243–S252.

[101] A. Califano, G. Stolovitzky, Y. Tu, Analysis of gene expression microarrays for phenotype classification, in: Proceedings of the 8th International Conference on Computational Molecular Biology, AAAI Press, 2000, pp. 75–85.

[102] J. Liu, W. Wang, Op-cluster: clustering by tendency in high dimensional space, in: ICDM, IEEE Computer Society, 2003, pp. 187–194.

[103] J. Nepomuceno, A. Troncoso, J. Aguilar-Ruiz, Biclustering of gene expression data by correlation-based scatter search, BioData Mining 4 (1) (2011) 3.

[104] S. Das, S.M. Idicula, in: H.R. Arabnia (Ed.), Advances in Computational Biology, vol. 680 of Advances in Experimental Medicine and Biology, Springer, New York, 2011, pp. 181–188, http://dx.doi.org/10.1007/978-1-4419-5913-3-21.

[105] w. Ayadi, M. Elloumi, J.-K. Hao, Iterated local search for biclustering of microarray data, in: T. Dijkstra, E. Tsivtsivadze, E. Marchiori, T. Heskes (Eds.), PRIB, vol. 6282 of Lect. Notes Comput. Sci., Springer, Nijmegen, The Netherlands, 2010, pp. 219–229.

[106] A. Bhattacharya, R.K. De, Bi-correlation clustering algorithm for determining a set of co-regulated genes, BMC Bioinform. 25 (21) (2009) 2795–2801.

[107] E. Yang, P.T. Foteinou, K.R. King, M.L. Yarmush, I.P. Androulakis, A novel non-overlapping bi-clustering algorithm for network generation using living cell array data, BMC Bioinform. 23 (17) (2007) 2306–2313.

[108] H.L. Turner, T.C. Bailey, W.J. Krzanowski, C.A. Hemingway, Biclustering models for structured microarray data, IEEE/ACM Lect. Notes. Comput. Sci. 2 (4) (2005) 316–329.

[109] M. Koyuturkand, W. Szpankowski, A. Grama, Biclustering gene-feature matrices for statistically significant patterns, in: Proceedings of the 2004 Computational Systems Bioinformatics Conference, vol. 16, IEEE Computer Society, 2004, pp. 480–484.

[110] I. Dhillon, S. Mallela, D. Modha, Information-theoretic co-clustering, in: Proceedings of The 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003, pp. 89–98.

[111] X. Gan, A. Liew, H. Yan, Discovering biclusters in gene expression data based on high-dimensional linear geometries, BMC Bioinform. 9 (1) (2008) 209.

[112] H. Zhao, A. Liew, X. Xie, H. Yan, A new geometric biclustering algorithm based on the Hough transform for analysis of large-scale microarray data, J. Theor. Biol. 251 (2) (2008) 264–274.

[113] D.Z. Wang, H. Yan, A graph spectrum based geometric biclustering algorithm, J. Theor. Biol. 317 (0) (2013) 200–211.