Computers in Biology and Medicine 42 (2012) 245–256

Contents lists available at SciVerse ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

An effective measure for assessing the quality of biclusters

Federico Divina [a], Beatriz Pontes [b,*], Raúl Giráldez [a], Jesús S. Aguilar-Ruiz [a]

[a] School of Engineering, Pablo de Olavide University, Ctra. Utrera s/n, 41013 Seville, Spain
[b] Department of Computer Science, University of Seville, Avda. Reina Mercedes s/n, 41012 Seville, Spain

ARTICLE INFO

Article history:
Received 17 March 2011
Accepted 26 November 2011

Keywords:
Biclustering
Gene expression data
Shifting and scaling patterns

ABSTRACT

Biclustering is becoming a popular technique for the study of gene expression data. This is mainly due to the capability of biclustering to address the data using various dimensions simultaneously, as opposed to clustering, which can use only one dimension at the time. Different heuristics have been proposed in order to discover interesting biclusters in data. Such heuristics have one common characteristic: they are guided by a measure that determines the quality of biclusters. It follows that defining such a measure is probably the most important aspect. One of the popular quality measure is the *mean squared residue* (MSR). However, it has been proven that MSR fails at identifying some kind of patterns. This motivates us to introduce a novel measure, called *virtual error* (VE), that overcomes this limitation. Results obtained by using VE confirm that it can identify interesting patterns that could not be found by MSR.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Microarray techniques allow to simultaneously measure the expression level of thousands of genes under different experimental conditions, producing in this way a huge amount of data. Microarray data is widely used in genomic research, and is usually organized in matrices. In such matrices, rows and columns may represent, for instance, experimental conditions and genes, respectively. Thus, an element of such expression matrix stands for the expression level of a given gene under a specific experimental condition.

Different techniques have been used in order to extract information from expression matrices. Among these, clustering has been widely used, with the main goal of finding groups of genes that present a similar variation of expression level under all the experimental conditions [1]. However, relevant genes are not necessarily related to every condition. In other words, genes might be relevant only for a subset of experimental conditions [2]. Thus, clustering should be performed not only on one dimension (genes) but on two dimensions (genes and conditions) simultaneously.

For this reason, biclustering techniques [3] are becoming popular due to the ability of simultaneously grouping both genes and conditions. The first approach applied to microarray analysis was proposed by Cheng and Church [4]. Biclustering bases its essential principle on clustering, from which differs in two main aspects. Considering a microarray of gene expression data, a typical clustering technique would build a set of clusters, where each gene belongs exactly to one single cluster. Nevertheless, many genes may be grouped into diverse clusters (or none of them) depending on their participation in different biological processes within the cell [5]. Another difference is found in the fact that biclustering aims at identifying genes that are co-expressed under a subsets of conditions. This is essential for numerous biological problems, such as the analysis of genes contributing to certain diseases [2], assigning biological functionalities to genes or when the conditions of a microarray are diverse.

Finding significant biclusters in a microarray has been proven to be a NP-hard problem [6] and much more complex than clustering [7]. Consequently, many of the proposed techniques are based on optimization procedures as the search heuristic. The development of a suitable heuristic is a critical factor for discovering interesting biclusters in an expression matrix. In order to guide a search heuristic, it is essential to define a measure or cost function for establishing the quality of bicluster. The use of a suitable measure is a key factor, as it determines the effectiveness of the heuristic. Moreover, such a measure can be used for comparing the performances of different search strategies.

As already stated, Cheng and Church [4] were the first in applying biclustering to microarray data. Their proposal was based on a greedy search heuristic based on a cost function, called *mean squared residue* (henceforth MSR). MSR measures the numerical coherence among the genes in a bicluster. Cheng and Church maximize the volume with an upper bound on MSR, since MSR tends to decrease as volume of bicluster increases. MSR has also been used as part of the cost function in some other works. Gremalschi and

* Corresponding author.
E-mail addresses: fdivina@upo.es (F. Divina), bepontes@us.es (B. Pontes), giraldez@upo.es (R. Giráldez), aguilar@upo.es (J.S. Aguilar-Ruiz).

0010-4825/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2011.11.015

Altun [8] proposed the opposite strategy to Cheng and Church that is to minimize MSR with a lower bound on volume of bicluster. In [9], the authors developed an iterative algorithm for finding a predefined number of biclusters. Bryan et al. [10] applied in their work a simulated annealing heuristic. A greedy strategy was proposed in [11], where the search starts from seed generated with the *k*-means clustering algorithm. Similar strategies were proposed in [12] and in [13], where a particle swap optimization (PSO) technique was used to refine the initial biclusters. A multi-objective PSO approach was proposed by Liu et al. [14], and in another work [15] they proposed a multiple objective ant colony optimization algorithm. An approach based on evolutionary computation was proposed by Divina and Aguilar [7] and Bleuler et al. [16], while other authors [17] based their proposal on fuzzy technology and spectral clustering.

Other biclustering approaches, which are not based on MSR, include the proposal by Tanay et al. [6], based on the use of bipartite graphs and probabilistic techniques. Sheng et al. [18] introduced the use of Gibbs sampling for finding biclusters. Carmona-Saez et al. [19] presented a new data mining technique, based on matrix factorization. Madeira and Oliveira [20] found all relevant biclusters in linear time on the size of the microarray. Ayadi et al. [21] proposed an enumeration algorithm which uses a tree structure to represent different biclusters discovered during the enumeration process. The same authors also proposed a hill climbing strategy [22]. Bicego et al. [23] rely on a probabilistic model, called topic model, to detect groups of highly correlated genes and conditions. Liu and Wang [24] developed a polynomial time algorithm, which searched for optimal biclusters with the maximum similarity score. Finally, Hanczar and Nadif [25,26] try to improve the performances of biclustering algorithms by using the ensemble approach.

Even if MSR has been used in many proposals for finding biclusters, it nevertheless presents some drawbacks that will be discussed in the next section. In this paper, we propose a measure, called *virtual error* (henceforth VE), as a novel cost function for evaluating biclusters based on the concept of behavioural pattern. Gene correlation in a bicluster can be represented by two distinct kind of patterns: shifting and scaling, being both of them formally described by Aguilar [27]. Taking into account the concept of pattern, it is possible to focus the analysis of expression data on the general behaviour that genes exhibit under subsets of conditions, instead of grouping genes with similar expression values. Shifting patterns represent groups of genes following exactly the same trends, i.e., parallel behaviour, but in different range of values. Scaling patterns represent genes in a bicluster fluctuating in unison, without presenting the same differences through the conditions, although conserving a multiplicative factor.

In order to test the effectiveness of VE, we incorporated it in a multi-objective evolutionary biclustering algorithm. Experiments show that VE yields the algorithm at finding interesting biclusters, confirming the validity of our proposal.

This paper is organized as follows. In the next section, we present the main motivations for this work. In Section 3 an analysis of the shifting and scaling patterns is given; we then provide a formal definition of VE in Section 4, followed by a description of the algorithm used in the experiments in Section 5. In Section 6 experimental results obtained from different datasets are presented and discussed. Finally, in Section 7, we summarize the main conclusions.

2. Motivation

Let, from now on, $\mathcal{B}$ be a bicluster containing I conditions and J genes, and let b_{ij} denote the elements of $\mathcal{B}$, where $1 \le i \le I$ and $1 \le j \le J$. Then $\mathcal{B}$ can be represented as follows:

$$\mathcal{B} = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1J} \\ b_{21} & b_{22} & \cdots & b_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ b_{I1} & b_{I2} & \cdots & b_{IJ} \end{pmatrix}$$

where rows are relative to conditions and columns to genes.

The MSR of a bicluster $\mathcal{B}$ is then given by Eq. (1)

$$MSR(\mathcal{B}) = \frac{1}{I \cdot J} \sum_{i=1}^{I} \sum_{j=1}^{J} (b_{ij} - \mu_{c_i} - \mu_{g_j} + \mu_{\mathcal{B}})^2 \quad (1)$$

where μ_{c_i} and μ_{g_j} are the means of ith row (condition c_i) and jth column (gene g_j), respectively; and $\mu_{\mathcal{B}}$ is the mean of the whole bicluster.

The lower the MSR, the better the numerical coherence among the genes is and, therefore, the better the quality of a bicluster seems to be. Thus, when the genes of a bicluster $\mathcal{B}$ show exactly the same shape, with the only difference that they started with different initial values, then the MSR of $\mathcal{B}$ is equal to 0 [27].

Nevertheless, biclusters with constant genes, i.e., that present a flat behaviour across all the experimental conditions, will also have MSR equal to 0. The same holds for a bicluster containing only one gene or condition. Thus, MSR equals to 0 does not always identify a good bicluster. Other measures, such as the volume and the gene variance of biclusters may be used in combination to the MSR, in order to reject trivial biclusters.

The volume is the number of rows multiplied by the number of columns ($I \cdot J$), and the gene variance of a bicluster $\mathcal{B}$ is given in Eq. (2)

$$var_{\mathcal{B}} = \frac{1}{I \cdot J} \sum_{i=1}^{I} \sum_{j=1}^{J} (b_{ij} - \mu_{g_j})^2 \quad (2)$$

If a bicluster presents a high gene variance, it means that its genes exhibit fluctuating trends under the same subset of conditions.

From the above considerations, it becomes evident that a criterion needs to be used in order to establish when the MSR of a bicluster can be considered low. In [4] a user parameter, denoted as δ, is used as threshold: biclusters with MSR higher than δ are rejected. Before applying biclustering, δ needs to be independently set for each dataset [28].

MSR has been proven to be inefficient for finding some kind of biclusters in microarray data, especially when they present strong scaling tendencies. In [27] an in-depth analysis on the consequences of using MSR as a quality measure for searching biclusters is proposed. One of the main conclusions is that MSR is not capable of assessing the quality of biclusters containing shifting trends, as shifting behaviour does not affect the MSR. Moreover, scaling patterns have an undesirable effect for evaluating biclusters: small scaling variations in data lead to great increases of MSR. Therefore, good biclusters may have a score greater than δ.

Fig. 1 shows an example of a bicluster discovered in the Human B-cells dataset [29]. This is a typical visualization of bicluster, where conditions are represented in the *X*-axis, the values of gene expression are represented in the *Y*-axis and each line is a gene. Cheng and Church [4] set δ to 1200 for this dataset. From a visual inspection of the bicluster, it can be seen that it is a quality bicluster, since the genes are highly co-expressed, presenting strong scaling trends. Nevertheless, the MSR for this example is 3470.15, almost three times the value of δ.

These observations motivate us to propose a novel approach for evaluating biclusters, taking into account the scaling behaviour inherent to gene data. This behaviour is more difficult to detect than the shifting one, but it is more probable in nature. Being able to find biclusters containing also scaling patterns

would be essential to help scientists obtain relevant information for numerous biological problems.

3. Patterns

The concepts of *shifting pattern* and *scaling pattern* were formally defined by Aguilar [27]. The main idea is that the expression values of the genes included in a bicluster have common components. These concepts are essential to understand the foundations of our approach VE.

Genes in a bicluster might present either one of these patterns or both of them simultaneously. In the following we provide formal definitions for the concepts of patterns, and provide some examples that helps to clarifying these concepts.

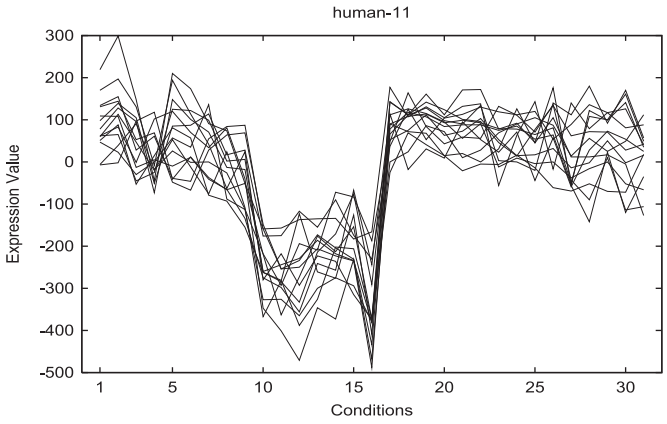


Fig. 1. Example of a quality bicluster with MSR = 3470.15.

3.1. Shifting pattern

Let us suppose each gene g_j (jth column) of the bicluster $\mathcal{B}$ has a typical value π_j, and that the expression values b_{ij} may be obtained by adding to π_j a value β_{ij}. Then, we can write any expression value as $b_{ij} = \pi_j + \beta_{ij}$. If for each condition c_i (ith row) these values β_{ij} are the same for all genes, then we can express them as β_i. We name β_i *shifting coefficient* for the ith condition.

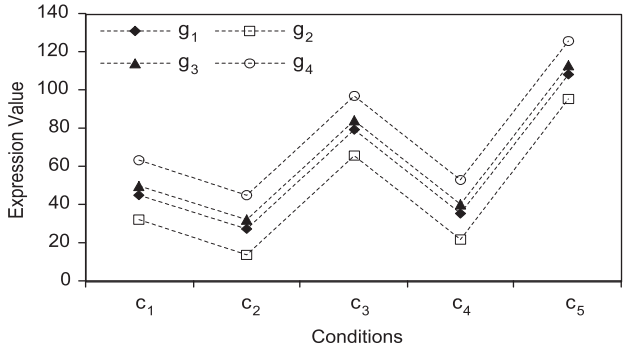
So we can express b_{ij} as

$$b_{ij} = \pi_j + \beta_i \tag{3}$$

When the expression values of a bicluster fulfil Eq. (3), such bicluster follows a *perfect shifting pattern*. Graphically, a perfect shifting pattern gives a parallel behaviour of the genes. Fig. 2 illustrates an example of bicluster presenting a perfect shifting pattern. This bicluster contains four genes g_j (with $1 \le j \le 4$) and five conditions c_i (with $1 \le i \le 5$). Below the graphic, the expression values (matrix on the left) are provided next to the typical pattern values (π_j) and shifting coefficients (β_i). As we can observe, π_j is constant for each gene (column), while the shifting coefficient β_i is constant for each condition (row). Furthermore, if we compute the MSR for the bicluster in Fig. 2, we can see that it is equal to 0.

3.2. Scaling pattern

As we discussed before, it is likely to find genes that follows a shifting pattern. Nevertheless, it is also interesting to discover genes that have the same behaviour, but not with the same scale. In the shifting case, the adjustment for condition c_i is obtained in an additive way. For the scaling pattern, this adjustment is obtained in multiplicative way. In this case, let us suppose the expression values b_{ij} can be obtained by multiplying π_j by a constant value for each condition c_i. We name such value as *scaling coefficient* and denote it by α_i. In such a case, a bicluster



$$\mathcal{B} = \begin{pmatrix} 45 & 32 & 50 & 63 \\ 27 & 14 & 32 & 45 \\ 79 & 66 & 84 & 97 \\ 35 & 22 & 40 & 53 \\ 108 & 95 & 113 & 126 \end{pmatrix} = \begin{pmatrix} 25+20 & 12+20 & 30+20 & 43+20 \\ 25+2 & 12+2 & 30+2 & 43+2 \\ 25+54 & 12+54 & 30+54 & 43+54 \\ 25+10 & 12+10 & 30+10 & 43+10 \\ 25+83 & 12+83 & 30+83 & 43+83 \end{pmatrix}$$

$$\{\pi_j\} = \begin{matrix} \pi_1 & \pi_2 & \pi_3 & \pi_4 \\ \{25 & 12 & 30 & 43\} \end{matrix}$$

$$\{\beta_i\} = \begin{matrix} \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 \\ \{20 & 2 & 54 & 10 & 83\} \end{matrix}$$

Fig. 2. Graphical representation of a bicluster containing a perfect shifting pattern.

shows a *perfect scaling pattern* when the values b_{ij} can be obtained by applying Eq. (4)

$$b_{ij} = \pi_j \times \alpha_i \tag{4}$$

In Fig. 3, an example of perfect scaling pattern is displayed. The bicluster contains four genes and five conditions. The expression values are also provided next to the pattern typical values and scaling coefficients, π_j and α_i, respectively. In this case, the genes do not follow a parallel tendency. Although the genes present the same behaviour with regard to the regulation, changes are more abrupt for some genes than for others.

In this case, the MSR for the bicluster in Fig. 3 is 423.98. It is a very high value, taking into account that the bicluster is very small. Thus, the scaling pattern leads to great increase of MSR.

3.3. Combined pattern: shifting and scaling

A bicluster may include some of the aforementioned patterns or even both of them, shifting and scaling, at the same time. In fact, it is the most probable case when real data are used. When both kind of patterns are included simultaneously, we say that the bicluster contains a *combined pattern*. In such a case, the expression value of the gene g_j and the condition c_i is calculated by multiplying and adding the scaling and shifting coefficients to the typical value π_j, respectively. Thus, we can obtain the following expression by combining Eqs. (3) and (4):

$$b_{ij} = \pi_j \times \alpha_i + \beta_i \tag{5}$$

Observe the example given in Fig. 4. This bicluster includes simultaneously the perfect patterns shown in Figs. 2 and 3, keeping the same scaling and shifting coefficients. Nevertheless, to visually identify that this bicluster follows a combined pattern is more difficult to find a single shifting or scaling pattern, since the effects of one has influence on the other.

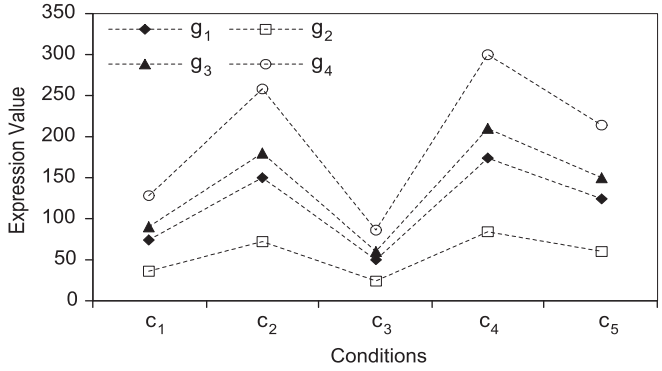
At first sight, genes g_1, g_3 and g_4 have similar behaviour, although g_4 differs for the last condition. However, gene g_2 seems independent of the other genes, since it has ascending tendency across every conditions, while the other genes presents a fluctuating behaviour. This fact happens when the shifting coefficients β_i are of the same magnitude that $\pi_j \times \alpha_i$. Note this aspect by observing the second column (gene g_2) of the matrix. It is also interesting that the shifting causes the genes g_1, g_2 and g_3 to significantly change for the last condition with regard to Fig. 3. Observe that the shifting coefficient for this condition (fifth row of the matrix), that is 83, has the same magnitude that $\pi_j \times \alpha_i$ for such genes. Therefore, when a bicluster includes shifting and scaling patterns simultaneously, identifying it as a good bicluster is a difficult task [30].

Regarding the MSR for the bicluster in Fig. 4, it is equal to 423.98 that is the same value than the scaling pattern case (Fig. 3). This fact reasserts that shifting behaviour does not affect the MSR and, furthermore, good biclusters may have a high MSR score.

4. Virtual error

In this section, we propose a new cost function, called *virtual error* (VE), for establishing the quality of biclusters. The basic idea behind VE is to measure how genes follow the general tendency within the bicluster. This is because if all the genes of a bicluster follow the same tendency under a given set of conditions, then it means that they are activated/deactivated under the same experimental conditions. If follows that such a bicluster may be potentially biologically interesting.

In order to catch the general tendency of the genes across the conditions contained in the bicluster, we first calculate a new

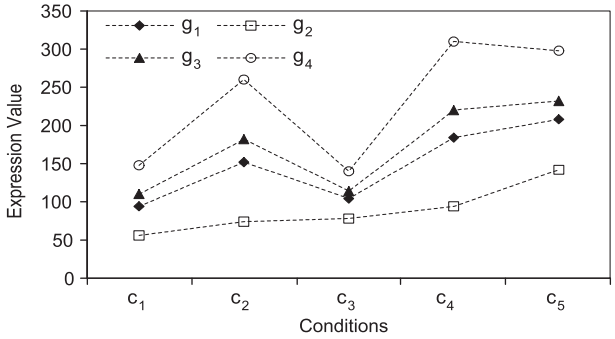


$$\mathcal{B} = \begin{pmatrix} 75 & 36 & 90 & 129 \\ 150 & 72 & 180 & 258 \\ 50 & 24 & 60 & 86 \\ 175 & 84 & 210 & 301 \\ 125 & 60 & 150 & 215 \end{pmatrix} = \begin{pmatrix} 25\times 3 & 12\times 3 & 30\times 3 & 43\times 3 \\ 25\times 6 & 12\times 6 & 30\times 6 & 43\times 6 \\ 25\times 2 & 12\times 2 & 30\times 2 & 43\times 2 \\ 25\times 7 & 12\times 7 & 30\times 7 & 43\times 7 \\ 25\times 5 & 12\times 5 & 30\times 5 & 43\times 5 \end{pmatrix}$$

$$\begin{matrix} & \pi_1 & \pi_2 & \pi_3 & \pi_4 \end{matrix}$$
$$\{\pi_j\} = \{25 \quad 12 \quad 30 \quad 43\}$$

$$\begin{matrix} & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 \end{matrix}$$
$$\{\alpha_i\} = \{3 \quad 6 \quad 2 \quad 7 \quad 5\}$$

Fig. 3. Graphical representation of a bicluster containing a perfect scaling pattern.



$$\mathcal{B} = \begin{pmatrix} 95 & 56 & 110 & 149 \\ 152 & 74 & 182 & 260 \\ 104 & 78 & 114 & 140 \\ 185 & 94 & 220 & 311 \\ 208 & 143 & 233 & 298 \end{pmatrix} = \begin{pmatrix} 25\times 3+20 & 12\times 3+20 & 30\times 3+20 & 43\times 3+20 \\ 25\times 6+2 & 12\times 6+2 & 30\times 6+2 & 43\times 6+2 \\ 25\times 2+54 & 12\times 2+54 & 30\times 2+54 & 43\times 2+54 \\ 25\times 7+10 & 12\times 7+10 & 30\times 7+10 & 43\times 7+10 \\ 25\times 5+83 & 12\times 5+83 & 30\times 5+83 & 43\times 5+83 \end{pmatrix}$$

$$\begin{array}{rccccc} & \pi_1 & \pi_2 & \pi_3 & \pi_4 & \\ \{\pi_j\} = & \{25 & 12 & 30 & 43\} & \\ & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 \\ \{\alpha_i\} = & \{3 & 6 & 2 & 7 & 5\} \\ & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 \\ \{\beta_i\} = & \{20 & 2 & 54 & 10 & 83\} \end{array}$$

Fig. 4. Graphical representation of a bicluster containing perfect shifting and scaling patterns.

column from the genes of the bicluster, called *virtual pattern*, defined as follows:

Definition 1 (*virtual pattern*). Given a bicluster $\mathcal{B}$, we define its virtual pattern ρ as the set of elements $\rho = \{\rho_1, \rho_2, \ldots, \rho_I\}$, where ρ_i, $1 \le i \le I$, is defined as the mean of the ith row:

$$\rho_i = \frac{1}{J}\sum_{j=1}^{J} b_{ij} \tag{6}$$

Each of the points of the virtual pattern represents the average value for all genes under a specific condition. Thus, if we graphically represent this values next to the real genes, the virtual pattern symbolizes the common tendency of the set of genes for the given bicluster.

Once the virtual pattern ρ has been computed, we can assess how well a specific gene g_j of the bicluster follows the general tendency. In order to do this, we compute the differences between the expression level values of g_j and the values of ρ for each experimental condition of the bicluster.

However, computing such differences using the original expression values can yield to a misclassification of the bicluster. In fact, the range of values of the expression values of the genes may be very far from each other. In order to remove or minimize these range differences, we will use the standardized gene expression values. In general, given a set of values $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$, the standardization of $\mathcal{V}$, that we denote as $\hat{\mathcal{V}}$, is the set $\hat{\mathcal{V}} = \{\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_N\}$ with $\hat{v}_k = ((v_k - \mu_{\mathcal{V}})/\sigma_{\mathcal{V}})$ (for $1 \le k \le N$), where $\mu_{\mathcal{V}}$ and $\sigma_{\mathcal{V}}$ are the mean and the standard deviation of the elements of $\mathcal{V}$, respectively. By using this numerical transformation, we standardize the expression values of every gene, including the virtual pattern. In this way, we scale the values to a common range.

We now define VE as the average value of all the differences between the standardized expression values and the standardized virtual pattern:

Definition 2 (*virtual error*). The virtual error of a bicluster $\mathcal{B}$, denoted by $\mathrm{VE}(\mathcal{B})$, is defined as

$$\mathrm{VE}(\mathcal{B}) = \frac{1}{I\cdot J}\sum_{i=1}^{I}\sum_{j=1}^{J} |\hat{b}_{ij} - \hat{\rho}_i| \tag{7}$$

where $\hat{b}_{ij}$ is the standardized expression value of the element $b_{ij} \in \mathcal{B}$, and $\hat{\rho}_i$ is the standardized value of the element ρ_i in the virtual pattern ρ.

VE computes the differences between the real genes and the virtual pattern, once they have been standardized. Therefore, the more similar the genes are, the lower the value for VE. In fact, if a bicluster follows a perfect shifting or scaling, VE is zero.[1] It follows that the lower the VE the better the bicluster.

A diagram on how VE is computed for any input bicluster is shown in Fig. 5. The whole process is comprised by four different steps: calculation of the virtual pattern, standardization of both the virtual pattern and the whole bicluster, and finally VE is given by the average of the differences between every standardized gene component and its corresponding standardized pattern element.

The smoothing effect of the standardization is clear in Fig. 6(b), where the range of values in the y-axis is significatively narrower than the original one shown in Fig. 6(a).

The bicluster in the example has a VE of 0.21, i.e., near to zero. This result shows that the genes follow a very similar behaviour

[1] Appendix A states two theorems and their corresponding proofs demonstrating that VE is zero when a bicluster follows a shifting or scaling pattern.

across the conditions, as Fig. 6(a) confirms. Therefore, this low value of VE indicates the good quality of the bicluster.

In general, we can conclude that VE is robust when a bicluster follows a shifting or scaling pattern since, in these cases, the value of VE is equals to zero. The same does not hold as far as MSR is concerned. In fact, it is true that when a bicluster presents a shifting pattern, its MSR is equals to zero. However, the MSR of a bicluster is not zero when the bicluster follows a perfect scaling pattern [27]. This is due to the fact that MSR is very sensitive to this kind of patterns, as shown in Section 3.2. This property demonstrates the effectiveness of the VE measure with respect to MSR.

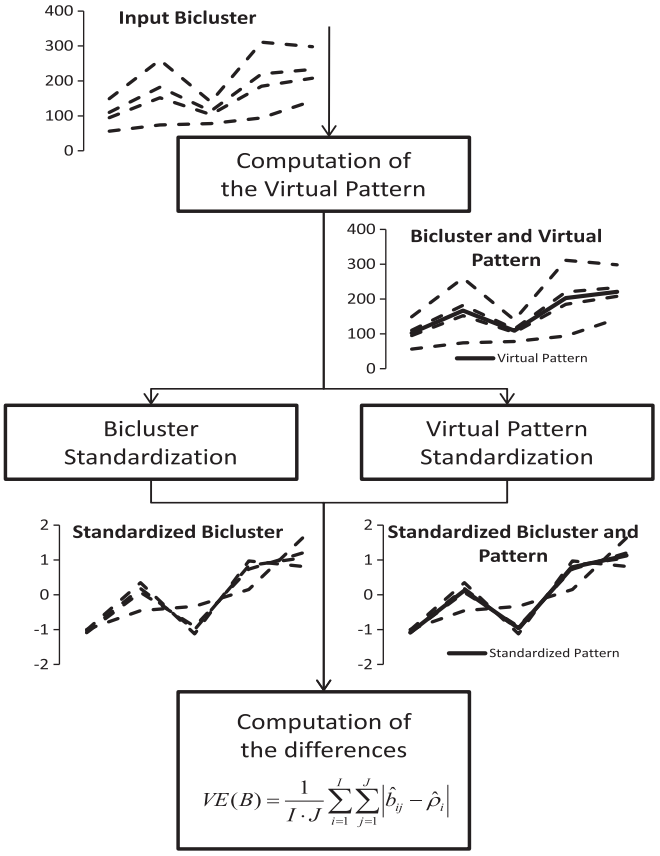


Fig. 5. Virtual error computation diagram.

5. Multi-objective evolutionary biclustering algorithm

In the previous section, VE was introduced as a quality measures that can be used to guide an optimization heuristic in order to discover biclusters in an expression matrix. However, the problem cannot be addressed by only optimizing the VE of biclusters. In fact, this approach may lead to the discovery of uninteresting biclusters. For instance, flat biclusters will have a low value of VE, or, again, biclusters containing few genes and conditions will typically have lower values of VE, if compared to biclusters characterized by higher volume. The same hold for the MSR. This is because the more genes or conditions are contained in a bicluster, the less likely the genes are to follow the same behaviour. Such biclusters are not very interesting, and, in order to solve this issue, other properties of the biclusters are usually optimized, e.g., the volume.

In particular, we are interested in finding biclusters with high volume, good quality (being quality measured by an appropriate metric such as VE or MSR) and relatively high gene variance. Thus, we can individuate at least three objectives to be optimized and these objectives are usually in conflict with each other.

For this reason, the problem of finding biclusters in an expression matrix can be straightforwardly seen as a multi-objective problem. Moreover, by addressing this problem as a multi-objective problem, it is not necessary to combine all the objectives into single cost function, which might become complicated, especially when both maximization and minimization are involved. Finding a way to combine the objectives in a single function can be problematic, and may require more parameters to set [31].

For these reasons, we incorporate VE into a multi-objective heuristic. In this section, before describing the algorithm used in this paper, we first provide a brief introduction to multi-objective optimization.

5.1. Multi-objective optimization

In a multi-objective optimization problem (MOP) [32], several objectives are to be optimized simultaneously. Often, these objectives are in conflict with each other. It becomes then difficult

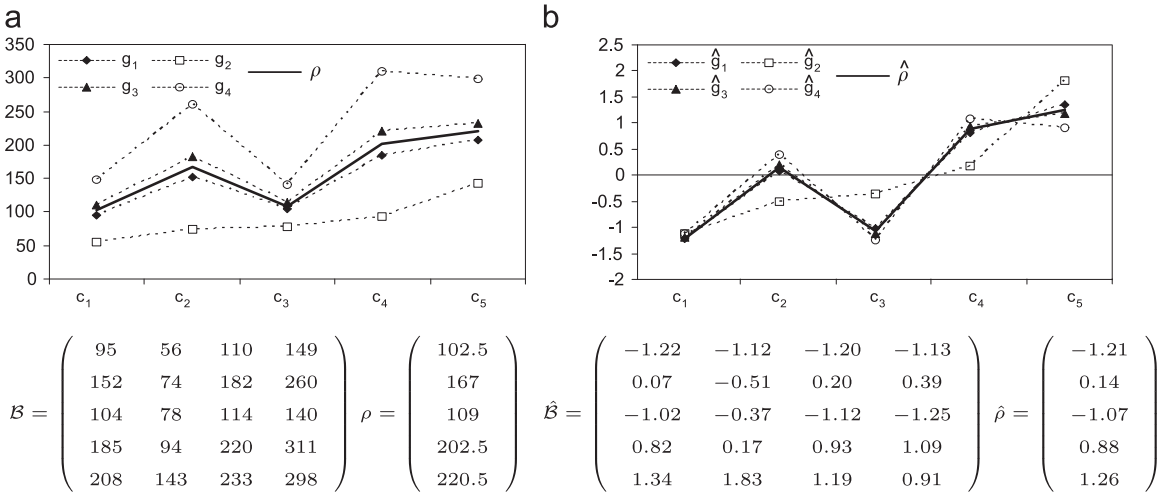


Fig. 6. Example to illustrate the virtual error. (a) Bicluster and the virtual pattern. (b) Standardized bicluster and the standardized virtual pattern.

to solve such problems, especially if all objectives are combined into a single objective to be optimized.

A solution to a MOP problem can be described by an *decision vector* $(x_1, x_2, \ldots, x_m)$ in the decision space X. A function $f : X \rightarrow Y$ evaluates the quality of a specific solution, by assigning it an *objective vector* $(y_1, y_2, \ldots, y_n)$ in the objective space Y. For instance, in the problem of finding biclusters, a decision vector could specify which genes and conditions belong to a bicluster, while the relative objective vector may specify the VE and volume of the bicluster.

If there was only one objective to be optimized, then assessing if a solution is better than another would be trivial. For instance, if the objective was to be maximize, then a solution $x_1 \in X$ would be better than $x_2 \in X$ is $f(x_1) > f(x_2)$, or equivalently if $y_1 > y_2$ ($y_1, y_2 \in Y$). In this case, there exists only a single optimum in the objective space.

If the objectives to be optimized are more than one, evaluating if x_1 is better than x_2 is more complex. Following the well known concept of Pareto dominance [33], we say that an objective vector y_1 dominates y_2 if no component of y_1 is worse than any component of y_2 and at least one component of y_1 is better than y_2. We can then say that a solution x_1 is better than x_2 if y_1 dominates y_2. We then say that x_1 dominates x_2. In contrast with single objective problems, with multiple objectives, there exist several optimal objective vectors representing different trade-off between all the objectives. The set of optimal solution in the decision space is called *Pareto set*, and its image in the objective space is called the *Pareto front*.

Evolutionary algorithms (EAs) are particularly suited for solving MOPs [33–35]. In fact they can approximate the Pareto front in a single optimization run. This is due to the ability of EAs to deal with a set of solutions simultaneously, and to exploit similarities among solutions by means of genetic operators. Moreover, EAs are less influenced by the shape of the Pareto front than other search techniques.

5.2. Sequential multi-objective biclustering

In this section we will give a brief description of the algorithm used in this paper. The algorithm is called SMOB (for sequential multi-objective biclustering). For a more detailed description of SMOB, we refer the reader to [36].

We individuate four objectives that are to be optimized: MSR, VE, the gene variance and the volume of the biclusters. However, as it will be described in Section 6, we will use three different settings of the algorithm, where only three objectives will be optimized at a time. The objectives that will always be considered are the volume and the gene variance.

SMOB adopts a sequential covering strategy. A MOEA is called n times, and each time a bicluster is returned. The returned bicluster is stored in a list L that contains all the biclusters found so far. When MSR is used as an objective, the returned bicluster is stored in the list only if its MSR is lower than the threshold δ. In order to avoid overlapping among biclusters, we associated a weight $w(e_{ij})$ with each element e_{ij} of the expression matrix. These weights are adjusted right after a bicluster is returned. $w(e_{ij})$ is equal to the number of biclusters stored in L that contain e_{ij}. When evaluating a bicluster the weights of its elements are used in order to penalize biclusters overlapping with elements of L, as it will be explained in the following. Notice that the sequential coverage strategy adopted is such that the order in which biclusters are discovered does not reflect their quality nor their biological relevance.

Each individual encodes a single bicluster. The encoding of biclusters is the one proposed in [7,37], where bit strings are evolved.

Tournament selection is used, and selected individuals undergo crossover and mutation. Elitism is applied by letting the non-dominated individuals survive to the next generation. The way individuals are evaluated differs slightly from [36]. In fact, the version of SMOB used in this paper adopts a strategy similar to NSGA [38]. Individuals are divided into different non-dominated fronts, and individuals belonging to the same front have the same starting fitness $rank(x)$. For instance, a non-dominated individual x_1 will have $rank(x_1) = 0$.

The fitness of an individual x is then defined as

$$f(x) = rank(x) + sh(x) + P(x) \tag{8}$$

where $sh(x)$ is the phenotypic sharing [39] and $P(x) = 1 - (V(x) - \sum_{i,j \in x} w(e_{ij}))/V(x)$, where $V(x)$ is the volume of x. In our implementation $sh(x)$ is the minimal euclidian distance, computed on the objectives to the other individuals. Thus, $\forall y \neq x : sh(x) = \min(\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2})$, where n is the number of objectives, y is an individual, and x_i, y_i are the objectives used.

As mentioned earlier in this section, $P(x)$ is used in order to avoid overlapping among biclusters. From the definition of $P(x)$, it follows that if a bicluster has low volume and it covers elements of the expression matrix that are already contained in many biclusters already found, $P(x)$ will be high. On the other hand, if the bicluster has a high volume and it overlaps with few biclusters, the penalty will be lower. If the bicluster x does not overlap with any bicluster then $P(x)$ is zero.

6. Experiments

In order to assess the validity our proposal, we conduct experiments on nine datasets shown in Table 1. The embryonal dataset was preprocessed as in [28], where each entry of the original dataset was substituted by its normalized value between 0 and 600. All the other datasets were preprocessed as in [4]. The most important preprocessing operation regards missing values. They are replaced with random values, although it is known the existing risk that these random numbers can affect the discovery

Table 1
Datasets used in the experimentation.

Dataset	Name	# Genes	# Conditions	Ref.
Yeast	Yeast *Saccharomyces cerevisiae* cell cycle	2884	17	[41]
Human	Human B-cells	4026	96	[29]
Colon	Colon cancer	2000	62	[41]
Malaria	Malaria *Plasmodium parasites* life cycle	3719	16	[42]
Embryonal	Embryonal tumours of the central nervous system	7129	60	[43]
Leukemia	Leukemia	7129	72	[44]
RatCNS	Rat central nervous system	112	9	[45]
Steminal	Steminal cells	26,127	30	[46]
PBM	Peripheral blood monocytes	2329	139	[47]

of biclusters [40]. The expectation was that these random values would not form recognizable patterns.

The aim of the experimentation is to assess the validity of VE as a quality measure. In particular, we aim to test whether VE can yield the discovery of biclusters characterized by higher gene variance and volume. Moreover, we are interested in comparing the results achieved by using VE as a measure of the quality of a bicluster against the results obtained when MSR is used. In order to do this, we compare the algorithm described in Section 5.2 against a version that uses as an objective MSR instead of VE. Another point we want to test, is whether the use of a too low threshold δ may prevent the algorithm from finding interesting biclusters, since the search is limited to biclusters with MSR lower than δ.

In the experimentation, we use three settings of the algorithm, that differ for the objectives subject of optimization:

- **SMOB**-δ. In this setting the objectives are: volume, gene variance and MSR, with δ shown in the second column of Table 2.
- **SMOB-VE**. In this setting the objectives are: volume, gene variance and VE.
- **SMOB**-Δ. In this setting the objectives are: volume, gene variance and MSR, with δ shown in the third column of Table 2.

In both SMOB-δ and SMOB-Δ, if the algorithm returns a bicluster whose MSR is higher than δ, such bicluster is rejected. As it can be noticed, the only difference between settings SMOB-Δ and SMOB-δ lies in the value used for the threshold δ. SMOB-Δ was included for testing the limitations, in terms of biclusters found, that the use of a low δ can impose on an algorithm.

In order to perform a fair comparison, we use the same parameter setting in all the versions of the algorithm. This setting is given in Table 3. The values of these parameters were obtained after a number of preliminary runs aimed at testing different parameter settings.

The values of δ used in SMOB-δ on the human and the yeast dataset were taken from [4], while for the other datasets they were established using a procedure suggested in [4]. The values of δ used in SMOB-Δ were determined as follows. For each dataset, we first run the experiments with SMOB–VE. We calculated the MSR of all the bicluster found, and then we selected the highest as δ to be used in SMOB-Δ. In this way we can test whether the use of δ prevented SMOB-δ from discovering some interesting biclusters, only because their MSR was higher than the used δ. Therefore, SMOB-Δ could obtain similar biclusters to those produced by SMOB–VE guaranteeing in this way a fair comparison.

On each dataset, we obtained 100 biclusters for each setting of the algorithm. The average MSR and VE are reported in Tables 4 and 5, respectively. Table 6 reports both the average gene variance and the average volume.

Standard deviation is reported next to each result. In order to test the statistical significance of the results, we applied Student's *T*-test of difference of means with confidence level of 1%. In the tables, a minus (plus) symbol next to a result indicates that the average is statistically significantly lower (higher) than the average obtained by SMOB–VE on a given dataset. So, for instance, the average MSR obtained by SMOB-δ on the yeast dataset is significantly lower than the average MSR obtained by SMOB–VE.

In order to present the results in a clearer way, we have not included in the tables information about the statistically significance of the differences of results obtained by SMOB-δ and SMOB-Δ. However, we can say that in Table 4, all the results obtained by SMOB-Δ, but the results obtained on the RatCNS datasets, are significantly higher than those obtained by SMOB-δ. The average VE of biclusters obtained by SMOB-Δ is significantly higher than those obtained by SMOB-δ on five datasets, and in particular on the yeast, human, colon, malaria and the leukemia dataset. As far as the gene variance is concerned, SMOB-Δ obtained significantly higher results on all the datasets but the RatCNS dataset. The average volume characterizing biclusters found by SMOB-Δ is significantly higher on five datasets: yeast, human, colon, malaria and leukemia.

As it can be noticed from Table 4, SMOB-δ obtains the lowest values of MSR. This result was expected, since MSR is one of the

Table 2
Values of δ used in the settings SMOB-δ and SMOB-Δ.

Dataset	δ for SMOB-δ	δ for SMOB-Δ
Yeast	300	3400
Human	1200	23,000
Colon	500	3300
Malaria	600	19,000
Embryonal	1800	10,000
Leukemia	1800	23,130,700
RatCNS	5	11
Steminal	10	130
PBM	0.3	1.3

Table 3
Parameter settings for the algorithm.

Parameter	Value
Generations	100
Population size	200
Crossover probability	0.85
Mutation probability	0.2
Tournament size	4

Table 4
Average MSR obtained on each dataset.

Dataset	MSR				
	SMOB–VE	SMOB-δ		SMOB-Δ	
Yeast	1419.6 $\pm$ 513.8	272.0 $\pm$ 24.9	–	1062.0 $\pm$ 274.8	–
Human	16441.1 $\pm$ 3323.2	1103.5 $\pm$ 86.7	–	10742.5 $\pm$ 2321.3	–
Colon	2491.0 $\pm$ 408.9	455.2 $\pm$ 45.5	–	2107.2 $\pm$ 306.3	–
Malaria	14456.1 $\pm$ 2227.3	449.7 $\pm$ 174.6	–	12095.52298.1	–
Embryonal	1506.0 $\pm$ 1932.6	398.2 $\pm$ 344.2	–	802.97 $\pm$ 1243.1	–
Leukemia	81.0e5 $\pm$ 45.0e5	1533.1 $\pm$ 169.2	–	49.2e5 $\pm$ 29.3e5	–
RatCNS	3.43 $\pm$ 2.6	1.63 $\pm$ 1.3	–	2.17 $\pm$ 2.0	–
Steminal	43.87 $\pm$ 37.2	4.07 $\pm$ 2.1	–	13.13 $\pm$ 12.9	–
PBM	0.47 $\pm$ 0.1	0.19 $\pm$ 0.1	–	0.30 $\pm$ 0.1	–

Table 5
Average VE obtained on each dataset.

Dataset	VE				
	SMOB–VE	SMOB-δ		SMOB-Δ	
Yeast	0.83 $\pm$ 0.05	0.73 $\pm$ 0.08	–	0.86 $\pm$ 0.05	+
Human	0.90 $\pm$ 0.04	0.82 $\pm$ 0.07	–	0.92 $\pm$ 0.04	+
Colon	0.53 $\pm$ 0.04	0.33 $\pm$ 0.07	–	0.53 $\pm$ 0.03	
Malaria	0.73 $\pm$ 0.06	0.48 $\pm$ 0.18	–	0.75 $\pm$ 0.05	+
Embryonal	0.70 $\pm$ 0.08	0.88 $\pm$ 0.07	+	0.89 $\pm$ 0.08	+
Leukemia	0.82 $\pm$ 0.08	0.72 $\pm$ 0.09	–	0.95 $\pm$ 0.06	+
RatCNS	0.58 $\pm$ 0.18	0.71 $\pm$ 0.17	+	0.73 $\pm$ 0.18	+
Steminal	0.70 $\pm$ 0.10	0.98 $\pm$ 0.10	+	1.0 $\pm$ 0.10	+
PBM	0.28 $\pm$ 0.10	0.34 $\pm$ 0.10	+	0.35 $\pm$ 0.10	+

objective subject to optimization in SMOB-δ. Moreover, in this setting a small threshold is used in order to reject biclusters. However, when this threshold is relaxed, as in SMOB-Δ the average MSR of the biclusters obtained by SMOB–VE is comparable, even if slightly higher in general. The fact that the MSR obtained by SMOB-Δ is lower than that obtained by SMOB–VE is due to the fact that SMOB-Δ considers the MSR as an objective to be optimized, whilst SMOB–VE does not. Moreover, the results obtained by SMOB-Δ are in general significantly higher than those obtained by SMOB-δ. Only on the RatCNS dataset the two algorithms obtain an average MSR that is not statistically significant. This result shows that a too low threshold restrict the search performed by the algorithm too tightly. This fact can be clearly seen by inspecting Table 6. This table presents on the left part the average row variance and on the right part the average volume obtained on the 100 biclusters found for each dataset. We can notice that SMOB-Δ obtains biclusters characterized by higher gene variance and volume than those discovered by SMOB-δ.

As Table 5 shows, SMOB-δ obtained the lowest values of VE on five datasets. This may seem odd, since SMOB-δ does not consider VE as an objective to be optimized. Nevertheless, this is explained by the fact that low values of MSR correspond to low values of VE. Moreover, in SMOB-δ a low threshold was used to limit the values of MSR. This kind of threshold is not used in SMOB–VE. Notice that VE is not the only objective being optimized by SMOB–VE, being the other ones the volume and the gene variance. This fact, in combination with the above considerations, explains why the values of VE obtained by SMOB–VE are, on average, higher than those obtained by SMOB-δ. It is worth to note that when the threshold δ is relaxed, the results obtained by SMOB–VE are significantly better than those provided by SMOB-Δ.

Due to the presence of shifting patterns in biclusters with low MSR, low values of MSR correspond to very low values of VE. However, the opposite is not true, i.e., low values of VE do not correspond to low values of MSR. This is explained by the presence of scaling patterns, which do not affect VE, but have the effect of remarkably increasing the values of MSR, as proven in [27].

From Table 6, we can notice that, on the two common objectives (i.e., gene variance and volume) subject of optimization by all the three settings of the algorithm, SMOB–VE obtains the best results. In particular the gene variance obtained by SMOB–VE is much higher than the gene variance of the biclusters found by SMOB-δ. This is because gene variance is in conflict with MSR. In fact, the presence of scaling patterns have the effect of incrementing the gene variance. On the other hand, MSR is also incremented by the presence of scaling patterns. Thus, since both the gene variance and the MSR are being optimized at the same time in SMOB-δ most scaling patterns are rejected. This is because such patterns would lead to biclusters with a MSR higher than the used threshold δ. It is also interesting to notice that SMOB–VE obtains biclusters with higher gene variance and volume than those obtained by SMOB-Δ. This confirms the fact that VE is more successful than MSR in guiding the search performed by the algorithm towards the discovery of more interesting biclusters. In fact, a low threshold imposed on MSR prevents the algorithm from finding biclusters containing certain patterns, and even if this threshold is increased, the algorithm obtains biclusters that are less interesting that those obtained when VE is used as one of the objectives.

6.1. Biological validation

In this section, we propose a biological validation of the results obtained by the three settings of the algorithm on three datasets: Embryonal, Leukemia and Steminal. We have selected these datasets because they were the ones having the highest percentage of gene names that could be found in Gene Ontology (GO) [48] and because of their significance. GO is a cross-species, controlled vocabulary describing three major functional features of gene products: molecular function (MF), cellular component (CC) and biological process (BP). In order to validate the results, we first use the gene functional dissimilarity (GFD) measure [49] and then, for each method, the number of significant biclusters, according to GO, is extracted.

GFD assigns a numerical value, between zero and one, to the gene set contained in a bicluster for each of the three GO sub-ontologies (MF, CC and BP). The value assigned to a biclusters represents the functional cohesion of the genes, where lower values represent higher functional similarity. We decided to use GFD since this measure presents the advantage that it can identify the most common function for all of the genes involved in a biological process.

Table 7 shows the average GFD for all the biclusters obtained on the three datasets by the three settings of the algorithm, according to each GO sub-ontology. Standard deviation is also reported next to the averages. In order to test whether the results are significantly different, we performed a two-tailed t-test. Results of this test, with confidence levels of 5% and 1%, are reported in the right part of the table. Thus, for instance, the average GFD obtained on the Embryonal dataset by SMOB–VE and SMOB-δ is significantly different for the MF and CC sub-ontologies with a confidence level of 1%, while for the BP ontology the difference is significant with a 5% confidence level. From the table, we can notice that SMOB–VE obtains the lowest GFD in all the cases but one: on the Steminal dataset for the BP ontology, where the three settings of the algorithm obtains basically the same results. This implies that, on average, the genes contained in the biclusters obtained on these datasets by using VE present a stronger functional similarity than those obtained by using MSR. It is also interesting to notice that in some cases, e.g., on the Leukemia dataset for the MF sub-ontology, SMOB-Δ obtains a lower value of GFD than SMOB-δ. This means that in such cases

Table 6
Average gene variance and volume obtained on each dataset.

Dataset	Gene variance					Volume				
	SMOB–VE	SMOB-δ		SMOB-Δ		SMOB–VE	SMOB-δ		SMOB-Δ	
Yeast	1661.7 ± 502.2	408.5 ± 77.2	–	1245.3 ± 294.5	–	606.4 ± 216.7	226.1 ± 77.3	–	461.4 ± 95.8	–
Human	17395.7 ± 3376.4	1412.2 ± 168.6	–	11412.4 ± 2476.8	–	1430.6 ± 473.2	362.6 ± 106.7	–	1128.3 ± 260.5	–
Colon	5597.6 ± 617.4	2836.7 ± 1026.7	–	4876.2 ± 548.6	–	1402.7 ± 514.4	197.5 ± 85.1	–	1172.4 ± 334.5	–
Malaria	20245.6 ± 2433.4	2667.3 ± 3765.7	–	16707.6 ± 2914.7	–	628.5 ± 179.2	54.6 ± 31.5	–	461.9 ± 108.2	–
Embryonal	1683.3 ± 2126.2	428.8 ± 365.9	–	849.4 ± 1304.1	–	1121.6 ± 401.7	1456.1 ± 577.3	+	1423.6 ± 519.26	+
Leukemia	8.94e6 ± 4.82e6	2391.3 ± 660.1	–	5.27e6 ± 3.10e6	–	1110.0 ± 348.0	441.7 ± 141.2	–	1132.2 ± 220.4	
RatCNS	4.6 ± 3.4	2.1 ± 1.6	–	2.7 ± 2.3	–	128.5 ± 63.4	128.5 ± 81.1		128.6 ± 69.5	
Steminal	50.8 ± 40.9	4.4 ± 2.3	–	13.8 ± 13.6	–	999.6 ± 336.0	1312.5 ± 371.8	+	1286.7 ± 399.4	+
PBM	1.3 ± 0.2	0.5 ± 0.2	–	0.8 ± 0.2	–	1427.9 ± 616.1	1659.7 ± 701.2		1744.8 ± 687.9	+

Table 7
Average GFD values for each of the three sub-ontologies of GO.

Dataset	GO's sub-ontology	Average GFD			Statistical significance (*T*-test)			
		SMOB–VE	SMOB-δ	SMOB-Δ	SMOB–VE vs SMOB-δ		SMOB–VE vs SMOB-Δ	
					<0.05	<0.01	<0.05	<0.01
Embryonal	MF	0.425 ± 0.100	0.464 ± 0.099	0.474 ± 0.106	×	×	×	×
	BP	0.603 ± 0.077	0.626 ± 0.068	0.633 ± 0.067	×		×	×
	CC	0.389 ± 0.098	0.446 ± 0.068	0.442 ± 0.083	×	×	×	×
Leukemia	MF	0.442 ± 0.098	0.505 ± 0.114	0.471 ± 0.074	×	×	×	
	BP	0.642 ± 0.050	0.648 ± 0.077	0.652 ± 0.050				
	CC	0.439 ± 0.081	0.473 ± 0.102	0.444 ± 0.067	×	×		
Steminal	MF	0.617 ± 0.044	0.632 ± 0.038	0.641 ± 0.036	×	×	×	×
	BP	0.710 ± 0.036	0.710 ± 0.023	0.715 ± 0.022				
	CC	0.492 ± 0.071	0.527 ± 0.052	0.528 ± 0.051	×	×	×	×

Table 8
Number of significant biclusters for the three GO ontologies, at two different levels.

Dataset	*p*-Value	Number of biclusters		
		SMOB–VE	SMOB-δ	SMOB-Δ
Embryonal	<0.01	1	2	5
	<0.05	16	6	10
Leukemia	<0.01	4	2	2
	<0.05	12	6	9
Steminal	<0.01	11	1	2
	<0.05	27	10	4

the average functional cohesion is stronger for the biclusters discovered by relaxing the threshold δ used with MSR. Thus, in some cases, a too strict threshold may prevent the algorithm from finding interesting biclusters.

To further analyse the results, we have used the ontologizer tool [50] to directly compute the most significantly enriched GO terms associated to the set of genes of every bicluster. For each bicluster, ontologizer provides a list with all the GO terms associated to the genes of the biclusters, detailing the GO category (BP, CC or MF) and the adjusted *p*-value (using the hypergeometric distribution and the Bonferroni multiple-hypothesis correction) for every GO term. Results of this analysis are reported in Table 8, which shows the number of biclusters with at least one significant GO term in any ontology, at two different levels: 1% and 5%. For Leukemia and Steminal datasets, SMOB–VE discovers more significant biclusters than SMOB-δ and SMOB-Δ with both levels. As far as the Embryonal dataset is concerned, SMOB–VE outperforms the other two versions of the algorithm when the *p*-value considered is lower than 0.05. In this case, of all the biclusters discovered by SMOB–VE, 16 are significant, while only 6 biclusters found by SMOB-δ are significant and SMOB-Δ finds 10 significant biclusters. If the threshold on the *p*-value is lowered to 0.01, only 1 bicluster is significant for SMOB–VE, whereas the other two settings finds 2 and 5 significant biclusters.

7. Conclusions

In this paper we introduce a novel measure for assessing the quality of biclusters, called *virtual error* (VE), which is based on the idea of measuring how well the genes in a bicluster follow the general tendency.

The main motivation for the developing of VE is to improve the performance MSR, more precisely with respect to the drawbacks of MSR at recognizing biclusters presenting scaling patterns. VE is, on the other hand, capable of dealing with both scaling and shifting patterns, being in this sense more robust than MSR. Moreover, when MSR is used, a user defined threshold must be supplied in order to reject non-interesting biclusters. Such threshold must be set for each dataset, and setting its value is not a trivial. VE does not need such a threshold, so that an algorithm using VE has less constraints through the search.

In order to assess the validity of VE, we conducted experiments on nine microarray datasets. For this, we have incorporated VE into a multi-objective evolutionary algorithm, called SMOB–VE, where the other objectives were the gene variance and the volume of the bicluster. We have compared the results of the previous algorithm with two other settings of the algorithm, called SMOB-δ and SMOB-Δ. These last two settings use MSR instead of VE. Specifically, the former uses a small threshold, while in the latter version this thresholds was removed. The last version of the algorithm was used in order to test whether the use of a too small threshold would prevent the algorithm from finding interesting biclusters.

From these experiments we can conclude that VE yields the algorithm obtaining good results on all the datasets. In general, biclusters found by VE are characterized by a greater volume and gene variance. An interesting result is that in the SMOB-δ setting of the algorithm, the average VE obtained is lower than in SMOB–VE. This is easily explained by the fact that low values of MSR correspond to low values of VE. We have also confirmed that the use of a threshold may prevent an algorithm from finding good results, when its value is too small.

We have also conducted a biological validation of the results obtained on three datasets. From this validation, it emerges that VE yields the discovery of biclusters whose genes have a stronger functional coherence. Moreover when the algorithm used VE, it discovers more significant biclusters, according to the adjusted *p*-value.

As for future developments, we intend to incorporate a biological evaluation of the biclusters in the algorithm. In fact, notice that at the moment, the sequential coverage strategy adopted is such that the order in which biclusters are discovered does not reflect their quality nor their biological relevance. We are also planning to investigate the effect of using the median instead of the mean in the functions used for computing VE. This may be beneficial if the data presents a condition imbalance. In this case, a preliminary analysis of the data could help to automatically decide whether to use the mean of the median.

In general, we can conclude that VE is an effective measure for assessing the quality of biclusters. In particular, VE is effective at recognizing biclusters containing both shifting and scaling patterns as quality biclusters. The same conclusions do not hold for MSR, which is negatively influenced by the presence of scaling patterns. It follows that VE can be used effectively within any heuristics for finding biclusters in gene expression data.

Conflict of interest statement

None declared.

Appendix A. Theorems

In this appendix we formally prove that a bicluster presenting either a shifting or a scaling pattern has VE equal to zero.

Theorem 1. *A bicluster presenting a perfect shifting pattern has virtual error equal to zero.*

Proof. If $\mathcal{B}$ contains a perfect shifting pattern, then we can represent each element as $b_{ij} = \pi_j + \beta_i$.

Applying two simple algebraic properties,[2] it is easy to obtain the mean and deviation of each gene g_j as

$$\mu_{g_j} = \pi_j + \mu_\beta$$

$$\sigma_{g_j} = \sigma_\beta$$

We use this results to standardize b_{ij}

$$\hat{b}_{ij} = \frac{b_{ij} - \mu_{g_j}}{\sigma_{g_j}} = \frac{\pi_j + \beta_i - \pi_j - \mu_\beta}{\sigma_\beta} = frac\beta_i - \mu_\beta \sigma_\beta \tag{A.1}$$

Combining the same properties (see footnote 1) it is easy to obtain the mean and standard deviation for the virtual pattern as

$$\mu_\rho = \mu_\pi + \mu_\beta$$

$$\sigma_\rho = \sigma_\beta$$

Finally, the standardized values of the virtual pattern are represented by

$$\hat{\rho}_i = \frac{\rho_i - \mu_\rho}{\sigma_\rho} = \frac{\mu_\pi + \beta_i - \mu_\pi - \mu_\beta}{\sigma_\beta} = \frac{\beta_i - \mu_\beta}{\sigma_\beta} = \hat{b}_{ij} \tag{A.2}$$

This result points out that when a bicluster follows a perfect shifting pattern, the virtual pattern is equal to all the real genes after standardizing them. This means that VE is equal to zero for every bicluster with a perfect shifting pattern. □

Theorem 2. *A bicluster presenting a perfect scaling pattern has virtual error equal to zero.*

Proof. If $\mathcal{B}$ contains a perfect scaling pattern, then we can represent each element as $b_{ij} = \pi_j \times \alpha_i$.

Following the same reasoning than for Theorem 1, the mean and deviation of each gene g_j are the following:

$$\mu_{g_j} = \pi_j \times \mu_\alpha$$

$$\sigma_{g_j} = \pi_j \times \sigma_\alpha$$

We use these results to standardize the values of the bicluster

$$\hat{b}_{ij} = \frac{b_{ij} - \mu_{g_j}}{\sigma_{g_j}} = \frac{\pi_j \times \alpha_i - \pi_j \times \mu_\alpha}{\pi_j \times \sigma_\alpha} = \frac{\alpha_i - \mu_\alpha}{\sigma_\alpha} \tag{A.3}$$

[2] Being $f(x) = g(x) \times c_1 + c_2$, we can enumerate these two properties related to the arithmetic mean ($\mu_{f(x)}$) and the standard deviation ($\sigma_{f(x)}$) of $f(x)$ as: $\mu_{f(x)} = \mu_{g(x)} \times c_1 + c_2$ and $\sigma_{f(x)} = \sigma_{g(x)} \times c_1$.

We next obtain the mean and standard deviation for the virtual pattern

$$\mu_\rho = \mu_\pi \times \mu_\alpha$$

$$\sigma_\rho = \mu_\pi \times \sigma_\beta$$

And the standardized values of the virtual pattern

$$\hat{\rho}_i = \frac{\rho_i - \mu_\rho}{\sigma_\rho} = \frac{\mu_\pi \times \alpha_i - \mu_\pi \times \mu_\alpha}{\mu_\pi \times \sigma_\alpha} = \frac{\alpha_i - \mu_\alpha}{\sigma_\alpha} = \hat{b}_{ij} \tag{A.4}$$

As we can observe, the result of the last equation shows that when a bicluster follows a perfect scaling pattern, the virtual pattern is equal to all the real genes after standardizing them. Therefore, we can state that VE is equal to zero for every bicluster with a perfect scaling pattern. □

References

[1] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. USA 95 (1998) 14863–14868.

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

[3] J. Hartigan, Direct clustering of a data matrix, J. Am. Stat. Assoc. 67 (337) (1972) 123–129.

[4] Y. Cheng, G.M. Church, Biclustering of expression data, in: Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology, La Jolla, CA, 2000, pp. 93–103.

[5] A.P. Gasch, M.B. Eisen, Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering, Genome Biol. 3 (11) research0059.1–0059.22.

[6] A. Tanay, R. Sharan, R. Shamir, Discovering statistically significant biclusters in gene expression data, Bioinformatics 18 (2002) 136–144.

[7] F. Divina, J.S. Aguilar-Ruiz, Biclustering of expression data with evolutionary computation, IEEE Trans. Knowl. Data Eng. 18 (5) (2006) 590–602.

[8] S. Gremalschi, G. Altun, Mean squared residue based biclustering algorithms. in: Bioinformatics Research and Applications, Fourth International Symposium, Lecture Notes on Computer Science, vol. 4983, 2008, pp. 232–243.

[9] J. Yang, H. Wang, W. Wang, P.S. Yu, An improved biclustering method for analyzing gene expression profiles, Int. J. Artif. Intell. Tools 14 (2005) 771–790.

[10] K. Bryan, P. Cunningham, N. Bolshakova, Application of simulated annealing to the biclustering of gene expression data, IEEE Trans. Inform. Technol. Biomed. 10 (3) (2006) 519–525.

[11] S. Das, S.M. Idicula, Application of cardinality based grasp to the biclustering of gene expression data, Int. J. Comput. Appl. 1 (1) (2010) 44–51.

[12] S. Das, S.M. Idicula, Greedy search-binary PSO hybrid for biclustering gene expression data, Int. J. Comput. Appl. 2 (3) (2010) 1–5.

[13] Y. Zhang, H. Wang, Z. Hu, A novel clustering and verification based microarray data bi-clustering method, in: Proc. Int. Conf. Swarm Intell., 2010, pp. 611–618.

[14] J. Liu, Z. Li, X. Hu, Y. Chen, Biclustering of microarray data with mospo based on crowding distance, BMC Bioinformatics 10 (S-4) (2009).

[15] J. Liu, Z. Li, X. Hu, Y. Chen, Moaco biclustering of gene expression data, I, J. Funct. Inf. Pers. Med. 3 (1) (2010) 58–72.

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

[17] C. Cano, L. Adarve, J. López, A. Blanco, Possibilistic approach for biclustering microarray data, Comput. Biol. Med. 37 (10) (2007) 1426–1436.

[18] Q. Sheng, Y. Moreau, B.D. Moor, Biclustering microarray data by Gibbs sampling, Bioinformatics 19 (2) (2003) 196–205.

[19] P. Carmona-Saez, R.D. Pascual-Marqui, F. Tirado, J.M. Carazo, A. Pascual-Montano, Biclustering of gene expression data by non-smooth non-negative matrix factorization, BMC Bioinformatics 7 (78) doi: 10.1186/1471–2105–7–78.

[20] S.C. Madeira, A.L. Oliveira, A linear time biclustering algorithm for time series gene expression data, in: Algorithms in Bioinformatics. International Workshop No. 5, Mallorca, Spain, 2005, pp. 39–52.

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

[22] W. Ayadi, M. Elloumi, J.-K. Hao, Iterated local search for biclustering of microarray data, in: Proceedings of the 5th International Conference on Pattern Recognition in Bioinformatics, PRIB, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 219–229.

[23] M. Bicego, P. Lovato, A. Ferrarini, M. Delledonne, Biclustering of expression microarray data with topic models, in: Proceedings of the 20th International Conference on Pattern Recognition, ICPR, IEEE Computer Society, Washington, DC, USA, 2010, pp. 2728–2731.
[24] X. Liu, L. Wang, Computing the maximum similarity bi-clusters of gene expression data, Bioinformatics 23 (2007) 50–56.
[25] B. Hanczar, M. Nadif, Bagging for biclustering: application to microarray data, in: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases, 2010, pp. 490–505.
[26] B. Hanczar, M. Nadif, Bagged biclustering for microarray data, in: Proceeding of the 19th European Conference on Artificial Intelligence, IOS Press, Amsterdam, The Netherlands, 2010, pp. 1131–1132 URL: http://portal.acm.org/citation.cfm?id=1860967.1861240.
[27] J.S. Aguilar-Ruiz, Shifting and scaling patterns from gene expression data, Bioinformatics 21 (2005) 3840–3845.
[28] S. Tavazoie, J. Hughes, M. Campbell, R. Cho, G. Church, Systematic determination of genetic network architecture, Nat. Genet. 22 (1999) 281–285.
[29] A.A. Alizadeh, M.B. Eisen, R.E. Davis, et al., Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling, Nature 403 (2000) 503–511.
[30] X. Xu, Y. Lu, A.K.H. Tung, W. Wang, Mining shifting-and-scaling co-regulation patterns on gene expression profiles, 2006, pp. 89–99.
[31] D. Corne, K. Deb, P.J. Fleming, The good of the many outweighs the good of the one: evolutionary multi-objective optimization, IEEE Connect. Newslett. 1 (1) (2003) 9–13.
[32] R. Steuer, Multiple Criteria Optimization: Theory, Computations, and Application, John Wiley & Sons, Inc., New York, 1986.
[33] C.A.C. Coello, A comprehensive survey of evolutionary-based multiobjective optimization techniques, Knowl. Inf. Syst. 1 (3) (1999) 129–156.
[34] C.A.C. Coello, G.B. Lamont, D.A.V. Veldhuizen, Evolutionary Algorithms for Solving Multi-Objective Problems (Genetic and Evolutionary Computation), Springer-Verlag, New York, Inc., Secaucus, NJ, USA, 2006.
[35] E. Zitzler, K. Deb, L. Thiele, Comparison of multi-objective evolutionary algorithms on test functions of different difficulty, in: A.S. Wu (Ed.), Proceedings of the 1999 Genetic and Evolutionary Computation Conference. Workshop Program, Orlando, Florida, 1999, pp. 121–122.
[36] F. Divina, J.S. Aguilar-Ruiz, A multi-objective approach to discover biclusters in microarray data, in: Proceedings of the 16th Genetic and Evolutionary Computation Conference (GECCO-2007), ACM, 2007, pp. 385–392.
[37] J.S. Aguilar-Ruiz, F. Divina, Evolutionary biclustering of microarray data, in: Proceedings of the 3rd European Workshop on Evolutionary Computation and Bioinformatics, Lecture Notes on Computer Science, 2005, pp. 1–10.
[38] N. Srinivas, K. Deb, Multiobjective optimization using nondominated sorting in genetic algorithms, Evol. Comput. 2 (3) (1994) 221–248.
[39] E. Zitzler, L. Thiele, Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto evolutionary algorithm, IEEE Trans. Evol. Comput. 3 (4) (1999) 257–271.
[40] J. Yang, W. Wang, H. Wang, P.S. Yu, δ-clusters: capturing subspace correlation in a large data set, in: Proceedings of the 18th IEEE Conference on Data Engineering, 2002, pp. 517–528.
[41] R. Cho, M. Campbell, E. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. Wolfsberg, A. Gabrielian, D. Landsman, D. Lockhart, R. Davis, A genome-wide transcriptional analysis of the mitotic cell cycle, Mol. Cell 2 (1998) 65–73.
[42] K.G. Le Roch, Y. Zhou, P.L. Blair, M. Grainger, K.J. Moch, D.J. Haynes, D. La, A.A. Holder, S. Batalov, D.J. Carucci, E.A. Winzeler, Discovery of gene function by expression profiling of the malaria parasite life cycle, Science (5639) (2003) 1503–1508. doi:10.1126/science.1087025.
[43] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, L.M. Sturla, M. Angelo, M.E. McLaughlin, J.Y.H. Kim, L.C. Goumnerova, P.M. Black, C. Lau, J.C. Allen, D. Zagzag, J.M. Olson, T. Curran, C. Wetmore, J.A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D.N. Louis, J.P. Mesirov, E.S. Lander, T.R. Golub, Prediction of central nervous system embryonal tumour outcome based on gene expression, Nature 415 (2002) 436–442.
[44] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, E. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.
[45] X.W. et al., Large-scale temporal gene expression mapping of central nervous system development, 1998, pp. 334–339.
[46] L. Boyer, K. Plath, J. Zeitlinger, T. Brambrink, L. Medeiros, T. Lee, S. Levine, M. Wernig, A. Tajonar, M. Ray, et al., Polycomb complexes repress developmental regulators in murine embryonic stem cells, Nature 441 (7091) (2006) 349–353.
[47] E. Hartuv, A. Schmitt, J. Lange, S. Meier-Ewert, H. Lehrach, R. Shamir, An algorithm for clustering cdna fingerprints, Genomics 66 (2000) 249–256.
[48] M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, et al., Gene ontology: tool for the unification of biology. The Gene Ontology, Nat. Genet. 25 (2000) 25–29.
[49] N. Diaz-Diaz, J.S. Aguilar-Ruiz, GO-based functional dissimilarity of gene sets, BMC Bioinformatics 12 (1) (2011) 360+ doi:10.1186/1471-2105-12-360.
[50] S. Bauer, S. Grossmann, M. Vingron, P.N. Robinson, Ontologizer 2.0—a multifunctional tool for go term enrichment analysis and data exploration, Bioinformatics 24 (14) (2008) 1650–1651.