Finding biologically significant biclusters: a new function for co-expression evaluation

J.E. Luna-Taylor[1], C.A. Brizuela[2], and I.N. Alvarado[1]

[1]Department of Systems and Computation, ITLP, La Paz, B.C.S., Mexico
[2]Computer Sciences Department, CICESE, Ensenada, B.C., Mexico

Corresponding author:
J.E. Luna-Taylor[1]

Email address: eluna@itlp.edu.mx

ABSTRACT

Analysis of DNA microarray data has been very useful for experimental molecular biology, as it provides unprecedented opportunities to study a wide variety of biological processes. As a part of this analysis, biclustering has been consolidated as one of the first steps in the discovery of new knowledge. Biclustering consists in identifying clusters of genes that present coherent behavior patterns for a subset of experimental conditions. The measure to assess this consistency is a key factor in the quality of discovered biclusters. In this paper, we propose a new function (*VF*) to evaluate the coherence of biclusters. This function recognizes shifting, and positive and negative scaling patterns, more efficiently than well-known reported functions with a similar purpose. Also, the *VF* function identifies positive and negative scaling subpatterns, which may be of biological interest and have not previously been discussed in the literature. To assess the performance of the *VF* function, a biclustering genetic algorithm (*BGA_VF*) was also designed, and tested on both synthetic and real data. The results show that the *BGA_VF* algorithm obtains high percentages of significant biclusters and recognizes all the analyzed combinations of coherence patterns.

INTRODUCTION

The analysis of DNA microarray data has been very useful for experimental molecular biology, as it provides unprecedented opportunities to study a wide variety of genes and their association with biological processes or metabolic functions (Fan and Ren, 2006; Hackl et al., 2004; Mischel et al., 2004). For instance, this kind of analysis has made possible to establish some correlations between genes expression and metabolic diseases, cancer, response to drug treatment, and response to different stress conditions in a specific organism (Macgregor, 2003).

Due to the vast amount of gene expression data produced during the last decades, and to the inherent complexity of their analysis, some computational techniques have been developed to assess their processing and interpretation (Raza, 2010; Slonim and Yanai, 2009). Biclustering is a widely used technique to analyze gene expression data, and it is one of the first stages in the gene expression analysis. It consists of finding groups of genes which are closely related under a subset of biological conditions, these groups are named biclusters (Cheng and Church, 2000).

To formally define a bicluster, let us consider an $n \times m$ matrix A, where each element a_{ij} represents the expression level of gene i under condition j. In general, the matrix A is regarded as a set of rows X and a set of columns Y, where an element a_{ij} corresponds to a real value that represents the relationship between row i and column j. A bicluster $A_{IJ} = (I,J)$ is a subset of rows $I \subseteq X$ and a subset of columns $J \subseteq Y$. Thus, the biclustering is defined as: given a matrix data A identify a set of biclusters $B_k = (I_k, J_k)$ such that each bicluster B_k fulfills some homogeneity constraints (Madeira and Oliveira, 2004).

Specific homogeneity characteristics of biclusters may vary from one problem statement to another. A good measure of homogeneity should be able to identify shifting and scaling patterns between the expressions levels of the genes that form part of a bicluster (Aguilar, 2005; Chen et al., 2015). A bicluster $B = (I,J)$ exhibits a shifting pattern if its element b_{ij} satisfies the condition:

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3110v2 | CC BY 4.0 Open Access | rec: 1 Aug 2017, publ: 1 Aug 2017

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

where π_j is the j_{th} base column value and β_i is the shifting factor to the i_{th} row. A bicluster $B = (I,J)$ displays a scaling pattern if the elements of the bicluster satisfy the condition:

$$b_{ij} = \alpha_i \pi_j$$

where π_j is the base value of the j_{th} column, and α_i is the scaling factor for the i_{th} row. If the concepts of shifting and scaling patterns are integrated, a bicluster $B = (I,J)$ shows a lineal pattern if every element b_{ij} satisfies the condition:

$$b_{ij} = \alpha_i \pi_j + \beta_i$$

where π_j is the base value of j_{th} column, while α_i and β_i are the scaling and shifting factors, respectively for the i_{th} row.

From year 2000, algorithms to generate significant biclusters have been developed. Cheng and Church (2000) were the pioneers in biclustering algorithms applied to gene expression data. Additionally, they proposed a measure known as MSR to evaluate biclusters coherence, measure that is widely used and analyzed in the literature (Pontes et al., 2015a). Prelic et al. (2006) introduced an evaluation and comparison for five outstanding methods: *CC* (Cheng and Church, 2000), *Samba* (Tanay et al., 2002), *OPSM* (Ben-Dor et al., 2004), *ISA* (Ihmels et al., 2002, 2004), and *xMotif* (Murali and Kasif, 2003), both real and simulated datasets were used to assess them. Regarding real data, biological information from GO annotations (Ashburner et al., 2000; Gasch et al., 2000), metabolic pathway maps (Gasch et al., 2000), and information about protein-protein interaction (Wille et al., 2004; Gasch et al., 2000) were used. Regarding the results presented in these studies, the approaches that obtained the best results were *ISA*, *Samba*, and *OPSM*.

Dharan and Nair (2009) developed *Reactive GRASP*, based on the generation of high-quality bicluster seeds by using the *k-means* algorithm (Hartigan and Wong, 1979), which evolved through restricted iterations. Pandey et al. (2009) introduced a method named *RAnge support Pattern* (*RAP*) based on a model of association pattern identification. This method uses a parameter referred to as *range support measure* to evaluate coherence among rows in a bicluster. Das and Idicula (2010) developed an algorithm based on greedy search mixed with the particle swarm optimization approach (*GS Binary PSO*). Further, Caldas and Kaski (2011) proposed a method based on a hierarchical model (*TreeBic*). The model assumes that microarray samples, or conditions, are grouped in a tree structure in which nodes correspond to hierarchical subsets. Nepomuceno et al. (2011) presented an approach based on an evolutionary computation technique (SScorr), and introduced a new fitness function based on the linear correlation among genes in a bicluster. Ayadi et al. (2012) proposed a pattern-driven neighborhood search algorithm (PDNS) that uses a bicluster pattern, both in its search space and in its neighborhood definition.

Evo-Bexpa is a proposed evolutionary algorithm (Pontes et al., 2013), which is able to discover shifting and positive scaling patterns in the behavior of genes in a bicluster. Based on the *NSGA-II* method (Deb et al., 2002) some algorithms (*MODPSFLB* (Liu et al., 2012), *PR-MOBI* (Seridi et al., 2013), and *eMOGB* (Brizuela et al., 2013)) model the biclustering as a multi-objective optimization problem (*MOO*). Although these algorithms are based on the same general strategy, they apply different heuristic techniques, such as evolutionary algorithms (*EA*), particle swarm optimization (*PSO*), and the shuffled frog-leaping algorithm (*SFL*).

The biclustering algorithms showed in literature used different search strategies which are guided by some functions to measure or evaluate the behavioral coherence of genes within biclusters. The kind of evaluation function used by the algorithms is a key factor in the quality of discovered biclusters. Some functions have been proposed more than a decade ago, and most of them are based on the identification of shifting and/or scaling patterns of the biclusters. A sumary of the ability to identify different patterns for some recognized functions are showed in Table 1 (Pontes et al., 2015b; Chen et al., 2015).

Only the *ACV* (Teng and Chan, 2008) and *MMSE* (Chen et al., 2015) functions are able to recognize perfect shifting and scaling patterns. However, a disadvantage for both functions is their computational complexity, which requires $O(|I|^2|J|)$ and $O(min(|I|,|J|)|I||J|)$ for *ACV* and *MMSE*, respectively. For the analysis of many biclusters or large sized biclusters, a large computation time is a clear disadvantage.

Function	A	B	C	D	E	F	Reference
MSR	√	√	X	X	X	X	(Cheng and Church, 2000)
ACV	√	√	√	√	√	√	(Teng and Chan, 2008)
ASR	√	√	√	√	X	X	(Ayadi et al., 2009)
VE	√	√	√	√	X	X	(Divina et al., 2012)
SMSR	√	X	√	X	√	X	(Mukhopadhyay et al., 2009)
MMSE	√	√	√	√	√	√	(Chen et al., 2015)

A. Perfect constant pattern, B. Perfect shifting pattern, C. Perfect scaling positive pattern, D. Perfect shifting and scaling positive pattern, E. Perfect scaling negative pattern, F. Perfect shifting and scaling negative pattern.

Table 1. Patterns identified by different bilcusters evaluation functions.

Based on the previous analysis of functions, we propose a new one to evaluate coherence within a bicluster. The function we propose (*VF*), is able not only to recognize shifting, and positive and negative scaling patterns, but also any combination of them. Furthermore, *VF* function identifies positive and negative scaling subpatterns. This means that *VF* identifies genes in a bicluster displaying a positive scaling pattern for a subset of experimental conditions and the same genes show a negative scaling pattern for a different subset of conditions. This behavioral pattern might have a biological meaning, and as far as we know it has not been considered in other functions. Another important characteristic of *VF* is the simplicity of its calculation, which only requires $O(|I||J|)$.

Besides, we designed a biclustering genetic algorithm (*BGA_VF*) to evaluate the biological significance of the identified biclusters when using *VF*. *BGA_VF* looks for the best bicluster according to the *VF* measure, given a range of desired gene number and conditions. The algorithm was tested with three real datasets: 1) Gasch's Yeast dataset (Gasch et al., 2000), 2) Leukemia dataset (Golub et al., 1999) and 3) Steminal dataset (Boyer et al., 2006). For all tests, the algorithm obtained high percentages of biclusters with statistical significance.

METHODS

In this work a new function to evaluate coherence within a bicluster is proposed. This function calculates a variation score of expression levels of genes in a bicluster. The function returns low scores for genes with similar expression pattern or higher values for non-similar ones. To test the performance of the proposed *VF* function, we also designed a biclustering genetic algorithm. This algorithm searches for biclusters with a minimum value of the variation function for any given pre-established range of numbers of genes and conditions.

The Proposed Variation Function

The proposed variation function (*VF*) takes into account the shifting patterns (additive model) as well as positives and negatives scaled patterns (multiplicative model). In other words, it considers that a set of genes has a similar behavior when despite the lack of identical expression values in the same subset of conditions, they show similar trends of under- and overexpression through such set of conditions. The *VF* function returns small values when the genes have similar expression levels.

Equation 1 shows the proposed variation function *VF* for a bicluster formed by a subset I of genes and a subset J of conditions. This formula is based on the ratio of change r_{ij} that is calculated by using Equation 2. The value r_{ij} represents the ratio of the change of expression level between conditions j and $j-1$ of gene i regarding the accumulated change of expression levels of all conditions of gene i. Where b_{ij} is the expression level of gene i under condition j. Equation 3 calculates r_{Ij} which is the mean of the ratios of change of all genes from condition j.

$$VF(I,J) = (|J|-1)\sum_{i\in I}\sum_{j\in J/\{1\}}\left|r_{ij}-r_{Ij}\right| \tag{1}$$

$$r_{ij} = \frac{\left|b_{ij}-b_{i(j-1)}\right|}{\sum_{j'\in J/\{1\}}\left|b_{ij'}-b_{i(j'-1)}\right|} \tag{2}$$

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3110v2 | CC BY 4.0 Open Access | rec: 1 Aug 2017, publ: 1 Aug 2017

$$r_{Ij} = \frac{1}{|I|} \sum_{i \in I} r_{ij} \tag{3}$$

The minimum possible value returned by the *VF* function is zero, which results for biclusters with perfect shifting and scaling patterns (see Appendix A). An example of a bicluster with a score *VF* equal to zero is shown in Fig. 1A. This bicluster has three genes that exhibit perfect shifting and scaling patterns with respect to each other. A small variation in the behavior pattern of some of the genes in the bicluster leads to a *VF* score greater than zero (Fig. 1B).

The maximum possible score calculated by the *VF* function for a bicluster is bounded by:

$$VF(I,J) \leq (2\,|I| - 2)(|J| - 1), \tag{4}$$

where $|I|$ is the number of genes and $|J|$ is the number of conditions in the bicluster (see Appendix B).

Algorithm and complexity

Algorithm 1 shows the calculation of the *VF* function for a bicluster. In the first block (lines 2-11), the calculation of ratio of change of expression (r_{ij}) is performed. In the second block (lines 13-19) the mean of the ratio of change for each condition (r_{Ij}) is obtained. In the last block, the final score is obtained from the double sum (lines 21-27). The computational cost for each of the three blocks is $O(|I||J|)$, and since they are independent, the computational time for the complete algorithm is also $O(|I||J|)$.

Algorithm 1. *VF* function calculation for a Bicluster.

Input: a matrix B of gene expression values of size $|I| \times |J|$.
Output: the *VF* score of the matrix B.

```
 1. //Calculation of ratio of change (Equation 2)
 2. for $i \leftarrow 1$ to $|I|$ do
 3.     $sum \leftarrow 0$
 4.     for $j \leftarrow 2$ to $|J|$ do
 5.         $d_{ij} \leftarrow \left| b_{ij} - b_{i(j-1)} \right|$
 6.         $sum \leftarrow sum + d_{ij}$
 7.     end
 8.     for $j \leftarrow 2$ to $|J|$ do
 9.         $r_{ij} \leftarrow \frac{d_{ij}}{sum}$
10.     end
11. end
12. //Mean ratio of change (Equation 3)
13. for $j \leftarrow 2$ to $|J|$ do
14.     $sum \leftarrow 0$
15.     for $i \leftarrow 1$ to $|I|$ do
16.         $sum \leftarrow sum + r_{ij}$
17.     end
18.     $r_{Ij} \leftarrow \frac{sum}{|I|}$
19. end
20. // VF final calculation (Equation 1)
21. $sum \leftarrow 0$
22. for $i \leftarrow 1$ to $|I|$ do
23.     for $j \leftarrow 2$ to $|J|$ do
24.         $sum = sum + \left| r_{ij} - r_{Ij} \right|$
25.     end
26. end
27. $VF \leftarrow (|J| - 1) \cdot sum$
```

Biclustering Genetic Algorithm

To test the performance of the *VF* function, we also propose a biclustering genetic algorithm. Following the idea of other evolutionary approaches (Mitra and Banka, 2006; Divina and Aguilar-Ruiz, 2006), a bicluster is represented as a two-section binary string where the first section corresponds to genes and the second section to conditions. If a given locus has an allele one, it indicates that its corresponding gene or condition is selected to be part of the bicluster.

The algorithm receives as input a gene expression matrix, a range of the expected number of genes and conditions, and a percentage of minimum quality accepted for the returned biclusters. These required values for the accepted biclusters are considered as hard constraints into the algorithm.

To generate the initial population each bicluster is constructed by performing a random selection of genes and conditions from the gene expression matrix. The parent selection process was made by applying binary tournament. In the binary tournament, a bicluster i is preferred to a bicluster j, if i fulfills the established restrictions and j does not, or if both fulfills the restrictions, but i has a lower *VF* score than j. The single-point crossover operator was used, applying it independently to the section of genes and to the section of conditions. For the mutation, a random position of the binary string is chosen, and its value is changed. Generational replacement with elitism was applied to generate the new population. The algorithm returns the discovered bicluster with the lowest *VF* score that also complies with the established constraints.

RESULTS AND DISCUSSIONS

Evaluation of the *VF* function with synthetic data

To evaluate the effectiveness of the *VF* function to recognize scaling and shifting patterns, six synthetic data sets proposed elsewhere (Chen et al., 2015; Teng and Chan, 2008; Ayadi et al., 2009) were used. Each of these data sets presents a different perfect pattern (A-F) (Table 2). Additionally, the results obtained using other functions evaluated in the work of Chen et al. (2015) are shown.

	Perfect Patterns						Optimal	
Function	**A**	**B**	**C**	**D**	**E**	**F**	**Values**	**Reference**
MSR	0.000	0.000	0.625	0.625	3.125	3.325	0	(Cheng and Church, 2000)
ACV	1.000	1.000	1.000	1.000	1.000	1.000	1	(Teng and Chan, 2008)
ASR	1.000	1.000	1.000	1.000	-0.200	-0.200	1, -1	(Ayadi et al., 2009)
VE	0.000	0.000	0.000	0.000	1.033	0.930	0	(Divina et al., 2012)
SMSR	0.000	0.089	0.000	0.021	0.000	3.458	0	(Mukhopadhyay et al., 2009)
MMSE	0.000	0,000	0.000	0.000	0.000	0.000	0	(Chen et al., 2015)
VF	0.000	0,000	0.000	0.000	0.000	0.000	0	This work

A. Constant, B. Shifting, C. Scaling positive, D. Shifting and scaling positive, E. Scaling negative, F. Shifting and scaling negative.

Table 2. Comparison of different evaluation functions of biclusters on synthetic test data.

These results show that the *VF* function is able to recognize different patterns of displacement, positive and negative scaling, as well as any combination of these. Of all the functions evaluated, only the *VF*, *ACV*, and *MMSE* functions recognized the six perfect pattern types. However, the *ACV* and *MMSE* functions have the disadvantage of having a higher computation cost, $O(|I|^2|J|)$ for *ACV*, and $O(min(|I|,|J|)|I||J|)$ for *MMSE*. The *VF* function has a simpler calculation, which requires an execution time of $O(|I||J|)$, that represents an important advantage when working with large volumes of biological data.

Test with real data

To evaluate and compare the effectiveness of the *BGA_VF* algorithm to recover significant or enriched biclusters for any GO category (Ashburner et al., 2000) three real gene expression datasets were analyzed. The Gasch's Yeast dataset (Gasch et al., 2000) which corresponds to expression levels of 2993 genes of Saccharomyces cerevisiae under 173 different stress conditions. The Leukemia dataset (Golub et al., 1999) containing the expression of 7129 genes from 25 patients suffering from acute myeloid leukemia (AML) and 47 suffering from acute lymphoblastic leukemia (ALL). Finally, we used the Steminal (Boyer et al., 2006) dataset that corresponds to the expression of 26127 genes for 30 time points of murine embryonic stem cells differentiation.

For each dataset, one hundred runs of the algorithm *BGA_VF* were made, selecting from each run the best bicluster, evaluated according to the *VF* function. The algorithm parameters were set as follows: a population size of 100 individuals, 1000 for the number of generations without an improvement on the best *VF* value found so far as stopping criterion, 90% rate for selecting the winner within the tournament, a 100% for the crossover rate, and 50% for the mutation rate. In each run of the algorithm, the biclusters were obtained within a range of 30 to 100 genes, a range of 5 to 20 experimental conditions, and 80% minimum for the quality of biclusters (a value for *VF* not greater than 20% of its upper bound). The details of the discovered biclusters are provided in the Supplemental Files S1-S7.

Evaluation of statistical significance of obtained biclusters with Yeast dataset

The obtained results for the algorithm *BGA_VF* from the yeast dataset were compared with the ones produced by other well-known biclustering algorithms: *OPSM* (Ben-Dor et al., 2004), *ISA* (Ihmels et al., 2002, 2004), *CC* (Cheng and Church, 2000), and *SScorr* (Nepomuceno et al., 2011). The results of algorithms *OPSM*, *ISA*, and *CC* were generated by using the Biclustering Analysis Toolbar (*BicAT*) software (Barkow et al., 2006). To evaluate the algorithms according to the percentage of significant biclusters recovered, the *AGO* tool (Akwaa and Kadah, 2009) was used. For this evaluation, *BGA_VF* accomplished between 89% and 100% of significant biclusters discovered for p-values in the range of $1e-5$ to $5e-2$ (Fig. 2). These percentages were higher than the percentages obtained by the other evaluated algorithms.

A comparison of the percentage of significant biclusters after filtering biclusters that did not overlap more than 25% was performed too. This filter is important due to the possibility to determine whether an algorithm can find diversity in the sets of discovered genes. On this comparison *SScorr* was not included since data about overlapping constraints are not available for this algorithm. For this evaluation *BGA_VF* obtained between 70% - 100% of significant biclusters for different p-values (Fig. 3). These percentages were higher than those obtained by *ISA* and *CC* algorithms. In this test, *OPSM* acquired 100% of significant biclusters; however, this percentage corresponds to just two biclusters generated without overlapping (Table 3).

Algorithm	Total Number of Biclusters	Biclusters Filtered	Percentage Biclusters Filtered	Reference
OPSM	19	2	10.5%	(Ben-Dor et al., 2004)
ISA	63	20	37.7%	(Ihmels et al., 2004)
CC	100	56	56%	(Cheng and Church, 2000)
BGA_VF	100	27	27%	This work

Table 3. Comparison of the quantity and percentage of biclusters without overlap found out by *OPSM*, *ISA*, *CC*, and *BGA_VF* algorithms.

Identified patterns for biclusters obtained on the Yeast dataset

Next, we wanted to identify the different types of patterns discovered by the *BGA_VF* algorithm in the yeast dataset. From the 100 generated biclusters by the *BGA_VF* we found biclusters with shifting and positive scaling patterns (Fig. 4A), shifting, and positive and negative scaling patterns (Fig. 4B), and interestingly, biclusters with positive and negative scaling subpatterns (i.e., patterns within a bicluster) where also identified (Figs. 4C and 4D). In the latter case, a bicluster showed one gene with a negative scaling pattern only in the stress condition 126; while the same gen showed a positive scaling pattern for the other conditions (Fig. 4C). In another case, a bicluster showed one gene with a negative scaling pattern regarding other genes for the 43, 62, and 64 stress conditions, and a positive scaling pattern for the other conditions (Fig. 4D).

Although finding subpatterns was not the goal of the designed function, our results suggest that the *VF* function can be useful to identifying related genes, in a same biological function or molecular process, that show different scaling (positives and negatives) subpatterns according to the evaluated experimental conditions. This behavioral might have a important biological meaning, and as far as we know it has not been considered in other functions.

Evaluation of statistical significance of biclusters obtained from Leukemia and Steminal datasets

To evaluate the statistical significance of the biclusters found by the *BGA_VF* on the Leukemia and Steminal datasets, the software *g:Profiler* (Reimand et al., 2016) with the Bonferroni correction was used. The results found by the *BGA_VF* were compared to results reported for the *SMOB* method (Fig. 5). The *SMOB* method achieves the coherence evaluation of the biclusters through the *VE* and *MSR* functions (Divina et al., 2012). With both datasets, the percentage of biclusters found by *BGA_VF* were higher than the one obtained by the *SMOB* algorithm by using *VE* and *MSR* functions independently.

The results obtained with the Yeast, Leukemia, and Steminal datasets showed that the algorithm *BGA_VF* is effective in the identification of biclusters with statistical significance. In all evaluated cases, the *BGA_VF* algorithm identified a higher percentage of significant biclusters than the other compared methods. These favorable results were maintained by considering only biclusters that do not overlap in more than 25% of the genes they contain. On the other hand, although the design of the *BGA_VF* algorithm did not focus on avoiding the overlap of biclusters, on the Leukemia and Steminal datasets, high percentages of biclusters without overlap (98% and 100%, respectively) were acquired.

CONCLUSIONS

In this work, a new function named *VF* to evaluate the coherence of biclusters, was proposed. The *VF* function identifies any combination of shifting and scaling patterns, both positive and negative, faster than functions reported in the literature for the same objective. Also, *VF* recognizes a new pattern not discussed in the literature, which may correspond to groups of related genes under the same biological function or molecular process. On the other hand, supported by the algorithm *BGA_VF*, the *VF* function is able to discover high percentages of biclusters with statistical significance, as well as high percentages of biclusters without overlap, especially for large databases.

We conclude that the *VF* function is effective because it obtains high percentages of significant biclusters and recognizes all combinations of discussed coherent patterns. Also, the *VF* function is efficient since it requires a small computation effort, which is a very important feature when it is required to process large volumes of expression data.

ACKNOWLEDGMENTS

We will like to thanks Andrea Murillo for her valuable support in the writing and translation of this paper.

APPENDICES

Appendix A. Optimal value of *VF* function

Proposition 1.

A bicluster that shows a shifting and/or scaling perfect pattern has a zero value of the *VF* function.

Proof:

We start by proving that for every bicluster of interest (with at least two experimental conditions) whose value *VF=0*, it follows that:

$$r_{ij} = r_{i'j}, \forall (i \neq i'), i' \in I, \forall j \in J/\{1\}.$$

Given the formula for calculating *VF*:

$$VF(I,J) = (|J|-1) \sum_{i \in I} \sum_{j \in J/\{1\}} \left| r_{ij} - r_{Ij} \right|,$$

the only way that *VF* equals zero is that the double sum is equal to zero, since $(|J|-1) > 0$ for any bicluster with at least two experimental conditions. And considering that only non-negative values are added, the only way that the double sum is zero is that all the summed values are zero, that is:

$$\left| r_{ij} - r_{Ij} \right| = 0, \forall i \in I, \forall j \in J/\{1\},$$

which implies that:

$$r_{ij} = r_{Ij}, \forall i \in I, \forall j \in J/\{1\},$$

and by transitivity we have to:

$$r_{ij} = r_{i'j}, \forall (i \neq i'), i' \in I, \forall j \in J/\{1\}.$$

On the other hand, we prove that by applying a scaling factor (either positive or negative) and/or an additive value to all levels of expression of a gene does not change the ratio r_{ij} of that gene.

Given the calculation of r_{ij}:

$$r_{ij} = \frac{\left|b_{ij} - b_{i(j-1)}\right|}{\sum_{j' \in J/\{1\}} \left|b_{ij'} - b_{i(j'-1)}\right|},$$

if we apply an arbitrary scaling factor c, and an additive value d also arbitrary, to each expression value of gene i, we have:

$$r_{ij} = \frac{\left|(c \cdot b_{ij} + d) - (c \cdot b_{i(j-1)} + d)\right|}{\sum_{j' \in J/\{1\}} \left|(c \cdot b_{ij'} + d) - (c \cdot b_{i(j'-1)} + d)\right|},$$

where the additive values cancel each other:

$$r_{ij} = \frac{\left|(c \cdot b_{ij} + \cancel{d}) - (c \cdot b_{i(j-1)} + \cancel{d})\right|}{\sum_{j' \in J/\{1\}} \left|(c \cdot b_{ij'} + \cancel{d}) - (c \cdot b_{i(j'-1)} + \cancel{d})\right|},$$

taking c as a common factor we have:

$$r_{ij} = \frac{\left|c \cdot (b_{ij} - b_{i(j-1)})\right|}{\sum_{j' \in J/\{1\}} \left|c \cdot (b_{ij'} - b_{i(j'-1)})\right|},$$

we extract c as positive value of the absolute operator, and being a constant value we can extract it from the summation:

$$r_{ij} = \frac{\cancel{c} \cdot \left|(b_{ij} - b_{i(j-1)})\right|}{\cancel{c} \cdot \sum_{j' \in J/\{1\}} \left|(b_{ij'} - b_{i(j'-1)})\right|},$$

resulting in the original formula for r_{ij}:

$$r_{ij} = \frac{\left|b_{ij} - b_{i(j-1)}\right|}{\sum_{j' \in J/\{1\}} \left|b_{ij'} - b_{i(j'-1)}\right|}.$$

The latter indicates that two i and i' genes with perfect scaling patterns and/or additives terms will have the same ratios of change for each experimental condition:

$$r_{ij} = r_{i'j}, \forall (i \neq i'), i' \in I, \forall j \in J/\{1\},$$

which, as previously proved, is the case when the VF function returns a zero. Therefore, a zero value returned by the VF function corresponds to perfect scaling patterns and/or additives of the behavior of the genes of a bicluster.

Appendix B. Upper bound for the *VF* function

Proposition 2.

$VF(I,J)$ is bounded as follows:

$$VF(I,J) \leq (|J| - 1)(2|I| - 2).$$

Proof:

We start by proving that for any gene i the sum of its ratios of changes is equal to 1:

$$\sum_{j\in J/\{1\}} r_{ij} = 1, \forall i \in I.$$

Given the formula of ratio of change for gene i in the condition j:

$$r_{ij} = \frac{\left|b_{ij} - b_{i(j-1)}\right|}{\sum_{j'\in J/\{1\}} \left|b_{ij'} - b_{i(j'-1)}\right|},$$

we have to:

$$\sum_{j\in J/\{1\}} r_{ij} = \sum_{j\in J/\{1\}} \frac{\left|b_{ij} - b_{i(j-1)}\right|}{\sum_{j'\in J/\{1\}} \left|b_{ij'} - b_{i(j'-1)}\right|},$$

$$\sum_{j\in J/\{1\}} r_{ij} = \frac{\sum_{j\in J/\{1\}} \left|b_{ij} - b_{i(j-1)}\right|}{\sum_{j'\in J/\{1\}} \left|b_{ij'} - b_{i(j'-1)}\right|} = 1,$$

since j and j' take the same set of values.

On the other hand, by developing the internal summation of the *VF* formula, we have for gene i:

$$\sum_{j\in J/\{1\}} \left|r_{ij} - r_{Ij}\right| = \sum_{j\in J/\{1\}} \left| r_{ij} - \frac{\sum_{i'\in I} r_{i'j}}{|I|} \right|$$

$$= \sum_{j\in J/\{1\}} \left| r_{ij} - \frac{(r_{1j} + r_{2j} + \ldots + r_{ij} + \ldots + r_{|I|j})}{|I|} \right|$$

$$= \sum_{j\in J/\{1\}} \left| r_{ij} - \frac{r_{1j}}{|I|} - \frac{r_{2j}}{|I|} - \ldots - \frac{r_{ij}}{|I|} - \ldots - \frac{r_{|I|j}}{|I|} \right|$$

$$= \sum_{j\in J/\{1\}} \left| (r_{ij} - \frac{r_{ij}}{|I|}) - \frac{r_{1j}}{|I|} - \frac{r_{2j}}{|I|} - \ldots - \frac{r_{|I|j}}{|I|} \right|$$

$$= \sum_{j\in J/\{1\}} \left| \frac{(|I|-1)r_{ij}}{|I|} - \frac{r_{1j}}{|I|} - \frac{r_{2j}}{|I|} - \ldots - \frac{r_{|I|j}}{|I|} \right|$$

$$= \sum_{j\in J/\{1\}} \left| \frac{(|I|-1)r_{ij} - r_{1j} - r_{2j} - \ldots - r_{|I|j}}{|I|} \right|$$

$$= \frac{1}{|I|} \sum_{j\in J/\{1\}} \left| (|I|-1)r_{ij} - r_{1j} - r_{2j} - \ldots - r_{|I|j} \right|$$

$$= \frac{1}{|I|} \sum_{j\in J/\{1\}} \left| (r_{ij} - r_{1j}) + (r_{ij} - r_{2j}) + \ldots + (r_{ij} - r_{|I|j}) \right|$$

$$= \frac{1}{|I|} \sum_{j\in J/\{1\}} \left| \sum_{i'\in I/\{i\}} (r_{ij} - r_{i'j}) \right|,$$

we take an upper bound:

$$\frac{1}{|I|} \sum_{j\in J/\{1\}} \left| \sum_{i'\in I/\{i\}} (r_{ij} - r_{i'j}) \right| \le \frac{1}{|I|} \sum_{j\in J/\{1\}} \left| \sum_{i'\in I/\{i\}} (\left|r_{ij}\right| + \left|r_{i'j}\right|) \right|,$$

then:

$$\sum_{j \in J/\{1\}} \left| r_{ij} - r_{Ij} \right| \leq \frac{1}{|I|} \sum_{j \in J/\{1\}} \left| \sum_{i' \in I/\{i\}} (\left| r_{ij} \right| + \left| r_{i'j} \right|) \right|,$$

and, based on its formula (Equation 2), we know that every value r_{ij} is always positive, so:

$$r_{ij} = \left| r_{ij} \right|, \forall i \in I, \forall j \in J/\{1\},$$

therefore:

$$\begin{aligned}\sum_{j \in J/\{1\}} \left| r_{ij} - r_{Ij} \right| &\leq \frac{1}{|I|} \sum_{j \in J/\{1\}} \sum_{i' \in I/\{i\}} (r_{ij} + r_{i'j}) \\ &= \frac{1}{|I|} \sum_{i' \in I/\{i\}} \sum_{j \in J/\{1\}} (r_{ij} + r_{i'j}) \\ &= \frac{1}{|I|} \sum_{i' \in I/\{i\}} (\sum_{j \in J/\{1\}} r_{ij} + \sum_{j \in J/\{1\}} r_{i'j}),\end{aligned}$$

previously it was demonstrated that:

$$\sum_{j \in J/\{1\}} r_{ij} = 1, \forall i \in I,$$

then, we have that:

$$\sum_{j \in J/\{1\}} \left| r_{ij} - r_{Ij} \right| \leq \frac{1}{|I|} (\sum_{i' \in I/\{i\}} 2) = \frac{1}{|I|} (|I| - 1)(2) = 2 - \frac{2}{|I|}.$$

taking this value as the upper bound for all bicluster genes, we have:

$$\sum_{i \in I} \sum_{j \in J/\{1\}} \left| r_{ij} - r_{Ij} \right| \leq |I| (2 - \frac{2}{|I|}) = 2|I| - 2.$$

therefore, this proves that an upper bound for the *VF* function is:

$$VF(I,J) = (|J| - 1) \sum_{i \in I} \sum_{j \in J/\{1\}} \left| r_{ij} - r_{Ij} \right| \leq (|J| - 1)(2|I| - 2).$$

In addition, it was proved experimentally that this bound is tight, since it was reached for certain biclusters (Fig. 6).

REFERENCES

Aguilar, J. (2005). Shifting and scaling patterns from gene expression data. *Bioinformatics*, 21:3840–3845.

Akwaa, F. A. and Kadah, Y. (2009). An automatic gene ontology software tool for bicluster and cluster comparisons. In *Proceedings of the 6th Annual IEEE conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB'09)*, pages 163–167.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25(1):25–29.

Ayadi, W., Elloumi, M., and Hao, J. (2009). A biclustering algorithm based on a bicluster enumeration tree: application to dna microarray data. *BioData Mining*, 2(9).

Ayadi, W., Elloumi, M., and Hao, J. (2012). Pattern-driven neighborhood search for biclustering of microarray data. *BMC Bioinformatics*, 13(Suppl 7):S11.

Barkow, S., Bleuler, S., Prelic, A., Zimmermann, P., and Zitzler, E. (2006). Bicat: a biclustering analysis toolbox. *Bioinformatics*, 22(10):1282–1283.

Ben-Dor, A., Chor, B., Karp, R., and Yakhini, Z. (2004). Discovering local structure in gene expression data: The order-preserving submatrix problem. *Journal of Computational Biology*, 10(3-4):373–384.

Boyer, L., Plath, K., Zeitlinger, J., Brambrink, T., and L. Medeiros, e. a. (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*, 441:349–353.

Brizuela, C. A., Luna-Taylor, J. E., Martinez-Perez, I., Guillen, H. A., Rodriguez, D. O., and Beltran-Verdugo, A. (2013). Improving an evolutionary multi-objective algorithm for the biclustering of gene expression data. In *proceedings of IEEE Congress on Evolutionary Computation*.

Caldas, J. and Kaski, S. (2011). Hierarchical generative biclustering for microrna expression analysis. *Journal of Computational Biology*, 18(3):251–261.

Chen, S., Liu, J., and Zeng, T. (2015). Measuring the quality of linear patterns in biclusters. *Methods*, 83:18–27.

Cheng, Y. and Church, G. M. (2000). Biclustering of expression data. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB'00)*, pages 93–103.

Das, S. and Idicula, S. M. (2010). Greedy search-binary pso hybrid for biclustering gene expression data. *International Journal of Computer Applications*, 2(3):0975–8887.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197.

Dharan, S. and Nair, A. S. (2009). Biclustering of gene expression data using reactive greedy randomized adaptive search procedure. *BMC Bioinformatics*, 10(Suppl. 1):S27.

Divina, F. and Aguilar-Ruiz, J. S. (2006). Biclustering of expression data with evolutionary computation. *IEEE Transactions on Knowledge and Data Engineering*, 18(5):590–602.

Divina, F., Pontes, B., Giraldez, R., and Aguilar-Ruiz, J. S. (2012). An effective measure for assessing the quality of biclusters. *Computers in Biology and Medicine*, 42:245–256.

Fan, J. and Ren, Y. (2006). Statistical analysis of dna microarray data in cancer research. *Clin Cancer Res*, 12(15):4469–73.

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., and Brown, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, 11:4241–4257.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., and et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537.

Hackl, H., Sanchez, C. F., Sturn, A., Wolkenhauer, O., and Trajanoski, Z. (2004). Analysis of dna microarray data. *Curr Top Med Chem*, 4(13):1257–70.

Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. *Applied Statistics*, 28:100–108.

Ihmels, J., Bergmann, S., and Barkai, N. (2004). Defining transcription modules using large-scale gene expression data. *Bioinformatics*, 20:1993–2003.

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., and Barkai, N. (2002). Revealing modular organization in the yeast transcriptional network. *Nature Genetics*, 31:370–377.

Liu, J., Li, Z., Hu, X., Chen, Y., and Liu, F. (2012). Multi-objetive dynamic population shuffled frog-leaping biclustering of microarray data. *BMC Genomics*, 13(Suppl. 3):S3–S6.

Macgregor, P. F. (2003). Gene expression in cancer: the application of microarrays. *Expert Rev. Mol. Diagn.*, 3(2):185–200.

Madeira, S. C. and Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1):24–45.

Mischel, P. S., Cloughesy, T. F., and Nelson, S. F. (2004). Dna-microarray analysis of brain cancer: molecular classification for therapy. *Nature Reviews Neuroscience*, 5:782–792.

Mitra, S. and Banka, H. (2006). Multi-objective evolutionary biclustering of gene expression data. *Journal of the Pattern Recognition Society*, 39:2464–2477.

Mukhopadhyay, A., Maulik, U., and Bandyopadhyay, S. (2009). A novel coherence measure for discovering scaling biclusters from gene expression data. *Journal of Bioinformatics and Computational Biology*, 7:853–868.

Murali, T. M. and Kasif, S. (2003). Extracting conserved gene expression motifs from gene expression data. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 8, pages 77–88.

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

Pandey, G., Atluri, G., Steinbach, M., Myers, C. L., and Kumar, V. (2009). An association analysis approach to biclustering. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Pontes, B., Giraldez, R., and Aguilar-Ruiz, J. S. (2015a). Biclustering on expression data: A review. *Journal of Biomedical Informatics*, 57:163–180.

Pontes, B., Girldez, R., and Aguilar-Ruiz, J. S. (2013). Configurable pattern-based evolutionary biclustering of gene expression data. *Algorithms for Molecular Biology*, (8):4.

Pontes, B., Girldez, R., and Aguilar-Ruiz, J. S. (2015b). Quality measures for gene expression biclusters. *PLoS ONE*, 10(3):1–24.

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

Raza, K. (2010). Application of data mining in bioinformatics. *Indian Journal of Computer Science and Engineering*, 1(2):114–118.

Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., and Vilo, J. (2016). g:profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Research*, 44(W1).

Seridi, K., Jourdan, L., and Talbi, E.-G. (2013). Multiobjective path relinking for biclustering: Application to microarray data. In *Evolutionary Multi-Criterion Optimization*, pages 200–214.

Slonim, D. and Yanai, I. (2009). Getting started in gene expression microarray analysis. *PLoS Comput Biol*, 5(10).

Tanay, A., oded Sharan, and Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18(Suppl. 1):S136–S144.

Teng, L. and Chan, L. (2008). Discovering biclusters by iteratively sorting with weighted correlation coefficient in gene expression data. *Journal of Signal Processing Systems*, 50(3):267–280.

Wille, A., Zimmermann, P., Vranova, E., Furholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W., and Buhlmann, P. (2004). Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome Biology*, 5:R92.

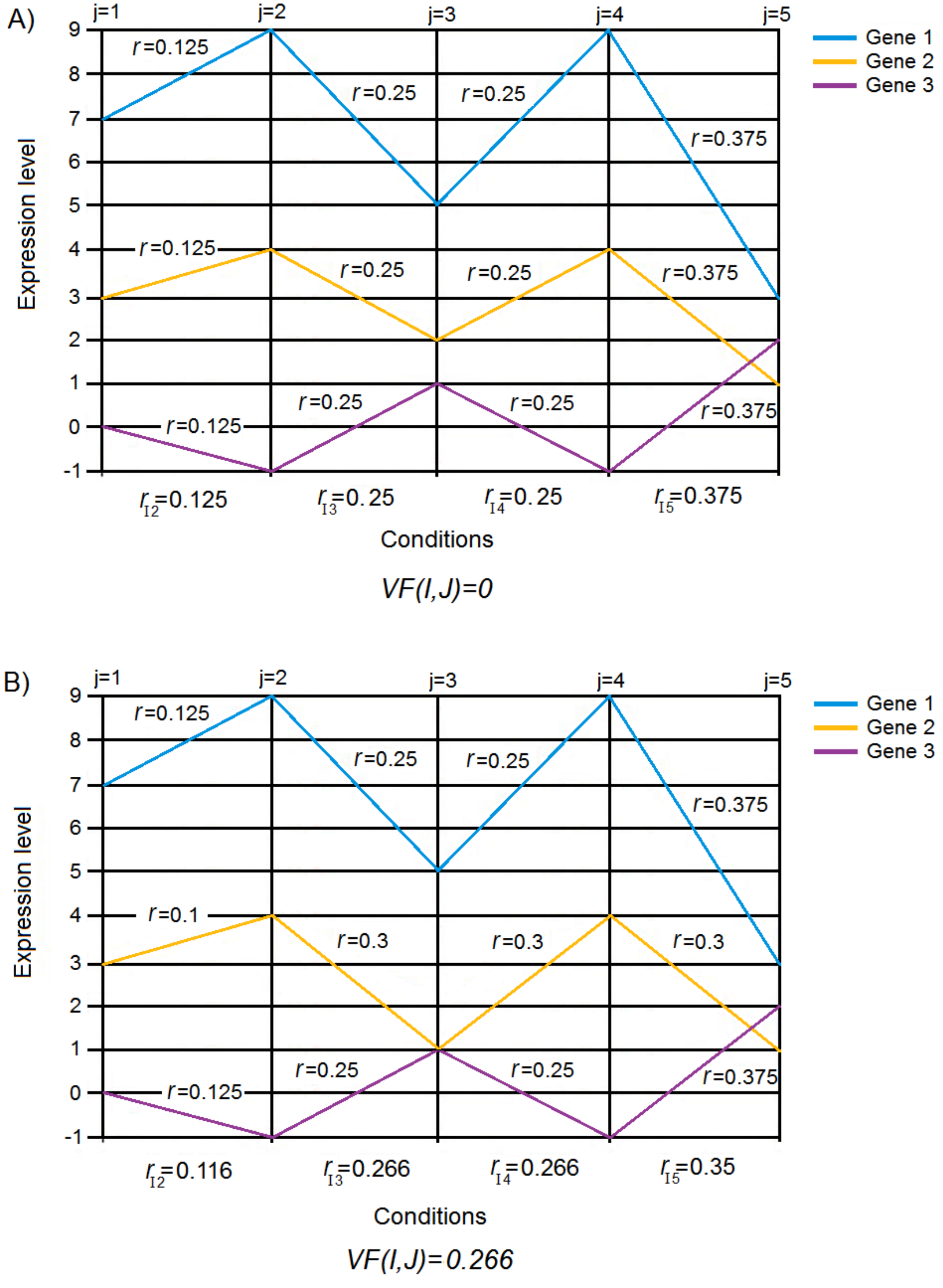


Figure 1. Examples of VF values for different expression patterns. A) Three genes with identical behavior, that show perfect scaling and shifting patterns with each other, whose calculation of score variation is equal to zero. B) Three genes with a similar behavior with a variation score slightly greater than zero. For this bicluster, the expression of gene 2 showed in A was modified from 2 to 1 in condition $j = 3$. Thus the perfect scaling and shifting pattern of this gene with respect to genes 1 and 3 is lost.

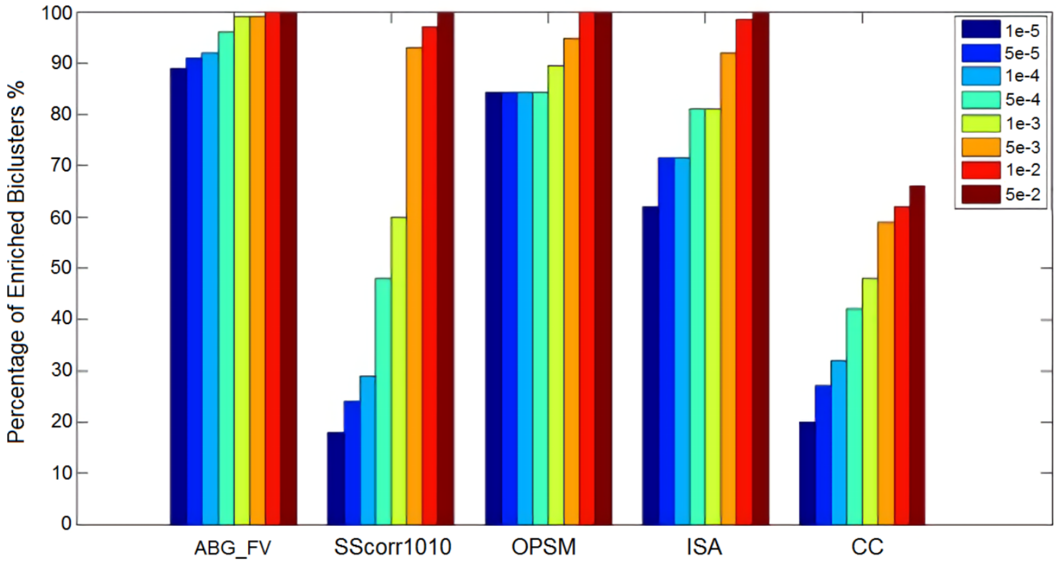


Figure 2. Comparison of percentage of significant biclusters found on the yeast dataset for different p-values. From the Yeast dataset one hundred biclusters were identified with the algorithm *BGA_VF*. The parameters used in *BicAT* were $l = 100$ for *OPSM*; $t_g = 2.0$, $t_c = 2.0$, $seeds = 500$ for *ISA*; and $\delta = 0.5$, $\alpha = 1.2$, and $M = 100$ for *CC* (Akwaa and Kadah, 2009). The results of algorithm *SScorr* were taken from (Nepomuceno et al., 2011) .

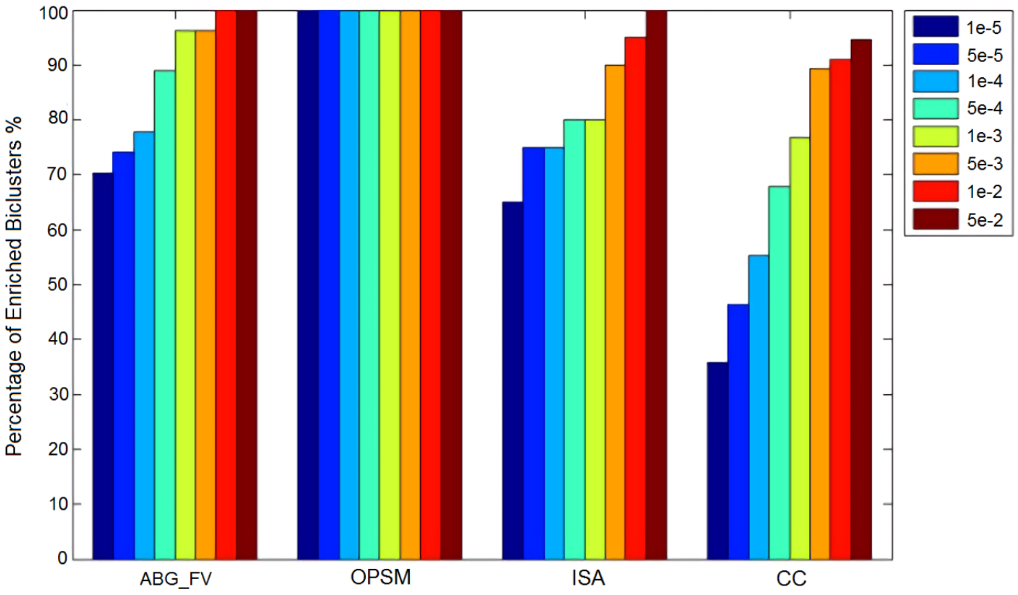


Figure 3. Comparison of the percentage of significant biclusters without overlap on the Yeast dataset for different p-values. From Yeast data base one hundred biclusters were identified through the algorithm *BGA_VF*. Following the parameters taken from (Akwaa and Kadah, 2009) the algorithms *OPSM*, *ISA*, and *CC* were executed. Subsequently, a filter was applied to all methods; only biclusters without an overlap higher than 25% of containing genes were kept.

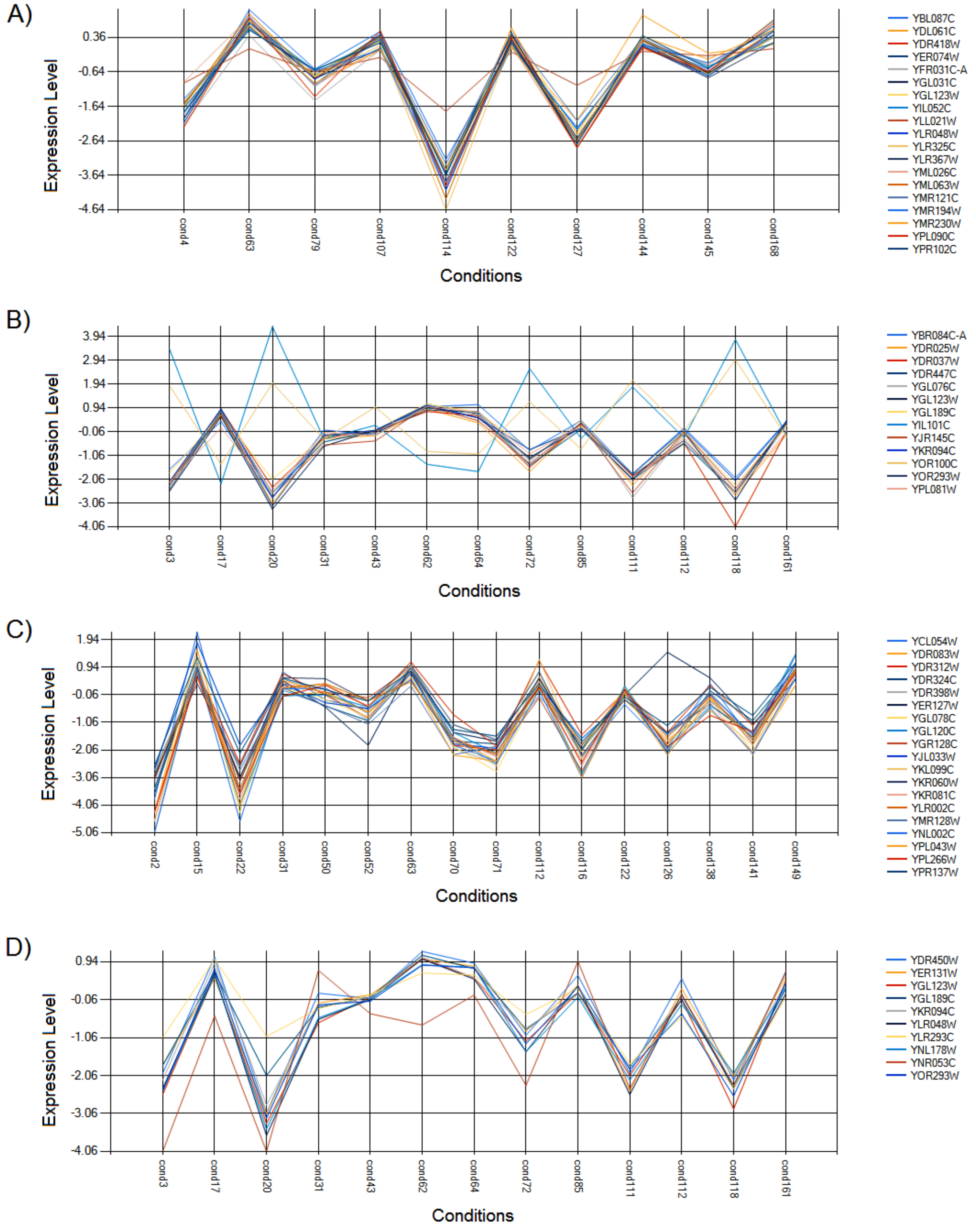


Figure 4. Identified patterns by the *BGA_VF* algorithms on the yeast dataset. One hundred biclusters were identified on the yeast dataset by the *BGA_VF* algorithm. Biclusters with shifting and positive scaling patterns (A), shifting, and positive and negative scaling patterns (B), as well as positive and negative subpatterns (C and D) were found. The figures correspond to four real biclusters discovered by the *BGA_VF* algorithm, for each one, the genes belonging to the same category are shown: GO:0005198 structural molecule activity (A), TF:M07442_0 Factor Rap1p (B), GO:0030684 preribosome (C) and GO:0071428 rRNA-containing ribonucleoprotein complex export from nucleus (D).

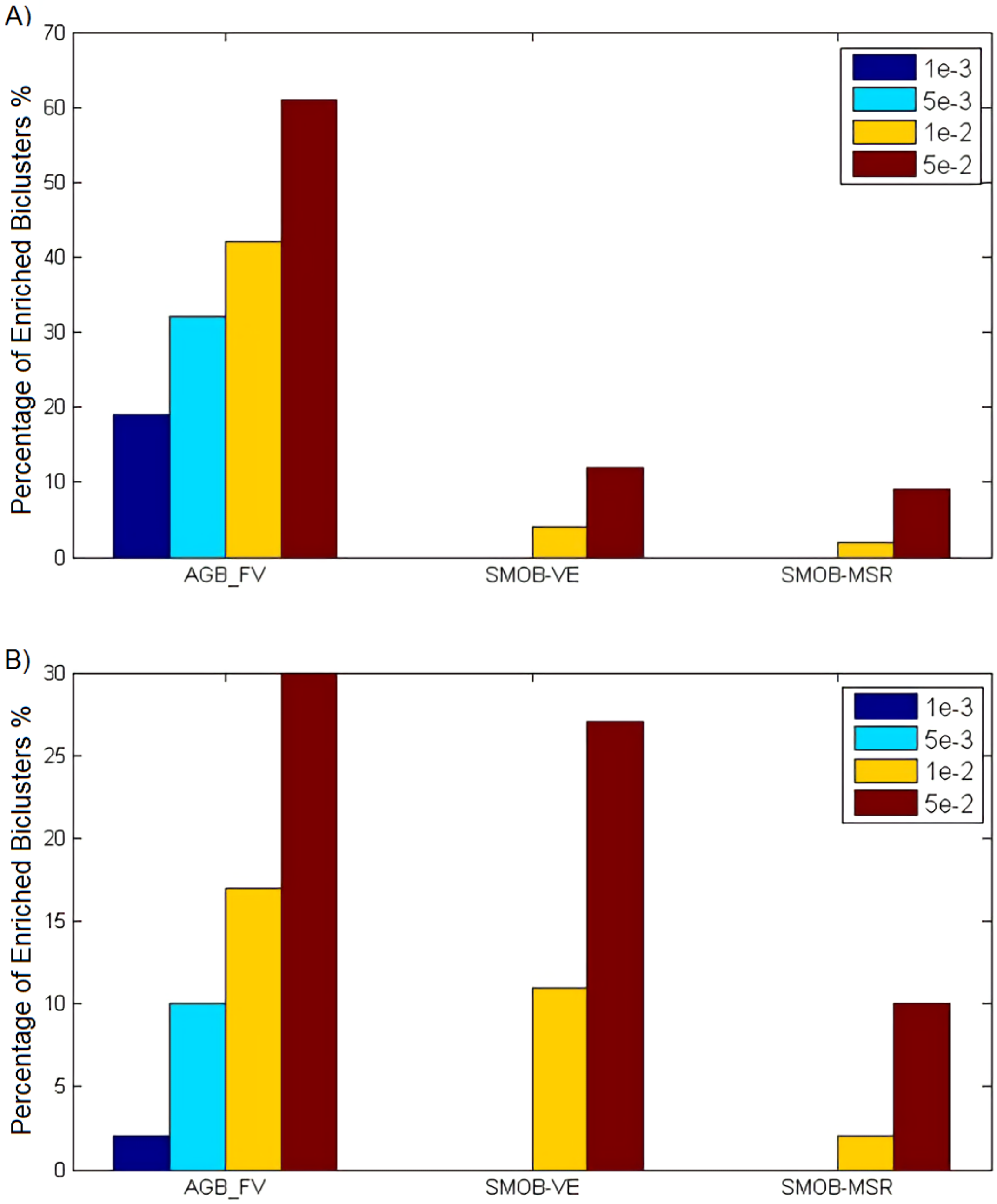
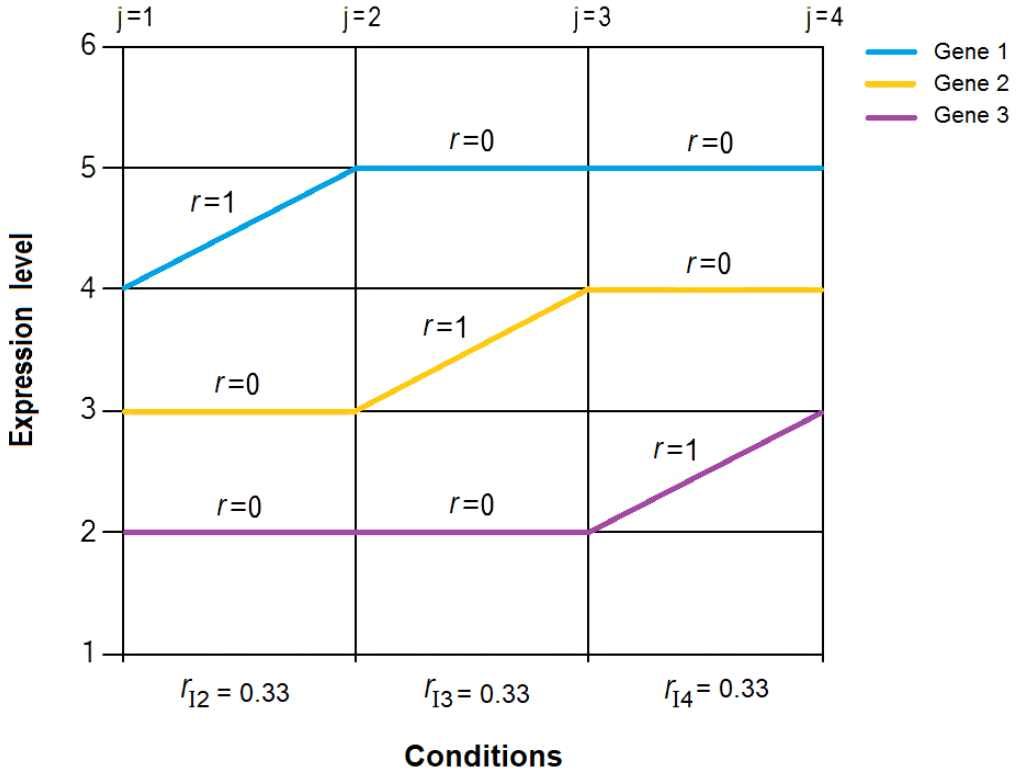


Figure 5. Comparison of acquired percentage of significant biclusters on Leukemia (A) and Steminal (B) datasets. On Leukemia and Steminal datasets one hundred biclusters were generated with the *BGA_VF* algorithm. For statistical analysis, 98 and 100 biclusters were considered from Leukemia and Steminal dataset, respectively. The analyzed biclusters did not show an overlap higher than 25%. The *SMOB-VE* and *SMOB-MSR* results for p-values 5e-3 and 1e-3 were not reported in the original work (Divina et al., 2012).



$$VF(I,J) = (|J|-1)\sum_{i\in I}\sum_{j\in J/\{1\}} |r_{ij} - r_{Ij}| = 12$$

$$(|J|-1)(2|I|-2) = (4\text{-}1)\,(2(3)\text{-}2) = 12$$

Figure 6. An example of a bicluster for which the value of *VF* function is equal to the upper bound. Each of the 3 genes presents a change in their level of expression for a different condition. This is an example of a bicluster for which the maximum value of the *VF* function is obtained, which is equal to the upper bound set for that number of genes and conditions.