Advances in Smart Systems Research ISSN 2050-8662
Vol. 6. No. 2 : pp.65-74 : k17is-218

Local search method based on biological knowledge for the biclustering of gene expression data

Ons Maâtouk[1,2], Wassim Ayadi[2,3], Hend Bouziri[1], and Beatrice Duval[2]

[1] LARODEC, Université de Tunis, 92 Boulevard 9 Avril, 1007 Tunis, Tunisie
[2] LERIA, Université d'Angers, Université Bretagne Loire, 2 Bd Lavoisier, 49045 Angers, France
[3] LaTICE, ENSIT, Université de Tunis, 92 Boulevard 9 Avril, 1007 Tunis, Tunisie

Abstract

Biclustering is a very interesting technique for unsupervised analysis of gene expression data. It aims to discover subsets of genes having similar behavior on a subset of conditions; such biclusters are related to close biological functions. The majority of existing biclustering algorithms are based on statistical criteria (e.g. size, coherence and structure...) to define the bicluster quality. However these measures are not directly related to biological knowledge and they may produce results that are difficult to interpret by a biologist. In fact, it is recognized that the integration of some biological information to guide the extraction of the biclusters ensures their relevance and their non-triviality. Therefore this work proposes a local search method that relies on ontological knowledge to cluster the genes while a correlation measure is used to cluster the conditions. An experimental study is performed using real microarray datasets. The results demonstrate the importance of the integration of the biological knowledge in the search process, to promote the discovery of non-trivial and biologically relevant biclusters.
keywords: *Biclustering; Local search method; Biological knowledge; Gene annotation; Gene expression; Computational Biology*

1 Introduction

Biological data are characterized by their wealth of information. Transcriptomic data provided by microarray or RNA-seq analysis are used to decipher the structural and functional relationships between the genes. Among the commonly used methods to understand and evaluate functional similarity between genes, there is the measure of their expression similarity [9]. It is assumed that a group of genes sharing a similar expression profile shares also the same regulatory regime and consequently the same functionalities. Transcriptomic data may be analyzed by a biclustering process [18, 15] that tries to cluster simultaneously the genes and the conditions; a bicluster is a subgroup of genes that exhibit a common behavior under a subgroup of conditions.

The biclustering aims to combine simultaneously the rows and the columns of a matrix to obtain consistent, homogeneous and stable biclusters. These biclusters represent a gene subset with same behavior under a condition subset. Each gene or each condition can participate in one or more biclusters.

Formally, the microarray data is represented as a data matrix $M(I, J)$, where I is a gene set and J a condition set, and the cell m_{ij} represents the expression level of the i^{th} gene relative to the j^{th} condition. A bicluster $B(G, C)$ associated with a data matrix $M(I, J)$ is a sub-matrix such that $G \subseteq I$ and $C \subseteq J$. The biclustering is an optimization problem aiming to extract and discover maximal biclusters with similar behavior genes and close biological functions.

The biclustering problem is a highly combinatorial problem with a search space size $O(2^{|I|+|J|})$ [31]. Moreover, in the general case, the biclustering problem is NP-hard [7], which explains why the majority of existing biclustering methods do not guarantee the optimality of their solutions.

Several algorithms [3, 4, 28, 17] and different measures have been proposed to extract a set of biclusters from a dataset. However, statistically significant solutions with good homogeneity are not necessarily biologically relevant. To obtain more relevant solutions, it would be interesting to incorporate some biological knowledge in the biclustering process. This knowledge must be used to guide the search of the biclusters and to ensure their biological relevance and their non-triviality. Despite that, only a few biclustering algorithms integrate biological knowledge into their research processes. AI-ISA [32], GenMiner [19] and scatter biclustering [26] algorithms annotate data with functional terms retrieved from ontologies repositories. They use these annotations to guide the search. Fang et al. algorithm [10], Data-Peeler [14] and bi-sets mining [5] are constraint-based algorithms enabling the discovery of dense biclusters and guaranteeing the satisfiability of constraints derived from background knowledge. The algorithms proposed by Menga et al. [20], Raya and Misra [29], Nepomuceno et al [24, 25] and Pio et al. [27] define a fitness function that aggregate statistical and biological similarity measures. These biclustering algorithms integrate the biological knowledge in a biological similarity measure and use a statistical measure to find nontrivial biclusters with relevant patterns.

In this context, our work proposes a local search method that integrates biological knowledge in the search process. By using a biological similarity measure, it extracts sets of genes with close biological functions. Then, by assigning conditions to the extracted genes, it captures relevant patterns based on the average correlation function. To evaluate the performance of the proposed method and to define the quality of the extracted biclusters, an experimental study is achieved on real microarray datasets. This paper is organized as follows. In section 2, the biclustering problem is defined and our local search algorithm is analyzed. Section 3 is dedicated to an experimental study of the proposed algorithm. Both statistical and biological evaluations are conducted. Conclusions are given in the last section.

2 Our LSM algorithm for the biclustering problem

Local search algorithms are widely used in difficult optimization problems, such as bioinformatics problems. They are characterized by their simplicity and their easy integration in other algorithms. The local search algorithm generally starts from initial reasonable solution and tries to get better solutions iteratively. At each iteration, series of local modifications are applied.

To ensure the quality of the biclusters, the initial solutions are constructed by the CC algorithm [7], not by a randomization process. Indeed, CC is recognized for its reasonable and time-efficient results and its almost total coverage of genes and conditions. For each bicluster of the initial solutions, the proposed Local Search Method (LSM) extracts genes with close biological functions based on the biological similarity measure. Then, it captures relevant patterns based on the average correlation function by assigning the right conditions to the extracted genes.

2.1 Data input

The input data of the algorithm are mainly the gene expression matrix M, the biological similarity matrix $M_{Bio.Sim}$ and the biclusters of the initial solutions. As mentioned previously, to ensure a reasonable quality and a high coverage of genes and conditions to the biclusters of the initial solution, these biclusters are extracted by the CC algorithm.

The rows and the columns of the gene expression matrix M correspond respectively to the genes and the conditions of the microarray data. A cell represents the expression level of a gene under a particular condition.

Regarding the biological similarity matrix $M_{Bio.Sim}$, it is calculated using the *GOSemSim* [33] which is an R package for computing semantic similarity among GO terms and gene clusters. It depends on the GO (Gene Ontology) annotations provided by Bioconductor [12] to obtain the GO terms and the relations between them. Several functions are provided by the GOSemSim package such as the *mgeneSim* function. It is designed for large-scale analysis to calculate the biological similarity between two gene sets. The *mgeneSim* function output is a biological similarity matrix $M_{Bio.Sim}$. Both rows and columns correspond to the genes of the microarray data and cells correspond to the pairwise GO semantic similarities of the two gene sets. Its values vary between 0 and 1. The higher the value obtained is, the higher is the similarity between the gene pair.

2.2 Objective function

The main biclustering goal is to extract biologically relevant biclusters of highly correlated genes. Thus, two criteria have to be optimized. A statistical criterion based on the coherence between bicluster genes and a biological criterion based on the biological relevance of these genes.

The majority of existing algorithms, such as those proposed by Menga et al. [20] and Nepomuceno et al. [24, 25], integrated the biological criterion in the fitness function by aggregating them with statistical criterion. However, even by giving the same importance to the two criteria, this aggregation does not reflect the real bicluster quality. For example, a bicluster with a medium quality does not necessarily have a medium value for the correlation measure and the biological similarity measure. It can be biologically relevant and does not have a good correlation or does not biologically relevant and have a good correlation. So, this aggregation may lead to not clearly distinguish the real bicluster quality.

Moreover, the gene annotation depends only on genes. So, the biological similarity measure evaluates the biological relevance of the bicluster genes but not the conditions. Contrary to the biological similarity measure, the correlation measure tries to find interesting patterns in the biclusters, considering both genes and conditions. Therefore two biclusters, with the same gene set and different condition sets, provide the same value for similarity biological measure and different value for the correlation measure.

For these reasons, we choose to use separate measures for each criterion (statistical and biological) to define the solution quality and guide the search to good solutions.

Correlation measure The average correlation function, proposed by Nepomuceno et al. [22], is used to evaluate the correlation between the bicluster genes. Its optimization allows the extraction of all biclusters' kind. The average correlation of the bicluster $B(G, C)$ is defined as follows:

$$\rho(B) = \frac{2}{|G|(|G|-1)} \sum_{i=1}^{|G|} \sum_{j=i+1}^{|G|} \left| \frac{cov(g_i, g_j)}{\sigma_{g_i} \sigma_{g_j}} \right| \tag{1}$$

where $cov(g_i, g_j)$ represents the covariance of the rows corresponding to the gene $g_i \in G$ and the gene $g_j \in G$.
σ_{g_i} (respectively σ_{g_j}) represents the standard deviations of the rows corresponding to the gene $g_i \in G$ (respectively the gene $g_j \in G$). This measure varies between 0 and 1. The higher $\rho(B)$ is, the higher is the correlation between the bicluster genes.

Biological similarity measure The matrix $M_{Bio.Sim}$, calculated by the *mgeneSim* function of the R package GOSemSim [33], is used as input data. It is used in the search process to measure biological similarity between the genes. The biological similarity of a bicluster $B(G, C)$ is calculated as follows:

$$BioSim(B) = \frac{2}{|G|(|G|-1)} \sum_{i=1}^{|G|} \sum_{j=i+1}^{|G|} M_{Bio.Sim}(g_i, g_j) \tag{2}$$

where $g_i \in G$ and $g_j \in G$. This measure varies between 0 and 1. The higher the $BioSim(B)$ value is, the higher is the similarity between the bicluster genes.

2.3 Description of the Local Search Method (LSM)

The proposed Local Search Method (LSM) tries to generate new biclusters by improving the biological and the statistical quality of the biclusters of the initial solutions. It extracts genes with close biological functions based on the values of biological similarity matrix $M_{Bio.Sim}$ and the biological similarity measure $BioSim$ (Equation 2). Then, it captures relevant patterns based on the average correlation function ρ (Equation 1) by assigning the right conditions to the determined genes. The search process treats the gene part and the condition part separately. As previously explained, contrary to the correlation average function, the biological similarity measure depends only on the gene set of the bicluster. It does not take into consideration its condition set.

For $B(G, C)$, a bicluster of the initial solution, the local search step constructs a bicluster $B'(G', C')$. In G', we keep only the genes of G that have a great number of genes very similar to them according to the biological similarity matrix $M_{Bio.Sim}$. This choice is controlled by a parameter $\%NbGene$ and depends on a threshold $Th_{Bio.Sim}$ defined as $BioSim(B)$ the average biological similarity between all the pairs of genes of G.

On the same way, C' keeps the conditions of C that are highly correlated according to the correlation matrix M_{Corr}. This matrix is calculated based on the average correlation function ρ. It presents the correlation between the condition pairs for the determined gene set G'. Both rows and columns correspond to the condition set C. Analogically, the choice is controlled by a parameter $\%NBCondition$ and depends on a threshold $Th_{Corr.Sim}$ defined as $\rho(B)$ the average correlation of the bicluster B.

For a bicluster $B(G, C)$, let us consider the biological similarity matrix $M_{Bio.Sim}$ and the gene expression matrix M as data input and let $\%NbGene$ and $\%NbCondition$ be chosen parameters:

- Compute $Th_{Bio.Sim}$ = $BioSim(B)$ (Eq. 2) and $Th_{Corr.Sim}$ = $\rho(B)$ (Eq. 1)
- For each gene $g \in G$
  - Define $Sim_1(g)$ = { $g' \in G$ / $M_{Bio.Sim}(g, g') \geq Th_{Bio.Sim}$} // $Sim_1(g)$ is therefore the set of the genes biologically similar to the gene g
- Sort the list of sets $Sim_1(g)$ according to their cardinality $|Sim_1(g)|$ // So that the next step checks firstly the similarity between the maximum number of genes
- For each gene $g' \in Sim_1(g)$ following the sorting order
  - Compute $Th_{Nb.Gene}$ = $\%NbGene \times |Sim_1(g)|$
  - Define $Sim_2(g')$ = { $g'' \in Sim_1(g)$ / $M_{Bio.Sim}(g', g'') \geq Th_{Bio.Sim}$}
  - If $|Sim_2(g')| > Th_{Nb.Gene}$: Add gene g' to the gene part G' of the bicluster B' // To ensure that the similarity of the bicluster B' is better than that of the bicluster B
- Compute the correlation matrix M_{Corr} where the cell $M_{Corr}(c, c')$ represents the average correlation of the genes of the gene part G' under the conditions $\{c, c'\}$ such as $c \in C$ and c' $\in C$.
- For each condition $c \in C$
  - Define $Sim_1(c)$ = { $c' \in C$ / $M_{Corr}(c, c') \geq Th_{Corr.Sim}$} // The set of the conditions correlated with the condition c
- Sort the list of sets $Sim_1(c)$ according to their cardinality $|Sim_1(c)|$
- For each condition $c' \in Sim_1(c)$ following the sorting order
  - Compute $Th_{Nb.Condition}$ = $\%NbCondition \times |Sim_1(c)|$
  - Define $Sim_2(c')$ = { $c'' \in Sim_1(c)$ / $M_{Corr}(c', c'') \geq Th_{Corr.Sim}$}
  - If $|Sim_2(c')| > Th_{Nb.Condition}$: Add condition c' to the condition part C' of the bicluster B' // To ensure the correlation between all the conditions of the output bicluster B'.

3 Experimental study

In order to evaluate the capacity of the proposed method to extract relevant biclusters and to analyze the influence of the integration of biological knowledge in the search process, an experimental study is performed using real data. Two gene expression datasets are considered. The first dataset is the *Yeast cell cycle* (2884 genes, 17 conditions) described by Tavazoie et al. [30] and then pretreated by Cheng and Church [7]. The second one is the *Saccharomyces cerevisiae* dataset (2993 genes, 173 conditions) described by Gasch et al. [11]. For both datasets, the assessment is based on common practice inferred from biclustering literature. The statistical assessment of the quality of a bicluster is defined by its average correlation value and its size (number of genes and number of conditions). Regarding the biological assessment, the percentage of enriched biclusters, one of the classic biological evaluation criteria commonly used in biclustering [24], and the number of enriched GO terms per bicluster are computed. In addition, the significant GO terms of a selected bicluster are presented.

The parameter settings have been fixed after several tests. Everytime, we use different parameters to keep those that provide the best results. The parameters $\%NbGene$ and $\%NbCondition$ are fixed to 0.25. In order to ensure the stability of the results, the process is repeated 30 times. The results presented in the following subsections are the average of the results obtained for the 30 runs. They are compared to the CC algorithm results (used as initial solutions). This allows to demonstrate the significant influence of the proposed method on the relevance of biclusters.

The LSM results are also compared to those of different state-of-the-art biclustering algorithms: BiMax [28], ISA [4], OPSM [3] and X-Motif [21]. The results of these algorithms are generated using the *Biclustering Analysis Toolbox-plus* (BicAT-plus) [1], a common biclustering analysis toolbox in which most important biclustering algorithms were implemented. The results were not compared to other algorithms that integrate biological knowledge (AI-ISA [32], Data-Peeler [14], bi-sets mining [5]), since these algorithms were not available and the results provided in the papers came from different datasets.

3.1 Statistical results

	Number of genes			Number of conditions			Average correlation		
	Avg.	Min.	Max.	Avg.	Min.	Max.	Avg.	Min.	Max.
CC	39.62	14	47	3.16	2	17	0.84	0.72	0.89
LSM	37.18	9	43	12.49	4	15	0.92	0.89	0.96

Table 1. Average, minimum and maximum of the number of genes, the number of conditions and the average correlation value of the extracted biclusters for the *Yeast Cell Cycle* dataset

	Number of genes			Number of conditions			Average correlation		
	Avg.	Min.	Max.	Avg.	Min.	Max.	Avg.	Min.	Max.
CC	81.11	39	258	19.64	9	23	0.33	0.14	0.43
LSM	79.52	33	244	18.98	11	21	0.81	0.58	0.96

Table 2. Average, minimum and maximum of the number of genes, the number of conditions and the average correlation value of the extracted biclusters for the *Saccharomyces cerevisiae* dataset

Table 1 and 2 summarize the statistical quality of the biclusters extracted by the proposed LSM algorithm and the CC algorithm, respectively, for the *Yeast Cell Cycle* and the *Saccharomyces cerevisiae* datasets. They present the average, minimum and maximum number of genes, number of conditions and average correlation value. It can be observed that the number of genes and the number of conditions decreased slightly by applying the proposed method LSM. The reason can be that the bicluster size depends on the size of the biclusters of the initial solutions (generated by the CC algorithm). The proposed method LSM retains only the biologically similar genes.

On the other hand, for the two datasets, the average correlation of the obtained biclusters is significantly higher than CC. It can be seen that for both datasets, the biclusters extracted by the two algorithms are highly correlated. Indeed, for the *Yeast Cell Cycle* dataset, the proposed LSM algorithm (respectively the CC algorithm) records an average correlation value equal to 0.92 (respectively 0.84). Contrariwise, for the *Saccharomyces cerevisiae* dataset, the average correlation value of the biclusters extracted by the CC algorithm does not exceed 0.43. In regards to the biclusters extracted by the proposed LSM algorithm, the average correlation values vary between 0.58 and 0.96. It can be noted that the correlation of the biclusters extracted by the LSM algorithm does not depend on the correlation of the initial solutions.

	BiMax	ISA	OPSM	X-Motif	CC	LSM
Number of genes	24.0	76.3	437.94	1.2	39.62	37.18
Number of conditions	3	8.7	9.5	11.4	3.16	12.49
Average correlation	0.66	0.50	0.91	0.71	0.84	0.92

Table 3. Average of the number of genes, the number of conditions and the average correlation value of the biclusters extracted by different biclustering algorithms for the *Yeast Cell Cycle* dataset

	BiMax	ISA	OPSM	X-Motif	CC	LSM
Number of genes	32.8	76.27	95.58	1.12	81.11	79.52
Number of conditions	3	8.71	12.5	34.52	19.64	18.98
Average correlation	0.68	0.59	0.87	0.97	0.33	0.81

Table 4. Average of the number of genes, the number of conditions and the average correlation value of the biclusters extracted by different biclustering algorithms for the *Saccharomyces cerevisiae* dataset

Table 3 and 4 compare the statistical results of several biclustering algorithms, respectively, for the *Yeast Cell Cycle* and the *Saccharomyces cerevisiae* datasets. It can be seen that the LSM algorithm competes favorably with several biclustering algorithms. For the *Yeast Cell Cycle* dataset (Table 3), the LSM algorithm outperforms other algorithms in terms of gene correlation with a reasonable bicluster size. As for the *Saccharomyces cerevisiae* dataset (Table 4), the best gene correlation is recorded for the biclusters extracted by the X-Motif algorithm with an average value equal to 0.97. It can be explained by the fact that its number of genes is very small (average number of genes equal to 1.12). It can be noted also that the LSM and the OPSM algorithms have relatively close results with high gene correlation and reasonable bicluster size. So, it can be inferred that the proposed LSM algorithm allows to capture relevant patterns and especially to extract biclusters of highly correlated genes, compared to the other biclustering algorithms.

3.2 Biological results

The biological significance of the obtained biclusters can be reflected by their functional enrichment based on the *Gene Ontology* [8] annotations. The gene products are described by three ontology structures provided by the GO in term of biological process, molecular function and cellular component.

In this context, the web tool *GOTermFinder* [6, 13] is used to seek significant GO and to annotate gene products in a given list. The enrichment degree of the biclusters is measured by their adjusted p-value [2, 16, 23]. The biclusters with a low p-value (lower than 5%) are considered as enriched. This means that the majority of the genes in this bicluster have common biological functions. The best biclusters have a p-value close to 0%.

	Perc. of enriched biclusters			Avg. nb. of GO terms per bicluster		
	B. process	M. function	C. component	B. process	M. function	C. component
CC	16 %	18 %	16 %	1.67	1.25	1.33
LSM	98 %	98 %	96 %	18.62	11.98	9.28

Table 5. Percentage of enriched biclusters and average number of enriched Go terms per bicluster for the *Yeast Cell Cycle* dataset

	Perc. of enriched biclusters			Avg. nb. of GO terms per bicluster		
	B. process	M. function	C. component	B. process	M. function	C. component
CC	20 %	22 %	18 %	1.8	1.67	1.25
LSM	100 %	97 %	95 %	39.55	8.23	17.46

Table 6. Percentage of enriched biclusters and average number of enriched Go terms per bicluster for the *Saccharomyces cerevisiae* dataset

Table 5 and 6 present the percentage of enriched biclusters with p-value lower than 0.1% and the average number of enriched Go terms per bicluster in the three ontology structures: B. process (Biological process), M. function (Molecular function) and C. component (Cellular component) of the proposed LSM algorithm and the CC algorithm, respectively, for the *Yeast Cell Cycle* and the *Saccharomyces cerevisiae* datasets.

It can be observed that the most enriched biclusters are those extracted by the proposed LSM algorithm. It can be seen in Table 6 that only 20%, 22% and 18% of the biclusters extracted by the CC algorithm are enriched for the *Saccharomyces cerevisiae* dataset in the ontology structures: biological process, molecular function and cellular component, respectively. Even for the *Yeast Cell Cycle* dataset (Table 5), only 16%, 18% and 16% of the biclusters are enriched respectively for the biological process, molecular function and cellular component. However, the LSM algorithm is able to extract many enriched biclusters. Indeed, 98%, 98% and 96% of the biclusters extracted by the LSM algorithm for the *Yeast Cell Cycle* dataset (Table 5) and 100%, 97% and 95% of them for the *Saccharomyces cerevisiae* dataset (Table 6), are enriched respectively for the biological process, molecular function and cellular component.

Moreover, it can be noted that the biclusters obtained by the proposed LSM algorithm have an average number of GO term higher than that obtained by the CC algorithm. The average GO term number per bicluster of the CC algorithm does not exceed 1.67 for the *Yeast Cell Cycle* dataset and 1.8 for the *Saccharomyces cerevisiae* dataset. It means that only some GO terms are shared by the genes of the obtained biclusters. That can be explained by the fact that the annotated genes are in the upper levels of the GO hierarchy.

Contrariwise, the LSM algorithm records respectively for the biological process, the molecular function and the cellular component, an average GO term number of 18.62, 11.98 and 9.28 for the *Yeast Cell Cycle* dataset and 29.55, 8.23 and 17.46 for the *Saccharomyces cerevisiae* dataset. It means that the obtained biclusters are composed of genes sharing a high number of GO annotations. So, more biological information related to the bicluster genes is provided.

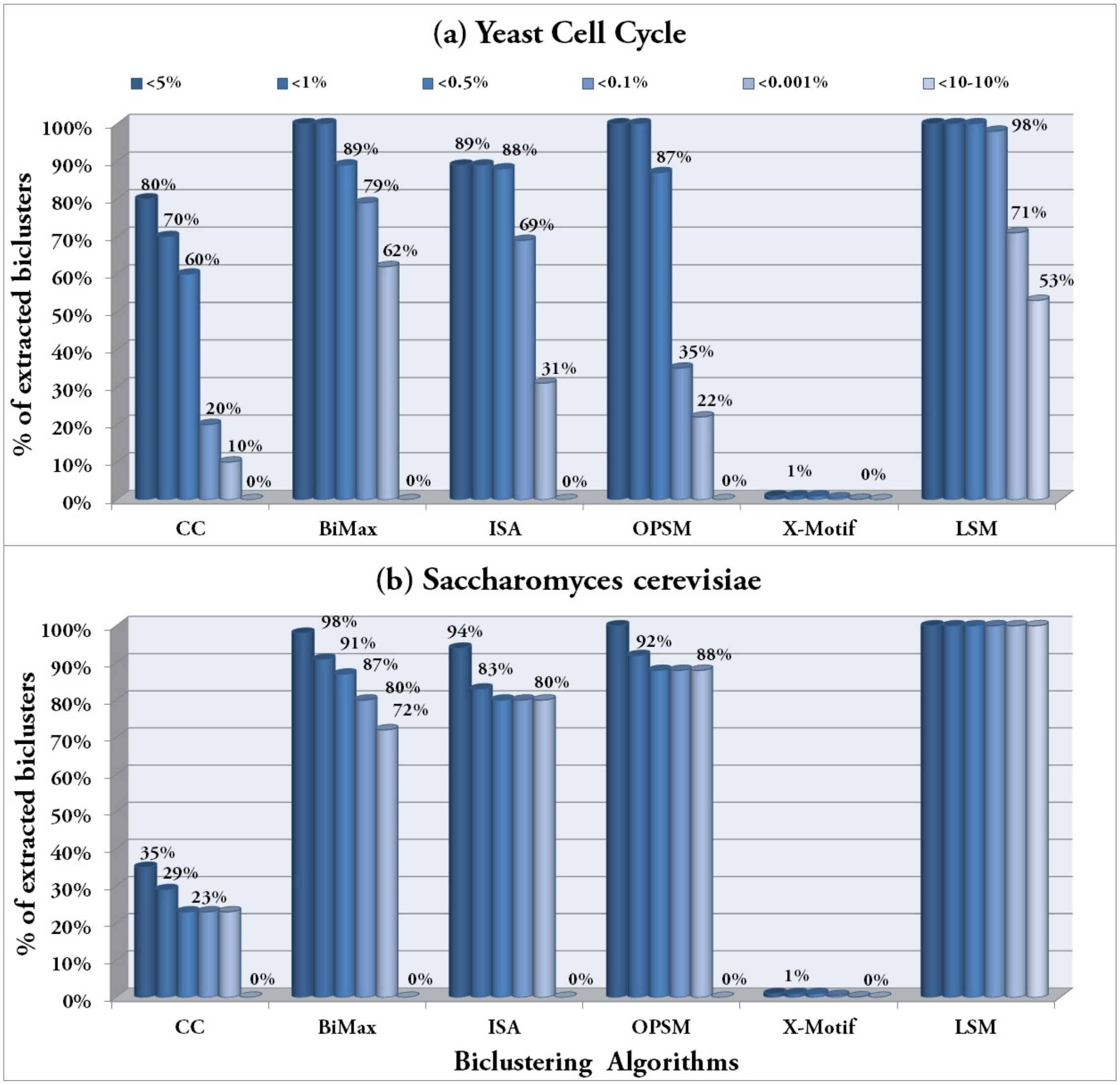

Figure. 1. Percentage of the biclusters extracted by different biclustering algorithms for different p-values for the two datasets : (a) *Yeast Cell Cycle* and (b) *Saccharomyces cerevisiae*

Figure 1 shows the percentage of extracted biclusters for different adjusted p-value (p=5%; 1%; 0.5%; 0.1%; 0.001% and 10^{-10}) for the *Yeast Cell Cycle* and the *Saccharomyces cerevisiae* datasets. The p-values (the enrichment degree) of the biclusters extracted by the LSM algorithm are measured using the web tool *GOTermFinder* [6, 13]. Regarding the other biclustering algorithms, the percentage of extracted biclusters were taken from [16, 23].

It can be noted that the majority of the algorithms have rather low percentage. For the *Yeast Cell Cycle* dataset (Figure 1 (a)), 53% of the biclusters extracted by LSM are statistically significant with a p-value lower than 10^{-10}%, while only 62%, 31% and 22% of the biclusters respectively extracted by Bimax, ISA and OPSM are statistically significant with a p-value lower than 0.001%. For the *Saccharomyces cerevisiae* dataset (Figure 1 (b)), 72%, 80% and 88% of the biclusters respectively extracted by Bimax, ISA and OPSM are statistically significant with a p-value lower than 0.001%. Only LSM reaches very low values. Indeed, 100% of the biclusters extracted by our algorithm are statistically significant with a p-value lower than 10^{-10}%.

Table 7 and 8 extend the previous analysis and present the most significant GO terms of a random bicluster extracted by the LSM algorithm, respectively for the *Yeast Cell Cycle* and the *Saccharomyces cerevisiae* datasets. The columns represent the ontology structures (Biological process, Molecular function and Cellular component). The rows show the significant GO terms, the involved number of genes and the biological enrichment degree provided by the p-value.

Table 7 presents same shared GO terms of a bicluster of 9 genes. Same for Table 8, it presents same shared GO terms of a bicluster of 155 genes. The first row of the first column in Table 7 indicate that all the bicluster genes (9 genes) belong to the *transport* biological process, and the statistical significance provided by the p-value is equal to 3.05e^{-06}.

The results show that there are, for the three GO structures, several processes identified by large groups of genes with high enrichment degree (low p-value). It can be deduced that the integration of the biological knowledge in the search process improves the performance of the proposed LSM algorithm to discover biologically relevant biclusters.

Biological process	Molecular function	Cellular component
transport 9 genes 3.05e^{-06}	**transmembrane transporter activity** 9 genes 7.30e^{-12}	**integral component of membrane** 9 genes 4.11e^{-07}
transmembrane transport 6 genes 3.99e^{-06}	**transporter activity** 9 genes 6.18e^{-11}	**intrinsic component of membrane** 9 genes 4.41e^{-07}
establishment of localization 9 genes 4.17e^{-06}	**substrate-specific transmembrane transporter activity** 7 genes 6.57e^{-08}	**membrane part** 9 genes 4.52e^{-06}
localization 9 genes 1.26e^{-05}	**substrate-specific transporter activity** 7 genes 2.24e^{-07}	**membrane** 9 genes 4.73e^{-05}
metal ion transport 4 genes 3.92e^{-05}		

Table 7. Significant GO terms in the three ontology structures of a bicluster extracted by the proposed LSM algorithm for the *Yeast Cell Cycle* dataset

Biological process	Molecular function	Cellular component
cytoplasmic translation 92 genes 1.51e^{-113}	**structural constituent of ribosome** 95 genes 3.46e^{-107}	**cytosolic ribosome** 95 genes 2.27e^{-120}
translation 99 genes 1.30e^{-58}	**structural molecule activity** 99 genes 2.37e^{-93}	**cytosolic part** 95 genes 4.19e^{-104}
peptide biosynthetic process 99 genes 2.25e^{-58}		**ribosomal subunit** 95 genes 1.18e^{-102}
peptide metabolic process 99 genes 7.35e^{-57}		**ribosome** 95 genes 1.36e^{-86}
organonitrogen compound biosynthetic process 101 genes 1.29e^{-44}		**intracellular ribonucleoprotein complex** 111 genes 8.41e^{-70}
ribosome biogenesis 70 genes 7.74e^{-44}		**cytosolic small ribosomal subunit** 49 genes 3.47e^{-69}

Table 8. Significant GO terms in the three ontology structures of a bicluster extracted by the proposed LSM algorithm for the *Saccharomyces cerevisiae* dataset

4 Conclusion

In the present work, a biclustering algorithm based on a local search method characterized by its simplicity is proposed. The algorithm input data are mainly the gene expression matrix and the biological similarity matrix. The gene expression matrix presents the expression level of the gene set under the condition set of a dataset. The biological similarity matrix is calculated by the mgeneSim function of the R package GOSemSim and represents the biological similarity measure values between the pairs of the gene set.

The main goal of the biclustering algorithm is to extract biologically relevant biclusters of highly correlated genes. Thus, two criteria have to be optimized. A statistical criterion based on the coherence between bicluster genes and a biological criterion based on the biological relevance of these genes. So, two distinct measures are considered. The biological similarity measure evaluates the biological relevance of the bicluster genes but not the conditions. Contrariwise, the correlation measure tries to find interesting patterns in the biclusters, considering both genes and conditions. For that, the search process treats separately the gene part and the condition part.

In order to evaluate the capacity of the proposed method to extract relevant biclusters and to analyze the influence of the integration of biological knowledge in the search process, an experimental study is performed using real data. These experimentations show that the LSM method extracts biclusters with a slightly smaller size than that of the initial solution biclusters. In contrast, the LSM method allows to improve the correlation between their genes. This method allows also to increase the percentage of enriched biclusters extracted as well as the number of enriched Go terms per bicluster.

As a consequence of the results, it can be deduced that the proposed LSM algorithm allows to extract biologically relevant biclusters with highly correlated genes. Hence, the importance of the integration of the biological knowledge in the search process to improve the performance of the algorithm is demonstrated.

Given to the simplicity of the proposed local search method, its ability to extract relevant biclusters and its easy integration in other algorithms, the future work will be focused on the study of a hybrid algorithm combining an evolutionary algorithm with this method.

References

[1] F. M. Al-Akwaa. Analysis and visualization of gene expression data using biclustering: A comparative study. *African Journal of Biotechnology*, 11(7):1744–1753, 2012.

[2] W. Ayadi and J. K. Hao. A memetic algorithm for discovering negative correlation biclusters of DNA microarray data. *Neurocomputing*, 145:14–22, 2014.

[3] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini. Discovering local structure in gene expression data : the order-preserving submatrix problem. *In RECOMB '02 : Proceedings of the Sixth Annual International Conference on Computational Biology*, pages 49–57, 2002.

[4] S. Bergmann, J. Ihmels, and N. Barka. Defining transcription modules using large-scale gene expression data. *Bioinformatics*, 20(13):1993–2003, 2004.

[5] J. Besson, C. Robardet, L. De Raedt, and J. F. Boulicaut. Mining bi-sets in numerical data. *Knowledge Discovery in Inductive Databases*, pages 11–23, 2007.

[6] E.I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J.M. Cherry, and G. Sherlock. GOTermFinder-open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715, 2004.

[7] Y. Cheng and G. M. Church. Biclustering of expression data. *In Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 93–103, 2000.

[8] Gene Ontology Consortium. Gene ontology : tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[9] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *National Academy of Sciences*, 95(25):14863–14868, 1998.

[10] G. Fang, M. Haznadar, W. Wang, H. Yu, M. Steinbach, T. R. Church, W. S. Oetting, B. V. Ness, and V. Kumar. High-order SNP combinations associated with complex diseases: Efficient discovery, statistical power and functional interactions. *Plos One*, 7(4), 2012.

[11] A.P. Gasch, P.T. Spellman, C.M. Kao, O. Carmel-Harel, M.B. Eisen, G. Storz, D. Botstein, and P.O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *In Molecular Biology of the Cell*, 11(12):4241–4257, 2000.

[12] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. C. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irrizary, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. H. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004.

[13] GOTermFinder. http ://db.yeastgenome.org/cgi-bin/GO/goTermFinde. 2004.

[14] I. Guerra., L. Cerf, J. Foscarini, M. Boaventura, and W. Meira. Constraint-based search of straddling biclusters and discriminative patterns. *Information and Data Management*, 4(2):114–123, 2013.

[15] R. Henriques, C. Antunesa, and S. C. Madeira. A structured view on pattern mining-based biclustering. *Pattern Recogn*, 48(12):3941–3958, 2015.

[16] X. Liu and L. Wang. Computing the maximum similarity bi-clusters of gene expression data. *Bioinformatics*, 23(1):50–56, 2007.

[17] Ons Maâtouk, Wassim Ayadi, Hend Bouziri, and Beatrice Duval. *Evolutionary Algorithm Based on New Crossover for the Biclustering of Gene Expression Data*, pages 48–59. Springer International Publishing, 2014.

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

[19] R. Martinez, N. Pasquier, and C. Pasquier. Genminer: mining informative association rules from genomic data. *IEEE international conference on Bioinformatics and Biomedecine*, pages 15–22, 2007.

[20] J. Menga, R. Lia, and Y. Luan. Classification by integrating plant stress response gene expression data with biological knowledge. *Mathematical Biosciences*, 266:65–72, 2015.

[21] T. M. Murali and S. Kasif. Extracting conserved gene expression motifs from gene expression data. *Pacific Symposium on Biocomputing*, 8:77–88, 2003.

[22] J. A. Nepomuceno, A. Troncoso, and J. S. Aguilar-Ruiz. Evolutionary metaheuristic for biclustering based on linear correlations among genes. *SAC'10, Sierre, Switzerland*, (5):1143–1147, 2010.

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

[24] J. A. Nepomuceno, A. Troncoso, I. A. Nepomuceno-Chamorro, and J. S. Aguilar-Ruiz. Integrating biological knowledge based on functional annotations for biclustering of gene expression data. *Computer Methods and Programs in Biomedicine*, 119(3):163–180, 2015.

[25] J. A. Nepomuceno, A. Troncoso, I. A. Nepomuceno-Chamorro, and J. S. Aguilar-Ruiz. Biclustering of gene expression data based on simUI semantic similarity measure. *Hybrid Artificial Intelligent Systems*, 9648:685–693, 2016.

[26] R. Peeters. The maximum edge biclique problem is NP-complete. *Discrete Applied Mathematics*, 131(3):651–651, 2003.

[27] G. Pio, M. Ceci, D. D'Elia, C. Loglisci, and D. Malerba. A novel biclustering algorithm for the discovery of meaningful biological correlations between microRNAs and their target genes. *BMC Bioinformatics*, 14(7):S8, 2013.

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

[29] S. S. Raya and S. Misra. A supervised weighted similarity measure for gene expressions using biological knowledge. *Gene*, 595(2):150–160, 2016.

[30] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22:281–285, 1999.

[31] A. Valente-Freitas, W. Ayadi, M. Elloumi, J. L. Oliveira, and J. K. Hao. A survey on biclustering of gene expression data. *In Biological Knowledge Discovery Handbook: Preprocessing, Mining and Postprocessing of Biological Data, Wiley Book Series on Bioinformatics: Computational Techniques and Engineering, Wiley-Blackwell, John Wiley and Sons Ltd., New Jersey, USA*, pages 591–608, 2013.

[32] A. Visconti, F. Cordero, and R. G. Pensa. Leveraging additional knowledge to support coherent bicluster discovery in gene expression data. *Intell Data Anal*, 18(5):837–855, 2014.

[33] G. Yu, F. Li, Y. Qin, X.Bo, Y. Wu, and S. Wang. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26(7):976–978, 2010.