Biclustering of Gene Expression Data Based on *SimUI* Semantic Similarity Measure

Juan A. Nepomuceno[1](✉), Alicia Troncoso[2],
Isabel A. Nepomuceno-Chamorro[1], and Jesús S. Aguilar–Ruiz[2]

[1] Departamento de Lenguajes y Sistemas Informáticos,
Universidad de Sevilla, Seville, Spain
janepo@us.es
[2] Área de Informática, Universidad Pablo de Olavide, Seville, Spain

Abstract. Biclustering is an unsupervised machine learning technique that simultaneously clusters genes and conditions in gene expression data. Gene Ontology (GO) is usually used in this context to validate the biological relevance of the results. However, although the integration of biological information from different sources is one of the research directions in Bioinformatics, GO is not used in biclustering as an input data. A scatter search-based algorithm that integrates GO information during the biclustering search process is presented in this paper. SimUI is a GO semantic similarity measure that defines a distance between two genes. The algorithm optimizes a fitness function that uses SimUI to integrate the biological information stored in GO. Experimental results analyze the effect of integration of the biological information through this measure. A SimUI fitness function configuration is experimentally studied in a scatter search-based biclustering algorithm.

Keywords: Biclustering of gene expression data · Gene pairwise GO measures · Scatter search metaheuristic

1 Introduction

Biclustering is an unsupervised machine learning technique that searches local patterns. Although it was previously known with names as subspace clustering or coclustering, most of the biclustering algorithms have been presented in the context of gene expression data due to the nature of these data. Gene expression data report the information of how thousand of genes are expressed along dozens of samples or experimental conditions. These kinds of datasets motivate the interest of grouping genes only in a subset of samples and not in all of them. Therefore, biclustering simultaneously clusters genes (features) and conditions (instances). Due to the NP-hard nature of the algorithmic problem, most of the algorithms use different heuristic strategies [1]. It must be highlighted a family of algorithms based on evolutionary computation and metaheuristics [2,3].

Gene Ontology (GO) is a dictionary where each gene is annotated with a specific set of biological terms. These terms are classified in three domains: Biological Process (BP), Molecular Functions (MF) and Cellular Components (CC).

The ontology has a tree structure. High levels in the tree structure have more general information and low levels have more specific information. Functional annotation files relate genes with their set of biological terms. Therefore, each gene is related with a set of GO terms. These kinds of files are usually used in biclustering in validation tasks to study the biological relevance of the results. The standard benchmark of comparison among biclustering algorithms is based on the use of functional annotation files [4]. However, although the integration of biological information is one of the research directions in Bioinformatics, this information is only used in validation but not as part of the search mechanism. Some new clustering [5] and biclustering [6] algorithms recently use GO annotation files to improve their searches processes in the context of gene expression data. These new search criteria use the information in gene expression matrices and in gene annotation files extracted from GO.

A concept of a distance among GO terms can be defined using a GO based-similarity measure. These measures evaluate the specificity of a GO term according the structure of the GO graph [7]. GO based-similarity measures can basically be classified in two groups: graph-based measures if they use the frequency of a term in the GO graph or (IC)-based measures if they use the depth in the GO graph. Otherwise, each gene in the gene annotation file is related with a set of terms. Therefore, a distance between two genes can be defined through a GO based-similarity measure.

SimUI measure is a graph-based GO similarity measure that evaluates the distance between two genes [8]. The idea of SimUI is counting the terms in the GO graph for each gene. This measure can be used to infer the biological relevance of a set of genes in a gene annotation file. The idea of this paper is to use the SimUI measure to introduce a biological bias during the search process in the biclustering algorithm. Therefore, the GO information is used as part of the search process to find the best results from a biological point of view. The proposed algorithm is a scatter search based-algorithm that sequentially finds each bicluster. The scatter search is a population based-metaheuristic where the evolutionary process is addressed by the evolution of a small set of solutions. This set of solutions is built according an equilibrium between an intensification and a diversity criterion. The proposed algorithm is based on the algorithm presented in [6]. Basically, a fitness function that evaluates the quality of each bicluster is optimized through a scatter search procedure. This merit function firstly evaluates the patterns in the bicluster using the gene expression matrix. Secondly, the biological relevance of the genes in the bicluster is also measured with SimUI that works with a gene annotation file.

The rest of the paper is organized as follows. The proposed algorithm is presented in Sect. 2. Firstly, the scatter search procedure is explained in Subsect. 2.1. Secondly, the SimUI based-fitness function is presented in Subsect. 2.2. Section 3 presents the experimental results to show that the performance of the algorithm is improved using SimUI. Finally, conclusions and future works are presented in Sect. 4.

2 Algorithm

The input data of the algorithm are basically a gene expression data matrix and a gene functional annotation file. Each row in the gene expression data matrix is the expression profile of a gene. Each column is a sample generated according a specific experimental condition. Gene functional annotation files provide biological information that is used to provide a biological bias during the bicluster search process. These files relate each gene with a set of biological terms as for example GO terms.

The proposed algorithm is a scatter search-based algorithm that optimizes a fitness function. This fitness function firstly uses the gene expression matrix information to define biclusters and secondly the functional annotation files to integrate the biological information. The search process is based on a scatter search metaheuristic procedure that is repeated to obtain each bicluster. It is important to note that the search procedure and the fitness function definition are independent.

2.1 Scatter Search Based-Procedure

Scatter search is a population based-metaheuristic where the evolution of a small and representative set of solutions is used to optimize a merit function. Each solution codifies a bicluster and at the end of the evolutionary process the best one represents the result. Each resultant bicluster is found using a scatter search procedure. Hence, the proposed algorithm repeats this procedure the number of times equal to the number of bicluster to discover. This procedure uses the same ideas that the algorithm presented in [9].

Solutions/biclusters are codified as two binary strings where the bits indicate their corresponding condition or gene in the gene expression matrix. The key idea in the scatter search process is the combination of intensification and diversification using a small set of solutions called the *reference set*. The intensification of solutions accelerates the search and the diversification avoids local minimum.

Figure 1 shows the scatter search scheme. The input basically is the gene expression matrix and the gene annotation file. Firstly, an *initial population* is built. The *diversification generation method* generates solutions as scatter as possible in order to contemplate different places in the search space. This method is based on the diversification method presented in [9], where a randomly seed solution is used to generate a set of diverse solutions. If x is a seed binary string, a new string x' is generated for each value of an integer $h = 1, 2, 3, \ldots, h_{max}$ as follows:

$$x'_{1+kh} = 1 - x_{1+kh} \quad \textbf{for} \quad k = 0, 1, 2, 3, \ldots, \lfloor n/h \rfloor \tag{1}$$

where $x = (x_1, \ldots, x_n)$, n is the number of bits, k takes values from 0 to the largest integer satisfying $k \leq n/h$, with the maximum value for h is $h_{max} = n/5$. All the remaining bits of x' are equal to those of x. Secondly, the *improvement method* changes each solution by other better solution from fitness function point

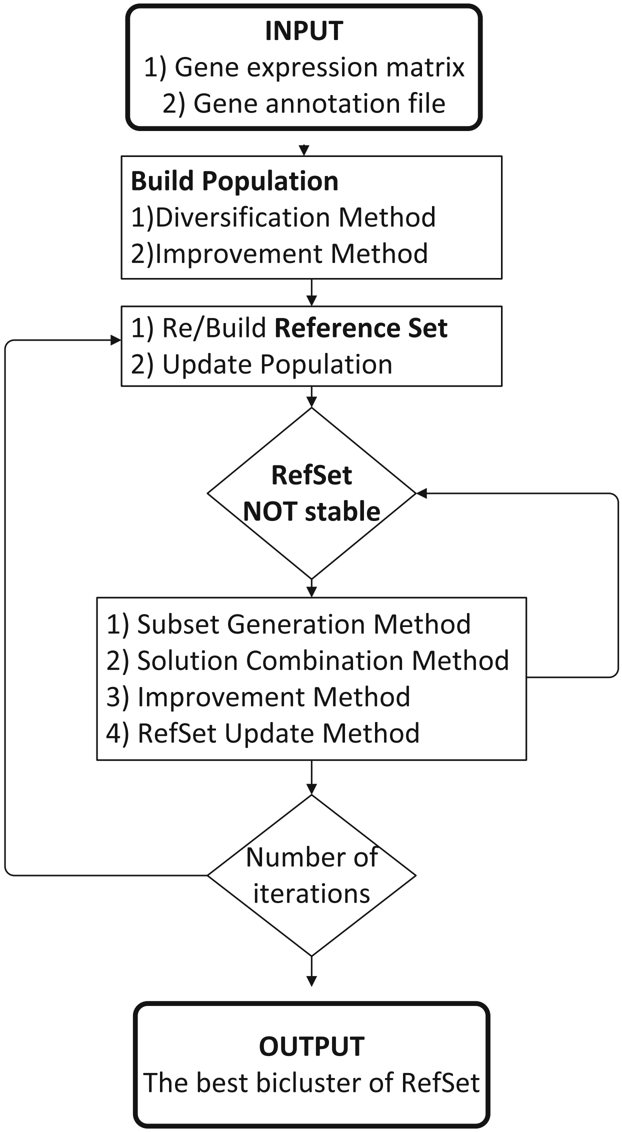

Fig. 1. Scatter search procedure to find a bicluster

of view. This method is a local search that intensifies the search process. Every improved solution is stored in the initial population.

The reference set is built according an intensification and a diversification criterion. Firstly, the half of the solutions in the reference set are most representative solutions from the initial population according to fitness function point of view. Secondly, the other half of solutions are the most scattered solutions respects the solutions previously considered. After that, the initial population is updated by removing these solutions in the reference set.

The reference set evolves until it is stable. Every new solution generated cannot improve from fitness point of view any solution previously stored in the reference set. Namely, every new solution is worse than the existent solutions in this set. New solutions are generated through the subset generation, solution combination and improvement method. The *subset generation method* builds new binary strings with uniform crossover operator usually used in genetics algorithms. The *solution combination method* builds new solutions combining with these strings. Each new solution is improved with the improvement method as it has previously commented. The *reference set update method* chooses the best solutions from the new generated solutions and the solutions in the reference set. This stability process of the reference set is repeated a number of times and the reference set is rebuilt. Finally, the best solution in the last reference set codifies the output bicluster. This procedure is repeated as many times as biclusters to discover. Note that the proposed algorithm obtains each bicluster through an independent scatter search procedure, which is a heuristic with a random nature. Therefore, the biclusters obtained are independent each other. The number of biclusters is an input parameter depending of the user needs or preferences.

2.2 *SimUI*-Based Fitness Function

A merit function is defined in order to characterize biclusters. The optimization of this function provides the resultant biclusters. Three different criteria are considered: the volume, the average correlation of the bicluster and finally, the *SimUI* measure of the set of genes. Given a bicluster composed by N genes and Q conditions, the fitness function is defined as follows:

$$f(B) = M_1 \cdot \frac{1}{N \cdot Q} + M_2 \cdot f_{corr}(B) + M_3 \cdot f_{SimUI}(B) \tag{2}$$

where the first term measures the volume, the second term measures the patterns in the gene expression matrix and the third one measures the information in GO of the set of genes according SimUI measure.

The average correlation is based on the correlation by pairs of genes. It has been previously used [9] and it is defined as:

$$f_{corr}(B) = 1 - \frac{1}{\binom{N}{2}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} |\rho_{ij}| \tag{3}$$

where ρ_{ij} is the pearson correlation coefficient between the genes g_i and g_j. Note that this correlation is calculated using the rows and the columns in the submatrix generated with the bicluster information in the gene expression matrix. The best value for the correlation is equal to 1, and the goal is to minimize the fitness function, therefore f_{corr} is modified to have the best value in 0. The absolute value is considered to capture positive and negative correlations.

The third term in the fitness function measures the functional similarities among genes using SimUI GO measure. The gene annotation file is used to introduce the biological information in GO. This term is defined as follows:

$$f_{SimUI}(B) = 1 - \frac{1}{\binom{N}{2}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} SimUI(g_i, g_j) \tag{4}$$

where $SimUI(g_i, g_j)$ is the SimUI measure for the genes g_i and g_j. Note that the best situation is a value equal to 1. For example, the measure of a gen with itself that is the maximum similarity. Hence, this term is also modified to find the optimum value in 0.

It is important to note that the first and the second term use the gene expression matrix and besides, the third term uses the gene functional annotation file. The parameters in Eq. (2) M_1, M_2 and M_3 control the importance of each term.

The **SimUI measure** is a gene pairwise GO measure based on counting in the graph of GO [7]. A distance between two genes is established according their functional similarity in GO. The gene annotation file is built such that each gene is related with a collection of GO terms. These terms include direct annotations of the genes and their ancestral terms up to the root node. Given two genes g_1 and g_2, the *SimUI* measure is defined as follows:

$$SimUI(g_1, g_2) = \frac{COUNT_{t \in GO(g_1) \cap GO(g_2)}}{COUNT_{t \in GO(g_1) \cup GO(g_2)}} \tag{5}$$

where *COUNT* is a function to count the number of GO terms. It must be noted that if there is not any annotation for g_1 and g_2, the measure is chosen equal to zero in order to represent a bad situation during the optimization process.

3 Experimental Results

The goal of the experiments is to study the influence of the SimUI measure as a mechanism to integrate biological information. The performance of the proposed algorithm is evaluated with and without any biological information integration. Note that if the third term (Eq. 2) is null there is not any biological information. In this case, the information captured by the gene annotation file is not used.

Table 1 presents the gene expression matrix size, the number of annotated genes in the matrix and a short description of the gene annotation file. The number of terms is the total number of GO terms in the gene annotation file. The average terms per gene is the average number of GO terms associated for each gene. A yeast dataset with accession number GDS1116 *GEO* repository has been used in these experiments. After being processed using *Babelomic* web tool, GDS116 gene expression matrix is composed by 882 genes and 131 conditions. The gene functional annotation file is a gene association file with the extension

Table 1. Functional annotations from GO for GDS1116 yeast dataset: input and validation data.

Dataset	Size	GO Functional annotation source	Number of genes	Number of terms	Avg. terms per gene
GDS1116	(882 × 131)	INPUT	798	1790	7.3
		BP domains	632	245	10.6
		MF domains	634	135	5.5
		CC domains	703	44	3.1

.goa_yeast downloaded from the additional information in the reference [8]. An extra file with the GO structure has been downloaded too. This file has the extension *.obo*.

The comparison criterion usually used in biclustering is the gene enrichment in the resultant biclusters [4]. A bicluster is said to be enriched if at least one enriched GO term is associated to it. The idea is that a group of genes is associated with any biological function in GO in this bicluster. Therefore it is biologically relevant. A ranking of algorithms can be defined depending on the percentage of enriched biclusters in their results. A gene annotation file is necessary to calculate the enrichment of biclusters but the same gene annotation file used as input could introduce a bias in the validation. Therefore, three gene annotation files have been generated for validation tasks different from the file used as input. The information of these files is also shown in Table 1. Three files for each GO domain have been generated using *Babelomic* web tool: for Biological Process (BP), Molecular Function (MF) and Cellular Component (CC).

The parameters configuration in the fitness function (Eq. 2) has been studied. The third term indicates SimUI term and it indicates if it is used to take the biological information from gene annotation file. Three parameter configurations are studied: (211), (212) and (221). Firstly, the fitness function configuration (211) gives the same importance to the average correlation (Eq. 3) that to the SimUI measure (Eq. 4). Secondly, the configuration (212) gives more importance to the SimUI over the average correlation. Finally, the average correlation is more relevant for the (221) configuration. If the parameter M_3 is equal to 0 there is only two possible configurations (210) and (220). It is important to note that the three terms in the fitness function vary between 0 to 1. Every configuration has the parameter M_1 equal to 2 due to the previous experience that shows that the volume must equal or more relevant that the average correlation in order to avoid trivial biclusters. It must be noted that several values varying from 1 to 2 for M_2 and M_3 were considered in previous studies [6] but the results did not show meaningful differences. (Namely, 1.0, 1.3, 1.5, 1.8 and 2).

Figure 2 shows a comparison among different fitness function configurations: (211), (212) and (221) for SimUI and (210) and (220) in the case that there is not any biological integration. The black bar, the grey and the white represent the BP, MF and CC GO domains respectively. They represents the percentage

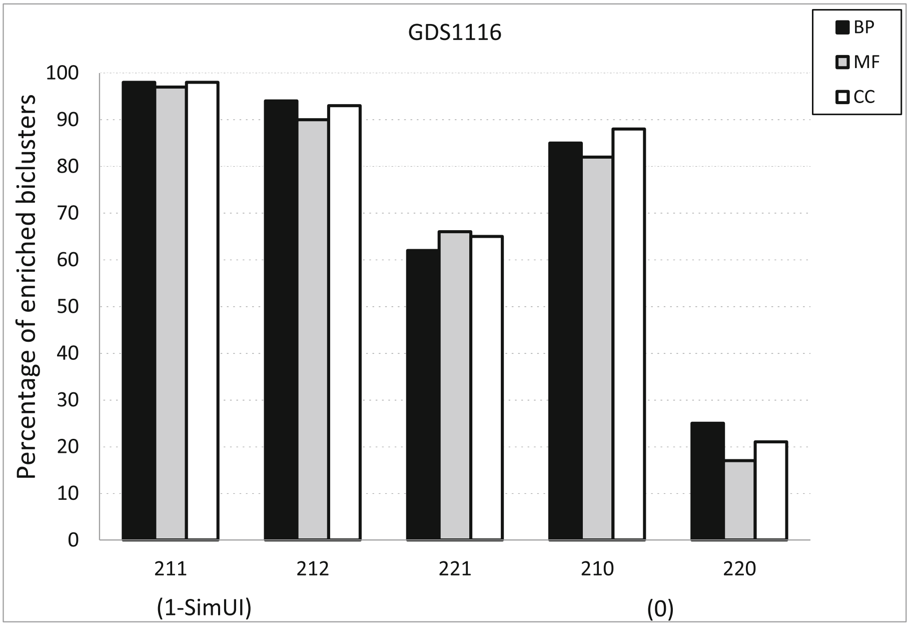

Fig. 2. Results with (1-SimUI) and without (0) biological information integration.

of enriched biclusters for each run of the algorithm. All runs have obtained 100 biclusters in order to have a high number of results to handle the random nature of the algorithm.

It can be observed in Fig. 2 that the best results are obtained when M_2 is equal to 1. In the case that SimUI is used, (211) and (221) has more percentage of enriched biclusters than (221). If there is not any biological integration, (210) obtains better results than (220). The configuration (211) obtains the best results and nearly the 100 % of biclusters are enriched. This case considers the average correlation (Eq. 3) and the SimUI measure (Eq. 4) with the same relevance in the fitness function (Eq. 2). The results for the three different domains BP, MF and CC present almost the same situation. These three domains has been considered in order to avoid any kind of bias in the validation. It is important to note that gene annotation file used as input is different from the files used to build Fig. 2.

The most relevant information that can be observed in Fig. 2 is that the integration of biological information using SimUI measure improves the results of the algorithm when there is not any kind of integration. (211) and (212) obtains better results than (210) and it also occurs with (221) respect (220).

4 Conclusions

A scatter search based-biclustering algorithm that integrates GO information has been presented in this paper. Each reported bicluster is the result of the optimization process of a fitness function. This fitness function evaluates the quality of each bicluster and SimUI similarity measure is used as a term to integrate the information stored in GO. The input data of the algorithm are the gene expression data and a gene annotation file. Thus, a biological bias is introduced in the search process to address the results to biologically relevant biclusters.

Experimental results have shown the effect of integration of the biological information through SimUI measure. Different parameters configurations have been studied and it can be concluded that SimUI based-fitness function helps to improve the algorithm performance. Future works will be focused on some improvements in the search procedure and the study of other gene pairwise GO measures.

Acknowledgements. We would like to thank Spanish Ministry of Science and Innovation, Junta de Andalucía and University Pablo de Olavide for the financial support under projects TIN2011-28956-C02-02, TIN2014-55894-C2-R, P12-TIC-1728 and APPB813097, respectively.

References

1. Eren, K., Deveci, M., Kucuktunc, O., Catalyurek, U.V.: A comparative analysis of biclustering algorithms for gene expression data. Briefings Bioinform. **14**(3), 279–292 (2013)
2. Divina, F., Aguilar-Ruiz, J.: A multi-objective approach to discover biclusters in microarray data. In: Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation, pp. 385–392. ACM, New York (2007)
3. Flores, J.L., Inza, I., Larrañaga, P., Calvo, B.: A new measure for gene expression biclustering based on non-parametric correlation. Comput. Methods Programs Biomed. **112**(3), 367–397 (2013)
4. Prelic, A., Bleuler, S., Zimmermann, P., Wille, A., Buhlmann, P., Gruissem, W., Hennig, L., Thiele, L., Zitzler, E.: A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics **22**(9), 1122–1129 (2006)
5. Verbanck, M., Le, S., Pages, J.: A new unsupervised gene clustering algorithm based on the integration of biological knowledge into expression data. BMC Bioinform. **14**(1), 42 (2013)
6. Nepomuceno, J.A., Troncoso, A., Nepomuceno-Chamorro, I.A., Aguilar-Ruiz, J.: Integrating biological knowledge based on functional annotations for biclustering of gene expression data. Comput. Methods Programs Biomed. **119**(3), 163–180 (2015)
7. Pesquita, C., Faria, D., Bastos, H., Ferreira, A., Falcao, A., Couto, F.: Metrics for go based protein semantic similarity: a systematic evaluation. BMC Bioinform. **9**(Suppl 5), S4 (2008)
8. Caniza, H., Romero, A.E., Heron, S., Yang, H., Devoto, A., Frasca, M., Mesiti, M., Valentini, G., Paccanaro, A.: Gossto: a stand-alone application and a web tool for calculating semantic similarities on the gene ontology. Bioinformatics **30**(15), 2235–2236 (2014)
9. Nepomuceno, J.A., Troncoso, A., Aguilar-Ruiz, J.: Biclustering of gene expression data by correlation-based scatter search. BioData Min. **4**(1), 3 (2011)