Applied Soft Computing 35 (2015) 637–651

Contents lists available at ScienceDirect

Applied Soft Computing

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

Scatter search-based identification of local patterns with positive and negative correlations in gene expression data

Juan A. Nepomuceno [a,*], Alicia Troncoso [b], Jesús S. Aguilar-Ruiz [b]

[a] Departamento de Lenguajes y Sistemas Informáticos, Universidad de Sevilla, Avd. Reina Mercedes s/n, 41012 Seville, Spain
[b] Department of Computer Science, School of Engineering, Pablo de Olavide University, Ctra. Utrera km. 1, 41013 Seville, Spain

ARTICLE INFO

Article history:
Received 15 April 2014
Received in revised form 22 May 2015
Accepted 18 June 2015
Available online 9 July 2015

Keywords:
Biclustering
Scatter search
Gene expression data

ABSTRACT

This paper presents a scatter search approach based on linear correlations among genes to find biclusters, which include both shifting and scaling patterns and negatively correlated patterns contrarily to most of correlation-based algorithms published in the literature. The methodology established here for comparison is based on a priori biological information stored in the well-known repository *Gene Ontology* (GO). In particular, the three existing categories in GO, *Biological Process*, *Cellular Components* and *Molecular Function*, have been used. The performance of the proposed algorithm has been compared to other benchmark biclustering algorithms, specifically a group of classical biclustering algorithms and two algorithms that use correlation-based merit functions. The proposed algorithm outperforms the benchmark algorithms and finds patterns based on negative correlations. Although these patterns contain important relationship among genes, they are not found by most of biclustering algorithms. The experimental study also shows the importance of the size in a bicluster in addition to the value of its correlation. In particular, the size of a bicluster has an influence over its enrichment in a GO term.

1. Introduction

Gene expression data provide the information that is collected from a group of microarray chips, each of which is built for a specific sample. The samples are generated according to a concrete experimental condition such as temperature, steps in the cell cycle or characterization of a patient. One single chip measures the expression level of thousands of genes in the sample under study [1]. After several preprocessing procedures, which comprise a process known as *low level microarray analysis*, the joining of the data from all of the samples constitutes the gene expression data to be analyzed. A gene expression matrix can be considered a two-dimensional numerical matrix, in which the rows are genes and the columns are the experimental conditions that are under study in each sample. A value in the matrix represents the gene expression value under a specific experimental condition. Data Mining techniques applied to infer knowledge from high-dimensional gene expression data sets comprise a process known as *high-level microarray analysis*. Most of these techniques are motivated by a simple idea, which is widely used in functional genomics: co-expression means co-regulation [2]. This assumption is called the guilt-by-association heuristic and is essential to study biological systems through "omic" data analysis.

Biclustering is a *unsupervised machine learning* technique that simultaneously clusters instances and features of the data set matrix. Unlike most of the *clustering* techniques, biclustering allows the overlapping among the results instead of making clusters which divide the data space. Therefore, the motivation is more to discover hidden information than to describe the data. Biclustering is a NP-hard problem considered first by Morgan and Sonquist [3], and later by Hartigan [4] and by Mirkin [5]. It can be found in the literature with other names, such as co-clustering [6] or subspace clustering [7]. In the context of gene expression data analysis, biclustering identifies patterns from gene expression data [8] and it was introduced by Cheng and Church [9].

Most of biclustering algorithms use the *mean squared residue* (MSR) measure [9] to obtain biclusters. Although scaling patterns are essential from a biological point of view, the MSR does not capture them when the gene variance values are high in the bicluster [10]. Recently, other measures based on correlations have been proposed to find biclusters. These measures are able to capture shifting and scaling patterns but do not obtain *activation–inhibition* expression patterns, which were presented in [11] and are a common feature in many molecular pathways [12,13].

* Corresponding author. Tel.: +34 954 559 769.
E-mail addresses: janepo@us.es (J.A. Nepomuceno), ali@upo.es (A. Troncoso), aguilar@upo.es (J.S. Aguilar-Ruiz).

http://dx.doi.org/10.1016/j.asoc.2015.06.019
1568-4946/

BIclustering algorithm based on a **S**catter **S**earch scheme, called BISS, is presented in this paper. The BISS approach attempts to overcome all the drawbacks of the existing biclustering algorithms to find biclusters including activation–inhibition patterns, in addition to both shifting and scaling patterns. The scatter search metaheuristic is a population-based evolutionary optimization method that emphasizes systematic processes against random procedures, in contrast to genetic algorithms [14]. The initial population of solutions is built using a diversification method, which non-randomly generates solutions with special characteristics. The fitness function is based on linear correlations among genes but also it considers negative correlations. The optimization process constitutes the evolution of a small set of solutions and includes a local search procedure, which intensifies the search without losing information from the scatter solutions of the problem. To evaluate the proposed algorithm two experiments have been carried out. As initial step, the proposed algorithm has been compared to classical algorithms of biclustering to analyze its potential to obtain biclusters. Secondly, other existing approaches based on correlations have been used with the purpose of comparing the kind of patterns discovered by the proposed algorithm.

The remainder of this paper is organized as follows. A review of the bibliography of biclustering is provided in Section 2. In Section 3, the proposed algorithm is presented to account for two aspects: first, how the search engine works, and second, the description of the merit function. Section 4 summarizes the experimental results, including a comparison of the performance of our approach to other biclustering methods. Finally, Section 5 is devoted to conclusions and future research.

2. Related research

Many biclustering algorithms have been proposed in recent years. These algorithms can be classified according to the type of patterns that are found, the size of the biclusters or the heuristic strategies that are used [8,15,16]. There is not a common criterion to compare different algorithms [17,18]. Some comparison methodologies are based on statistical metrics [19] or on the study of the behaviour over known synthetic data sets [20]. However, a sufficiently accepted methodology is based on enrichment analyses and uses a priori biological information that is stored in known repositories, such as GO [21].

The first proposals for solving the problem of searching for local patterns in gene expression data stem from the clustering field. An iterative hierarchical clustering is separately applied to genes and conditions, and the resultant biclusters are the combination of obtained clusters for each dimension in [22]. In 2000, Cheng and Church were the first to consider the biclustering in the context of gene expression data [9]. The Cheng and Church algorithm (CHCH) is a deterministic greedy iterative search method that obtains biclusters with a low MSR. If I and J are the sets of rows (genes) and columns (conditions) in a specific bicluster, respectively, then the MSR is defined as:

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

where a_{ij} is the expression value in row i and column j, a_{iJ} is the mean of the expression values in row i, a_{Ij} is the mean of the expression values in column j and a_{IJ} is the mean of the complete bicluster. The algorithm begins with the whole data matrix, and it iteratively adds or removes rows and columns until it finds a bicluster with a residue that is less than a given threshold. The process is repeated until the required number of biclusters is obtained. The number of biclusters to be obtained is an input parameter. The FLexible Overlapped biClustering (FLOC) algorithm [23] improves CHCH by obtaining simultaneously a set of biclusters and incorporating a strategy for addressing missing values in the process.

During recent years, several algorithms based on different techniques have been proposed. The Iterative Signature Algorithm (ISA) [24] is a nondeterministic greedy algorithm that finds up- and down-regulated biclusters. The algorithm starts with a random set of rows, and it iteratively updates columns and rows until convergence. Specifically, each column and row in a bicluster must have an average value that is less than several parameters, which measure symmetric requirements to obtain up- or down-regulated biclusters. This process is repeated using different seeds. The Order Preserving Submatrix (OPSM) algorithm [25] is a deterministic greedy algorithm that searches for biclusters according to a model that is based on linear ordering among rows. Most of the interesting patterns, such as constant values, shifting or scaling, are captured by this model. The biclusters are built through a scoring system, and the best bicluster is selected for each iteration of the algorithm. The Statistical-Algorithmic Method for Bicluster Analysis (SAMBA) [26] is a greedy algorithm that is based on an exhaustive bicluster enumeration using a bipartite graph model. It adds or removes nodes to find maximum weight subgraphs. The Plaid Model [27] algorithm is a statistical modelling approach that represents the input matrix as a superposition of layers, where each layer corresponds to a bicluster. It iteratively adjusts the parameters of each layer to handle its MSR. Spectral biclustering [28] identifies biclusters using techniques from linear algebra, especially eigenvector calculus. The idea comprises capturing up- or down-regulated biclusters with a variance that is lower than a given threshold. The characterization of the biclusters as hyperplanes in a high-dimensional space is the goal of several algorithms, which use image processing techniques [29] or Hough transform-based hyperplane detection algorithms [30].

The combination of a search engine and a measure characterizing the patterns that are sought is the methodology followed by a broad family of biclustering algorithms. These algorithms are optimization metaheuristics which are adapted to gene expression data, such as evolutionary approaches [31–33], multiobjective evolutionary approaches [34,35], greedy randomized adaptive search [36], simulated annealing [37], particle swarm Optimization [38] or estimation of distribution algorithms [39]. Most of these algorithms use the MSR as part of their merit function to characterize the types of patterns that are relevant to be found.

Recently, several biclustering algorithms using measures based on correlations have been proposed to obtain co-expressed genes [40–46]. In particular, BCCA [44] is a nondeterministic greedy algorithm which builds an initial bicluster composed of a pair of genes by removing experimental conditions while the Pearson correlation coefficient is lesser than a given threshold, and later, all the genes maintaining the correlation are added to this bicluster. In the same way, BICLIC [45] obtains a seed bicluster by applying clustering to each dimension separately, and second, the biclusters are expanded during the search process. The algorithm presented in [46] is based on a scatter search scheme. Although this algorithm could be considered to belong to the family of biclustering algorithms based on optimization metaheuristics, it uses the correlation instead of the MSR as a merit function. Although these methods use measures based on correlations, neither of them except for BICLIC captures negative correlations among genes.

Several biclustering algorithms which are specifically designed for binary datasets [21,47] or time series gene expression data [48,49] can be found in the literature. In the case of algorithms for binary datasets, a discretization step is necessary before the algorithm is applied, and therefore, a preprocessed matrix is used instead of the original gene expression data matrix. Thus, the preprocessing step is essential in this type of approach. In the case of time series data, the biclusters must have contiguous