Neurocomputing 145 (2014) 14–22

Contents lists available at ScienceDirect

Neurocomputing

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

A memetic algorithm for discovering negative correlation biclusters of DNA microarray data

Wassim Ayadi [a,b], Jin-Kao Hao [b,*]

[a] LaTICE, School of Science and Technology of Tunis, University of Tunis, 1008 Tunis, Tunisia
[b] LERIA, University of Angers, 2 Boulevard Lavoisier, 49045 Angers, France

ARTICLE INFO

Article history:
Received 18 January 2014
Received in revised form
6 May 2014
Accepted 17 May 2014
Available online 11 July 2014

Keywords:
Biclustering
Microarrays data
Negative correlations
Memetic algorithm

ABSTRACT

Most biclustering algorithms for microarrays data analysis focus on positive correlations of genes. However, recent studies demonstrate that groups of biologically significant genes can show negative correlations as well. So, discovering negatively correlated patterns from microarrays data represents a real need. In this paper, we propose a Memetic Biclustering Algorithm (MBA) which is able to detect negatively correlated biclusters. The performance of the method is evaluated based on two well-known microarray datasets (*Yeast cell cycle* and *Saccharomyces cerevisiae*), showing that MBA is able to obtain statistically and biologically significant biclusters.

1. Introduction

DNA microarray technology permits to measure simultaneously the expression levels of thousands of genes under diverse experimental conditions. This technology typically generates large amounts of raw data that need to be analyzed to draw useful information for specific biological studies and medical applications. In this context, biclustering of DNA microarray data is a particularly interesting approach since it allows the simultaneous identification of groups of genes that show highly correlated expression patterns through groups of experimental conditions (samples) [8,31,25,24].

Gene expressions from DNA microarrays are usually represented by an $n \times m$ data matrix $M(I,J)$ where n and m are respectively the number of measured genes and the number of conditions (or time points). Each cell $M[i,j]$ $(i \in I = \{1,2,\ldots,n\}, j \in J = \{1,2,\ldots,m\})$ represents the expression level of the ith gene under the jth condition. A bicluster is a subset of genes associated with a subset of conditions, i.e., a couple (I',J') such that $I' \subseteq I$ and $J' \subseteq J$.

Given a data matrix $M(I,J)$, the biclustering problem consists in extracting from $M(I,J)$ a group of coherent and significant biclusters of large size. In its general form, the biclustering problem is NP-hard [17,31].

Existing biclustering algorithms can be grouped into two large classes [7]: those that adopt a systematic search approach and those that adopt a stochastic search framework, also called heuristic or metaheuristic approach. Representative systematic search algorithms include greedy [5,10,17,15,30,39], divide-and-conquer algorithms [27,35] and enumeration algorithms [4,3,29,37]. Stochastic search algorithms include neighborhood-based algorithms [6,14], GRASP [19,20] and evolutionary algorithms [13,21,22,32]. A recent review of various biclustering algorithms for biological data analysis is provided by Valente-Freitas et al. [40].

A majority of existing biclustering algorithms extract only positive correlated genes. However, recent studies show that a group of biologically significant genes can present negative correlations. Fig. 1 shows an example of these correlations. Contrary to the case of a positive correlation where genes present similar patterns, in a negative correlation, genes present opposite patterns.

For example, in their study on the development of expression patterns for Arabidopsis thaliana, Schmid et al. [36] found that two groups of genes show negative correlations from an early seed development stage to a late stage. In Zhao et al. [43], the authors considered the negative correlated genes. They found that genes YLR367W and YKR057W of the *Yeast* data share the same pattern, while they have a negative correlated pattern against gene YML009C under eight conditions. These genes are grouped into the same bicluster because they are involved in protein translation and translocation.

In this paper, we address the issue of finding negative correlations based on local pattern of gene expression profiles. The key originality of our MBA method concerns the use of *positive and negative bicluster patterns* both in its search strategies and neighborhood definition. Bicluster pattern is a characteristic representation of a bicluster and is

* Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2014.05.074
0925-2312/

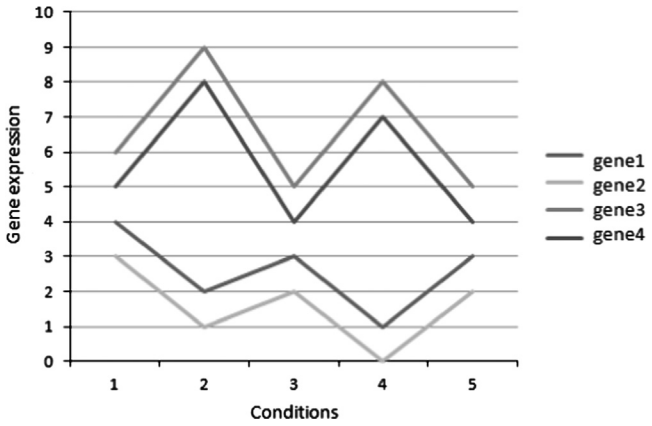

Fig. 1. Genes 1 and 2 are negatively correlated with the genes 3 and 4.

used to evaluate genes/conditions of biclusters. Positive bicluster pattern is used to improve the quality of a given initial positive bicluster, while the negative bicluster pattern is used to add negative correlation genes to the same bicluster. In the general case, if the absolute value of the correlation is considered, positive and negative correlation biclusters can be extracted without distinguishing the two types of correlations. However, the goal of our algorithm is to build the negative correlation biclusters.

The remainder of the paper is organized as follows: in Section 2, we present the *Average Spearman's Rho* (ASR) evaluation function. In Section 3, we describe the proposed MBA algorithm. In Section 4, experimental studies of MBA on real DNA microarray datasets are presented. Moreover, we illustrate a biological validation of some extracted biclusters via two web-tools, *FuncAssociate* [12] and *GOTermFinder*.[1] Conclusions are given in the last section.

2. The ASR evaluation function

Many evaluation functions exist for bicluster evaluation such as Euclidean distance, Pearson correlation and Mean Squared Residue (MSR). Among these measures, MSR is the most popular evaluation function [17]. It has been used by several biclustering algorithms [2,13,15,20,32,41,42]. However, MSR is deficient to assess correctly the quality of certain types of biclusters like multiplicative models [1,15,34,39].

In Ayadi et al. [4], the authors have proposed another evaluation function, called *Average Spearman's Rho* (ASR). Let (I',J') be a bicluster in a data matrix $M(I,J)$, the ASR evaluation function is then defined by

$$ASR(I',J') = 2*\max\left\{ \frac{\sum_{i\in I'}\sum_{j\in I';j\geq i+1}\rho_{ij}}{|I'|(|I'|-1)}, \quad \frac{\sum_{k\in J'}\sum_{l\in J';l\geq k+1}\rho_{kl}}{|J'|(|J'|-1)} \right\} \tag{1}$$

where ρ_{ij} $(i\neq j)$ is Spearman's rank correlation [28] associated with the row indices i and j in the bicluster (I',J'), ρ_{kl} $(k\neq l)$ is Spearman's rank correlation associated with the column indices k and l in the bicluster (I',J') and $ASR(I',J')\in[-1..1]$.

A high (resp. low) ASR value, close to 1 (resp. close to -1), indicates that the genes/conditions of the bicluster are positively (resp. negatively) correlated. ASR can thus be used to measure effectively both positive and negative correlations.

In the next section, we describe the proposed memetic biclustering algorithm MBA which uses the ASR measure.

[1] http://db.yeastgenome.org/cgi-bin/GO/goTermFinder.

3. The MBA algorithm

3.1. Memetic algorithm

A memetic algorithm (MA) is typically based on the population-based search and neighborhood-based local search [33]. The basic rationale behind a MA is to combine these two different search methods in order to take advantage of their complementary search strategies. Indeed, it is generally believed that the population-based search framework offers more facilities for exploration while neighborhood search provides more capabilities for exploitation. If they are combined in a suitable way, the resulting hybrid method can then offer a good balance between exploitation and exploration, assuring a high search performance [26].

Mitra and Banka [32] present a *Multi-Objective Evolutionary Algorithm* (MOEA) based on Pareto dominance. The authors try to find biclusters with maximum size and homogeneity by using a multi-objective genetic algorithm called *Non-dominated Sorting Genetic Algorithm* (NSGA-II) [18] in combination with a local search procedure. Gallo et al. [22] illustrate another MOEA algorithm combined with a local search strategy. They extract biclusters with multiple criteria like maximum rows, columns, homogeneity and row variance.

3.2. Preprocessing of gene expression matrix

Our algorithm applies a preprocessing step to transform the input data matrix M to a *Behavior Matrix* M'. This preprocessing step aims to highlight the trajectory patterns of genes. Indeed, according to Schmid et al. [36] and Zhao [43], a group of genes are considered to be biologically significant if they present negative correlations. Within the transformed matrix M', each row represents the trajectory pattern of a gene across all the combined conditions while each column represents the trajectory pattern of all the genes under a pair of particular conditions in the data matrix M. The whole matrix M' provides thus useful information for the identification of relevant correlation biclusters.

Formally, the behavior matrix M' is constructed progressively by merging pairs of columns (conditions) from the input data matrix M. Since M has n rows and m columns, there is $m(m-1)/2$ distinct combinations between columns, represented by J''. So, M' has n rows and $m(m-1)/2$ columns. M' is defined as follows:

$$M'[i,l] = \begin{cases} 1 & \text{if } M[i,k] < M[i,q] \\ -1 & \text{if } M[i,k] > M[i,q] \\ 0 & \text{if } M[i,k] = M[i,q] \end{cases} \tag{2}$$

with $i\in[1..n]$, $l\in[1..J'']$, $k\in[1..m-1]$, $q\in[2..m]$ and $q\geq k+1$.

Fig. 2 shows an illustrative example. We can observe, by considering each row of M', the trajectory (or behavior) pattern of each gene through all the combined conditions, i.e., up (1), down (-1) and no change (0). This figure also shows the trajectory of all rows (genes) over combined columns (combined conditions). Similarly, the combinations of all the paired conditions give useful information since a bicluster may be composed of a subset of non-contiguous conditions. Our MBA algorithm uses M' to define its search space as well as its neighborhood that is critical for the search process.

3.3. General procedure of MBA

The key originality of MBA concerns the use of *positive and negative bicluster patterns* both in its search strategies and neighborhood definition. The bicluster pattern is a characteristic representation of a bicluster. It can be used to evaluate genes/conditions of biclusters. The positive bicluster pattern is used to improve the quality of the positive bicluster B, and the negative bicluster