Information Sciences 181 (2011) 435–449

Contents lists available at ScienceDirect

Information Sciences

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

Noise-robust algorithm for identifying functionally associated biclusters from gene expression data ☆

Jaegyoon Ahn [a], Youngmi Yoon [b], Sanghyun Park [a,*]

[a] Computer Science Department, Yonsei University, South Korea
[b] Division of Information Technology, Gachon University of Medicine and Science, South Korea

ARTICLE INFO

Article history:
Received 17 February 2010
Received in revised form 11 September 2010
Accepted 6 October 2010

Index Terms:
Knowledge discovery
Data mining
Biclustering
Gene expression data analysis
Microarray analysis

ABSTRACT

Biclusters are subsets of genes that exhibit similar behavior over a set of conditions. A biclustering algorithm is a useful tool for uncovering groups of genes involved in the same cellular processes and groups of conditions under which these processes take place. In this paper, we propose a polynomial time algorithm to identify functionally highly correlated biclusters. Our algorithm identifies (1) gene sets that simultaneously exhibit additive, multiplicative, and combined patterns and allow high levels of noise, (2) multiple, possibly overlapped, and diverse gene sets, (3) biclusters that simultaneously exhibit negatively and positively correlated gene sets, and (4) gene sets for which the functional association is very high. We validate the level of functional association in our method by using the GO database, protein–protein interactions and KEGG pathways.

1. Introduction

Finding sets of co-regulated genes can lead to identification of their functionality and eventually the genetic pathways involved. Clustering co-regulated genes from a microarray dataset is performed by examining the expression values of the genes under various conditions. Each condition, also called a sample, denotes the state to which the gene was exposed (e.g., temperature) or the stage of an arbitrary cellular processes. Note that not all samples are observed in a particular cellular process, nor do all genes in a microarray dataset participate in it. We can expect subsets of genes to be co-regulated under certain experimental conditions, but to behave almost independently under other conditions [7]. The data mining technique used to find a submatrix of a functionally coherent gene and sample set in a microarray is known as biclustering, as presented by Cheng and Church [9]. A set of genes in a bicluster exhibits several patterns of expression values. The data matrix in Fig. 1(a) exhibits no obvious pattern when plotted in Fig. 1(b). Biclusters that exhibit various patterns can be identified from the data matrix with various biclustering algorithms, as indicated in Fig. 1(c)–(e).

Many biclustering algorithms have been introduced [17], and most variants of the biclustering problem have been shown to be NP-hard [9], so all of these algorithms use heuristic methods or probabilistic approximations. Accordingly, these algorithms have various strengths and weaknesses and each identifies different patterns. Biclustering algorithms can be generally divided into two groups according to the type of patterns:

☆ This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No. 2010-0008639).

* Corresponding author. Tel.: +82 2 2123 5714; fax: +82 2 365 2579.
E-mail addresses: ajk@cs.yonsei.ac.kr (J. Ahn), ymyoon@gachon.ac.kr (Y. Yoon), sanghyun@cs.yonsei.ac.kr (S. Park).

0020-0255/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2010.10.013