RN-Cluster: Discovering coherent biclusters which is Robust to Noise

Jaegyoon Ahn[1], Youngmi Yoon[1,2], and Sanghyun Park[1]
1. Computer Science Department, Yonsei University, South Korea
2. Information Technology Department, Gachon University of Medicine and Science, South Korea
{ajk, amyyoon, sanghyun} @cs.yonsei.ac.kr

Abstract

A bicluster is a subset of genes that show similar behavior within a subset of conditions. Biclustering algorithm is a useful tool to uncover groups of genes involved in the same cellular process and groups of conditions which take place in this process. We are proposing a polynomial time algorithm to identify functionally highly correlated biclusters. Our algorithm identifies 1) the gene set that follows additive, multiplicative, and combined patterns simultaneously that allow high level of noise, 2) the multiple, possibly overlapped, and diverse gene sets, 3) biclusters with negatively correlated as well as positively correlated gene set simultaneously, and 4) gene sets whose functional association is strongly high. We validated the level of functional association of our method, and compared with current methods using GO.

1. Introduction

Not all the genes in microarray dataset participate in a particular cellular process, and not all samples can be observed in a particular cellular process. We can expect subsets of genes to be co-regulated under certain experimental conditions, but to behave almost independently under other conditions [1]. Finding the set of co-regulated genes can lead to identify the functionality of the group of genes and eventually find the genetic pathways. Cheng [2] named the data mining technique that finds a submatrix of coherent gene set and sample set in a microarray as biclustering.

Many biclustering algorithms have been introduced, and biclustering is proven to be a NP-hard problem [2], so all these algorithms used heuristic methods or probabilistic approximation. Accordingly strengths and weaknesses of each algorithm are various and the patterns that each biclustering algorithm identified are also various. Biclustering algorithms could be divided into two groups largely by the patterns they find.

- *Algorithms that find additive or multiplicative patterns:*

δ-biclustering [2] uses mean squared residue of a submatrix to find biclusters. As a result, it finds additive or multiplicative co-regulation patterns. One weakness of δ-biclustering is that it allows only a small degree of noise. Thus it can identify strict patterns only. Also it can easily miss overlapping clusters due to the random value substitutions once a bicluster is identified.

p-Cluster [3] first scans the dataset to find all column-pair and row-pair maximal clusters called MDS. Then it does the pruning in turn using the row-pair MDS and the column-pair MDS. It then mines the final clusters based on a prefix tree. However, p-Cluster is not robust to noise, either.

Tri-Cluster [4] is the first algorithm that mines 3 dimensional microarray dataset. It makes a DFS(Depth First Search) tree whose node is the genes which show same range of fluctuation within user specified threshold ε. If ε is too big, DFS tree could grow too deep to complete the mining. However, Tri-Cluster with small ε does not allow high degree of noise. Moreover, its time complexity is exponential to the number of samples.

reg-Cluster [5] mines additive and multiplicative co-regulation patterns. It defines d_{ij} as a difference of the gene expression value between conditions c_i and c_j. Then it finds the gene set whose ratio of d01 and dij is within ε. Although the idea of mining additive and multiplicative patterns together is novel, it has a few problems. First, finding a proper ε is very difficult job. If ε is too big, the gene set in a bicluster would have many false positive genes. And if ε is too small, the gene set in a bicluster would have many false negative genes. Second, it constructs a DFS tree like Tri-Cluster. Thus it has same problems with Tri-Cluster.

This work was supported by the Korea Science and Engineering Foundation(KOSEF) grant funded by the Korea government(MOST) (No. R01-2006-000-11106-0).

978-0-7695-3191-5/08 $25.00 © 2008 IEEE
DOI 10.1109/BIOTECHNO.2008.8