DSS: A biclustering method to identify diverse and state specific gene modules in gene expression data

Jungrim Kim*, Yunku Yeu*, Jeongwoo Kim*, Youngmi Yoon[x] and Sanghyun Park*[†]
*Department of Computer Science, Yonsei University
Seoul, South Korea
Email: see http://delab.yonsei.ac.kr/eng/member
[x]Department of Computer Science, Gachon University
Seongnam, South Korea
Email: ymyoon@gachon.ac.kr
[y]Corresponding Author

Abstract The biclustering method is a useful co-clustering technique to identify biologically relevant gene modules. In this **paper, we propose a novel method to find not only functionally-related gene modules but also state specific gene modules by** applying a genetic algorithm to gene expression data. To identify **these gene modules, the proposed method finds biclusters in** which genes are statistically overexpressed or under expressed, and are differentially-expressed in the samples in the bicluster compared to the samples not in the bicluster. In addition, we improve the genetic algorithm by adding a selection pool for preserving the diversity of the population. The resulting gene modules exhibit better performances than comparative methods in the GO (Gene Ontology) term enrichment test and an analysis connection between gene modules and disease. This is especially the case with gene modules that receive the highest score in the breast cancer dataset; they are closely linked to the ribosome pathway. Recent studies show that dysregulation of ribosome biogenesis is associated with breast tumor progression.

I. Introduction

Due to the availability of large amounts of biological data, there have been many research studies for identifying new biologically valuable knowledge. Clustering is a technique for finding a gene module that shows a similar expression pattern across the set of all samples. Although it is a very useful technique for identifying relevant knowledge, it is difficult to find a disease-related gene module because disease-related gene modules do not affect the whole process of the disease progression [28]. As an example, subtypes of a heterogeneous disease like cancer are characterized by distinct genetic alteration. To overcome this limitation, a biclustering technique can be used. Biclustering is a co-clustering technique which allows simultaneous clustering of the genes and samples in order to find a gene module which shows crucial expression on specific samples. It requires more complicated calculations than the one-way clustering technique; accordingly, it has problems with time complexity. Thus, most of the research uses a heuristic method or probability approximation for finding gene modules.

Since Cheng and Church [31] have introduced a biclustering method to analyze gene expression data, many researchers have used biclustering methods [6], [7], [17], [23], [27] for analyzing gene expression data. To find biclusters, the CC (Cheng and Church) method calculates the mean squared residue score of candidate biclusters that exhibit an additive pattern. If this score is close to zero, it means that the bicluster becomes optimized. The order-preserving submatrix (OPSM) method [1] finds order-preserving submatrices (bicluster). Order-preserving submatrices have a permutation pattern where the columns of matrices are in non-decreasing order. The Iterative Signature Algorithm (ISA) [22] finds cis-regulatory biclusters of which the gene expression is high (at the gene and the sample) by giving a high weight to the gene and sample. The Debi (Differentially Expressed BIclusters) method [2] finds a bicluster that exhibits a statistical difference in the expression between the samples in the bicluster and the samples not in the bicluster, using a frequent item set approach. QUBIC (QUalitative BIClustering) algorithm [12] is a method which identifies biclusters efficiently with 'scaling patterns' utilizing a graph technique. Chakraborty et al [3] use a genetic algorithm to find biclusters without the threshold of the maximum allowable dissimilarity in gene expression data. Nepomuceno et al [18] present a scatter search approach to find biclusters based on linear correlations.

In this paper, we propose a new biclustering method which aims to find not only functionally-related gene modules, but also state specific gene modules. To find state specific gene modules, the proposed method finds biclusters where gene modules are statistically overexpressed or under expressed. Besides, gene modules in biclusters are differentially expressed between the samples in the bicluster and the samples not in the bicluster. Figure 1 shows an example of the biclusters that we want to find. In this figure, gene expression data are z-scored in each sample. Rows represent genes and columns represent samples. Colors of rectangles represent the degree of the gene expression. The more that gene expressions are statistically high, the closer this color is to green; the more that gene expressions are statistically low, the closer this color is to red. In addition, bolded rectangles represent biclusters. As shown in Figure 1, we find biclusters according to the following two rules: i) genes in a bicluster should be statistically coexpressed

978-1-5090-1897-0/16/$31.00