Subspace Clustering for Microarray Data Analysis: Multiple Criteria and Significance Assessment

[1]Hui Fang, [1]Chengxiang Zhai, [2]Lei Liu, and [1]Jiong Yang

[1]Department of Computer Science
[2]W. M. Keck Center for Comparative and Functional Genomics
University of Illinois, Urbana, 61801

As one of the latest breakthroughs in experimental molecular biology, microarray technology provides a powerful tool for monitoring the expression patterns of thousands of genes simultaneously, producing huge amounts of valuable gene expression data. Gene expression data are organized as matrices --- tables where rows represent genes, columns represent various samples such as tissues or experimental conditions, and a cell number indicates the expression level of a particular gene in a particular sample.

The goal of microarray data analysis is to identify a subset of genes whose expression levels rise and fall coherently under a subset of conditions, that is, they exhibit fluctuation of a similar shape when conditions change. Typically, clustering algorithms are first used to group together genes with similar patterns of expression. After clustering, people would usually inspect the clusters manually with annotation of genes and assign functions to clusters. Two major challenges are thus (1) to find gene clusters that are truly meaningful biologically; and (2) to rank and prioritize the clusters in some reasonable order for human inspection. We present a new clustering method that addresses both challenges.

Although many software tools for microarray analysis exist, they are inadequate for revealing all interesting gene clusters because they compute gene similarities based on all conditions/samples. As a result, only genes are similar in all conditions can be discovered. However, some patterns may only occur in a subset of conditions because the design of experiment conditions is often based on little knowledge of gene functions. Another major deficiency of the existing methods is that no confidence values are assigned to the discovered clusters. Therefore, it is difficult for biologists to prioritize their inspection of clusters. Analysis and interpretation of the microarray data is now a major bottleneck in utilization of microarray technology.

In this poster, we present a new clustering method for analyzing microarray data that improves existing approaches in three aspects -- capturing gene similarities under a subset of gene expression conditions, combining multiple criteria to capture trend similarity, and assigning statistical significance to detected clusters.

Our algorithm aims at discovering non-traditional subspace clusters. A subspace cluster is a subset of genes that exhibit similar expression patterns over a subset of conditions. To capture biologically meaningful clusters, we apply multiple criteria for constraining a subspace cluster. Specifically, we define our subspace cluster as a submatrix satisfying two distinct constraints -- fluctuation constraint and trend constraint. The fluctuation constraint requires that for all genes in a cluster, the difference of expression levels between two conditions need to be similar. The trend constraint captures the correlation between genes, i.e. when the expression level of one gene goes up under some conditions, the expression level of the correlated genes should also go up accordingly. In general, we are advocating constraining a cluster with multiple criteria capturing biological requirements from different perspectives.

Assessing the clustering results and interpreting the clusters are as important as generating the clusters. In this poster, we propose two quantitative ways for evaluating clusters. The first is to exploit the Gene Ontology (GO) tree-like structure. We use the depth of the common parent nodes shared by the genes in a cluster to assess the quality of the cluster. A deeper common parent is more specific and thus reveals more biological meaning. By using this measure, we show that our multi-criteria subspace clustering method can discover more coherent gene clusters than existing clustering algorithms on two different microarray data sets.

Our second method is to exploit the original sample data points to assess the statistical significance of the discovered clusters. Many steps in the process of generating microarray data can introduce variations. In the existing approaches, each cell in the microarray data matrix is usually the mean of several replicates. The variation in the original replicates is thus largely ignored. However, such variation information can help us assess the significance of a cluster. Intuitively, a gene may fall into a cluster purely because of the high variance in the replicates rather than biological relevance. As far as we know, no previous work has ever considered the variances in either clustering or evaluating clusters. In this poster, we propose a method to compute the confidence level for each generated cluster based on the original variances of cell values. The clusters can then be ranked according to their confidence levels. The proposed new method has a great potential for helping biologists discover meaningful gene clusters through generating more coherent clusters and ranking clusters based on statistical significance.