Trends in Bioinformatics - Presentation
“Biclustering using Biological Context Information”

With recent advances in sequencing technology, RNA sequencing has become the tool of choice to obtain gene expression data. In comparison to the traditional microarray technology, it yields a higher specificity and sensitivity and it supports detecting weakly expressed genes as well. This allows to study unknown gene functions. The underlying assumption is that genes sharing similar expression patterns also contribute to similar biological processes.

An approach to group genes into collections of akin patterns is clustering. The corresponding algorithms make it possible to discover disease-specific genes. However, studies have shown that a gene contributes to ten different biological processes on average[1]. Due to this fact, single clustering algorithms are not suitable to capture this property.

In contrast, biclustering approaches assign genes and samples to multiple clusters. Their main advantage is that they support finding overlapping clusters, a characteristic that reflects the behavior of genes very well. They also simultaneously assign both genes and samples to clusters. This allows interpreting genes of a cluster as contributors to biological processes that are shared across a specific set of samples only.

However, most biclustering approaches are limited as they only rely on patterns that can be found in the data itself. Hidden biological processes that are not represented by the sampled data can not be discovered. On the other hand, the algorithms can find structures that are not of interest. Including biological context information can support the algorithm in finding patterns that contribute to the same processes. By adjusting the added context information, one can regularize the algorithm to find structures that are of interest only. The results of such analyses can highly improve the plausibility of the found patterns since they no more only depend on statistical properties of the expression data.

We propose an approach that extends the biclustering algorithm BiMax[2] by including additional biological context information. The problem of finding biclusters is modeled as a bipartite graph problem. In the graph, one set of vertices represents the samples while the second set of vertices represents the genes. An edge between a gene and a sample is drawn if the gene is significantly expressed in the sample. We connect two nodes within the set of genes based on their biological interaction to introduce context information. We then search for the biggest cliques using the Bron-Kerbosch[3] algorithm to find biclusters.

We plan to evaluate the approach on a widely used dataset of The Cancer Genome Atlas (TCGA) project. It will be compared with state of the art algorithms with respect to quality and quantity. We expect our solution to stand out in finding new patterns that have a high biological relevance.

[1] The role of the genome project in determining gene function: insights from model organisms, Arnone & Davidson 1996
[2] A systematic comparison and evaluation of biclustering methods for gene expression data, Prelic et al. 2006
[3] Finding All Cliques of an Undirected Graph, Bron & Kerbosch 1973