Gene Expression Analysis Tools

Brodsky, L., Kositsky, M., Leontovich, A., Kalaidzidis, Y., Safro, I., Shtutman, M. and Feinstein, E.

Q.B.I. Enterprises Ltd., Ness-Ziona, Israel

We have developed a software package for microarray analysis, named GEA (for **G**ene **E**xpression **A**nalysis). Its core resides on three basic approaches: nested (hierarchical) clustering algorithm, "main vector" algorithm (correspondence of gene clusters to the matrix of distances between samples), and hierarchical bayesian network of dependencies between genes or gene clusters. The package also contains a proprietary module for quality control of cDNA microarray hybridizations.

Nested clustering provides the understanding of the geometry of distribution of gene expression patterns. It distinguishes the genes smoothly spread within a big area from those ones packed in compact clusters. The basic steps of plain clustering on every level of hierarchy are the following:

- Detection of the best surrounding for every gene according to the probability of its density under hypothesis of the uniformity of the distribution of genes in a wider area.
- Extraction of clusters from surroundings by combination of greedy method and k-means algorithm.

The same two steps are applied to gene-representatives of clusters to produce the next level of hierarchy. Some sub-clusters of an upper level cluster are merged if their separation is not statistically valid.

The aim of "main vector" algorithm is to detect which clusters of genes provide the major input into proximity of probes. Genes as vectors in R^k space (k – number of samples) are transformed into $R^{k(k-1)/2}$ space, where half-matrix of distances between samples is also presented as a vector of this space ("main vector"). The sum of all gene-vectors in this space is the main vector, and thus the clusters of genes making an essential input in the matrix of distances between samples can be distinguished.

This technique can be applied for filtering of genes according to some biological hypothesis (for instance, grouping of samples through the arrangement of the hypothetical matrix of distances between them), and to two way clustering of filtered genes and samples. Unlike Principal Component analysis, "main vector" algorithm can work not only with correlation (covariation) distances between samples, but with Euclidean distance and several other metrics as well.

Finally, hierarchical bayesian network with nodes as vectors of $R^{k(k-1)/2}$ space presents dependencies between nodes, which consists either of genes, or clusters of genes and different types of main vectors arranged for a given population of genes.