Mining literature to improve biological knowledge extraction by microarray transcriptional profiling

R.A. Calogero, S. Motta, G. Pedrazzi, S. Rago, E. Rossi, G. Iazzetti, R.Turra

Dip. Scienze Cliniche e Biologiche, Università di Torino

DNA microarray technology is a high throughput method for gaining information on gene function. It allows the simultaneous collection of quantitative data about the differential expression of thousands of genes at a time. This large amount of data is analysed to identify clusters of genes that share common expression characteristics, but the obtained results provide no information regarding the biological similarities of genes within clusters. The published literature, on the other hand, provides a potential source of information to assist in interpretation of clustering results.

We present a tool that enables the validation and improves the comprehension of the experimental results by identifying the main functions among a group of genes, as are reported in Medline documents. It also makes possible the identification of disease-specific genes, thus simplifying the design of specially-devoted microarray.

The tool relies on two components: a gene name extractor and a mining algorithm. The name extractor is based on existing dictionaries of gene names and aliases. The mining algorithm analyses the co-occurances of words in the selected documents in order to automatically interpret the context where the gene names appear and to map documents (and genes) into functional classes.