

CoExpress: a tool for an effective co-expression analysis of large microarray data sets

Petr V. Nazarov¹, Arnaud Muller¹, Viktor Khutko², Loïc Couder¹ and Laurent Vallar¹

¹) Microarray Center, CRP-Sante, Luxembourg

²) SUSS MicroTec Test Systems GmbH, Germany

Co-expression (CE) analysis of microarray data may provide interesting insights in understanding the gene and transcript level regulations in biological samples. It allows gene-networks reconstruction, disease pattern recognition, inferring of causal genes, etc. However, due to high computational costs and memory limitations, there is still a need in effective and user-friendly tools for the analysis of CE.

Here we propose a stand-alone software tool CoExpress for the interactive CE analysis of microarray data. The software is a user-friendly and allows on-the-fly study of CE, including (a) expression data preprocessing (b) building and visualization of CE matrix using correlation or mutual information metrics, (c) clustering, visualization and filtering of CE profiles, (d) visualization of co-expression networks for genes of interest. The possibility of the user-defined data processing using R-scripting is realized, providing a powerful tool for advanced users.

Due to its efficiency of memory usage and algorithm optimizations, a Windows version of CoExpress allows simultaneous analysis of CE for up-to 30000 genes or transcripts, measured on a hundred of arrays, in a reasonable time even on a standard PC. For a more time-consuming analysis, i.e. when working with thousands of experiments or/and using mutual information as a metric, a multi-thread command-line version has been developed that can be run on Linux multi-CPU systems. Due to the specificity of the CE calculation, the growth of productivity is almost linear with the increase of number of CPUs.

The work of the software was tested using data from 2428 Affymetrix HGU133plus2 array experiments, downloaded from public repositories and preprocessed using R/Bioconductor. Data were normalized using RMA and then summarized, using gene symbols as indexes. The resulting data matrix, containing measurements for 19894 unique gene symbols, were analyzed using the multi-thread version of CoExpress. The analysis revealed that 2812 genes are co-expressed with at least one other gene with the absolute correlation higher than 0.8.

The validation of the resulting network was performed using STRING service at string.embl.de. The gene sets with the same co-expression profile were compared with a set of genes randomly selected genes and showed significantly higher level of connectivity.

The up-to-date version of CoExpress and its multi-thread module are freely available for downloading from www.bioinformatics.lu. The multi-thread module is distributed together with its source code under the GPL, which allows to modify, recompile and run it under various OS.