CENTRE DE RECHERCHE PUBLIC SANTÉ

MICROARRAY CENTER

CoExpress: a Tool for an Effective Co-Expression Analysis of Large Microarray Data Sets

Petr V. Nazarov[1], Arnaud Muller[1], Viktar Khutko[2], Loïc Couderc[1] and Laurent Vallar[1]

Introduction

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.

Tool

Here we propose a free, 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:

- expression data **normalization** and R-based preprocessing (optionally);
- building and visualization of **CE matrix** using correlation or mutual information metrics;
- clustering, visualization and filtering of **CE profiles**;
- visualization of **co-expression networks** for genes of interest.

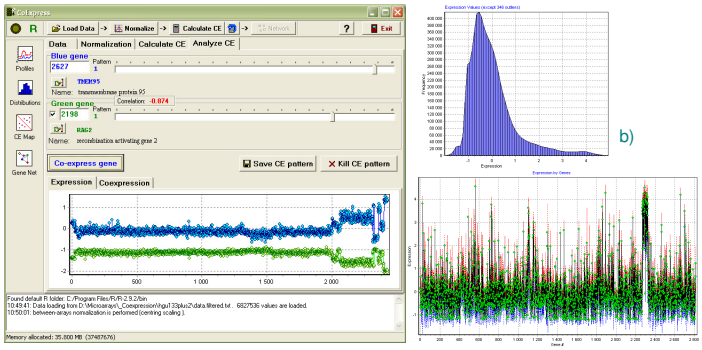

Fig. 1. CoExpress interface during manual analysis of co-expressed genes (a), together with additional information about a data set: distribution (b), and averaged expression profile by genes (c)

Technical Notes

The software tool exists in two versions:

- **Windows-based version** for an interactive data analysis and visualization
- **Linux command line version** for multithreading analysis of big datasets

The properties and comparative description of both variants are given below:

Table 1. Comparative description of two versions of CoExpress

Parameters	Windows	Linux
Maximum genes	~ 30 000	> 60 000
Maximum arrays	< 1 000	> 1 000
Multi-CPU support	-	+
Graphical User Inerface	+	-
Compiller	bcc32	gcc
Time for CE calculation on a big dataset*		
1 CPU	3h 45m	55 m
8 CPU	n/a	7m
Time for CE calculation on a small dataset**		
1 CPU	1m 26s	1m 13s

(*) 2428 Affymetrix arrays with 19894 genes were used

(**) 17 Agilent two-color arrays with 15375 genes were used

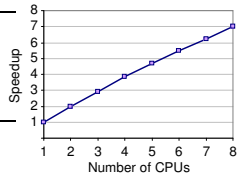

Fig. 2. Speedup with the increase of number of CPUs (on a big data set*)

The efficiency of the parallelization in the Linux version of CoExpress is demonstrated by Fig. 2: the growth of a speedup is almost linear with the increase of number of available processors.

The GUI for the Windows version was developed using Borland C++ Builder (Codegear).

Public Data Analysis

Software's performance 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 Linux version of CoExpress. The analysis revealed that **2812 genes are co-expressed** (each has at least one other gene with the absolute correlation $|r| \geq 0.8$).

The expression values for these 2812 genes where exported into a Windows-based CoExpress and further analyzed (see Fig.3 for their CE matrix). A major common network containing 2617 genes was detected, together with 67 smaller networks with no interconnection.

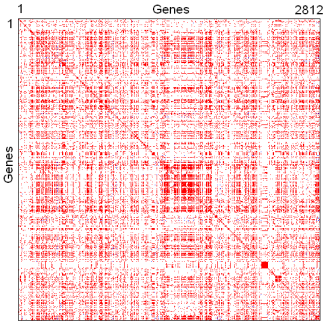

Fig. 3. Co-expression matrix for 2812 genes found after analysis of public Affymetrix data. Red dots show genes with r > 0.8, blue – with r < -0.8

To improve the relevance of the network we performed a linear between-array normalization on all 2812 co-expressed genes. This significantly increased the resolution of the analysis by revealing 139 smaller networks easier to handle and to study.

The validation of the obtained co-expression networks were performed using **STRING 8.2** (http://string.embl.de) – a service, public database and web resource dedicated to protein-protein interaction. This database integrates information coming from experiments, databases, text mining, etc. Two sets of genes, the genes connected by CoExpress into a network and randomly selected genes, where uploaded to STRING. The protein-protein interaction networks for these two data sets are presented in Fig.4. The connectivity of the inferred network is significantly higher than of a random network, suggesting that the data provided by CoExpress are in concordance with known biology.

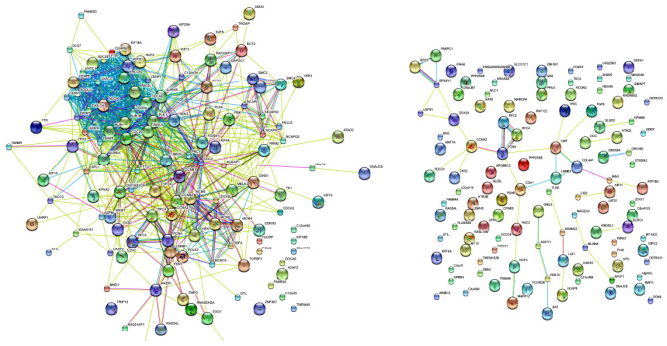

Fig. 4. Validation of CoExpress results using STRING 8.2: (a) protein-protein network for 127 genes from the same co-expression pattern, (b) "network" built on the same amount of randomly selected genes.

Concluding Remarks

CoExpress will be further developed towards introducing advanced network reconstruction methods and integration with public databases.

The **current version** of CoExpress and its multi-thread Linux version 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.

We would like to thank those, whose comments transform CoExpress into its current shape: Francisco Azuaje, Isabel Nepomuceno and Etienne Moussay.

1. Microarray Center, CRP-Santé, 84, Val Fleuri, L-1130, Luxembourg
2. SUSS MicroTec Test Systems GmbH, Germany

petr.nazarov@crp-sante.lu

MICROARRAY CENTER
PUBLIC RESEARCH CENTER FOR HEALTH (CRP-SANTÉ), LUXEMBOURG