CO-EXPRESSION ANALYSIS OF LARGE MICROARRAY DATA SETS USING COEXPRESS SOFTWARE TOOL

Petr V. Nazarov, Arnaud Muller, Viktar Khutko and Laurent Vallar

CRP-Santé, Luxembourg,
84 Val Fleuri, L-1526 Luxembourg,
petr.nazarov@crp-sante.lu

ABSTRACT

Here we propose a stand-alone software tool named CoExpress for the fast interactive co-expression (CE) analysis of microarray data. The software is user-friendly and allows on-the-fly study of CE, including microarray data preprocessing, building and visualization of CE matrix using correlation or mutual information metrics, clustering, visualizing and filtering of CE profiles, basic topological analysis, visualization and export of CE networks. The performance of the software was validated using simulated data and public data from a set of Affymetrix HGU133plus2 arrays.

1. INTRODUCTION

Gene regulatory networks (GRN) in living cells can be considered as extremely complex information processing systems. One of the main features of the GRN is their robustness and ability to form a proper biochemical respond to a wide range of extracellular conditions. The knowledge about the part of GRN related to a specific biofunction of cellular process is of extreme importance for controlling them. However, being a reverse-engineering task, the GRN reconstruction is highly challenging, and requires analysis of large sets of experimental data. One of the straightest ways to reconstruct GRN is based on co-expression (CE) analysis of transcriptomic data from cDNA microarrays. Two significantly co-expressed genes have the same or inverted expression profile over a number of experiments. Biologically this is a good evidence for either a direct interaction between the genes or their mutual participation in the same biofunction.

Despite the fact that co-expression-based methods of GRN prediction are widely used during the last few years, there is still a need for effective and user-friendly tools for the analysis of CE. The absence of such tools can be partially explained by high computational costs of the analysis and memory limitation of standard PCs.

Here we propose a stand-alone software tool CoExpress for the fast interactive CE analysis of microarray data.

Number of features distinguishes this tool from similar reported recently [1, 2]. (a) It allows an interactive data analysis. (b) A researcher can work on his own data or on a specifically selected subset of public data. (c) The possibility of the user-defined data processing by R-scripting provides a powerful tool for advanced users. (d) Visual version of CoExpress allows 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. (e) For a more time-consuming analysis (thousands of experiments) a multi-thread command-line version has been developed that can be run on Linux 64 bit multi-CPU systems.

2. METHODS AND TOOL

2.1. Methods

The simplest measure of CE is based on Pearson correlation coefficient (r) and can be calculated as:

$$C_{ij} = r^p = \left(\sum_{k=1}^{n_a} \frac{(x_{ik} - m_i)(x_{jk} - m_j)}{s_i s_j} \right)^p \qquad (1)$$

where i, j – indexes of considered genes, n_a – number of microarrays, m_i, s_i – mean and standard deviation for the expression of the i-th gene, p – weighting power.

Mutual information is the second wide spread method for CE estimation [1]. It can be effectively calculated as described in [3].

To build a CE network (undirected graph), only the CEs with absolute values higher than a specified threshold were considered.

In this paper we performed the analysis using correlation-based method because it is faster and more straight forward. Similar validation could be done for mutual information measure as well.

As the size of the CE matrix growth quadratically with increasing of gene number, it is a needed to optimize its allocation in the memory. After consideration of two alternatives: sparse matrix and approximate complete matrix, we decided to store the complete matrix as it is time-efficient. The values of CE were transformed into integers between –100 and 100 for memory optimization and stored into a triangular matrix.

2.2. Input data format

The expression data should be given to the software in the form tables stored as tab-separated text files. The layout of such a table is given in Table 1. First row is a header, the names of the first two columns should be "ID" and "Name" – they contain gene annotation. Other columns with arbitrary names contain expression values. The expression values should be log-transformed.

Table 1. The structure of input expression data table.

	ID	Name	<names of the arrays>		
Header					
Data section	<id of gene>	<name of gene>	<log2 expression values>		
	…	…	…	…	…
	TP53	tumour protein p53	4.92	5.04	…
	…	…	…	…	…

2.3. Software tool CoExpress

The software tool exists in two versions: Windows-based version for an interactive data analysis and visualization; and high throughput command line version (available in Linux and Windows, including 64 bit systems) for multithread analysis of big datasets. The properties and comparative description of both variants are given in Table 2. The GUI for the Windows version was developed using Borland C++ Builder (see Figure 1).

Table 2. Comparison 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

2.3.1. Windows GUI version

The analysis starts with importing of the data into the program. The imported data can be visualized using gene and array expression profiles and distributions.

The second step – data preprocessing can be performed using simple linear normalization within- or between arrays. As an alternative – the preprocessing can be performed using R-scripting. The script is automatically launched to modify the data.

The third step is the most time consuming and includes building of CE matrix and detection of groups of co-expressed genes (CE patterns).

At forth step the investigator can interactively check the expression profiles of genes of interest, save entire or a part of CE network, export original data only for relevant genes (data filtering) and visualize CE matrix and sub-networks.

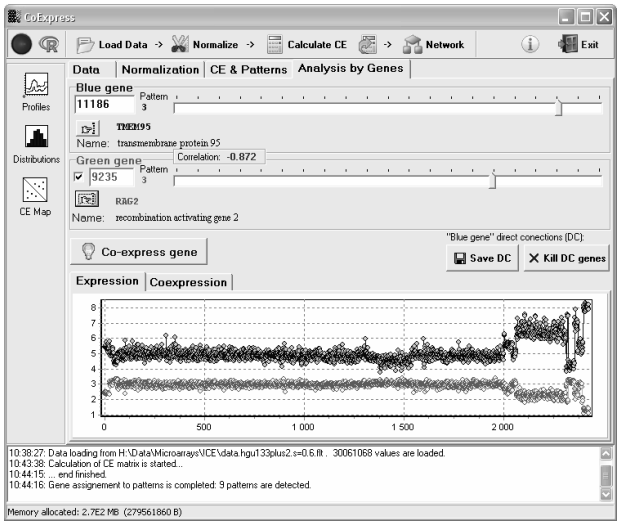

Figure 1. User interface of CoExpress during investigation of 2 co-expressed genes.

2.3.2. Multithread command-line version

The multithread version of CoExpressed is designed for high-throughput analysis. It is a console application, which can be recompiled for Linux and Windows systems. Multithreading is realized using Pthreads (POSIX Threads) library, existing for both Linux and Windows OS. Due to the specificity of the CE calculation, the growth of productivity is almost linear with the increase of number of CPUs: the 7x speed-up have been reached on an 8 CPU system under Linux.

The standard console command for correlation-based CE analysis is:

```
ce_calc.exe -t number_of_threads
      -p power -s threshold -i input_data_file
      -o output_CE_file -f output_filtered_file
```

3. RESULTS

3.1. Public data preprocessing

CoExpress was applied to experimental data from Affymetrix HGU133plus2 arrays, downloaded from thr public repository ArrayExpress [4]. These data are related to the analysis of samples from various human tissues.

Quality control was performed by R/Biocondictor package **simpleaffy** in order to detect and remove low quality and outlier arrays. As a result 2428 good quality arrays were considered for CE network reconstruction.

Background correction and data normalization was done by RMA algorithm [5] realized in R/Bioconductor (package **affy**). Then the data were summarized, using gene symbols as indexes. The poor annotated probe-sets were removed. The resulting data matrix contained measurements for 19894 unique gene symbols.

3.2. Co-expression analysis

Processed public data were analyzed by 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 $|r| > 0.8$, and 12 468 genes (63% of total number) having at least one $|r| > 0.6$.

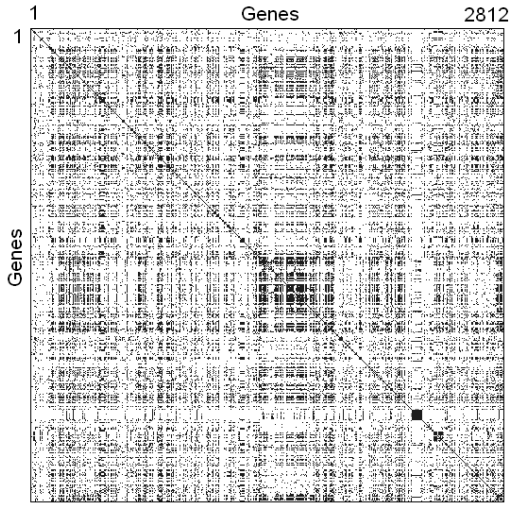

Figure 2. Co-expression matrix for 2812 genes found after analysis of public Affymetrix data. Dots show CE events for genes with $|r| > 0.8$.

The expression values for these 2812 genes where exported into a Windows-based CoExpress and further analyzed (see Figure 2 for their CE matrix). A major common network containing 2617 genes was detected, together with 67 smaller networks with no interconnection.

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 independent smaller networks easier to handle and to study.

4. DISCUSSION

4.1. Non-random CE patterns

As can be seen from Figure 2, the CE matrix obtained upon the analysis of public data discovers a huge set of ce-expressed gene. To exclude the possibility of array-specific effects we performed the study on 100 arrays (subset of lung-related arrays of original public dataset). As a control, we used randomized data, where the expression values where randomly mixed inside each array. The resulting distributions are given in Figure 3. The distribution for randomized arrays appeared significantly narrower than experimental distribution. In addition, experimental distribution is lightly skewed into the positive correlation domain, suggesting that positive interactions are much more common than negative interactions (inhibition) as was already shown in [6].

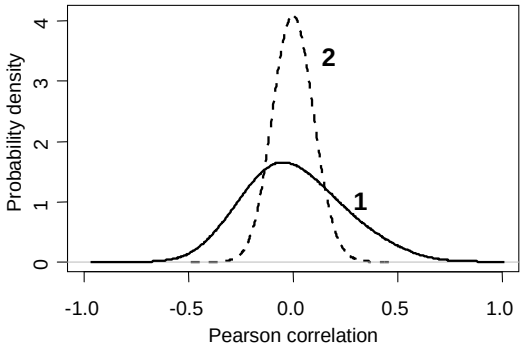

Figure 3. Distributions of correlation coefficients for experimental data (curve 1) and randomized data (curve 2).

4.2. Validation on simulated data

Validation of a method on simulated data is the most precise way of benchmarking, because it allows to compare the found outcomes to initially known ones. Here we have generated a mixture of random and co-expressed genes with different levels of signal-to-noise ratio. We considered 10000 genes measured over 100 microarrays. 20 genes where selected as the "core genes", which defined expression patterns. For each of those, 100 other co-expressed genes where generated using the simple equation:

$$e_{ij} = (1-a)e_{cj} + a\varepsilon_{ij}, \qquad 2)$$

where e_{ij} – expression for i-th gene on j-th microarray, e_{cj} – expression for the core gene, ε_{ij} – normal random value with the same mean and variance as e_{cj}, a – a noise fraction, varied for 100 genes from 0 to 1. The resulting behavior of true positives (TP) and false positives (FP) with respect to the specified cutting threshold is shown in Figure 4. It was found, that the minimal threshold value which do not introduce FP is 0.55. For this threshold we were able to detect on average 60 genes per co-expression pattern, which corresponds to the noise fraction $a < 0.6$. Therefore, the method can correctly detect even significantly distorted interactions.

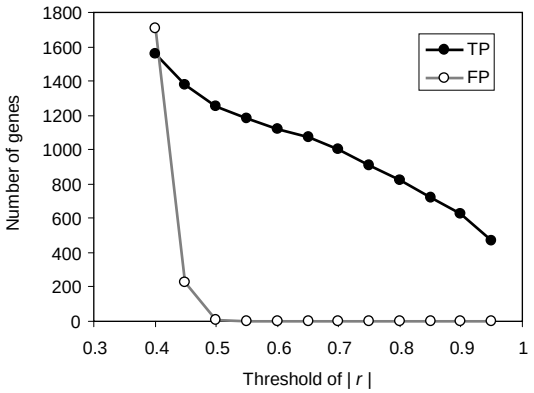

Figure 4. Behavior of true (TP) and false positives (FP) with respect to $|r|$ threshold.

4.3. Validation on experimental data

Next, we validated CoExpress on experimental data set. We have performed a bootstrapping experiment, during which the following actions were performed iteratively: (a) 10% randomly selected experimental arrays were excluded from the processed data set (243 of 2428); CE analysis is performed on the rest 90%; (c) the connections between genes with $|r|$ higher then the specified threshold were recorded. The lists of connections obtained after 100 runs were compared, resulting in concordance between reconstructed CE networks of 95% or 94% for $|r| > 0.6$ or 0.8, respectively. Thus we can conclude that this method of CE network reconstruction is robust.

4.4. Validation using STRING service

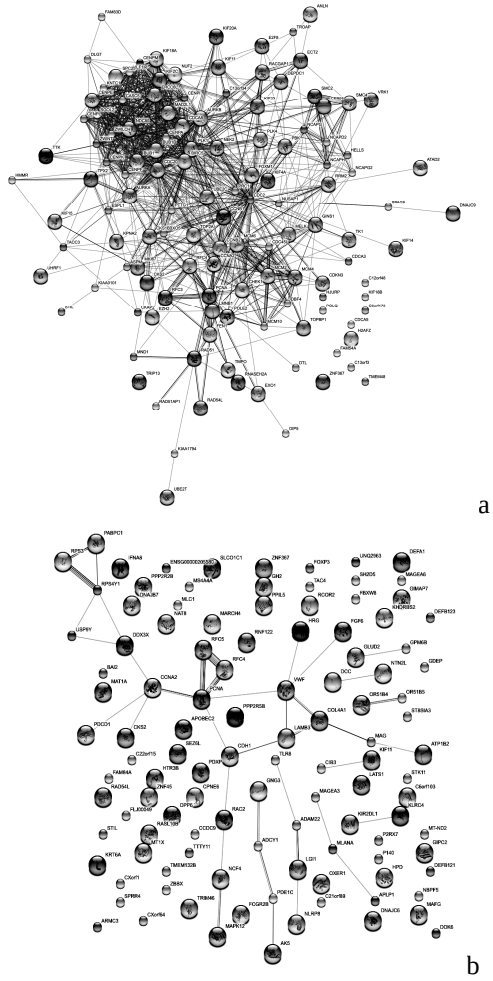

Figure 5. Validation of CoExpress results using STRING 8.2: (a) protein-protein network for 127 genes from the same co-expression pattern, (b) "network" built on the same amount of randomly selected genes.

Finally, CoExpress was validated using STRING 8.2 [7] – a service, public database and web resource, giving access to knowledge about protein-protein interaction. This database integrates information coming from various sources including experiments, databases and text mining. Two sets of genes where uploaded to STRING: the first set containing the genes connected by CoExpress into a network and the second one with genes randomly selected. The protein-protein interaction networks for these two data sets are presented in Figure 5. 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. Similar results were obtained when the random gene set was selected and validated by STRING 10 more times, suggesting that our result is not a coincidence.

5. CONCLUSION

As it was shown, both versions of the tool are able to work with big data sets. The validation using simulated data

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.

CoExpress will be further developed towards more advanced topological analysis, incorporation *a priory* knowledge about genes into CE network reconstruction and introducing more advanced network reconstruction methods, such as regression-based methods.

6. ACKNOWLEDGMENTS

We would like to thank Francisco Azuaje, Isabel Nepomuceno and Etienne Moussay for their useful comments about CoExpress.

7. REFERENCES

[1] A. A. Margolin, et al., "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinformatics*, vol. 7 Suppl 1, p. S7, 2006.

[2] D. Jupiter, H. Chen, and V. VanBuren, "STARNET 2: a web-based tool for accelerating discovery of gene regulatory networks using microarray co-expression data," *BMC Bioinformatics*, vol. 10, p. 332, 2009.

[3] P. Qiu, A. J. Gentles, and S. K. Plevritis, "Fast calculation of pairwise mutual information for gene regulatory network reconstruction," *Comput Methods Programs Biomed*, vol. 94, pp. 177-80, May 2009.

[4] ArrayExpress. http://www.ebi.ac.uk/microarray-as/ae/

[5] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed, "Summaries of Affymetrix GeneChip probe level data," *Nucleic Acids Res*, vol. 31, p. e15, Feb 15 2003.

[6] H. K. Lee, A. K. Hsu, J. Sajdak, J. Qin, and P. Pavlidis, "Coexpression analysis of human genes across many microarray data sets," *Genome Res*, vol. 14, pp. 1085-94, Jun 2004.

[7] STRING. http://string.embl.de