CORRELATION-BASED ANALYSIS OF LARGE TRANSCRIPTOMIC DATA SETS

P.V. Nazarov[1], V.V. Khutko[2], M.M. Yatskou[3], L. Vallar[1]

[1]CRP-Sante, Luxembourg; [2]SUSS MicroTec Test Systems GmbH, Germany;
[3]Belarusian State University, Belarus

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 bio-function.

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.

METHODS AND TOOL

Methods. Correlation-based CE for a pair of genes is calculated as a simple Pearson correlation coefficient between genes profiles. Often to increase the contrast of high correlation coefficients, correlation measure is powered (the most common power used is 6). 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 straightforward. 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 in a triangular matrix.

Input data format. The expression data should be given to the software in the form of a table stored as tab-separated text file. 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.

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 GUI for the Windows version was developed using Borland C++ Builder.

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 fourth 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.

Multithread command-line version. The multithread version of CoExpress 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

RESULTS

CoExpress was applied to experimental data from Affymetrix HGU133plus2 arrays, downloaded from the public repository ArrayExpress [3]. These data are related to the analysis of samples from various human tissues and contained 2428 good quality arrays. Normalized by RMA algorithm [4] form R/Bioconductor (package affy), public 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.

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 $|r| > 0.6$. The expression values for these 2812 genes where exported into a Windows-based CoExpress and further analyzed. 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.

DISCUSSION

Non-random CE patterns. The CE matrix obtained upon the analysis of public data discovers a huge set of co-expressed gene. To exclude the possibility of array-specific effects we performed the study on 100 selected arrays of this set. As a control, we used randomized data, where the expression values where randomly mixed inside each array. The resulting distributions are given in Figure 1a. 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 [5].

Validation on simulated data. Validation of a method on simulated data is the most precise way of benchmarking, because it allows comparing 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 de-fined expression patterns. For each of those, 100 other co-expressed genes where generated with various noise fraction *a,* varied 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 1b. 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.

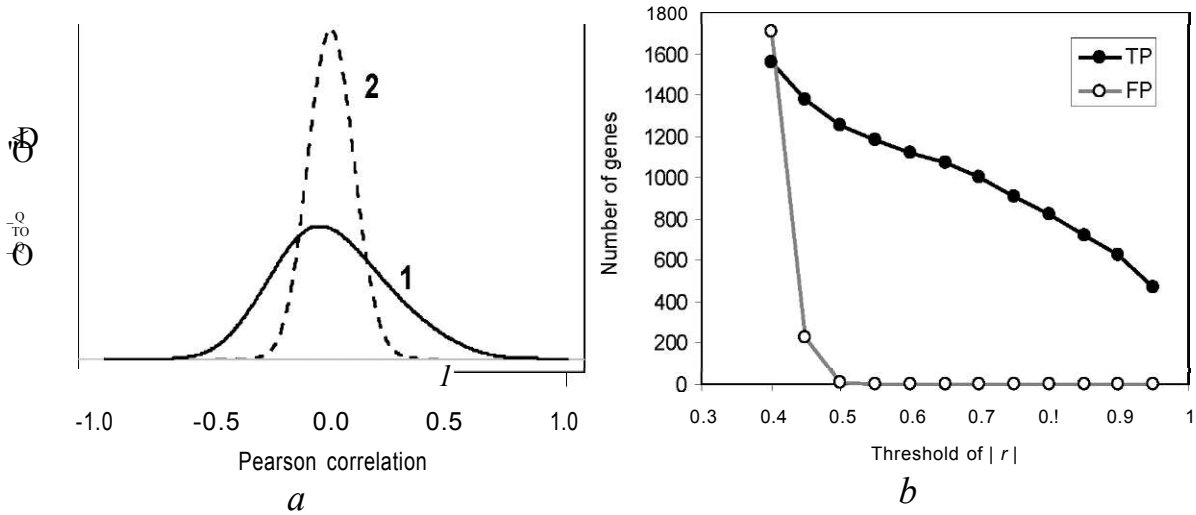

Fig. 1a- distributions of correlation coefficients for experimental data (curve 1) and randomized data (curve 2);
1b- behavior of true (TP) and false positives (FP) with respect to \r\ threshold

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.

Validation using STRING service. Finally, CoExpress was validated using STRING 8.2 [6] - 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 connectivity of the inferred network was 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.

CONCLUSION

As it was shown, both versions of the tool are able to work with big data sets. The validation using simulated data showed the precision and robustness of the approach. The current version of CoExpress and its multi-thread Linux version are freely available for downloading from www.sablab.net/coexpress. The multi-thread module is distributed together with its source code under the GPL, which allows modifying, recompiling and running it under various OS.

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

We would like to thank Arnaud Muller, Francisco Azuaje, Isabel Nepomuceno for their useful comments about CoExpress and its validation.

References

1. *Margolin, A. A.* ARACNE : an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context / A. A. Margolin et al. // BMC Bioinformatics. 2006. Vol. 7. Sup. 1. P. S7.
2. *Jupiter, D.* STARNET 2: a web-based tool for accelerating discovery of gene regulatory networks using microarray co-expression data / D. Jupiter, H. Chen, V. VanBuren // BMC Bioinformatics. 2009. Vol. 10. P. 332.
3. Internet-adress: http://www.ebi.ac.uk/microarray-as/ae/.
4. *Irizarry, R A.* Summaries of Affymetrix GeneChip probe level data / R. A. Irizarry et al. // Nucleic Acids Res. 2003. Vol. 31. P. 15.
5. *Lee, H. K.* Coexpression analysis of human genes across many microarray data sets / H. K. Lee, et al. // Genome Res. 2004. Vol. 14. P. 1085-94.
6. Internet-adress: http://string.embl.de.