Computer tool for gene expression data processing and correlation analysis

Anastasia M. Spitsina,

Novosibirsk State University, 630090, Novosibirks, 2 Pirogova Str., anastasia.spitsina@gmail.com

Natalya N. Podkolodnaya, Vadim M. Efimov, Vladimir N. Babenko, Yuriy L. Orlov

The Institute of Cytology and Genetics, 630090, Novosibirsk, Lavrentyeva 10,

{nata,efimov,bob,orlov}@bionet.nsc.ru

In recent years various databases (BioGPS (http://biogps.org/), GEO NCBI (http://www.ncbi.nlm.nih.gov/geo/)) has accumulated a large body of experimental data obtained using DNA microarrays. Such data have great importance for applied medical research as well as for fundamental statistical analysis of gene expression in mammalian organisms across tissues and cell types. Thus, development of high-throughput computer tools for gene expression analysis should meet fast growth of available gene expression data [1,2]. In this work we used Affymetrix expression microarray data for human, mouse *Mus musclus* and rat *Rattus norvegicus*.

Affymetrix U133 microarray set — a suite consisting of two arrays. It contains about 45 000 sets of samples representing more than 39,000 transcripts obtained from the approximately 33,000 genes annotated in the human genome. Though sequencing technologies (RNA-seq) provide new data on gene expression, large collection of human clinical data keep importance of Affymetrix data analysis, especially in many tissues and organs, including brain (BioGPS).

Technical complexity of the work with data in this form consists in large volume - 45 000 lines and 80 columns, there are no a uniform (some parts of the table - text names and other parts - numerical data), and each line is a separate element with a set of parameters (e.g. gene probe, its expression and its identifier) [1]. Some genes have several rows corresponding to them - sample- repeats, which make analysis difficult. So the aim of this work was to develop and improve own program for data processing and analysis of gene expression. Software package must meet the following requirements:

— storing a database in memory for working it consistently, without reading them each time from a file. This will save time access to the database and will allow to work with the changed data (for example, after the unification of multiple databases);

— simple and intuitive interface.

In addition, the software package must be compatible with previously developed in the ICG SB RAS tools [3] and computational modules (such as JACOBI 4).

Developed earlier C ++ software package has options for statistical analysis and data pre-processing, such as the calculation and construction of tissue-specific profiles, filtering genes according to available information on the location on the chromosomes [2].

Also in our work we implemented a program for analysis of expression correlations (across samples and tissues), which will simplify the identification of structural features of gene networks. The program performs the analysis of the expression of human genes, the study of the quality of the measurement signal on the microarray, analysis of tissue-specific gene expression, visualization of gene communication using correlation coefficients (linear Pearson and rank Spearman coefficients) with a simple and intuitive interface. Visualization of the relationships between genes is constructed in the form of a gene network (using a script developed in Java language). The program is applicable to not only the microarray data, but also RNA-seq data (SRA (http://www.ncbi.nlm.nih.gov/sra), ArrayExpress (http://www.ebi.ac.uk/arrayexpress/).

A comparative analysis of expression data for human genes whose expression is increased in the brain tissue was done by the software developed. The expression patterns of pairs of transcripts, co-localized in the genome were analyzed [3]. Using RefSeq and BioGPS databases genes with high expression were identified, gene networks of interactions of these genes were constructed, correlation matrix and tissue specificity profiles of sample data were calculated. The structural features of genes with high expression were identified (e.g., number of exons, relationship with alternative splicing).

We studied correlation of gene expression contained in gene networks of circadian rhythm and cholesterol regulation, and well as genes responsible for aggressive behavior in mice (by previously annotated data). Also we made comparative analysis of the obtained results using program and information about co-expression of genes in the STRING database (http://string-db.org/). The effectiveness of gene networks reconstruction from the examined samples was analyzed. We confirmed many annotated gene network structures for curated gene sets. Overall the software developed extend analysis of the packages developed earlier [3,4], more flexible in data formats and meet demands for co-expression analysis.

This tool will be integrated with the software and algorithmic complex for multivariate analysis of microarray data JACOBI 4 designed for similar data processing to the same algorithm [5]. JACOBI Package 4 is a set of programs for multivariate analysis with open source, which is equally suitable for use by users with any experience with PCs. Project JACOBI 4 develops to support new technology (which was developed in the ICG SB RAS) of the search of candidate genes in the gene networks, and to expand its functionality requires integration tools for Affymetrix data processing.

The work has been supported by RFBR (№14-04-01906) and ICG SB RAS budget project VI.61.1.2. Software was tested at Siberian Supercomputer Center SB RAS (SSCC).

References:

1. A.M. Spitsina, V.M. Efimov, V.N. Babenko, Y.L. Orlov (2014) Computer analysis of human gene expression data using biogps database of microarray Affymetrix U133 // In: Proceedings of the Ninth International Conference on Bioinformatics of Genome Regulation and Structure\Systems Biology (BGRS\SB-2014). Novosibirsk, Russia, June 23–28, 2014. Publishing House SB RAS. P. 154.

2. A.M. Spitsina et al. (2015) Supercomputer analysis of genomics and transcriptomics data revealed by high-throughput DNA sequencing, *Program systems: theory and applications* (http://psta.psiras.ru/) (In Russian) (*In press*).

3. Y.L. Orlov et al. (2012) ICGenomics: Program complex for analysis of symbol genomic sequences, *Vavilov Journal of Genetics and Breeding*, **16:**732–741. (In Russian).

4. Y.L. Orlov et al. (2007) Quality assessment of the Affymetrix U133A&B probesets by target sequence mapping and expression data analysis, *In Silico Biol.*, **7**(3):241-60.

5. D.A. Polunin, I. A. Shtaiger, V.M. Efimov (2014) Development of software system JACOBI 4 for multivariate analysis of microarray data, *Vestnik NSU. Information Technology*, **V12, № 2:**90–98. (In Russian).