White paper

zRMicroArray: a flexible environment for analysis of mRNA and miRNA microarray data

Zlatomir Zlatev[1], Ivan Ivanov[2]

[1]Information Systems Department, Sofia University, Bulgaria
[2]Department of Physiology and Pharmacology, Texas A&M University

Corresponding author: Zlatomir Zlatev, z.zlatev.z@gmail.com

June 2010

Abstract

Background

cDNA high density microarrays have become a widely accepted tool for simultaneously interrogating gene expression. Gene expression analysis holds the promise to help in understanding and modeling of complex cell regulatory mechanisms. Ultimately, the goal is to develop sets of molecular biomarkers for diagnostic purposes or for designing of therapeutic intervention strategies. At the same time, the currently available high density microarray platforms require specialized approaches for data preprocessing, normalization, transformation, and subsequent analyses. It has become apparent that scientists from different fields of life sciences have a need for software packages that not only offer such functionalities but are also user friendly, flexible, freely available and open to further development.

Results

zRMicroArray software implements all of the functionalities, needed for preprocessing and statistically analyzing single-color cDNA microarray data. It allows for parallel execution of the implemented algorithms which makes it appropriate for multiprocessor and multi-core computer systems. The Graphical User Interface (GUI) is easy and intuitive to use, which makes it particularly suitable for users,

who do not have specialized programming or statistical background. The software is designed on Microsoft Windows platform, and requires Microsoft Windows 98 or newer operating system.

Conclusions

zRMicroArray software has a modular structure which allows for further development, e.g. data sources from two-color microarray platforms, next generation sequencing technologies and implementation of additional statistical methods for data analyses.

Background

The cDNA microarray technology has transformed both the basic life sciences and the translational biomedical research for the past 10-15 years [1]. The modern high-throughput parallel data collection platforms facilitate disease prognosis and classification, identification of gene function and selection of drug targets, and advance our understanding of the related cellular processes and pathways. At the same time, high dimensionality and size (hundreds or more gigabytes) of the collected data sets turns data preprocessing, normalization, transformation and statistical analyses into challenging tasks. The scientists who design and perform the experiments in the wet laboratories often feel intimidated by the complexity of these issues, and attempt to outsource the data analyses or purchase expensive commercial

software packages. This not only creates a bottleneck in the pipeline for data processing and analysis but also forces the laboratories to often "cut corners" and settle for processing smaller number of samples in order to reduce their costs. Thus, there is a need for development of freely available software packages that streamline the raw data processing and initial statistical analysis. However, the scientists from the life or biomedical sciences usually do not have mathematical, statistical or programming background, and this determines a set of important requirements that need to be satisfied by such software packages.

- Graphical user interface (GUI) which is intuitive and easy to use.
- Parallel execution of the implemented algorithms which allows for the implementation on currently available multicore or multiprocessor desktop computers.
- Executables that are available free of charge and easy to install.
- Modular structure which allows for further and tailored to the specific laboratory needs development of the package.

This article describes the package ZRMicraArray which successfully addresses the aforementioned requirements and in its present version provides the following functionalities:

- Data import
- Data grouping, according to experimental treatments.

- Initial analysis of data quality and statistical properties.
- Data filtration and normalization.
- Statistical analysis of the data in order to find differentially expressed genes.
- Results export and comparisons with previous analyses.

ZRMicroArray was developed as part of the requirements for the M.S. degree of the author.

Implementation

zRMicroArray software is written in Microsoft Visual FoxPro (VFP) programming language [5] and use relational databases to store its data. Some of the computations, especially statistical ones are written in R, [6]. The software uses R-(D)COM, [7], to access the R calculation engine. Two C DLLs [4] are used from the VFP GUI and from R to speed up some time consuming routines. Computations are distributed for parallel execution according to the number of processors or cores available, Figure 2. The data is stored as VFP relational database. CSV files are created to pass large data blocks to the R functions. JPG files are created for graphical data representations. Data is exported in CSV, XLS or DBF file formats.

Results and Discussion

The workflow of the data processing implemented in ZRMicroarray is outlined in **Figure 1.**

Project

For each study, the user creates a project. The project serves as a container for the data and its analysis. Data manipulations, grouping and analyses are automatically saved to the project after each processing step is completed.

Data import

A blank project is created and data is imported into it. The acceptable import data formats are either .txt or .xls files. Other data sources should be manually converted to one of the supported formats and then imported to the project. After data from the experiment is imported the user has the option to examine various data distributions and statistics.

Grouping

Data grouping is organized into two levels: *factors* and *factor values.* The factors and their values represent the experimental design used to produce the data. Each factor could have two or more values and there should be at least one factor presented in a given project. The combination of factors and their values groups the microarrays from an experiment. The user has the option to examine and finalize the grouping.

Initial analysis

The initial data analysis calculates important statistics and provides graphical representations that illustrate the calculations. The imported gene expression values are tested for normality of their distributions in each one of the experimental groups – an important assumption for the subsequent statistical analysis. Furthermore, the user has the option for data imputation which generates gene expression values which could be used for replacing missing data. Such expression values are generated only for genes, that are close to normally distributed within the respective groups, and in those cases, values are generated from the group distribution, using the *rnorm()* R function. The normality test implemented in the current version is the Shapiro – Wilk normality test [2]. The user could specify either to use or not to use the generated values.

Data filtering and transformations

At this stage of data analysis, the user has the option to choose which genes are going to be subjected to statistical analysis. The decision is based on the information and statistics about the data obtained from the preceding steps in data processing. All the expression values that correspond to quality control spots on the microarrays are also removed. zRMicroArray current version features commonly accepted data transformations: median, quantile and logarithmic. The user could apply as many filters, transformations, and combinations of those as desired. The

outcomes of those filters and transformations are all saved within the project and are available during the subsequent statistical analysis.

Statistical analysis

The statistical analysis workflow is illustrated on Error! Reference source not found.. First, the program creates tasks for parallel execution, according to the number of available processors or cores. Each task is designed to perform the designated statistical analyses on a portion of the genes as follows:

- Each gene distribution within a group is tested for normality using Shapiro - Wilk normality test.
- The genes which pass the normality test are subjected to Analysis of Variance (ANOVA) and respective p-values are computed.
- Data is regrouped according to the factors and their values. Step 6.1 and 6.2 are repeated till all factors interactions, and base effects are tested.
- The results from all of the parallel tasks are summarized.
- *False discovery rate* (FDR) correction is applied, based on the computed in step 6.2 p-values [3].

Other functionalities currently implemented

- Normalized data distributions (overall and by genes)
- Normalized data export
- Fold change computations and export

- Results agreement table
- Results comparisons
- Differentially expressed genes export

Conclusions

zRMicroArray software has a modular structure which allows for further development, e.g. data sources from two-color microarray platforms, next generation sequencing technologies and implementation of additional statistical methods for data analyses. The software has user friendly GUI suitable for users, who do not have specialized programming or statistical background and is vary fast due to the parallel distributed tasks and coding stile. A new version is under development which will allow the user to interactively modify and create data normalization and statistical analysis code.

Availability and requirements

Project name: zRMicroArray

Operating system(s): Microsoft Windows 98 or later

Programming language: Visual Fox Pro, R, C

Other requirements: R 2.10.1 or higher, rscproxy 1.3 or higher (R package), preprocessCore 1.8 or higher (R package), fdrtool 1.2.6 or higher (R package), R Scilab DCOM 3.0 server or higher.

Authors contributions

Zlatomir Zlatev: design of the software, methods and algorithms selection, the GUI design, the manuscript, and all of the coding.

Ivan Ivanov: advice and guidance, protocols to be followed, manuscript correction, proof reading.

Acknowledgements

The author thanks for the advice and guidance received from Prof. Ivan Ivanov, Department of Physiology and Pharmacology, Texas A&M University, Prof. Antony Popov, Department of Information Technologies, Faculty of Mathematics and Informatics, University of Sofia, Bulgaria, and the team of the Quantitative Biology Core, The Center for Translational Environmental Health Research, Texas A&M University.

References

1. Hardiman, G.: **Microarray platforms – comparisons and contrasts.** *Pharmacogenomics* 2004, **5**(5): 487 – 502.

2. Shapiro, S. S. and Wilk, M. B.: **An analysis of variance test for normality (complete samples).** *Biometrika* 1965, **52**(3, 4): 591–611.
3. Korbinian Strimmer: **A unified approach to false discovery rate estimation** 2008.
4. Sigal Blay: **Calling C code from R, Dept. of Statistics and Actuarial Science.** Simon Fraser University, October 2004.
5. *Microsoft Visual FoxPro Programmer's Guide*, Microsoft Press, 1998
6. The R Project for Statistical Computing.
7. Thomas Baier, R/Scilab (D)COM Server.

Figures

Additional files

ReadMe.txt – zRMicroArray installation instructions

zRMicroArray_Distribute.exe – zRMicroArray software package archive

PresentationEN_new.mov – zRMicroArray software package basic overview video