MICROARRAY DATA ANALYSIS PLUGIN FOR BIOUML

I.N. Kiselev[1,], A.A. Shadrin[1], Y.V. Kondrakhin[1,2], F.A. Kolpakov[1,2]*

[1]Technological Institute of Digital Techniques SB RAS, Novosibirsk, Russia

[2]Institute of Systems Biology, Novosibirsk, Russia

*Corresponding author: axec@developmentontheedge.com

Motivation and Aim

Huge amount of microarray data has become available during the past decade. However, dealing with large tables of information and extracting useful data from them is complicated. The aim of this work was to develop software to automate and simplify the solution of problems such as finding up- and down-regulated and coexpressing genes, discovering gene expression differences between microarrays data sets, analysis of gene expression patterns and finding putative regulators.

BioUML

BioUML (http://www.biouml.org/) is open source Java framework for formal description and modeling of biological systems. It has plugin-based architecture, so that adding new plugins to it is natural and enhances abilities for analysis.

Implemented analysis tools:

Possible scenario of microarray data processing using our plugin presented in **fig. 1**. Next methods were implemented for this purpose:

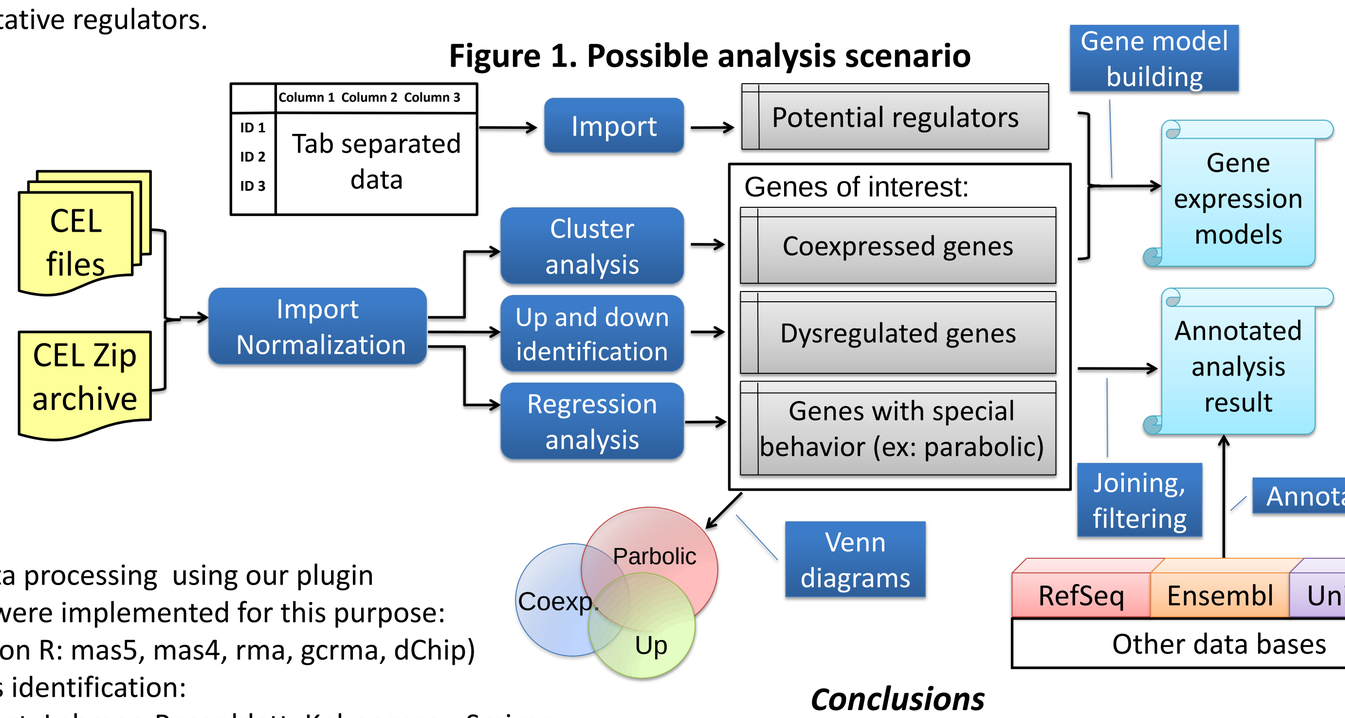
- CEL files normalization (based on R: mas5, mas4, rma, gcrma, dChip)
- Up- and down-regulated genes identification:
  - Student t-test, Wilkokson test, Lehman-Rosenblatt, Kolmogorov-Smirnov tests.
  - Hypergeometric analysis and meta-analysis [1].
- Other statistical tools and supplement:
  - Polynomial regression.
  - Cluster analysis (Chinese Restaurant algorithm, R) [2].
  - Joining datasets (inner, outer, left, right and symmetric difference).
  - Venn diagrams building.
  - Annotation through remote databases such as Ensembl, Unigene, etc.
- Gene expression model building:
  - Linear model with time delay.
  - Nonlinear model [3].
  - Stochastic Differential Equation (SDE) model [4].

An interface general for all methods is presented in **fig. 2** and **fig. 3**.

Up and down regulated genes identification:

For this purpose we implemented classic methods and method specially designed for statistical analysis of microarray data Hypergeometric analysis [1]. Meta-analysis based on hypergeometric test is also included in plugin. The general purpose of meta-analysis is to obtain reliable results by processing integrated data sets derived from independent experimental studies. All methods return score = lg(P-value). All methods also support FDR (False Discovery Rate) calculating which is based on input dataset stochastic permutations.

Figure 1. Possible analysis scenario



Other statistical tools and supplement

For time-course gene sets analyzing we included polynomial regression. To find similarly expressed genes we ported to Java Chinese Restaurant Cluster algorithm [2] capable of finding not only synexpressed genes but also genes with shifted or inverted expression profiles and classic K-means cluster algorithm based on R. For user convenience we included data-manipulating tools such as joining datasets, building Venn diagrams and annotating data sets with gene titles, descriptions, functional roles and so on from databases included in BioUML.

Gene expression model building

After finding up or down regulated genes another task is to find possible regulators which affect target gene. To accomplish this task we implemented methods for gene model expression building i.e. choosing regulators from pool and estimating parameters considering different regulation functions :

- $\frac{dY}{dt} = a + b * X(t + \Delta)$
- $\frac{dY}{dt} = a + b * f(X(t)) - c * Y$
- $\frac{dY}{dt} = a + b * f(X(t)) + d * Random$

Where a, b, c, d – estimated parameters

Conclusions

Java-based plugin was developed and integrated into both BioUML workbench and web editions. Analysis methods can be started using interface or console with java-script code. Software allows complete chain of microarray tools from import and normalization to finding regulated genes and possible regulators. It was successfully tested and is now used for data analysis within EU grants FP6 №037590 “Net2Drug” and FP7 №202272 “LipidomicNet”. Plugin architecture allows simple extension with new methods of analysis for further development.

Availability

Software is freely available as a part of BioUML on websites:

- http://www.biouml.org/ (workbench version).
- http://server.biouml.org/bioumlweb/index.html# (web version).

Acknowledgements

This work was supported by EU grants FP6 №037590 “Net2Drug”, FP7 №202272 “LipidomicNet”, and Integrated Project SBRAS 17.

References

1. Y.V.Kondrakhin, R.N.Sharipov, A.E.Kel, F.A.Kolpakov. (2008) Identification of Differentially Expressed Genes by Meta-Analysis of Microarray Data on Breast Cancer, In Silico Biology, **8**: 383-411.
2. Z.S.Qin. (2006) Clustering microarray gene expression data using weighted Chinese restaurant process. Bioinformatics, 22:1988-1997.
3. K.Chen et al. (2005) A stochastic differential equation model for quantifying transcriptional regulatory network in Saccharomyces cerevisiae. Bioinformatics, **21**: 2883-1890.
4. J.Vohradsky, T.T.Vu. (2007) Nonlinear differential equation model for quantification of transcriptional regulation applied to microarray data of Saccharomyces cerevisiae. Nucleic Acids Research, **35**: 279-287.

Figure 2. User interface of microarray analysis plugin for BioUML workbench with running Hypergeometric analysis.

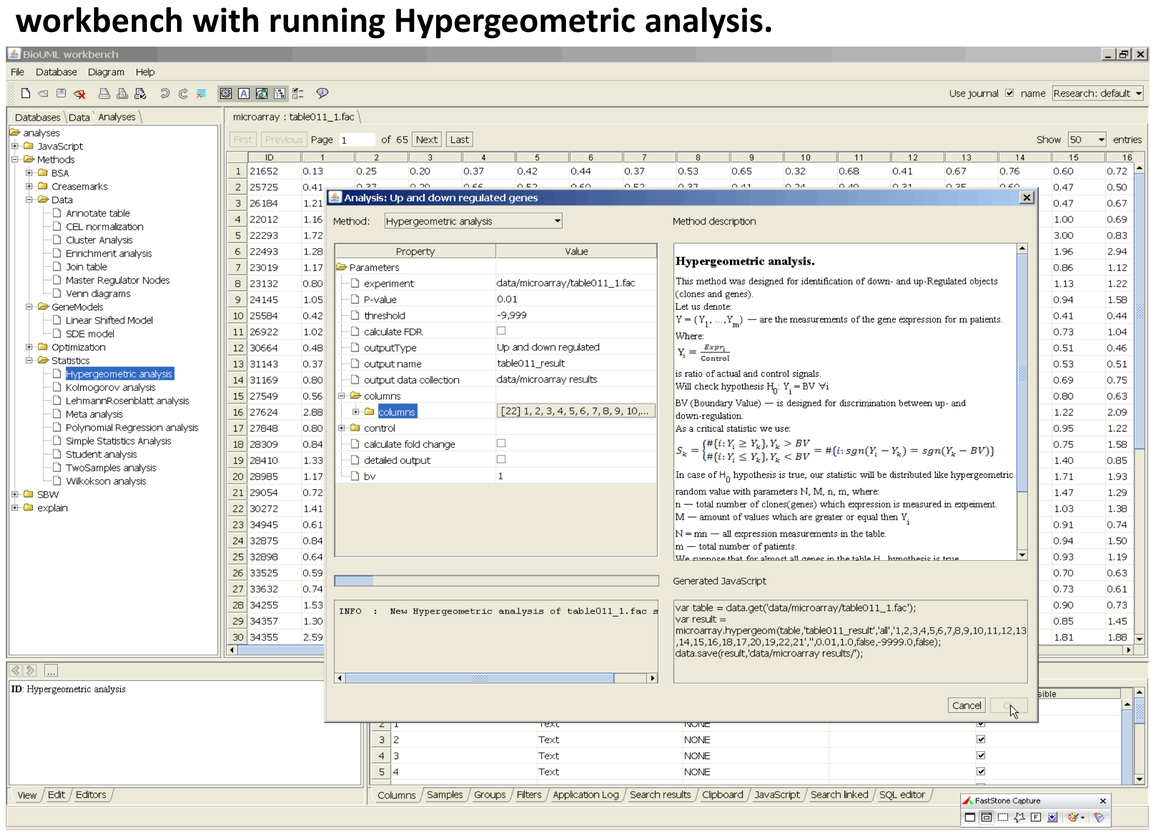


Figure 3. User interface of microarray analysis plugin for BioUML web edition with stochastic gene model building and result graphics.