BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 19 2007, pages 2631–2632
doi:10.1093/bioinformatics/btm333

Gene expression

Eu.Gene Analyzer a tool for integrating gene expression data with pathway databases

Duccio Cavalieri[1,*], Cinzia Castagnini[1,†], Simona Toti[1,2,†], Karolina Maciag[1], Thomas Kelder[3], Luca Gambineri[4], Samuele Angioli[4] and Piero Dolara[1]

[1]Department of Preclinical and Clinical Pharmacology, [2]Department of Statistics, University of Florence, Florence, Italy, [3]BIGCAT Bioinformatics, University of Maastricht, Maastricht, The Netherlands and [4]Inspect.it, Capolona (Arezzo), Italy

Received on April 19, 2007; revised on June 15, 2007; accepted on June 17, 2007

Advance Access publication June 28, 2007

Associate Editor: David Rocke

ABSTRACT

Motivation: Eu.Gene Analyzer is an easy-to-use, stand-alone application that allows rapid and powerful microarray data analysis in the context of biological pathways. Its intuitive graphical user interface makes it an easy and flexible tool, even for the first-time user. Eu.Gene supports a variety of array platforms, organisms and pathway ontologies, transparently deals with multiple nomenclature systems and seamlessly integrates data from different sources. Two different statistical methods, the Fisher Exact Test and the Gene Set Enrichment Analysis (GSEA), are implemented to identify biological pathways transcriptionally affected under experimental conditions. A suite of tools is offered to define, visualize and share custom non-redundant pathway sets.

In conclusion, Eu.Gene Analyzer is a new software application that takes advantage of information from multiple pathway databases to build a comprehensive interpretation of experimental results in a simple, intuitive environment.

Availability: Download of Eu.Gene Analyzer Java version is available free of charge for academic users. Please visit the web page: http://www.ducciocavalieri.org/bio/Eugene.htm

Contact: duccio.cavalieri@unifi.it

Supplementary information: http://www.ducciocavalieri.org/bio/Eugene/Suppl_Inf

1 INTRODUCTION

Pathway-based microarray analysis methods look for patterns of gene expression variation in any predefined set of genes. While the effect of each individual gene can be subtle, a coordinated change among many gene products can produce potent biological effects. Eu.Gene explores this type of multi-gene effects.

2 FEATURES

Many collections of pathways and other meaningful gene sets are publically available. Eu.Gene loads and stores pathway definitions from the latest update of multiple public databases, KEGG (Kanehisa *et al*., 2006), Reactome (Joshi-Tope *et al*., 2005) and GenMAPP (Dahlquist *et al*., 2002), within the stand-alone application.

One obstacle to concurrent use of different pathway databases is their use of diverse types of identifiers. That is, different databases use various types of identifiers, or primary names, to refer to the same entities. The same problem appears among different microarray platforms. Eu.Gene resolves the nomenclature problems by handling conversion among various annotation and nomenclature systems transparently to the user (see Table 1a, Supplementary Material). To do so, a built-in conversion map converts all identifiers to a common format, Ensembl Gene and Transcript IDs (Hubbard *et al*., 2007). Thus, Eu.Gene users can load data from different array platforms (see Tables 1d and 2b, Supplementary Material) and different pathway sources using the original nomenclature. The entire conversion map is also available directly, as a Conversion Tool that easily converts user-generated lists of gene names among the different nomenclature systems. In some cases, the size of the conversion map is very huge and the analysis tasks become memory demanding (see Table 2a, Supplementary Material).

The available pathways can be browsed to select which to include in a custom pathway set. Browsing based on pathway source database is easy and convenient. Alternately, Eu.Gene provides a search function to identify and select pathways involving specific key words in the name. Thus, a pathway set may represent the space of biological categories of interest and may conveniently be stored for future use and shared with others.

The use of several pathway sources often results in the occurrence of redundant pathways with a high degree of overlap (Cary *et al*., 2005). The 'Entity Affinity Filter' tool helps to navigate the territory and create robust, non-redundant pathway sets. Users select threshold values for maximum tolerated overlap among pathways to include in the selected pathway set. A dendrogram visualization tool guides

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the second and third authors should be regarded as joint First Authors.

the user in the selection of redundancy threshold values. The dendrogram illustrates hierarchical trees of the loaded pathways, ordered according to the mutual overlap. To minimize overlap among pathways, users can employ the ‘SuperSetMode’ to collapse all available pathways into a super-pathway that can be stored and used for further analysis.

Pathways frequently include a number of genes not represented in the microarray. User can exclude from a pathway set the pathways which do not meet a threshold value of percentage of which are present on the microarray.

Eu.Gene Analyzer implements two different statistical methods to evaluate which pathways are most affected by differences in gene expression observed in a functional genomic experiment: the one-tailed Fisher Exact Test (FET) (Grosu *et al*., 2002) and Gene Set Enrichment Analysis (GSEA) (Subramanian *et al*., 2005). The FET requires a user-defined threshold for gene expression to identify pathways enriched in both over- and under-expressed genes. The method GSEA does not use thresholds and is suited to the detection of coordinated, modest changes in transcriptional activity of many genes in a pathway. The output provided by the program is a text file with the relevant analytical information. For the FET, output values include signed and unsigned *P*-values. The *P*-value is a measure of the significance of a pathway’s enrichment in transcriptionally altered genes; its sign reflects the relative number of up-regulated genes with respect to down-regulated genes for each pathway. GSEA analysis output files contains enrichment scores (ES) and empirical *P*-values for each pathway. The output is exportable in HTML or MS-Excel format.

3 CONCLUSION

The integration of pathway browsing and statistical tools for microarray analysis in a single, stand-alone application enables the user to perform cycles of analysis and fine-tuning to construct a biologically meaningful pathway set. Results from multiple experiments can be used for further analysis, such as clustering. The pathwayset and analysis criteria used for an experimental group can be saved as a project and used to analyze additional data. This feature allows the comparison between multiple experiments and keeps track of the analysis performed.

The significant improvement provided by Eu.Gene is represented by the power and flexibility offered to biologist in selecting and understanding the set of pathways analyzed and by the possibility to apply and compare the results from different statistical methods. The number of pathways, genomes, platform and statistical methods implemented in Eu.Gene Analyzer can be easily expanded in the future versions of the program.

ACKNOWLEDGEMENTS

This work was funded by the European Networks of Excellence NuGO (FOOD-2003-5063-60), Dc Thera (LSHB-CT-2004-512074), by a year 2007 grant from Cassa di Risparmio di Firenze at the ‘Noi per Voi’ association and by a grant from AIRC Milan, Italy. We thank Prof. Bruce Conklin, Dr Chris Evelo and Dr Andre Boorsma for the useful discussions.

Conflict of Interest: none declared.

REFERENCES

Cary,M.P. *et al*. (2005) Pathway information for systems biology. *FEBS Lett*., **579**, 1815–1820.

Dahlquist,K.D. *et al*. (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet*., **31**, 19–20.

Grosu,P. *et al*. (2002) Pathway processor: a tool for integrating whole-genome expression results into metabolic networks. *Genome Res*., **12**, 1121–1126.

Hubbard,T.J.P. *et al*. (2007) Ensembl 2007. *Nucleic Acids Res*., **35**, D610–D617.

Joshi-Tope,G. *et al*. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*., **33**, D428–D432.

Kanehisa,M. *et al*. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*., **34**, D354–D357.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.