

A bioinformatic workbench for the analysis of microarray data based on open source software and public databases.

Margherita Mutarelli^{1,2}, Giovanni Colonna² and Angelo Facchiano^{1,2}

¹ Istituto di Scienze dell'Alimentazione CNR, via Roma 52A/C, 83100 Avellino, Italy

² Centro di Ricerca Interdipartimentale di Scienze Computazionali e Biotecnologiche, Seconda Università di Napoli, via de Crecchio, 80138 Napoli Italy

DNA microarray analysis [1, 9] has become the most widely used technique for generating genome-wide expression profiles and represents the first practical application that allows the results of whole-genome sequencing projects to be used effectively to address biologically relevant questions. As the microarray technology matures, our ability to generate large numbers of high-quality assays is quickly and efficiently accelerating, so one of the greatest challenges in working with microarrays is collecting, managing, and analyzing data. In fact, one can no longer expect findings from microarray studies to be accepted without adequate replication and sampling to assure that the results reflect the underlying biological processes [4]. Studies that previously would have been done with a small number of hybridizations now involve tens or hundreds of assays, and the challenge is moving from generating data to collecting, managing, and analyzing the data to identify statistically and biologically significant patterns of gene expression. Furthermore, the identification and estimation of systematic errors and sources of variation in the data become fundamental to obtain a sensible interpretation of the results [11]. Achieving these goals requires each laboratory's set up of a comprehensive system of tools for data management and analysis. The rapid growth of this field has prompted the increasing number of commercial solutions to various aspects of this problem, each of which has particular strengths and weaknesses. However, many of the most innovative approaches to data analysis have been developed in academic laboratories practicing the technology [6, 10], and a long time often elapses between when these approaches are first described to when they appear in commercial products. In response to both the rapid growth of the field and the need for affordable, state-of-the-art tools, several efforts have recently sought to develop open source software for expression analysis, with the availability of both the program source code and well-defined standards for adding functionality to the software [5]. Projects that aim to create comprehensive yet flexible systems for microarray data analysis are gaining an increasing presence in the field.

The aim of our work is to create a workbench of bioinformatic tools by collecting the available software and integrating it when needed by creating user-friendly interfaces and tools for the easy exchange of data among the programs. Our project is based on the utilization of open source software and public libraries, being convinced that the free availability of the program source code and the availability of a clear, well-defined application program interface (API) allows developers to integrate the software with other systems and to add new functionality (see the licensing agreements that can be found at <http://www.opensource.org> and <http://www.gnu.org> for an explanation of the open source philosophy).

This approach gives the benefits of creating a community resource that can advance the field, but also there are several advantages to an open source approach to software development in a scientific environment, including:

- full access to the algorithms and their implementation, which allows users to understand what they are doing when they run a particular analysis;
- the ability to fix bugs and extend and improve the supplied software;
- encouraging good scientific computing and statistical practice by providing appropriate tools, instruction, and documentation;

- providing a workbench of tools that allow researchers to explore and expand the methods used to analyze biological data;
- ensuring that the international scientific community is the owner of the software tools needed to carry out research;
- promoting reproducible research by providing open and accessible tools to carry out that research (reproducible research as distinct from independent verification).

It is our hope that by helping to create an environment that encourages scientists to create new applications and to make them accessible to laboratory biologists, we can create the same sort of community-based effort to drive the development of software and, in doing so, advance the general state of the art in functional genomics.

In more details, our workbench is presently based on some of the most widely used and comprehensive systems for statistical and functional analysis:

- the statistical analysis tools written in R through the Bioconductor project (<http://www.bioconductor.org>);
- the Java[®]-based TM4 software system available from The Institute for Genomic Research (<http://www.tigr.org/software>);
- the BRB Array Tools package from the Biometric Research Branch (BRB) of the NCI (<http://linus.nci.nih.gov/BRB-ArrayTools.html>);
- GenMAPP and related software, developed at the Gladstone Institutes of San Francisco University (<http://www.genmapp.org>).

The software revised in our work represent different approaches to the same problem, and each has its advantages and disadvantages. As an example, Bioconductor builds on the existing power of the R statistical analysis tool development community (<http://cran.r-project.org>) and allows for the rapid implementation and dissemination of new methods. However, the R command-line environment and language complexity can be discouraging to first-time users. Moreover, several efforts are underway to simplify and enhance the user interface. TM4 gives users a graphical interface that is easy to navigate and the architecture provides great flexibility for development. However, implementation of new statistical tools requires the creation of new analysis libraries and users have to install new software releases, thus it is not easily customizable or expandable. The BRB Array Tools package is not exactly open source, but it is freely available for academics and represents a compromise between the two: in fact it has the possibility to use powerful and sophisticated R analysis by using the most user-friendly interface that is Excel, widely used by biologists without requiring the user to interact with R programming environment. GenMAPP [3], on the other side, is a visualization tool that allows to view expression data from different experiments on the same group of genes and look for global changes, thus it helps the final step that is biological interpretation of data.

The integration of data and their interpretation can also be helped by using public databases and flexible query systems. We have implemented in our web server (<http://bioinformatica.isa.cnr.it>) a mirror site of the GeneCards databases [8] and a local installation of the SRS query system [7] with a growing number of databases including EMBL, Unigene, LocusLink, GeneOntology and GeneOntologyAnnotations (GO and GOA), Pathway, OMIM.

Most of the described projects have in common the availability of their software source code, which allows users to modify the program to both meet the local needs in each laboratory and to continue to expand its functionality. The recent establishment of the MAGE-ML standard [2] for representing microarray data promises to provide a means by which these and other systems can communicate and exchange data and results. One might hope that the software development efforts described here and other projects will converge and that their integration will result in set of tools that has the advantages of all of these without their limitations. The possibility that this will happen

depends on open access to the source code, which will allow the community to leverage our collective expertise to the benefit of everyone working in gene expression analysis and related areas.

Acknowledgements

This research is partially supported by FIRB Post-genomica (Grant RBNE0157EH_003). M. Mutarelli is a PhD student of Dottorato di Biologia Computazionale (XVIII ciclo), Seconda Università degli Studi di Napoli.

Bibliography

1. AA. VV. The Chipping Forecast II *Nat. Genet. Suppl.* **32**:461-552 (2002).
2. Brazma, A., *et al.* Minimum information about a microarray experiment (MIAME) – towards standards for microarray data. *Nature Genet.* **29**:365-371 (2001).
3. Dahlquist, K.D., Salomonis, N., Vranizan, K., Lawlor, S.C. & Conklin, B.R. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet.* **31**:19-20 (2002).
4. Dudoit, S., Yang, Y.H., Callow, M.J. & Speed T. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Technical Report, Statistical Dept. of California, Berkeley* (2000).
5. Dudoit, S., Gentleman, R.C. & Quackenbush, J. Open source software for the analysis of microarray data. *Biotechniques* **34**:S45-S51 (2003)
6. Eisen, M., Spellman, P.T., Botstein, D. & Brown, P.O. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* **95**:14863-14867 (1998).
7. Etzold, T., Ulyanov, A. & Argos, P. SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.* **266**:114-128 (1996).
8. Rebhan, M., Chalifa-Caspi, V., Prilusky, J., Lancet, D. GeneCards: encyclopedia for genes, proteins and diseases. *Weizmann Institute of Science, Bioinformatics Unit and Genome Center* (1997).
9. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**:467-470 (1995).
10. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci.* **98**:5116-5121 (2001).
11. Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J. & Speed, T.P. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* **30**:e15 (2002).