AMDA, an automated pipeline for Affymetrix data analysis: an update

Kapetis D(1), Vitulli F(1), Tuana G(1), Pelizzola M(2), Ricciardi-Castagnoli P(3),
Maria F(4) and Zolezzi F(1)

(1) Genopolis Consurtium, Department of Biotechnology and Biosciences,
University of Milano-Bicocca, Piazza della Scienza 2,20126 Milan, Italy
(2) Department of Epidemiology and Public Health, Yale University School of Medicine,
New Haven, Connecticut 06520, USA
(3) Singapore Immunology Network (SIgN), Biomedical Sciences Institutes,
Agency for Science, Technology and Research (A*STAR),
8A Biomedical Grove, IMMUNOS,138648,Singapore
(4) Department of Biotechnology and Biosciences, University of Milano-Bicocca,
Piazza della Scienza 2,20126 Milan, Italy

Motivation

Expression profiling using microarrays has become a widely used method for the study of gene-expression patterns. We published in 2006, the Automated Microarray Data Analysis (AMDA, version 2.3.5) pipeline [1] that was developed under R version 2.4 and Bioconductor 1.9 release. In this work we present the updated version of AMDA: AMDA 2.8.0 that has been implemented in R version 2.8.1 and Bioconductor 2.3 release. Furthermore, additional quality controls metrics, gene filtering and gene selection approaches have been added. To improve the understanding of the biological data, differentially expressed genes (DEGs) have been mapped into the KEGG pathways. The new updated design of AMDA intends to better respond to various experimental designs that can occur in Microarray experiments as well as to deliver more insightful biological understanding and up to date annotations.

Methods

AMDA code has been implemented by using Bioconductor's function repositories. Relative log expression (RLE) and the normalized unscaled standard error (NUSE) package has been added as statistical quality controls from affyPLM package [2]. Filtering method using Inter-Quartile Range (IQR) function has been added for gene selection (genefilter R package). Wrapper function [1] code has been further extended to include the new RankProduct and Limma Paired functions. Conversion and labeling of KEGG graphical presentation is possible thanks to ImageMagick R Software Suite.

Results

Quality Controls RLE plots the deviation of each gene from its median across arrays, which is 0. NUSE plots the standardized standard error estimates from the Probe Level Models fit; so, the median standard error across arrays is 1 for each gene. These quality controls make very easy the identification of microarray that stand out from the group due to poor RNA quality or failed hybridization to the microarray. IQR Based function to pre-filter the Expression Set A more flexible and experimental design oriented dataset filtering method using IQR function has been added for gene selection. The choice of IQR implementation is based on its computational ease and ability to remove uninformative probe sets – that do not vary in the expression set. The algorithm computes the IQR value for each probe set, and sets automatically the optimal IQR as the threshold by which filtering the probe set intensities. Rank Product Function Rank Product (RP,[2]) provides a straightforward and statistically meaningful way to determine the significance of differential expression for each gene. The RP approach is powerful for both identifying biologically relevant expression changes and controlling the false discovery rate (FDR). It shows to be reliable in highly noisy data performing a permutation test on the set of replicates. Limma Paired Sample Function Experimental design has been further customized with Paired Limma function. This method performs comparison between a common baseline and an experimental condition of a dataset containig paired samples, as occurring in datasets where sets of patients have been examined under different conditions. The implementation of a paired samples test makes AMDA useful also for analyzing data from clinical studies (e.g: lymphocytes before and after drug treatment) KEGG pathway Maps Enrichment of DEGs in KEGG gene sets and their mapping in the pathway diagram assesses the potential functional convergence of gene-signatures on basis the of the KEGG pathway modules. DEGs that are significantly enriched in a pathway are mapped on the KEGG graphical representation of the pathway (blue down- and red up-regulated genes). The new improved version of AMDA includes several new features. This release shows a pipeline more oriented towards gene selection for DEGs identification and their biological annotation. Future work will include the addition of more tools and the improvement of the connectivity with other software packages.

References

[1] Pelizzola M, Pavelka N, Foti M and Ricciardi-Castagnoli P. AMDA: an R package for the automated microarray data analysis. BMC Bioinformatics 2006, 7:335

[2] Gentleman RC, Carey VJ, Bates DJ, Bolstad BM, Dettling M, Dudoit S, Ellis B, Gautier L; Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth GK, Tierney L Yang YH, Zhang J. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology. 2004;5

Contact : dimos.kapetis@gmail.com