Merging R into the web: DNMAD & Tnasas, web-based resources for microarray data analysis.

Juan M. Vaquerizas, Joaquín Dopazo and Ramón Díaz-Uriarte
Bioinformatics Unit, CNIO (Spanish National Cancer Centre)
Melchor Fernández Almagro 3, 28029 Madrid, Spain.

April, 2004

Abstract

DNMAD and Tnasas are two web-based resources for microarray data analysis (normalization and class prediction), that implement R code into a web server configuration. These tools are included in GEPAS, a suite for gene expression data analysis [4].

The first tool we present is called DNMAD that stands for "**D**iagnosis and **N**ormalization for **M**icro**A**rray **D**ata". DNMAD is essentially a web-based interface for the Bioconductor package limma [7]. DNMAD has been designed to provide a simple and at the same time robust and fast way for normalization of microarray data. We provide global and print-tip loess normalization [9]. We also offer options for the use of quality flags, background adjustment, input of compressed files with many of arrays, and slide-scale normalization. Finally, to enable a fully featured pipeline of analysis, direct links to the central hub of GEPAS, the Preprocessor [5] are provided. The tool, help files, and tutorials are available at `http://dnmad.bioinfo.cnio.es`.

The second tool we present is called Tnasas for "**T**his is **n**ot **a** **s**ubstitute for **a** **s**tatistician". Tnasas performs class prediction (using either SVM, KNN, DLDA, or Random Forest) together with gene selection, including cross-validation of the gene selection process to account for "selection bias" [1], and cross-validation of the process of selecting the number of genes that yields the smallest error. This tool has mainly an exploratory and pedagogical purpose: to make users aware of the widespread problem of selection bias and to provide a benchmark against some (overly) optimistic claims that occasionally are attached to new methods and algorithms; Tnasas can also be used as a "quick and dirty" way of building a predictor. Tnasas uses packages e1071 [2], randomForest [6], and class (part of the VR bundle [8]). The tool, help files, and tutorials are available at `http://tnasas.bioinfo.cnio.es`.

Both tools use a Perl-CGI script that checks all the data input by the user and, once checked, sends it to an R script that performs the actions requested by the user. To merge the Perl-CGI script with the script in R we use the R package CGIwithR [3]. Source code for both applications will be released under the GNU GPL license.

References

[1] C. Ambroise and G.J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci USA*, 99: 6562–6566, 2002.

[2] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingess. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien.
`http://cran.r-project.org/src/contrib/PACKAGES.html#e1071`

[3] D. Firth. CGIwithR: Facilities for the use of R to write CGI scripts.
`http://cran.r-project.org/src/contrib/PACKAGES.html#CGIwithR`

[4] J. Herrero, F. Al-Shahrour, R. Díaz-Uriarte, Á. Mateos, J. M. Vaquerizas, J. Santoyo, and J. Dopazo. GEPAS: A web-based resource for microarray gene expression data analysis. *Nucleic Acids Res*, 31:3461–3467, 2003.

[5] J. Herrero, R. Díaz-Uriarte, and J. Dopazo. Gene expression data preprocessing. *Bioinformatics*, 19:655–656, 2003.

[6] A. Liaw, M. Wiener, with original Fortran code from L. Breiman and A. Cutler. randomForest: Breiman's random forest for classification and regression.
`http://cran.r-project.org/src/contrib/PACKAGES.html#randomForest`

[7] G. K. Smyth, M. Ritchie, J. Wettenhall, and N. Thorne. limma: Data analysis, linear models and differential expression for microarray data.
`http://www.bioconductor.org/repository/release1.3/package/html/limma.html`

[8] B. D. Ripley and W. Venables. VR: Functions and datasets to support Venables and Ripley, 'Modern Applied Statistics with S' (4th edition). `http://cran.r-project.org/src/contrib/PACKAGES.html#VR`

[9] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 30:e15, 2002.