W308–W314 Nucleic Acids Research, 2008, Vol. 36, Web Server issue Published online 27 May 2008
doi:10.1093/nar/gkn303

GEPAS, a web-based tool for microarray data analysis and interpretation

Joaquín Tárraga[1,2], Ignacio Medina[1,3], José Carbonell[1], Jaime Huerta-Cepas[1,2], Pablo Minguez[1,3], Eva Alloza[1], Fátima Al-Shahrour[1], Susana Vegas-Azcárate[4], Stefan Goetz[1,3], Pablo Escobar[1], Francisco Garcia-Garcia[1,2], Ana Conesa[1], David Montaner[1,2] and Joaquín Dopazo[1,2,3,*]

[1]Bioinformatics Department, Centro de Investigación Príncipe Felipe (CIPF), Autopista del Saler 16, E46013, [2]Functional Genomics Node, INB, CIPF, Autopista del Saler 16, E46013, [3]Center for Biomedical Research on Rare Diseases (CIBERER), CIPF, Autopista del Saler 16, E46013, Valencia and [4]Department of Statistics, Universidad Carlos III, E28903, Madrid, Spain

Received January 31, 2008; Revised April 29, 2008; Accepted April 30, 2008

ABSTRACT

Gene Expression Profile Analysis Suite (GEPAS) is one of the most complete and extensively used web-based packages for microarray data analysis. During its more than 5 years of activity it has continuously been updated to keep pace with the state-of-the-art in the changing microarray data analysis arena. GEPAS offers diverse analysis options that include well established as well as novel algorithms for normalization, gene selection, class prediction, clustering and functional profiling of the experiment. New options for time-course (or dose-response) experiments, microarray-based class prediction, new clustering methods and new tests for differential expression have been included. The new *pipeliner* module allows automating the execution of sequential analysis steps by means of a simple but powerful graphic interface. An extensive re-engineering of GEPAS has been carried out which includes the use of web services and Web 2.0 technology features, a new user interface with persistent sessions and a new extended database of gene identifiers. GEPAS is nowadays the most quoted web tool in its field and it is extensively used by researchers of many countries and its records indicate an average usage rate of 500 experiments per day. GEPAS, is available at http://www.gepas.org.

INTRODUCTION

Since its introduction in the mid 1990s (1), microarrays have revolutionized the way in which the research community addresses biological problems. Its success relays on its application to classify types of tumours (2), predicting disease outcome (3) or even the response to treatments (4). These practical applications of microarrays, despite them not being free of criticisms (5), have definitively fuelled the use of the methodology. In this scenario, the real bottleneck in the use of microarray technologies comes from the data analysis step (6). The web-based package Gene Expression Profile Analysis Suite (GEPAS) has been growing during the last 5 years (7–10) trying to keep pace with the state-of-the-art in algorithms for high-throughput gene expression data analysis as well as responding to the demands of the microarray community.

Although originally designed to analyse microarray data, the most important modules of GEPAS are not tied to the technology or to the microarray platforms used to extract the data on gene expression. GEPAS is rather oriented to analyse high-throughput gene expression data and to test different types of genome-scale hypotheses.

GEPAS is not a web server of a simple tool, but it constitutes one of the largest resources for integrated microarray data analysis available over the web. GEPAS is used by researchers worldwide as can be seen in the usage map, where all the sessions are mapped to its geographic location (http://bioinfo.cipf.es/access_map/map.html). By the end of year 2007, an average of 500 experiments per day were being analysed in GEPAS. The recent release 4.0 presented here includes new modules, new tests in already existent modules, technical improvements (GEPAS is now based on web services technology and includes Web 2.0 features) and a more powerful and intuitive interface which includes graphical tools to define workflows and persistent private sessions.

*To whom correspondence should be addressed. Tel: +34 96 328 96 80; Fax: +34 96 328 97 01; Email: jdopazo@cipf.es

GENERAL OVERVIEW

GEPAS has been designated for the analysis of high-throughput gene expression data. Obviously, today this means microarray data analysis, but this situation might change in the future and the data could come from different platforms or technologies. Although some of their modules are platform dependent, the core of GEPAS aims to analyse and test hypothesis using gene expression data in a simple but rigorous way.

Many different biological questions can be addressed through gene-expression experiments, nevertheless, there are usually three types of objectives in this context: ‘class comparison’, ‘class prediction’ and ‘class discovery’ (6). The first two objectives fall into the category of supervised methods and usually involve the application of tests to define differentially expressed genes, or the use of different procedures to predict class membership on the basis of the values observed for a number of ‘key’ genes. Clustering methods belong to the last category, also known as unsupervised analysis, because no previous information about the class structure of the data set is used in the study. Thus, GEPAS is composed by the following modules:

Normalization and pre-processing

GEPAS implements normalization facilities for both two-colour and Affymetrix arrays. Normalization in two-colour arrays is performed using print-tip loess (11) with a number of different options. Affymetrix CEL files using standard bioconductor (12) tools, in particular the package affy (13). Besides its friendly web interface we provide the user with the speed and above all, the physical memory available in our server. In addition, the *pre-processor* (14) module performs some pre-processing of the data (log-transformations, standardizations, imputation of missing values, etc.).

Class discovery

Clustering techniques are used for class discovery either in genes or in experiments. GEPAS includes the best performing clustering methods according to different independent benchmarkings (15,16). There are obviously more methods but among the most extensively used for gene expression data clustering we can highlight: hierarchical clustering (17), SOM (18), *SOTA* (19) and *K-means* (20). It is worth mentioning that the version of SOM implemented here can automatically find the optimal number of clusters (21).The evaluation of cluster quality, a barely addressed issue, has been implemented here using the silhouette method (22), which presents an optimal performance in noisy situations, such as microarray data (23), along with some descriptive measures for each cluster partition (average profiles, standard deviation profiles, inter- and intra-cluster distances).

Differential gene expression

GEPAS implements tests for finding genes with significant differences in expression between two or more classes, related to a continuous experimental factor (e.g. the concentration of a metabolite) or to survival data. For **two-class** comparisons, GEPAS implements the popular *t-test*, the *empirical Bayes test* (24), the *CLEAR-test* that combines differential expression and variability (25), the *data-adaptive test* (26) and the *SAM test* (27). For comparisons involving **more than two classes** GEPAS uses the classical *ANOVA*. In order to find genes whose expression is significantly **correlated to a continuous variable** (e.g. the level of a metabolite), regression analysis and estimates of Pearson’s and Spearman’s correlation co-efficients can be obtained. Finally, for finding genes whose expression is related to **survival** times GEPAS estimates a *Cox proportional hazards regression model* (28). Right censored data is allowed as well as replicates in the survival times. Censoring variables should be provided by the researcher together with survival times that may be replicated.

When appropriate, P values adjusted for multiple testing are provided. Three methodologies are implemented. One of them controls the FWER (family-wise error rate) (29) while the others control the FDR (false discovery rate) (30).

Predictors

A new module for **class prediction** (31) has been implemented. The module includes different classifiers, such as diagonal linear discriminant analysis (DLDA) (32), k-nearest neighbour (KNN) (33), support vector machines (SVM) (34), SOM (18) and shrunken centroids (PAM) (35) of well-known efficiency as class predictors using microarray data (32). Cross-validation error is calculated in such a way as to avoid the well-known selection bias problem (36). See ref. (31) for details. Once the model has been trained it can be used for further prediction of new samples. This implementation is unique among similar programmes.

Time-course and dose–response gene-expression experiments

A new module for the analysis of multi-series time-course and dose–response microarray experiments has been added. In this type of experiments, the researcher aims to study gene expression changes across time or across dosages and to evaluate trend differences between the various experimental groups (37).

This module implements and extends the maSigPro statistical approach for the study of gene expression changes along time and the specific trend differences between various experimental groups (38). The method is a two-regression step approach where individual series are identified by dummy variables. The procedure first adjusts a global regression model which considers all experiment series and a maximum complexity in the time/dosage-dependent response. This first step indentifies differentially expressed genes at a given false positive control rate. In the second step, a variable selection method is applied to find the best model for each gene and to analyse particular significant profile differences between series. Finally, significant genes are clustered and displayed showing these trend differences.

Functional profiling

There are many available tools that make use of gene functional annotations to provide an interpretation for the observed global changes in gene expression in microarray experiments (39). Probably, one of the most complete packages for functional profiling analysis is the Babelomics suite (40,41). This suite of programs for functional annotation of genome-scale experiments has undergone a deep modification described in detail elsewhere (Al-Shahrour, submitted to this issue). Babelomics performs functional enrichment analysis, that is, comparing two lists of genes and testing simultaneously in order to find significant over-abundance of diverse biologically relevant terms that would define functional modules such as GO, KEGG pathways, Interpro motifs or regulatory modules such as Transfac® motifs, CisRed motifs, miRNA binding motifs or other types of modules such as the ones defined by relative abundance in tissues and bioentities extracted from PubMed. All the tests are further adjusted for multiple testing effects (42,43). Additionally, gene set enrichment analysis can be performed using different algorithms (44,45) using several sources of information (46). The Babelomics suite is fully integrated into GEPAS. Gene expression analyses resulting in lists of genes to be compared (different clusters, genes differentially expressed, etc.) can be submitted to Babelomics for functional enrichment analysis. Moreover, arrangements of genes according to, for example, differential expression or other criteria can be sent to Babelomics to be studied by gene set enrichment analysis. This allows discovering pathways or functional modules of genes that are coordinately activated or deactivated in the experiment studied.

Entry points and data formats

There are two entry points to GEPAS: platform dependent and platform independent. GEPAS accepts and normalizes different types of microarray data which include Affymetrix CEL files and 13 different two-channel arrays including Agilent, Genepix and other. Once the files are normalized any type of analysis can be applied. On the other hand, there is another simple format by means of which data from other platforms, other technologies (e.g. SAGE) and even other nature (e.g. proteomics, Chip-on-chip data) can be input in any of the GEPAS modules. A very simple text file with the numeric gene expression values are in the format of a tabulator-delimited matrix, in which rows make reference to gene identifiers and columns to experiments, can be used for this purpose. Information on the experiments can be stored in the first rows starting by a # symbol. The first column contains the gene identifiers.

WHAT IS NEW IN VERSION 4.0?

The novelties added to this version have been described in more detail above, in the general overview of the programme. Summarizing, we have implemented a number of new tests, inexistent in previous versions, apart from new whole modules. Thus, much more options for normalization have been added (support for 12 more formats). New tests for differential expression such as an improved version of clear (25) test or the popular SAM test (27) were implemented. The module for cluster visualization has also been extensively improved. Much work has been invested if implementing an improved tool for protein and gene ID conversion which includes a large number of species and databases. Now, the converter tool supports more than 10 species and more than 40 gene ID references for human [including single nucleotide polymorphism (SNP) and orthologous information]. In general, almost all the modules of GEPAS have undergone improvements to some extent. We have included a new complete module that allows the analysis of multi-series time-course and dose–response microarray experiments. The module is an implementation of the maSigPro statistical approach for the study of gene expression changes along time and the specific trend differences between various experimental groups (38). Another new module is the clustering by a version of SOM (21) that automatically finds the number of clusters. Obviously, the Babelomics has its own catalogue of novelties that are described in an accompanying paper.

In addition, there are technical novelties such as the re-engineering to web services, the inclusion of Web 2.0 technology features, the new interface of sessions and the pipeliner, which are described below.

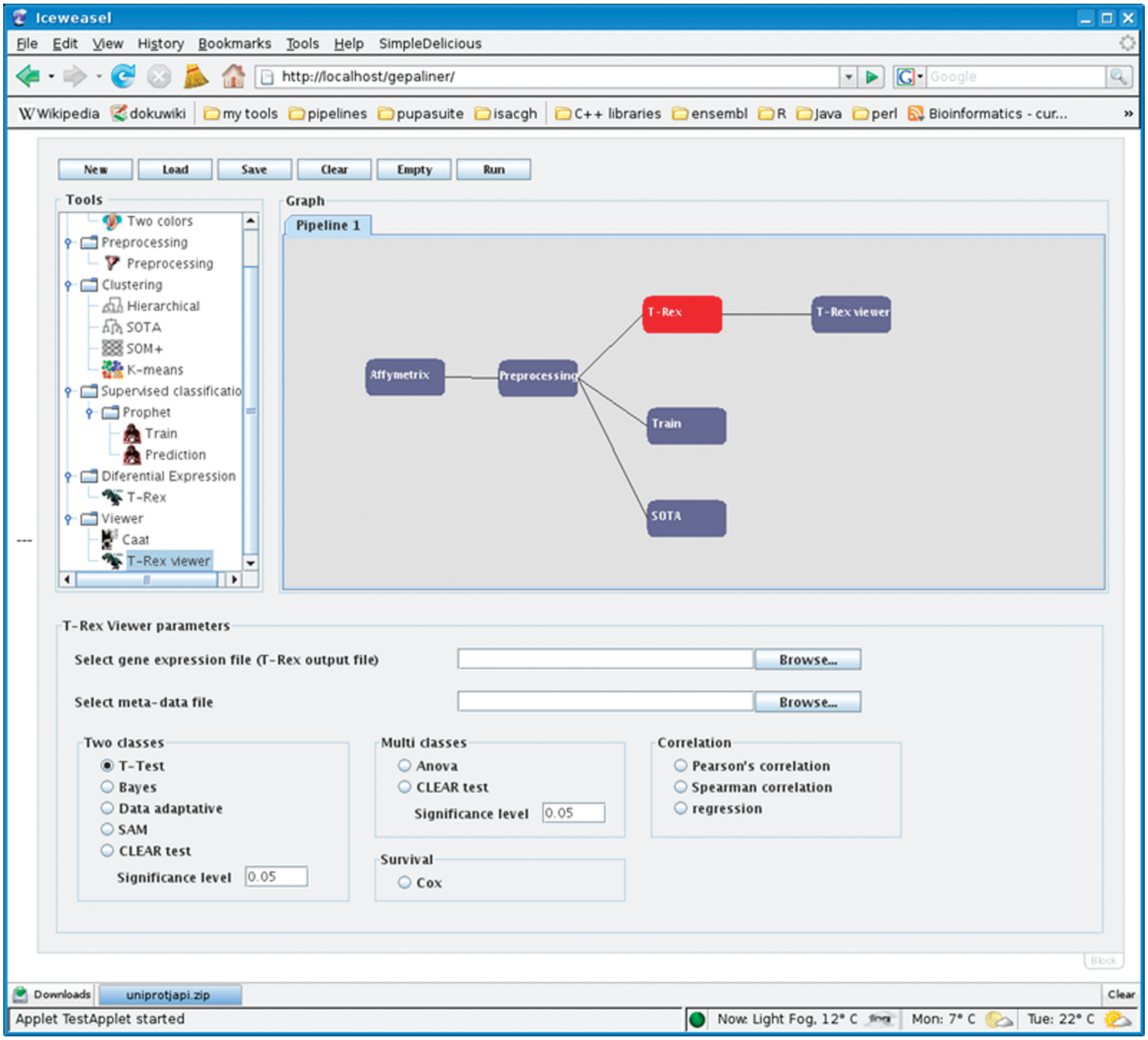
All the novelties included in GEPAS are, in terms of resources invested, far beyond the work demanded by a conventional web server that offers a unique facility.

The pipeliner: a graphic module for easy implementation of workflows

Microarray data analysis consists of a series of steps that can be carried out by sequentially running different GEPAS modules (e.g. normalization + pre-processing + gene selection + functional profiling of significant genes). If some of these steps have to be repeated systematically many times (which would happen, for example in a microarray core facility) it is easier to have the possibility of saving the sequence of operations as a workflow and using it in future analysis. The possibility of saving and storing operations is also useful when a researcher uses a non-default set of parameters in the tools. The advanced 'pipeliner' module allows users to define workflows, for repetitive tasks, in a completely visual manner by choosing, dragging and dropping icons representing the different modules in the package (without the need of any scripting skills). Figure 1 shows the graphic interface that allows defining sequences of operations as well as setting the parameter used in these. The workflows so defined by this Java applet can be stored in the sessions and can be further loaded from them.

Internal re-engineering, technological improvements and the session interface

GEPAS has been completely re-engineered and now it is based on SOAP web services and on new Web 2.0 technology features such as AJAX. This has facilitated the design of a new interface that allows asynchronous use,



T-Rex Viewer parameters

Select gene expression file (T-Rex output file) Browse...

Select meta-data file Browse...

Two classes
- ● T-Test
- ○ Bayes
- ○ Data adaptative
- ○ SAM
- ○ CLEAR test

Significance level 0.05

Multi classes
- ○ Anova
- ○ CLEAR test

Significance level 0.05

Survival
- ○ Cox

Correlation
- ○ Pearson's correlation
- ○ Spearman correlation
- ○ regression

Figure 1. The pipeliner interface with the available modules on the left and the customization options window below. Modules can be dragged and dropped on the screen and the sequence of execution is defined by linking them. Clicking on a module brings about the corresponding parameters' window below. Workflows defined in this straightforward manner can be stored in the session manager and used in future sessions.

as well as projects, jobs and user management. Thus, the users can choose between the traditional anonymous sessions without loging in (as in previous versions) or to log into the new environment with username and password. This new environment offers persistent sessions in which data is kept stored as well as different facilities for tracking of the operations performed. Both options are free.

GEPAS is now running in a high-end cluster with 10 dedicated Intel XEON Quad-Core CPUs at 2.0 GHz (summing up a total of 40 cores) with a large amount of RAM (total 60 GB). In this way we can offer a high computer power to end users.

An improved module for protein and gene ID conversion including a large number of species and databases is used behind the scene. This module allows importing any microarray file regardless of the IDs used in the platform. More species and gene references have been added and now the converter module supports more than 10 species and more than 40 ID references for human (including SNP and orthologous information). This module has been implemented in Java to speed up the performance. Besides the web interface a public web service Application Programming Interface is provided, allowing anyone to access the data from their code.

Related training activities

In addition, there is a teaching programme related to GEPAS (http://bioinfo.cipf.es/docus/courses/courses.html) with on-line tutorials that can be freely used (http://bioinfo.cipf.es/docus/courses/on-line.html).

GEPAS usage

The impact over the user's community has been estimated by the corresponding number of Scholar Google citations. According to the number of citations, GEPAS is by far the most popular web resource in its category with 196 citations [252 if the citations of the SOTA (19) are included]. The updated citations for the web-tools with a significant presence in the scientific community can be found at: http://bioinfo.cipf.es/docus/tools-citations/microarrays. GEPAS is used by a broad research community of many countries and its records indicate an average usage rate of around 500 users per day. The geographical distribution of users can be monitored in real time at: http://bioinfo.cipf.es/access_map/map.html. The web-based pipeline for microarray gene expression data, GEPAS, is available at http://www.gepas.org.

Future plans

We are working on several improvements that will be released in an upcoming version. These include normalization for one channel Agilent arrays, for exon arrays (both Agilent and Affymetrix), for tiling arrays and for Illumina arrays. New tests for differential expression will be included. A new version of the predictor with more predictor tools and new cross-validation methods will also be implemented. The ISACHG (47) for array-CGH analysis will be fully integrated in GEPAS and interfaces to databases such as ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) or Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) will be provided.

DISCUSSION

GEPAS is a long-term, ongoing ambitious project that aims to provide the scientific community with an advanced set of tools for high-throughput gene expression data analysis, without renouncing to an easy and intuitive use. Since its official release in 2003 (7), GEPAS has been running uninterruptedly and has grown-up to include more tools to keep pace with the novelties in the microarray data analysis arena (7–9). GEPAS has the vocation of being a consistent set of both state-of-the-art and widely established algorithms, instead of a simple collection of as-much-as-possible tools. In fact, any new tool which has been included in the package has been the response to a new or emerging requirement requested by our users. As the Functional Genomics node of the Spanish Institute of Bioinformatics (INB; http://www.inab.org) and being part of the Spanish Network of Cancer (RTICC; http://www.rticcc.org) and the Network of Centres for Research in Rare Diseases (CIBERER, http://www.ciberer.es), we have a direct contact with researchers from which we get much of the feedback necessary to build up a useful tool. We are also integrated in the EMERALD project (http://www.microarray-quality.org/), where we will provide input in the data mining methodologies such as clustering, gene selection or predictors, to assess the implications of QA/QC.

GEPAS, integrated with the Babelomics suite (40,41), offers all the necessary methods in order to perform the most common analysis of microarray data. GEPAS has been designed to take full advantage of the properties of the web: connectivity, cross-platform functionality and remote usage. Its modular architecture based on web services allows easy implementation of new tools and facilitates the connectivity of GEPAS from and to other web-based tools.

It cannot be discarded that the technologies and the platforms will change in the future. Such foreseeable changes can only affect the entry point and the technology-related part of GEPAS (that is, the normalization). The important contribution of GEPAS is its potential for analyzing high-throughput gene expression data and for testing different types of hypotheses in this context, regardless the technology that has produced such results.

The step of functional interpretation is typically made by studying the enrichment in pre-defined modules of genes related among them by any interesting biological property (common function, regulation, chromosomal location, etc.) as a function of some parameter derived from the experiment. Thus, functional enrichment methods (39) are used to find gene modules significantly over-represented among the relevant genes selected in the experiment. Over-representation of a given gene module means that genes with a particular property have been activated or deactivated in the experiment. Recently, gene set enrichment methods are superseding conventional functional enrichment methods for the functional interpretation of high-throughput gene-expression data, given their higher sensitivity (39,48,49). Both families of methods along with several definitions of modules (functional, transcriptional, text-mining based and phenotypical and tissues based) are implemented in the Babelomics module, fully integrated in GEPAS.

GEPAS is now running in a high-end cluster that offers high computer power. This allows using tools (for example normalization tools are highly RAM-consuming) that are usually beyond the capabilities of the hardware available to many end users.

Although there are many alternatives for microarray data analysis, there is no other similar resource over the web with the number of possibilities offered by GEPAS.

ACKNOWLEDGEMENTS

This work is supported by grants from the Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER) ISCIII, and projects BIO2005-01078 from the Spanish Ministry of Education and Science, EMERALD from the EU and the National Institute of Bioinformatics (www.inab.org), a platform of Genoma España. Funding to pay the Open Access publication charges for this article was provided by project

BIO2005-01078 fom the Spanish Ministry of Education and Science.

Conflict of interest statement. None declared.

REFERENCES

1. Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.
2. Sorlie,T., Perou,C.M., Tibshirani,R., Aas,T., Geisler,S., Johnsen,H., Hastie,T., Eisen,M.B., van de Rijn,M., Jeffrey,S.S. *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA*, **98**, 10869–10874.
3. Beer,D.G., Kardia,S.L., Huang,C.C., Giordano,T.J., Levin,A.M., Misek,D.E., Lin,L., Chen,G., Gharib,T.G., Thomas,D.G. *et al.* (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, **8**, 816–824.
4. van 't Veer,L.J., Dai,H., van de Vijver,M.J., He,Y.D., Hart,A.A., Mao,M., Peterse,H.L., van der Kooy,K., Marton,M.J., Witteveen,A.T. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.
5. Simon,R. (2005) Roadmap for developing and validating therapeutically relevant genomic classifiers. *J. Clin. Oncol.*, **23**, 7332–7341.
6. Allison,D.B., Cui,X., Page,G.P. and Sabripour,M. (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.*, **7**, 55–65.
7. Herrero,J., Al-Shahrour,F., Diaz-Uriarte,R., Mateos,A., Vaquerizas,J.M., Santoyo,J. and Dopazo,J. (2003) GEPAS: a web-based resource for microarray gene expression data analysis. *Nucleic Acids Res.*, **31**, 3461–3467.
8. Herrero,J., Vaquerizas,J.M., Al-Shahrour,F., Conde,L., Mateos,A., Diaz-Uriarte,J.S. and Dopazo,J. (2004) New challenges in gene expression data analysis and the extended GEPAS. *Nucleic Acids Res.*, **32**, W485–W491.
9. Vaquerizas,J.M., Conde,L., Yankilevich,P., Cabezon,A., Minguez,P., Diaz-Uriarte,R., Al-Shahrour,F., Herrero,J. and Dopazo,J. (2005) GEPAS, an experiment-oriented pipeline for the analysis of microarray gene expression data. *Nucleic Acids Res.*, **33**, W616–W620.
10. Montaner,D., Tarraga,J., Huerta-Cepas,J., Burguet,J., Vaquerizas,J.M., Conde,L., Minguez,P., Vera,J., Mukherjee,S., Valls,J. *et al.* (2006) Next station in microarray data analysis: GEPAS. *Nucleic Acids Res.*, **34**, W486–W491.
11. Smyth,G., Yang,Y. and Speed,T. (2003) Statistical issues in microarray data analysis. In Brownstein,M. and Khodursky,A. (eds), *Functional Genomics: Methods and Protocols*, Vol. 224. Humana Press, Totowa, NJ, pp. 111–136.
12. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
13. Gautier,L., Cope,L., Bolstad,B.M. and Irizarry,R.A. (2004) affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307–315.
14. Flicek,P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2008) Ensembl 2008. *Nucleic Acids Res.*, **36**, D707–D714.
15. Handl,J., Knowles,J. and Kell,D.B. (2005) Computational cluster validation in post-genomic data analysis. *Bioinformatics*, **21**, 3201–3212.
16. Datta,S. and Datta,S. (2006) Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC Bioinformatics*, **7**, 397.
17. Sneath,P. and Sokal,R. (1973) *Numerical Taxonomy*. W. H. Freeman, San Francisco.
18. Kohonen,T. (1997) *Self-organizing Maps*. Springer, Berlin.
19. Herrero,J., Valencia,A. and Dopazo,J. (2001) A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, **17**, 126–136.
20. Hartigan,J. and Wong,M. (1979) A k-means clustering algorithm. *Appl. Stat.*, **28**, 100–108.
21. Vegas-Azcárate,S. and Muruzábal,J. (2005) *Biosignal Processing and Classification*, Vol. 1. Insticc Press, Setubal, pp. 50–59.
22. Rousseeuw,P. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.
23. Azuaje,F. (2002) A cluster validity framework for genome expression data. *Bioinformatics*, **18**, 319–320.
24. Kendziorski,C.M., Newton,M.A., Lan,H. and Gould,M.N. (2003) On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Stat. Med.*, **22**, 3899–3914.
25. Valls,J., Grau,M., Sole,X., Hernandez,P., Montaner,D., Dopazo,J., Peinado,M.A., Capella,G., Moreno,V. and Pujana,M.A. (2007) CLEAR-test: combining inference for differential expression and variability in microarray data analysis. *J. Biomed. Inform.*, **41**, 33–45.
26. Mukherjee,S., Roberts,S.J. and van der Laan,M.J. (2005) Data-adaptive test statistics for microarray data. *Bioinformatics*, **21 (Suppl. 2)**, ii108–ii114.
27. Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.
28. Klein,J.P. and Moeschberger,M.L. (2003) *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York.
29. Holm,S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, **6**, 65–70.
30. Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.
31. Medina,I., Montaner,D., Tarraga,J. and Dopazo,J. (2007) Prophet, a web-based tool for class prediction using microarray data. *Bioinformatics*, **23**, 390–391.
32. Dudoit,S., Fridlyand,J. and Speed,T. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77–87.
33. Ripley,B. (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.
34. Vapnik,V. (1998) *Statistical Learning Theory*. Wiley, New York.
35. Tibshirani,R., Hastie,T., Narasimhan,B. and Chu,G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.
36. Ambroise,C. and McLachlan,G.J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA*, **99**, 6562–6566.
37. Simon,I., Siegfried,Z., Ernst,J. and Bar-Joseph,Z. (2005) Combined static and dynamic analysis for determining the quality of time-series expression profiles. *Nat. Biotechnol.*, **23**, 1503–1508.
38. Conesa,A., Nueda,M.J., Ferrer,A. and Talon,M. (2006) maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics*, **22**, 1096–1102.
39. Dopazo,J. (2006) Functional interpretation of microarray experiments. *OMICS*, **10**, 398–410.
40. Al-Shahrour,F., Minguez,P., Tarraga,J., Montaner,D., Alloza,E., Vaquerizas,J.M., Conde,L., Blaschke,C., Vera,J. and Dopazo,J. (2006) BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res.*, **34**, W472–W476.
41. Al-Shahrour,F., Minguez,P., Vaquerizas,J.M., Conde,L. and Dopazo,J. (2005) BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucleic Acids Res.*, **33**, W460–W464.
42. Al-Shahrour,F., Diaz-Uriarte,R. and Dopazo,J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.
43. Al-Shahrour,F., Minguez,P., Tarraga,J., Medina,I., Alloza,E., Montaner,D. and Dopazo,J. (2007) FatiGO+: a functional profiling tool for genomic data. Integration of functional annotation,

regulatory motifs and interaction data with microarray experiments. *Nucleic Acids Res.*, **35**, W91–W96.

44. Al-Shahrour,F., Diaz-Uriarte,R. and Dopazo,J. (2005) Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics*, **21**, 2988–2993.
45. Al-Shahrour,F., Arbiza,L., Dopazo,H., Huerta-Cepas,J., Minguez,P., Montaner,D. and Dopazo,J. (2007) From genes to functional classes in the study of biological systems. *BMC Bioinformatics*, **8**, 114.
46. Minguez,P., Al-Shahrour,F., Montaner,D. and Dopazo,J. (2007) Functional profiling of microarray experiments using text-mining derived bioentities. *Bioinformatics*, **23**, 3098–3099.
47. Conde,L., Montaner,D., Burguet-Castell,J., Tarraga,J., Medina,I., Al-Shahrour,F. and Dopazo,J. (2007) ISACGH: a web-based environment for the analysis of Array CGH and gene expression which includes functional profiling. *Nucleic Acids Res.*, **35**, W81–W85.
48. Goeman,J.J. and Buhlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.
49. Nam,D. and Kim,S.Y. (2008) Gene-set approach for expression pattern analysis. *Brief. Bioinform.*, **9**, 189–197.