Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W91–W96***
doi:10.1093/nar/gkm260

FatiGO+: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments

Fátima Al-Shahrour[1], Pablo Minguez[1], Joaquín Tárraga[1,2], Ignacio Medina[1], Eva Alloza[1], David Montaner[1,2] and Joaquín Dopazo[1,2,*]

[1]Bioinformatics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia 46013, Spain and [2]Functional Genomics Node, INB, CIPF, Valencia 46013, Spain

Received January 30, 2007; Revised March 27, 2007; Accepted April 8, 2007

ABSTRACT

The ultimate goal of any genome-scale experiment is to provide a functional interpretation of the data, relating the available information with the hypotheses that originated the experiment. Thus, functional profiling methods have become essential in diverse scenarios such as microarray experiments, proteomics, etc. We present the FatiGO+, a web-based tool for the functional profiling of genome-scale experiments, specially oriented to the interpretation of microarray experiments. In addition to different functional annotations (gene ontology, KEGG pathways, Interpro motifs, Swissprot keywords and text-mining based bioentities related to diseases and chemical compounds) FatiGO+ includes, as a novelty, regulatory and structural information. The regulatory information used includes predictions of targets for distinct regulatory elements (obtained from the Transfac and CisRed databases). Additionally FatiGO+ uses predictions of target motifs of miRNA to infer which of these can be activated or deactivated in the sample of genes studied. Finally, properties of gene products related to their relative location and connections in the interactome have also been used. Also, enrichment of any of these functional terms can be directly analysed on chromosomal coordinates. FatiGO+ can be found at: http://www.fatigoplus.org and within the Babelomics environment http://www.babelomics.org

INTRODUCTION

It is well-known that genes do not operate alone within the cell, but in an intricate network of interactions that we only recently start to understand (1–3). These observations question the validity of the traditional reductionistic vision, in which one or a few key genes would be the causative factors of phenotypes or diseases (4), and urges to take into consideration the functional dimension in the interpretation of genome-scale experiments. In this new scenario, the deregulation of blocks of functionally related genes would be behind the disease phenotype (5). Thus, there is a clear necessity for methods and tools to assist in the functional interpretation of genome-scale experiments such as microarrays, and to formulate genome-scale hypothesis from a systems biology perspective (6–8) in a way that the collective properties of groups of genes are taken into account. Thus, a number of tools, pioneered by Onto-Express (9), or FatiGO (10) that used multiple-testing correction for the first time, GOMiner (11), etc., can be considered representatives of a family of methods that make use of different functional annotations (typically gene ontology (GO) or KEGG pathways) to find significant functional enrichments that might be useful in the interpretation of the results of microarray experiments (8,12). Functional enrichment methods are applied *a posteriori* on a group of genes of interest that have been selected in a previous step on the basis of their experimental values. For example, common criteria for selecting groups of genes in the context of microarray experiments would be their differential expression between two classes of experiments. By means of this simple two-step approach of gene selection followed by the functional enrichment analysis, a reasonable biological interpretation of a microarray experiment can be achieved. Nevertheless, it has been noted that the artificial imposition of a threshold in the gene selection step, which ignores the cooperative behaviour among genes, can have arbitrary consequences on the proper interpretation of the experiment (13). Thus, a new generation of methods which directly test the coordinate behaviour of blocks of functionally related genes has recently been proposed (14,15).

*To whom correspondence should be addressed. Tel: +34 963289680; Fax: +34 963289701; Email: jdopazo@cipf.es

There are, however, situations in which functional enrichment methods have a domain of application because the groups of genes to study are defined as discrete classes without the necessity of applying any arbitrary threshold. For example, many clustering methods define non-overlapping groups (clusters) of genes in a deterministic way. Other experiments such as ChIP-on-Chip can be used to define groups of genes under the control of one or more transcription factors, undergoing methylation, etc., and such groups are also real discrete classes.

We present here the FatiGO+, an advanced version of the original FatiGO program (10). FatiGO+ is part of a new generation of functional enrichment tools characterized by the integration of heterogeneous, biologically relevant information. FatiGO+ includes new functional annotations, more regulatory information, structural data on protein interactions, new tests and new facilities for the representation of the results, such as displaying enrichment values onto pathway representations. Additional Table 1 shows a comparison between FatiGO+ and other web-based tools for functional enrichment analysis. New aspects such as the use of the interactome scaffold (not merely the list of known complexes as in other tools (16,17)), the use of new regulatory information (e.g. miRNAs) or the use of text-mining derived functional terms, just to cite the most prominent features, represent a novelty of FatiGO+ with respect to other tools.

Program characteristics and functionality

The original FatiGO program (10) was published in 2004 and, since then it has become a extensively used tool (FatiGO/FatiGO+ have had an average of more than 100 experiments analysed per day and a cumulated total of more than 37 000 worldwide uses during the last year; see a map with the geographical distribution of users here: http://bioinfo.cipf.es/access_map/map.html). FatiGO+ constitutes the extension of the concept of functional enrichment not only to new functional annotations but also to other relevant classes of biological concepts that can be assigned to genes, or their corresponding gene products, such as regulation and protein interaction data.

Purpose of the analysis. The main purpose of FatiGO+ is to check for significant enrichments of the functional characteristics selected by the user in one of the lists of genes with respect to the other one. Typically this operation represents the comparison of a pre-selected list of genes with respect to the genes of reference (usually the rest of genes involved in the experiment). Functional enrichment is carried out by many different tools already available (8), nevertheless, FatiGO+ can be used for other purposes. The main screen displays four tabs. By default, the enrichment analysis option is selected. There two lists of genes (pre-selected and reference groups) can be input for their comparison. Another tab offer the possibility of carrying out enrichment analysis using functional terms defined by the user. Additionally, other tab can be used just for listing functional terms or characteristics present in a list of genes. Finally, another tab presents a different interface, in which enrichment analysis can be directly performed on chromosomal regions. Thus, the pre-selected group of genes is defined by delimiting the coordinates of a chromosomal region and the reference group is the rest of the genes in the genome. This is very useful to study the functional impact of groups of genes that cluster in close positions in the genome in pathologies that occur with chromosomal copy number alterations.

Input data format. Input data format are simple text files of genes (although data can also be pasted within the corresponding boxes). The most common gene or protein identifiers are accepted. This is achieved by using Ensembl (version 42) identifiers as universal cross-references. The following model organisms have been included in FatiGO+: *Homo sapiens, Mus musculus, Gallus gallus, Rattus norvegicus, Drosophila melanogaster, Danio rerio, Caenorhabditis elegans, Saccharomyces cerevisiae, Arabidopsis thaliana* and *Streptomyces coelicolor*. Data can also be directly imported from the GEPAS suite of microarray data analysis (http://www.gepas.org). In this way, it is possible from GEPAS to directly study functional enrichment of genes contained in a cluster or co-expression, genes differentially expressed, genes selected in a predictor, genes contained in a given chromosomal region, etc. When the option for functional enrichment using user-defined functional categories, in addition to the lists of genes a table of correspondence gene-annotation must be provided. Again this is a very simple text file with two tab-delimited columns in which the first one contains gene identifiers and the second one the functional annotation.

Understanding the differences between two lists of genes in the light of functional annotations. We have included in FatiGO+ different functional annotations available for genes. Probably the most widely used is GO (18). GO represents the biological knowledge as a tree (more precisely as a directed acyclic graph, DAG, in which a node can have more that one parent) where elements near the root of the tree make reference to more general concepts while deeper elements near the leaves of the tree make reference to more specific concepts. Conceptually we can consider that, if a gene is annotated to a given level then it is automatically annotated at all the upper levels (all the parent levels) up to the root. Since genes are annotated at different levels of the GO hierarchy, it is common to use this abstraction to choose a pre-defined level in the hierarchy instead of using directly the original levels of annotation of the genes. This strategy is used to produce GOSlim (see http://www.geneontology.org/GO.slims.shtml), which are cut-down versions of the GO hierarchy containing subsets of the terms in the whole GO at different levels. GoSlim levels give a broad overview of the ontology content without the detail of the specific fine grained terms. Choosing a GOSlim level of GO increases the power of the enrichment tests (7,8,12). What FatiGO+ does is the enrichment analysis at different GOSlim levels and reports the deepest (the more detailed definition) level at which significance is found. This strategy is known as nested inclusive analysis (NIA) (19).

The KEGG pathways database (20) is another well-known source for functional annotation. We have also included functional motifs mapped to proteins by the Interpro database (21) and the keywords in the Swissprot entries (22). All these functional annotations can be considered discrete classes (that is, a gene has or has not a given functional annotation) so, enrichment is tested by means of Fisher's exact test for 2×2 contingency tables (10,23,24). For each functional annotation the genes belonging to the two groups compared are distributed in a 2×2 contingency table, where rows account for presence/absence of the annotation, and columns represent each of the two clusters.

Additionally, we have used text-mining methods (25) to extract other functional aspects of the genes beyond the ones covered by the 'traditional' repositories described before. In the approach described here we have used two types of specific words (bioentities): those referred to chemical products and those related to diseases. The bioentities were extracted from PubMed, and are related to human genes by a score derived from the frequency of gene-bioentity co-occurrences and depending on their proximity within the text. The scores are based on how unlikely it is to observe a certain level of co-occurrences to happen by chance (26). The gene-bioentity correspondence tables with the respective scores were obtained using the AKS software (available at: http://www.bioalma.com/aks2/). Contrarily to the case of GO and other similar functional categories, bioentities are not discrete classes. Thus, the membership of a gene to a given bioentity is conditioned through the scores. Then, instead of the usual Fisher's (or equivalent) test, we use a Kolmogorov–Smirnov test to check for bioentity enrichments.

Finally, any other functional term can be used providing it can be represented as a table of correspondence between genes and instances of the term. Any user-defined table of correspondence gene-term can be input by FatiGO+ and functional enrichment with a Fisher's exact test on contingency tables can be conducted.

Understanding the differences between two lists of genes using regulatory information. Information on the regulation of gene expression can also be used within the context of functional enrichment tests. It can be useful to determine if, for example, a set of genes pre-selected because they co-express across time are under the control of the same transcription factor. Different databases containing transcription factor binding sites and other regulatory motifs are available so as promoter regions of the genes can be scanned for the possible presence of these target motifs. We have used information on diverse regulatory motifs from the CisRed (27) and Transfac (28) databases. These motifs can be studied at different locations in the promoters (in the first 1kb, in 5kb or up to 10kb).

Also other levels of regulation beyond the transcription factors or other common regulatory motifs can be considered by including information on microRNAs targets. It has recently been demonstrated the role of miRNAs in the negative regulation of the expression of their target genes (29). Actually, miRNAs have gained importance as cancer markers as they have demonstrated to be deregulated in some cancers (30). In the context of gene expression one might be interested in knowing whether there is some enrichment in miRNA target sequences in genes that have been deregulated in a particular cancer when compared to their non-deregulated counterparts. A significant enrichment would clearly point to the presence or absence of one or several miRNAs. Data on miRNA targets have been taken from the miRBase (31).

Understanding the differences between two lists of genes over the interactome scaffold. Data on protein–protein interaction allows to understand the physical interplay between blocks of genes. Two common cases of obvious biological relevance, in which we can expect interactions among genes, are protein complexes (or other protein aggregates) and signalling cascades. For example, in the context of gene expression it can be extremely useful to study the effect of gene expression changes within a pre-defined scaffold of interactions.

The parameter used here is the number of partners with which a protein interacts. FatiGO+ will check, by means of a Kolmogorov–Smirnov test for abnormally high (significant) number of interactions in one group of pre-selected genes with respect to the reference group. That would point to the possible existence of one or more complexes (or at least to the existence of a complex interaction scheme) within the group of pre-selected genes.

We use the modules graph and RBGL (32) from the bioconductor package to manage the interactome and calculate the parameters. The interactome database has been built using the PIANA (33) program, that allows integrating into one unique database different sources of protein–protein interactions. The databases included were: BIND (34), DIP (35), HPRD (36), MIPS (37), as well as interactions from different publications (2,3,38).

Comparing a set of genes to lists of gene expression in tissues and diseases. A quite common necessity in drug discovery or when testing animal models is to know whether the effect of a drug is restoring the normal function of a tissue or if a given model can be considered to represent a given disease. A powerful method to check the real impact of any treatment is to compare the expression of the genes with the corresponding profiles of gene expression in the situation aimed.

FatiGO+ uses two repositories containing information of gene expression in different tissues. One of them was obtained from the SAGE Tag libraries (Cancer Genome Anatomy Project). A total of 279 human libraries that belong to 29 different tissues and 190 mouse libraries from 26 tissues have been used (the data were obtained from: http://cgap.nci.nih.gov/SAGE). The other repository contains gene expression data obtained from microarray experiments available from the Genomics Institute of the Novartis Foundation data. A total of 79 human tissues and 61 mouse tissues with normal histology were obtained from http://wombat.gnf.org/index.html.

Multiple testing correction. All the *p*-values obtained for all the different functional annotations, regulatory elements, and the interactions according to the interactome scaffold, are adjusted using the False Discovery Rate (39) method. This produces a final report in which the FDR-adjusted *p*-values are provided for all the individual tests conducted.

Representation of the results. Different enrichment analysis are conducted through distinct tests and the results are represented in a number of intuitive graphical representations. The main results page contains a summary of all the tests performed. The user can follow the links to find details on each enrichment test conducted. Figure 1 shows some of the graphical representation of results that can be found for different tests. For example, Figure 1A shows the result of the NIA on the GO hierarchy. Many GO levels have been analysed and the deepest significant terms are reported. The user can also obtain the list of results at all the levels (Figure 1B). When a graphical representation of the relationships among genes exist, as is the case of KEGG pathways, these can be represented to visualize the possible impact of the genes in the structure. Figure 1C shows a pathway for which a significant enrichment was found, with the pre-selected genes in red and the genes in the reference group in green. The pathway images are obtained by using the public web services from KEGG at http://soap.genome.jp/KEGG.wsdl. When the biological information does not define a discrete class we use a Kolmogorov–Smirnov test. Figure 1D shows a different graphical representation for the result of one of such tests (in this case the number of protein interaction partners).

CONCLUSIONS

Despite new strategies oriented to the study of differential gene expression or other situations in which a threshold must be imposed (8,14,15), there are still many scenarios where discrete classes must be interpreted, in which the functional enrichment analysis is required (e.g. functional analysis of clustering results (40), ChIP-on-Chip experiments, massive KOs, etc.)

Since functional enrichment analysis is a 'classic' among the methods for the biological interpretation of lists of genes there are numerous tools available for such purpose (not all of them are web servers). Reviews can be found in (8,12). A common criticism to many of these tools is that

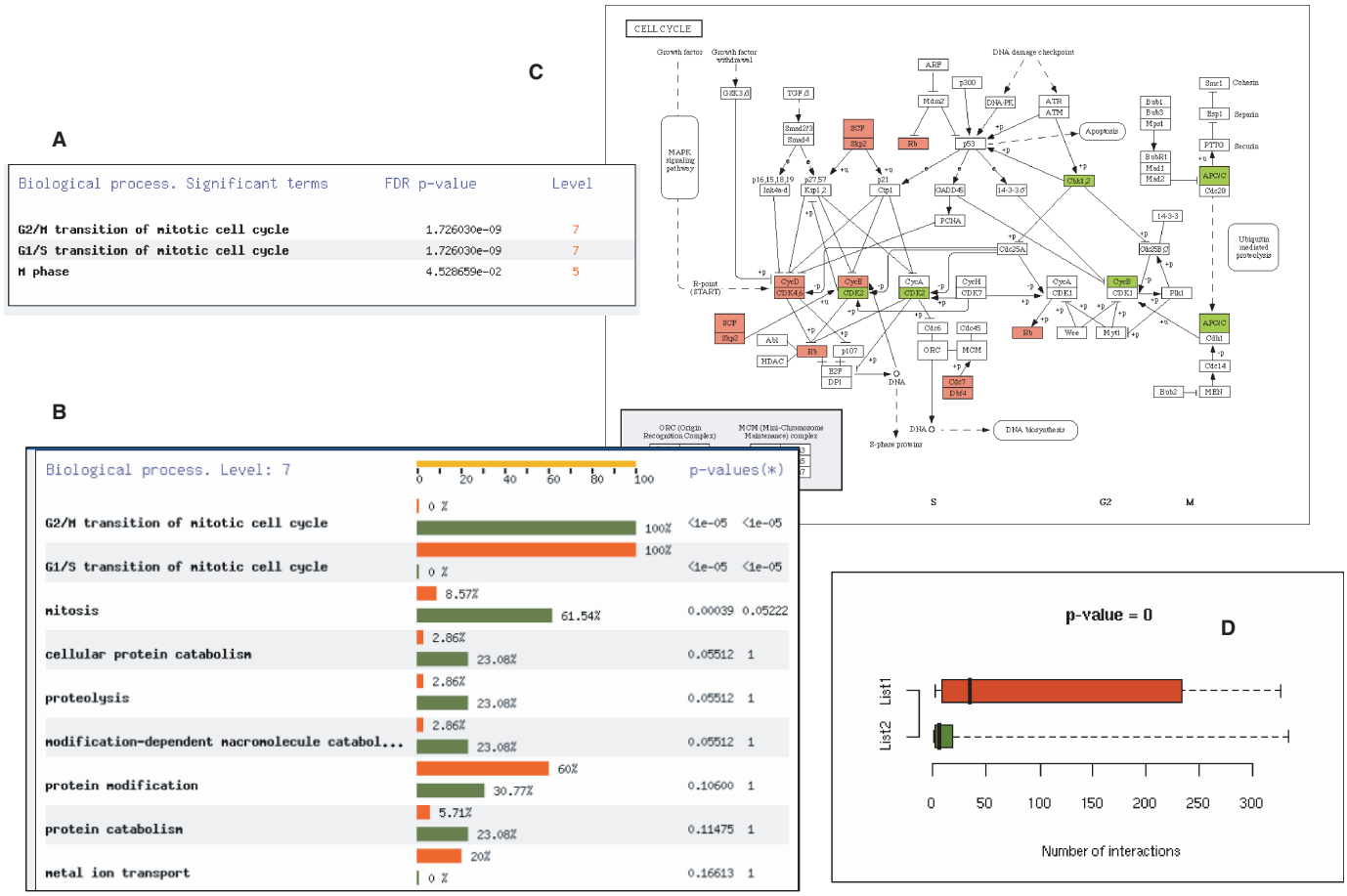

Figure 1. Different graphical representation of the results of the tests. **(A)** Deepest levels at which significance was found in the GO hierarchy. **(B)** All the GO annotations tested. Both, uncorrected and FDR-corrected *P*-values are shown. **(C)** A KEGG pathway with the genes in the pre-selected group labelled in red and the genes in the reference group labelled in green. **(D)** The result of a test that compares the degree of protein interactions in the pre-selected group of genes with respect to the reference group.

many of them offer almost exactly the same options with slight differences in the tests or in the strategies for testing (12). Programs for functional enrichment analysis must improve their performances by expanding the biologically relevant concepts to be tested beyond the common functional annotations which, in some cases, implies the implementation of different types of tests. It is of special importance the study of orthogonal concepts such as function, regulation, physical interactions, etc., given that it allows to detect genes with common annotations (e.g. genes involved in a given pathway and under the control of the same transcription factor), which provide a more complete biological interpretation of the experiment studied.

Given the number of distinct functional, regulatory, structural and other biologically relevant concepts that have been included in the FatiGO+ it can be considered one of the most complete tools for studying functional enrichment to our knowledge.

Since the biological interpretation of microarray experiments represents a major demand of functional enrichment methods we have connected FatiGO+ to our GEPAS (41,42) suite for microarray data analysis. Nevertheless, despite this primary use, it is important to remark that FatiGO+ can be applied to any type of genome-scale experiment (e.g. proteomics, massive KO with siRNAs, whole genome genotyping, etc.) or even to check genome-scale hypothesis (in evolution, population genetics, development, etc.)

ACKNOWLEDGEMENTS

This work is supported by grants from the Spanish ministry of education and science (BIO 2005-01078) 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 Genoma España.

Conflict of interest statement. None declared.

REFERENCES

1. Hallikas,O., Palin,K., Sinjushina,N., Rautiainen,R., Partanen,J., Ukkonen,E. and Taipale,J. (2006) Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell*, **124**, 47–59.
2. Rual,J.F., Venkatesan,K., Hao,T., Hirozane-Kishikawa,T., Dricot,A., Li,N., Berriz,G.F., Gibbons,F.D., Dreze,M. *et al*. (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173–1178.
3. Stelzl,U., Worm,U., Lalowski,M., Haenig,C., Brembeck,F.H., Goehler,H., Stroedicke,M., Zenkner,M., Schoenherr,A. *et al*. (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968.
4. Pinkel,D. and Albertson,D.G. (2005) Array comparative genomic hybridization and its applications in cancer. *Nat. Genet*., 37 Suppl, S11–S17.
5. Westerhoff,H.V. and Palsson,B.O. (2004) The evolution of molecular biology into systems biology. *Nat. Biotechnol*., **22**, 1249–1252.
6. Butcher,E.C., Berg,E.L. and Kunkel,E.J. (2004) Systems biology in drug discovery. *Nat. Biotechnol*., **22**, 1253–1259.
7. Al-Shahrour,F. and Dopazo,J. (2005).Ontologies and functional genomics. In: Azuaje,F. and Dopazo,J. (eds), *Data Analysis and Visualization in Genomics and Proteomics*, Wiley, West Sussex. pp. 99–112.
8. Dopazo,J. (2006) Functional interpretation of microarray experiments. *Omics*, **10**, 398–410.
9. Khatri,P., Draghici,S., Ostermeier,G.C. and Krawetz,S.A. (2002) Profiling gene expression using onto-express. *Genomics*, **79**, 266–270.
10. 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.
11. Zeeberg,B.R., Feng,W., Wang,G., Wang,M.D., Fojo,A.T., Sunshine,M., Narasimhan,S., Kane,D.W., Reinhold,W.C. *et al*. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*., **4**, R28.
12. Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.
13. Pan,K.H., Lih,C.J. and Cohen,S.N. (2005) Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays. *Proc. Natl Acad. Sci. U S A*, **102**, 8961–8965.
14. 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.
15. Mootha,V.K., Lindgren,C.M., Eriksson,K.F., Subramanian,A., Sihag,S., Lehar,J., Puigserver,P., Carlsson,E., Ridderstrale,M. *et al*. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet*., **34**, 267–273.
16. Dennis,G.Jr., Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C. and Lempicki,R.A. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*., **4**, P3.
17. Hosack,D.A., Dennis,G.Jr., Sherman,B.T., Lane,H.C. and Lempicki,R.A. (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol*., **4**, R70.
18. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.
19. Al-Shahrour,F., Minguez,P., Tarraga,J., Montaner,D., Alloza,E., Vaquerizas,J.M., Conde,L., Blaschke,C., Vera,J. *et al*. (2006) BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res*., **34**, W472–W476.
20. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.
21. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P. *et al*. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res*., **33**, D201–D205.
22. Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H. *et al*. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*., **34**, D187–D191.
23. Fisher,L. and Van Belle,G. (1993) *Biostatistics: A Methodology for the Health Sciences* Wiley, New York.
24. Draghici,S., Khatri,P., Martins,R.P., Ostermeier,G.C. and Krawetz,S.A. (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.
25. Krallinger,M. and Valencia,A. (2005) Text-mining and information-retrieval services for molecular biology. *Genome Biol*., **6**, 224.
26. Andrade,M.A. and Valencia,A. (1998) Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics*, **14**, 600–607.
27. Robertson,G., Bilenky,M., Lin,K., He,A., Yuen,W., Dagpinar,M., Varhol,R., Teague,K., Griffith,O.L. *et al*. (2006) cisRED: a database system for genome-scale computational discovery of regulatory elements. *Nucleic Acids Res*., **34**, D68–D73.
28. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Pruss,M., Reuter,I. *et al*. (2000) TRANSFAC: an

integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.
29. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
30. Lu,J., Getz,G., Miska,E.A., Alvarez-Saavedra,E., Lamb,J., Peck,D., Sweet-Cordero,A., Ebert,B.L., Mak,R.H. *et al.* (2005) MicroRNA expression profiles classify human cancers. *Nature*, **435**, 834–838.
31. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
32. Carey,V.J., Gentry,J., Whalen,E. and Gentleman,R. (2005) Network structures and algorithms in Bioconductor. *Bioinformatics*, **21**, 135–136.
33. Aragues,R., Jaeggi,D. and Oliva,B. (2006) PIANA: protein interactions and network analysis. *Bioinformatics*, **22**, 1015–1017.
34. Bader,G.D., Betel,D. and Hogue,C.W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **31**, 248–250.
35. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
36. Mishra,G.R., Suresh,M., Kumaran,K., Kannabiran,N., Suresh,S., Bala,P., Shivakumar,K., Anuradha,N., Reddy,R. *et al.* (2006) Human protein reference database—2006 update. *Nucleic Acids Res.*, **34**, D411–D414.
37. Mewes,H.W., Amid,C., Arnold,R., Frishman,D., Guldener,U., Mannhaupt,G., Munsterkotter,M., Pagel,P., Strack,N. *et al.* (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.*, **32**, D41–44.
38. Espadaler,J., Romero-Isart,O., Jackson,R.M. and Oliva,B. (2005) Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships. *Bioinformatics*, **21**, 3360–3368.
39. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B*, **57**, 289–300.
40. 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.
41. 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.
42. Montaner,D., Tarraga,J., Huerta-Cepas,J., Burguet,J., Vaquerizas,J.M., Conde,L., Minguez,P., Vera,J., Mukherjee,S. *et al.* (2006) Next station in microarray data analysis: GEPAS. *Nucleic Acids Res.*, **34**, W486–W491.