W510–W515 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl329

JProGO: a novel tool for the functional interpretation of prokaryotic microarray data using Gene Ontology information

Maurice Scheer[1,3], Frank Klawonn[3], Richard Münch[1], Andreas Grote[1,2], Karsten Hiller[1], Claudia Choi[1], Ina Koch[4], Max Schobert[1], Elisabeth Härtig[1], Ulrich Klages[3] and Dieter Jahn[1,*]

[1]Institute of Microbiology, Technische Universität Braunschweig, Spielmannstrasse 7, Braunschweig, 38106, Germany, [2]Institute of Biochemical Engineering, Technische Universität Braunschweig, Gaussstrasse 17, Braunschweig, 38106, Germany, [3]Department of Computer Science, Fachhochschule Wolfenbüttel, Am Exer 2, Wolfenbüttel, 38302, Germany and [4]Department of Bioinformatics, Technische Fachhochschule Berlin, Seestrasse 64, Berlin, 13347, Germany

Received February 14, 2006; Revised March 15, 2006; Accepted April 14, 2006

ABSTRACT

A novel program suite was implemented for the functional interpretation of high-throughput gene expression data based on the identification of Gene Ontology (GO) nodes. The focus of the analysis lies on the interpretation of microarray data from prokaryotes. The three well established statistical methods of the threshold value-based Fisher's exact test, as well as the threshold value-independent Kolmogorov–Smirnov and Student's *t*-test were employed in order to identify the groups of genes with a significantly altered expression profile. Furthermore, we provide the application of the rank-based unpaired Wilcoxon's test for a GO-based microarray data interpretation. Further features of the program include recognition of the alternative gene names and the correction for multiple testing. Obtained results are visualized interactively both as a table and as a GO subgraph including all significant nodes. Currently, JProGO enables the analysis of microarray data from more than 20 different prokaryotic species, including all important model organisms, and thus constitutes a useful web service for the microbial research community. JProGO is freely accessible via the web at the following address: http://www.jprogo.de

INTRODUCTION

High-throughput technologies, such as DNA microarray-based gene expression profiling provide valuable information on the transcription of all genes of a single microorganism. Usually, the first step for the interpretation of a microarray experiment is the statistical processing of the huge amount of data, including background substraction and normalization. This procedure usually generates long lists of interesting genes, which were found to be differentially expressed between the two conditions tested. The challenging biological interpretation of these results follows. For the automatization of this process on the basis of the available functional information about the proteins encoded by the genes of interest, a sophisticated and well structured functional classification system is required. Gene Ontology (GO) (1), which is based on a solid ontology and is applicable for every organism, fits these needs. It is organized as a directed acyclic graph (DAG) whose nodes represent the strict vocabulary that describes key biological functions and processes.

Here, we present JProGO (**J**ava Tool for the Functional Analysis of **Pro**karyotic Microarray Data using the **G**ene **O**ntology), which allows for the functional interpretation of gene expression data based on GO. For related purposes some programs and algorithms have already been developed (2–14). In general, these tools perform a thorough analysis based on a threshold value dependent or independent method, representing the obtained results as a table, denormalized tree, or as a subgraph of GO.

*To whom correspondence should be addressed. Tel: +49 531 391 5801; Fax: +49 531 391 5854; Email: d.jahn@tu-bs.de

Table 1. List of all 23 prokaryotic species currently available in JProGO together with the number of genes

Organism	Gene number
Bacillus cereus (strain ATCC 10987)	5603
Bacillus subtilis (strain 168)	4101
Caulobacter crescentus (strain CB15)	3737
Clostridium tetani (strain Massachusetts)	2373
Corynebacterium glutamicum (strain DSM 20300 [Nakagawa])	3099
Escherichia coli (strain K12)	4291
Helicobacter pylori (strain ATCC 700392)	1580
Listeria innocua (serovar 6a, strain CLIP 11262)	2981
Listeria monocytogenes (serovar 1/2a, strain EGD-e)	2855
Methanococcus jannaschii (strain JAL-1)	1770
Mycobacterium tuberculosis (strain H37Rv)	3918
Mycobacterium tuberculosis (strain Oshkosh)	4187
Mycoplasma genitalium (G-37)	480
Mycoplasma pneumoniae (strain M129)	688
Pseudomonas aeruginosa (strain PAO1)	5573
Pseudomonas putida (strain KT2440)	5351
Rickettsia conorii (strain Malish 7)	1374
Rickettsia prowazekii (strain Madrid E)	834
Salmonella typhimurium (strain ATCC 700720)	4412
Staphylococcus aureus (strain N315)	2595
Streptococcus pneumoniae (strain TIGR4)	2094
Streptococcus pyogenes (serovar M1, strain SF370)	1697
Yersinia pestis (biovar Mediaevalis, strain KIM5)	4090

Only chromosomal protein-coding genes are considered; data were imported from PRODORIC database (January 2006) (19).

JProGO offers an integrative platform for the majority of statistical methods commonly used for the determination of the significant GO nodes. They calculate the probability that the expression profile of each particular GO node deviates randomly from its environment (alpha error, *P*-value).

The user can choose between four well established methods comprising the threshold value-based Fisher's exact test (2,12) as well as the threshold value independent Kolmogorov–Smirnov test (4), Student's *t*-test (5) and unpaired Wilcoxon's test (3). For the Fisher's exact test we offer both two-sided and one-sided versions. The latter are identical to the hypergeometric test as implemented by related analysis tools (6,13).

Amongst the variety of different methods of analysis, JProGO especially facilitates the analysis of microarray data for a broad range of prokaryotic species (more than 20 species), including all important model organisms (see Table 1).

METHODS AND IMPLEMENTATION

Main program and statistical methods

The main program was written in Java (http://java.sun.com). It manages the invocation of the different algorithms that were implemented for the identification of GO nodes that significantly differ in their gene expression profile from their environment. For the algorithms based on Fisher's exact test, Kolmogorov–Smirnov test, Student's *t*-test and unpaired Wilcoxon's test, the *P*-value calculation is performed with the free statistical language R (http://www.r-project.org/) (15). In this setup, R runs in a server mode using Rserve. The Java progam sends requests to and obtains results from R via TCP/IP using Java-based JRClient classes. Rserve plugin for R and JRClient are both freely available from http://stats.math.uni-augsburg.de/Rserve. Statistical testing is restricted to GO nodes to which a sufficient number of genes is assigned (in this case at least four genes the expression of which has been measured). Correction for multiple testing is per default done by the false discovery rate (FDR) method (16). In addition, the more conservative Bonferroni method was implemented (17).

For the construction of the GO graph, recent versions of GO (December 2005) (1) and GO Annotation (November 2005) (18) were imported into the relational database PRODORIC (19), which contains information on the genes and the gene products of all publicly available sequenced prokaryotic genomes. PRODORIC was then used to calculate organism-specific GO graphs in which each gene of the considered organism was annotated to the most detailed GO node available and—due to the true path rule—to all its parent nodes. These organism specifically annotated GO graphs were then stored as XML files in order to enable their fast subsequent reconstruction for the web service. For that purpose the Sax Parser is used.

Detection of alternative gene names

For each prokaryotic organism of JProGO, official short names, ORF ID's (genome project numbers) and synonymous gene names were obtained from the PRODORIC database (19) and from a recent Uniprot database release (November 2005) (20). To avoid ambiguities, redundant synonyms are excluded from matching. For the recognition of the alternative gene names the following priority order is maintained: The gene names from the microarray datasets are first matched with the official short names. Subsequently ORF ID's and the remaining non-redundant synonyms are taken into account.

Web interface and visualization

For the smooth integration of the Java classes of the main progam into the web interface, we employed Java Servlet Technology. Jakarata Tomcat runs as the servlet container and web server on a Linux platform (http://jakarta.apache.org/tomcat). For visualization of the expression profile of the genes belonging to a specific GO node and those representing the background distribution, the Java package JFreeChart is used (http://jfree.org/jfreechart). For representation of GO subgraphs that contain the significant GO nodes and all their parent nodes up to the root standard graph traversal algorithms, such as breadth first search are used. Subgraphs are visualized with the dot tool (GraphViz package, http://www.graphviz.org).

RESULTS AND DISCUSSION

Rationale of the approach and features of the tool

Our tool enables the automatic interpretation of DNA microarray gene expression data by combining a priori biological knowledge with statistical evaluation of expression profiles of predefined groups of genes. It recognizes alterations in biological functions and processes via the identification of corresponding GO terms that are significantly changed in their gene expression profile under the two conditions compared.

For this purpose we offer a comprehensive range of different statistical methods which are either threshold-based (Fisher's exact test) or threshold independent (Kolmogorov–Smirnov test, Student's *t*-test and unpaired Wilcoxon's test).

Threshold-free methods are more appropriate for the identification of functions and processes involved in cellular adaptation processes, which are characterized by their continuous nature. The major advantage is that the user does not have to define a somewhat arbitrary threshold value, which usually has a strong influence on the outcome of the analysis. With the threshold-value independent Student's *t*-test we have included a parametric test into JProGO. In contrast, the other two threshold independent tests, the Kolmogorov–Smirnov test and the unpaired Wilcoxon's test, are non-parametric and therefore do not assume a normal distribution for the expression profile of each tested GO node.

Since for each GO node a separate statistical test has to be performed, a multiple testing problem arises. Therefore, we have included the control of the FDR (16), which probably belongs to the most accepted methods of correction. One advantage of the FDR method in comparison to family-wise error rate (FWER) controlling methods, such as Bonferroni correction (17) is its greater statistical power.

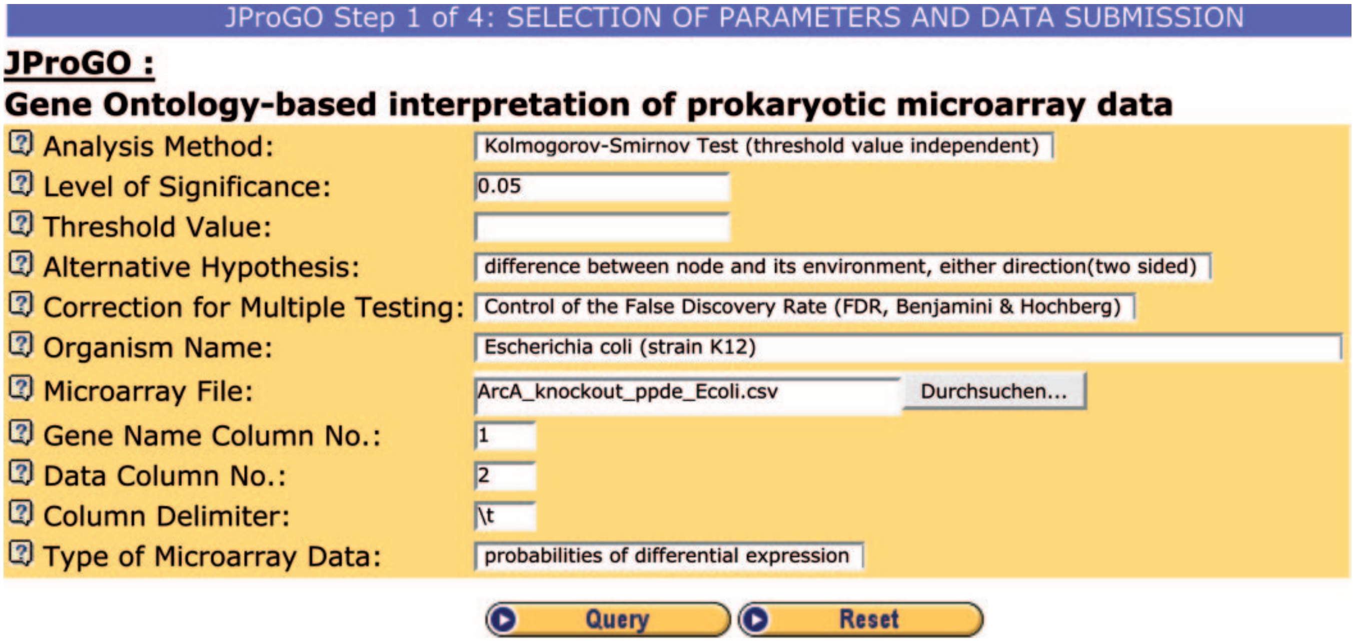
Another advantage of our tool is that it supports the user-friendly immediate functional evaluation of expression data from more than 20 prokaroytic species (see Table 1). Currently available tools mainly focus on the analysis of microarray data from eukaryotes, centering on model organisms, such as man, mouse, rat, *Arabidopsis* or yeast. For an analysis with the JProGO program the user only has to submit the expression data. The input of microarray data both as expression ratios and as probabilities of differential expression is supported. No additional effort for the generation of an assignment between the GO nodes and corresponding genes or the mapping to alternative gene names is required. We provide an automated recognition of alternative gene names.

An important feature of JProGO is that it visualizes the obtained results both as a table with numerous sorting and filter functions and as a subgraph of GO, containing all significant GO nodes and their paths up to the root node (see Table 1 and Figure 3). The distribution of expression values of all genes assigned to a selected GO node and that of the background distribution can be visualized. A list of genes assigned to a particular GO node of interest can be obtained together with the corresponding expression values. Links to PRODORIC database allow direct access to in-depth information on the selected genes.

Use of the web interface

Input. In order to start the functional interpretation of prokaryotic gene expression data, the microarray data are uploaded (Figures 1 and 2). Expression data should be saved in a text file that contains in each row the short name of a particular gene and its corresponding value. After upload of the data, the prokaryotic organism from which the data have been derived is specified. The user has to state the type of data, either expression ratios or probabilities of differential expression. As further options the method of analysis can be chosen and the significance level should be determined (default 0.05). If a threshold-based method is selected, the threshold value is specified.

After selection of these options all necessary settings are made. Then, the proof of validity of the submitted microarray data starts; this includes checking for correct number formats of the expression values and the matching of gene names (Figure 2). A summary of this check is shown on a second



JProGO Step 1 of 4: SELECTION OF PARAMETERS AND DATA SUBMISSION

JProGO :

Gene Ontology-based interpretation of prokaryotic microarray data

Analysis Method:	Kolmogorov-Smirnov Test (threshold value independent)
Level of Significance:	0.05
Threshold Value:	
Alternative Hypothesis:	difference between node and its environment, either direction(two sided)
Correction for Multiple Testing:	Control of the False Discovery Rate (FDR, Benjamini & Hochberg)
Organism Name:	Escherichia coli (strain K12)
Microarray File:	ArcA_knockout_ppde_Ecoli.csv
Gene Name Column No.:	1
Data Column No.:	2
Column Delimiter:	\t
Type of Microarray Data:	probabilities of differential expression

Figure 1. JProGO web interface data submission form.

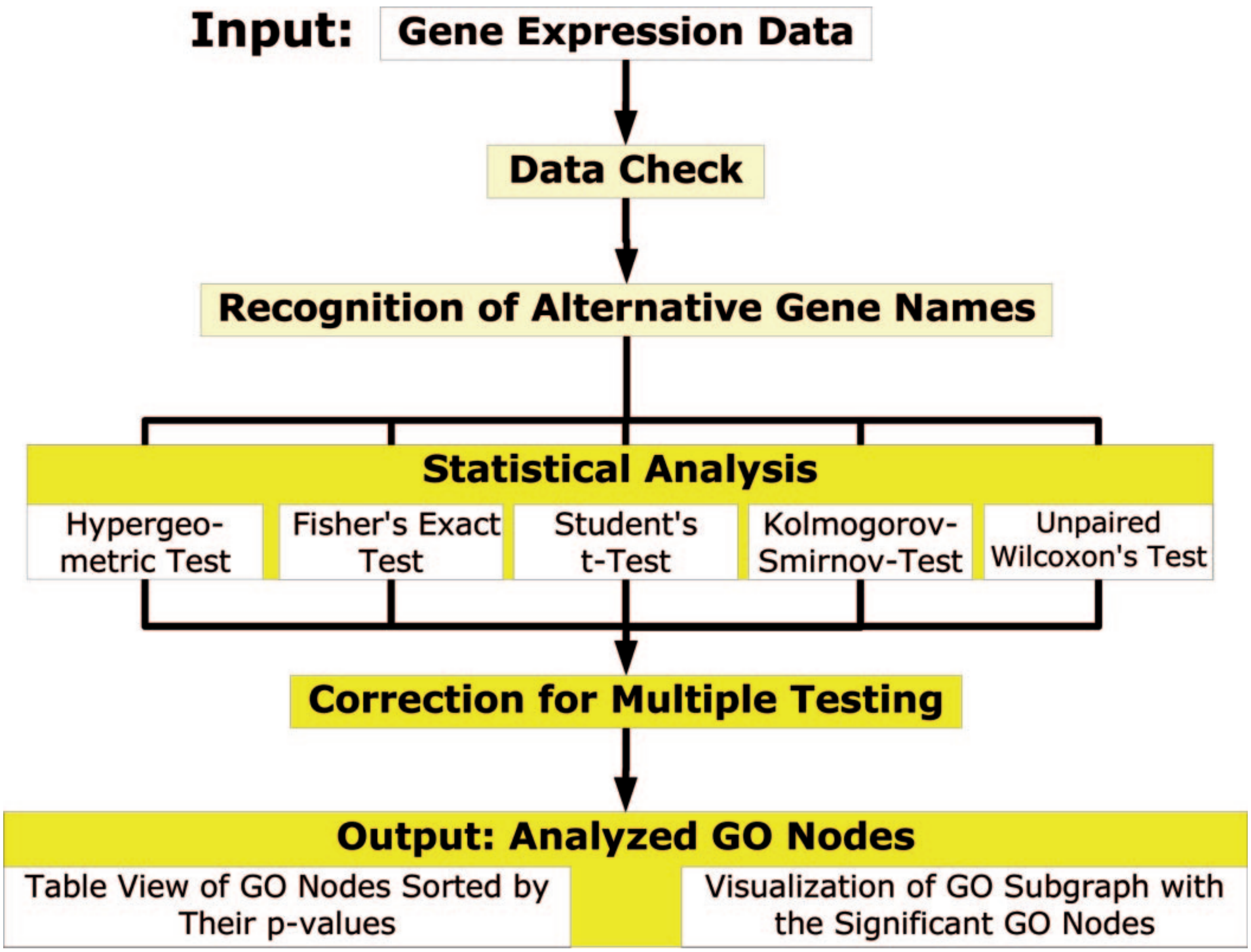

Input:
Gene Expression Data
Data Check
Recognition of Alternative Gene Names
Statistical Analysis
Hypergeo-metric Test
Fisher's Exact Test
Student's t-Test
Kolmogorov-Smirnov-Test
Unpaired Wilcoxon's Test
Correction for Multiple Testing
Output: Analyzed GO Nodes
Table View of GO Nodes Sorted by Their p-values
Visualization of GO Subgraph with the Significant GO Nodes

Figure 2. Workflow of the analysis process of JProGO.

Table 2. Example for the application of JProGO for a microarray dataset from *E.coli* (21) that investigates the influence of ArcA, a global transcriptional regulator on genes of aerobic function (compare to Figure 3)

GO category	GO accession nos	GO name	*P*-value
Molecular function	GO:0005215	Transporter activity	1.8845E-6
Biological process	GO:0005996	Monosaccharide metabolism	4.71 97E-6
Biological process	GO:0019318	Hexose metabolism	1.2557E-5
Biological process	GO:0006066	Alcohol metabolism	1.3639E-5
Biological process	GO:0006810	Transport	1.5379E-5
Biological process	GO:0006351	Transcription, DNA-dependent	2.1493E-5
Biological process	GO:0051179	Localization	2.2139E-5
Biological process	GO:0051234	Establishment of localization	2.2139E-5
Biological process	GO:0015980	Energy derivation by oxidation of organic compounds	4.5974E-5
Molecular function	GO:0003676	Nucleic acid binding	4.9689E-5

Table view showing the significant GO nodes with their *P*-values.

web page. The user can now decide whether to proceed with the analysis or to return to the data submission form.

Calculation is done on the fly, so no waiting queue exists but results are immediately available to the user. Altogether a complete analysis takes about 3–5 min, assuming that the web server is not fully loaded.

Output. The analysed GO nodes are displayed as an interactive table. They are sorted by their calculated *P*-values by default, thus biologically most relevant nodes can be found at the top of the table (see Table 2). The user can choose between several sorting and search options, such as searching for GO nodes that contain a specific phrase in their names.

In addition, the user can select nodes the *P*-value of which is above or below a certain value or within a certain range. Displayed GO nodes can be restricted to one or two of the three sub-ontologies: molecular function, biological process and cellular component. GO nodes that meet any of the selected filter criteria are high-lighted in the corresponding table.

In addition to the table view, the high-lighted nodes—by default GO nodes with a significant *P*-value—are visualized as a GO subgraph. The selected GO nodes and all their parents up to the root node are shown. Here, the GO nodes' size and colour reflect a certain *P*-value. Thus, the user can easily recognize those functions and processes which are affected and classify them in the context of the hierarchy of GO. The table can be downloaded as a tab-delimited text file and the relevant GO subgraph as a pdf or png file.

In both, table and GO subgraph view, clicking on a node generates another web page that shows a table with all the genes that are assigned to this node and their corresponding expression values. Each gene offers a hyperlink to PRODORIC database (19) where more detailed functional and regulatory information is provided. In addition, the expression profile of the genes belonging to the GO node and those of the background distribution are visualized as histograms. A guided tour for the JProGO web interface can be found on our web page (http://www.jprogo.de/tour.jsp).

EXAMPLE APPLICATION

Table 1 and Figure 3 show the results of a typical JProGO analysis. We investigated a knockout of the global transcriptional regulator ArcA, which regulates the expression of genes in response to changes in oxygen tension in *Escherichia coli* (21). The posterior probabilities of differential expression for the microarray dataset were taken from http://www.jbc.org/cgi/content/full/M414030200/DC1.

For example, threshold-free Wilcoxon's test (two-sided, alpha = 0.05, FDR correction) yielded 10 significant GO nodes, representing oxidative energy derivation (e.g. 'energy derivation by oxidation of organic compounds') and central energy-coupled processes, such as 'hexose metabolism' whose activities are typically altered under anaerobic life conditions (Table 2 and Figure 3).

These results fit well with the biological expectation.

In this example both the FDR correction and the Bonferroni method yield the same results. Supplementary Data and further examples are available on our web page (http://www.jprogo.de/examples.jsp).

CONCLUSION AND OUTLOOK

We provide the user with an integrative program suite which allows the easy and intuitive functional interpretation of mass gene expression data, such as microarray data. For this purpose, the user has the opportunity to choose between threshold-value based or threshold-value independent statistical methods. In addition, we provide the user with a correction method for multiple testing, the well established FDR method (16). We also plan to offer a permutation-based approach, which would be an appealing alternative since permutation-based procedures (22–24) are well suitable for correcting multiple dependent tests, although they require considerably longer running times. In order to integrate this feature, expanding JProGO towards an asynchronous web service would be appropriate.

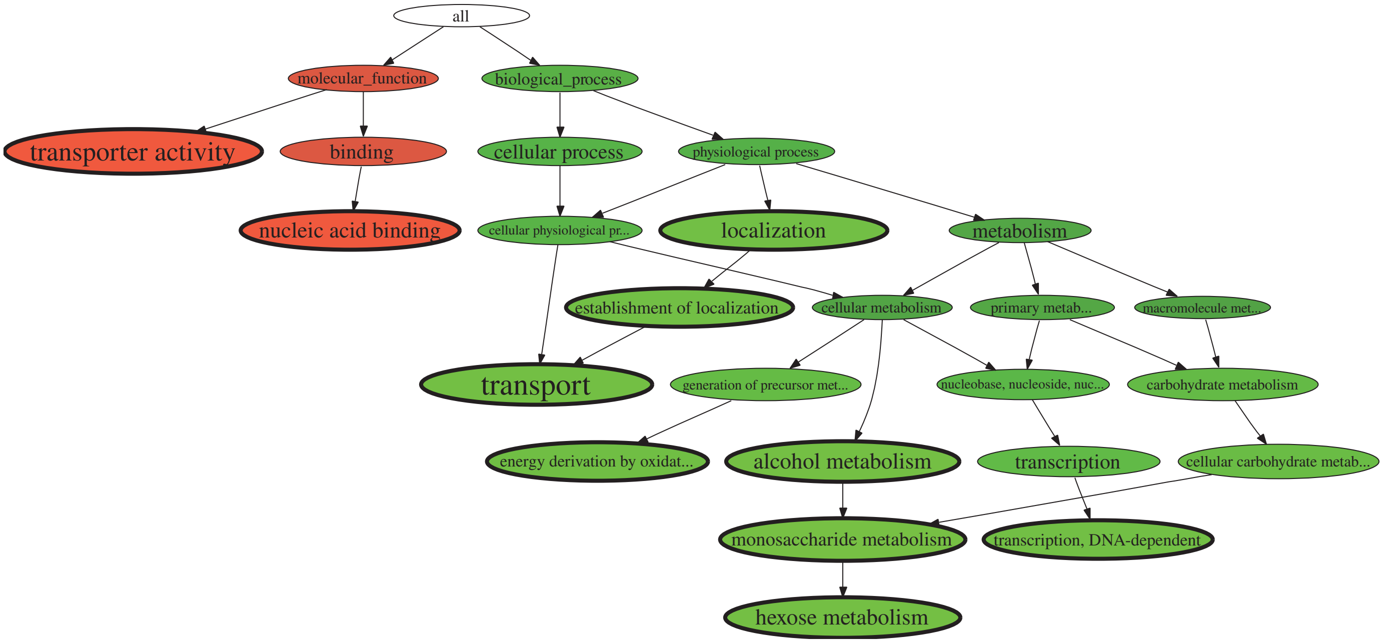


Figure 3. Example for the application of JProGO for a microarray dataset from *E.coli* (21) that investigates the influence of ArcA, a global transcriptional regulator on genes of aerobic function. View of the subgraph induced by the significant GO nodes (thick border, see also Table 1) and the root node ('all'). The node's *P*-value is reflected by its size and brightness whereas nodes with lower *P*-values are larger and brighter. In addition, the node's colour represents its GO sub-ontology (category) which is either molecular function (red), biological process (green) or cellular component (blue).

In summary, JProGO enables a comparative exploratory analysis of the microarray data using the strengths of the selected statistical methods. Because of focussing on well investigated bacterial species, we think the tool is especially suitable for reseachers in the field of prokaryotic gene expression.

In the near future, JProGO will be expanded to support analysis of other species in order to include all sequenced prokaryotes. Moreover, in addition to GO further groupings of genes will be included, such as operons or regulons which comprise genes regulated by the same transcription factor. Since PRODORIC database (19), which is a rich source of operons, regulons and regulatory motifs, is one of the main data sources for JProGO its expansion towards containing such gene groupings can be smoothly achieved and this will constitute a valuable feature for the web service.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Barbara Schulz for critical proofreading of the manuscript and Albert Gevorgyan for sharing his knowledge in Java programming. This work was funded by the German Federal Ministry of Education and Research for the Bioinformatics Competence Center ‘Intergenomics’ (Grant No. 031U210D) and the Sonderforschungsbereich 573 (From Gene to Product). Funding to pay the Open Access publication charges for this article was provided by the Technical University of Braunschweig/Institute of Microbiology.

Conflict of interest statement. None declared.

REFERENCES

1. Ashburner,M., Ball,C., Blake,J., Botstein,D., Butler,H., Cherry,J., Davis,A., Dolinsky,K., Dwight,S., Eppig,J. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nature Genet.*, **79**, 266–270.
2. 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.
3. Barry,W.T., Nobel,A.B. and Wright,F.A. (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, **21**, 1943–1949.
4. Ben-Shaul,Y., Bergman,H. and Soreq,H. (2005) Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression. *Bioinformatics*, **21**, 1129–1137.
5. Boorsma,A., Foat,B.C., Vis,D., Klis,F. and Bussemaker,H.J. (2005) Tprofiler: scoring the activity of predefined groups of genes using gene expression data. *Nucleic Acids Res.*, **34**, W592–W595.
6. Boyle,E.I., Weng,S., Gollub,J., Jin,H., Botstein,D., Cherry,J.M. and Sherlock,G. (2004) GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.
7. Kim,S.Y. and Volsky,D.J. (2005) PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, **6**, 144.
8. Mootha,V.K., Lindgren,C.M., Eriksson,K.F., Subramanian,A., Sihag,S., Lehar,J., Puigserver,P., Carlsson,E., Ridderstrale,M., Laurila,E., Houstis,N. *et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genet.*, **34**, 267–273.
9. Smid,M. and Dorssers,L.C. (2004) GO-Mapper: functional analysis of gene expression data using the expression level as a score to evaluate Gene Ontology terms. *Bioinformatics*, **20**, 2618–2625.
10. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. and Mesirov,J.P. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
11. Volinia,S., Evangelisti,R., Francioso,F., Arcelli,D., Carella,M. and Gasparini,P. (2004) GOAL: automated Gene Ontology analysis of expression profiles. *Nucleic Acids Res.*, **32**, W492–W499.
12. Zeeberg,B.R., Feng,W., Wang,G., Wang,M.D., Fojo,A.T., Sunshine,M., Narasimhan,S., Kane,D.W., Reinhold,W.C., Lababidi,S., Bussey,K.J., Riss,J., Barrett,J.C. and Weinstein,J.N. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, **4**, R28.
13. Zhang,B., Schmoyer,D., Kirov,S. and Snoddy,J. (2004) GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics*, **5**, 16.
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. R Development Core Team (2004) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
16. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, **57**, 289–300.
17. Bland,J.M. and Altman,D.G. (1995) Multiple significance tests: the Bonferroni method. *Biomedical J.*, **310**, 170.
18. Camon,E., Magrane,M., Barrell,D., Binns,D., Fleischmann,W., Kersey,P., Mulder,N., Oinn,T., Maslen,J., Cox,A. and Apweiler,R. (2003) The Gene Ontology annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res.*, **13**, 662–672.
19. Münch,R., Hiller,K., Barg,H., Heldt,D., Linz,S., Wingender,E. and Jahn,D. (2003) PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Res.*, **33**, 266–269.
20. Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.
21. Salmon,K.A., Hung,S.P., Steffen,N.R., Krupp,R., Baldi,P., Hatfield,G.W. and Gunsalus,R.P. (2005) Global gene expression profiling in *Escherichia coli K12*: effects of oxygen availability and ArcA. *J. Biol. Chem.*, **280**, 15084–15096.
22. Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, **29**, 1165–1188.
23. Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.
24. 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.