Published online 7 May 2008 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* ***W347–W351***
doi:10.1093/nar/gkn239

ProfCom: a web tool for profiling the complex functionality of gene groups identified from high-throughput data

Alexey V. Antonov[1], Thorsten Schmidt[2,*], Yu Wang[1] and Hans W. Mewes[1,2]

[1]Helmholtz Zentrum Munich, National Research Center for Environment and Health, Institute for Bioinformatics, Ingolstädter Landstraße 1, D-85764 Neuherberg and [2]Department of Genome-Oriented Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, 85350 Freising, Germany

Received February 1, 2008; Revised April 3, 2008; Accepted April 15, 2008

ABSTRACT

ProfCom is a web-based tool for the functional interpretation of a gene list that was identified to be related by experiments. A trait which makes ProfCom a unique tool is an ability to profile enrichments of not only available Gene Ontology (GO) terms but also of 'complex functions'. A 'Complex function' is constructed as Boolean combination of available GO terms. The complex functions inferred by ProfCom are more specific in comparison to single terms and describe more accurately the functional role of genes. ProfCom provides a user friendly dialog-driven web page submission available for several model organisms and supports most available gene identifiers. In addition, the web service interface allows the submission of any kind of annotation data. ProfCom is freely available at http://webclu.bio.wzw.tum.de/profcom/.

INTRODUCTION

Relating experimental data to biological knowledge is a necessity to cope with the data avalanches emerging from recent developments in high-throughput technologies. Automatic functional profiling has become the *de facto* approach for the secondary analysis of high-throughput data. A number of tools employing available gene functional annotations as well as pathway databases have been developed (1–18). The advantages and limitations of most of these tools are reviewed in ref. (19).

An important aspect of standard functional profiling methodology is inability to overcome the limits of employed annotation vocabularies. Do current annotation vocabularies cover all possible biological functions? Can they cover them in the future? The space of possible biological functions is almost infinite. However, to control it one does not need an infinite number of functional terms. Consider a very direct analogy. Human language contains a limited number of words but through grammar rules these words can be transformed into an almost infinite number of sentences, which allow the expression of almost any idea. In our previous paper (20), we proposed to construct new functional terms (referred to as 'complex functions'). A 'complex function' is constructed as a combination of available terms. The three Boolean operations ('AND', 'OR', 'NOT') play the role of grammar rules and resulting space of 'complex functions' covers an almost infinite number of possible biological functions.

The present article describes ProfCom, a web tool for functional profiling based on the concept described previously (20). ProfCom supports automatic analyses for several model organisms as well as provides a web service interface, which allows the submission of any kind of annotation data. For each organism, ProfCom provides analysis of different annotations, including Gene Ontology (GO) (21), FunCat (22) and InterPro Motifs (23). ProfCom currently offers automatic analyses for *Homo sapiens, Mus musculus, Rattus norvegicus, Caenorhabditis elegans, Drosophila melanogaster* and *Saccharomyces cerevisiae*. In addition, any organism and annotation can be analyzed by ProfCom using Web service interface.

MATERIALS AND METHODS

Statistical analysis and ProfCom profiling engine

A standard tool for automatic functional profiling accepts a query list of genes (referred to as set A, usually the set of genes experimentally identified to be related to the studied biological phenomena) and a reference set (referred to as set B, usually the set of all genes from the analyzed organism). Then, for each attribute f from the set F (f is usually a functional term from the employed annotation vocabulary F, i.e. GO, FunCat, etc.) the number a_f genes

*To whom correspondence should be addressed. Tel: +49 8161 71 21 33; Fax: +49 8161 71 21 86; Email: t.schmidt@wzw.tum.de

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

Table 1. Types of gene identifiers recognized by ProfCom and data sources used for Id mapping

Type of Ids	File used
'Gene Symbol', 'Ensembl', 'LocusTag'	ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene_info.gz
'RefSeq Protein ID', 'RefSeq Transcript ID'	ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2refseq.gz
'UniProt/Swiss-Prot'	ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene_refseq_uniprotkb_collab.gz
'UniGene'	ftp://ftp.ncbi.nlm.nih.go/gene/DAT/gene2unigene
'Affymetrix probe codes'	http://www.affymetrix.com/Annotation files

in set A and b_f genes in set B that have been annotated with f is counted. In the next step, the null hypothesis H_0 (genes that belong to the set A are independent of having attribute f) is tested. Hypergeometric, binomial or χ^2-tests are usually employed to find over/under represented attributes (19).

Unlike most currently available web tools for functional profiling, ProfCom implements different profiling paradigms. Along with standard profiling of functional terms f (referred to as 'base' categories) from annotation vocabularies it also searches for the enrichment related to 'complex functions', which are defined as any Boolean combination of 'base' categories (for example, a new 'complex function' w may define the set of genes that belongs simultaneously to the 'base' categories f_1 and f_2). We consider intersection, union and difference operations. For example, intersection of two categories f_1 and f_2 is formally defined as 'complex function' $w = f_1 \cap f_2$. In other words, w corresponds to the set of genes that belong to both categories f_1 and f_2. The union of two categories f_1 and f_2 is formally defined as $w = f_1 \cup f_2$. In this case, w corresponds to the set of genes that belong either to category f_1 or f_2. The difference between two categories f_1 and f_2 is formally defined as $w = f_1 \backslash f_2$; 'complex function' w corresponds to the set of genes from category f_1 excluding those that simultaneously belong to category f_2.

Each 'complex function' is characterized by the number of base categories required to construct it. We will refer to this characteristic as degree. For example, the base categories can be defined as 'complex functions' of the first degree, the category $w = f_1 \cap f_2$ is a 'complex function' of the second degree (intersection).

Consideration of all possible 'complex functions' leads to combinatorial complexity. To analyze enrichments for all possible combinations of degree higher than 2 is computationally infeasible. For this reason, a search algorithm should be used. ProfCom employs the algorithm based on greedy heuristics (20). Greedy heuristics does not guarantee to find the optimal solution in every case but significantly reduce the computational complexity. To adjust P-values for multiple testing ProfCom uses the Monte–Carlo simulation approach. The estimated P-value corresponds exactly to the definition of an experiment-wise Westfall and Young P-value (3,20,24). More details on the searching algorithm and P-value adjustment can be found in Supplementary Materials.

Automatically supported annotations and gene Ids

As input ProfCom accepts several types of gene or protein identifiers. For example, for the human genome ProfCom supports identifiers from 'Entrez Gene' (25), 'UniProt/Swiss-Prot', 'Gene Symbol' (25,26), 'UniGene' (25), 'Ensembl' (27), 'RefSeq Protein ID', 'RefSeq Transcript ID' (28) and 'Affymetrix probe codes' (29). Additionally, a mixture of several identifier types is possible.

Table 2. Data file used by ProfCom to automatically retrieve annotations

Annotation	File used
Gene Ontology	ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go
InterPro Motifs	ftp://ftp.ebi.ac.uk/pub/databases/interpro/protein2ipr.dat
FunCat	http://mips.gsf.de/

In the first step, user-supplied gene Ids are mapped to 'Entrez Gene' identifiers. For this purpose, files from NCBI and Affymetrix websites are used. Detailed information on data sources used by ProfCom is in Table 1.

The user gets full information on mapping of the supplied gene ids. It includes four tables along with the ProfCom results online. Table 1 reports full mapping details of recognized gene Ids. It includes the informational source used as well as a possible multiple mapping of the user supplied Ids to 'Entrez Gene' Ids. Table 2 reports unrecognized gene Ids. Table 3 reports the final mapping (one-to-one mapping), which is used in subsequent analyses. ProfCom implements simple heuristics to resolve multiple mapping issues. If it is possible to map a particular gene Id to several 'Entrez Gene' Ids, the Id which has the most abundant annotation is selected. However, if the user finds this mapping to be incorrect (Table 3) he/she can simply resubmit the data by substituting those ambiguous gene Ids with 'Entrez Gene' Ids considered to be correct. On the other hand, if several supplied gene Ids are mapped to the same 'Entrez Gene' Id then they are considered as belonging to one gene and the Ids are reported concatenated together by a semicolon (';'). Table 4 reports all such cases.

We would like to point out that protein and gene identifiers can be highly ambiguous (30) with multiple synonymous variants. For this reason, the quality of the retrieved annotation can be different for different types of identifiers. Several powerful recourses to map different type of gene Ids exist (http://beta.uniprot.org/). To escape multiple mapping issues, we recommend submitting 'Entrez Gene' identifies to ProfCom.

ProfCom automatically supports several annotations. Currently, they include GO (21), FunCat (22) and InterPro Motifs (23). Detailed information on data sources used to retrieve each annotation is presented in

N	Simulation Distribution, P-value	Hypergeometric Distribution, P-value	SET A statistics	SET A size	SET B statistics	SET B size	Category Code	Category Name	Genes
1	0.1	0.00056	5	44	278	17615	GO:0007267	cell-cell signaling(BP)	genes
2	0.04	0.00016	8	44	676	17615	GO:0005576	extracellular region(CC)	genes
3	0.13	0.0009	8	44	894	17615	GO:0005509	calcium ion binding(MF)	genes

Figure 1. ProfCom output table 'Top enriched categories of degree 1' for the considered example.

the Table 2. The ProfCom web interface allows the user to use all annotations simultaneously or combine them.

In addition to the interactive web-submissions, custom annotation data can be analyzed using the ProfCom Web service. This allows the use of ProfCom for almost any problem domain, e.g. different annotation types or organisms. Furthermore, web services enable one to run ProfCom analyses in pipelines or automated workflows from most systems. This ensures a fast and convenient usage for a broad range of use cases: starting from a quick hypothesis evaluation to detailed high-quality annotations.

Implementation

ProfCom runs on a standard Apache/Tomcat web server. The actual profiling algorithm is implemented in Java and C for platform independence and high performance. The computation is distributed on Linux workstations utilizing a Sun Grid engine and thus ensures scalability. A ProfCom analysis starts by user-friendly dialog-driven web form. In the first step, the model organism is chosen and the list of gene or protein names of interest is uploaded. Optionally, the reference set of genes can be uploaded. By default, the set of all annotated genes ('Entrez Gene' Ids) from the chosen organism is used as the reference set. Depending on the chosen organism the ProfCom web page automatically shows all available annotations.

Illustration of ProfCom model inference process

Here, we present one example of analyses of real data by ProfCom to illustrate it novelties and utilities in comparison to existing related tools. More examples can be found in Supplementary Materials, where we bring together several independent studies that performed gene expression analyses to identify over/under expressed genes in different cancer types. We collect a set of differentially expressed genes originally identified in each study (we refer to each of these sets as set A and the set of all human genes is referred to as set B).

In ref. (31), microarray experiments were done to compare gene expression in 50 ovarian cancer specimens, including all four histotypes to gene expression in five pools of normal ovarian surface epithelial cells. Data were analyzed to determine whether changes in gene expression correlated with different histotypes, grade or stage.

Several set of genes that show the greatest ability to differentiate between considered cancer subtypes were originally identified. For example, 47 selected genes were 2-fold differentially expressed in mucinous ovarian cancers compared to other histotypes and with normal ovarian surface epithelial cells. Standard functional profiling reveals several GO term significantly overrepresented. It is widely known that the processes of Ca^{++} homeostasis are often disordered in many cancer types (32). Therefore, the presence of GO term '*calcium-ion binding*' among top enriched GO terms is of particular interest. Eight genes (MRC1, EFHD2, PLS1, ANXA10, LDLR, MMP1, S100P, THBS2) from the set A are related by this term (Figure 1). On the other hand, there are 894 genes in the whole human genome classified as 'calcium-ion binding'. Using conventional GO terms vocabularies, standard profiling procedure is not able to supply evidences that would discriminate these eight genes (from all human 894 'calcium-ion binding') and, thus, to clarify molecular mechanism involved.

The complex function '*calcium-ion binding EXCLUDING integral to membrane EXCLUDING hydrolase activity*' inferred by ProfCom (Figure 2) relates all 'calcium-ion binding' genes from the set A and is more specific in comparison to a single GO term, i.e. only 533 genes (compared to 894) in the human genome are classified by this complex function. It is not only better from statistical viewpoint (equal selectivity with ~1-fold increase in specificity), but also supplies valuable biological information which can be helpful for making biological conclusions about molecular mechanisms involved in the considered cancer type.

CONCLUSION

Automatic functional profiling becomes the *de facto* approach for the secondary analysis of high-throughput data. A number of tools employing available gene functional annotations have been developed. However, most of these tools are limited by available annotation vocabularies and may fail to provide full interpretation of biological relationships in a set of genes involved in complex biological phenomena. Here, we present ProfCom, a web-based tool that implements the new profiling paradigm for the interpretation of functional relations between genes. ProfCom profiling engine employs three logical operations ('AND', 'OR', 'NOT') to provide complex functions that classify more specifically the biological role of a gene group.

As been demonstrated, in many cases, complex functions provide better understanding of molecular mechanisms involved for the phenomena under study. On the

N	Simulation Distribution, P-value	Hypergeometric Distribution, P-value	SET A statistics	SET A size	SET B statistics	SET B size	Category Code	Category Name	Genes
1	0.03	8.1e-05	5	44	179	17615	((((GO:0007267) EXCLUDING (GO:0007275))) EXCLUDING (GO:0016020))	((((cell-cell signaling(BP)) EXCLUDING (multicellular organismal development(BP)))) EXCLUDING (membrane(CC)))	genes
2	0.02	3.5e-05	8	44	533	17615	((((GO:0005509) EXCLUDING (GO:0016021))) EXCLUDING (GO:0016787))	((((calcium ion binding(MF)) EXCLUDING (integral to membrane(CC)))) EXCLUDING (hydrolase activity(MF)))	genes
3	0.05	6.3e-05	7	44	421	17615	((((GO:0005509) EXCLUDING (GO:0016021))) EXCLUDING (GO:0005515))	((((calcium ion binding(MF)) EXCLUDING (integral to membrane(CC)))) EXCLUDING (protein binding(MF)))	genes
4	0.02	4.9e-05	8	44	560	17615	((((GO:0005576) EXCLUDING (GO:0007275))) EXCLUDING (GO:0005737))	((((extracellular region(CC)) EXCLUDING (multicellular organismal development(BP)))) EXCLUDING (cytoplasm(CC)))	genes

Figure 2. ProfCom output table 'Top enriched categories of degree 3' for the considered example.

other hand, in some cases, relative GO terms can form many redundant complex functions and may complicate the manual analyses of the ProfCom results. This may be considered as a potential disadvantage. One potential way to resolve redundancy problem is the inclusion of methodologies that group related sets of annotations before the analyses (18,33,34), in the future.

ProfCom provides technical support to the user that corresponds to the best currently available standards in the field. It has a dialog-driven web page for submission that covers several mostly exploited model organisms. In addition, the web service interface allows one submitting any kind of annotation data and is not limited to a particular organism or problem domain. This property significantly simplifies the procedure of data analyses and increases the spectrum of gene sets that can be analyzed. These features make ProfCom an attractive practical tool for biologists interpreting new experimental data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Ulrich Gueldener, Philip Wong for helpful discussions and Michael Strasser for initial technical help at the project beginning. T.S. was supported by the DFG Program 'Bioinformatics Initiative Munich'. Funding to pay the Open Access publication charges for this article was provided by Helmholtz Zentrum Munich, German Research Center for Environmental Health.

Conflict of interest statement. None declared.

REFERENCES

1. Antonov,A.V. and Mewes,H.W. (2006) BIOREL: the benchmark resource to estimate the relevance of the gene networks. *FEBS Lett*., **580**, 844–848.
2. Antonov,A.V., Tetko,I.V. and Mewes,H.W. (2006) A systematic approach to infer biological relevance and biases of gene network structures. *Nucleic Acids Res*., **34**, e6.
3. Berriz,G.F., King,O.D., Bryant,B., Sander,C. and Roth,F.P. (2003) Characterizing gene sets with FuncAssociate. *Bioinformatics*, **19**, 2502–2504.
4. Carmona-Saez,P., Chagoyen,M., Tirado,F., Carazo,J.M. and Pascual-Montano,A. (2007) GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol*., **8**, R3.
5. Draghici,S., Khatri,P., Tarca,A.L., Amin,K., Done,A., Voichita,C., Georgescu,C. and Romero,R. (2007) A systems biology approach for pathway level analysis. *Genome Res*., **17**, 1537–1545.
6. Khatri,P., Draghici,S., Ostermeier,G.C. and Krawetz,S.A. (2002) Profiling gene expression using onto-express. *Genomics*, **79**, 266–270.
7. Khatri,P., Bhavsar,P., Bawa,G. and Draghici,S. (2004) Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments. *Nucleic Acids Res*., **32**, W449–W456.
8. Khatri,P., Voichita,C., Kattan,K., Ansari,N., Khatri,A., Georgescu,C., Tarca,A.L. and Draghici,S. (2007) Onto-Tools: new additions and improvements in 2006. *Nucleic Acids Res*., **35**, W206–W211.
9. Martin,D., Brun,C., Remy,E., Mouren,P., Thieffry,D. and Jacq,B. (2004) GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol*., **5**, R101.

10. Masseroli,M., Martucci,D. and Pinciroli,F. (2004) GFINDer: genome function integrated discoverer through dynamic annotation, statistical analysis, and mining. *Nucleic Acids Res.*, **32**, W293–W300.
11. 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. *et al.* (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, **4**, R28.
12. Zhang,B., Kirov,S. and Snoddy,J. (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.*, **33**, W741–W748.
13. Al-Shahrour,F., az-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.
14. 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.
15. 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.
16. 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.
17. Goffard,N. and Weiller,G. (2007) PathExpress: a web-based tool to identify relevant pathways in gene expression data. *Nucleic Acids Res.*, **35**, W176–W181.
18. Reimand,J., Kull,M., Peterson,H., Hansen,J. and Vilo,J. (2007) g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.*, **35**, W193–W200.
19. Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.
20. Antonov,A.V. and Mewes,H.W. (2006) Complex functionality of gene groups identified from high-throughput data. *J. Mol. Biol.*, **363**, 289–296.
21. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
22. Mewes,H.W., Amid,C., Arnold,R., Frishman,D., Guldener,U., Mannhaupt,G., Munsterkotter,M., Pagel,P., Strack,N., Stumpflen,V. *et al.* (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.*, **32**, D41–D44.
23. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.
24. Westfall,P.N. and Young,S.S. (1993) *Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment*. John Wiley & Sons, New York.
25. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2006) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **34**, D173–D180.
26. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **35**, D5–D12.
27. Birney,E., Andrews,D., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cox,T., Cunningham,F., Curwen,V., Cutts,T. *et al.* (2006) Ensembl 2006. *Nucleic Acids Res.*, **34**, D556–D561.
28. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
29. Liu,G., Loraine,A.E., Shigeta,R., Cline,M., Cheng,J., Valmeekam,V., Sun,S., Kulp,D. and Siani-Rose,M.A. (2003) NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res.*, **31**, 82–86.
30. Draghici,S., Sellamuthu,S. and Khatri,P. (2006) Babel's tower revisited: a universal resource for cross-referencing across annotation databases. *Bioinformatics*, **22**, 2934–2939.
31. Marquez,R.T., Baggerly,K.A., Patterson,A.P., Liu,J., Broaddus,R., Frumovitz,M., Atkinson,E.N., Smith,D.I., Hartmann,L., Fishman,D. *et al.* (2005) Patterns of gene expression in different histotypes of epithelial ovarian cancer correlate with those in normal fallopian tube, endometrium, and colon. *Clin. Cancer Res.*, **11**, 6116–6126.
32. Revankar,C.M., Advani,S.H. and Naik,N.R. (2006) Altered Ca2+ homeostasis in polymorphonuclear leukocytes from chronic myeloid leukaemia patients. *Mol. Cancer*, **5**, 65.
33. Alexa,A., Rahnenfuhrer,J. and Lengauer,T. (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, **22**, 1600–1607.
34. Grossmann,S., Bauer,S., Robinson,P.N. and Vingron,M. (2007) Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. *Bioinformatics*, **23**, 3024–3031.