Published online 23 May 2011 *Nucleic Acids Research, 2011, Vol. 39, Web Server issue* ***W323–W327***
doi:10.1093/nar/gkr372

BioProfiling.de: analytical web portal for high-throughput cell biology

Alexey V. Antonov*

Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Institute for Bioinformatics and Systems Biology, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany

Received February 7, 2011; Revised April 21, 2011; Accepted April 29, 2011

ABSTRACT

BioProfiling.de provides a comprehensive analytical toolkit for the interpretation gene/protein lists. As input, BioProfiling.de accepts a gene/protein list. As output, in one submission, the gene list is analyzed by a collection of tools which employs advanced enrichment or network-based statistical frameworks. The gene list is profiled with respect to the most information available regarding gene function, protein interactions, pathway relationships, *in silico* predicted microRNA to gene associations, as well as, information collected by text mining. BioProfiling.de provides a user friendly dialog-driven web interface for several model organisms and supports most available gene identifiers. The web portal is freely available at http://www.BioProfiling.de/gene_list.

INTRODUCTION

The development of high-throughput technologies has a dramatic impact on modern biology. Although being different technically, the experimental output of 'omics' technologies in the majority of cases is reduced to a list of genes/proteins. Genes or proteins that are differentially expressed or co-expressed across varying cellular conditions or have different epigenetic or mutational status are commonly delivered in many biological and clinically related studies. Functional profiling had become the *de facto* standard approach for the analysis of high-throughput data (1). Functional profiling can be generally defined as a statistical procedure to understand functional context of the gene/protein list using prior knowledge of gene properties and interactions (1–5). The most widespread example of functional profiling is enrichment analysis of Gene Ontology (GO) terms (6–10).

Recently, we have introduced several web tools, which employ either an advance enrichment profiling schema [ProfCom (11), GeneSet2MiRNA (12), PLIPS (13), CCancer (14)] or a network-based statistical framework [KEGG spider (15), PPI spider (16), *R* spider (17)] for the interpretation of gene/protein lists based on available prior knowledge stored in public databases. BioProfiling.de provides experimentalists with an efficient interface to these tools: in one submission, the gene list is profiled with respect to the most information available regarding gene function [GO(18)], pathway relations [KEGG database (19), Reactome knowledgebase (20)], protein interactions [IntAct (21)], *in silico* predicted gene to MiRNA associations [GeneSet2MiRNA (12)] and information collected by text mining [PLIPS&CCancer (13,14)].

BioProfiling.de is not only a common interface for the collection of recently developed tools but also a pipeline for the fast implementation of new tools capable of exploring novel biological principles to group genes into functional classes or to associate genes into a global gene network. For example, ProfCom_PROT_MOTIFS is a new tool implemented within BioProfiling.de pipeline. In this case, genes are grouped into functional classes based on amino acid triplet composition of their protein products. ProfCom_PROT_MOTIFS employs the 'ProfCom' statistical framework to identify 'amino acid triplets' and logical combinations of 'amino acid triplets' overrepresented in the submitted gene/protein list.

BioProfiling.de provides a user-friendly dialog-driven web interface and supports most available gene/protein identifiers. BioProfiling.de provides analyses for the six organisms: *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*.

MATERIALS AND METHODS

Statistical frameworks

The prior knowledge about gene/protein function and interactions is commonly reduced to two data models, either grouping genes into classes based on the shared feature (Type 1) or connecting a pair of genes

*To whom correspondence should be addressed. Tel: +49 (0) 89 3187 2788; Fax: +49 (0) 89 3187 3585;
Email: alexey.antonov@helmholtz-muenchen.de

by edges (Type 2). The GO (18) database is an example of Type 1 data, while IntAct (21) database of protein–protein interactions is an example of Type 2 data. BioProfiling.de implements two different statistical frameworks to deal with both types of prior knowledge. The first statistical framework, referred to as ProfCom, is related to the Type 1 data and represents advanced enrichment schema. The second statistical framework, referred to as Global Network, was recently introduced to deal with Type 2 data.

ProfCom

In this case, the prior knowledge represents grouping genes into functional classes (GO terms) or grouping genes based on whether or not they are regulated by the same microRNA. Let us denote each class (i.e. GO term, microRNA, 'amino acid triplet') as f and the set of all available classes as F. In a standard enrichment schema, a query list of genes (referred to as list A) and a reference list (referred to as list B, usually all genes from the genome) are compared. For each class f from the set F, the number a of genes in the list A and the number b of genes in the list B that have been annotated with f are 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.

ProfCom extends the standard enrichment schema by construction 'complex classes', which are Boolean combination of the available classes of F. ProfCom uses two Boolean operations: intersection and difference. For example, intersection (AND operator) of two categories f_1 and f_2 is formally defined by the set of genes that belong to both classes f_1 and f_2. The difference (NOT operator) between two classes f_1 and f_2 is formally defined as the set of genes from f_1 which are not in f_2.

Unlike the standard enrichment schema, which is limited to the set F, ProfCom tests all possible pairwise combinations joined by logical operators AND, NOT from the set F. Next, ProfCom employs the algorithm based on greedy heuristics to search for the most enriched triplet and quadruplet combinations. In the case of triplet and quadruplet combinations, the use of greedy heuristics does not guarantee finding the optimal solution in every case but does significantly reduce the computational complexity. To adjust P-values for multiple testing ProfCom uses both Bonferroni correction and the Monte–Carlo simulation approach.

Global network (spider tools)

In this case, pairwise gene associations of any biological essence are used as prior knowledge in the form of a global gene network (reference gene network). The sub-network inference procedure is based on natural assumptions:

- most genes from the input list are related and
- most genes that are not from the input list are unrelated.

These assumptions can be reformulated as standard optimization principle:

- to find a gene sub-network with maximal number of input genes connected by a minimal number of missing genes (genes that are not from the input list).

To realize this optimization principle, a network inference algorithm was recently proposed (15–17). A parameter m is introduced which fixes the maximal number of missing genes between any two input genes to be connected by edge in the output network model. The model is inferred in three steps by fixing m to be 0, 1, 2. At each step, any two input genes are connected by edge if they have less then or equal to m genes in between with respect to the reference gene network. At each step ($m = 0$, 1, 2), a connected sub-network component with maximal number of input genes is inferred and referred to as model D1, D2, D3, accordingly. It is clear that given a reference network and any input gene list (even randomly generated gene list), some genes from the input list might be connected into sub-network just by chance, in particular, when parameter m is equal 2. All spider tools implement robust statistical framework to estimate P-value of the inferred models. More details can be found in the original publications (15–17,22).

BioProfiling.de tools

BioProfiling.de provides a common interface for the collection of recently developed tools. The summary of currently available tools is presented in Table 1. Description and details of the tools can be found in original publications. Here, we provide a short description of the novel (recently unpublished) tools implemented within the BioProfiling.de analytical pipeline.

ProfCom PROT_MOTIFS

ProfCom PROT_MOTIFS implements the 'ProfCom' statistical framework to identify amino acid triplets or logical combinations of 'amino acid triplets' overrepresented in the submitted list (genes are mapped to corresponding proteins). In the case, every 'amino acid triplet' represent a functional class (equivalent to GO category) and genes are grouped into the same class if the corresponding protein(s) have the same 'amino acid triplet'. Single, pair, triplet or quadruplet combinations of amino acid triplets are considered (joined by 'AND', 'NOT' logical operators) and the ones which mostly discriminate the input list from all other genes are identified.

CCancer spider

CCancer spider implements the 'Global Network' statistical framework to analyze gene list using as reference knowledge the global gene association network derived from CCancer&PLIPS database. In total, CCancer&PLIPS database has 5238 gene/protein lists reported in various functional context by independent studies. For each gene pair, the number of times *k12* they are reported together (in the same gene/protein list)

Table 1. A Summary of currently available BioProfiling.de tools for the interpretation of gene/protein list

Tool name	Statistical framework	Database (prior knowledge)
ProfCom_GO	ProfCom	GO
ProfCom_InerPro	ProfCom	InterPro database
ProfCom_GO_not_IEA	ProfCom	GO
KEGG spider	Global Network	KEGG
PPI spider	Global Network	IntAct
GeneSet2MiRNA	ProfCom	*In silico* predicted gene to MiRNA regulatory relations
R spider	Global Network	Reactome and KEGG
CCancer&PLIPS[a]	Standard Enrichment	CCancer and PLIPS databases
ProfCom_PROT_MOTIFS	ProfCom	Protein sequences (amino acid triplets)
CCancer spider[a]	Global Network	CCancer and PLIPS databases

[a]Available only for human genome.

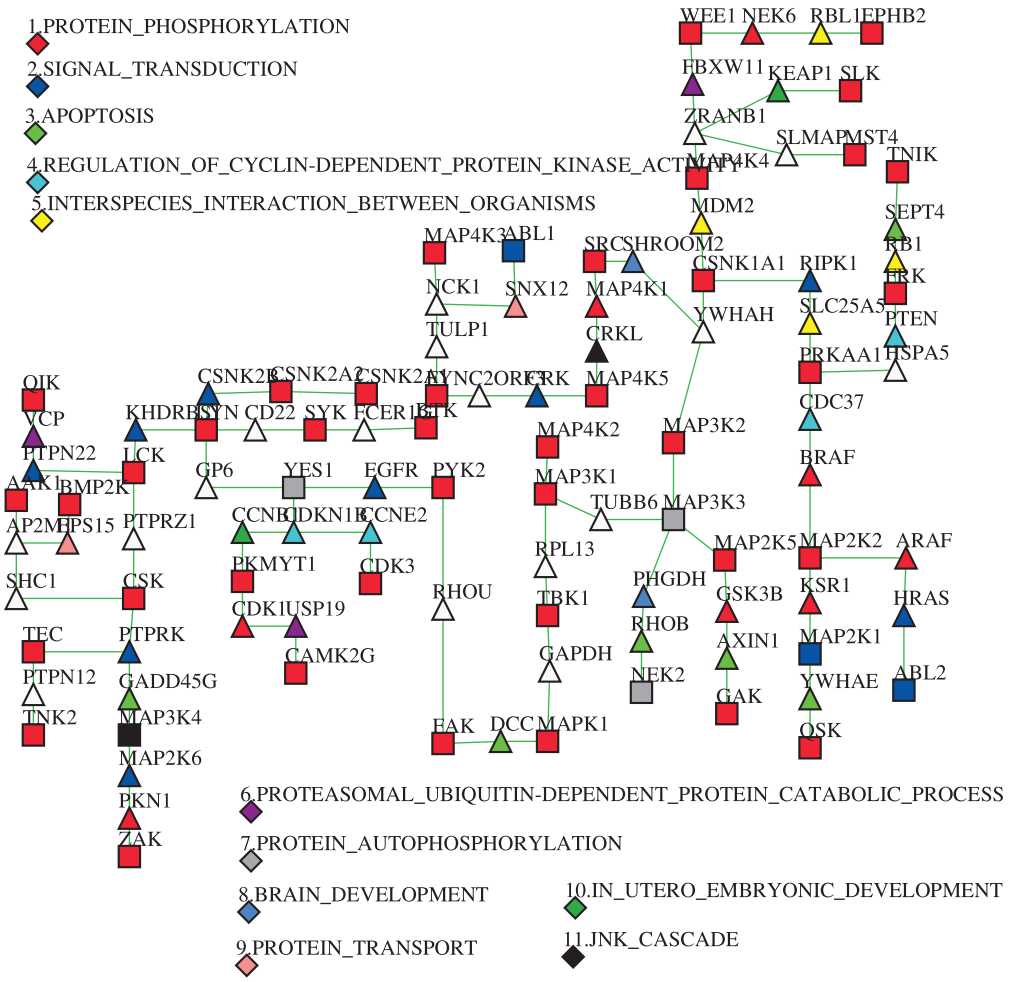

Figure 1. According to the global PPI network, all **47** Bosutinib targets (rectangles), which can be mapped to the global PPI network can be connected into sub-network with maximum two missing genes (triangles) in between. The *P*-value estimated by Monte–Carlo simulation is < 0.005.

is counted, as well as, the number of times each gene is reported alone (*k1*, *k2*). The standard urn schema is used to derive significantly associated gene pairs. Let us denote the total number of gene/protein lists in CCancer&PLIPS database as *N* (5 238 at the moment). The value *k12* follows a hypergeometric distribution with parameters *N*, *k1* and *k2* (*k1* balls were drawn without replacement from an urn containing '***N***' balls in total, *k2* of which are white). The *P*-value need to be adjusted for multiple testing (each gene is tested versus all other genes). Bonferroni correction for multiple testing is used. Two genes are connected by edge in resulting global gene network used by CCancer spider if the significance of their association is <0.01.

RESULTS

BioProfiling.de (http://www.BioProfiling.de/gene_list) is a freely available analytical web portal, which provides a comprehensive analytical toolkit for the interpretation gene/protein lists. In one submission, the gene list is analyzed by a collection of tools. BioProfiling.de has a simple user-friendly interface. As input, it accepts several types of gene or protein identifiers, such as 'Entrez Gene', 'Gene Symbols', 'UniProt/Swiss-Prot' (23), 'IPI - International Protein Index', 'UniGene', 'Ensembl' and 'RefSeq'.

Data submission

To start the analyses, the user needs to upload a text file with gene/protein identifiers and select an organism. After data submission, a link is provided to the 'Main Result page'. As soon as computations are finished, the results will be available there. The user can either bookmark this page and return to it in 2–3 h or periodically refresh it.

The submitted gene/protein Ids are automatically mapped to the 'Entrez Gene' ids. Gene Id mapping is an inherently difficult problem. To escape errors in results related to mapping issues, we recommend submitting 'Entrez Gene' identifiers. We also suggest several resources (6,24,25), which primarily concentrated to solve Gene Id mapping problem.

The mapping report is provided first. If the number of recognized gene/protein ids is less than 10 then the user will get an error message. Next, the table with a short description of the tools available for the submission is provided. Each line of the table corresponds to one tool. The first column of the table specifies the tool name, the second provides the status of the computations (or a link to the results of the tool, in the case the computations are finished). The third column provides a short summary of the tool: the statistical framework, the database of prior knowledge and the total number of gene covered/annotated in the database for the selected genome.

After the computations are finished, the status 'in progress' is substituted with a link to the tool results (second column of the summary table). The structure of the output is the same for all 'spider' tools as well as for all 'ProfCom' tools. In the case of the 'spider' tools, the main output summarized in the table 'Enriched sub-networks', where the details of the best sub-network models (D1, D2, D3) inferred from the submitted gene list are provided. In the case of the 'ProfCom' tools, the user initially gets a short summary table which reports the top enriched complex classes of degree 0, 1, 2, 3. The last column in the table ('full report') provides links to the detailed reports of the 'complex class' of a given degree.

Example: Bosutinib protein targets

BioProfiling.de provides a comprehensive functional profiling of a gene/protein list from various biological perspectives. The next example aims to demonstrate a wide spectrum of biological insights that one can get by using BioProfiling.de. Bosutinib is a novel drug (promiscuous kinase inhibitor). The whole proteome binding spectra of Bosutinib was identified by chemical proteomics (26), in total 55 proteins were reported to be direct Bosutinib interactors. Here, we used BioProfiling.de to understand properties of Bosutinib protein targets. As one might expect, the list of Bosutinib protein targets was significantly enriched from many functional perspectives. Particularly, interesting are results produced by ProfCom_PROT_MOTIF, a new tool in BioProfiling.de collection. In this case, the logical combinations of amino acid triplets highly discriminative between the list of Bosutinib protein targets and the whole-human proteome are reported. For example, logical pattern '**((DFG and HRD) not (LPY, HEE))**' was present in 50 (out of 55) Bosutinib protein targets while only 305 (out of approximately 25 000) proteins in the whole genome comply with the pattern. The *P*-value of the enrichment adjusted by Bonferroni correction for multiple testing is 1.6e-77. In addition, results by spider tools (PPI spider, R spider) suggest that Bosutinib protein targets form densely interaction pattern. The result supports the novel 'network pharmacology' paradigm (27) in drug discovery: to be effective the drug should target multiple functionally dependent targets.

CONCLUSIONS

BioProfiling.de provides experimentalists a comprehensive toolkit for gene/protein list interpretation. In one submission, the gene list is profiled with respect to the most information available regarding gene function (GO), pathway relations (KEGG database, Reactome knowledgebase), protein interactions (IntAct), *in silico* predicted gene to MiRNA associations (GeneSet2MiRNA), information collected by text mining (PLIPS and CCancer) and protein 'amino acid triplets' composition.

BioProfiling.de implements two statistical frameworks ('ProfCom' and 'Global Network'), which allow fast implementation of new tools capable to explore novel biological principles (as prior knowledge) to group genes into functional classes or to associate genes by edge into global gene network. In the future, the collection of tools is going to expand to cover novel biological principles to profile gene/protein list using either 'ProfCom' or 'Global network' statistical framework.

We also would like to point out that both statistical frameworks ('ProfCom', 'Global Network') are implemented only at BioProfiling.de tools. Although, there are many tools for the functional profiling of gene/protein lists, there several features in both frameworks which make BioProfiling.de distinguishable. Therefore, BioProfiling.de provides a combination of traits that makes it different among other resources available.

FUNDING

This work was supported by the Helmholtz Association ''Impuls und Vernetzungsfonds'' (Systems Biology Alliance). Funding for open access charge: Helmholtz Zentrum München.

Conflict of interest statement. None declared.

REFERENCES

1. Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.
2. Berger,S.I., Posner,J.M. and Ma'ayan,A. (2007) Genes2Networks: connecting lists of gene symbols using mammalian protein interactions databases. *BMC Bioinformatics*, **8**, 372.
3. 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.
4. Reimand,J., Tooming,L., Peterson,H., Adler,P. and Vilo,J. (2008) GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. *Nucleic Acids Res.*, **36(Suppl. 2)**, W452–W459.
5. 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. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
6. 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.
7. 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.
8. Khatri,P., Draghici,S., Ostermeier,G.C. and Krawetz,S.A. (2002) Profiling gene expression using onto-express. *Genomics*, **79**, 266–270.
9. 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.
10. 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.
11. Antonov,A.V., Schmidt,T., Wang,Y. and Mewes,H.W. (2008) ProfCom: a web tool for profiling the complex functionality of gene groups identified from high-throughput data. *Nucleic Acids Res.*, **36**, W347–W351.
12. Antonov,A.V., Dietmann,S., Wong,P., Lutter,D. and Mewes,H.W. (2009) GeneSet2miRNA: finding the signature of cooperative miRNA activities in the gene lists 1. *Nucleic Acids Res.*, **37**, W323–W328.
13. Antonov,A.V., Dietmann,S., Wong,P., Igor,R. and Mewes,H.W. (2009) PLIPS, an automatically collected database of protein lists reported by proteomics studies. *J. Proteome Res.*, **8**, 1193–1197.
14. Dietmann,S., Lee,W., Wong,P., Rodchenkov,I. and Antonov,A.V. (2010) CCancer: a bird's eye view on gene lists reported in cancer-related studies 1. *Nucleic Acids Res.*, **38(Suppl.)**, W118–W123.
15. Antonov,A.V., Dietmann,S. and Mewes,H.W. (2008) KEGG spider: interpretation of genomics data in the context of the global gene metabolic network. *Genome Biol.*, **9**, R179.
16. Antonov,A.V., Dietmann,S., Rodchenkov,I. and Mewes,H.W. (2009) PPI spider: a tool for the interpretation of proteomics data in the context of protein-protein interaction networks. *Proteomics*, **9**, 2740–2749.
17. Antonov,A.V., Schmidt,E.E., Dietmann,S., Krestyaninova,M. and Hermjakob,H. (2010) R spider: a network-based analysis of gene lists by combining signaling and metabolic pathways from Reactome and KEGG databases 2. *Nucleic Acids Res.*, **38(Suppl.)**, W78–W83.
18. 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.
19. Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. and Kanehisa,M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **27**, 29–34.
20. Vastrik,I., D'Eustachio,P., Schmidt,E., Gopinath,G., Croft,D., de Bono,B., Gillespie,M., Jassal,B., Lewis,S., Matthews,L. *et al.* (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.*, **8**, R39.
21. Aranda,B., Achuthan,P., am-Faruque,Y., Armean,I., Bridge,A., Derow,C., Feuermann,M., Ghanbarian,A.T., Kerrien,S., Khadake,J. *et al.* (2010) The IntAct molecular interaction database in 2010 5. *Nucleic Acids Res.*, **38**, D525–D531.
22. Antonov,A.V., Dietmann,S., Wong,P. and Mewes,H.W. (2009) TICL–a web tool for network-based interpretation of compound lists inferred by high-throughput metabolomics. *FEBS J.*, **276**, 2084–2094.
23. Boutet,E., Lieberherr,D., Tognolli,M., Schneider,M. and Bairoch,A. (2007) UniProtKB/Swiss-Prot. *Methods Mol. Biol.*, **406**, 89–112.
24. Berriz,G.F. and Roth,F.P. (2008) The Synergizer service for translating gene, protein and other biological identifiers 3. *Bioinformatics*, **24**, 2272–2273.
25. Baron,D., Bihouee,A., Teusan,R., Dubois,E., Savagner,F., Steenman,M., Houlgatte,R. and Ramstein,G. (2011) MADGene: retrieval and processing of gene identifier lists for the analysis of heterogeneous microarray datasets 2. *Bioinformatics*, **27**, 725–726.
26. Fernbach,N.V., Planyavsky,M., Muller,A., Breitwieser,F.P., Colinge,J., Rix,U. and Bennett,K.L. (2009) Acid elution and one-dimensional shotgun analysis on an Orbitrap mass spectrometer: an application to drug affinity chromatography 2. *J. Proteome Res.*, **8**, 4753–4765.
27. Hopkins,A.L. (2008) Network pharmacology: the next paradigm in drug discovery 3. *Nat. Chem. Biol.*, **4**, 682–690.