W406–W410 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* *Published online 28 April 2008*
doi:10.1093/nar/gkn215

CoPub: a literature-based keyword enrichment tool for microarray data analysis

Raoul Frijters[1], Bart Heupers[2], Pieter van Beek[2], Maurice Bouwhuis[2], René van Schaik[3], Jacob de Vlieg[1,3], Jan Polman[3] and Wynand Alkema[3,*]

[1]Computational Drug Discovery (CDD),, Nijmegen Centre for Molecular Life Sciences (NCMLS), Radboud University Nijmegen Medical Centre, Nijmegen, [2]SARA Computing and Network Services, Amsterdam and [3]Department of Molecular Design & Informatics, Organon, part of Schering-Plough Corporation, Oss, The Netherlands

Received January 30, 2008; Revised April 4, 2008; Accepted April 9, 2008

ABSTRACT

Medline is a rich information source, from which links between genes and keywords describing biological processes, pathways, drugs, pathologies and diseases can be extracted. We developed a publicly available tool called CoPub that uses the information in the Medline database for the biological interpretation of microarray data. CoPub allows batch input of multiple human, mouse or rat genes and produces lists of keywords from several biomedical thesauri that are significantly correlated with the set of input genes. These lists link to Medline abstracts in which the co-occurring input genes and correlated keywords are highlighted. Furthermore, CoPub can graphically visualize differentially expressed genes and over-represented keywords in a network, providing detailed insight in the relationships between genes and keywords, and revealing the most influential genes as highly connected hubs. CoPub is freely accessible at http://services.nbic.nl/cgi-bin/copub/CoPub.pl.

INTRODUCTION

Analysis and interpretation of microarray data is not a trivial task. Many public and commercial bioinformatics tools have been developed to help scientists interpret the lists of differentially expressed genes that are the result of microarray experiments. For example, Gene Ontology and Pathway Mapping tools (1–4) allow batch input of genes and produce lists of GO terms or pathways that are significantly correlated with the input set of genes (5–13).

The outcome of these tools is based on well-established relationships between the genes and biological processes in which they participate. However, the primary literature contains much more information about the functions of genes than is captured in structured vocabularies or canonical pathways. To extract this additional information on gene function from literature, we used thesaurus-based keyword matching in Medline abstracts to link human, mouse and rat genes to biomedical concepts describing liver pathologies, pathways, GO terms, diseases, drugs and tissues (Table 1). This approach builds on the assumption that co-occurrence of a gene and a biomedical concept in the same abstract is an indication of a functional link between the gene and the concept.

In this article, we describe a tool named CoPub that calculates keyword over-representation for a set of regulated genes in a similar fashion to general GO term over-representation tools, but where the over-represented keywords for the gene set are retrieved directly from Medline by text mining. Several text mining methods for the analysis of microarray data have been published that annotate clustered sets of regulated genes based on their literature profile (14–17), or on their expression profile, often based on subsets of the total Medline repository (18–21). CoPub uses the entire Medline library to calculate robust statistics for gene-keyword co-occurrence, and is not dependent on pre-clustered gene sets to calculate significance for keyword over-representation. In addition to calculating over-represented keywords, CoPub also shows the results graphically in an interactive network, providing an additional level of insight into the biological mechanisms related to a set of regulated genes.

CoPub has two other features: the Gene search and the BioConcept search. The Gene search and the BioConcept search options identify genes and keywords that share occurrences in Medline abstracts with a gene or keyword of interest, which provides a kind of annotation for the gene or keyword of interest.

In an earlier study (22), we successfully applied CoPub for compound toxicity evaluation of a variety of compounds, which shows that CoPub is a useful additional

*To whom correspondence should be addressed. Tel: +31 (0)412 663678; Fax: +31 (0)412 662553; Email: wynand.alkema@organon.com

Table 1. Overview of the 11 thesauri that were generated to search Medline

Thesaurus category	Number of keywords	Source	URL
Genes[a]			
Human	122 425 (24 876 genes)	NCBI's Entrez Gene Database	http://www.ncbi.nlm.nih.gov/sites/entrez
Mouse	130 759 (22 593 genes)	"	"
Rat	73 572 (12 296 genes)	"	"
Gene Ontology (GO)			
Biological processes	3621	Gene Ontology Database	http://www.geneontology.org
Molecular functions	961	"	"
Cellular components	216	"	"
Liver pathologies	489	Textbooks	–
Pathways	817	KEGG, Reactome, Encyclopedia of Human Genes and Metabolism DB	http://www.genome.jp/kegg/, http://www.genomeknowledge.org, http://humancyc.org
Diseases	4164	Karolinska Institutet	http://www.mic.ki.se/Diseases/Alphalist.html
Drugs	5796	RxList database	http://www.rxlist.com/top200.htm
Tissues	1112	ExPASy Proteomics server	http://www.expasy.org/cgi-bin/lists?tisslist.txt

[a] Full gene names, gene symbols, alternative gene names/symbols.

bioinformatics tool for microarray data analysis. CoPub is freely accessible at http://services.nbic.nl/cgi-bin/copub/CoPub.pl.

METHODS

Text mining Medline abstracts

Eleven thesauri were generated to search Medline (Table 1). These thesauri describe genes (human, mouse and rat), Gene Ontology terms, diseases, pathways, drugs, tissues and liver pathologies. The keyword thesauri are based on biological items, which represent an instance of a biological concept (e.g. a gene, a pathway), and may contain one or more keywords (e.g. a gene is assigned a full gene name as well as a gene symbol and gene aliases).

The full Medline baseline XML files (1966 to February 2008) were obtained from the NCBI website (http://www.nlm.nih.gov/bsd/licensee/2008_stats/baseline_doc.html) and extracted to small text files containing title, abstract and substances.

Regular expressions were used to search the compiled Medline text files for the presence of all keywords (~250 000) from the biological concept thesauri, as described by Alako *et al.* (23). Keywords that generated a hit in a Medline abstract were stored, together with the PubMed identifiers (IDs) of the Medline records in which the hit occurred. For every biological item, the hits were made non-redundant (note: multiple keywords of a biological item can occur in the same Medline abstract), resulting in a PubMed ID-biological item list. Gene symbols were curated for ambiguity and gene hits of orthologous genes were combined to make the keyword search more comprehensive.

Co-publication of biological items (e.g. a gene with a pathology term) was retrieved from the database by matching common Medline abstract occurrences. For every biological item pair, an *R*-scaled score, which describes the strength of a co-citation between two keywords given their individual frequencies of occurrence (23), and the literature count, which is the number of co-publications between every biological item pair, was calculated. Both measures were used to describe the strength of the relationship between two keywords.

To link gene expression data to literature data, mappings of Affymetrix probe set identifiers to Entrez Gene identifiers and orthology information were retrieved from Affymetrix human, mouse, rat GeneChip Genome Array annotation files (http://www.affymetrix.com). Mappings of Ensembl identifiers to Entrez Gene identifiers were retrieved from BioMart (http://www.biomart.org).

Keyword enrichment calculation

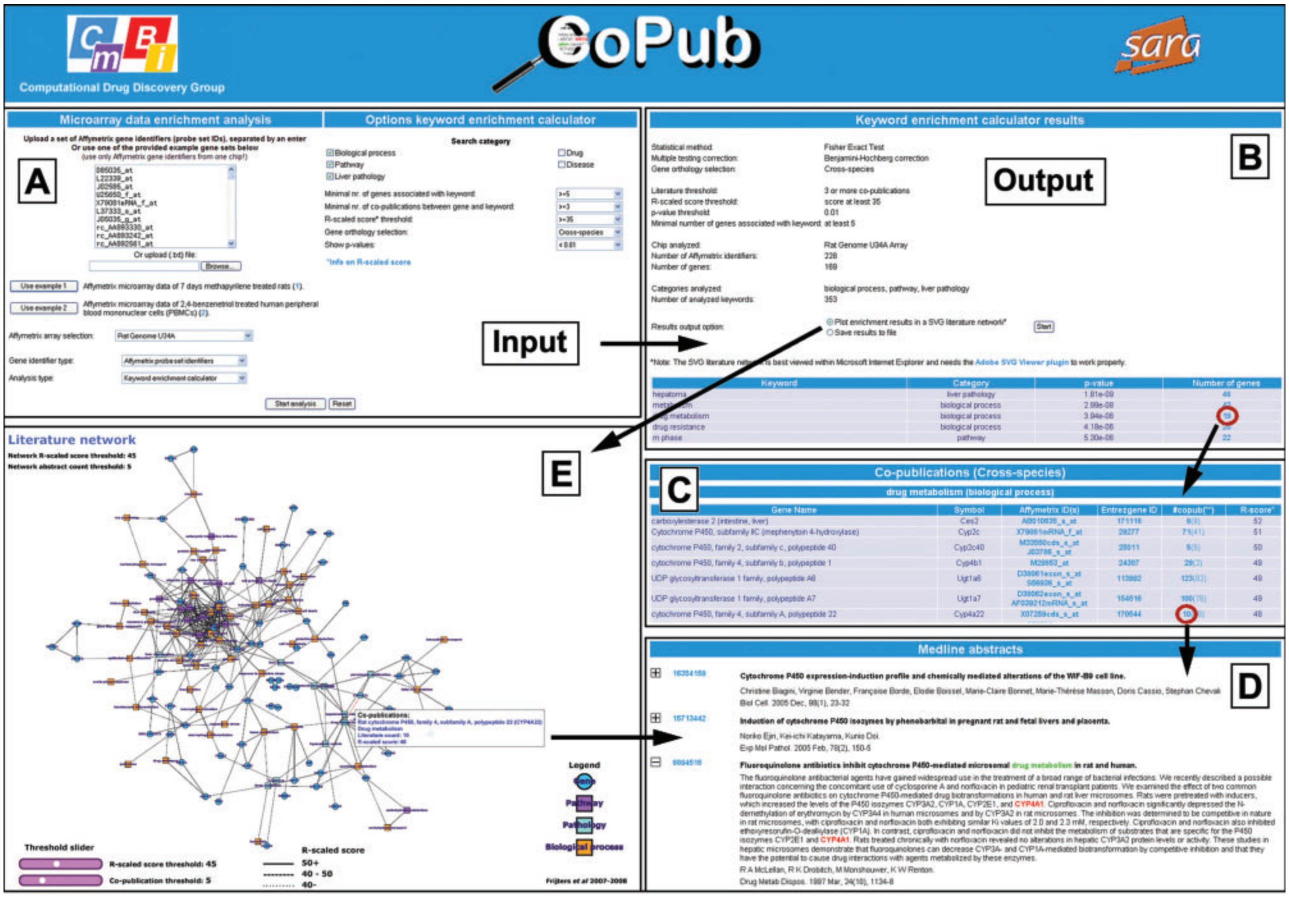
Keyword enrichment calculation is performed using the Fisher exact test, in which the association of a given keyword with a set of regulated genes (i.e. co-publications in Medline abstracts) is statistically tested against a background set; the set of unchanged genes on the microarray in case Affymetrix probe set identifiers are uploaded, or all other genes in the genome when Entrez Gene identifiers or Ensembl identifiers are uploaded. The calculated *P*-values are corrected using the Benjamini–Hochberg multiple testing correction method. All statistical tests are done using the *R* Statistics package (http://www.r-project.org).

To generate literature networks CoPub uses GraphViz (http://www.graphviz.org) to calculate the graph layout (neato), and for exporting the literature network in a scalable vector graphic (SVG) format. Interactivity in generated SVG networks is implemented using Perl and JavaScript.

CoPub WEB SERVER

The CoPub user-interface offers three analysis methods: the Microarray data analysis, the Gene search and the BioConcept search.

The Microarray data analysis option calculates keyword over-representation for a set of differentially expressed genes, and offers graphical visualization of the analysis results in a literature-based network. Screenshots and



Microarray data enrichment analysis

Upload a set of Affymetrix gene identifiers (probe set IDs), separated by an enter
Or use one of the provided example gene sets below
(use only Affymetrix gene identifiers from one chip!)

```
D85035_at
L22339_at
J02585_at
U25650_f_at
X79081mRNA_f_at
L37333_s_at
J05035_g_at
rc_AA893330_at
rc_AA893242_at
rc_AA892561_at
```

Or upload (.txt) file: [Browse...]

[Use example 1] Affymetrix microarray data of 7 days methapyrilene treated rats (1).

[Use example 2] Affymetrix microarray data of 2,4-benzenetriol treated human peripheral blood mononuclear cells (PBMCs) (2).

Affymetrix array selection: Rat Genome U34A

Gene identifier type: Affymetrix probe set identifiers

Analysis type: Keyword enrichment calculator

[Start analysis] [Reset]

Options keyword enrichment calculator

Search category

☑ Biological process ☐ Drug
☑ Pathway ☐ Disease
☑ Liver pathology

Minimal nr. of genes associated with keyword: >=5
Minimal nr. of co-publications between gene and keyword: >=3
R-scaled score* threshold: >=35
Gene orthology selection: Cross-species
Show p-values: < 0.01

*Info on R-scaled score

Keyword enrichment calculator results

Statistical method: Fisher Exact Test
Multiple testing correction: Benjamini-Hochberg correction
Gene orthology selection: Cross-species

Literature threshold: 3 or more co-publications
R-scaled score threshold: score at least 35
p-value threshold: 0.01
Minimal number of genes associated with keyword: at least 5

Chip analyzed: Rat Genome U34A Array
Number of Affymetrix identifiers: 228
Number of genes: 168

Categories analyzed: biological process, pathway, liver pathology
Number of analyzed keywords: 353

Results output option: ● Plot enrichment results in a SVG literature network* ○ Save results to file [Start]

*Note: The SVG literature network is best viewed within Microsoft Internet Explorer and needs the Adobe SVG Viewer plugin to work properly.

Keyword	Category	p-value	Number of genes
hepatoma	liver pathology	1.81e-08	48
metabolism	biological process	2.99e-08	47
drug metabolism	biological process	3.94e-06	19
drug resistance	biological process	4.18e-06	20
m phase	pathway	5.30e-06	22

Co-publications (Cross-species)

drug metabolism (biological process)

Gene Name	Symbol	Affymetrix ID(s)	Entrezgene ID	#copub(**)	R-score*
carboxylesterase 2 (intestine, liver)	Ces2	AB010635_s_at	171118	8(8)	52
Cytochrome P450, subfamily IIC (mephenytoin 4-hydroxylase)	Cyp2c	X79081mRNA_f_at	29277	71(41)	51
cytochrome P450, family 2, subfamily c, polypeptide 40	Cyp2c40	M33550cds_s_at J03786_s_at	29011	6(5)	50
cytochrome P450, family 4, subfamily b, polypeptide 1	Cyp4b1	M29853_at	24307	29(2)	49
UDP glycosyltransferase 1 family, polypeptide A6	Ugt1a6	D38061exon_s_at S56936_s_at	113992	123(82)	49
UDP glycosyltransferase 1 family, polypeptide A7	Ugt1a7	D38062exon_s_at AF039212mRNA_s_at	154516	100(76)	49
cytochrome P450, family 4, subfamily A, polypeptide 22	Cyp4a22	X07259cds_s_at	170644	10([illegible])	48

Medline abstracts

⊞ 16354159 **Cytochrome P450 expression-induction profile and chemically mediated alterations of the WIF-B9 cell line.**
Christine Biagini, Virginie Bender, Françoise Borde, Elodie Boissel, Marie-Claire Bonnet, Marie-Thérèse Masson, Doris Cassio, Stephan Chevalier
Biol Cell. 2006 Dec, 98(1), 23-32

⊞ 15713442 **Induction of cytochrome P450 isozymes by phenobarbital in pregnant rat and fetal livers and placenta.**
Noriko Ejiri, Kei-ichi Katayama, Kunio Doi
Exp Mol Pathol. 2005 Feb, 78(2), 150-5

⊟ 8894516 **Fluoroquinolone antibiotics inhibit cytochrome P450-mediated microsomal drug metabolism in rat and human.**
The fluoroquinolone antibacterial agents have gained widespread use in the treatment of a broad range of bacterial infections. We recently described a possible interaction concerning the concomitant use of cyclosporine A and norfloxacin in pediatric renal transplant patients. We examined the effect of two common fluoroquinolone antibiotics on cytochrome P450-mediated drug biotransformations in human and rat liver microsomes. Rats were pretreated with inducers, which increased the levels of the P450 isozymes CYP3A2, CYP1A, CYP2E1, and CYP4A1. Ciprofloxacin and norfloxacin significantly depressed the N-demethylation of erythromycin by CYP3A4 in human microsomes and by CYP3A2 in rat microsomes. The inhibition was determined to be competitive in nature in rat microsomes, with ciprofloxacin and norfloxacin both exhibiting similar Ki values of 2.0 and 2.3 mM, respectively. Ciprofloxacin and norfloxacin also inhibited ethoxyresorufin-O-dealkylase (CYP1A). In contrast, ciprofloxacin and norfloxacin did not inhibit the metabolism of substrates that are specific for the P450 isozymes CYP2E1 and CYP4A1. Rats treated chronically with norfloxacin revealed no alterations in hepatic CYP3A2 protein levels or activity. These studies in hepatic microsomes demonstrate that fluoroquinolones can decrease CYP3A- and CYP1A-mediated biotransformation by competitive inhibition and that they have the potential to cause drug interactions with agents metabolized by these enzymes.
R A McLellan, R K Drobitch, M Monshouwer, K W Renton
Drug Metab Dispos. 1997 Mar, 24(10), 1134-8

Literature network

Network R-scaled score threshold: 45
Network abstract count threshold: 5

Co-publications:
Rat cytochrome P450, family 4, subfamily A, polypeptide 22 (CYP4A22)
Drug metabolism
Literature count: 10
R-scaled score: 48

Figure 1. Screenshots and workflow of the Microarray data analysis. (**A**) Input screen for uploading gene identifiers (Affymetrix probe set identifiers, Entrez Gene identifiers or Ensembl identifiers), selection of keyword categories and to specify thresholds (e.g. *P*-value significance level), with which the keyword over-representation analysis will be performed (sensible defaults are provided). (**B**) Output screen which reports on significantly linked keywords to the set of submitted genes, ranked on *P*-values after multiple testing correction. The number of genes that are significantly associated with the analyzed keyword, links to an overview of uploaded genes that share co-publications with the analyzed keyword (**C**), which provides access to highlighted Medline abstracts in which they co-occur (**D**). (**E**) Visualization of the keyword over-representation results in an interactive literature network (as SVG), in which nodes represent genes and keywords, and edges represent links in Medline abstracts. Clicking on an edge retrieves highlighted Medline abstracts in which genes and keywords co-occur (D).

workflow of the Microarray data analysis are shown in Figure 1 and described subsequently in more detail.

The Gene search identifies genes and keywords that share co-occurrences in Medline abstracts with a gene of interest. It provides answers on a question like; 'Which diseases and drugs are strongly connected to the gene p53'? In a similar manner, the BioConcept search identifies genes and keywords that share co-occurrences in Medline abstracts with a keyword of interest and provides answers on a question like, 'Which pathways are associated with Alzheimer's disease?' Screenshots and the workflows of the Gene search and the BioConcept search are shown in Figure 2 and described below in more detail.

Microarray data analysis

Input. The user can select one of two analysis modes for microarray data analysis: the keyword enrichment calculator or the matrix generator. For each of the two analysis modes, the user needs to upload a list of gene identifiers (Affymetrix probe set identifiers, Entrez Gene identifiers or Ensembl identifiers), either by copy–paste or as a text file. Following this, the user must specify the correct Affymetrix microarray chip or species and the categories of keywords used for analysis (Figure 1A). Example gene sets are provided.

For the keyword enrichment calculator, thresholds need to be specified for the *P*-value significance level, the minimal number of co-publications, the minimal *R*-scaled score and the minimal number of submitted genes that have linkage with the analyzed keyword in literature (i.e. share abstract occurrences). For the threshold settings, sensible defaults are provided.

Both analysis modes can either be performed with 'species-specific' or 'cross-species' gene information. In the 'cross-species' mode, full gene names as well as gene symbols of human, mouse and rat orthologous genes are combined. This is in contrast to the 'species-specific' mode, in which species-specific gene names and symbols are used for analysis.

Output. The matrix generator produces a tab-delimited file in which all co-publication information between the uploaded set of genes and the selected keywords are

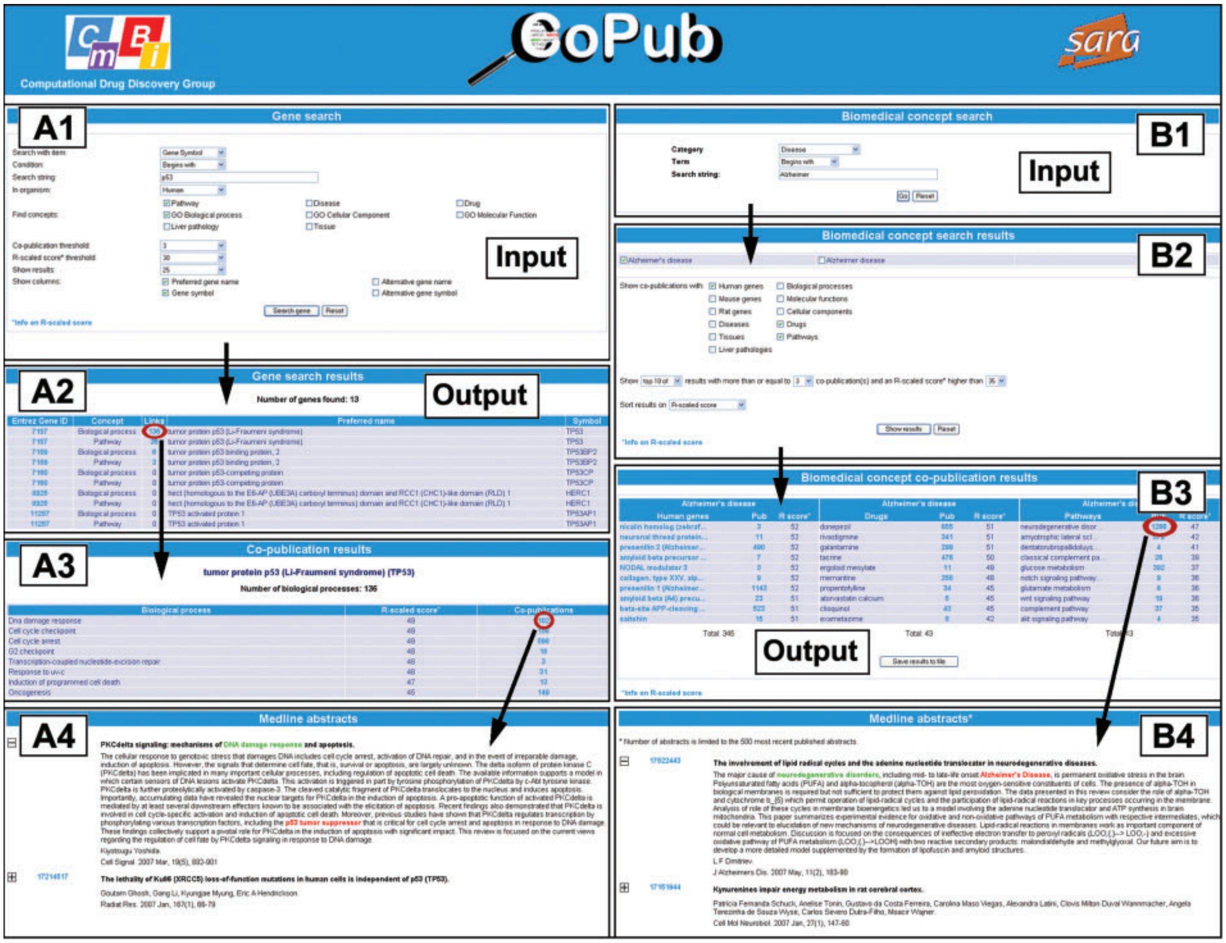

CMBI
Computational Drug Discovery Group
CoPub
sara
A1
Gene search
Input
A2
Gene search results
Number of genes found: 13
Output
A3
Co-publication results
tumor protein p53 (Li-Fraumeni syndrome) (TP53)
Number of biological processes: 136
A4
Medline abstracts
B1
Biomedical concept search
Input
B2
Biomedical concept search results
B3
Biomedical concept co-publication results
Output
B4
Medline abstracts*

Figure 2. Screenshots and workflows of the Gene search (**A**) and the BioConcept search (**B**). (A1) Input screen of the Gene search, which requires a single gene name as input. Furthermore, the categories of keywords need to be specified for which co-occurrences in literature with the gene of interest will be matched and retrieved. (A2) Output screen of the Gene search, which reports on the number of keywords that co-occur with the gene of interest, and links to an overview of the keywords (A3) and to Medline abstracts in which they co-occur (A4). (B1) Input screen of the BioConcept search, which requires a single keyword as input. (B2) Page to specify the categories of genes and keywords for which co-occurrences in literature with the keyword of interest will be matched and retrieved. (B3) Output screen of the BioConcept search, which reports on genes and keywords that co-occur with the keyword of interest in Medline abstracts, and with links to these abstracts (B4).

presented in a matrix format. The values in the matrix can either be the absolute number of co-publications or the *R*-scaled score between a gene and a keyword. This matrix is provided as a flat-format text file, which can be used for any kind of follow-up analysis, such as clustering the genes on basis of their keyword profile.

The keyword enrichment calculator produces a list of keywords, ranked on *P*-values (Figure 1B). This list already provides a first impression of the biological processes related to the gene set. The user can drilldown into these results by clicking on the hyperlinked number of genes that are significantly associated with the analyzed keyword (Figure 1B). It links to an overview of uploaded genes that share co-publications with the analyzed keyword (Figure 1C), and provides access to highlighted Medline abstracts in which they co-occur (Figure 1D).

CoPub can also visualize the keyword enrichment results in a SVG format (Figure 1E). In this interactive network, nodes represent over-represented keywords and differentially expressed genes and edges represent literature links. The nodes and edges link to the relevant Medline abstracts in which co-occurring genes and keywords are highlighted. This allows for quick retrieval of relevant literature and interpretation of the data. Threshold sliders for the *R*-scaled score and the literature count can be used to reduce the size of the network interactively. Alternatively, the literature-network can be re-calculated with new threshold values.

Gene search and BioConcept search

Input. The Gene search option requires a single gene name or symbol, or for the BioConcept search, a single keyword as input (Figure 2). Furthermore, for both the Gene search and the BioConcept search, the categories of keywords need to be specified for which co-occurrences in literature with the gene or keyword of interest will be matched and retrieved.

In addition, thresholds for the minimal number of co-publications and the minimal *R*-scaled score between keywords/genes can be specified, for which sensible defaults are provided.

Output. Both the Gene search and the BioConcept search return lists of genes and keywords that are strongly linked with the input gene or keyword of interest.

The results on these pages are all hyperlinked. This enables the user to navigate through various pages that report on how many times genes and keywords co-occurred in literature, and provide access to the relevant Medline abstracts (Figure 2).

DISCUSSION AND CONCLUSION

We have developed CoPub, a web-based tool for calculating keyword enrichment in sets of regulated genes. It detects keywords that are significantly linked to a set of genes, using robust co-occurrence statistics of genes and keywords in Medline abstracts. In a study in which gene expression profiles induced by 11 distinct hepatotoxicants were analyzed with CoPub, we were able to accurately describe histopathological findings and the mode of toxicity of these compounds (22). This shows that CoPub is a useful additional tool in the toolbox for the analysis of microarray experiments.

We intend to further develop CoPub by updating its Medline content on a regular basis (once every 2 months), and by adding new and improved keyword thesauri. Furthermore, we will broaden the scope of the keyword over-representation analysis by allowing identifiers from other microarray platforms as input for the keyword over-representation analysis. On the output side, options for multiple graphical output formats, such as png and jpg as well as a connection to Cytoscape, an open source network analysis tool, will be offered, allowing for improved downstream analysis of the results generated by CoPub.

ACKNOWLEDGEMENTS

The authors thank SARA Computing and Networking Services (Amsterdam, The Netherlands) for updating and maintaining our CoPub database and web server, which was made possible by grants received from the Netherlands Bioinformatics Centre (NBIC) under the BioAssist program and from Organon, part of Schering-Plough corporation.

Conflict of interest statement. None declared.

REFERENCES

1. Cheng,J., Sun,S., Tracy,A., Hubbell,E., Morris,J., Valmeekam,V., Kimbrough,A., Cline,M.S., Liu,G., Shigeta,R. *et al*. (2004) NetAffx gene ontology mining tool: a visual approach for microarray data analysis. *Bioinformatics*, **20**, 1462–1463.
2. Nakao,M., Bono,H., Kawashima,S., Kamiya,T., Sato,K., Goto,S. and Kanehisa,M. (1999) Genome-scale gene expression analysis and pathway reconstruction in KEGG. *Genome Inform. Ser. Workshop Genome Inform*., **10**, 94–103.
3. Dahlquist,K.D., Salomonis,N., Vranizan,K., Lawlor,S.C. and Conklin,B.R. (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet*., **31**, 19–20.
4. Mlecnik,B., Scheideler,M., Hackl,H., Hartler,J., Sanchez-Cabo,F. and Trajanoski,Z. (2005) PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. *Nucleic Acids Res*., **33**, W633–W637.
5. 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.
6. Chung,H.J., Park,C.H., Han,M.R., Lee,S., Ohn,J.H., Kim,J. and Kim,J.H. (2005) ArrayXPath II: mapping and visualizing microarray gene-expression data with biomedical ontologies and integrated biological pathway resources using scalable vector graphics. *Nucleic Acids Res*., **33**, W621–W626.
7. Pandey,R., Guru,R.K. and Mount,D.W. (2004) Pathway miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data. *Bioinformatics*, **20**, 2156–2158.
8. 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.
9. 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.
10. Wu,J., Mao,X., Cai,T., Luo,J. and Wei,L. (2006) KOBAS server: a web-based platform for automated annotation and pathway identification. *Nucleic Acids Res*., **34**, W720–W724.
11. Backes,C., Keller,A., Kuentzer,J., Kneissl,B., Comtesse,N., Elnakady,Y.A., Muller,R., Meese,E. and Lenhof,H.P. (2007) GeneTrail—advanced gene set enrichment analysis. *Nucleic Acids Res*., **35**, W186–W192.
12. Beissbarth,T. and Speed,T.P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.
13. Huang da,W., Sherman,B.T., Tan,Q., Kir,J., Liu,D., Bryant,D., Guo,Y., Stephens,R., Baseler,M.W., Lane,H.C. *et al*. (2007) DAVID bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res*., **35**, W169–W175.
14. Jelier,R., Jenster,G., Dorssers,L.C., Wouters,B.J., Hendriksen,P.J., Mons,B., Delwel,R. and Kors,J.A. (2007) Text-derived concept profiles support assessment of DNA microarray data for acute myeloid leukemia and for androgen receptor stimulation. *BMC Bioinform*., **8**, 14.
15. Chaussabel,D. and Sher,A. (2002) Mining microarray expression data by literature profiling. *Genome Biol*., **3**, RESEARCH0055.
16. Chagoyen,M., Carmona-Saez,P., Shatkay,H., Carazo,J.M. and Pascual-Montano,A. (2006) Discovering semantic features in the literature: a foundation for building functional associations. *BMC Bioinform*., **7**, 41.
17. Kuffner,R., Fundel,K. and Zimmer,R. (2005) Expert knowledge without the expert: integrated analysis of gene expression and literature to derive active functional contexts. *Bioinformatics*, **21 (Suppl 2)**, ii259–ii267.
18. Jenssen,T.K., Laegreid,A., Komorowski,J. and Hovig,E. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet*., **28**, 21–28.
19. Rubinstein,R. and Simon,I. (2005) MILANO—custom annotation of microarray results using automatic literature searches. *BMC Bioinform*., **6**, 12.
20. Blaschke,C., Oliveros,J.C. and Valencia,A. (2001) Mining functional information associated with expression arrays. *Funct. Integr. Genomics*, **1**, 256–268.
21. Burkart,M.F., Wren,J.D., Herschkowitz,J.I., Perou,C.M. and Garner,H.R. (2007) Clustering microarray-derived gene lists through implicit literature relationships. *Bioinformatics*, **23**, 1995–2003.
22. Frijters,R., Verhoeven,S., Alkema,W., van Schaik,R. and Polman,J. (2007) Literature-based compound profiling: application to toxicogenomics. *Pharmacogenomics*, **8**, 1521–1534.
23. Alako,B.T., Veldhoven,A., van Baal,S., Jelier,R., Verhoeven,S., Rullmann,T., Polman,J. and Jenster,G. (2005) CoPub Mapper: mining MEDLINE based on search term co-publication. *BMC Bioinformatics*, **6**, 51.