CoPub Manual (version 2.3 alpha 2008-06-12)

1. Introduction

CoPub is a text mining tool that detects co-occurring biomedical concepts in abstracts from the MedLine literature database. The biomedical concepts included in CoPub are **all** human, mouse and rat genes, furthermore biological processes, molecular functions and cellular components from Gene Ontology, and also diseases, drugs, liver pathology and pathways. Altogether more than 260,000 search strings are linked with CoPub.
Special attention was given to genes and proteins. For all human, mouse and rat genes not only long forms of names were used, but also their symbols and aliases, which increases recall. Symbols not referring to genes or proteins are a well known problem, but sophisticated scripts detect these homonyms and neglect the abstracts in which they occur, thereby increasing precision.
CoPub is especially useful for microarray data analysis. It is often difficult to grasp the meaning of lists of differentially expressed genes. Mining MedLine with gene names one by one is laborious and tedious, if not impossible, and many relevant abstracts will be missed. With CoPub it is now possible to upload a list of Affymetrix identifiers and find biomedical concepts from MedLine that are significantly linked to the gene set. From every retrieved biomedical concept it is only one mouse click to co-published genes and another one to the relevant abstracts.

2. Access to CoPub

CoPub is hosted and updated by SARA Computing and Networking Services and access is possible via http://services.nbic.nl/cgi-bin/copub/CoPub.pl.

3. General Features of CoPub

With CoPub three types of searches are possible:

- Gene Search
  With **Gene search** links for your gene of interest with bioconcepts from several biomedical thesauri are detected and shown. The number of significant bioconcepts is shown as a hyperlink. Following the hyperlink the significant bioconcepts are provided in a table with a link to the number of co-publications. This link leads to the references and finally the abstracts.
  In the current version of CoPub (version 2.3 alpha 2008-06-12) bioconcepts are biological processes, molecular functions, cellular components, pathways, tissues, diseases, liver pathology and drugs.
- Bioconcept Search
  The **Bioconcept search** actually works the same as the Gene search, but in this case the input is a single bioconcept which can be chosen from the biomedical thesauri mentioned above. The output is a list of genes or bioconcepts with links to references and abstracts in which the bioconcept of interest is co-published with a gene or bioconcept.
- Multiple Gene Search
  The **Microarray Data Analysis** option allows the input of many genes as Affymetrix identifiers, EntrezGene identifiers or Ensembl Gene identifiers and will provide as output a list of significant keywords from selected categories with associated p-values and the number of genes responsible for the alert. For every keyword the number of genes is a hyperlink to a table with the gene names and the actual number of co-publications. This number is hyperlinked to the references and the abstracts.

A network can be made from the input list of genes and the output list of bioconcepts and subsequently visualized in SVG format.

Figure 1 Homepage of CoPub

3.1 Gene Search

The Gene Search option provides links in MedLine from a gene with keywords from biological processes, molecular functions, cellular components, diseases, liver pathology, pathways or drugs.

3.1.a How to perform a Gene Search

- Decide whether you want to search with a gene name or a gene symbol with ***Search with item***.
- Select a condition with ***condition***: “contains”, “matches exactly”, “begins with” or “ends with”.
- In open box after ***Search string*** type the string (or part of string) you want to find.
- ***In organism*** select the species: “human”, “mouse”, “rat” or “all”. “All” means the combined results of human, mouse and rat.
- With ***Find Concepts*** the category/categories of keywords is chosen: “biological processes”, “molecular functions”, “cellular components”, “pathways”, “tissues”, “diseases”, “liver pathology” or “drugs”.
- With ***Co-publication threshold*** select the minimum number of abstracts in which gene and keyword co-occur.
- With ***R-scaled score threshold*** select the minimum relative score between gene and keyword. The higher the score the more stringent the search will be. Explanation about this R-scaled score is given via a hyperlink at he bottom of the page.

- ***Show results*** allows the output of either 25, 50, 150 or all results.
- ***Show columns*** determines the output columns.
- ***Search gene*** starts the search.
- ***Reset*** clears all previous settings.

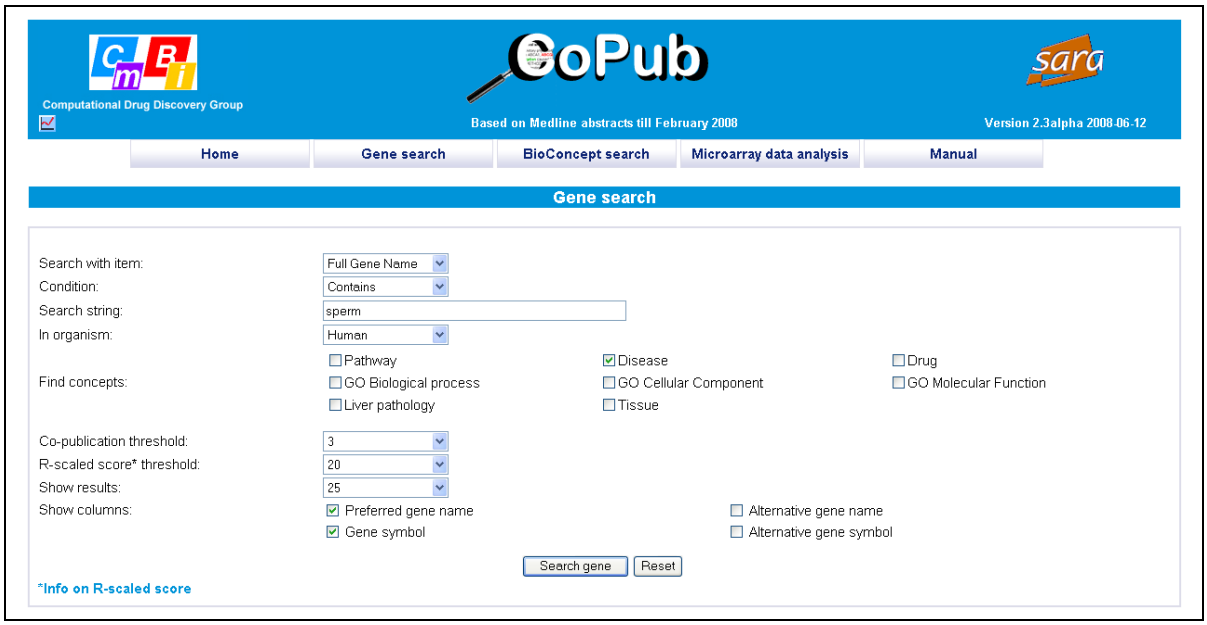


Figure 2 Gene Search: search for co-occurrences of human genes – with "sperm" in the gene name – and diseases.

3.1.b Interpretation of gene search results

The search described in the previous section "**How to perform a Gene Search"** will yield 153 genes with a gene name (or alternative name) containing "sperm". Only the first 6 genes are displayed below. The first column contains links to gene specific information via Entrez Gene. The second column shows the category from which gene-concept pairs are found. The third column gives the number of biomedical concepts for the respective gene. The following columns show preferred gene names and symbols. Sometimes results are displayed in which the original query string is not present. In these cases the search string is present in *alternative gene name* or *gene alias*.

Computational Drug Discovery Group

CoPub

Based on Medline abstracts till February 2008

Version 2.3alpha 2008-06-12

Home | Gene search | BioConcept search | Microarray data analysis | Manual

Gene search results

Number of genes found: 153

Entrez Gene ID	Concept	Links	Preferred name	Alternative name	Symbol
831	Disease	61	calpastatin	calpain inhibitor sperm BS-17 component heart-type calpastatin sperm BS-17 component heart-type calpastatin calpain inhibitor	CAST
1538	Disease	0	cylicin, basic protein of sperm head cytoskeleton 1	cylicin 1	CYLC1
1539	Disease	0	cylicin, basic protein of sperm head cytoskeleton 2	cylicin 2	CYLC2
1617	Disease	4	deleted in azoospermia 1	deleted in azoospermia	DAZ1
1618	Disease	2	deleted in azoospermia-like	deleted in azoospermia-like autosomal germline specific RNA binding protein spermatogenesis gene on the Y-like autosomal spermatogenesis gene on the Y-like autosomal germline specific RNA binding protein deleted in azoospermia-like autosomal	DAZL
2821	Disease	55	glucose phosphate isomerase	neuroleukin oxoisomerase sperm antigen-36 phosphohexomutase phosphosaccharomutase phosphohexose isomerase phosphoglucose isomerase autocrine motility factor hexosephosphate isomerase glucose-6-phosphate isomerase hexose monophosphate isomerase sperm antigen-36 phosphosaccharomutase	GPI

Figure 3 Results of gene search with “sperm” and disease.

Clicking the hyperlink “4” for the fourth gene (DAZ1) displays a table with disease terms co-published with the DAZ1 gene. Ranking is based either on decreasing R-scaled score or decreasing number of co-publications. There are 63 abstracts mentioning both “infertility, male” and “DAZ1”. The hyperlink “63” associated with “infertility” brings you to the references and the abstracts.

Computational Drug Discovery Group

CoPub

Based on Medline abstracts till February 2008

Version 2.3alpha 2008-06-12

Co-publication results

deleted in azoospermia 1 (DAZ1)

Number of diseases: 4

Disease	R-scaled score*	Co-publications
Infertility, male	56	63
Sterility	51	82
Klinefelter's syndrome	50	4
Testis, undescended	49	6

*Info on R-scaled score

Figure 4 Results for DAZ1 and diseases.

3.2 BioConcept Search

The BioConcept Search option provides links in MedLine for a keyword from categories *GO biological processes, GO molecular functions, GO cellular components, diseases, pathways, tissues, liver pathology* or *drugs* with *human, mouse, rat or orthologous genes* and keywords from the following categories: *GO biological processes, diseases, pathways, drugs, GO molecular functions, GO cellular components, liver pathology* and *tissues.*

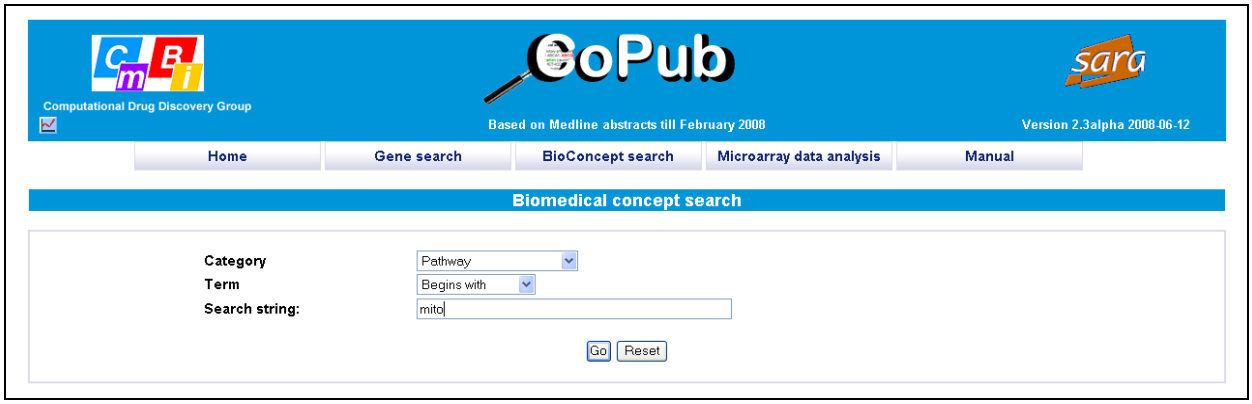


Figure 5 BioConcept search.

3.2.a How to perform a BioConcept Search

- Press button BioConcept Search
- In the BioConcept Search window select the ***category*** from which you would like to select a keyword. Possible categories are "GO biological Processes", "GO molecular functions", "GO cellular components", "tissues", "liver pathology", "Pathways", "Drugs" and "Diseases".
- With ***Term*** select "Begins with", "Contains", "Ends with" or "Matches exactly".
- In the empty box after ***Search string*** enter a search string or a part of a string.
- Press ***Go*** button.

3.2.b Interpretation of BioConcept search results

A BioConcept search for e.g. pathways beginning with "mito" results in a list/table of 9 pathways beginning with "mito". Every pathway has a checkbox attached to it. After selection of the appropriate pathway(s) a selection should be made for the categories from which keywords are co-published with the selected pathway(s). Select one or more categories.
Decide how many results should be shown, how results should be sorted and what the minimal number of abstracts should be with co-publication of the selected pathway(s) and other keywords.

Figure 6 BioConcept search with pathways beginning with "mito".

Select e.g. pathway "mitochondrial fatty acid beta-oxidation" and categories "human genes" and "diseases" and subsequently define the number of links you want to see and the minimum number of co-publications per link and also whether results should be sorted based on R-scaled score or on co-publication number.

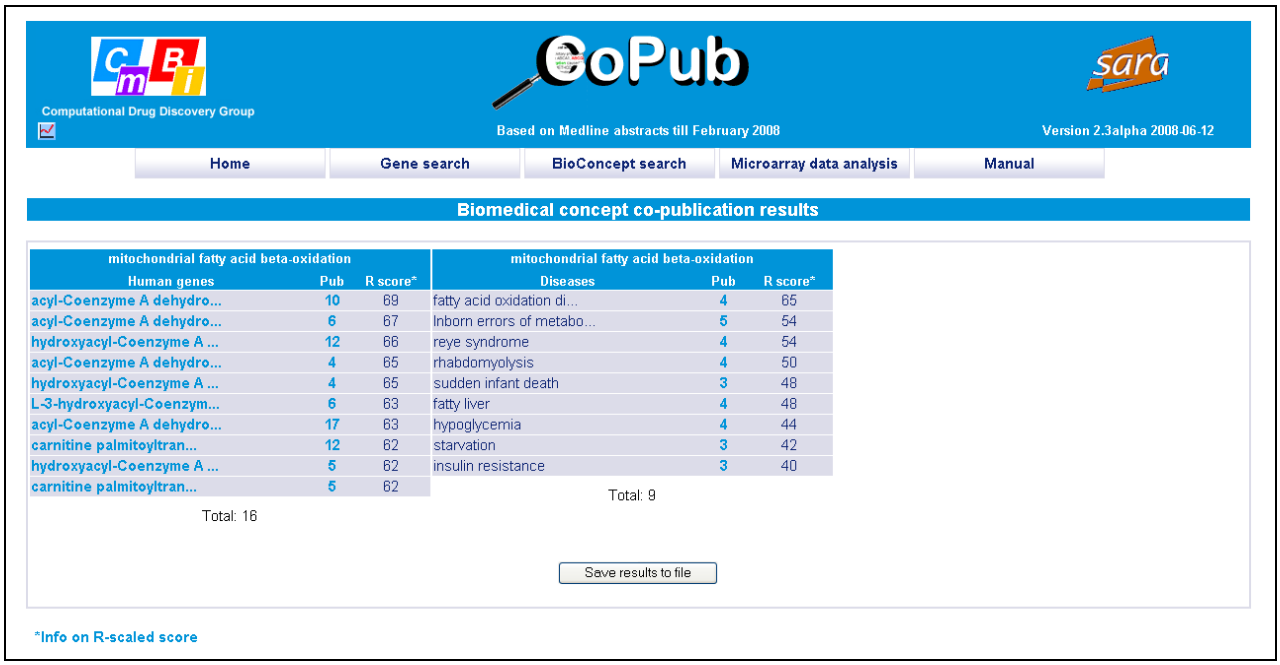


Figure 7 Results for "mitochondrial fatty acid beta oxidation" and "human genes" and "mitochondrial fatty acid beta oxidation" and "diseases".

A list of human genes and diseases is shown sorted by decreasing R-scaled score. The R score in the most right column indicates a relative score taking into account

the absolute number of publications for keyword A and B separately and together. Furthermore R is log-transformed and scaled between 0 and 100. More information is given through the hyperlink at he bottom of the page or here.
Gene names may be truncated -this depends on the column width- but full gene names are shown by mouseover.
The column "Pub" gives the absolute number of co-publications between keyword A and B as a hyperlink. The hyperlink brings you to a list of abstracts. Clicking the "+" preceding every reference shows the abstract with keywords A and B highlighted. The hyperlinked PubMed Identifier is a link to the abstract at EBI (http://srs.ebi.ac.uk/).

Queried strings

Queried pathway term:	Mitochondrial fatty acid beta-oxidation
Queried human gene (Entrez Gene ID):	37
Gene Name:	acyl-Coenzyme A dehydrogenase, very long chain
Gene Symbol:	ACADVL (alias: ACAD6 / LCACD / VLCAD / LCACD / ACAD6)
Number of co-publications:	10
R-scaled score	69

Click on the + image to retrieve the abstract from EMBL-EBI and highlight the terms, or click the PubMed ID to open the abstract in a external window for EMBL-EBI.

Medline abstracts

Please be patient when clicking on the + image to retrieve Medline abstract from EMBL-EBI.

⊟ 17999356

Genetic basis for correction of very-long-chain acyl-coenzyme A dehydrogenase deficiency by bezafibrate in patient fibroblasts: toward a genotype-based therapy.

Very-long-chain acyl-coenzyme A dehydrogenase (VLCAD) deficiency is an inborn mitochondrial fatty-acid beta-oxidation (FAO) defect associated with a broad mutational spectrum, with phenotypes ranging from fatal cardiopathy in infancy to adolescent-onset myopathy, and for which there is no established treatment. Recent data suggest that bezafibrate could improve the FAO capacities in beta-oxidation-deficient cells, by enhancing the residual level of mutant enzyme activity via gene-expression stimulation. Since VLCAD-deficient patients frequently harbor missense mutations with unpredictable effects on enzyme activity, we investigated the response to bezafibrate as a function of genotype in 33 VLCAD-deficient fibroblasts representing 45 different mutations. Treatment with bezafibrate (400 microM for 48 h) resulted in a marked increase in FAO capacities, often leading to restoration of normal values, for 21 genotypes that mainly corresponded to patients with the myopathic phenotype. In contrast, bezafibrate induced no changes in FAO for 11 genotypes corresponding to severe neonatal or infantile phenotypes. This pattern of response was not due to differential inductions of VLCAD messenger RNA, as shown by quantitative real-time polymerase chain reaction, but reflected variable increases in measured VLCAD residual enzyme activity in response to bezafibrate. Genotype cross-analysis allowed the identification of alleles carrying missense mutations, which could account for these different pharmacological profiles and, on this basis, led to the characterization of 9 mild and 11 severe missense mutations. Altogether, the responses to bezafibrate reflected the severity of the metabolic blockage in various genotypes, which appeared to be correlated with the phenotype, thus providing a new approach for analysis of genetic heterogeneity. Finally, this study emphasizes the potential of bezafibrate, a widely prescribed hypolipidemic drug, for the correction of VLCAD deficiency and exemplifies the integration of molecular information in a therapeutic strategy.

S Gobin-Limballe, F Djouadi, F Aubey, S Olpin, B S Andresen, S Yamaguchi, H Mandel, T Fukao, J P N Ruiter, R J A Wanders, R McAndrew, J J Kim, J Bastin.
Am J Hum Genet. 2007 Nov, 81(6), 1133-43

⊞ 15862275

Synergistic heterozygosity in mice with inherited enzyme deficiencies of mitochondrial fatty acid beta-oxidation.

A Michele Schuler, Barbara A Gower, Dietrich Matern, Piero Rinaldo, Jerry Vockley, Philip A Wood.
Mol Genet Metab. 2005 May, 85(1), 7-11

⊞ 15347768

Mitochondrial fatty acid beta-oxidation in the human eye and brain: implications for the retinopathy of long-chain 3-hydroxyacyl-CoA dehydrogenase deficiency.

Figure 8 References and abstracts for "Mitochondrial fatty acid beta oxidation" and "Acyl-Coenzyme A dehydrogenase, very long chain".

3.3 Microarray Data Analysis

The **Microarray Data Analysis** option calculates co-occurrences between input genes and biomedical concepts from different categories. Not every biomedical concept for which co-occurrences are found will be shown, only those keywords that are co-published with the genes from the gene set more than by change alone. A Fisher Exact test will calculate the p-values for every keyword using all genes from the used Affymetrix GeneChip as a background gene set. P-values are adjusted by applying Benjamini-Hochberg multiple testing correction.
Every biomedical concept co-published with a gene from the gene set is shown with the absolute number of abstracts in which gene and keyword are co-published, a relative score R and an adjusted p-value.

3.3.a How to perform Microarray Data Analysis

- To upload a gene set either copy/paste a list of Affymetrix, EntrezGene or Ensembl identifiers in the empty window or define a tab-delimited text file with Affymetrix, EntrezGene or Ensembl identifiers (without a header). We have provided two example gene sets to test CoPub Microarray Data Analysis.
- With ***Gene identifier type*** select the appropriate identifier type.
- With ***Species selection*** select the appropriate species.
- Select the correct Affymetrix GeneChip. This version of CoPub only allows input of human, mouse or rat genes. Therefore GeneChips for only these three species can be selected.
- With ***Analysis type*** decide whether you want to see only significant biomedical concepts (Enrichment calculator) or all biomedical concepts (Matrix generator). Selecting Matrix Generator will show you a different right panel that allows settings for a matrix of co-publications.
- For Enrichment calculator select the categories from which keywords will be displayed with ***Search category***.
- With ***Minimal number of genes associated with concept*** set a threshold for the number of genes co-published with a biomedical concept. In general the more genes are co-mentioned with a keyword the higher the significance of that keyword.
- The minimal number of co-publications of a gene with a keyword may be set with ***Minimal nr. of co-publications between gene and keyword***. Only one or a few co-publications may me regarded as insignificant or coincidence.
- With ***R-scaled score threshold*** the relative score R may be set to a certain value. The higher R the more stringent the search.
- With ***Gene Orthology Selection*** decide on a *species-specific* or a *cross species* search. The species-specific option will use gene names and symbols from the selected species. The cross-species option combines gene names and symbols from human, mouse and rat.
- The Fisher Exact Test and subsequent Benjamini-Hochberg algorithm calculates p-values for every keyword. With ***Show p-values*** the p-value cut-off for the output is determined.
- Start the analysis.
  NB: Sensible default values are given based on extensive use of CoPub, however they may be changed to make searches more or less stringent.

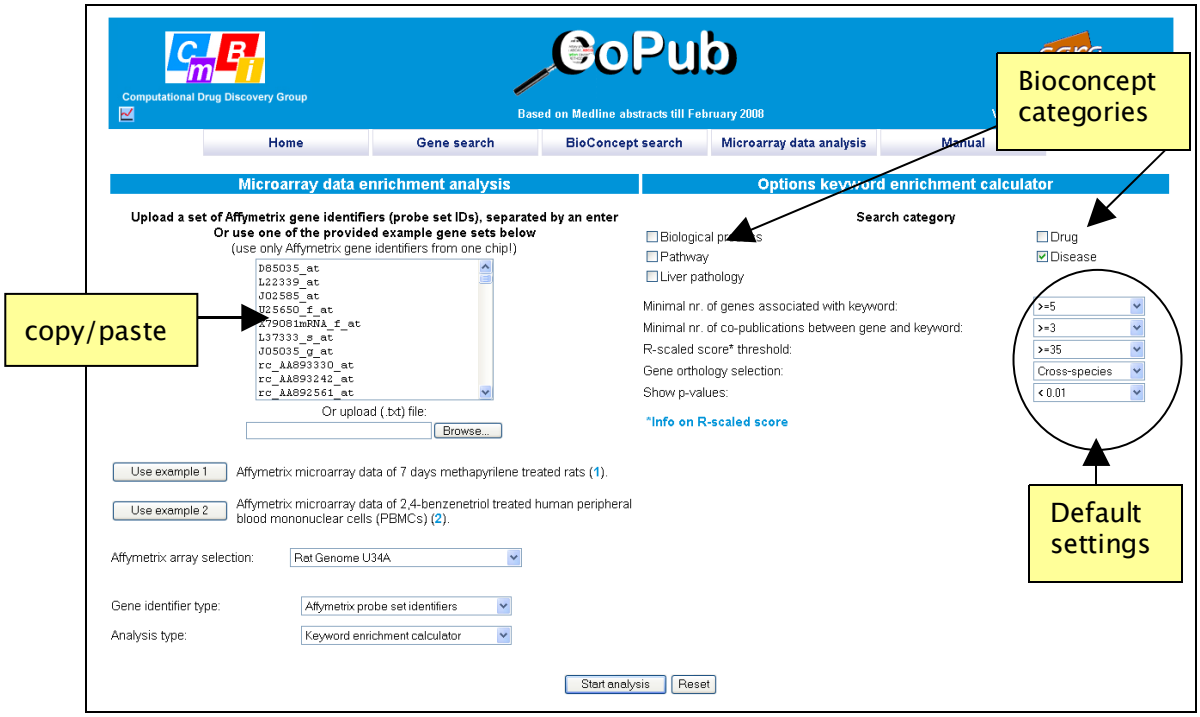
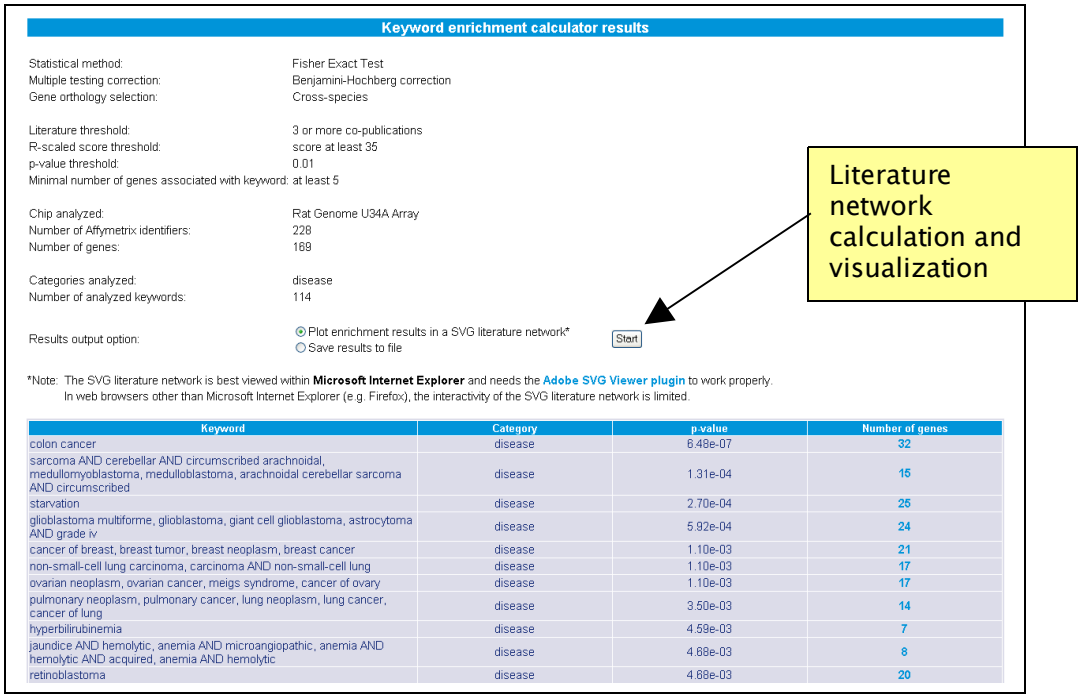


Figure 9 Microarray data analysis.

3.3.b Interpretation of Microarray Analysis results

Below results are shown with Example set1 and category "disease".



Keyword	Category	p-value	Number of genes
colon cancer	disease	6.48e-07	32
sarcoma AND cerebellar AND circumscribed arachnoidal, medullomyoblastoma, medulloblastoma, arachnoidal cerebellar sarcoma AND circumscribed	disease	1.31e-04	15
starvation	disease	2.70e-04	25
glioblastoma multiforme, glioblastoma, giant cell glioblastoma, astrocytoma AND grade iv	disease	5.92e-04	24
cancer of breast, breast tumor, breast neoplasm, breast cancer	disease	1.10e-03	21
non-small-cell lung carcinoma, carcinoma AND non-small-cell lung	disease	1.10e-03	17
ovarian neoplasm, ovarian cancer, meigs syndrome, cancer of ovary	disease	1.10e-03	17
pulmonary neoplasm, pulmonary cancer, lung neoplasm, lung cancer, cancer of lung	disease	3.50e-03	14
hyperbilirubinemia	disease	4.59e-03	7
jaundice AND hemolytic, anemia AND microangiopathic, anemia AND hemolytic AND acquired, anemia AND hemolytic	disease	4.68e-03	8
retinoblastoma	disease	4.68e-03	20

Figure 10 Microarray data analysis results for example gene set 1 and category "disease".

The output of a multiple gene search first shows the settings of the search. In particular the name of the input file (in case a file was uploaded), the p-value threshold, the minimum number of co-publications, the minimum number of genes per keyword, the R-scaled threshold, chip type, the number of Affymetrix identifiers in the input gene set, the number of genes based on Entrez Gene identifiers in the input gene set, the categories searched and the number of keywords in these categories.
The output is a table with 4 columns. The first column lists the biomedical concepts, the second column the category from which this keyword is derived, the third column the adjusted p-value and the last column the number of genes from the input list that was found co-published with that particular bioconcept. The table is sorted by increasing p-value. The number in the last column is a hyperlink to a list of co-published genes. The table may be stored as a tab-delimited txt file.
The first column shows the gene name, the second column the gene symbol, the next column all Affymetrix identifiers from the selected chip type associated to that gene, the fourth column a link to LocusLink/Entrez Gene and the last column shows the number of co-publications for the keyword and the gene. The number between parenthesis shows the number of abstracts in case a species-specific search would have been done. This may give an indication about species differences.
Clicking the link in the last column brings up the references and from there it is one mouse click to the abstracts.
The results from the microarray data analysis may be saved to a file or a literature network may be calculated and visualized.

Organon Schering-Plough — CoPub — sara

Version 2.1alpha 2008-01-16 — Based on Medline abstracts till May 2007

Co-publications (Cross-species)

Starvation (disease)

Gene Name	Symbol	Affymetrix ID(s)	Entrezgene ID	#copub(**)	R-score*
tribbles homolog 3 (Drosophila)	Trib3	rc_H31287_g_at	246273	3(0)	47
asparagine synthetase	Asns	U07201_at	25612	20(19)	46
serine dehydratase	Sds	X13119cds_s_at J03863_at	25044	10(10)	42
DNA-damage inducible transcript 3	Ddit3	U30186_at	29467	9(5)	41
pyruvate kinase, liver and RBC	Pklr	X05684_at	24651	8(6)	40
activating transcription factor 3	Atf3	M63282_at	25389	3(3)	39
glucose-6-phosphatase, catalytic	G6pc	L37333_s_at	25634	41(41)	39
ATP citrate lyase	Acly	J05210_g_at	24159	6(6)	39
growth arrest and DNA-damage-inducible 45 alpha	Gadd45a	L32591mRNA_at L32591mRNA_g_at rc_AI070295_at rc_AI070295_g_at	25112	7(7)	39
acyl-CoA synthetase long-chain family member 1	Acsl1	rc_AA893242_at	25288	11(7)	38
aldolase A	Aldoa	M12919mRNA#2_at rc_AA924326_s_at	24189	3(3)	37
ornithine decarboxylase 1	Odc1	J04791_s_at	24609	28(28)	37
cell division cycle 2 homolog A (S. pombe)	Cdc2a	X60767mRNA_s_at	54237	36(0)	37
eukaryotic translation elongation factor 1 alpha 1	Eef1a1	rc_AI008852_at	171361	14(1)	37
stearoyl-Coenzyme A desaturase 1	Scd1	rc_AI175764_s_at J02585_at	246074	8(2)	36
topoisomerase (DNA) 2 alpha	Top2a	rc_AA899854_at	360243	3(3)	36
glutamate-cysteine ligase, catalytic subunit	Gclc	J05181_at	25283	7(1)	35

Figure 11 Results for co-occurrence of "starvation" and regulated genes from example data set 1.

3.3.c Literature network calculation and visualization

Apart from the table output as shown above in figure 10 there is also the possibility to calculate and visualize a network in SVG format. Just hit the start button (see figure 10). This option will include all the genes used as input and the bioconcepts shown as output.
The SVG literature network is best viewed within Microsoft Internet Explorer and needs the **Adobe SVG Viewer plugin** to work properly.
To generate literature-networks CoPub uses GraphViz (http://www.graphviz.org) to calculate the graph layout (neato), and for producing the literature-network in a scalable vector graphic (SVG) format. Interactivity in generated SVG networks is implemented using Perl and JavaScript.

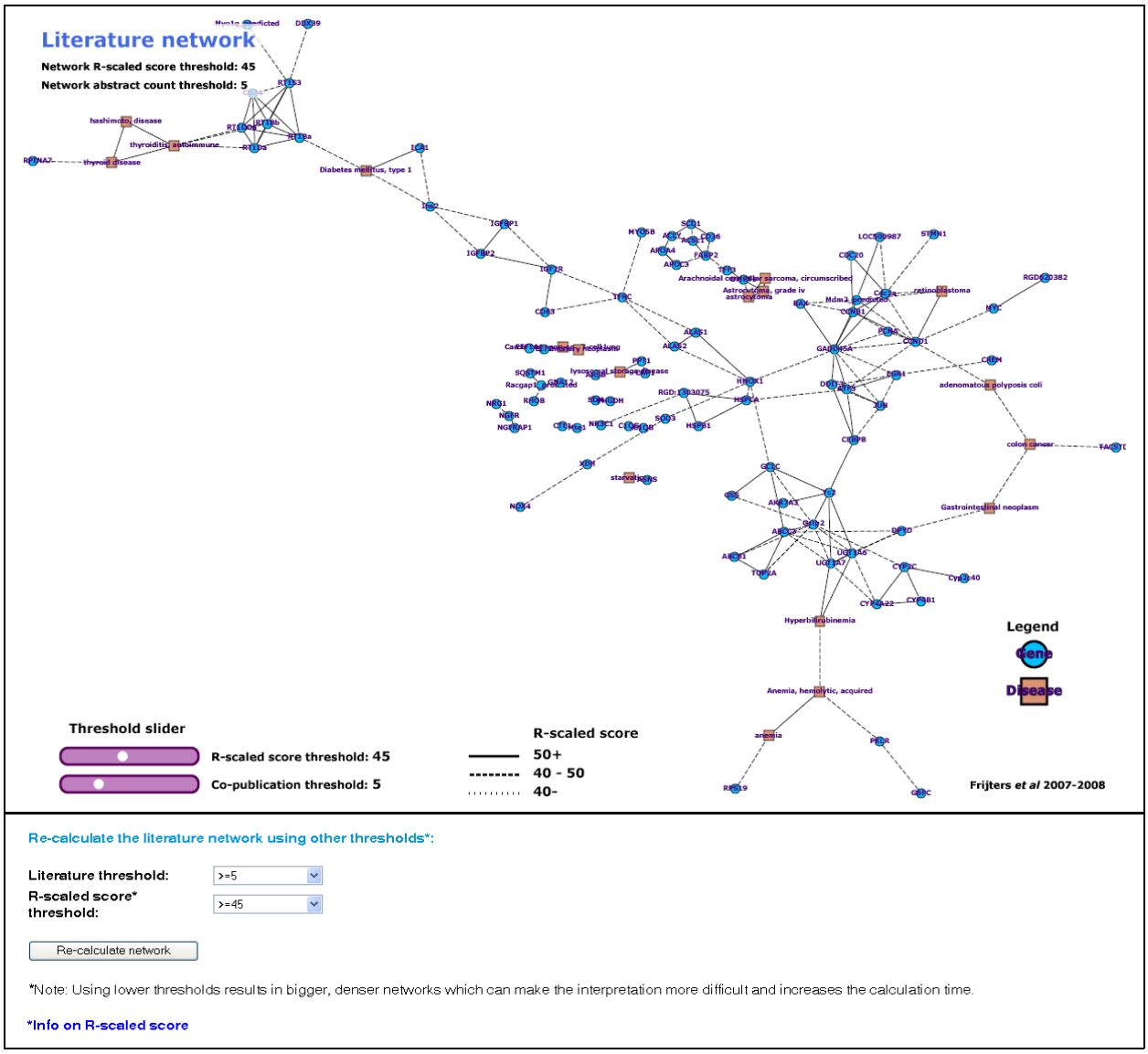


Figure 12 Literature network for genes from example dataset 1 and significantly associated genes

In this figure the nodes represent input genes or keywords from the chosen categories. The edges represent the number of co-occurrences between the linked nodes. Nodes are hyperlinked to the abstracts found for that node (gene/keyword). Edges are hyperlinked to abstracts in which both nodes occur.

Sometimes a network is too dense because of the number of genes and keywords. In these cases it is possible to recalculate the network using more stringent settings. The dropdown boxes “Literature threshold” and “R-scaled score threshold” allow for a more stringent or relaxed settings. After changing the setting just click the ***Re-calculate network*** button.

4 Thesauri

.a Genes

Human, mouse and rat gene thesauri were compiled from NCBI Entrez Gene database (release of December 2005) (Maglott *et al*, 2007, NAR, vol 35). In order to search Medline with one or more full gene names, gene symbols and aliases, the gene name thesauri were processed as described by Alako *et al* (2005, BMC Bioinformatics, vol 11, nr 6). Furthermore, gene names and gene symbols of orthologous genes were combined to make the keyword search in Medline more comprehensive.
Currently the gene thesaurus contains 25,083 human genes, 35,944 mouse genes and 24,427 rat genes.

.b GO Biological Processes, GO Molecular functions, CO Cellular Components

The GO biological process, molecular function and cellular components thesauri were compiled from the Gene Ontology database.

.c Drugs

The drugs thesaurus is a compilation of 5795 drugs (brand names and general names) from RxList from 1995-2005.

.d Pathways

The pathway thesaurus was compiled from the KEGG database, the encyclopedia of human genes and metabolism database and the Reactome database , containing 817 pathway names.

.e Diseases

The disease thesaurus was made using the Karolinska Institute Diseases and Disorders Database supplemented with disease names from Wikipedia (May 23rd 2007) (http://en.wikipedia.org/wiki/Disease) in particular from sections: childhood diseases, eponymous diseases, diseases caused by insects and infectious diseases.

.f Liver pathology

Developed at Organon from pathology textbooks in collaboration with pathologists.

5 Example Data Sets

Example 1: Microarray data (Rat Genome U34A chip) of 7 days methapyrilene treated rats (1).
Ellinger-Ziegelbauer et al, Mutat Res. 2005 Aug 4;575(1-2):61-84
Comparison of the expression profiles induced by genotoxic and nongenotoxic carcinogens in rat liver.

Example 2: Microarray data (Human Genome U133A 2.0 Array) of 2,4-benzenetriol treated human peripheral blood mononuclear cells (PBMCs) (2).
Gilles B et al, Genomics. 2007 Sep;90(3):324-33. Epub 2007 Jun 15