
SciMiner

User's Manual

Copyright 2008 Junguk Hur. All rights reserved.

Bioinformatics Program

University of Michigan

Ann Arbor, MI 48109, USA

Email: juhur@umich.edu

Homepage: <http://jdrf.neurology.med.umich.edu/SciMiner/>

I. Introduction to SciMiner

- A. SciMiner is a web-based biomedical literature mining and analysis tool aiming at identifying targets (genes / proteins) in user's interested topics by NLP (Natural Language Processing)
- B. Refer to the SciMiner homepage for details.

II. Menu Bar in SciMiner

Home
Introduction
Supplementary
Run SciMiner
Post-Mining Analysis
Merge Queries
Completed
Download
Contact

- i. **Run SciMiner:** For submission of a PubMed query or PMIDs to initiate mining process.
- ii. **Post-Mining Analysis:** Functional enrichment analyses of five areas; targets, Gene Ontology, MeSH, pathway, and Protein-Protein Interaction.
- iii. **Merge Queries:** Integrate multiple search results.
- iv. **Completed:** Retrieve completed SciMiner mining (Run SciMiner) or analysis (Post-Mining Analysis) jobs.
- v. **Download:** Standalone version is available here.

III. Log-In

User should have a valid account to use the SciMiner. If you do not have one, you can easily create one online. “Run SciMiner”, “Post-Mining Analysis”, “Merge Queries” and “Completed” menus require user authentication.

The screenshot shows the SciMiner login interface. At the top, the SciMiner logo is displayed alongside the text "An online literature mining tool for target identification and functional enrichment analysis" and the NCBI logo. A blue sidebar on the left contains navigation links: Home, Introduction, Supplementary, Run SciMiner, Post-Mining Analysis, Merge Queries, Completed, Download, and Contact. The main content area is white and contains a login form with the following text: "Enter your email address and password.", "If you don't have a SciMiner account, Click [HERE](#) to create one.", and "Email to your [system administrator](#) for any question." The form includes an "E-mail:" field with the value "useremail@school.edu", a "Password:" field with masked characters, and "Login" and "Reset" buttons.

The public version of SciMiner has a default of 1000 documents (500 new documents) limit per query. If users need more documents per query, they can either download the standalone version or contact the author (juhur@umich.edu) to request an increase. For standalone version, we generally recommend 20,000 documents per 1GB of RAM available. Going above this limit might result in abrupt halting of SciMiner due to a memory allocation error.

IV. Submitting a SciMiner query

Users can submit a simple text query or a set of PMIDs to SciMiner. Go to 'Run SciMiner' menu in the menu bar.

A. Starting a SciMiner mining process (query)

The screenshot shows the SciMiner web interface. The header includes the University of Michigan Medical School logo, the SciMiner title, the tagline "An online literature mining tool for target identification and functional enrichment analysis", and the NCBI logo. A left sidebar contains navigation links: Home, Introduction, Supplementary, Run SciMiner, Post-Mining Analysis, Merge Queries, Completed, Download, and Contact. The main content area is titled "SciMiner Query Submission" and contains the following instructions and form fields:

(Instruction): Click [HERE](#) for detailed help or move mouse pointer around to get a quick tip.

- 1) Name your query and enter your Pubmed search terms or provide PMIDs directly.
- 2) Submit your query, review the result, and correct wrongfully identified targets if necessary.
- 3) Re-run the same query to get an updated target identification results, if necessary with modified advanced options and user filters.

(Section1) Name your query.
Query name :

(Section2) Enter your query or provide PMIDs
1) NCBI Entrez PubMed search term(s)
 [PREVIEW](#) (ex) "Amyotrophic lateral sclerosis"[MeSH] AND "Reactive Oxygen Species"[MeSH]

[SHOW advanced options!](#)

* Supplementary files . You may need these files to create your own filter lists. (HUGO IDs are essential.)
Better save the text files and load them in a text editor like [UltraEdit](#) or [EditPlus](#)

- 1) Full HUGO Content ([EXCEL](#), [TXT](#))
- 2) Unique symbols ([TXT](#))
- 3) Unique names ([TXT](#))

[Logout](#)

- i. Give a name to your query.
- ii. Provide any of the following three
 1. NCBI PubMed query string or simple text strings
 2. List of PMIDs (in the advanced options)
 3. A file containing a list of PMIDs (in the advanced options)
- iii. If you provided a PubMed query string, please click 'PREVIEW' button to see how many papers you should expect. For most users, 1000 is the default threshold for total number of document with a maximum of 500 new documents.

iv. (Optional) Modify the ‘advanced options’ with section 3

(Section3) SciMiner Mining Mode (*[NCBI Gene2PubMed](#) and [GeneRIF](#) involve no text mining)

Mining mode	SciMiner text mining	Species extension by HomoloGene	Extend by HomoloGene
-------------	----------------------	---------------------------------	----------------------

1. Mining Mode

- A. **SciMiner text mining:** This option applies to the standard SciMiner text mining to the documents. With this mining mode, the ‘Species extension by HomoloGene’ mode is always ‘Extend by HomoloGene’ even though users change to ‘Only explicit human targets’
- B. **NCBI Gene2PubMed:** This option make SciMiner use the NCBI’s Gene2PubMed mapping information. Please refer to the [NCBI Help document](#) for detail.
- C. **NCBI GeneRIF:** This open make SciMiner use the NCBI’s GeneRIF information. They are manually entered review sentences about genes. Refer to the above [NCBI Help document](#) for more detail.

v. (Optional) Modify the ‘advanced options’ with section 4

(Section4) Additional options for SciMiner text mining

<input type="checkbox"/>	Names to be ignored	<input type="text"/>	<input type="button" value="Browse..."/>	(Default by SciMiner v2.2)
<input type="checkbox"/>	Symbols to be excluded	<input type="text"/>	<input type="button" value="Browse..."/>	(Default by SciMiner v2.2)
<input type="checkbox"/>	Symbols to be included	<input type="text"/>	<input type="button" value="Browse..."/>	(Default by SciMiner v2.2)
<input checked="" type="checkbox"/>	Include any gene symbol longer than	<input type="text" value="6"/>	, unless specified by the above filters	
<input checked="" type="checkbox"/>	SciMiner confidence score threshold	<input type="text" value="0.1"/>	(0.1: minimum, ≥0.3: moderate, ≥0.6: high) more on scores	
<input checked="" type="checkbox"/>	Do not include phenotype only genes	(ex) IDDM2 (Insulin-dependent diabetes mellitus 2), SCZD1 (Schizophrenia disorder 1) (Full list EXCEL , TXT)		

- 1. Specify custom filters (IGNORE, EXCLUDE, INCLUDE) if you have.
- 2. Include any gene symbol longer than
 - A. You can specify if you want any acronym longer than some threshold even without positive confidence score.
- 3. Change score threshold.
 - A. For more information on scoring systems, click ‘more on scores’ or refer to the supplementary materials.
- 4. Do not include phenotype only genes
 - A. There are 676 phenotype only genes like IDDM2 (Insuline-dependent diabetes mellitus 2). You can exclude these genes. Default option is

‘ON’.

- vi. Click ‘**Submit Query**’ button and wait the process is completed. An email will be automatically sent to your registered email account with the result URL link.

Tip: It is strongly recommended that users first should review their SciMiner results and create own IGNORE, EXCLUDE, INCLUDE filters.

The IGNORE list may contain entities to be ignored. The INCLUDE and EXCLUDE lists of acronyms (or symbols) are included or excluded when conditions are met. For example, the default SciMiner EXCLUDE list has ‘SDS’ and ‘sodium dodecyl sulfate’ as its condition. Identification of ‘SDS’ in a text as ‘serine dehydratase’ will be excluded if there is an occurrence of ‘sodium dodecyl sulfate’ in the same document. In order to further improve the accuracy of mined targets, SciMiner allows users to manually edit target identifications on the mining result pages.

Once review is over, users are recommended to apply their own filters and run SciMiner with same query again to get an updated result.

B. Understanding the query result page











- i. Sections: The result pages will usually have 6 different sections.
 1. Search option summary: About your submitted query

SciMiner Result Report Page	
Congratulation! Your query with SciMiner has been successfully completed. Your query and search options were	
[HIDE search option summary]	
Query ID #	7
User name	Junguk Hur
Email	juhur@umich.edu
Query name	ALS_ROS
Query string	"Amyotrophic Lateral Sclerosis"[MeSH] AND "Reactive Oxygen Species"[MeSH]
PMID count #	111
Score threshold	0.1
Other	Remove_phenotype_gene, Use_symbol_length_option, Always_accept_symbol_longer_than: 6

2. Summary: Number of identified targets in a table

Summary			
SciMiner has analyzed (111) articles. [Documents Summary] [PMIDs]			
	Identified Targets		
	Before filtering	After filtering	Unique targets
Symbol-based mining	6135	4490	133
Name-based mining	737	645	99
Merged	6872	5135	187 (detail)

3. Top 10 most frequent targets: By number of papers

Top 10 most frequent targets ... (All 190 genes)						
Here are the top 10 most frequently found targets from your query. They are sorted primarily by the number of paper and then by the number of occurrence.						
[HIDE Top 10 most frequent targets]						
Rank	HUGO	Symbol	Target Name	#Occur	#Paper	MiMI
1	11179	SOD1	superoxide dismutase 1, soluble (amyotrophic lateral sclerosis 1 (adult))	2682	84	
2	19986	CYCS	cytochrome c, somatic	51	12	
3	1516	CAT	catalase	41	12	
4	11180	SOD2	superoxide dismutase 2, mitochondrial	34	11	
5	7873	NOS2A	nitric oxide synthase 2A (inducible, hepatocytes)	114	9	
6	990	BCL2	B-cell CLL/lymphoma 2	101	8	
7	399	ALB	albumin	16	8	
8	7872	NOS1	nitric oxide synthase 1 (neuronal)	182	6	
9	620	APP	amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease)	20	6	
10	12805	XDH	xanthine dehydrogenase	13	6	

- View the interaction network (MiMI): Build interaction networks among the targets. Clicking “Generate_Network” will launch Cytoscape and display interaction networks among the targets. Users can limit the targets to be displayed by applying additional filters.

View the interaction network (MiMI data)

You can create a interaction network of the targets found above. Interaction data is based on the MiMI (Michigan Molecular Interactions) database. The network will be visualized by Cytoscape with the support of MiMI plugin.

Minimum # Paper [Generate Network](#) [Download Symbols](#)

5. Top 10 most enriched MeSH terms (sorted by p-values)

Top 10 most enriched MeSH terms (sorted by p-value)

Here are the MeSH terms that are highly enriched in your query(corpus) when compared to all other PubMed documents. The comparison was done by Fishers Exact test. The default significance level is **0.05** and any p-value smaller than **0.05** is shown in red. Use the EXCEL file below to sort by enrichment fold (t-ratio / b-ratio).

[[HIDE Top 10 most enriched MeSH terms](#)]

Corpus size 112 Background(MedLine) 17760657 [Switch table detail](#)

Rank	MeSH Term	t+	p-value	t-ratio	enrichment folds
1	Amyotrophic Lateral Sclerosis	111	0	0.99	2111.3
2	Superoxide Dismutase	81	2.10e-202	0.72	490.3
3	Reactive Oxygen Species	59	2.81e-137	0.53	386.6
4	Oxidative Stress	44	1.57e-87	0.39	191.1
5	Motor Neurons	29	9.11e-55	0.26	157.5
6	Hydrogen Peroxide	28	1.42e-52	0.25	153.7
7	Superoxides	24	4.44e-49	0.21	226.1
8	Mutation	41	2.20e-47	0.37	27.8
9	Mice, Transgenic	22	7.31e-35	0.20	77.5
10	Free Radicals	18	2.01e-31	0.16	110.9

Full result ([HTML](#), [TXT](#), [EXCEL](#))

6. Result files: Contains the raw result files

Result files

Here are the raw result files of SciMiner mining. If you are interested in seeing all the individual identification results, these files are useful to you. These files can also be useful in your generating customary filters. For example, check the **Filtered out** results, to see any targets you are quite certain that should be included in the result. Put such targets into your **INCLUDE** filter list and run the same query again.

These files are usually large and we recommend you save these files and load them in a text editor or in EXCEL. Check [File Format](#) of these files.

	Before filtering	After filtering	Filtered out	Unique targets
Symbol-based mining	1MB	1MB	502KB	13KB
Name-based mining	220KB	193KB	27KB	10KB
Merged	1MB	1MB	19KB	19KB

!. Note that passed targets are those whose SciMiner score is above the specified threshold.

C. Two detailed result pages

There are two additional detailed pages, which includes a document-oriented or a target-oriented.

i. [Documents Summary]

This link takes you to a document based summary page.

PMID	TITLE	# of Genes	Links	Journal
15536073	The hyperglycemia-induced inflammatory response in adipocytes: the role of reactive oxygen species.	31	M E H P	J Biol Chem
18218985	Extracellular signal-regulated kinase 5 SUMOylation antagonizes shear stress-induced antiinflammatory response and endothelial nitric oxide synthase expression in endothelial cells.	28	M E H P	Circ Res

1. **# of Genes:** The total number of targets found in this article.

2. **Links:**




- A. **M:** MEDLINE (text format)
- B. **E:** Export EndNote citation
- C. **H:** HTML (full text) file
- D. **P:** PubMed summary

3. At the bottom of the page, there is a clickable link for downloading the citations (EndNote) of all the papers.

ii. [All *** genes]

This link takes you to a target based summary page. All of the found targets are displayed with the most frequent one showing at the top.

Here are the targets (genes/proteins) that JUMiner has mined.

Rank	HUGO	Symbol	Target Name	#Occur	#Paper	Matched_Terms	MiMI
1	6081	INS	insulin	305	21	insulin	
2	11892	TNF	tumor necrosis factor (TNF superfamily, member 2)	138	13	TNF-treated TNF- amp #x3b1 TNF-alpha tumor necrosis factor alpha tnf alpha	
3	6018	IL6	interleukin 6 (interferon, beta 2)	138	11	IL-6 interleukin 6 il 6	

1. **Field description**

- A. **Rank**
 - B. **HUGO:** HGNC (Human Gene Nomenclature Consortium) HUGO ID
 - C. **Symbol:** Official HUGO Symbol
 - D. **Target Name:** Official HUGO description
 - E. **#Occur:** Total number of occurrence
 - F. **#Paper:** Total number of papers in which the target has been found
 - G. **Matched_terms:** Actual matched terms in the articles
-
-

H. **MiMI:** JAVA Web Start link to launch Cytoscape with MiMI plugin.

2. Further links

A. **#Paper:** List all documents with the specific target found in

B. **Matched_Terms:** List all the incidences of actual matching with up to 100 characters at each side. (flanking text)

















D. Review the results and find possible error.

- i. Review the result thoroughly.
- ii. Usually frequently found targets are true and correctly identified targets. If the frequency is very low, it is possible that the identified target could be false-positive or wrong identification.
- iii. Flanking texts might be useful to determine if the identified targets are true genes/proteins.
- iv. If you have a list of wrong identification (false-positive), you can now improve the overall mining results by using customary filters. These filters can be used from your next queries.
- v. Users can also directly make change through ‘Edit’ button. Users should be very cautious in using this edit-feature not to make any incorrect change.

E. Manual Correction of identification

SciMiner allows users to correct or delete any wrong identification that users may notice. Any correction will be applied to the mining queries after the correction was made. This feature is available from detailed target list page.

- i. Go to the detailed target list page (Click All xxx genes)

154	10945	SLC1A7	solute carrier family 1 (glutamate transporter), member 7	11	1	EAAT5	
155	10939	SLC1A1	solute carrier family 1 (neuronal/epithelial high affinity glutamate transporter, system Xag), member 1	10	1	EAAC1 EAAT3 EAAC1-knockout	
156	11121	SMPD2	sphingomyelin phosphodiesterase 2, neutral membrane (neutral sphingomyelinase)	10	1	nMase	
157	2728	DDOST	dolichyl-diphosphooligosaccharide-protein glycosyltransferase	9	1	AGE-R1 OST48 age r1	
158	4311	GCLC	glutamate-cysteine ligase, catalytic subunit	9	1	GCLC glutamate cysteine ligase glutamate cysteine ligase catalytic subunit	
159	270	PARP1	poly (ADP-ribose) polymerase family, member 1	9	1	PARP-1 poly adp ribose polymerase poly adp ribose polymerase 1	
160	1062	BLVRA	biliverdin reductase A	8	1	BVR	
161	7211	MPG	N-methylpurine-DNA glycosylase	8	1	MPG n methylpurine dna glycosylase n methylpurine dna glycosylase mpg	
162	9237	PPARGC1A	peroxisome proliferator-activated receptor gamma, coactivator 1 alpha	8	1	PGC-1 pgc 1	
163	9816	RAD50	RAD50 homolog (S. cerevisiae)	8	1	Rad50	
164	10840	SHC1	SHC (Src homology 2 domain containing) transforming protein 1	8	1	SHC p66Shc p66SHC p66	
165	2859	DHCR24	24-dehydrocholesterol reductase	7	1	seladin-1 seladin 1	
166	4312	GCLM	glutamate-cysteine ligase, modifier subunit	7	1	GCLM GCL glutamate cysteine ligase modifier subunit	
167	7230	MRE11A	MRE11 meiotic recombination 11 homolog A (S. cerevisiae)	7	1	Mre11	
168	7660	NCF1	neutrophil cytosolic factor 1, (chronic granulomatous disease, autosomal 1)	7	1	p47	
169	10944	SLC1A6	solute carrier family 1 (high affinity aspartate/glutamate transporter), member 6	7	1	EAAT4	

- ii. Click 'Matched_Terms' column to see the all of the mined target details.

HUGO ID Detailed Result 7660						
HUGO ID	7660					
Symbol	NCF1					
Name	neutrophil cytosolic factor 1, (chronic granulomatous disease, autosomal 1)					
#Occurrence	7					
#Paper	1					

PMID	Match String	Actual String	Score	Flanking text	Edited by	Edit
18219386	p47	p47	0.6	protein p22 phox and recruitment of the activating cytosolic components p47 phox p67 phox and p40 phox are needed for function	Junguk Hur	EDIT
18219386	p47	p47	0.6	Upon cell activation p47 phox is phosphorylated thereby initiating translocation of the p47 phox	Junguk Hur	EDIT
18219386	p47	p47	0.6	activation p47 phox is phosphorylated thereby initiating translocation of the p47 phox /p67 p67 phox /p40 p40 phox complex to the	Junguk Hur	EDIT
18219386	p47	p47	0.6	phox /p40 p40 phox complex to the membrane where phosphorylated p47 phox binds to p22 phox	Junguk Hur	EDIT
18219386	p47	p47	0.6	Rac acts to coordinate the translocation of the p47 phox /p67 p67 phox /p40 p40 phox complex and is	Junguk Hur	EDIT
18219386	p47	p47	0.6	the same interacting partners but Nox1 and Nox2 are both p47 phox and Rac dependent (15	Junguk Hur	EDIT
18219386	p47	p47	0.6	Apocynin is believed to block the translocation of p47 phox /p67 p67 phox to Nox (23 and would	Junguk Hur	EDIT

- iii. 'Edited by' column will show any previous manual correction by users. Click 'EDIT' icon to change the current identification.

HUGO ID	Approved Symbol	Approved Name	Entrez Gene ID
10576	CLEC11A	C-type lectin domain family 11, member A	6320
9070	PLEK	pleckstrin	5341
15912	NSFL1C	NSFL1 (p97) cofactor (p47)	5908
6062	ING1	inhibitor of growth family, member 1	3621
7660	NCF1	neutrophil cytosolic factor 1, (chronic granulomatous disease, autosomal 1)	65361

Current Entry
 You have evoked this script for the following identification result. Note that **Match String** is the form in SciMiner dictionary, while **Actual String** is the actual form found in the document. **Actual String** has more variable forms.
 PMID: 18219386 HUGO ID: [7660](#) Match String: **p47** Actual String: **p47**

Modification Scope

**! MAKE SURE THAT YOU COMPLETELY UNDERSTAND WHAT YOU ARE DOING HERE !
 ! CHECK THE IDENTIFICATION RESULTS THOROUGHLY !
 ! USE HUGO ID FOR NEW ASSIGNMENT !**

If you are sure what you are going to do, select your modification option.

This specific finding

[Search NCBI Entrez Gene and find a relevant HUGO ID \(HGNC ID\)](#)

Enter new HUGO ID

[UPDATE](#) [DELETE](#)

If SciMiner found conflicting symbols, it displays such list. For example in the figure above, it shows five possible genes for the selected term 'p47'. Users are recommended to use NCBI Entrez Gene database (available on the same page) to check details of genes matching the term (here as 'p47').

The change can be made in five different levels:

This specific finding: Change is only applied to this identification.

Within this document: Change is only applied to the document for any identification with the exactly same HUGO ID, Match String, and Actual String.

Within this document - All by the matchString: Change is applied to only to the current document (PMID)for any identification has the exactly same HUGO ID and Match String.(Actual String is NOT checked.)

Within the query corpus: Change is applied to every document in the query for any identification with the exactly same HUGO ID, Match String, and Actual String.

Within the query corpus - All by the matchString: Change is applied to every document in the query for any identification has the exactly same HUGO ID and Match String.(Actual String is NOT checked.)

! Users should be completely certain what they are changing since this may affect the result of other users as well. We believe accumulated knowledge would greatly enhance the quality of SciMiner text-mining results in the long run. !

- iv. To update the identification, enter correct **HUGO ID** (Not Entrez Gene ID) in the box and click 'UPDATE'. Click 'Search NCBI Entrez Gene and find a relevant HUGO ID (HGNC ID)' to search in the Entrez Gene database. Detailed pages for each Entrez Gene record has HUGO ID as HGNC ID in the primary source.
- v. It should be noted that update or deletion only applied to the same 'Match String', 'Actual String' and HUGO ID. The case of the actual string also matters.

F. Custom filters

- i. **IGNORE:** any matching term to be ignored in name matching. (ex: cytoplasmic domain, protein, or too broad non-specific terms)
- ii. **INCLUDE:** any symbol to be included on given conditions in symbol matching. (ex: NF-KB => Then any case of NF-KB will be reported regardless of the SciMiner score)
- iii. **EXCLUDE:** any symbol to be excluded on given conditions in symbol matching.
(ex: SDS sodium dodecyl sulfate => If SDS is identified as 'serine dehydratase' and 'sodium dodecyl sulfate' is also in the same document, then this SDS is excluded from the target list.

G. Run SciMiner again

- i. If you have compiled your own custom filters or gone through the manual correction step, you need to run the same query. Make sure custom filters are correctly specified if you have ones.

V. Post-Mining Analysis

Post-mining analysis can only be run with any existing SciMiner mining result. Remember that analysis is done by comparing target set (query) with background set (query).

SciMiner Post-Mining Analysis

(Instruction): Click [HERE](#) for detailed help or move mouse pointer around to get a quick tip.

- 1) Click **'Retrieve Completed Query'** to retrieve currently available SciMiner results.
- 2) Enter query ID for tested (and background set)
- 3) Select analysis modules and click **'Start Analysis'** to submit the analysis.

(Section_1) Search your SciMiner mining results by

Query name :

(Section_2) Here are the list of your previous SciMiner results. (completed only)

Query Num	Mode	Name	Date	PMIDs	Targets	Link	Del

Query result to be tested Limit targets by and by list

Query result as background Limit targets by and by list

P-value for significance test Significant p-values will be in **red** in result.

(Section3) Select functional analysis modules to run.

<input type="checkbox"/>	Analysis module	Method	Test against	Note
<input type="checkbox"/>	Gene (Name) Enrichment	Fisher's exact	<input type="text" value="Selected Background Above"/>	
<input type="checkbox"/>	Gene Ontology (GO) Enrichment	Fisher's exact	<input type="text" value="Selected Background Above"/>	
<input type="checkbox"/>	MeSH Term Enrichment	Fisher's exact	<input type="text" value="Selected Background Above"/>	
<input type="checkbox"/>	Pathway Enrichment	Fisher's exact	<input type="text" value="Selected Background Above"/>	
<input type="checkbox"/>	Protein-Protein Interaction network of targets	T-test, Z-score (100 repetitions)	<input type="text" value="All HUGO Genes in SciMinerDB"/>	Based on Gaussian distribution

A. Introduction to three sections

- Section1:** Retrieve currently available (meaning completed query). Query name can be specified. No need to be the full name. Case insensitive.
 - Section2:** Specify target and background set you want to compare. You can use a subset of targets from the selected queries by changing the threshold.
 - Section3:** Select analysis modes and specify background set. Background set could be either the selected one in the above section2 or full documents in SciMiner or other whole set of all of Gene Ontology or PubMed.
-

* **Note** that the minimum threshold for the background set in section2 applies any background set defined by Section3. It is NOT limited only to the ‘Selected Background Above’.

B. Understanding the result

- i. **The test performed here is fisher’s exact test, which statistically assess whether two sets are different or not. From a 2x2 contingency table**

T: target set, **B**: background set

T+	T-
B+	B-

Consider the following case. SOD1 is found at

20 papers (total 25) from target

20 papers (total 50) from background

Then the founding of SOD1 in the two sets is significantly different?

20	5
20	30

This would give p-value of 0.001 (two-tailed)

- ii. **Most of the tables in the analysis result page have similar columns.**

1. T+, T-, B+, B- are explained above.
2. T-ratio = T+ / T-
3. B-ratio = B+ / B-

iii. Target Enrichment

In target enrichment result section, a brief summary of targets from tested and background sets is given, which is followed by the top 10 most enriched (by the p-value) targets are shown in a table. In the example below, **IL6** is the top 1 most significantly enriched target. The color represents in which set the target is enriched. **Red** color represents enrichment in the tested set and **green** implies enrichment in the background set. This color scheme applied to all other sections as well (GO, MeSH, Pathway enrichment tests)

Target Enrichment

Here is summary of the targets (genes and proteins) from each query (search) result. Please note that **Tested set** refers to the identified targets from the **Target Query** and **Background set** refers to the identified targets from the **Background Query**. If no background query number is specified above, then the background set is full HUGO gene set (for Gene Ontology terms, pathways, and protein-protein interactions) and full PubMed (for MeSH terms)

	Corpus size	# of targets (total)	# of targets (used)
Tested set	172	399	399
Background set	168	555	555

Top 10 most enriched Targets (sorted by p-value)


Here are the top 10 most enriched targets in the tested set.

[Switch table detail](#)

Rank	HUGOID	Symbol	Name	t+	p-value	t-ratio	enrichment folds
1	6018	IL6	interleukin 6 (interferon, beta 2)	1	1.30e-07	0.01	0.0
2	11892	TNF	tumor necrosis factor (TNF superfamily, member 2)	10	3.90e-07	0.06	0.2
3	9605	PTGS2	prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)	2	5.10e-07	0.01	0.1
4	1516	CAT	catalase	30	8.70e-06	0.17	5.9
5	4235	GFAP	glial fibrillary acidic protein	6	1.36e-03	0.03	0.3
6	31395	CDX8B	cytochrome c oxidase, subunit 8B pseudogene	13	1.51e-03	0.08	12.7
7	5992	IL1B	interleukin 1, beta	4	1.70e-03	0.02	0.2
8	5438	IFNG	interferon, gamma	3	1.78e-03	0.02	0.2
9	11180	SOD2	superoxide dismutase 2, mitochondrial	19	1.88e-03	0.11	4.6
10	11179	SOD1	superoxide dismutase 1, soluble (amyotrophic lateral sclerosis 1 (adult))	91	3.30e-03	0.53	1.4

Full result ([HTML](#), [TXT](#), [EXCEL](#)) [Back to top](#)

iv. GO Enrichment results



Gene Ontology Enrichment

Download a compressed file ([GO_Compressed.zip \(7.0 MB\)](#)) including all of the individual GO files. This zip file contains every raw and intermediate files, some of which are available on this web page.




	Corpus size	# of targets (total)	# of targets (used)
Tested set	172	399	399
Background set	168	555	555

Gene Ontology Enrichment Section is composed of 9 results. For the three Gene Ontology categories (**Biological Processes, Molecular Functions, and Cellular Components**), three different sets of Gene Ontology terms are used.

Explicit GO: Only use GO terms that are explicitly assigned to the target gene.

Full GO: Use both explicitly assigned GO terms and implicitly assigned terms. Implicit terms are all the parent terms of explicit GO terms from the GO tree structure.

Level 2~5: Use only the highest 4 levels of GO terms. These can generally provide functional categories due to their broad meanings. GO assignments are based on the full GO.

-  **Biological Processes**
-  **Molecular Functions**
-  **Cellular Components**

* Gene Ontology terms are obtained from external resources for each target.

- GO enrichment are performed in three different ways for each of three categories
(biological function, molecular mechanism, and cellular components)
 - Level 2~5:** Only GO terms between the top level 2 and 5 are used.
 - Explicit GO:** Use all GO terms regardless of the level. But only use GO terms explicitly assigned to each target (gene)
 - Implicit GO:** Include any implicitly inferred GO terms in calculating enrichment score. Not clear? Recall the structure of Gene Ontology hierarchical tree.
Suppose Gene-A has GO:0004601 (peroxidase activity). Then we implicitly assign the parent term GO:0016209 (antioxidant activity) to the Gene-A and use such information in enrichment analysis.
- Suppose that you are very interested in GO:0004601 (peroxidase activity) and you want to know what targets has or has not this GO from your

mined target sets. This is available in the compressed file ([GO Compressed.zip \(9.1 MB\)](#)) including all of the individual GO files.

A. Download and unzip this file. You can navigate to find such individual GO information with target lists.

v. MeSH Enrichment

MeSH Terms Enrichment

Here are the top 10 most enriched MeSH terms in the target query corpus. You may click each MeSH term to view the details at the NCBI MeSH browser.

Top 10 most enriched MeSH terms (sorted by p-value)

corpus size : 172 Background : 168 [Switch table detail](#)

Rank	MeSH Term	t+	p-value	t-ratio	enrichment folds
1	Reactive Oxygen Species	104	1.63e-39	0.60	101.6
2	Mitochondria	33	5.80e-08	0.19	10.7
3	Oxidation-Reduction	27	3.60e-06	0.16	8.8
4	Oxidative Stress	59	3.99e-06	0.34	2.6
5	Superoxide Dismutase	69	1.52e-04	0.40	1.9
6	Microglia	4	1.59e-04	0.02	0.2
7	Lipid Peroxidation	13	1.95e-04	0.08	bZero
8	Inflammation Mediators	1	6.97e-04	0.01	0.1
9	Disease Models, Animal	12	1.74e-03	0.07	0.4
10	Antioxidants	24	2.16e-03	0.14	3.3

Full result ([HTML](#), [TXT](#), [EXCEL](#)) [Back to top](#)

* MeSH terms are obtained from the MEDLINE records of the documents in the tested query and background set.

vi. Pathway Enrichment

The following table summarizes those targets that were used in this pathway enrichment analysis. Pathway information are derived from [KEGG](#) and [Reactome](#) databases for each target.

	# of targets	# of Targets with pathway info	# of unique pathways
Tested set	399	276	534
Background set	555	392	496

More full results are available.

Test set (all in text file format)	Pathway assignment	Matrix_Type1	Matrix_Type2
Background set (all in text file format)	Pathway assignment	Matrix_Type1	Matrix_Type2

[Back to top](#)

Top 10 most enriched Pathways (sorted by p-value)

Here are the top 10 most enriched pathways in the target query corpus. You may click each pathway ID or name to view the details at either [KEGG](#) or [Reactome](#) databases.

# of targets in test set		# of targets in background set				
399		555		Switch table detail		
Rank	PathwayID	Title	t+	p-value	t-ratio	
1	73894	DNA Repair	18	1.42e-06	0.05	
2	73890	Double-Strand Break Repair	8	8.98e-04	0.02	
3	hsa04080	Neuroactive ligand-receptor interaction - Homo sapiens (human)	8	2.08e-03	0.02	
4	hsa04060	Cytokine-cytokine receptor interaction - Homo sapiens (human)	20	3.61e-03	0.05	
5	73951	Homologous recombination repair of replication-independent double-strand breaks	6	5.24e-03	0.02	
6	157579	Telomere Maintenance	6	5.24e-03	0.02	
7	73888	Homologous Recombination Repair	6	5.24e-03	0.02	
8	hsa04514	Cell adhesion molecules (CAMs) - Homo sapiens (human)	3	1.95e-02	0.01	
9	hsa00620	Pyruvate metabolism - Homo sapiens (human)	8	2.05e-02	0.02	
10	hsa04610	Complement and coagulation cascades - Homo sapiens (human)	1	3.05e-02	0.00	

Pathway information is obtained from external resources for each target in the tested and background set. Matrix formatted files can be loaded into Excel for better display.

			Pathway ID; Name					
#				# of genes	hsa04210; Apoptosis - Homo sapiens (human)	hsa05222; Small cell lung cancer - Homo sapiens (human)	hsa05220; Chronic myeloid leukemia - Homo sapiens (human)	hsa05212; Pancreatic cancer - Homo sapiens (human)
HUGO	Symbol	Name	# of pathways					
1787	CDKN2A	cyclin-dependent kinase inhibitor 2A (m	8			*	*	*
6482	LAMA2	laminin, alpha 2 (merosin, congenital m	4		*			
7211	MPG	N-methylpurine-DNA glycosylase	1					
9907	RBMS1	RNA binding motif, single stranded inte	0					
1473	CANX	calnexin	1					
7553	MYC	v-myc myelocytomatosis viral oncogene	12		*	*		
20418	PDS5B	PDS5, regulator of cohesion maintenanc	0					
1535	RUNX1T1	runt-related transcription factor 1; tran	1					
11998	TP53	tumor protein p53	24	*	*	*	*	*
132	ACTB	actin, beta	8					
2339	CRABP2	cellular retinoic acid binding protein 2	0					
3437	ERCC5	excision repair cross-complementing ro	8					
9884	RB1	retinoblastoma 1 (including osteosarcon	25		*	*	*	*

vii. Interaction enrichment

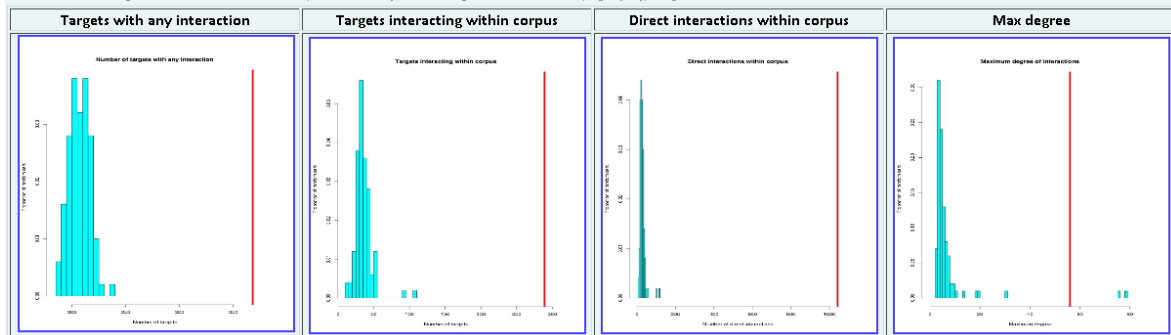
1. Network Significance Test

- A. This is to show that the mined targets are closely related with respect to your topic. Basically, we assume that closely related targets would highly interact with each other.
- B. Red bars represent the values of the tested set targets, while blue histograms are from randomly generated networks of the same size of the tested set.
 - i. **Targets with any interaction:** The number of targets that have any interaction data from the MiMI (Michigan Molecular Interactions) database.
 - ii. **Targets interacting within corpus:** The number of targets that have direct interactions with other targets in the given tested set. The higher this number is, the more closely related to each others.
 - iii. **Direct interactions within corpus:** The number of direct interactions among the tested targets.
 - iv. **Max degree:** The highest degree of maximally interacting partner.
- C. **100** random sets (**the size of target set**) are generated based on the frequency data in SciMiner database. The distribution of these 100 sets is compared with the target set data. The **red bar** in the figures represents the target set. A table of summary is followed by histograms of randomly generated network data as in the following figures.

This **Network Significance Test** section will test the integrity of the Protein-Protein Interaction network among the test set targets. This is based on the assumption that targets commonly related to a certain topic will be more likely to have frequent protein-protein interaction with each others. Therefore, if a set of targets identified by text-mining (SciMiner) from a certain query have more frequent direct interactions among the targets compared to randomly generated sets, it can, in part, support the validity of using text-mining method to identify related targets from a set of related papers. **100** random network of the same number of target in the test set will be generated from the background set. The test set had **399** targets. Thus 399 are randomly selected from the background which had a total of **25254** targets. If the background set was **full HUGO** set, all the targets will have an equal chance of being selected. If the background set was either **Selected background above** or **Whole document in SciMinerDB**, the probability of each target begin selected is determined by the **observed frequency** of each target as in (# of papers with the target / # of all papers). Two statistical measures are given below, standard Z-test and one-sample T-test.

	Targets with any interaction	Targets interacting within corpus	Direct interactions within corpus	Max degree
Tested	368	289	1042	56
Mean	206.56	34.91	27.18	6.35
STDEV	9.41	12.46	14.51	10.74
Z-Score	17.2	20.4	69.9	4.6
P-value(Z)	0.0e+00 (0.000)	0.0e+00 (0.000)	0.0e+00 (0.000)	1.9e-06 (0.000)
T-Stat	-171.6	-203.9	-699.5	-46.2
P-value(T)	2.5e-124 (0.000)	9.9e-132 (0.000)	1.1e-184 (0.000)	7.9e-69 (0.000)

This section will histograms of the random network samples. **Red bars** represent the targets from the tested set (target query) being tested.



-- END OF THE USER MANUAL --

If you have any question or comment, please email to Junguk Hur (juhur@umich.edu).