# SciMiner: Web-based literature mining tool for target identification and functional enrichment analysis

Authors: Junguk Hur[1*], Adam D. Schuyler[2], David J. States[3], and Eva L. Feldman[1, 2]

Affiliations: 1) Bioinformatics Program
University of Michigan
Ann Arbor, MI 48109, USA

2) Department of Neurology
University of Michigan
Ann Arbor, MI 48109, USA

3) School of Health Information Science
University of Texas at Houston
Houston, TX 77030, USA

* Corresponding Author
Junguk Hur juhur@umich.edu

# Supplementary Materials

## S1. Comparison with Other Tools

## S2. Full Text Sources

## S3. Annotation Resources

## S4. Dictionary Compilation and Expansion

## S5. Mining Rules

## S6. Confidence Scoring System

## S7. Performance Evaluation by BioCreAtIvE II Gene Normalization Task

## S1. Comparison with Other Tools

Currently available web-based biomedical text mining tools include EBIMed (Rebholz-Schuhmann, et al., 2007), ALI BABA (Plake, et al., 2006), and PolySearch (Cheng, et al., 2008). EBIMed provides a simple interface for identifying associations among named entities (genes/proteins, gene ontologies, drug names, and species). Ali Baba visualizes associations in a graphical way. And PolySearch provides more than 50 different classes of queries against various types of text, scientific abstract or biomedical databases.

However, these methods are limited in that (i) they only access MEDLINE abstracts as their literature data source, (ii) they do not allow users to edit the mining results, and (iii) they are unable to perform comparisons between search results of multiple queries.

Table 1 gives a detailed comparison of SciMiner features with the above mentioned tools. Compared features are as following.

1) **MEDLINE Abstracts**: MEDLINE abstracts are used as the source of the text data
2) **Full Text HTMLs**: Full text HTML documents are used as the source of the text data if available.
3) **Search terms as in PubMed**: Supports search terms as being used in PubMed query.
4) **Structured Query**: Available in PolySearch such as “Given X condition, find every Y”
5) **Accepting PMID list as input**: A list of PMIDs can be used as a search query
6) **Limit of document number**: The maximum number of documents per query.
7) **Genes and protein recognition**: Tool identifies genes and proteins names and symbols.
8) **Ambiguity (conflict) resolution**: Conflicting symbols are resolved.
9) **Other named entity recognition**: Other named entities like GO terms and drug names are identified by text-mining.

10) **User editability to increase accuracy**: Users can manually edit individual target identification results and additional filters of IGNORE, EXCLUDE, and INCLUDE may be used.
11) **Documents processed on the fly**: Documents are processed depending on users' queries. Previously processed documents by any previous queries would not need to be reprocessed, though.
12) **Minimum number of citations per target**: Number of associated documents per target in users' search can be specified to show such targets with a specified number of documents associated.
13) **Minimum score**: Certain score thresholds can be specified as an additional filter.
14) **Comparison among search results**: Comparisons can be performed among different search results (or different queries) in terms of target lists, GO terms, MeSH terms, and pathways.
15) **Functional enrichment (GO, Pathway, MeSH) of queries**: Enriched biological features can be identified by Fisher's exact test in comparisons among different search results.
16) **Result notification by email**: An email notice will be sent out to users when the results are ready.
17) **Downloadable results**: Mining and analysis results are available for download.
18) **Results are cached**: Processed documents will be kept in database to provide a quicker result in later queries. Users can maintain their search and analysis results in users' account.
19) **Highlighted target in abstract**: Identified targets are color highlighted. In SciMiner, this is limited to those targets from abstracts. However, SciMiner still lists the targets identified from full texts as well with some flanking texts for each identified target.
20) **Visualization of interacting targets**: Interactions (Protein-Protein Interaction) are visualized in Cytoscape based on the protein-protein interaction data from the MiMI database. Note that other tools will shows interactions among identified entities but these are not based on the actual molecular interaction data. They are based on co-occurrence in the processed documents.
21) **Links to PubMed**: Provides a URL link to the PubMed AbstractPlus page for each document.

22) **Links to journal HTML (publisher)**: Provides a URL link to the full text page for each document. These links take users to the publishers' webpage.

23) **Links to other databases**: Provides links to external databases including NCBI Entrez Gene, HGNC (HUGO Nomenclature, MiMI (Michigan Molecular Interactions), QuickGO, KEGG (Kyoto Encyclopedia of Genes and Genomes), Reactome, DrugBank, UniProt, HMDB (Human Metabolome Database), HPRD (Human Protein Reference Database), and GAD (Genetic Association Databsae)

24) **Document based summary**: Provides a document centric summary page for each processed document. SciMiner provides detailed list targets and related information, while others only highlight identified entities in the abstract.

25) **Execution time estimation**: Provides an estimated time for each query and analysis job.

26) **Standalone version**: Downloadable standalone package is available.

27) **Bulk data set**: Provides annotation and gene / protein details in a downloadable format. SciMiner provides some of these data on the public web version. All of the other data are available in the standalone package.

28) **User account**: Allows users to manage their previous search and analysis results. User account is essential since SciMiner allows user to merge and compare multiple search results.

29) **EndNote citation export for any documents**: Provides EndNote citation data for every processed document. Users can import these citation data directly from the SciMiner webpage without visiting PubMed or publishers' websites to download such citation files.

| | Features | SciMiner | EBIMed | Ali Baba | PolySearch |
|---|---|---|---|---|---|
| Text Data Source | MEDLINE Abstracts[1)] | O | O | O | O |
| | Full text HTMLs[2)] | O | X | X | X |
| Query Input | Search terms as in PubMed[3)] | O | O | O | X |
| | Structured query[4)] | X | X | X | O |
| | Accepting PMID list as input[5)] | O | O | O | X |
| | Limit of document number[6)] | 500-Unlimited[a)] | 10,000 | 10,000 | 500-Unlimited |
| Document processing | Genes and protein recognition[7)] | O | O | O | O |
| | Ambiguity (conflict) resolution[8)] | O | O | X | ? |
| | Other named entity recognition[9)] | X[b)] | drug, species, cells, diseases | GO, drug, species | Drug, disease, metabolite, tissue, cell |
| | User editability to increase accuracy[10)] | O | X | X | X |
| | Document processed on the fly[11)] | O | O | O | O |
| Filtering | Minimum number of citations per target[12)] | O | X | X | O |
| | Minimum score[13)] | O | X | X | O |
| Enrichment Analysis | Comparison among search results[14)] | O | X | X | X |
| | Functional enrichment (GO, Pathway, MeSH) of queries[15)] | O | X | X | Δ |
| Result | Result notification by email[16)] | O | X | X | O |
| | Downloadable result[17)] | O | Graph | X | X |
| | Results are cached[18)] | O | X | X | O |
| | Highlighted target in abstract[19)] | O | O | O | O |
| | Visualization of interacting targets[20)] | Via Cytoscape | O | X | X |
| | Links to PubMed[21)] | O | O | X | O |
| | Links to journal HTML (publishers)[22)] | O | X | X | X |
| | Links to other databases[23)] | NCBI Gene, HGNC, MiMI, QuickGO, KEGG, Reactome, NCBI MeSH | PubMed, MeSH, DrugBank, UniProt | QuickGO, NCBI Taxonomy viewer, DrugBank | PubMed, OMIM, DrugBank, UniProt, HMDB, HPRD, GAD |
| | Document based summary[24)] (Highlight in abstract) | O | O | O | X |
| Other features | Execution time estimation[25)] | O | O | O | X |
| | Standalone version[26)] | O | X | X | X |
| | Bulk data set[27)] | Δ[c)] | X | X | O |
| | User account[28)] | O | X | X | Δ[d)] |
| | EndNote citation export for any documents[29)] | O | X | X | X |

Table 1. Comparison of SciMiner features with other web-based literature mining tools

**Notes:**

a) The public version of SciMiner has a default of 500 document 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.

b) GO terms, MeSH terms, pathways, protein-protein interactions are not directly identified from the text data. Instead, SciMiner uses external annotation resources to associate identified targets to these entities. Thus these biological entities for each search will be summarized based on the targets identified.

c) Some of data are available on the web and full data are obtainable in the standalone distribution packages.

d) Though PolySearch does not employ user account system, it provides a job ID for each submission, so that it can be retrieved later.

## S2. Full Text Sources

SciMiner searches MEDLINE abstracts and available full text HTML documents. Figure 1 illustrates how users' queries are processed to retrieve literature data. Once a query is submitted to SciMiner, it is sent to the NCBI PubMed server to retrieve all of the resulting PMIDs (PubMed Unique Identifiers). The PMID list is compared to the current SciMiner database to determine which documents need to be retrieved and processed. Note that any documents that have been previously processed by other queries will not need to be reprocessed.

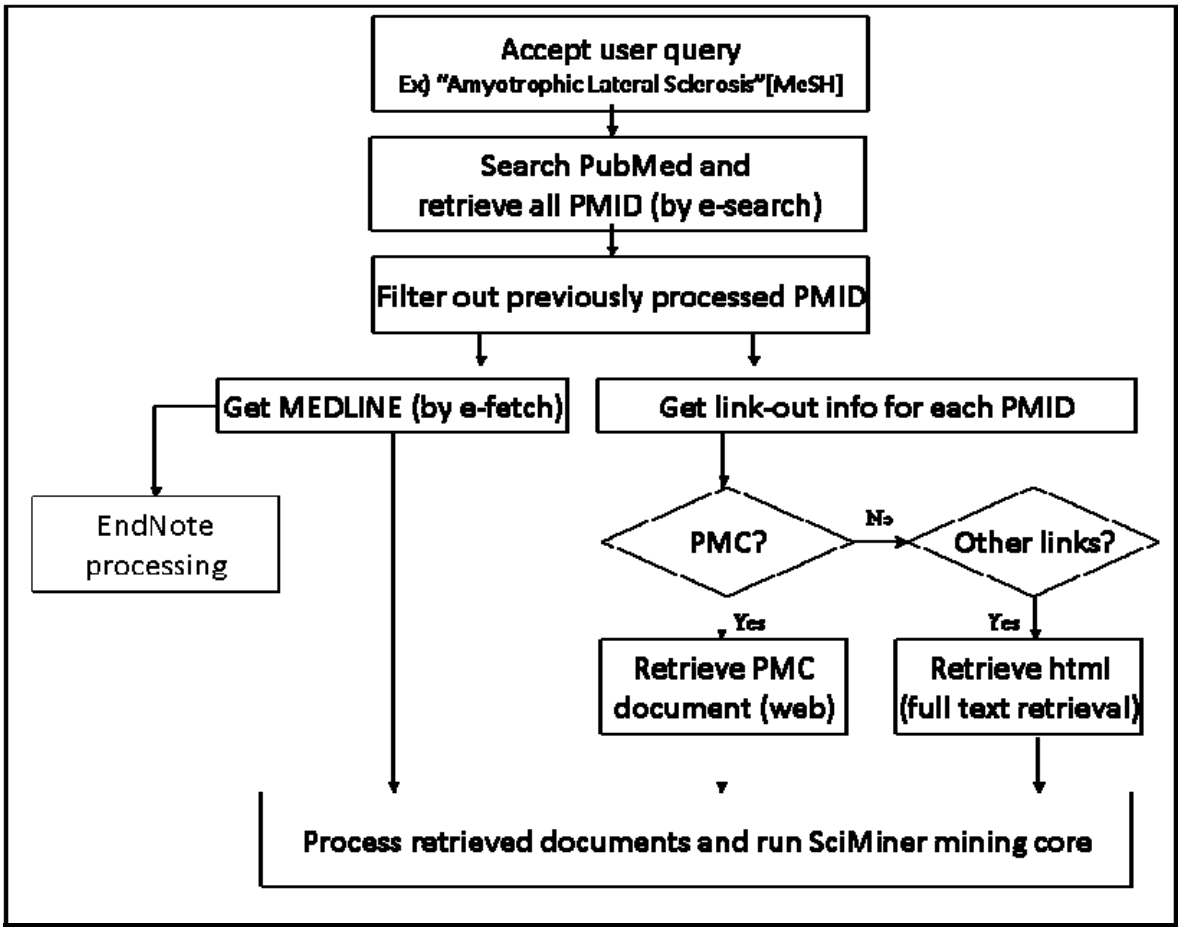


Figure 1. Schematic diagram of SciMiner query process and document retrieval

* PMC: PubMed Central

Each document's MEDLINE record is retrieved by NCBI e-fetch utility and processed accordingly. In order to retrieve available full text HTML, 'NCBI PubMed's link-out to journal' information is also fetched for each document to acquire the corresponding journal's URL. If multiple links are available, the PubMed Central (PMC) gets the highest priority. For URLs directing to an abstract page or a service provider selection page, SciMiner automatically tries to locate possible full text URLs and retrieve the full text HTML. Such publishers are noted as 'Multi' in the 'Retrieval Steps' column in Table 2. Depending on the subscription status of users' institutional library, full text availability is variable for the standalone local installation of SciMiner. Table 2 shows the list of journal publishers that are currently supported by SciMiner system. Note that PDF documents are not processed by SciMiner.

| **Publishers** | **Full Text** | | | **MEDLINE Abstract** | **Note** |
|---|---|---|---|---|---|
| | **Available** | **Supported by SciMiner** | **Retrieval Steps[1)]** | | |
| PubMedCentral[2)] | O | O | Single | O | |
| Nature | O | O | Multi | O | |
| Science | O | O | Single | O | |
| Elsevier (ScienceDirect) | O | O | Single | O | |
| Highwire | O | O | Single | O | |
| Blackwell synergy | O | O | Multi | O | |
| Informaworld | O | O | Multi | O | |
| Springer link | O | O | Multi | O | |
| Ovid | O | O | Multi | O | |
| Karger | O | O | Multi | O | |
| PortlandPress | O | O | Single | O | |
| Generic[3)] | O, X | O, X | Single | O | |
| Wiley Science | O | X | | O | PDF |
| Libert online | O | X | | O | PDF |
| Ingenta | O | X | | O | |

Table 2. Full text sources by Publishers

- O: available or supported, X: unavailable or unsupported

1) Retrieval steps
   A. Single: the NCBI Link-out URL directly points to the full text html.
   B. Multi: the NCBI Link-out URL is processed in multiple steps to get the full text html.
2) PubMed Central always gets the highest priority as the source of full text.
3) Generic implies all other publishers than those specified in Table 2.

Table 3 shows the current statistics in the SciMiner database as of 11/10/2008 with 92,089 documents processed. It should be noted that these statistics do not necessarily represent the whole literature available in PubMed due to its very small number of documents.

| Document Source Type | Publisher/Journal Type | count | percentage |
|---|---|---|---|
| Full text (if available) AND MEDLINE Abstract | Elsevier (ScienceDirect) | 19654 | 21.3% |
| | HighWire | 9672 | 10.5% |
| | Other (Generic) | 6333 | 6.9% |
| | PubMed Central | 6218 | 6.8% |
| | Wiley Science | 4874 | 5.3% |
| | Springer link | 2287 | 2.5% |
| | Ingenta | 1156 | 1.3% |
| | Nature | 1122 | 1.2% |
| | Ovid | 1066 | 1.2% |
| | Informaworld | 898 | 1.0% |
| | Libert online | 595 | 0.6% |
| | Karger | 559 | 0.6% |
| | Science | 159 | 0.2% |
| | Portland Press | 120 | 0.1% |
| | BlackWell Synergy | 78 | 0.1% |
| MEDLINE Abstract Only | NCBI (MEDLINE only) | 25312 | 27.5% |
| | PDFONLY | 4942 | 5.4% |
| | Failed | 444 | 0.5% |

Table 3. Proportion of publishers in the SciMiner database ordered by the percentage (a snapshot at 11/10/2008 with 92,089 documents processed)

Here are the web addresses for the publishers mentioned above (in alphabetical order). Refer to each publisher for more detail of their lists of published journals.

| Publisher | URL |
|---|---|
| Blackwell Synergy | http://www.blackwellpublishing.com/ |
| Elsevier (ScienceDirect) | http://www.elsevier.com/ |
| HighWire | http://highwire.stanford.edu/ |
| Informaworld | http://www.informaworld.com/ |
| Ingenta | http://www.ingenta.com/ |
| Karger | http://www.karger.com/ |
| Libert online | http://www.liebertonline.com/ |
| Nature | http://www.nature.com/ |
| Ovid | http://www.ovid.com/ |
| Portland Press | http://www.portlandpress.com |
| PubMed Central | http://www.pubmedcentral.nih.gov/ |
| Science | http://www.sciencemag.org/ |
| SpringerLink | http://www.springerlink.com |
| Wiley Science | http://www.interscience.wiley.com |

Table 4. Journal publishers' web addresses.

## S3. Annotation Resources

Targets (genes/proteins) used in the SciMiner system are based on the HUGO system, meaning that SciMiner reports are based on **human targets**. It should be noted that SciMiner system does NOT distinguish human genes from other species genes. Even though the actual biological detail might be different from species to species, SciMiner assumes that their overall biological functions are relatively well conserved among species if they are using same symbol or acronyms and this is enough to get an overview of biological functions. Thus, if a paper mentions superoxide dismutase 1 (Sod1) from mouse, SciMiner assigns this to SOD1 and disregards taxonomy information.

Annotation information for each HUGO entry is collected from the following resources. It should be noted that from the literature text data, SciMiner tries to identify only targets (genes and proteins). Other data types (pathway, Gene Ontology terms, MeSH, and Protein-Protein interactions) for identified targets are collected through external annotation resources listed below in Table 5.

| Type | Data Source Name | URL |
|---|---|---|
| Gene Protein | HGNC (HUGO Gene Nomenclature) | http://www.genenames.org/ |
| | NCBI Entrez Gene | http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene |
| Pathway | KEGG Pathway | http://www.genome.ad.jp/kegg/pathway.html |
| | Reactome Pathway | http://www.reactome.org/ |
| | NCBI Entrez Gene | http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene |
| PPI | MiMI (Michigan Molecular Interactions) | http://mimi.ncibi.org/ |
| MeSH | PubMed | http://www.ncbi.nlm.nih.gov/pubmed/ |
| Gene Ontology | Gene Ontology Consortium | http://www.geneontology.org |
| | NCBI Entrez Gene | http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene |

Table 5. External annotation databases

Mapping among database entries have been made available by in-house Perl scripts and the Clone/Gene ID converter {Alibes, 2007 #3}.

**S4. Dictionary Compilation and Expansion**

SciMiner uses two dictionaries, referred to as 'Symbol' and 'Name', compiled from the HGNC (HUGO Gene Nomenclature) database (http://www.genenames.org) and the NCBI Entrez Gene database (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene) previously known as LocusLink. The Symbol dictionary holds single word acronyms, while the Name dictionary contains longer descriptions (at least two words) of targets. In the current version (SciMiner 2.2), the Symbol dictionary has 83,735 unique entries and the Name dictionary has 136,827 unique entries. These dictionaries are extended to 87,014 and 263,304 entries, respectively, via the SciMiner dictionary expansion rules, which include relaxed special character handling and Greek character conversions such as TNF-alpha to TNF-A and TNFA.

- **Resources**
  - HUGO Gene Nomenclature (HGNC): http://www.genenames.org/. The following data columns from the raw data file of HGNC are used.
    - Approved Symbols
    - Approved Names
    - Previous Symbols
    - Previous Names
  - NCBI Entrez Gene: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA. The following data columns from the raw data file of HGNC are used.
    - OFFICIAL_SYMBOL
    - ALIAS_SYMBOL
    - OFFICIAL_GENE_NAME
    - ALIAS_PRODUCT
    - ALIAS_PROT

- For any conflicting symbols and names, the above mentioned 'Symbol' and 'Name' dictionaries contain a default assignment. Full conflict information is kept in separate files and used during SciMiner mining process. The following order of preference is

used for the default assignment of symbols or names to HUGO IDs.

- For Symbol dictionary:
  HUGO Approved Symbol > NCBI Official Symbol > HUGO Previous Symbols > NCBI Alias Symbol
- For Name dictionary:
  HUGO Approved Name > NCBI Official Gene Name > HUGO Previous Names > NCBI Alias Product > NCBI Alias Prot

- **Dictionary Expansions Rules**
  - For entries in the symbol dictionary:
    Greek words are replaced by their corresponding single characters if no such symbol already exists in the dictionary. These include Alpha, Beta, Gamma, and Kappa.
    - Dash '-' will be removed and the resulting symbol will be used as a new expanded entry unless there is already a preexisting entry assigned to different target in the dictionary.
      - Ex) TNF-alpha (HGNC ID: 11892)
        TNF-A and TNFA are added to the dictionary terms for HGNC ID:11892
  - For entries in the name dictionary:
    Special characters are removed and resulting names are added to the dictionary. Additional rules are applied as follows.
    - Phrase in parenthesis is removed.
      - Ex) AFG3 ATPase family gene 3-like 1 (yeast)
        ==> AFG3 ATPase family gene 3-like 1
    - Phrases before the first comma.
      - Ex) 39S ribosomal protein L10, mitochondrial
        ==> 39S ribosomal protein L10
    - Whole phrases after removing commas.
      - Ex) ANKRD26-like family C, member 1A
        ==> ANKRD26-like family C member 1A

- **English Dictionary**
  - General English words are not included in the SciMiner mining process. A dictionary was obtained from http://vburton.ncsa.uiuc.edu/wordlist.txt, which contains 135,000 words. Any word ending with '~ase' or those that are frequently used for gene symbols such as ski and Jun are removed from this dictionary.
  - This will further screen possible false positive findings by Symbol match method. For example, the gene 'CLOCK' will not be accepted if it is found by relaxed match as 'Clock' in text. For any symbols having same alphabets with symbol dictionary, the case should also be matched.

## S5. Mining Rules

### S5.1. Overall Target Recognition (Mining) Process

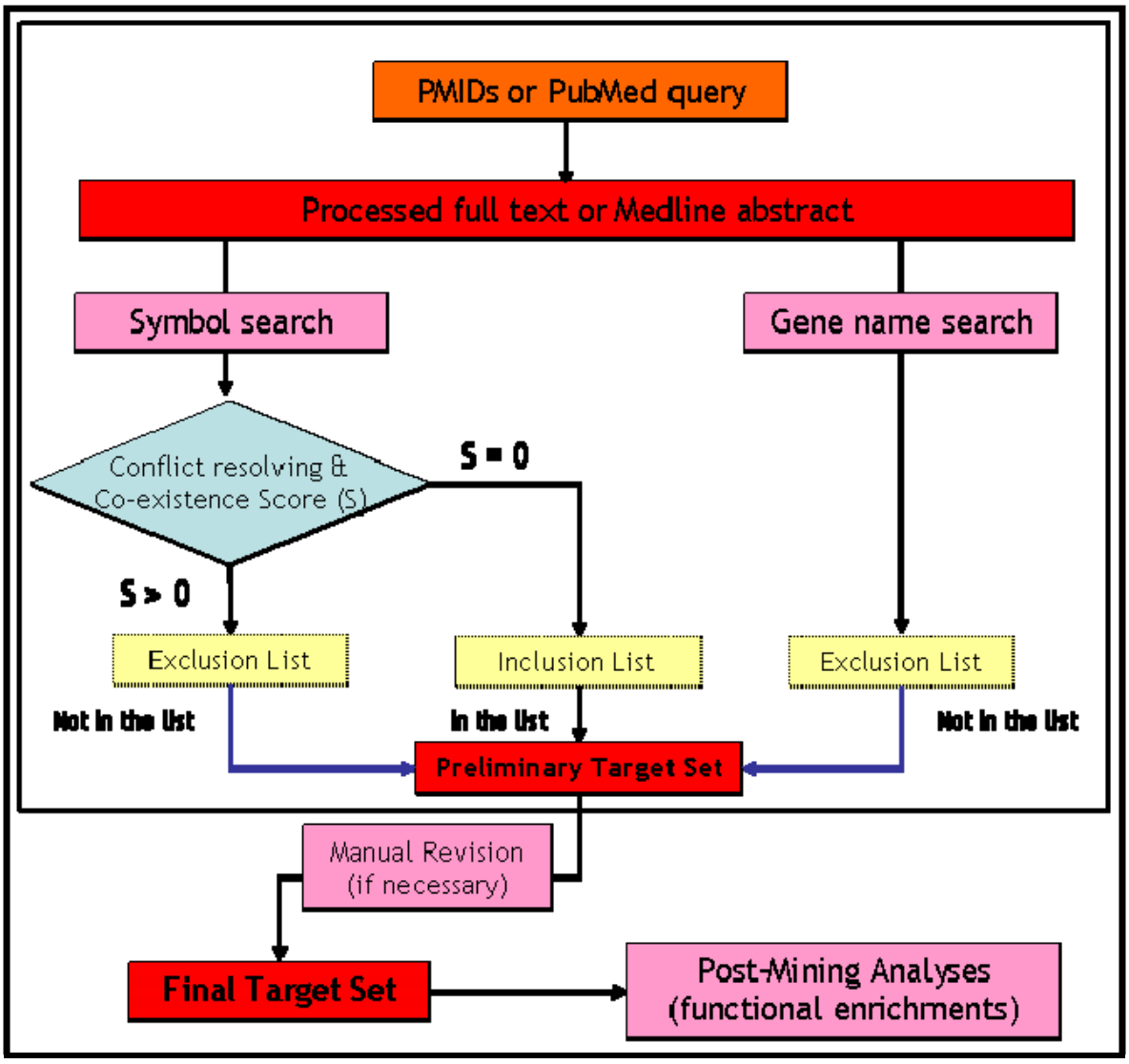


Figure 2. Schematic diagram of SciMiner Target Recognition Process

Retrieved documents (both MEDLINE records and full text HTMLs) are preprocessed for removal of unnecessary hyperlinks and some UTF-8 characters in text, and then are split by an in-house sentence-splitter into individual sentences. Sentences undergo the target recognition process via ‘Symbol search’ and ‘Gene name search’ depending on the base dictionary as introduced in ‘S2. Dictionary Compilation and Expansion’. In the current version of SciMiner v2.2, targets identified by Symbol search undergo a name-resolving

process, while those targets identified by Name search do not. This is because single-word symbols or acronyms can have many different meanings, while multi-word names are usually specific.

The following subsections S5.1.1 and S5.1.2 describe detailed steps and mining rules implemented in Symbol and Name searches.

S5.1.1. Symbol search

Key parts of the Symbol search are summarized below. There are some variations in the actual scripts. Those who are interested in the full detail are referred to 'SciMinerMining.pm' module which is available in the standalone version or upon request.

1) INCLUDE/EXCLUDE/IGNORE filters are loaded into a new hash table.
2) Symbol Dictionary loading.

   Pre-compiled symbol dictionary is loaded into two hash tables; one with keeping the case and another with making all in lower case from the second character. For example, SOD1 is hashed as 'SOD1' in the first hash and 'Sod1' in the second hash. The following rules are used to create new symbol candidate terms.

   A) Anything ending with –ALPHA, -BETA, -GAMMA, and –KAPPA to –A or A, -B or B, -G or G, and –K or K. Ex) PI4K-BETA => PI4K-B, PI4KB
   B) [All non-numeric characters]-[Numbers] to [All non-numeric characters][Numbers] without '-dash' in-between.
   C) Any symbol entry is filtered by the IGNORE list as well as the general English dictionary as introduced in the 'S4. Dictionary Compilation and Expansion' section above.
3) HUGO and NCBI Gene data are loaded into a new hash table.
4) Sentences are further split by a space ' ' and each word is checked against the hash tables of symbol dictionary. A word containing special characters like a slash '/' or a parentis '(' will be checked as a whole as well as split forms by such special characters. EX) Smad3/Smad4 => Smad3 Smad4
5) Checking against dictionary hash tables and further rules

A) Check the word keeping the case.

B) If no match has been found above, try to convert characters from the second position to the end into a lower case, while keeping the first character as it is and check hash table if there is a match. Ex) SOD1 -> Sod1

C) If no match has been found above and the word has the following pattern (^[h|m]([A-Z].*)), then [h|m] is removed and the remaining is checked. Ex) hPop1 => POP1, or mSin3a => SIN3A

D) If no match has been found above and the word contains a dash '-',

i. Starting or ending dash will be removed.
Ex) SOD1- => SOD1

ii. Ending –receptor will be tried as –R or R.
Ex) INS-receptor => INS-R, INSR

iii. Ending –(alpha|beta|gamma|kappa) will be tried as –A or A and etc.
Ex) TNF-alpha => TNF-A, TNFA

iv. Anything ending with the followings will be truncated
(1) like|dependent|specific|receptor|staining|induced|inducible|activated|repressed|stimulated|controlled|enhanced|mediated
Ex) SOD1-induced => SOD1

v. If no match has been found above and the word is in the following Perl pattern
/^(alpha|beta|gamma)(\d+)-(\S+)&/, it will be manipulated as in the example.
Ex) beta1-syntrophin => syntrophin beta 1

vi. Dashes are simply removed unless it is in (number)-dash-(number) pattern.

E) If no match has been found above and the word ends (alpha|beta|gamma|kappa), single green characters are replaced.
Ex) TNFgamma => TNF-G, TNFG

F) If no match has been found above and the word ends with a lower 's' which is a probable plural form.

i. If the word is in all lower case, then it is ignored.

ii. Otherwise, the word is checked against the dictionary alone.

iii. Accompanying words are also checked for any possible expansion forms.
Ex) SMADs 3 and 4 => SMAD3 and SMAD4

G) If no match has been found above and the word contains at least one upper case character, then the word is converted to all upper case and checked. If the word is in all lower case, then it is discarded.

H) Further expansions rules are applied if a match has been found from above (at $i^{th}$ position) and accompanying word ($i+1^{th}$ position) is either 'and' or 'to'.
Ex) SMAD3 and 4 => SMAD3 SMAD4

6) Confidence scoring calculation – Positive score will be assigned if there exist longer descriptions of the word (acronym) being tested in the same document (not only in the same sentence). Longer descriptions refer to the entries in the name dictionary and expanded names.

7) Conflict resolution – If the identified symbol has a conflict (belonging to a precompiled symbol-conflict set), confidence scores will be calculated for all of the possible candidates specified by the symbol-conflict set. Only the top-scoring will be chosen and reported.

8) Acronyms which are possibly not gene symbols will be further checked and filtered out.

A) Anything followed by (et al, buffer, score, version, medium, media, cell, software, program, algorithm, system, test, company, agent) and their plural forms if available. Ex) SPSS program or MES buffer

B) Anything that has 'acknowledge' or 'thank' in its flanking text. Usually identified words with these words are author names matched usually by the 5)B) step above. Sentences from the acknowledgement section are excluded in SciMiner process. However, such sentences are not corrected filtered out thus go through the mining process.

C) Anything that has a pattern of / ^[AGCTU]{3}$/i and the flanking text also has /([ACGTU]{3}\s{0,1}){2,}/i. This will filter out codons such as AGT, AUG, and etc.

9) Score boosting. Partial positive scores will be given if zero-scored match meet any of the following condition. Keep that in mind any legitimate symbol can have a zero-confidence score if there is no supporting longer form of gene names. This happens quite often where only an abstract is available.
   A) The matched word comes from a same block with a positive scored word.
      Ex) SOD1/NOS1 where scores were 0 for SOD1 and 0.5 for NOS1. Then SOD1 gets a partial score of 0.2.
      Our assumption is that if part of word being tested contains a positive scored match, the remaining would also be probably a gene symbol (as long as it is found as a match from the mining steps above).
   B) The matched word has one or more positive scored neighbors within two word distances.
      Ex) 'SOD1 TP53 NOX3 were up-regulated' where scores were 0 for SOD1, 2.5 for TP53, and 0.9 for NOX3. Then SOD1 gets a partial score of 0.2.
10) Any remaining matches will go through EXCLUDE/INCLUDE step.
   A) Any positive scored matches will be checked against EXCLUDE list. If it belongs to EXCLUDE list and a corresponding condition is found in the same document, then this positive scored match will be marked as 'EXCLUDED'.
   B) Any zero scored matches will also be checked against INCLUDE list. If it belongs to the INCLUDE list unless it also belongs to EXCLUDE list and its corresponding condition is met.

S5.1.2. Gene name search

Key parts of the Gene name search are summarized below. Compared to the Symbol search above, the Gene name search is much simpler in general. There are some variations in the actual scripts, thus those who are interested in the full detail are recommended to look in the 'SciMinerMining.pm' module which is available in the standalone version or upon request.

1) Dictionary loading. Pre-compiled Gene name dictionary is loaded into an array (@UNIQNAME) after the following processing.

A. Replacement of any special characters with a blank space.
B. Removal any multiple consecutive spaces into a single space.
C. Conversion into lower case.
D. Gene name entry should be at least 4-character long.
E. Additional hash of partial gene names. The official HUGO gene names are processed and used for confidence-scoring purpose only. These are NOT used as individual gene name identification but only used during confidence scoring calculation for Symbol search above. Anything that has a match in IGNORE list or English dictionary will not be included. Minimum word length is 3 to be included.

   Ex) FAS: Fas (TNF receptor superfamily, member 6)

   Fas: Not to be included since it's identical to the main entry.

   TNF: Included with a partial score of 0.3

   receptor: Not included since it is filtered by English dictionary.

   superfamily: Not included since it is filtered by English dictionary.

   member: Not included since it is filtered by English dictionary.

   Fas TNF: Included with a partial score of 0.3

   Fas TNF receptor: Included with a partial score of 0.3

   Fas TNF receptor superfamily: Included with a partial score of 0.3

   Fas TNF receptor superfamily member: Included with a partial score of 0.3

F. The main gene name array is sorted in an alphabetical order.
G. First four characters of each gene name are collected and the following two hash tables are generated. This is used to reduce the search space during gene name searching. Actual application is illustrated in the accompanying section. Four character threshold has been empirically chosen.
   i. %first4codeStart: Contain the starting index of @UNIQNAME for the entries starting with the given 4-character.
   ii. %first4codeEnd: Contain the last index of @UNIQNAME for the entries starting with the given 4-character.

2) Sentences are further split by a space ' ' and any special characters are removed.

3) The first four characters of every single word in the document are collected.
4) For each four-character code, candidate gene names are obtained from the @UNIQNAME array using %first4codeStart (as a starting index) and %first4codeEnd (as an ending index). These names are used in Perl regular expression to identify any occurrence in the full document.
   A. Since gene names are so variable in length, we cannot use hash-table based approach as in the Symbol search.
5) Any identification made above will be checked against the EXCLUDE filter.
6) Remaining identifications will be reported and the confidence score of 1 is assigned.

In the final step, Symbol based identification result and Gene name based identification result will be integrated into one based on HUGO.

## S6. Confidence Scoring System

The same acronym can be shared by multiple distinct targets, which becomes a major obstacle in correctly recognizing abbreviated forms of target names. This ambiguity is resolved with a confidence scoring scheme based on the co-occurrence of abbreviated symbols and longer descriptions in the same document. Unlike other system employing co-occurrence based approaches such as ProMiner (Hanisch, et al., 2005), SciMiner extends the co-occurrence search scope to the MEDLINE MeSH records and further allows partial name matches. This becomes particularly useful when only an abstract is available.

This score is used to 1) resolve the name conflict and 2) increase the precision.

Matches through the name dictionary are given a score of ‘1’ and do not go through the name resolution process. Matches through the symbol dictionary are checked for co-occurrence of longer descriptions (entries in the name dictionary and expanded forms). Scores are assigned to each match according to the following rules:

1) A score of ‘0.5’ is given for perfect matches with a longer description from the unique name dictionary.
2) A score of ‘0.3’ is given for a partial match to the approved name of the corresponding HUGO symbol or full match to expanded names.

The following rules are used to increase the overall accuracy of the mining result by minimizing possible false positives.

1) A score of 0.5 is given for a match preceding or followed by the following terms; gene(s), protein(s), mRNA(s)
2) A score of 0.3 is given for a match within in a same block of positively scored target (a single word in the original text but only separated by special characters) like ‘Bcl-xL/Bad’
3) A score of 0.2 is given for a match one word apart from other positive matches.
4) A score of 0.1 is given for a match two words apart from other positive matches.

5) Any sentence from Acknowledgement and authors will be excluded since they are very prone to have false positive matching of symbols to author names.

## S7. Performance evaluation by BioCreAtIvE II Gene Normalization Task

SciMiner's text mining performance was evaluated using the BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) II (Year 2006) Gene Normalization (GN) Task as a gold standard (Morgan, et al., 2008). The Gene Normalization task is to correctly identify the unique identifiers of genes and proteins mentioned in literature data. BioCreAtIvE II focused on identification of human genes and linking them to the NCBI Entrez Gene database. The gold standard set containing 785 human gene identifiers in a corpus of 262 abstracts was compiled by human experts for the task.

When the scoring scheme disabled, SciMiner identified 1,114 human gene identifiers from the same 262 abstracts. 677 identifiers were matched to the gold standard, indicating 86.2% recall, 60.8% precision, and 71.3% F-measure. With scoring scheme enabled and a score threshold of zero, SciMiner identified a total of 1,092 identifiers with 684 correct identifications showing 87.1% recall, 62.6% precision, and 72.9% F-measure. This implies that using scoring scheme based improves the precision of the target identification. It should be noted that without scoring scheme, an acronym conflict is not resolved and the default entry in SciMiner dictionary was simply reported. Unresolved conflicts have contributed to a slight increase in the total number of identifications (1,114 vs 1,092).

For example, the document of PMID 10072587 (titled as "Cloning of a novel gene (ING1L) homologous to ING1, a candidate tumor suppressor") includes the following sentence, "The ING1 gene encodes p33(ING1), a putative tumor suppressor for neuroblastomas and breast cancers, which has been shown to cooperate with p53 in controlling cell proliferation.". With the scoring scheme on, SciMiner correctly identified p33 as ING1, but without the scoring scheme, SciMiner incorrectly identified p33 as LTB (lymphotoxin beta TNF superfamily, member 3), which had 'p33' as one of its synonyms.

Table 6 shows the performance summary of SciMiner with various scoring thresholds applied and Figure 3 illustrates how these measures vary with different score thresholds. Recall decreases as score threshold increases. Precision improves as the score threshold increases to a value of 0.7, and then decreases slightly. This indicates that utilizing the scoring scheme increases the overall precision, but further optimization is required.

Utilizing the SciMiner scoring scheme and optimally tuning the score threshold parameter for each of the evaluation measures result in maximum values of 87.1% recall (at score threshold of zero), 71.3% precision (at score threshold of 0.7), and 75.8% F-measure (at score threshold of 0.3). Compared to the 54 BioCreAtIvE II Gene Normalization Task results posted by 20 groups (Morgan, et al., 2008), SciMiner's recall, precision and F-measure rank 2$^{nd}$, 34$^{th}$, 19$^{th}$, respectively

At low or zero confidence score thresholds, SciMiner shows very high recall rates, but also registers high numbers of false positive identifications leading to relatively low precision. However, it should be noted that SciMiner provides users with opportunities to improve the overall accuracy. Users are allowed to edit the identification results if they find any misidentified targets and they can also use custom filters (IGNORE, INCLUDE, and EXCLUDE) to improve accuracy.

| Score Threshold | Total Identification | True Positive | False Positive | False Negative | Recall | Precision | F-measure |
|---|---|---|---|---|---|---|---|
| No Scoring Scheme | 1114 | 677 | 437 | 108 | 0.862 | 0.608 | 0.713 |
| 0 | 1092 | 684 | 408 | 101 | **0.871** | 0.626 | 0.729 |
| 0.1 | 956 | 659 | 297 | 126 | 0.839 | 0.689 | 0.757 |
| 0.2 | 937 | 652 | 285 | 133 | 0.831 | 0.696 | 0.757 |
| 0.3 | 929 | 650 | 279 | 135 | 0.828 | 0.700 | **0.758** |
| 0.4 | 842 | 597 | 245 | 188 | 0.761 | 0.709 | 0.734 |
| 0.5 | 838 | 595 | 243 | 190 | 0.758 | 0.710 | 0.733 |
| 0.6 | 816 | 581 | 235 | 204 | 0.740 | 0.712 | 0.726 |
| 0.7 | 764 | 545 | 219 | 240 | 0.694 | **0.713** | 0.704 |
| 0.8 | 764 | 545 | 219 | 240 | 0.694 | 0.713 | 0.704 |
| 0.9 | 738 | 521 | 217 | 264 | 0.664 | 0.706 | 0.684 |
| 1 | 699 | 490 | 209 | 295 | 0.624 | 0.701 | 0.660 |
| 1.1 | 689 | 484 | 205 | 301 | 0.617 | 0.702 | 0.657 |
| 1.2 | 670 | 468 | 202 | 317 | 0.596 | 0.699 | 0.643 |
| 1.3 | 643 | 446 | 197 | 339 | 0.568 | 0.694 | 0.625 |
| 1.4 | 637 | 440 | 197 | 345 | 0.561 | 0.691 | 0.619 |
| 1.5 | 610 | 417 | 193 | 368 | 0.531 | 0.684 | 0.598 |

Table 6. Recall, precision, and F-score for multiple SciMiner confidence score thresholds.

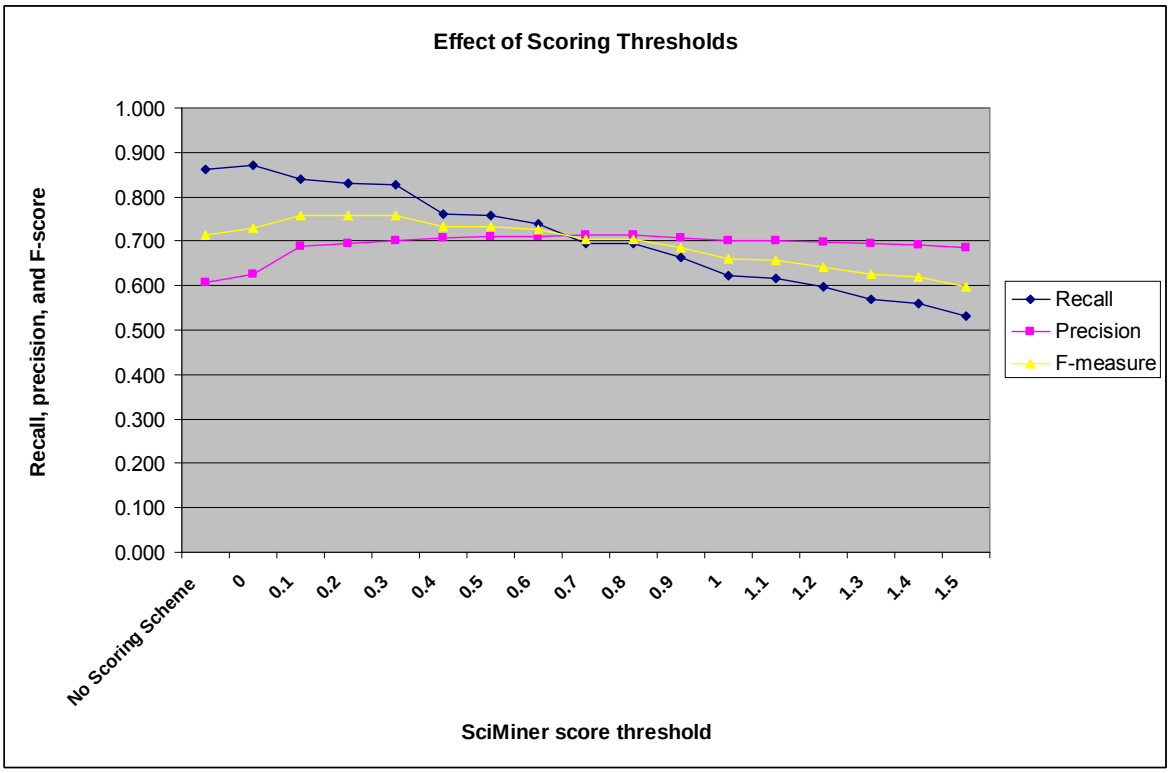


Figure 3. SciMiner Recall, Precision, and F-measure by different score threshold

After running the BioCreAtIvE test, additional rules were incorporated into SciMiner. The improved scoring results are shown in Table 7.

| Score threshold | | Total Identification | True Positive | Recall | Precision | F-score |
|---|---|---|---|---|---|---|
| SciMiner current version 2.2 | 0 | 1116 | 705 | **0.898** | 0.632 | 0.742 |
| | 0.1 | 977 | 679 | 0.865 | 0.695 | 0.771 |
| | 0.2 | 958 | 673 | 0.857 | 0.703 | 0.772 |
| | 0.3 | 951 | 672 | 0.856 | 0.707 | **0.774** |
| | 0.4 | 859 | 613 | 0.781 | 0.714 | 0.746 |
| | 0.5 | 854 | 611 | 0.778 | 0.715 | 0.746 |
| | 0.6 | 832 | 597 | 0.761 | **0.718** | 0.738 |
| | 0.7 | 777 | 557 | 0.710 | 0.717 | 0.713 |
| | 0.8 | 777 | 557 | 0.710 | 0.717 | 0.713 |
| | 0.9 | 750 | 532 | 0.678 | 0.709 | 0.693 |
| | 1 | 705 | 495 | 0.631 | 0.702 | 0.664 |
| | 1.1 | 695 | 489 | 0.623 | 0.704 | 0.661 |
| | 1.2 | 676 | 473 | 0.603 | 0.700 | 0.648 |
| | 1.3 | 645 | 447 | 0.569 | 0.693 | 0.625 |
| | 1.4 | 639 | 440 | 0.561 | 0.689 | 0.618 |
| | 1.5 | 611 | 417 | 0.531 | 0.682 | 0.597 |

Table 7. The performance of the current version of SciMiner v2.2.

## REFERENCES

Cheng, D., Knox, C., Young, N., Stothard, P., Damaraju, S. and Wishart, D.S. (2008) PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites, *Nucleic Acids Res*, **36**, W399-405.

Hanisch, D., Fundel, K., Mevissen, H.T., Zimmer, R. and Fluck, J. (2005) ProMiner: rule-based protein and gene entity recognition, *BMC Bioinformatics*, **6 Suppl 1**, S14.

Morgan, A., Lu, Z., Wang, X., Cohen, A., Fluck, J., Ruch, P., Divoli, A., Fundel, K., Leaman, R., Hakenberg, J., Sun, C., Liu, H.-h., Torres, R., Krauthammer, M., Lau, W., Liu, H., Hsu, C.-N., Schuemie, M., Cohen, K.B. and Hirschman, L. (2008) Overview of BioCreAtIvE II gene normalization, *Genome Biology*, **9**, S3.

Plake, C., Schiemann, T., Pankalla, M., Hakenberg, J. and Leser, U. (2006) AliBaba: PubMed as a graph, *Bioinformatics*, **22**, 2444-2445.

Rebholz-Schuhmann, D., Kirsch, H., Arregui, M., Gaudan, S., Riethoven, M. and Stoehr, P. (2007) EBIMed--text crunching to gather facts for proteins from Medline, *Bioinformatics*, **23**, e237-244.