BIOINFORMATICS APPLICATIONS NOTE
Vol. 25 no. 6 2009, pages 838–840
doi:10.1093/bioinformatics/btp049

Data and text mining

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

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

[1]Bioinformatics Program, [2]Department of Neurology, University of Michigan, Ann Arbor, MI 48109 and [3]School of Health Information Science, University of Texas at Houston, Houston, TX 77030, USA

Received on September 6, 2008; revised on January 9, 2009; accepted on January 21, 2009

Advance Access publication February 2, 2009

Associate Editor: Jonathan Wren

ABSTRACT

Summary: SciMiner is a web-based literature mining and functional analysis tool that identifies genes and proteins using a context specific analysis of MEDLINE abstracts and full texts. SciMiner accepts a free text query (PubMed Entrez search) or a list of PubMed identifiers as input. SciMiner uses both regular expression patterns and dictionaries of gene symbols and names compiled from multiple sources. Ambiguous acronyms are resolved by a scoring scheme based on the co-occurrence of acronyms and corresponding description terms, which incorporates optional user-defined filters. Functional enrichment analyses are used to identify highly relevant targets (genes and proteins), GO (Gene Ontology) terms, MeSH (Medical Subject Headings) terms, pathways and protein–protein interaction networks by comparing identified targets from one search result with those from other searches or to the full HGNC [HUGO (Human Genome Organization) Gene Nomenclature Committee] gene set. The performance of gene/protein name identification was evaluated using the BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) version 2 (Year 2006) Gene Normalization Task as a gold standard. SciMiner achieved 87.1% recall, 71.3% precision and 75.8% F-measure. SciMiner's literature mining performance coupled with functional enrichment analyses provides an efficient platform for retrieval and summary of rich biological information from corpora of users' interests.

Availability: http://jdrf.neurology.med.umich.edu/SciMiner/.
A server version of the SciMiner is also available for download and enables users to utilize their institution's journal subscriptions.

Contact: juhur@umich.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The PubMed database maintained by the National Center for Biotechnology Information (NCBI) is a key resource for biomedical science. It is a large and rapidly expanding dataset; more than 18 millions records from over 22 000 journals are indexed by PubMed today. With the increasing volume of the published biomedical literature, text mining has emerged as an increasingly important technology. The goal of biomedical text mining is to aid researchers in identifying relevant information more efficiently by having computers read the literature.

Currently available web-based biomedical text mining tools include ALI BABA (Plake *et al*., 2006), EBIMed (Rebholz-Schuhmann *et al*., 2007) and PolySearch (Cheng *et al*., 2008). 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 (see the Supplementary Material S1 for more details).

Here we present SciMiner, a web-based literature mining tool, which automatically collects MEDLINE records and available full text documents (see the Supplementary Material S2 for more details). Targets (genes and proteins) are extracted and ranked by the number of documents in which they appear. To provide an overall summary of their biological functions, these targets are further analyzed for their enrichments in Gene Ontology (GO) terms, pathways, Medical Subject Heading (MeSH) terms and protein–protein interaction networks based on external annotation resources (more details are available in the Supplementary Material S3). The performance of gene/protein name identification is assessed using the BioCreAtIvE II (Critical Assessment of Information Extraction systems in Biology) Gene Normalization Task (Morgan *et al*., 2008).

2 IMPLEMENTATION

2.1 Data management

SciMiner is implemented in Perl and uses a MySQL database to store compiled dictionaries and identified targets. Server–client communication is handled by CGI (Common Gateway Interface) scripts.

2.2 Target recognition

SciMiner uses a dictionary and rule-based approach to identify targets in the literature. The dictionaries, referred to as 'Symbol' and 'Name', are compiled from the HGNC [HUGO (Human Genome Organization) Gene Nomenclature Committee] genes and the NCBI Entrez Gene database entries annotated as human. The Symbol dictionary holds single word acronyms, while the Name dictionary contains longer descriptions of each target. SciMiner employs

*To whom correspondence should be addressed.

dictionary expansion rules, which include relaxed special character handling and Greek character conversions such as tumor necrosis factor (TNF)-alpha to TNF-A and TNFA. Detailed descriptions of the dictionary compilation and the expansion rules are available in the Supplementary Material S4. The target recognition rules defined on the above dictionaries are described in the Supplementary Material S5.

2.2.1 Confidence scoring scheme 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. Compared with other systems employing co-occurrence based approaches [e.g. 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 the full text bodies are not available. More details are given in the Supplementary Material S6.

2.2.2 User-provided filters and manual correction SciMiner accuracy is increased by allowing users to provide their own 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 identified targets on the mining result pages.

2.3 Post-mining analysis

Functional enrichment analyses are performed by comparing the identified targets of one search with those of other search results. Fisher's exact test (Fisher, 1922) is used to identify statistically significant overrepresentations of target list entries, GO terms, MeSH terms and pathways. This post-mining analysis step provides a simple but intuitive way to understand overrepresented biological functions.

2.4 Visualization

A summary is provided for the Target Recognition and post-mining analysis results and the full results are available as a web-page, a simple text file and an Excel file. The molecular interaction networks of the targets can be visualized in Cytoscape (Shannon *et al*., 2003) by following links from the Target Recognition result page. This functionality is enabled by the Cytoscape MiMI plug-in (Gao *et al*., 2009) and Java Web Start (http://java.sun.com/).

3 RESULTS AND DISCUSSION

3.1 Performance evaluation on BioCreAtIvE corpus

The performance of gene/protein name identification was evaluated using the BioCreAtIvE II (Year 2006) Gene Normalization Task as a gold standard, which tests for the correct identification of human genes against the NCBI Entrez Gene database. The gold standard set contains 785 human gene identifiers in a corpus of 262 abstracts.

With the confidence scoring scheme disabled, SciMiner identified 1114 human gene identifiers of which 677 matched the gold standard set. This corresponds to 86.2% recall, 60.8% precision and 71.3% F-measure. Utilizing the SciMiner scoring scheme and optimally tuning the score threshold parameter for each of the evaluation measures resulted in maximum values of 87.1% recall (at score threshold of 0), 71.3% precision (at score threshold of 0.7) and 75.8% F-measure (at score threshold of 0.3). Compared with the 54 BioCreAtIvE II Gene Normalization Task results posted by 20 groups (Morgan *et al*., 2008), SciMiner's recall, precision and F-measure rank second, 34th and 19th, respectively. A complete table of results is shown in the Supplementary Material S7.

3.2 Application

SciMiner was run on a query of 'Amyotrophic Lateral Sclerosis' and found 3226 targets from 10 625 documents as of August 31, 2008. The most frequently found target was superoxide dismutase 1 (SOD1) from 2198 papers, followed by amyloid beta (APP), ubiquitin (RPS27A), microtubule-associated protein tau (MAPT). Post-mining analysis identified 183 enriched pathways in these targets ($P < 0.001$). They include Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways of amyotrophic lateral sclerosis, apoptosis and signaling pathways (e.g. MAPK and JAK-STAT).

Post-mining analysis was performed between two subsets of the above corpus; Query1 ('Amyotrophic Lateral Sclerosis' and 'Reactive Oxygen Species') and Query2 ('Amyotrophic Lateral Sclerosis' and 'Inflammation'). This comparison identifies targets that are overrepresented in either 'Reactive Oxygen Species' or 'Inflammation' in the domain of 'Amyotrophic Lateral Sclerosis'. Query1 found 401 targets from 172 documents, while Query2 found 561 from 168 documents. Catalase (CAT) and SOD1 were highly overrepresented in the Query1 result, while TNF and interleukin-6 were highly overrepresented in the Query2 result. The pathway enrichment analysis further found that 'DNA repair' and 'cytokine-cytokine receptor interactions' were the most significantly enriched pathways from the targets of Query1 and Query2, respectively.

3.3 Discussion

SciMiner provides a convenient web-based platform for mining targets (genes and proteins) from the biomedical literature with the capacity for functional enrichment analyses. SciMiner performs well compared with other methods, but is unique in that it (i) searches full text documents (not just abstracts); (ii) allows users to directly edit the mining results; and (iii) allows comparisons to be made between search results of multiple queries.

Funding: National Institutes of Health (R01-LM008106 to D.J.S., U54-DA021519 (NCIBI) to D.J.S. and E.L.F., partial); Program for Neurology Research and Discovery and the Bioinformatics Graduate Program.

Conflict of Interest: none declared.

REFERENCES

Cheng,D. *et al*. (2008) PolySearch: a web-based text mining system for extracting relationships between human diseases. genes, mutations, drugs and metabolites. *Nucleic Acids Res.*, **36**, W399–W405.

Fisher,R.A. (1922) On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. *J. R. Stat. Soc.*, **85**, 87–94.

Gao,J. *et al*. (2009) Integrating and annotating the interactome using the MiMI plugin for cytoscape. *Bioinformatics*, **25**, 137–138.

Hanisch,D. *et al*. (2005) ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics*, **6**(Suppl 1), S14.

Morgan,A. *et al*. (2008) Overview of BioCreative II gene normalization. *Genome Biol.*, **9**, S3.

Plake,C. *et al*. (2006) AliBaba: PubMed as a graph. *Bioinformatics*, **22**, 2444–2445.

Rebholz-Schuhmann,D. *et al*. (2007) EBIMed–text crunching to gather facts for proteins from Medline. *Bioinformatics*, **23**, e237–e244.

Shannon,P. *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.