BIOINFORMATICS APPLICATIONS NOTE
Vol. 24 no. 11 2008, pages 1410–1412
doi:10.1093/bioinformatics/btn117

*Data and text mining*

# MedEvi: Retrieving textual evidence of relations between biomedical concepts from Medline

Jung-jae Kim*, Piotr Pęzik and Dietrich Rebholz-Schuhmann
EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received on November 8, 2007; revised on January 18, 2008; accepted on March 31, 2008
Advance Access publication April 9, 2008
Associate Editor: Alfonso Valencia

## ABSTRACT

**Summary:** Search engines running on MEDLINE abstracts have been widely used by biologists to find publications that are related to their research. The existing search engines such as PubMed, however, have limitations when applied for the task of seeking textual evidence of relations between given concepts. The limitations are mainly due to the problem that the search engines do not effectively deal with multi-term queries which may imply semantic relations between the terms. To address this problem, we present MedEvi, a novel search engine that imposes positional restriction on occurrences matching multi-term queries, based on the observation that terms with semantic relations which are explicitly stated in text are not found too far from each other. MedEvi further identifies additional keywords of biological and statistical significance from local context of matching occurrences in order to help users reformulate their queries for better results.

**Availability:** http://www.ebi.ac.uk/tc-test/textmining/medevi/

**Contact:** kim@ebi.ac.uk

## 1 INTRODUCTION

When exploring biomedical literature for information relevant to our research, we heavily rely on search engines (e.g. PubMed) which deliver us documents that match keyword-based queries. In the case of a query consisting of multiple keywords or terms, there is a need for restricting positional distance between occurrences of the terms in a document. If the terms are found too far from each other in the text, it is very likely that the text does not, at least not explicitly by means of the terms given, describe any relationship between concepts denoted by the terms. We regard this positional restriction as crucial in seeking relational information, for example, when users attempt to find textual evidence of relations between given concepts in the literature. We provide a novel tool to address this need with a special focus on the biomedical domain.

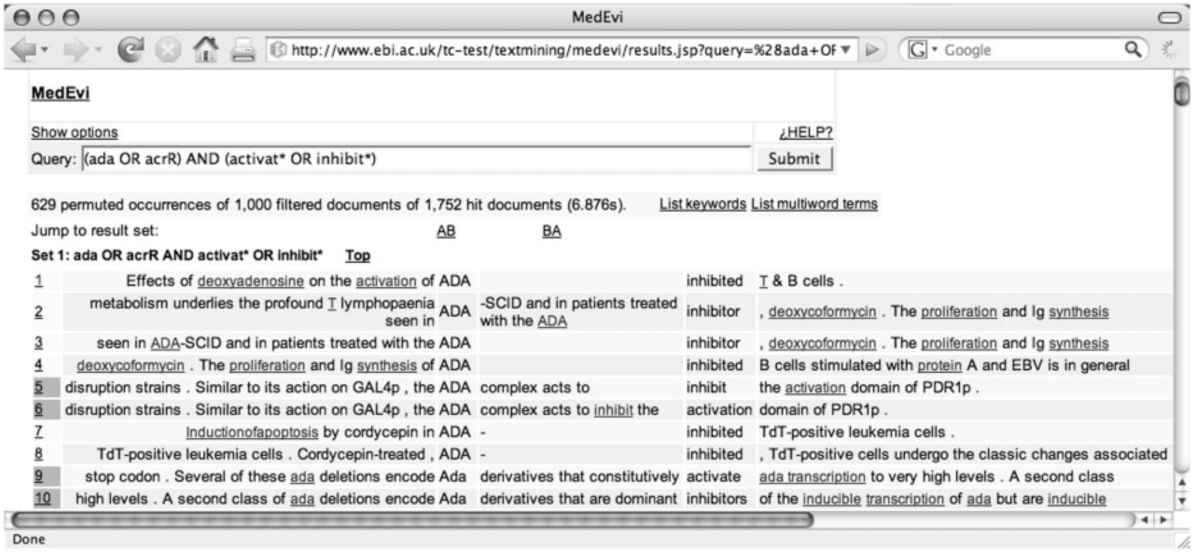
The tool presented here, named MedEvi, is a search engine that retrieves occurrences matching a given query with their local context. It is inspired by keywords-in-context (KWIC) concordancers, which have over the last few decades revolutionized the field of lexicography where different senses of lexical entries of dictionaries have to be defined in their authentic usage context (Sinclair, 1991). We believe that a concordancer is a good candidate to meet the above-mentioned tasks of information seeking, since it innately deals with the local context of matching occurrences where the evidence being searched is much more likely found than in other parts of the retrieved documents.

The common limitation of existing concordancers, however, is that they consider only single-term queries. To deal with multiple-term queries effectively, we implement the positional restriction on top of a concordancer. This feature of MedEvi is similar to the concept of proximity query (Baeza-Yates and Ribeiro-Neto, 1999), for example, as implemented in the proximity search of Lucene queries and the defined adjacency operator of OVID database queries. The difference between them is that while the latter is explicitly stated, if any, in query strings (e.g. 'A ADJn B'), the former is compulsorily applied to all queries where the distance between query terms, similar to 'n' of 'ADJn', can be adjusted by users.

MedEvi allows multi-term queries, composed with BOOLEAN operators (e.g. AND, OR). It is different from other existing search engines that also allow multi-term queries [e.g. PubMed (http://www.ncbi.nlm.nih.gov/sites/entrez), HubMed (http://www.hubmed.org)]. While the other search engines produce as results a list of MEDLINE abstracts, MedEvi directly browses text fragments that may eventually show semantic relations between given terms. It is different from other text mining tools that also browse text fragments, mostly sentences [e.g. iHOP (http://www.ihop-net.org/UniPub/iHOP/), MEDIE (http://www-tsujii.is.s.u-tokyo.ac.jp/medie/)]. While the text mining tools focus on certain biological entities like proteins (iHOP) (Hoffman and Valencia, 2005) and certain grammatical structures like subject-verb-object (MEDIE), MedEvi does not impose any syntactic or semantic restrictions, thus being widely used in any biomedical domains. We explain the features of MedEvi in the next section.

Users of MedEvi have found the tool useful to find evidence from the literature, for example, to see whether candidate chemicals are involved in a metabolic pathway, to identify the proteins that regulate given proteins, and to find whether a multi-term ontology concept actually appears in the literature even with a high degree of syntactic variations. Note that the applications above are generally concerned of semantic relations between biomedical concepts. Selected example queries can be found on the web page of MedEvi.

*To whom correspondence should be addressed.



MedEvi
http://www.ebi.ac.uk/tc-test/textmining/medevi/results.jsp?query=%28ada+OF
Google
MedEvi
Show options
¿HELP?
Submit
Done

Query: (ada OR acrR) AND (activat* OR inhibit*)

629 permuted occurrences of 1,000 filtered documents of 1,752 hit documents (6.876s). List keywords List multiword terms

Jump to result set: AB BA

**Set 1: ada OR acrR AND activat* OR inhibit*** Top

| | | | | | |
|---|---|---|---|---|---|
| 1 | Effects of deoxyadenosine on the activation of | ADA | | inhibited | T & B cells . |
| 2 | metabolism underlies the profound T lymphopaenia seen in | ADA | -SCID and in patients treated with the ADA | inhibitor | , deoxycoformycin . The proliferation and Ig synthesis |
| 3 | seen in ADA-SCID and in patients treated with the | ADA | | inhibitor | , deoxycoformycin . The proliferation and Ig synthesis |
| 4 | deoxycoformycin . The proliferation and Ig synthesis of | ADA | | inhibited | B cells stimulated with protein A and EBV is in general |
| 5 | disruption strains . Similar to its action on GAL4p , the | ADA | complex acts to | inhibit | the activation domain of PDR1p . |
| 6 | disruption strains . Similar to its action on GAL4p , the | ADA | complex acts to inhibit the | activation | domain of PDR1p . |
| 7 | Inductionofapoptosis by cordycepin in | ADA | - | inhibited | TdT-positive leukemia cells . |
| 8 | TdT-positive leukemia cells . Cordycepin-treated , | ADA | - | inhibited | , TdT-positive cells undergo the classic changes associated |
| 9 | stop codon . Several of these ada deletions encode | Ada | derivatives that constitutively | activate | ada transcription to very high levels . A second class |
| 10 | high levels . A second class of ada deletions encode | Ada | derivatives that are dominant | inhibitors | of the inducible transcription of ada but are inducible |

**Fig. 1.** Screen shot of MedEvi result set

## 2 SOFTWARE FEATURES

MedEvi receives a query either through the standard user interface in the entry page or via the advanced user interface available. It retrieves MEDLINE abstracts relevant to the query by using an Apache Lucene index (http://lucene.apache.org) that covers the whole set of MEDLINE abstracts and is updated on a bi-monthly basis. It then outputs hypertext that consists of aligned occurrences matching the query with hyperlinks attached to additional candidate keywords. Figure 1 shows an example output with the top 10 occurrences of the query "(ada OR acrR) AND (activat* OR inhibit*)".

### 2.1 Query syntax

The query syntax of MedEvi is based on the Lucene query syntax. Like Lucene, MedEvi allows both single terms and phrases, concept variables (see Section 2.3 for details), wildcards (i.e. $^*$, ?), BOOLEAN operators (only AND, OR) and escaping of special characters. It does not support field search and fuzzy search. Grouping in MedEvi is restricted to OR operators [e.g. '(Ada OR acrR) AND (activat$^*$ OR inhibit$^*$)'], while grouping for AND operators [e.g. '(Ada AND activat$^*$) OR (acrR AND inhibit$^*$)'] is not allowed. This restriction enables MedEvi to align occurrences of queries by the keywords in the queries, as exemplified in Figure 1.

### 2.2 Advanced search options

If a query string is successfully validated against the syntax, MedEvi searches the Lucene index with the query to retrieve MEDLINE abstracts. It then filters out abstracts that do not meet the default options or the options set through the advanced user interface. The options include maximum distance between keywords, range of publication dates of retrieved abstracts, maximum number of retrieved abstracts, criteria for sorting query occurrences. MedEvi also allows users to limit the search for occurrences of queries within sentence boundaries, as the sentence boundaries are often critical in relation extraction (Ding *et al*., 2002). Notice, however, that the experimental results of Ding and coworkers also support the necessity of positional restrictions that are narrower than sentence boundaries, for high precision of relation extraction. The details about the default offset of the search options, which were empirically chosen, are available on the help page of the MedEvi website.

### 2.3 Support of concept variables and database identifiers as query terms

MedEvi provides 10 variables for prevailing types of biomedical entities (e.g. cell, disease, drug, gene) to apply semantic restrictions to the search results. The functionality is inspired by the question-answering task of the Genomics Track in TREC 2007. For example of the question 'What serum [PROTEINS] change expression in association with high disease activity in lupus?', we may create a query like 'serum and [gene] and expression and lupus' for MedEvi to collect gene and protein names, which may be the answers of the question, into a column dedicated for the variable (i.e. [gene]). The details of the variables are available on the help page of the MedEvi website.

MedEvi also recognizes query terms that are UniProt accession numbers (e.g. P06134 for 'Ada'), and it automatically expands them to sets of synonymous terms, so that instead of specifying a set of names denoting a protein, one can use a UniProt accession number to locate strings associated with this accession number. The estimated precision and recall of the module for recognizing gene/protein names are 91.5% and 94%, respectively, when we accept nested terms as correct matches (Rebholz-Schuhmann *et al*., 2007).

### 2.4 Grouping of occurrences of queries

In the case of multi-term queries, MedEvi groups their occurrences by the order of query terms in the occurrences. For the example query of Figure 1, occurrences in the order of 'ada' and 'activat*' (i.e. AB in the display) are displayed before those in order of 'activat*' and 'ada' (i.e. BA).

### 2.5 Identification of additional candidate keywords

MedEvi automatically identifies additional candidate keywords which can be adopted by users for further narrowing the search results. As candidate keywords, it first recognizes gene and protein names, species names, drug names and Gene Ontology terms in the local context of the query occurrences. The estimated precision of the modules for the named entity recognition varies between 75% and 95% according to the types of named entities (Rebholz-Schuhmann *et al*., 2007). It then identifies nouns and verbs in the local contexts and scores them based on their frequencies in the results and in the whole set of MEDLINE abstracts by utilizing keyword extraction statistics (Oakes, 1998).

MedEvi provides three links for each additional candidate keyword to help users expand their queries: a link to add the keyword to the old query, another to replace the old query with the new keyword, and the other to show information of the keyword from well-known databases (e.g. UniProt, Gene Ontology).

### 2.6 Generating output pages

MedEvi outputs the result of a query in the form of an aligned hypertext. In the case of multi-term queries, the hypertext has a section for each permutation of terms. Each section has one or more rows that correspond to string occurrences matching the query. The occurrences are sorted by the relevance scores of their source documents, which are generated by the Lucene index according to the given query. If a document has multiple occurrences, they are displayed in adjacent rows whose index cells are uniformly coloured. The index column has links to PubMed web pages that have actual citation information for the source documents, while the citation information can be displayed in a pop-up window if the mouse cursor is placed onto the index column.

## 3 CONCLUSION

MedEvi is supplementary to existing search engines and text mining tools in the biomedical domain. It shows significant improvements in the presentation of results which offer new information seeking capabilities, by the combination of different search techniques such as concordance, positional restriction, semantic restriction and keyword lookup.

## ACKNOWLEDGEMENTS

Medline abstracts are provided from the NLM (Bethesda, MD, USA) and PubMed (www.pubmed.org) is the premier Web portal to access the data. Antonio Jimeno Yepes contributed to the improvement of MedEvi's semantic search capabilities. This work has been inspired by his contributions to the TREC Genomics Track competition 2007.

*Funding*: This research was sponsored by the EC STREP project 'BOOTStrep' (FP6-028099, www.bootstrep.org).

*Conflict of Interest*: none declared.

## REFERENCES

Baeza–Yates,R. and Ribeiro–Neto,B. (1999) *Modern Information Retrieval*. Addison-Wesley, Wokingham, UK.

Ding,J. *et al*. (2002) Mining MEDLINE: abstracts, sentences, or phrases? *In proceedings of Pacific Symposium on Biocomputing*. pp. 326–337, World Scientific Publishing Company, Singapore.

Hoffmann,R. and Valencia,A. (2005) Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, **21**(Suppl. 2), ii252–ii258.

Oakes,M.P. (1998) *Statistics for Corpus Linguistics*. Edinburgh University Press, Edinburgh, UK.

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

Sinclair,J.M. (1991) *Corpus, Concordance, and Collocation*. Oxford University Press, Oxford, UK.