BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 21 2008, pages 2559–2560
doi:10.1093/bioinformatics/btn469

*Data and text mining*

# FACTA: a text search engine for finding associated biomedical concepts

Yoshimasa Tsuruoka[1,2,*], Jun'ichi Tsujii[1,2,3] and Sophia Ananiadou[1,2]

[1]School of Computer Science, The University of Manchester, [2]National Centre for Text Mining (NaCTeM), Manchester, UK and [3]Department of Computer Science, The University of Tokyo, Japan

Received on June 13, 2008; revised on August 9, 2008; accepted on August 29, 2008

Advance Access publication September 4, 2008

Associate Editor: Jonathan Wren

**ABSTRACT**

**Summary:** FACTA is a text search engine for MEDLINE abstracts, which is designed particularly to help users browse biomedical concepts (e.g. genes/proteins, diseases, enzymes and chemical compounds) appearing in the documents retrieved by the query. The concepts are presented to the user in a tabular format and ranked based on the co-occurrence statistics. Unlike existing systems that provide similar functionality, FACTA pre-indexes not only the words but also the concepts mentioned in the documents, which enables the user to issue a flexible query (e.g. free keywords or Boolean combinations of keywords/concepts) and receive the results immediately even when the number of the documents that match the query is very large. The user can also view snippets from MEDLINE to get textual evidence of associations between the query terms and the concepts. The concept IDs and their names/synonyms for building the indexes were collected from several biomedical databases and thesauri, such as UniProt, BioThesaurus, UMLS, KEGG and DrugBank.

**Availability:** The system is available at http://www.nactem.ac.uk/software/facta/

**Contact:** yoshimasa.tsuruoka@manchester.ac.uk

## 1 INTRODUCTION

Information about pairwise association between biomedical concepts, such as genes, proteins, diseases and chemical compounds constitutes an important part of biomedical knowledge.[1] It is common for a researcher to need answers to questions like 'What diseases are relevant to a particular gene?' or 'What chemical compounds are relevant to a particular disease?' Text mining complements biomedical databases by providing researchers with a convenient way to find such information from the literature.

There are a number of web-based text mining applications which can be used for this purpose. EBIMed (Rebholz-Schuhmann *et al.*, 2007) receives a PubMed-style query from the user and analyzes the matched documents to recognize protein/gene names, GO annotations, drugs and species mentioned. Frequently occurring concepts are shown in a table, and the user can view the sentences corresponding to the associations. PolySearch (Cheng *et al.*, 2008) can produce a list of concepts which are relevant to the user's query by analyzing multiple information sources including PubMed, OMIM, DrugBank and Swiss-Prot. It covers many types of biomedical concepts including diseases, genes/proteins, drugs, metabolites, SNPs, pathways and tissues. Systems that provide similar functionality include XplorMed (Perez-Iratxeta *et al.*, 2003), MedlineR (Lin *et al.*, 2004), LitMiner (Maier *et al.*, 2005) and Anii (Jelier *et al.*, 2008).

Although these applications are useful in exploring such information in the literature, not many of them provide real-time responses—the users often have to wait for several minutes (or even hours) before they receive the results. Some of the systems provide reasonably quick responses by limiting the number of documents to be analyzed to a very small number (e.g. 500 abstracts), but such limitation leads to a significant deterioration of the coverage. LitMiner and Anii are exceptions in that they can return the result immediately, presumably thanks to pre-computed association statistics between the concepts. However, they do not accept a flexible query (e.g. free keywords or Boolean combinations of keywords/concepts), hence the concepts that can be specified by the user's query are limited to predefined ones.

To complement existing applications, we have developed FACTA, which is a text search engine for browsing biomedical concepts that are potentially relevant to a query. The distinct advantage of FACTA is that it delivers real-time responses while being able to accept flexible queries. This is achieved by online computation of association statistics—FACTA analyzes the documents retrieved by the query dynamically, using pre-indexed words and concepts.

## 2 SOFTWARE FEATURES

FACTA receives a query from the user as the input. A query can be a word (e.g. 'p53'), a concept ID (e.g. 'UNIPROT:P04637'), or a combination of these [e.g. '(UNIPROT:P04637 AND (lung OR gastric))']. The system then retrieves all the documents that match the query from MEDLINE using word/concept indexes. The concepts contained in the documents are then counted and ranked

*To whom correspondence should be addressed.

[1]In this article, a biomedical concept refers to a conceptual entity which is normally grounded to a record in a biomedical database. In text, the same concept (e.g. UniProt:O00203) may be represented by different terms (e.g. 'AP-3 complex subunit beta-1' or 'Beta3A-adaptin'). Note also that the same term may represent different concepts depending on the context, although this problem is currently not resolved in FACTA.

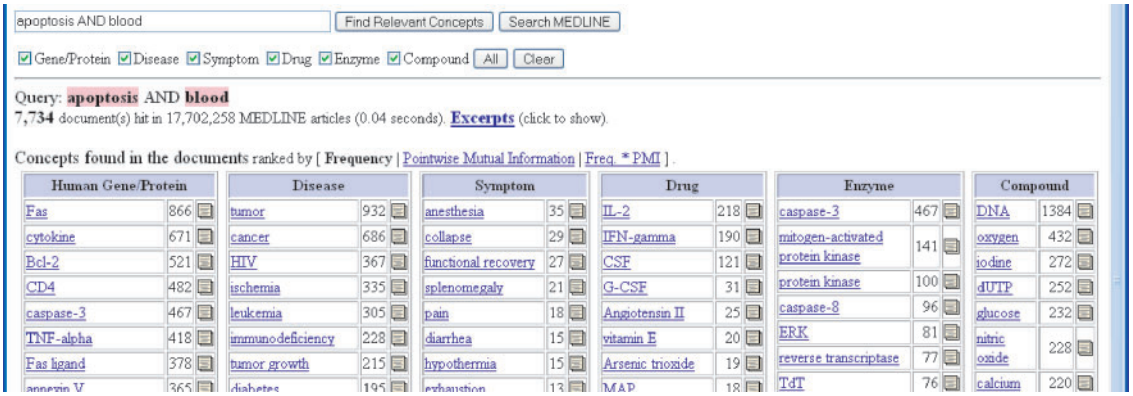


**Fig. 1.** A screenshot of FACTA search results.

according to their relevance to the query. The results are presented to the user in a tabular format.

Figure 1 shows an example of the search result. For the input query 'apoptosis AND blood', the system retrieved 7734 documents from MEDLINE in 0.04 s. The relevant concepts of six categories are displayed in a table and ranked by their frequencies. The document icon next to each concept name in the table allows the user to view snippets from MEDLINE and see textual evidence of the association. The user can also invoke another search by clicking a concept name in the table. This allows the user to explore associations between many different concepts in a highly interactive manner.

### 2.1 Indexing

FACTA's real-time responses to the queries are made possible by the use of its own indexing scheme and implementation of the analysis engines in C++. It uses two indexes built offline—one for the words and the other for the concepts. Both indexes are stored in memory to achieve quick responses, while the actual sentences of MEDLINE abstracts are stored on external storage. The system runs on a generic Linux server with 2.2 GHz AMD Opteron processors and 16 GB memory.

Currently, FACTA covers six categories of biomedical concepts: human genes/proteins, diseases, symptoms, drugs, enzymes and chemical compounds. The concepts appearing in the documents are recognized by dictionary matching. In total, 80 260 unique concepts are indexed. We used UniProt accession numbers as the concept IDs for genes/proteins and collected their names and synonyms from BioThesaurus (Liu *et al.*, 2006). We used UMLS (Humphreys and Lindberg, 1989) for diseases and symptoms. The concept IDs and names for drugs, enzymes and chemical compounds were collected from several databases including HMDB, KEGG and DrugBank.

Ambiguity causes problems in indexing. For example, the term 'collapse' is not necessarily used as a symptom name in the documents that produced the results shown in Figure 1, so ideally such occurrences should be disambiguated using the context and excluded from the counting for the category. There is also intra-category ambiguity, e.g. some protein synonyms can be mapped to multiple gene/protein IDs. These problems are currently not addressed in FACTA.

### 2.2 Ranking

Since the number of the concepts contained in the documents is usually very large, it is important that the concepts are properly ranked when presented to the user. Although frequencies are normally a good indicator of the relevance of a concept, they tend to overestimate the importance of common concepts. FACTA can also rank the concepts by using pointwise mutual information, which is defined as $\log p(x,y)/(p(x)p(y))$, where $p(x)$ is the proportion of the documents that match the query, $p(y)$ is the proportion of the documents that contain the concept, and $p(x,y)$ is the proportion of the documents that match the query and contain the concept. Pointwise mutual information gives an indication of how much more the query and concept co-occur than we expect by chance. For example, if their occurrences are completely independent (i.e. $p(x,y)=p(x)p(y)$), the measure gives a value of zero.

## ACKNOWLEDGEMENTS

The research team is hosted by the JISC/BBSRC/EPSRC sponsored National Centre for Text Mining.

*Funding*: Biotechnology and Biological Sciences Research Council (grant code BB/E004431/1).

*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.

Humphreys,B.L. and Lindberg,D.A.B. (1989) Building the unified medical language system. In *Proceedings of the 13th SCAMC*. pp. 475–480.

Jelier,R. *et al.* (2008) Anni 2.0: a multipurpose text-mining tool for the life sciences. *Genome Biol.*, **9**.

Lin,S.M. *et al.* (2004) MedlineR: an open source library in R for Medline literature data mining. *Bioinformatics*, **20**, 3659–3661.

Liu,H. *et al.* (2006) BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*, **22**, 103–105.

Maier,H. *et al.* (2005) LitMiner and WikiGene: identifying problem-related key players of gene regulation using publication abstracts. *Nucleic Acids Res.*, **33**, W779–W782.

Perez-Iratxeta,C. *et al.* (2003) Update on XplorMed: a web server for exploring scientific literature. *Nucleic Acids Res.*, **31**, 3866–3868.

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