Bioinformatics Advance Access published June 1, 2007

BIOINFORMATICS

# BioText Search Engine: beyond abstract search

Marti A. Hearst,[a], Anna Divoli [a], Harendra Guturu [a], Alex Ksikes [b], Preslav Nakov [a], Michael A. Wooldridge [c], and Jerry Ye [a]

[a]University of California, Berkeley, CA, [b]University of Cambridge, UK, [c]California Digital Library

Associate Editor: Prof. Alfonso Valencia

**ABSTRACT**

**Summary:** The BioText Search Engine is a freely available Web-based application that provides biologists with new ways to access the scientific literature. One novel feature is the ability to search and browse article figures and their captions. A grid view juxtaposes many different figures associated with the same keywords, providing new insight into the literature. An abstract/title search and list view shows at a glance many of the figures associated with each article. The interface is carefully designed according to usability principles and techniques. The search engine is a work in progress, and more functionality will be added over time.

**Availability:** http://biosearch.berkeley.edu

**Contact:** hearst@ischool.berkeley.edu, divoli@ischool.berkeley.edu

## 1 INTRODUCTION

Literature search is an important part of bioresearchers' work, both for keeping up with the latest developments in their area of expertise and for investigating new areas. A typical search starts at PubMed or other services such as BIOSIS, OVID, EMBASE, or using the search services of a specific journal or a publishing group's web site. Most of these search over title, abstract and document metadata, without making use of the full text. Alternative tools for searching MEDLINE abstracts have been developed; for instance HubMed, a simpler interface to PubMed (Eaton, 2006), eTBLAST, which returns abstracts similar to user-input text (Lewis *et al.*, 2006), and GoPubMed, which performs PubMed keyword-type search but classifies the returned abstracts using Gene Ontology (GO) terms (Doms and Schroeder, 2005).

On the Web, searching within the full text of documents has been standard for more than a decade, and much progress has been made on how to do this well. Full-text search of biology articles is often offered on a small subset of articles by publishing groups (e.g., Nature, Science, Highwire, Science Direct), and recently Google Scholar has begun offering search over the full text of journal articles, but with no special consideration for the needs of biologists.

Although researchers in the area of text mining have started investigating approaches for full-text analysis (e.g., BioCreative (Hirschman *et al.*, 2005) and TREC genomics (Hersh *et al.*, 2006)), the intellectual property restrictions until recently have made it impossible for any real advances in search interfaces for full text journal articles. However, the PubMedCentral Open Access journal collection now provides a substantial and unrestricted source for scientists to experiment with for providing full text search.

In this article, we present the BioText Search Engine, a freely available Web-based application that allows biologists to search over abstracts and figure captions of Open Access Journals, retrieving figures as well as their associated text. This idea is based on the observation, noted by our own group as well as many others, that when reading bioscience articles, researchers tend to start by looking at the title, abstract, figures, and captions. Figure captions can be especially useful for locating information about experimental results – a prominent example of this was seen in the 2002 KDD competition (Yeh *et al.*, 2003). Allowing search over captions in biology articles has been attempted before by FigSearch but in a very restricted manner, and in the form of a prototype (Liu *et al.*, 2004). Another project links the figures of a journal article to the corresponding sentence(s) from the abstract (Yu and Lee, 2006).

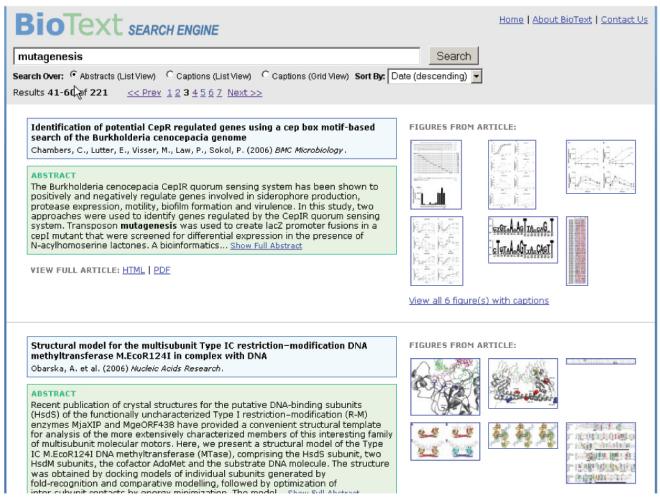


**Fig. 1.** For the query *mutagenesis*, results of searching over the titles and abstracts, showing many of the article's figures alongside its abstract.

## 2 SYSTEM DESCRIPTION

### 2.1 Design

We employ the principles of human-computer interaction for the design and development of the interface, meaning we solicit reactions from biologists both in person and remotely. We prototype, test, and revise the design based on user response, and we apply user interface design guidelines and principles (Hearst *et al.*, 2007).

The current design consists of an interaction flow in which users can search over either the text of abstracts (plus titles, author names, and other metdata), see Figure 1, or search over the text of the captions, see Figure 2. The results can be viewed either in a list view (in the case of abstract search and caption search) or in a grid view (in the case of caption search), see Figure 3.

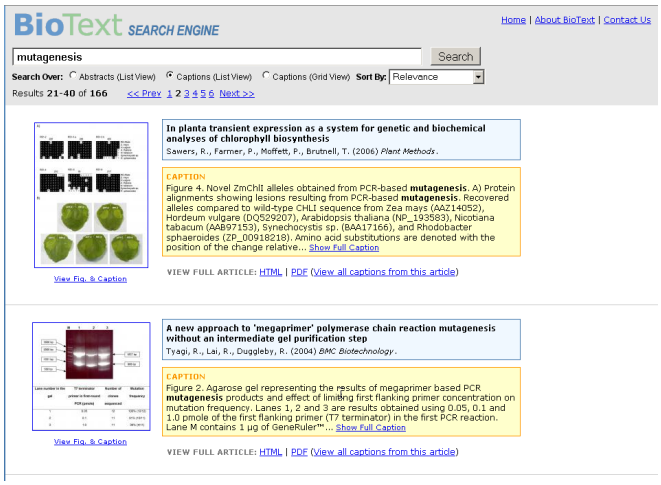


**Fig. 2.** For the query *mutagenesis* results of searching over the captions, showing the corresponding figures.

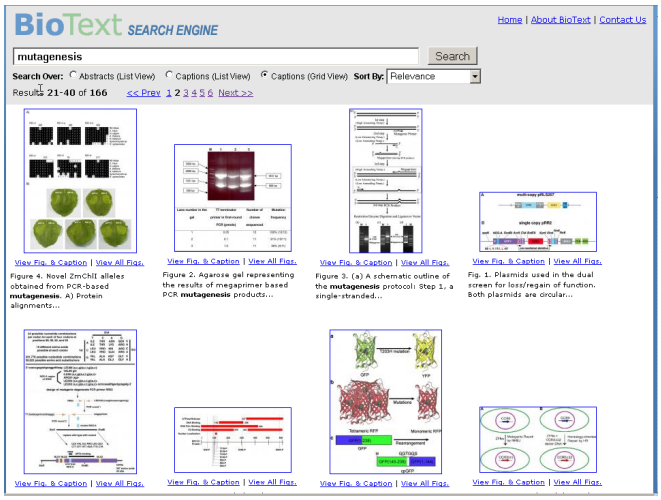


**Fig. 3.** For the query *mutagenesis*, results of searching over the captions, showing the corresponding figures in the grid view.

## 2.2 Functionality

As mentioned above, figure captions contain important information about experimental methods. For example, searching on "Western Blot" in the current collection produces few results when run only over title and abstract text, but returns more than a thousand results in caption search (note that caption search does not currently also search over abstracts). Similar behavior is seen for the queries *PCR*, "*phylogenetic tree*", and "*sequence alignment*". The grid view may be especially useful for seeing commonalities among topics, such as all the phylogenetic trees that include a given gene, or seeing all images of embryo development of some species.

## 2.3 Implementation

The current system indexes all Open Access articles available at PubMedCentral. This collection consists of more than 150 journals, 20,000 articles, and 80,000 figures (new articles are downloaded daily). The figures are stored locally, in order to be able to present thumbnails quickly. The Lucene open source search engine is used to index, retrieve, and rank the text (using the default statistical ranking). Publication date is stored as a separate field and can also be used to sort the results). For tokenization, the standard analysis settings for Lucene are used: words are split at punctuation characters and hyphens, unless there is a number in the token, and uses lowercasing, simple stemming, and stopword removal. The interface is web-based and is implemented in python and PHP. Logs and other information are stored using MySQL.

# 3 FUTURE WORK

In the near future we will provide full-text search, but since the usability of different ranking functions for biology articles is still not well-understood, we plan to do extensive usability testing before supporting this feature. One issue is whether or not different sections should be weighted differently for different query types, (e.g., Shah *et al.*, 2003). We are also investigating how best to show excerpts or summaries from full text.

We also plan to augment the caption search by indexing the parts of the full text that refer to the caption, and to provide search over table captions, to complement the figure caption search. We will also incorporate filtering by metadata such as author and journal name, and topical features such as genes/proteins, organisms, and species.

For the grid view, we plan to provide grouping according to categories that are of interest to biologists, such as sequence alignments and phylogenetic trees. To this end, we are in the process of building a classifier for figures and their captions, in order to allow for grouping by type. We have developed an image annotation interface and are soliciting help with hand-labeling mated caption classifier.

Additional future developments on the BioText search engine will depend on feedback and requests we receive from users, and the results of usability testing.

*Acknowledgements:* This work was funded in part by NSF DBI-0317510.

# REFERENCES

Doms, A. and Schroeder, M. (2005). GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Research*, **33**, 783–786.

Eaton, A. (2006). HubMed: a web-based biomedical literature search interface. *Nucleic Acids Research*, **34**(Web Server issue), W745.

Hearst, M. A., Divoli, A., Ye, J., and Wooldridge, M. A. (2007). Exploring the efficacy of caption search for bioscience journal search interfaces. In *ACL 2007 Workshop on BioNLP*.

Hersh, W., Cohen, A., Roberts, P., and K., R. H. (2006). TREC 2006 Genomics Track Overview. *The Fifteenth Text Retrieval Conference*.

Hirschman, L., Yeh, A., Blaschke, C., and Valencia, A. (2005). Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, **6**, 1.

Lewis, J., Ossowski, S., Hicks, J., Errami, M., and Garner, H. (2006). Text similarity: an alternative way to search MEDLINE. *Bioinformatics*, **22**(18), 2298.

Liu, F., Jenssen, T.-K., Nygaard, V., Sack, J., and Hovig, E. (2004). FigSearch: a figure legend indexing and classification system. *Bioinformatics*, **20**(16), 2880–2882.

Shah, P., Perez-Iratxeta, C., Bork, P., and M.A., A. (2003). Information extraction from full text scientific articles: where are the keywords? *BMC Bioinformatics*, **4**(20).

Yeh, A., Hirschman, L., and Morgan, A. (2003). Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *BIOINFORMATICS*, **19**(1), i331–i339.

Yu, H. and Lee, M. (2006). Accessing bioscience images from abstract sentences. *Bioinformatics*, **22**(14), e547.