



# BioKI:Enzymes — an adaptable system to locate low-frequency information in full-text proteomics articles

Sabine Bergler, Jonathan Schuman, Julien Dubuc, Alexandr Lebedev

bioki@cse.concordia.ca

## 1 Goals

BioKI:Enzymes locates low-frequency information in full texts with convenient display of the context at different levels of detail. Designed as a two-step process, BioKI:Enzymes allows the user a high degree of control over a keyword-based ranking of articles retrieved by the PubMed Central search engine.

Explicit control parameters displayed on the user interface for direct manipulation by the user is an important step in transparency and adaptability. A “Refine search” button on the results page brings the user back to all the previous settings, allowing exploration of the retrieved PMC subcorpus by successively more accurate rerankings through keyphrase refinement, until the desired goal is reached.

**What is low-frequency information?** Information that is not redundant, that is which does not necessarily occur in many articles, and within an article, may be expressed only once (most likely in the body of the article, not the abstract) is considered *low-frequency* information. It contrasts with the information usually queried for QA systems, where the answer might be found in any of a number of documents and it is in fact the redundancy that identifies the most likely answer.

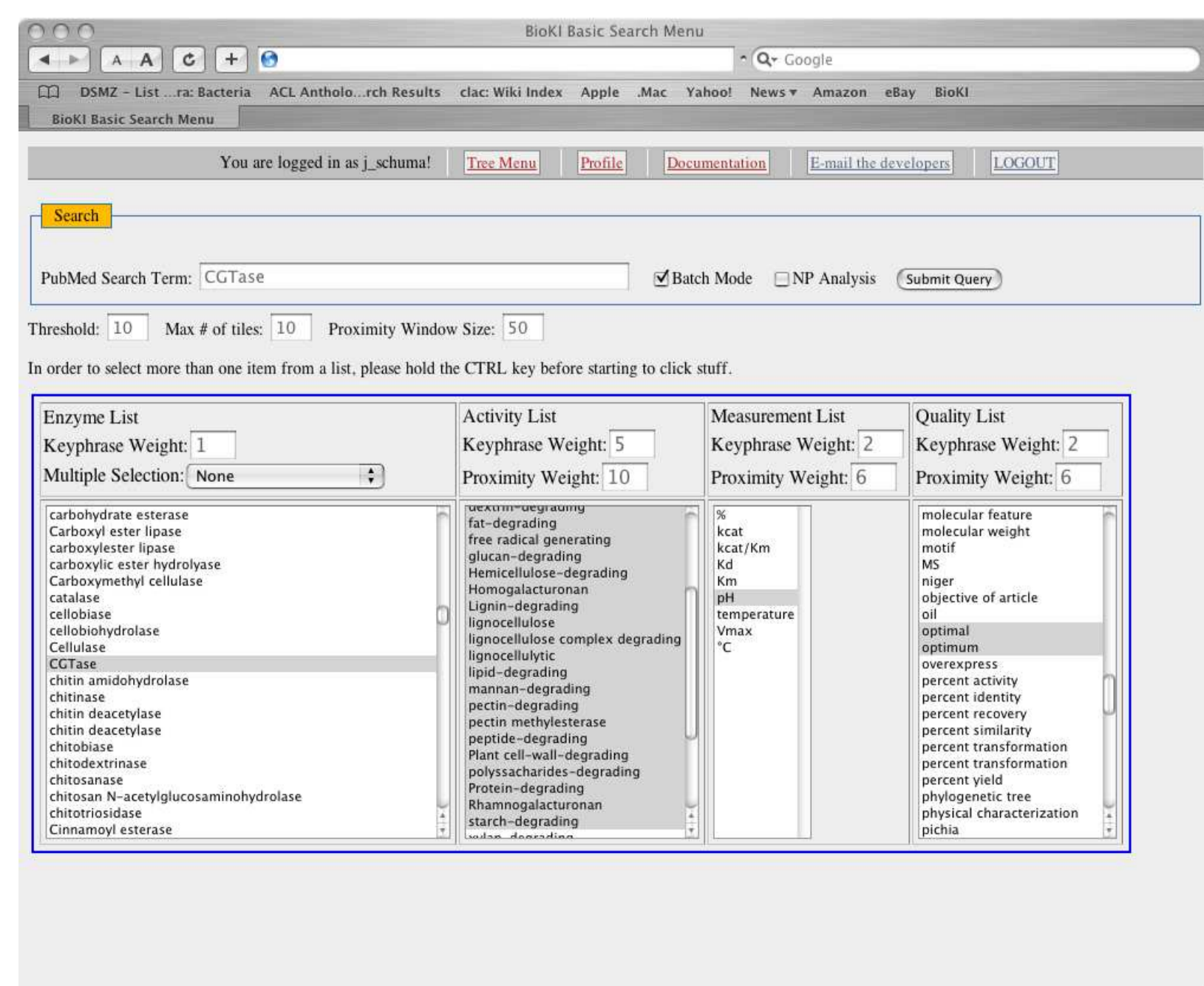
**Why full text?** An example for low-frequency information are temperature and pH optima for certain catalytic reactions, which may be available only in the “Materials and Methods” or “Results” section of an article, and not in the abstract. The entire body of the article has to be processed to locate this information.

**What are the data?** BioKI:Enzymes navigates full-text articles from PubMed Central(PMC). It uses a two-step process:

1. full-text articles are retrieved from PubMed Central (PMC)
2. for each article
  - (a) the most relevant passages are identified according to a set of user selected keywords
  - (b) the articles are ranked according to the pertinence of the representative passages

## 2 Interaction

To assist the user in keyword selection, BioKI:Enzymes offers four lists of keywords. Any number of keywords can be selected (and added) in each list and keyphrases from one list all receive the same weight.



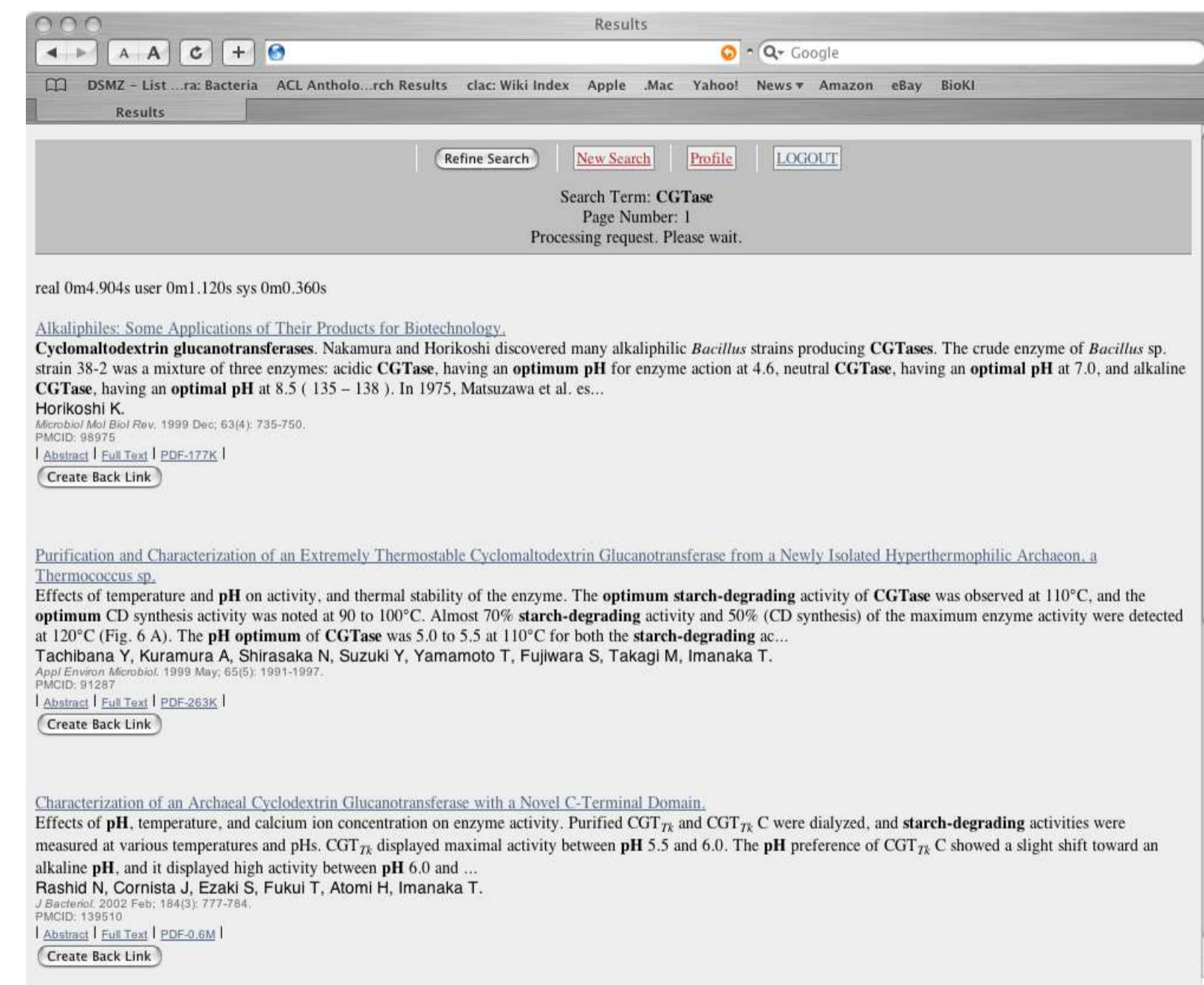
The lists represent

- enzymes
- their activities (such as *carbohydrate degrading*)
- their qualities (such as *maximum activity*)
- measurements (such as *pH*)

The word lists are convenient for selecting alternate spellings that might be hard to enter ( $\alpha$ -*amylase*) and for setting up keyphrase templates in a *profile*, which can be stored under a name and reused. Stemming and the equivalent treatment of Greek characters and their different transliterations generates additional, synonymous keywords.

**Treeview of ontology terms** Modeling the EC number taxonomy (<http://www.chem.qmul.ac.uk/iubmb/enzyme/>) allows the user to select individual enzymes or propagate node selection automatically down the hierarchy from selected concepts.

**What is adaptable?** The user has direct control over the ranking modality (“and” or “or”) and the weight of keywords in the ranking. Each of the four keyword categories has a weight associated with it. A threshold bonus can be assigned for keywords that co-occur closer than a user-determined threshold.

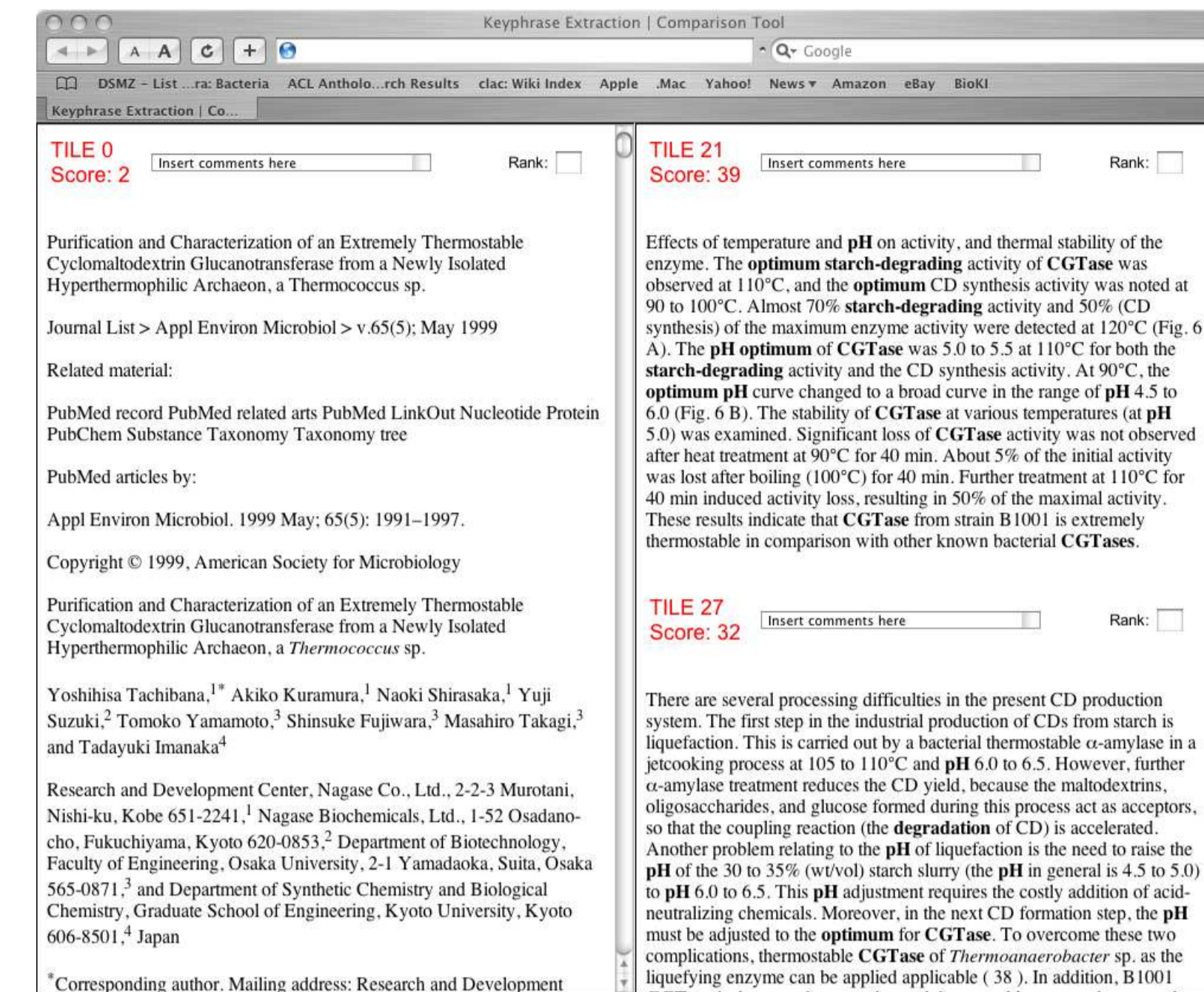


## 2.1 General Search

**“Or” ranking** Adds weights of every keyphrase occurrence and the co-occurrence bonus (only for terms from different lists occurring closer than the threshold).

The basic assumption underlying our tool is that the user can specify a general search term for PMC, but wishes the results to be ordered according to the keyphrases selected. Thus under the PMC search term *CGTase*, for instance, he or she might select all activities on our activities list in combination with *pH* and *optimum*, *optimal* (see left).

**Control parameters** While the search space is too big for a scientist to control all these degrees of freedom without support, our initial experiments have shown that we could control the ranking behavior with repeated refinements of the weight settings. Current research aims to identify a number of useful template weight settings.



## 2.2 Targeted Search

The general mode allows partial matching and counting of repeated keyphrases, thus tiles where all targeted keyphrases occur can loose ground to those where few occur repeatedly. A targeted query mode that requires all categories of keyphrases to be present and only counts unique occurrences is required for some searches.

**“And” ranking** The weight settings are ignored and the text segment that has all required keywords (one from each list as specified) closest together will be ranked highest. This is the mode of choice for a targeted search for specific information, like “pH optima” in a PMC subcorpus for *amylase*.

For the selection of keyphrases *a-amylase*,  $\alpha$ -*amylase*, and *alpha-amylase*, *pH*, *optimum*, *optimal* the selected keywords of the top tile of the highest ranked article were the bold items in the two sentences:

*Igarashi et al. isolated a novel liquefying  $\alpha$ -amylase (LAMY) from cultures of an alkaliphilic Bacillus isolate, strain KSM-1378 ( 73 ). The enzyme had a pH optimum of 8.0 to 8.5 and displayed maximum activity at 55° C.*

This screenshot shows how the context helps interpret the keyphrase occurrences: the anaphor *the enzyme* has been successfully framed with the relevant information, the user can verify the correctness of this hit and any additional information at a glance.

## 3 Validation through Different Contexts

Scientists face two major obstacles in using IR and IE technology: how to select the best keywords for an intended search and how to assess the validity and relevance of the extracted information.

**Validity of extracted information** BioKI provides convenient access to different degrees of context for the user to assess validity or relevance:

- a ranked list of articles provides the first five lines of the most pertinent text segment selected by BioKI
- a side-by-side view of
  - the full-text article as retrieved through PMC on the left

– the different text segments (we use TextTiler (Hearst 1997) to segment the article), ordered by relevance to the keywords, on the right

The user can thus assess the information in the context of the text segment first, and in the original context, if desired. The different details provided allow the user evaluation at a glance.

## 4 Assessment and Future Work

**Reranking through keyphrase refinement** The strength of BioKI lies in its adaptability to user queries. In this it contrasts with template-based IE systems like BioRAT (Corney et al 2004), which extract information from full-length articles, but use handcoded templates to do so. It also contrasts with other literature navigation systems such as GoPubMed (Doms and Schroeder 2005). GoPubMed retrieves abstracts using PubMed search, and then organizes and ranks the articles according to terms that occur in the Gene Ontology (<http://www.geneontology.org>). GoPubMed results depend only on the selected keywords and the GO hierarchy of terms. Like GoPubMed, BioKI is not specific to an information need, but in contrast is meant to give more control to the user. To this end, the same PMC search results can be reordered by successively refining the selected BioKI keywords and weights until more desirable texts appear at the top. This behavior is modeled after frequent behavior using search engines such as Google, where often the first search serves to better select keywords for a subsequent, better targeted search.

**Performance** Reranking based on keyword refinement can be done almost instantaneously (20 sec for 480 keyphrases on 161 articles), since the downloaded texts from PMC are cached, and since the system spends most of its runtime downloading and storing the articles from PMC. This is currently a feasibility study, targeted to eventually become a Web service. Performance still needs to be improved (3:14 min for 1 keyphrase on 161 articles, including downloading), but the quality of the ranking and variable context views might still entice users to wait for them.

In conclusion, it is feasible to develop a highly user-adaptable passage highlighting system over full-text articles that focuses on low-frequency information. Adaptability is provided both through increased user control of the ranking parameters and through presentation of results in different contexts which at the same time justify the ranking and authenticate keyword occurrences in their source text.

This tool is among the first to attempt literature navigation capabilities on full-length articles from PMC. Explicit control over the ranking parameters make it highly user adaptable.

### Acknowledgments

The first prototype of BioKI was implemented by Evan Desai. We thank our domain experts Justin Powlowski, Emma Masters, and Regis-Olivier Benech. Work funded by Genome Quebec.

D. P. A. Corney, B.F. Buxton, W.B. Langdon, and D.T. Jones. 2004. BioRAT: Extracting biological information from full-length papers. *Bioinformatics*, 20(17):3206–3213.

Andreas Doms and Michael Schroeder. 2005. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Research*, 33:W783–W786. Web Server issue.

M.A. Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):34–64.