BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 13 2006, pages 1668–1669
doi:10.1093/bioinformatics/btl159

*Data and text mining*

# An online literature mining tool for protein phosphorylation

X. Yuan[1,†], Z. Z. Hu[1,*,†], H. T. Wu[1], M. Torii[2], M. Narayanaswamy[3], K. E. Ravikumar[3], K. Vijay-Shanker[2] and C. H. Wu[1]

[1]Protein Information Resource, Department of Biochemistry and Molecular and Cellular Biology, Georgetown University Medical Center, Washington, DC 20007, USA, [2]Department of Computer and Information Sciences, University of Delaware, Newark, DE 19716, USA and [3]AU-KBC Research Centre, Anna University, Chennai, India

Received on February 3, 2006; revised on March 21, 2006; accepted on April 21, 2006
Advance Access publication April 27, 2006
Associate Editor: Thomas Lengauer

**ABSTRACT**

A web-based version of the RLIMS-P literature mining system was developed for online mining of protein phosphorylation information from MEDLINE abstracts. The online tool presents extracted phosphorylation objects (phosphorylated proteins, phosphorylation sites and protein kinases) in summary tables and full reports with evidence-tagged abstracts. The tool further allows mapping of phosphorylated proteins to protein entries in the UniProt Knowledgebase based on PubMed ID and/or protein name. The literature mining, coupled with database association, allows retrieval of rich biological information for the phosphorylated proteins and facilitates database annotation of phosphorylation features.

**Availability:** The online RLIMS-P is freely accessible at http://pir.georgetown.edu/iprolink/rlimsp

**Contact:** zh9@georgetown.edu

**Supplementary Information:** http://pir.georgetown.edu/iprolink/rlimsp/supplement/

## INTRODUCTION

With increasing volume of scientific literature now available electronically, efficient text mining tools will greatly facilitate the extraction of information buried in free text and will assist in database annotation. Many methods, including natural language processing, machine learning and rule-based approaches, have been employed for biological literature mining, especially in the areas of biological entity recognition (Mika and Rost, 2004; Zhou *et al*., 2004; Yeh *et al*., 2005) and citation mapping and information extraction (Shatkay and Feldman, 2003). Protein phosphorylation is a fundamental molecular event essential to cellular processes (Hunter, 2000). A rule-based text mining system, RLIMS-P (Rule-based LIterature Mining System for Protein Phosphorylation), for extracting protein phosphorylation information from MEDLINE abstracts was previously developed and benchmarked with excellent performance (Hu *et al*., 2005; Narayanaswamy *et al*., 2005). The system extracts three objects involved in protein phosphorylation—the protein kinase, the protein substrate (phosphorylated protein) and the residue/position being phosphorylated (phosphorylation site). Here we have developed a web-based version of the RLIMS-P literature mining tool for easy accessibility and enhanced functionality. The online RLIMS-P provides a user-friendly interface for phosphorylation information mining, evidence tagging and protein mapping to UniProt Knowledgebase (UniProtKB) (Wu *et al*., 2006); thereby facilitating biological studies of phosphorylated proteins and database annotation of phosphorylation features.

## DEVELOPMENT OF ONLINE RLIMS-P

The RLIMS-P system utilizes shallow parsing and extracts phosphorylation objects by matching text with manually developed rule-patterns (Narayanaswamy *et al*., 2005). Built upon the RLIMS-P system, the online tool allows users to determine whether a MEDLINE abstract contains protein phosphorylation information and to extract protein kinase, phosphorylated protein and phosphorylation site from the abstract and its title. Furthermore, the online RLIMS-P post-processes the text output from the server program with three additional functionalities—phosphorylation annotation ranking, evidence tagging and protein entity mapping (Supplementary Figure S1).

### Phosphorylation annotation ranking

For any given phosphorylation-related abstract, RLIMS-P produces one or multiple annotation results, each consisting of up to three phosphorylation objects (kinase, phosphorylated protein and site). The multiple annotation results are ranked based on the number of objects and sites extracted (Supplementary Figure S1B). Annotations with phosphorylated protein information take precedence over protein kinase information. Therefore, annotation with three objects and with the most sites will rank first, while annotation with site and phosphorylated protein will precede one with site and kinase.

### Evidence tagging

To provide evidence attribution for the annotation results, the online RLIMS-P tags each individual object with an internal identifier during the shallow parsing and uses it subsequently for color-tagging phosphorylation objects in the abstract for web display.

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

 

### Protein entity mapping

The online RLIMS-P provides semi-automatic mapping of phosphorylated proteins to UniProtKB entries based on PubMed ID (PMID) and/or proteins names.

(1) *PMID mapping*. UniProtKB contains extensive bibliographic information for each protein entry via the UniProtKB-PMID mapping that includes reference citations annotated in UniProtKB and collected from curated databases such as SGD, MGD and GeneRIF (Wu *et al*., 2006). For an abstract that already exists in the PMID mapping, the online RLIMS-P automatically associates the phosphorylated protein with the UniProtKB protein entry.

(2) *Name mapping*. If the abstract has no PMID mapping in UniProtKB, the phosphorylated protein can be mapped based on protein names using the web-based BioThesaurus (Liu *et al*., 2006). BioThesaurus provides a comprehensive collection of gene/protein names from multiple molecular databases with associations to UniProtKB protein entries. The online BioThesaurus allows the retrieval of synonyms of a given protein and the identification of protein entries sharing a given name. The online RLIMS-P integrates these BioThesaurus functions, allowing users to select the corresponding protein entry for the phosphorylated protein from a list of UniProtKB entries retrieved by BioThesaurus (Supplementary Figure S2).

## RLIMS-P WEBSITE

The online RLIMS-P web site accepts user submission of PMIDs as input and returns a summary table for all PMIDs (Supplementary Figure S1A) with links to full reports (Figure S1B). The summary table lists the PMID of each phosphorylation-related abstract along with its top-ranking annotation result, followed by a list of remaining PMIDs for abstracts containing no phosphorylation information. Full reports can be retrieved from the summary table using hypertext links (from 'text evidence') or by selecting one or more PMID(s) in the list.

The full RLIMS-P report contains five sections (Supplementary Figure S1B): (1) PubMed citation information (publication date, authors, journal); (2) PMID mapping to UniProtKB, consisting of the accession, ID, protein name, organism and protein family of the mapped entry, with links to UniProtKB and iProClass (Wu *et al*., 2004) protein reports containing rich biological and functional information; (3) name mapping to UniProtKB, including options to use either names appeared in the abstract or user-specified names for searching online BioThesaurus; (4) annotation with a ranked list of RLIMS-P extraction results for each set of the three phosphorylation objects and (5) text evidence showing the original abstract and title, with extracted objects tagged in different colors to distinguish protein kinases, phosphorylated proteins and phosphorylated residues/positions. An option is provided for turning on/off the color-tagging for each type of object in the abstract for visual inspection.

## CONCLUSION

The online RLIMS-P text mining tool allows users to mine protein phosphorylation information from MEDLINE abstracts with ease. The user-friendly interface provides functionalities for viewing the extracted phosphorylation information and mapping the phosphorylated proteins to UniProtKB entries. The text mining and rich biological information will facilitate scientific studies of phosphorylated proteins by research biologists. The online tool can also assist database curators to annotate phosphorylation information and is being used for literature-based curation of protein phosphorylation features in UniProtKB. Several enhancements of the system are planned, including (1) allowing other types of input query, such as free-text abstracts, in addition to PMID and (2) extending text mining to full-length articles.

## ACKNOWLEDGEMENT

This project is supported in part by grant U01-HG02712 from the National Institutes of Health, USA.

*Conflict of Interest*: none declared.

## REFERENCES

Hu,Z.Z. *et al*. (2005) Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics*, **21**, 2759–2765.

Hunter,T. (2000) Signaling—2000 and beyond. *Cell*, **100**, 113–127.

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

Mika,S. and Rost,B. (2004) Protein names precisely peeled off free text. *Bioinformatics*, **20** (Suppl. 1), I241–I247.

Narayanaswamy,M. *et al*. (2005) Beyond the clause: extraction of phosphorylation information from medline abstracts. *Bioinformatics*, **21** (Suppl. 1), i319–i327.

Shatkay,H. and Feldman,R. (2003) Mining the biomedical literature in the genomic era: an overview. *J. Comput. Biol.*, **10**, 821–855.

Wu,C.H. *et al*. (2004) The iProClass integrated database for protein functional analysis. *Comput. Biol. Chem.*, **28**, 87–96.

Wu,C.H. *et al*. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34** (Database issue), D187–D191.

Yeh,A. *et al*. (2005) BioCreAtIvE task 1A: gene mention finding evaluation. *BMC Bioinformatics*, **6** (Suppl. 1), S2.

Zhou,G. *et al*. (2004) Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, **20**, 1178–1190.