RLIMS-P: Literature-based curation of protein phosphorylation information

Manabu Torii[1,2]*, Gang Li[1,2], Zhiwen Li[1,2], Irem Çelen[1], Francesca Diella[4], Rose Oughtred[5], Cecilia Arighi[1,2], Hongzhan Huang[1,2], K. Vijay-Shanker[2], Cathy H. Wu[1,2,3]

[1]Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE 19711, USA, [2]Department of Computer and Information Sciences, University of Delaware, Newark, DE 19711, USA, [3]Protein Information Resource, Department of Biochemistry, Molecular and Cellular Biology, Georgetown University Medical Center, Washington, DC 20007, USA, [4]EMBL, Germany, [5]Department of Genomics, Princeton University, Princeton, NJ, USA.

*Corresponding author: Tel: 302 831 6162, E-mail: torii@udel.edu

Abstract

Annotation of protein phosphorylation information has been the focus of many biological knowledge bases. To support the literature-based curation of phosphorylation information, an information extraction (IE) system, named RLIMS-P, has been developed, which extracts protein phosphorylation information from biomedical literature. The system has been recently redesigned as RLIMS-P v2 and a new online curator website has been developed. The new website offers improvements for curation functionalities, including PubMed-style keyword search of extracted information, multiple views of retrieved information and their downloading, editing of automatically gathered information, and entity normalization. Curators from Phospho.ELM, Protein Ontology (PRO), and BioGrid were recruited to test the website in the BioCreative Track 5 - User Interactive Task (IAT). We expect the new website can be a useful tool for biocurators to search relevant literature and annotate phosphorylation information. Final results from the current evaluation test will be presented at the workshop.

Introduction

The reversible phosphorylation of proteins is central to the regulation of most aspects of cell function. The flow of molecular information through signaling pathways frequently depends on protein phosphorylation mediated by specific kinases that recognize and phosphorylate specific sites in the target proteins (1). In many cases, deregulation of the kinase-substrate network has been linked to disease, including cancer (2). Given its relevancy, protein phosphorylation has been an active research area as well as the focus of curation in multiple knowledgebase, such as

Protein Ontology (PRO)[1], PhosphoSitePlus[2], Phospho.ELM[3], and UniProt Knowledgebase (UniProtKB)[4]. To support review of relevant literature by biocurators, a rule-based information extraction (IE) system, named RLIMS-P, has been developed in our group (3,4). The system is designed to identify protein phosphorylation information reported in biomedical literature and it extracts entities involved in the phosphorylation event (kinase, substrate, and site). Recently the system has been revised as RLIMS-P v2 and applied to the entire MEDLINE. To make the large amount of information extracted from MEDLINE, a web interface for biocurators has also been redesigned. The new web interface allows users to search, retrieve, edit, and manage protein phosphorylation information online. In addition, we have integrated gene normalization results obtained with GenNorm (5) and the bibliography mapping information available in UniProtKB (6) in this web interface.

Based on the new interface design, we set up a curator website for BioCreative Track 5 - User Interactive Task (IAT). In this report, we describe this curator website, and introduce the data and the curation tasks considered for the IAT task.

Material and Methods

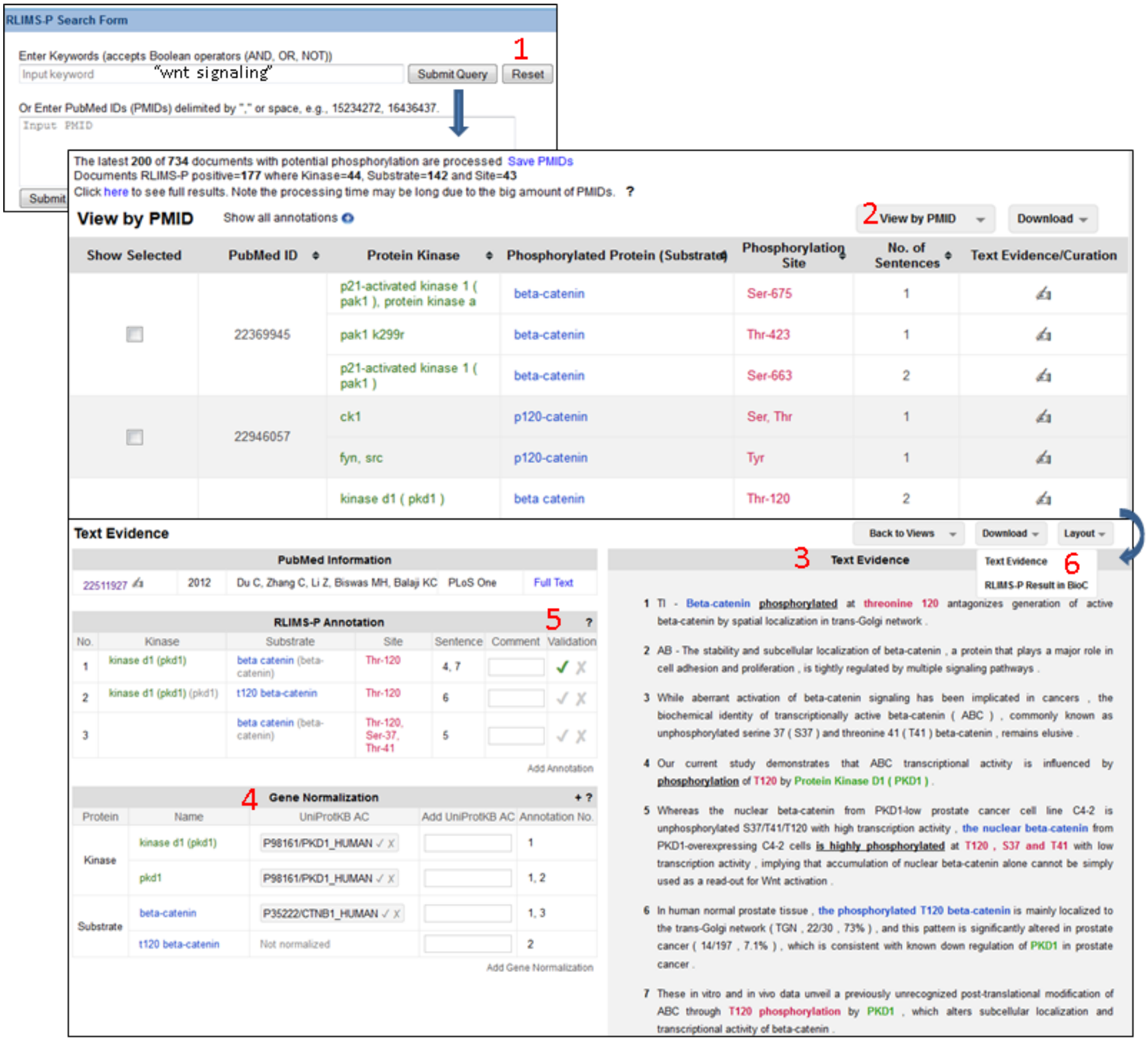
RLIMS-P system

RLIMS-P is a rule-based IE system designed to extract a kinase, a substrate, and a site that are involved in a phosphorylation event. The system consists of several text processing modules, including (i) a shallow parser that syntactically analyzes input sentences, (ii) a term classifier that identifies semantic categories of phrases, e.g., identification of protein names, (iii) a pattern-based IE engine that extracts entities involved in the target event, and (iv) an additional IE component that identifies an event reported across multiple sentences. This system has been recently redesigned as RLIMS-P v2 (7). One of the enhancements in the new system includes a design of the IE engine that eases management of extraction patterns. The new system can cover a large number of extraction patterns through combination of pattern fragments, instead of requiring a large set of complex patterns. Sentence simplification techniques in the original system, which improve pattern matching, were extended for the new design, based on the recent work in the group (8). RLIMS-P v2 was evaluated in different settings and F-scores for the extraction task were over 90%. For further information about RLIMS-P v2 and its evaluation results, readers may refer to (7).

[1] http://pir.georgetown.edu/pro
[2] http://www.phosphosite.org
[3] http://phospho.elm.eu.org
[4] http://www.uniprot.org

The database

Phosphorylation information extracted from the MEDLINE archive using RLIMS v2 is stored in a database. Normalization of protein names obtained using GenNorm (5) is integrated in this database. In addition, the bibliography mapping service of PIR/UniProt is used to associate extracted information with UniProtKB entries. The resulting database is incrementally updated weekly in synch with MEDLINE citations in PubMed. The database initially built using the 2013 release of the MEDLINE archive contains phosphorylation information extracted from 165,840 abstracts, and links to 43,329 UniProtKB entries.



The latest **200** of **734** documents with potential phosphorylation are processed Save PMIDs
Documents RLIMS-P positive=**177** where Kinase=**44**, Substrate=**142** and Site=**43**
Click here to see full results. Note the processing time may be long due to the big amount of PMIDs. ?

View by PMID Show all annotations

Show Selected	PubMed ID	Protein Kinase	Phosphorylated Protein (Substrate)	Phosphorylation Site	No. of Sentences	Text Evidence/Curation
□	22369945	p21-activated kinase 1 (pak1), protein kinase a	beta-catenin	Ser-675	1	✍
		pak1 k299r	beta-catenin	Thr-423	1	✍
		p21-activated kinase 1 (pak1)	beta-catenin	Ser-663	2	✍
□	22946057	ck1	p120-catenin	Ser, Thr	1	✍
		fyn, src	p120-catenin	Tyr	1	✍
		kinase d1 (pkd1)	beta catenin	Thr-120	2	✍

Text Evidence

PubMed Information

22511927 ✍	2012	Du C, Zhang C, Li Z, Biswas MH, Balaji KC	PLoS One	Full Text

RLIMS-P Annotation

No.	Kinase	Substrate	Site	Sentence	Comment	Validation
1	kinase d1 (pkd1)	beta catenin (beta-catenin)	Thr-120	4, 7		✓ ✗
2	kinase d1 (pkd1) (pkd1)	t120 beta-catenin	Thr-120	6		✓ ✗
3		beta catenin (beta-catenin)	Thr-120, Ser-37, Thr-41	5		✓ ✗

Gene Normalization

Protein	Name	UniProtKB AC	Add UniProtKB AC	Annotation No.
Kinase	kinase d1 (pkd1)	P98161/PKD1_HUMAN ✓ ✗		1
	pkd1	P98161/PKD1_HUMAN ✓ ✗		1, 2
Substrate	beta-catenin	P35222/CTNB1_HUMAN ✓ ✗		1, 3
	t120 beta-catenin	Not normalized		2

Text Evidence

1 TI - Beta-catenin phosphorylated at threonine 120 antagonizes generation of active beta-catenin by spatial localization in trans-Golgi network .

2 AB - The stability and subcellular localization of beta-catenin , a protein that plays a major role in cell adhesion and proliferation , is tightly regulated by multiple signaling pathways .

3 While aberrant activation of beta-catenin signaling has been implicated in cancers , the biochemical identity of transcriptionally active beta-catenin (ABC) , commonly known as unphosphorylated serine 37 (S37) and threonine 41 (T41) beta-catenin , remains elusive .

4 Our current study demonstrates that ABC transcriptional activity is influenced by phosphorylation of T120 by Protein Kinase D1 (PKD1) .

5 Whereas the nuclear beta-catenin from PKD1-low prostate cancer cell line C4-2 is unphosphorylated S37/T41/T120 with high transcription activity , the nuclear beta-catenin from PKD1-overexpressing C4-2 cells is highly phosphorylated at T120 , S37 and T41 with low transcription activity , implying that accumulation of nuclear beta-catenin alone cannot be simply used as a read-out for Wnt activation .

6 In human normal prostate tissue , the phosphorylated T120 beta-catenin is mainly localized to the trans-Golgi network (TGN , 22/30 , 73%) , and this pattern is significantly altered in prostate cancer (14/197 , 7.1%) , which is consistent with known down regulation of PKD1 in prostate cancer .

7 These in vitro and in vivo data unveil a previously unrecognized post-translational modification of ABC through T120 phosphorylation by PKD1 , which alters subcellular localization and transcriptional activity of beta-catenin .

Figure 1-Snapshot of the main pages in RLIMS-P website; namely search, result table, and text evidence. 1-6 refer to the functionalities listed in the main text.

The web interface

For BioCreative IV - IAT task, a new curator website has been set up (http://research.bioinformatics.udel.edu/text_mining/rlimsp2/). The website (Figure 1) supports the following functionalities:

1. Search and retrieval of phosphorylation information gathered by RLIMS-P using the PubMed-style query, as well as the query by PMIDs;
2. Display of a query result (a table of kinase, substrate, and site) with different 'view' options (e.g., group by kinase, substrate, or PMID) as well as sorting options;
3. Display of text evidence (MEDLINE abstracts with highlighted entities);
4. Provision of protein normalization information for kinases and substrates using GenNorm, a state-of-the-art normalization tool (5);
5. A user login for editing, saving and exporting curated annotations;
6. Downloading of phosphorylation information in the CSV format, and that of evidence text in the BioC format (9);
7. Support of different browsers: Google Chrome, Mozilla Firefox, Internet Explorer 9, and Safari.

These functionalities as well as the usage of the website are described in a help document (http://research.bioinformatics.udel.edu/text_mining/rlimsp2/files/RLIMSP_help.pdf). The website has been developed with the help of PRO curators, but it is intended for a broader curation community, not limited to PRO curation.

The BioCuration task

Three curators were recruited to test the RLIMS-P website. They are domain experts with experience on kinase-substrate event annotation, specifically annotation for Phospho.ELM, PRO, and PhosphoGRID/BioGRID databases. Curation guidelines were developed and they describe which entities should be captured (kinase, substrate and site) and how they should be normalized, along with exercises to get familiar with the curation criteria and the interface (http://research.bioinformatics.udel.edu/text_mining/rlimsp2/files/RLIMSP_guidelines.pdf). The curation task requested is summarized below:

1. Given a set of 50 PMIDs, fill in the tuples of kinase, substrate and site with normalization information. Perform this task on a half of this collection using the curator website and on the other half without using it. The curator records the annotation results along with UniProtKB identifiers where possible. The curator will record the time spent.
2. Complete the user survey (http://ir.cis.udel.edu/biocreative/survey2.html).

All the annotation results will be reviewed by a senior PRO curator and the performance of the RLIMS-P system will be measured using standard performance measures, such as precision and recall. We should also examine the time spent for the manual curation and that for the RLIMS-P-assisted curation.

The datasets

Three datasets tailored to the participating curators were prepared as below.

Dataset 1

The first dataset was prepared for the PhosphoGRID/BioGRID curator. This dataset includes articles with phosphorylation information on yeast, published between 2012 and 2013. The selection of 50 PMIDs was based on the PubMed query: `("2012/01/01"[Date - Publication] : "3000"[Date - Publication]) AND (saccharomyces OR yeast) AND phosphory*`. The retrieved results were inspected to confirm that the contents were appropriate for the curation task.

Dataset 2

The second dataset was prepared for the Phospho.ELM curator. This dataset was compiled for any kinase-substrates relation reported in articles published in 2013. The selection of 50 PMIDs were based on the PubMed query: `("2013/01/01"[Date - Publication] : "3000"[Date - Publication]) AND kinase AND phosphory*`. Again, the retrieved results were inspected so that the contents were appropriate.

Dataset 3

The third dataset was prepared for the PRO curator. This set contained a subset of abstracts from Dataset 1 (11) and a subset from Dataset 2 (36), and the remaining ones were collected from literature related to transient potential receptors (TRP).

Results and Discussion

The original RLIMS-P system had been used for PRO curation of protein forms (10,11), Phopho.ELM curation (12), and pathway curation (13). It had also been used to provide information for another text-mining system, eFIP, which extracts functional impact of phosphorylation events (14,15). We expect that the enhancements in RLIMS-P v2 and the new curator website described in this report can further help curators to annotate phosphorylation information or a text mining tool based on RLISM-P to extract biomedical knowledge. Final results from the current evaluation test will be closely examined and our analyses as well as the obtained results will be presented at the workshop.

Funding

This work was supported by the National Science Foundation [ABI-1062520 to M.T., C.A., H.H., K.V., and C.H.] and the National Library of Medicine of the National Institutes of Health [G08LM010720 to C.A., H.H., K.V., and C.H.].

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Acknowledgement

We acknowledge the work of Drs. Narayanaswamy and Ravikumar, who played an integral role in developing the original RLIMS-P system. We would like to thank the IAT Track organizers Drs. Phoebe Roberts, Sherry Mathis, and Catalina Oana Tudor for their assistance during RLIMS-P evaluation.

Conflict of Interest: none declared.

References

1. Pawson, T. and M. Kofler. 2009. Kinome signaling through regulated protein-protein interactions in normal and cancer cells. Curr Opin Cell Biol *21*:147-153.
2. Zhang, L. and R.J. Daly. 2012. Targeting the human kinome for cancer therapy: current perspectives. Crit Rev Oncog *17*:233-246.
3. Narayanaswamy, M., K.E. Ravikumar, and K. Vijay-Shanker. 2005. Beyond the clause: extraction of phosphorylation information from medline abstracts. Bioinformatics (Oxford, England) *21 Suppl 1*:i319.
4. Hu, Z.Z., M. Narayanaswamy, K.E. Ravikumar, K. Vijay-Shanker, and C.H. Wu. 2005. Literature mining and database annotation of protein phosphorylation using a rule-based system. Bioinformatics (Oxford, England) *21*:2759.
5. Wei, C.-H. and H.-Y. Kao. 2011. Cross-species gene normalization by species inference. BMC bioinformatics *12 Suppl 8*.
6. The UniProt Consortium. 2012. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic acids research *40*:D71.
7. Torii, M., C.N. Arighi, Q. Wang, C.H. Wu, and K. Vijay-Shanker. 2013. Text Mining of Protein Phosphorylation Information Using a Generalizable Rule-Based Approach, ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics (ACM-BCB) 2013.
8. Peng, Y., Tudor, C.O., Torii, M., Wu, C. H. and Vijay-Shanker, K. 2012. iSimp: A sentence simplification system for biomedicail text, IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2012), Philadelphia, USA.
9. Comeau, D.C., R. Islamaj Dogan, P. Ciccarese, K.B. Cohen, M. Krallinger, F. Leitner, Z. Lu, Y. Peng, et al. 2013. BioC: a minimalist approach to interoperability for biomedical text processing. Database (Oxford) *2013*:bat064.
10. Ross, K.E., C.N. Arighi, J. Ren, H. Huang, and C.H. Wu. 2013. Construction of Protein Phosphorylation Networks by Data Mining, Text Mining, and Ontology Integration: Analysis of the Spindle Checkpoint. *in press*.
11. Ross, K.E., C.N. Arighi, J. Ren, D.A. Natale, H. Huang, and C.H. Wu. 2013. Use of the protein ontology for multi-faceted analysis of biological processes: a case study of the spindle checkpoint. Frontiers in genetics *4*.
12. Dinkel, H., C. Chica, A. Via, C.M. Gould, L.J. Jensen, T.J. Gibson, and F. Diella. 2011. Phospho.ELM: a database of phosphorylation sites--update 2011. Nucleic acids research *39*:D261.

13. Schmidt, C.J., L. Sun, C.N. Arighi, K. Decker, K. Vijay-Shanker, M. Torii, C.O. Tudor, C. Wu, and P. D'Eustachio. 2012. Pathway curation: Application of text-mining tools eGIFT and RLIMS-P, p. 523. IEEE.
14. Arighi, C.N., A.Y. Siu, C.O. Tudor, J.A. Nchoutmboube, C.H. Wu, and V.K. Shanker. 2011. eFIP: a tool for mining functional impact of phosphorylation from literature. Methods in molecular biology (Clifton, N.J.) *694*:63.
15. Tudor, C.O., C.N. Arighi, Q. Wang, C.H. Wu, and K. Vijay-Shanker. 2012. The eFIP system for text mining of protein interaction networks of phosphorylated proteins. Database (Oxford) *2012*:bas044.