BioCreative IV IAT Task: Literature-based Curation of Protein Phosphorylation Information

University of Delaware, Newark, DE USA
Team #240

Background

Annotation of protein phosphorylation information has been the focus of many biological knowledge bases, such as Protein Ontology[1], PhosphoSitePlus[2], Phospho.ELM[3], and UniProt Knowledgebase (UniProtKB)[4]. To support such annotation effort by helping biocurators reviewing literature, a rule-based information extraction (IE) system, named RLIMS-P, has been developed [1, 2], which identifies protein phosphorylation information (kinase, substrate, and site) reported in biomedical literature. With a modern powerful server, application of RLIMS-P to the entire MEDLINE has become feasible and large amount of automatically extracted information has resulted. To make the extracted information readily usable for biocurators, we designed and developed a web environment for biocurators to search, retrieve, edit, and manage protein phosphorylation information extracted from literature.

RLIMS-P: Rule-based information extraction system for protein phosphorylation information

RLIMS-P is a rule-based IE system that extracts entities involved in a protein phosphorylation event, namely, kinase, substrate, and site. Each of these three entities bears significant biological information. For instance, a kinase may only recognize specific sites in the substrate protein, and different biological responses may be elicited depending on the sites. Accordingly, RLIMS-P was developed to extract all three types of entities.

RLIMS-P consists of several customized text mining components, including (i) a shallow parser that syntactically analyzes input sentences, (ii) a term classifier that identifies semantic categories of phrases in text, e.g., identification of protein names, (iii) a pattern-based IE engine that extracts target entities, and (iv) an additional IE component that extracts entities across multiple sentences. Recently, the system has been redesigned for improved modularity and generalizability, which can ease further enhancement of system components and adoption of new technology and resources in the system [3].

We also developed a web site for biologists to search and retrieve information extracted from Medline. Our goal in participating in the User Interactive Task is to enrich the current web interface and make it as a web-based environment for curating protein phosphorylation information, as discussed in the following section. The current RLIMS-P website for searching and browsing phosphorylation information gathered from Medline is available at: http://research.bioinformatics.udel.edu/rlimsp/.

A web-based environment for protein phosphorylation information curation

The curation task we target is annotation of substrate proteins along with kinase and site information. As stated before, there are several databases and knowledge bases that collect and record protein phosphorylation information and, hence, there are several teams working on literature-based curation of phosphorylation information. Among them, we work closely with curators of the Protein Ontology project, and follow the requirements set forth by them as well as the technical requirements of the BioCreative IAT task. Specifically, we aim to provide the following functionalities in the proposed system:

- A mechanism for curators to query and retrieve phosphorylation information gathered by RLIMS-P using the same query style as PubMed;
- Display of information (a table of kinase, substrate, and site) with different 'view' options (group by kinase, substrate, or PMID).

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

- Provision of protein normalization information for kinases and substrates using GenNorm, an existing high-performance normalization tool [4];
- Display of text evidence for extracted information (Medline abstracts with highlighted entities);
- A mechanism for logged-in users to edit, save and manage extracted entities and protein normalization information;
- Downloading of retrieved and edited phosphorylation information in a simple CSV format, and availability of evidence text in the BioC format;
- Support of different browsers: Google Chrome, Mozilla Firefox, Internet Explorer 9, and Safari.

These requirements/specifications have been proposed and examined with the help of the Protein Ontology curators. The systems are highly generic and practical and they could be used in a broader curation community concerning protein phosphorylation information, not limited to curators of the Protein Ontology project.

Status of the system development

We have developed all the modules described above and they can be made accessible to the reviewer. The website currently supports the PubMed-style query to retrieve phosphorylation information extracted from MEDLINE as well as the PMID query, the display of the extraction summary with sorting options, the link to evidence sentences, and download of retrieved information (See Figure 1 and 2). In addition, the website provides suggestion of UniProtKB entries for extracted proteins via the bibliography mapping service at PIR. An example of existing editing capabilities can be found at: http://annotation.dbi.udel.edu/text_mining/rlimsp/ (At the "Login" found at the upper right corner of this page, use the login name: `RLIMSP@biocreative.org` and try, for example, the query PMIDs: `2108025, 16436437`). The result table of the web site allows for annotation of correct and wrong

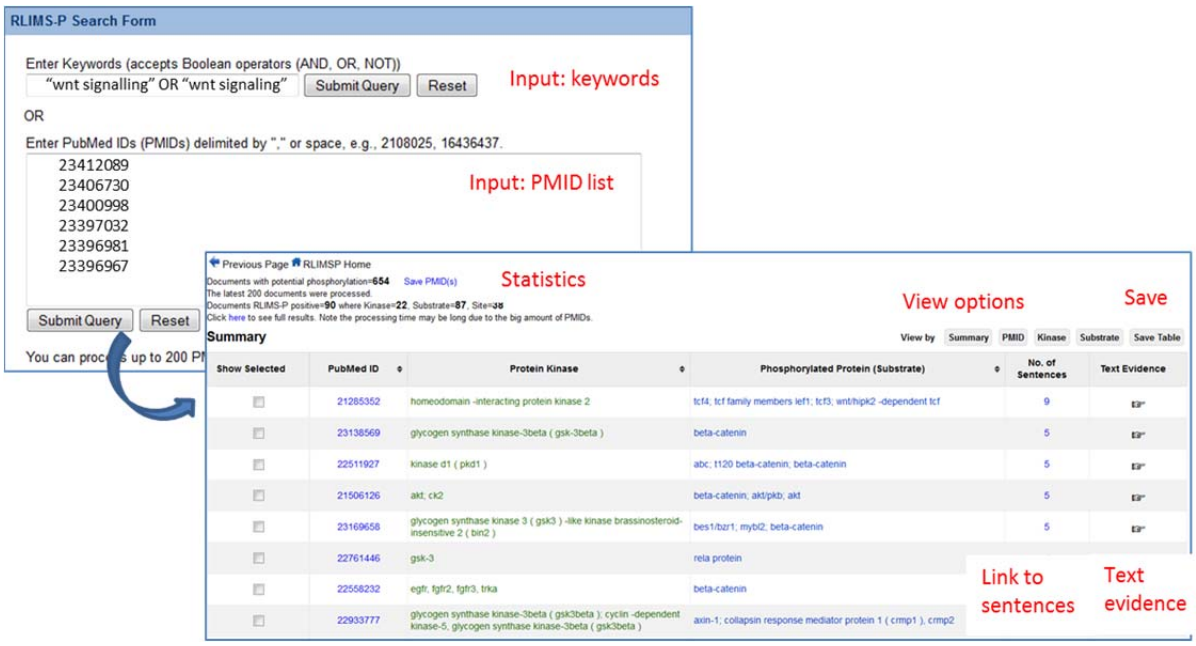

Figure 1-RLIMS-P website. The search form allows querying using keywords or a list of PMIDs. The result page default view displays the statistics on phosphorylation information based on query input, the summary view of kinases/substrates mentioned per abstract along with the number of sentences that are evidence for such information extraction. The result table offers different views to display data in the most convenient way for the user.

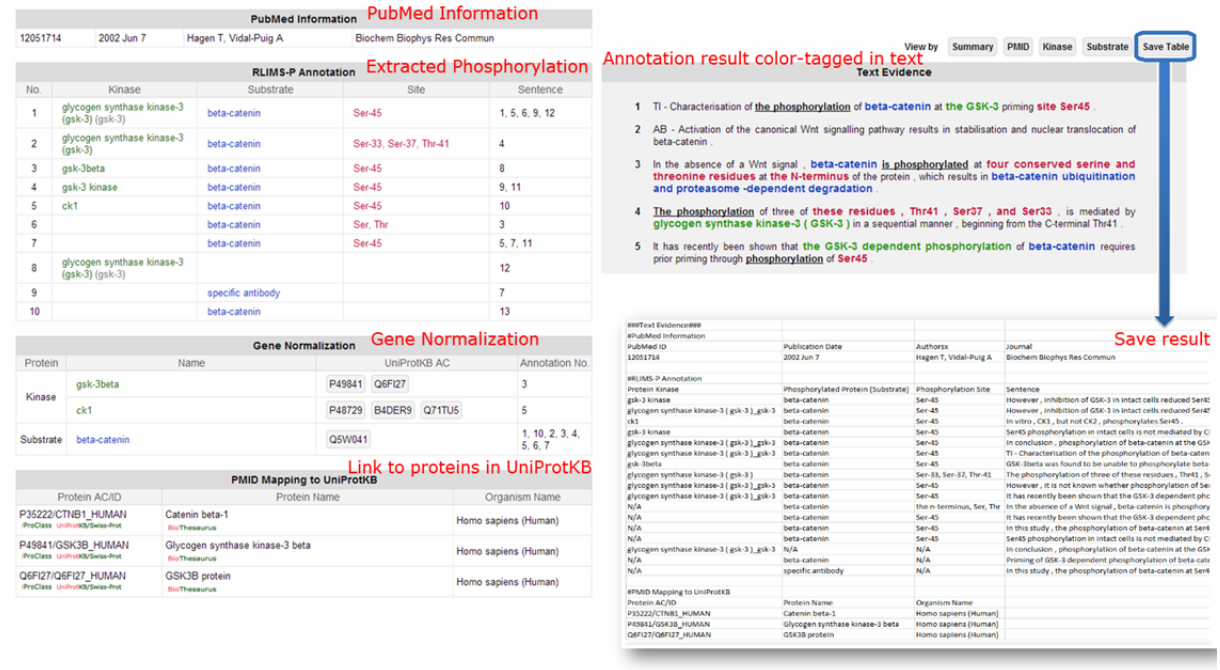

Figure 2-RLIMS-P Text evidence page. The text evidence page displays the general information about the PMID, the results of RLIMS-P extraction from such abstract, the abstract with color-tagged annotation types, the gene normalization information, and the link to UniProtKB entries that are mentioned in the abstract. All this information can be saved in CVS format.

extracted information. Note that for the task at BioCreative we will only request the validation at the abstract level as that is what biocurators normally do. The option of editing extraction results will be integrated in the interface. For gene/protein normalization, we use the GeneNorm tool. Examples of the normalization result can be found at: http://annotation.dbi.udel.edu/text_mining/rlimsp/gennorm-demo.html.

All the parts are developed and will be integrated into the new interface by the end of July as the training will start on August. Updates will be submitted to the reviewer accordingly.

RLIMS-P Benchmarking

RLIMS-P is currently being evaluated on different document sets. On a diverse document set consisting of 60 Medline abstracts sampled among citations in the Phospho.ELM database, the system achieved F-scores of 0.91, 0.93, and 0.96 respectively for kinase, substrate, and sites in the abstract-level evaluation (i.e., redundant tuples extracted from the same abstract are aggregated as one instance, unlike the case of the trigger-level evaluation, where the extraction status is evaluated without aggregating them; in this particular evaluation setting, we included non-normalizable entity mentions, besides normalizable ones, so as to evaluate the system performance apart from the normalization status of entity mentions). The collection of the 60 PMIDs will be posted on the project website.

User Community

RLIMS-P has been used for PRO curation of protein phosphorylated forms [5, 6], pathway curation [7], and Phopho.ELM curation [8], and also for providing information for another text mining tool named eFIP [9]. The website of RLIMS-P is linked from the Phospho.ELM web page[5] and it is also listed in

[5] http://phospho.elm.eu.org/links.html

iProLink in the Protein Information Resource (PIR) website [6]. We would like other biocuration communities to learn and potentially adopt RLIMS-P to retrieve relevant articles for phosphorylation, and curate phosphorylation information. We have engaged two users to help in this task.

The BioCuration task

We plan to request biocurators to perform two tasks:

1. Given a set of PMIDs (to be selected based on the interest of biocurators involved), obtain the tuples of kinase, substrate and site with normalization information. Perform this task on a part of this collection using RLIMS-P and on the other part without using the system. The curator will record the time spent, besides providing curation results.
2. Search information using the PubMed-style query on the website, and validate the retrieved results for selected substrates. The curator will report the curation procedure and the steps involved for searching and annotating the information, which will be examined for the analysis of the system usability. The curation results will also be recorded as in the previous task.

Metrics

For both tasks, 1 and 2, the performance of automated extraction will be measured using the standard metrics, such as precision. For task 1 we will calculate the performance of the manual vs. system-assisted curation. We will also compared annotated result vs. a reference set annotated by the PRO curator.
For task 2 we will compare the functionalities used by the user for the different task steps to the available web functionalities.

References

1. Hu ZZ, Narayanaswamy M, Ravikumar KE, Vijay-Shanker K, Wu CH: **Literature mining and database annotation of protein phosphorylation using a rule-based system**. *Bioinforma. Oxf. Engl.* 2005, **21**:2759–2765.
2. Narayanaswamy M, Ravikumar KE, Vijay-Shanker K: **Beyond the clause: extraction of phosphorylation information from medline abstracts**. *Bioinforma. Oxf. Engl.* 2005, **21 Suppl 1**:i319–327.
3. Torii M, Arighi CN, Wang Q, Wu CH, Vijay-Shanker K: **Text Mining of Protein Phosphorylation Information Using a Generalizable Rule-Based Approach**. In *Proc. Acm Conf. Bioinforma. Comput. Biol. Biomed. Informatics (ACM-BCB)*. (To appear).
4. Wei C-H, Kao H-Y: **Cross-species gene normalization by species inference**. *BMC Bioinformatics* 2011, **12 Suppl 8**:S5.
5. Ross KE, Arighi CN, Ren J, Natale DA, Huang H, Wu CH: **Use of the protein ontology for multi-faceted analysis of biological processes: a case study of the spindle checkpoint**. *Front. Genet.* 2013, **4**:62.
6. Ross KE, Arighi CN, Ren J, Huang H, Wu CH: **Construction of Protein Phosphorylation Networks by Data Mining, Text Mining, and Ontology Integration: Analysis of the Spindle Checkpoint**. 2013, **in press**.
7. Schmidt CJ, Sun L, Arighi CN, Decker K, Vijay-Shanker K, Torii M, Tudor CO, Wu C, D'Eustachio P: **Pathway curation: Application of text-mining tools eGIFT and RLIMS-P**. In IEEE; 2012:523–528.
8. Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, Gibson TJ, Diella F: **Phospho.ELM: a database of phosphorylation sites--update 2011**. *Nucleic Acids Res.* 2011, **39**:D261–267.
9. Tudor CO, Arighi CN, Wang Q, Wu CH, Vijay-Shanker K: **The eFIP system for text mining of protein interaction networks of phosphorylated proteins**. *Database J. Biol. Databases Curation* 2012, **2012**:bas044.

[6] http://pir.georgetown.edu/pirwww/iprolink/