Scalable Text Mining Assisted Curation of Post-Translationally Modified Proteoforms in the Protein Ontology

Karen E. Ross and Darren A. Natale
Protein Information Resource
Georgetown University Medical Center
Washington, DC, USA
E-mail: ker25@georgetown.edu

Cecilia Arighi, Sheng-Chih Chen, Hongzhan Huang, Gang Li, Jia Ren, Michael Wang, K. Vijay-Shanker and Cathy H. Wu
Center for Bioinformatics and Computational Biology
University of Delaware
Newark, DE, USA

The Protein Ontology Consortium

Abstract—**The Protein Ontology (PRO) defines protein classes and their interrelationships from the family to the protein form (proteoform) level within and across species. One of the unique contributions of PRO is its representation of post-translationally modified (PTM) proteoforms. However, progress in adding PTM proteoform classes to PRO has been relatively slow due to the extensive manual curation effort required. Here we report an automated pipeline for creation of PTM proteoform classes that leverages two phosphorylation-focused text mining tools (RLIMS-P, which detects mentions of kinases, substrates, and phosphorylation sites, and eFIP, which detects phosphorylation-dependent protein-protein interactions (PPIs)) and our integrated PTM database, iPTMnet. By applying this pipeline, we obtained a set of ~820 substrate-site pairs that are suitable for automated PRO term generation with literature-based evidence attribution. Inclusion of these terms in PRO will increase PRO coverage of species-specific PTM proteoforms by 50%. Many of these new proteoforms also have associated kinase and/or PPI information. Finally, we show a phosphorylation network for the human and mouse peptidyl-prolyl cis-trans isomerase (PIN1/Pin1) derived from our dataset that demonstrates the biological complexity of the information we have extracted. Our approach addresses scalability in PRO curation and will be further expanded to advance PRO representation of phosphorylated proteoforms.**

Keywords—Protein Ontology (PRO), text mining, post-translational modification, proteoform, phosphorylation

I. Introduction

The Protein Ontology (PRO) (proconsortium.org) [1] is an OBO Foundry ontology that defines classes of proteins and protein complexes and indicates how these classes interrelate. Classes defined in PRO can be either organism-independent or organism-specific and range in granularity from more general protein family classes to more specific proteoform classes (which account for the precise molecular form of a protein, including specification of sequence or splice variant and any post-translational modification [PTM]) [2]. It has long been appreciated that PTMs play a pivotal role in protein function, regulating activity, localization, and protein-protein interactions (PPIs), and that disruptions in PTM can lead to disease [3]. Recent advances in proteomics have revealed that the majority of human proteins undergo PTM, often on many sites [3]. The ability of PRO to represent the full variety of PTM proteoforms for each gene product, including proteoforms with combinations of multiple modifications, makes it an ideal resource for understanding PTM cross-talk and PTM-regulated functions. Thus, a major focus of the PRO curation effort is to represent and annotate PTM proteoforms and identify corresponding proteoforms across species (ortho-proteoforms).

There are currently three curation pipelines for creation of proteoform classes in PRO: (1) bulk import of data from other projects that characterize PTM proteoforms, including Reactome [4] and the Consortium for Top-Down Proteomics [5]; (2) requests for individual terms needed for Gene Ontology annotation in model organism databases (e.g, Mouse Genome Database [6]) or for semantic tagging (e.g., Alzforum [7]); and (3) in-house literature-based curation using a text mining assisted workflow [8]. The need for extensive manual review by domain experts has proved to be a major bottleneck in PRO curation. Moreover, coverage of PTM proteoforms in PRO reflects the organisms and pathways of interest to individual users. PRO presently contains ~2,550 PTM proteoform classes, including 1,700 organism-specific terms and 850 organism-independent parent classes. Of the organism-specific terms, about half were created via bulk data import while the remainder were created on an individual basis.

We have previously used two PTM-focused text mining tools to assist with manual curation of PTM proteoforms. The first tool, RLIMS-P [9] detects mentions of kinase, substrate, and phosphorylation site in free text; the second, eFIP [10], detects causal relationships between phosphorylation and PPIs (e.g., the binding between Bad

pSer-136 and 14-3-3 in the sentence: *Akt phosphorylates Bad at Ser136 and promotes the association of Bad with 14-3-3*. PMID: 17342096). Although these tools have considerably speeded up expert curation by pinpointing relevant information in the literature, they have an untapped potential in further automation of the curation process.

Concurrent with our text mining work, we have developed iPTMnet, an integrated resource for PTM network analysis (http://research.bioinformatics.udel.edu/iptmnet/; [11]. iPTMnet integrates text mining results from RLIMS-P and eFIP that have been automatically normalized (i.e., the proteins detected in text have been mapped to their corresponding UniProtKB identifiers) with data from multiple high-quality PTM resources (e.g., PhosphoSitePlus [3] and PhosphoGrid [12]), covering organisms from human to yeast.

Here we describe an automated workflow for creation of PTM proteoforms in PRO that takes advantage of the information we have integrated in the iPTMnet database. Key components of the workflow include i) full scale PubMed text mining using RLIMS-P/eFIP; ii) automatic normalization of protein entities in the text mining output; iii) validation of the text mining results by comparing to information in expert curated PTM resources; and iv) automatic generation of PRO terms, including logical and textual definitions, based on a standardized template. In our first application of this approach, we identified ~820 proteoforms with a single phosphorylation site that can be included in PRO. For many of these terms, we also automatically extracted kinase and/or interactant information, which can be used to annotate the terms. This work reflects a significant advance in our efforts to represent the landscape of PTM proteoforms in PRO.

II. Approach

A. Full Scale Text Mining and Entity Normalization

We have developed the text mining tools RLIMS-P [9] and eFIP [10] to mine kinase-substrate-site relationships and phosphorylation-dependent PPIs, respectively, from free text. The rule-based RLIMS-P has achieved F-scores (harmonic mean between precision and recall [13]) of 0.91, 0.92, and 0.95 for kinases, substrates, and sites, respectively, based on a corpus of PubMed abstracts [9]. It has been evaluated in the BioCreative Interactive Text Mining Task for usability and utility [14] and is being adopted for computer-assisted literature-based curation by several databases. eFIP employs RLIMS-P to detect mentions of phosphorylation and then examines one or two consecutive sentences for any mention of proteins that interact with the substrate. The textual position of this information relative to phosphorylation is then used to assess whether the phosphorylation event has a direct effect (positive or negative) on the interaction. In an evaluation on 100 sections of full-length articles from the PMC Open Access collection, eFIP achieved an F-score of 84% [10]. Results of full-scale RLIMS-P/eFIP mining of PubMed abstracts and PubMed Central Open Access (PMC) articles are stored in a local database. The stored information includes entities, relations, and evidence attribution.

Funding: NSF (ABI-1062520), NIH (R01GM080646), Delaware INBRE (P20GM103446), and institutional resources of Center for Bioinformatics and Computational Biology at University of Delaware.

To normalize the gene/protein names in the text mining results to UniProtKB accession numbers (ACs), we use PubTator [15] and the UniProt ID mapping service [16]. PubTator is a web interface that provides RESTful APIs to retrieve gene normalization results generated by GenNorm [17]. For each PMID, a list of gene mentions and their normalized Entrez IDs is retrieved. The Entrez IDs are then mapped to UniProtKB ACs using mapping information retrieved from the UniProt website. Any Entrez IDs that cannot be mapped to a UniProtKB AC are discarded. To improve data quality, we perform two integrity checks on the normalized results: (1) for substrates, we confirm that the mapped protein sequence has the correct residue at the position that is reported to be phosphorylated (e.g, if the phosphorylation site is Ser-100, we confirm that position 100 of the mapped sequence is a serine); and (2) for kinases, we check whether the corresponding UniProtKB record contains the keyword "kinase."

B. Integration of Text Mining Results with PTM Database Information: iPTMnet

iPTMnet (Fig. 1) integrates normalized results of full-scale text mining from RLIMS-P and eFIP with PTM data from several expert curated PTM resources for visualization and analysis of PTM networks. Underlying iPTMnet is an Oracle (11g release 2) database. The text mining results that are consumed by iPTMnet are the normalized RLIMS-P results from all PubMed abstracts and the normalized eFIP results from all PubMed abstracts and PMC full-length articles. For data integration, gene/protein names from the source databases, which are represented in a variety of formats, are mapped to UniProtKB ACs. We used the iPTMnet database as the source of PTM information for PRO proteoform term curation (see below).

C. Selection of PTM Proteoforms for Automated PRO Curation

To select PTM proteoforms for PRO curation (Fig. 1) we:

- Retrieved from the iPTMnet database all substrate-site pairs that were captured by RLIMS-P and at least one PTM database based on the same PMID(s) (Fig. 1, Step 1a). We excluded PMIDs where multiple phosphorylation sites were detected by RLIMS-P or by the corroborating database(s) because of the difficulty of automatically determining whether a combinatoric PTM proteoform (simultaneous phosphorylation on multiple sites) or independent singly phosphorylated proteofoms were being described. We also discarded cases with conflicts between the text mined and database information (e.g., due to errors in automated species assignment).
- Obtained normalized kinase and phosphorylation-dependent interactant information from the iPTMnet database for the selected substrate-site pairs (Fig. 1, Step 1b). After manual validation, this information can potentially be used to associate annotation with the PRO terms.
- Excluded PMIDs where the abstract contains language that suggests that PTMs other than

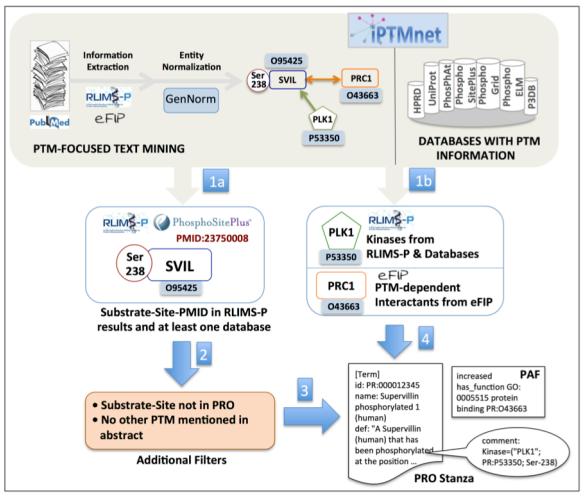

Fig. 1. Workflow for automated generation of PRO terms for PTM proteoforms. Substrate-site pairs identified by RLIMS-P and supported by at least one other resource are retrieved from the iPTMnet database (1a) along with any pertinent kinase or PPI information (1b). Additional filtering to remove cases that are already in PRO or are likely to be part of multiply modified proteoforms is performed (2). PRO stanzas are created based on a template (3) and annotation (e.g. PPIs) is added to the PRO Annotation File (PAF; 4).

phosphorylation are described (e.g., ubiquitin* and acetyl*). This check reduced the likelihood that the proteoform has other PTMs in addition to the single phosphorylation site (Fig 1, Step 2)

- Excluded cases where the substrate-site pair is already in PRO, either as a singly phosphorylated proteoform or as part of a multiply modified form (Fig. 1, Step 2). In addition, we excluded results that were extracted from PMIDs that were already curated by PRO as we reasoned that all proteoforms that are supported by those PMIDs are likely to have been identified in the expert curation process.

D. *Automated Generation of PRO Stanzas*

PRO terms can be created for PTM proteoforms that pass all data integrity checks using a template (Fig. 1 Step 3). If the substrate is mapped to a specific isoform of a protein, the name and text definition will additionally include the isoform number and the parent will be the organism-specific isoform. Associated kinase and/or PTM-dependent interactant information (i.e., eFIP results) will be prioritized for expert review. Kinase information will be added to the stanza comment line and interactant information will be added to the PRO Annotation File (PAF) following standard PRO curation procedures (Fig 1, Step 4)[1].

III. Results and Discussion

A. *Identification of PTM Proteoforms for Automated PRO Curation.*

From full-scale text mining of 25 million PubMed abstracts with RLIMS-P, we identified ~185,000 papers with kinase, substrate, and/or site information. After normalization of protein entities, we obtained ~5,300 normalized substrate-site pairs and ~1,550 kinase-substrate-site triples. Mining of PubMed abstracts and PMC full-length articles with eFIP identified ~8,500 articles with PTM-dependent PPI information; after normalization, we obtained ~770 substrate-site-interactant triples.

Of the ~5,300 substrate-site pairs from RLIMS-P, 1,033 were curated by another resource in the iPTMnet database based on the same PMID(s). Of these, we eliminated 94 because there was a conflict between the text mining results and the curated resource usually related to species assignment, 84 because the abstracts they were extracted from mentioned other PTMs and 78 because the site and/or PMID(s) were already in PRO (Note: some substrate-site pairs were eliminated for more than one of these reasons.) After these filtering steps, we obtained 818 substrate-site pairs [2] potentially suitable for automated PRO term generation. Of these, 731 (89%) have kinase information, including 285 (35%) with kinase information from RLIMS-P, and 93 (11%) have PPI information (from eFIP), which can be added to PRO as annotation after expert review.

Two curators manually reviewed the full-text articles for 91 substrate-site pairs randomly chosen from the list of 818 results. The number of results reviewed was determined by the time available to the curators. In 83 cases (91%), the evidence supported the existence of the singly-phosphorylated PTM proteoform identified by our automated approach. Of the remaining eight pairs, there was one case where the species was assigned incorrectly by all sources (text-mining and two databases) and seven cases where the article suggested that the proteoform had multiple phosphorylation sites, even though only a single site was captured by all sources. In one of the seven cases, the phosphorylation required prior phosphorylation on another site; thus, the singly phosphorylated form we proposed is unlikely to exist. Using the RLIMS-P web interface [9], we performed a keyword search for "priming", a term commonly used to describe sequential phosphorylation events, and found ~600 results (only 0.3% of total RLIMS-P results); also, our pipeline will filter out any of these cases where multiple sites are mentioned in the abstract. Therefore, we think that this type of error will be relatively rare. In the other six cases, the existence of the singly phosphorylated form was not ruled out; moreover, it is acceptable to create a PRO term that names only a subset of the modification sites in a multiply modified proteoform because, conformant to the Open World Assumption [18], PRO does not make any assertions about sites that are not explicitly named. PRO only asserts what is known based on the experimental results. Because the existence of other site modifications cannot be excluded, PRO definitions imply only that at least the explicit modifications have to be present. Thus, our evaluation indicates that our dataset is highly enriched for well-supported singly phosphorylated forms while containing very few errors (2/91 (2%)).

[1] PRO curation guidelines can be found on the PRO website (http://proconsortium.org).

[2] List available at: http://www.proteininformationresource.org/pro/iptmnet2pro.html

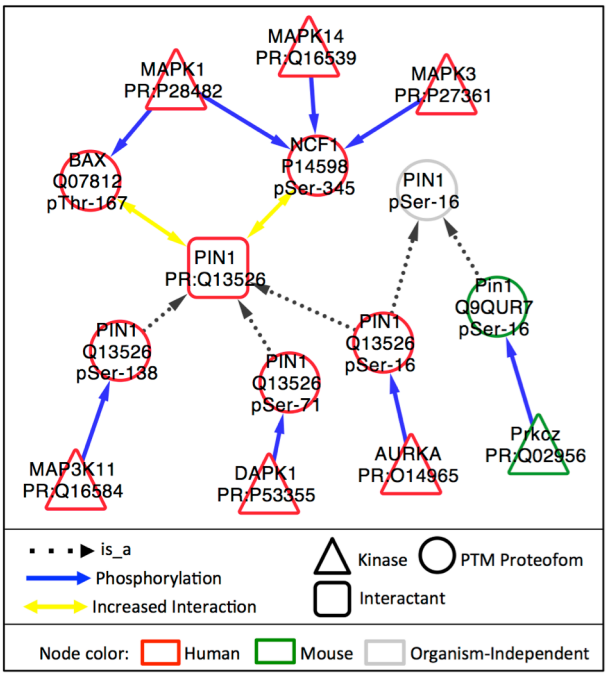

Fig. 2. Network of PIN1 kinases, PTM proteoforms, and phosphorylated interacting proteins from human and mouse.

B. Use Case: PIN1 Phosphorylation Network

Fig. 2 shows a network centered on the peptidyl-prolyl cis-trans isomerase PIN1/Pin1 (human/mouse) that illustrates the potential richness of the PTM information in our dataset and the advantages of using an ontological representation of PTM proteoforms. PIN1/Pin1 recognizes a phosphorylated motif in its binding partners and induces a conformational change [19]. Currently, the information in PRO about PIN1/Pin1 is limited—no PTM proteoforms of PIN1/Pin1 are described and only one case of PIN1 binding to a phosphoprotein (CCNE1 pSer-384, PR:000025637) is annotated. In our dataset, we found two human PTM proteoforms (NFC1 pSer-345 and BAX pThr-167) that bind to PIN1 in a phospho-dependent manner. Several kinases for these proteoforms were identified, including MAPK1, which phosphorylates both. In turn, we found three PTM proteoforms of PIN1 (pSer-16, pSer-71, and pSer-138), phosphorylated by multiple kinases. Interestingly, we also found a Ser-16 phosphorylated proteoform of mouse Pin 1. The human and mouse pSer-16 proteoforms can be connected at the ortho-proteoform level in the PRO hierarchy (Fig 2, grey node).

C. Conclusions and Future Work

Here we describe a workflow for automatic generation of PRO terms for PTM proteoforms based on text mining results with direct literature evidence attribution. When developing an automated curation pipeline, it is important to minimize inclusion of erroneous information; thus, we used stringent filtering criteria at the cost of discarding a great majority (~85%) of our normalized substrate-site pairs. Even with strict filters in place, we will be able to create ~820 new organism-specific PRO terms for PTM proteoforms, a 50% increase over the number of species specific PTM forms currently curated by PRO. As the use case demonstrates, this approach can provide rich information on PTM sites, PTM enzymes, biological consequences of PTM (i.e. PTM-dependent PPI), and orthologous proteoforms across species. At the same time, the automatic detection and normalization of kinase and PPI information will greatly reduce the manual effort required for annotation of the automatically created PRO terms.

In this study, we focused exclusively on data supported by text mining results; however, our approach could be applied to substrate-site pairs that are reported in any two resources in the iPTMnet database. We also plan to identify proteoform candidates from full-text RLIMS-P results. It has been observed that ~90% of phosphorylation sites are mentioned only in the body of an article (not the abstract) [9, 20] so full-text mining should greatly increase our yield of proteoforms as well as improve data integrity. Finally, we are considering approaches for automated detection of proteoforms with multiple PTMs. It is often very challenging for a curator, let alone an automated system, to determine whether experimental evidence supports the existence of a proteoform with multiple PTMs as opposed to a population of proteins with individual modifications. One possibility would be to make use of PTM proteomic data. Bottom-up proteomic data is usually not useful for detecting PTM combinations because the proteins are cleaved into short peptides before identification. If a protein has several phosphorylated residues, they will typically be separated across multiple peptides, making it impossible determine whether they were orignally present on the same protein molecule. However, if two phosphorylation sites on a protein are close enough, they could potentially be found on the same peptide. In these cases, proteomic data could be used as evidence in support of the multiply modified proteoform.

In conclusion, we have implemented an automated workflow using text mining results and curated database information to create new PRO terms for PTM proteoforms. This approach, which can achieve large gains in curation efficiency without compromising quality, can significantly expand the ontological representation of PTM.

REFERENCES

[1] D.A. Natale, et al., "Protein Ontology: a controlled structured network of protein entities," Nucleic Acids Res, vol. 42, pp. D415-421, 2014.

[2] L.M. Smith, N.L. Kelleher, and P. Consortium for Top Down, "Proteoform: a single term describing protein complexity," Nat Methods, vol. 10, pp. 186-187, 2013.

[3] P.V. Hornbeck, et al., "PhosphoSitePlus, 2014: mutations, PTMs and recalibrations," Nucleic Acids Res, vol. 43, pp. D512-520, 2015.

[4] A. Fabregat, et al., "The Reactome pathway Knowledgebase," Nucleic Acids Res, vol. 44, pp. D481-487, 2016.

[5] X. Dang, et al., "The first pilot project of the consortium for top-down proteomics: a status report," Proteomics, vol. 14, pp. 1130-1140, 2014.

[6] C.J. Bult, et al., "The Mouse Genome Database: enhancements and updates," Nucleic Acids Res, vol. 38, pp. D586-592, 2010.

[7] J. Kinoshita and T. Clark, "Alzforum," Methods Mol Biol, vol. 401, pp. 365-381, 2007.

[8] K.E. Ross, et al., "Construction of protein phosphorylation networks by data mining, text mining and ontology integration: analysis of the spindle checkpoint," Database (Oxford), vol. 2013, pp. bat038, 2013.

[9] M. Torii, et al., "RLIMS-P 2.0: A Generalizable Rule-Based Information Extraction System for Literature Mining of Protein Phosphorylation Information," IEEE/ACM Trans Comput Biol Bioinform, vol. 12, pp. 17-29, 2015.

[10] C.O. Tudor, et al., "Construction of phosphorylation interaction networks by text mining of full-length articles using the eFIP system," Database (Oxford), vol. 2015, 2015.

[11] K.E. Ross, et al., "iPTMnet: Integrative Bioinformatics for Studying PTM Networks," Methods Mol Biol, vol. in press.

[12] I. Sadowski, et al., "The PhosphoGRID Saccharomyces cerevisiae protein phosphorylation site database: version 2.0 update," Database (Oxford), vol. 2013, pp. bat026, 2013.

[13] R. Rodriguez-Esteban, "Biomedical text mining and its applications," PLoS Comput Biol, vol. 5, pp. e1000597, 2009.

[14] C.N. Arighi, et al., "An overview of the BioCreative 2012 Workshop Track III: interactive text mining task," Database (Oxford), vol. 2013, pp. bas056, 2013.

[15] C.H. Wei, H.Y. Kao, and Z. Lu, "PubTator: a web-based text mining tool for assisting biocuration," Nucleic Acids Res, vol. 41, pp. W518-522, 2013.

[16] C. UniProt, "Update on activities at the Universal Protein Resource (UniProt) in 2013," Nucleic Acids Res, vol. 41, pp. D43-47, 2013.

[17] C.H. Wei and H.Y. Kao, "Cross-species gene normalization by species inference," BMC Bioinformatics, vol. 12 Suppl 8, pp. S5, 2011.

[18] R. Stevens, et al., "Using OWL to model biological knowledge," International Journal of Human-Computer Studies, vol. 65, pp. 583-594, 2007.

[19] T.H. Lee, et al., "Death-associated protein kinase 1 phosphorylates Pin1 and inhibits its prolyl isomerase activity and cellular function," Mol Cell, vol. 42, pp. 147-159, 2011.

[20] A.L. Veuthey, et al., "Application of text-mining for updating protein post-translational modification annotation in UniProtKB," BMC Bioinformatics, vol. 14, pp. 104, 2013.