Knowledge Representation of Protein PTMs and Complexes in the Protein Ontology: Application to Multi-Faceted Disease Analysis

Karen E. Ross[1], Catalina O. Tudor[1], Gang Li[1], Ruoyao Ding[1], Irem Celen[1], Julie Cowart[1], Cecilia N. Arighi[1], Darren A. Natale[2] and Cathy H. Wu[1,2]

[1]Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, USA
[2]Protein Information Resource, Georgetown University Medical Center, Washington, DC, USA
E-mail: ross@dbi.udel.edu

Abstract—**Alterations in protein post-translational modification (PTM) and PTM cross-talk are increasingly being appreciated as driving mechanisms of human disease. The Protein Ontology (PRO) is a valuable resource to study the relation between PTM and disease because it represents individual proteins and protein complex subunits at the proteoform (e.g., isoform, PTM form, and sequence variant) level, with links to their functional properties. We constructed a multi-relation network that represents knowledge obtained from large scale text-mining for phosphorylation-dependent protein-protein interactions (PPIs) and their disease associations, built on the PRO framework for representation of PTM forms, complexes, and protein families, as well as their attributes and relationships. We then conducted two case studies that demonstrate the use of PRO in disease analysis. (i) We performed cross-species comparisons of two glioma-associated phosphorylated proteoforms of the human DNMT1 methylase, which revealed that the forms are not strictly conserved in mouse, a frequently used glioma model system. (ii) We used PRO-defined proteoforms of the oncoprotein beta-catenin phosphorylated on various combinations of the N-terminal sites, Ser-33, Ser-37, Thr-41, and Ser-45, to interpret a hierarchical clustering analysis of cancer types based on their pattern of mutations in these sites. The cancers formed two major clusters: one with mutations in Ser-33/Ser-37/Thr-41 and the other with mutations in Thr-41/Ser-45. Proteoform-specific annotation in PRO suggests that stabilization of beta-catenin may play a role in oncogenesis in the first group, whereas alterations in beta-catenin transcriptional or cell adhesion activity may play a more important role in the second group. Together, these scenarios illustrate the general applicability of PRO to disease understanding. Future plans include the integration of PRO with other semantic resources to increase our ability to address these problems with computational reasoning.**

Keywords—Protein Ontology; phosphorylation; text mining; beta-catenin; cancer

I. INTRODUCTION

Aberrations in protein post-translational modification (PTM), resulting from genetic variations that affect individual PTM sites as well as alterations in PTM enzyme activity that have global effects on the balance of PTM forms in the cell, have been implicated in many diseases [1, 2]. Protein phosphorylation, in particular, has been recognized as a central disease-driving mechanism, leading to the development of kinase inhibitors as therapeutic agents [1]. With state-of-the-art text mining tools, it is now possible to extract detailed information about proteins, PTMs, and diseases from the literature on a large scale. Tools such as RLIMS-P [3] and eFIP [4] have captured a wealth of information on phosphorylated protein forms, their modifying enzymes, and the functional impact of phosphorylation, with a minimum of manual curator effort.

Despite these advances, representing this information in a form that is useful for both human interpretation and computational reasoning is challenging. It is important not only to link genetic variant information with its effects on protein sequence and function but also to capture the impact of imbalances of particular PTM forms and complexes on disease.

Used in conjunction with other bioinformatic resources, the Protein Ontology (http://pir.georgetown.edu/pro/pro.shtml; PRO [5]) enables the structured representation and interpretation of this information. PRO, a member of the Open Biomedical Ontologies (OBO) foundry, represents proteoforms (e.g., isoforms, PTM forms, and sequence variants) [6] and protein complexes and their relationships within and across species. Once a proteoform or complex is defined in PRO it can be annotated with functional and/or disease information derived from the scientific literature or bioinformatic databases. This framework can then support the analysis of biological processes in health and disease.

Mouse models have been critical for understanding human diseases. These models rely on the high degree of conservation of proteins and pathways between human and mouse. Although protein phosphorylation sites in human are also highly conserved in mouse (92% conserved in one study [7]), conservation at the proteoform level, which can have significant functional consequences in a disease model has not been assessed. With its emphasis on representation of proteoforms and cross-species relationships, PRO is a valuable resource for this type of analysis.

In this article, we leverage PRO to: (i) create a phosphorylation network based on large-scale text mining of phosphorylation-dependent PPIs that impact disease; (ii) facilitate cross-species comparison of proteoforms of the DNMT1 methylase that are associated with glioblastoma in humans; and (iii) interpret patterns of mutations in the beta-catenin oncoprotein observed in different cancer types. These examples illustrate the variety of ways in which PRO can be used to represent disease knowledge and gain insight into disease mechanisms.

II. Methods

To identify phosphorylation-dependent PPIs described in literature, all PubMed abstracts and PubMedCentral (PMC) Open Access full-length articles were processed using eFIP, which identifies mentions of phosphorylation-dependent PPIs. The kinases, phosphorylated proteins (substrates), and interacting partners (interactants) were mapped to UniProtKB identifiers, when possible, using GeneNorm [8]. Disease mentions in article titles or abstracts were computationally detected by matching to a custom dictionary of disease terms based on MeSH and MedlinePlus. PRO terms for proteoforms and complexes were created as described in [9]. Term annotation such as binding partners and disease association is stored in the PRO annotation file (PAF), and PTM enzyme information is recorded in the comments section of the OBO stanza using structured vocabulary. All terms and annotations are available upon request and will be made public on the PRO website and in downloadable files as part of PRO release 43 (September 2014). The network was constructed using Cytoscape 3.0 [10]. Sequence alignment was performed using Clustal Omega [11] and visualized with Jalview 2.8.1 [12]. Cancer-associated mutations in beta-catenin were obtained from Catalog of Somatic Mutations in Cancer (COSMIC) [13]. Tumor tissue types that had at least 10 mutations in the beta-catenin N-terminal phosphorylation sites Ser-33, Ser-37, Thr-41, and Ser-45 were used. For each tissue the proportion of mutations at each site was calculated relative to the total number of mutations at all four sites. The heatmap was constructed using the heatmap.2 function of R (version 3.0.2; http://www.r-project.org/) with default parameters.

III. Results and discussion

A. Network Representation of Disease-Associated Phoshorylation-Dependent Protein-Protein Interactions

Text-mining of over 23 million PubMed abstracts and 800,000 PMC open access full-length articles using eFIP identified sentences describing PPIs that were dependent on the phosphorylation state of one of the interactants in over 13,000 articles. About 500 articles also had phosphorylation site information, UniProtKB-mapped substrates and interactants, and mention of disease in the title or abstract. Through manual curation of the 109 most recent articles, we found 52 disease-associated phospho-dependent PPIs in 39 articles. (In the remaining articles the disease mention was not causally related to the PPI.) A multi-relation network based on some of these results, including phospho-dependent PPIs, PTM enzyme-PTM form relationships, proteoform/complex-disease relationships, and relations among PRO terms, is shown in Fig. 1.

This network illustrates several ways in which the PRO framework allows curators to precisely represent complex biological information. First, because PRO treats each proteoform as a separate entity, multiple forms of a protein can be defined and individually annotated. For example, the Tyr-357-phosphorylated form (PR:000037508) and the Ser-127 phosphorylated form (PR:000037510) of YAP1 differ in their ability to bind 14-3-3 proteins (PR:000003237) and the apoptosis regulatory protein p73 (PR:O15350) and exhibit different associations with cancer and Alzheimer's Disease. PRO can also represent forms with multiple types of modification, facilitating the description of PTM cross-talk (e.g., CCND1 Thr-286-phosphorylated and ubiquitinated form (PR:000037512)). Second, PRO represents protein complexes as distinct entities to which complex-specific annotation can be attached. Moreover, complex subunits are defined using PRO terms, enabling the specification of which particular proteoforms (e.g., phosphorylated forms) are part of the complex. For example, the proteoform of the DNMT1 DNA methlyase that lacks phosphorylation on Ser-127 and Ser-143 (PR:000037504) forms a complex with the DNA-associated factors PCNA (PR:P12004) and UHRF1 (PR:Q96T88). This complex (PR:000037517) has been associated with tumor suppression [14]. Third, protein terms in PRO are defined at multiple levels of granularity from the family level down to the isoform and/or modification level. Thus, when describing a biological relationship involving a protein, the term that is most appropriate given the current state of knowledge can be used. For example, because 14-3-3 proteins are encoded by several genes, and the protein products of these genes are not always distinguishable in experimental assays, 14-3-3 proteins are represented by the class PR:000003237 that encompasses the protein products of all 14-3-3 family genes. Similarly, when the protein is known to be the product of a particular gene, but no isoform information is available, a gene-level PRO term that encompasses all protein products of a gene is used (e.g., TP73 (PR:O15350)). Integration of PRO terms into a multi-relation network context further allows identification of proteoforms sharing common PTM enzymes (e.g., AKT (PR:000029189)) or interacting partners (e.g., 14-3-3 (PR:000003237)) or implicated in the same diseases.

B. Cross-Species Comparison of Proteoforms of the Glioma-Associated DNA Methylase, DNMT1

DNMT1 phosphorylation on Ser-127 or Ser-127/Ser-143 and the concomitant reduction in binding to UHRF1 and PCNA (Fig. 1) has been associated with glioma in humans [14]. Because the mouse is often used as a model system for studying glioma, we investigated whether the glioma-associated DNMT1 proteoforms are conserved in mouse as well as in several other mammals (Fig. 2). PRO representation of proteoforms enables the cross-species comparison of PTM at the PTM-form level, which is more likely to reflect functional conservation than comparisons of the individual sites alone. While human Ser-143 is conserved across all species, Ser-127 is found only in other primates (red/pink

This work has been supported by the National Science Foundation [ABI-1062520] and the National Institutes of Health [5R01GM080646-08].

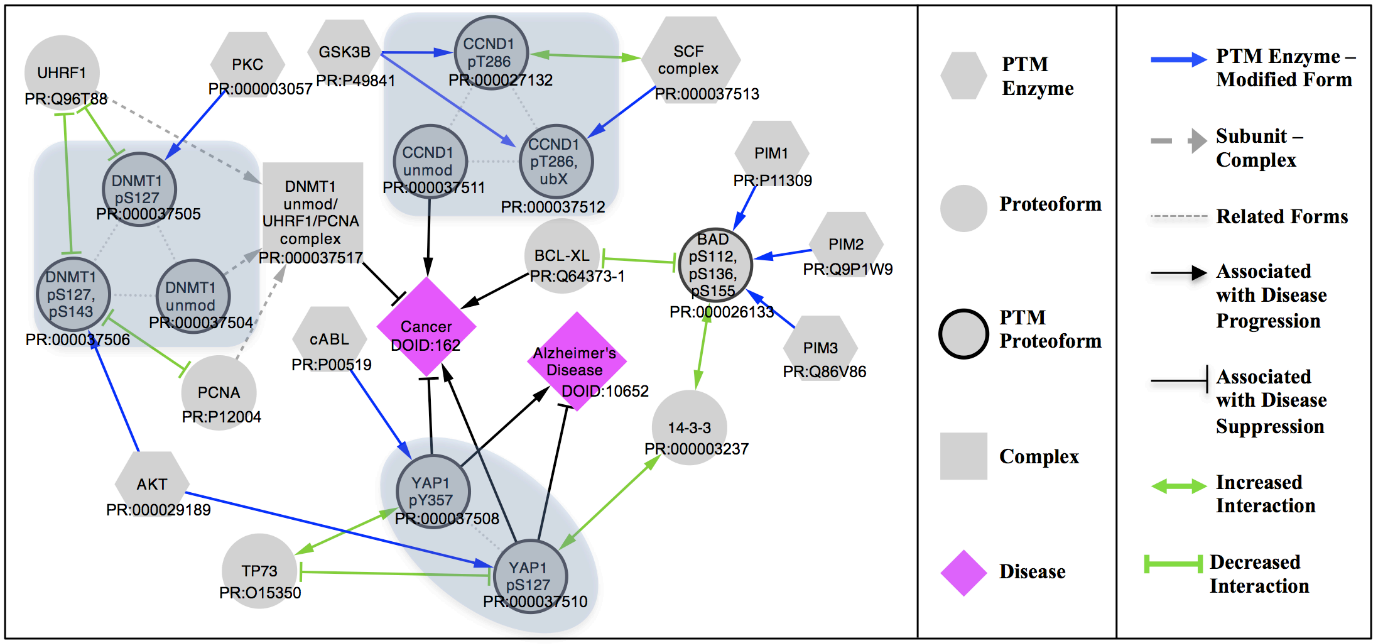
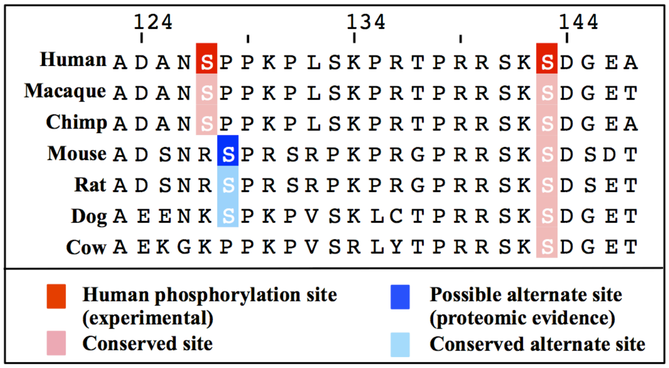


Fig. 1. Multi-relation network showing partial text mining results for disease-associated phosphorylation-dependent protein-protein interactions. PRO terms for proteoforms and complexes and Disease Ontology terms for diseases are indicated.

residues). Thus, neither the Ser-127 phosphorylated form nor the Ser-127/Ser-143 phosphorylated form of DNMT1 (Fig. 1) is strictly conserved in mouse, rat, dog, or cow, suggesting that non-primates might use a different mechanism for regulating DNMT1 interaction with PCNA and UHRF1. However, mouse, rat, and dog (but not cow) have a serine at the adjacent position (blue/light blue residues), which could potentially fulfill the same role as human Ser-127. This serine has been shown to be phosphorylated in mouse in a high-throughput phospho-proteomic study [15]. Further studies are needed to clarify the role of DNMT1 phosphorylation in a mouse glioma model system.


```
                124                  134                 144
Human   A D A N S P P K P L S K P R T P R R S K S D G E A
Macaque A D A N S P P K P L S K P R T P R R S K S D G E T
Chimp   A D A N S P P K P L S K P R T P R R S K S D G E A
Mouse   A D S N R S P R S R P K P R G P R R S K S D S D T
Rat     A D S N R S P R S R P K P R G P R R S K S D S E T
Dog     A E E N K S P K P V S K L C T P R R S K S D G E T
Cow     A E K G K P P K P V S R L Y T P R R S K S D G E T
```

🟥 Human phosphorylation site (experimental)
🟦 Possible alternate site (proteomic evidence)
Conserved site
Conserved alternate site

Fig. 2. Partial sequence alignment of the DNMT1 DNA methylase from several mammalian species showing degree of conservation of the human phosphorylation sites Ser-127 and Ser-143.

C. *Analysis of Cancer-Associated Mutations in beta-catenin Phosphorylation Sites*

Beta-catenin is a multi-functional protein involved in cell-cell adhesion and transcriptional regulation [16]. Several of its key transcriptional targets drive cell proliferation, and excessive beta-catenin transcriptional activity is oncogenic. Beta-catenin stability is regulated by phosphorylation of four residues in the N-terminus. Casein kinase I phosphorylates Ser-45 of beta-catenin, which promotes the sequential phosphorylation of Thr-41, Ser-37, and Ser-33 by GSK3-beta. Phosphorylation at Ser-37 and Ser-33 enables recognition of beta-catenin by the ubiquitin ligase beta-TrCP, which targets it for degradation by the proteosome. Mutations in the four phosphorylation sites stabilize the protein and have been associated with cancer.

Through text mining with RLIMS-P and eFIP, we defined four beta-catenin proteoforms phosphorylated at different combinations of the N-terminal sites, with distinct sub-cellular localizations, binding partners, and activities (Fig. 3A). To gain insight into the role of these proteoforms in cancer, we used data from COSMIC on cancer-associated mutations in these sites to perform hierarchical clustering of different cancer types (Fig. 3B). The cancers fall into two major clusters with different mutation patterns. Cluster 1 cancers (pink box) are characterized by mutations at Ser-33 and Ser-37, with few mutations at Ser-45. Conversely, Cluster 2 cancers (blue box) are predominantly mutated at Ser-45. Both clusters show intermediate levels of Thr-41 mutations. The Ser-33/Ser-37 mutation the pattern in Cluster 1 suggests that oncogenesis in these cancer types is related to the lack of proteoforms 1 and 2. Both of these forms are unstable due to their association with

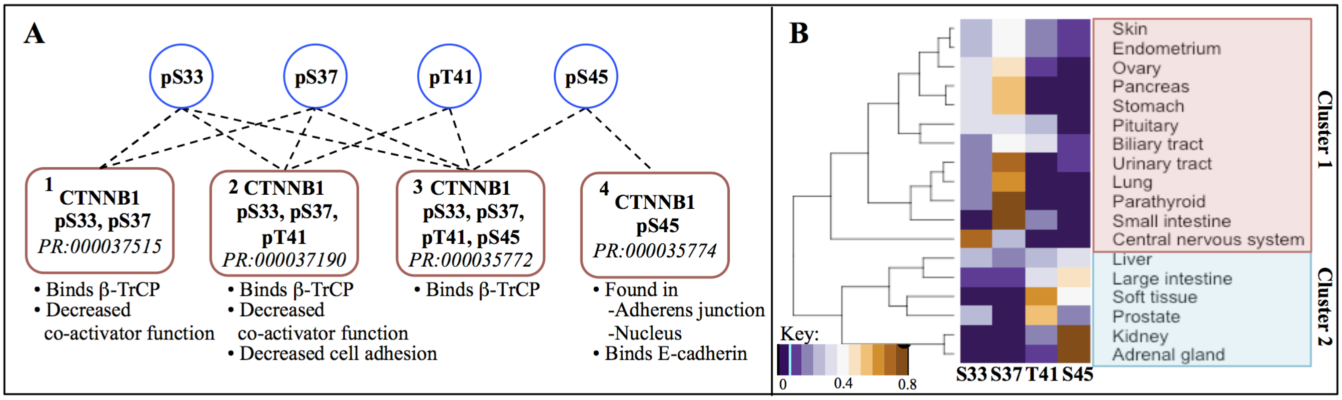


Fig. 3. (A) Proteoforms of beta-catenin phosphorylated on several combinations of the N-terminal phosphorylation sites Ser-33, Ser-37, Thr-41, and Ser-45 with partial functional annotation. (B) Hierarchical clustering of cancer types based on their pattern of mutations in these phosphorylation sites.

the ubiquitin ligase beta-TrCP. Thus, beta-catenin stabilization may be playing an important role in these cancers. Kinases, such as HIPK2, that can phosphorylate Ser-33 and Ser-37 without prior phosphorylation of Ser-45 may be regulating beta-catenin stability in these tissues [17]. Cluster 2 cancers have relatively few mutations in the residues that bind beta-TrCP; instead, these cancers are associated with lack of Ser-45 phosphorylated proteoform 3. Unlike other beta-catenin proteoforms, proteoform 3 is found in the nucleus and may be a key transcriptionally active form of beta-catenin [18]; this proteoform can also bind to the adhesion molecule E-cadherin [19]. Thus, alterations in beta-catenin transcriptional and cell adhesion activity independent of beta-catenin levels may contribute to Cluster 2 cancers. This example highlights the value of integrating experimental disease information from bioinformatic resources such as COSMIC with PRO representation of proteoforms to gain new insight into disease.

IV. Conclusions and Future Work

Through the structured representation of proteoforms and complexes PRO facilitates: (i) representation of proteoform-disease relations identified by large-scale text mining; (ii) cross-species comparisons at the proteoform level for evaluation of the relevance of animal models of disease; and (iii) interpretation of disease-associated mutation patterns. Currently, the PRO terms curated in this project can be viewed on the PRO website by biologists interested in PTM, PPI, and disease relationships. We are working toward formalizing these relationships and disseminating them in standard semantic web format (e.g. RDF/XML) to enable computational reasoning and hypothesis generation.

References

[1] L.M. Graves, J.S. Duncan, M.C. Whittle, and G.L. Johnson, "The dynamic nature of the kinome," Biochem J, vol. 450, pp. 1-8, 2013.

[2] B.M. Kessler, "Ubiquitin - omics reveals novel networks and associations with human disease," Curr Opin Chem Biol, vol. 17, pp. 59-65, 2013.

[3] X. Yuan, et al., "An online literature mining tool for protein phosphorylation," Bioinformatics, vol. 22, pp. 1668-1669, 2006.

[4] C.O. Tudor, et al., "The eFIP system for text mining of protein interaction networks of phosphorylated proteins," Database (Oxford), vol. 2012, pp. bas044, 2012.

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

[6] 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.

[7] R. Malik, E.A. Nigg, and R. Korner, "Comparative conservation analysis of the human mitotic phosphoproteome," Bioinformatics, vol. 24, pp. 1426-1432, 2008.

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

[9] 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.

[10] M.E. Smoot, et al., "Cytoscape 2.8: new features for data integration and network visualization," Bioinformatics, vol. 27, pp. 431-432, 2011.

[11] F. Sievers and D.G. Higgins, "Clustal Omega, accurate alignment of very large numbers of sequences," Methods Mol Biol, vol. 1079, pp. 105-116, 2014.

[12] A.M. Waterhouse, et al., "Jalview Version 2--a multiple sequence alignment editor and analysis workbench," Bioinformatics, vol. 25, pp. 1189-1191, 2009.

[13] S.A. Forbes, et al., "COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer," Nucleic Acids Res, vol. 39, pp. D945-950, 2011.

[14] E. Hervouet, et al., "Disruption of Dnmt1/PCNA/UHRF1 interactions promotes tumorigenesis from human and mice glial cells," PLoS One, vol. 5, pp. e11333, 2010.

[15] M. Trost, et al., "The phagosomal proteome in interferon-gamma-activated macrophages," Immunity, vol. 30, pp. 143-154, 2009.

[16] T. Valenta, G. Hausmann, and K. Basler, "The many faces and functions of beta-catenin," EMBO J, vol. 31, pp. 2714-2736, 2012.

[17] E.A. Kim, et al., "Homeodomain-interacting protein kinase 2 (HIPK2) targets beta-catenin for phosphorylation and proteasomal degradation," Biochem Biophys Res Commun, vol. 394, pp. 966-971, 2010.

[18] M.T. Maher, et al., "Beta-catenin phosphorylated at serine 45 is spatially uncoupled from beta-catenin phosphorylated in the GSK3 domain: implications for signaling," PLoS One, vol. 5, pp. e10184, 2010.

[19] M.C. Faux, et al., "Independent interactions of phosphorylated beta-catenin with E-cadherin at cell-cell contacts and APC at cell protrusions," PLoS One, vol. 5, pp. e14127, 2010.