Representing Modification Sites in PRO

Jonathan P. Bona
University at Buffalo
Buffalo, NY
Email: jpbona@buffalo.edu

Jenny Rouleau
Canisius College
Buffalo, NY
Email: rouleauj@canisius.edu

Alan Ruttenberg
University at Buffalo
Buffalo, NY
Email: alanrutt@buffalo.edu

***Abstract*—This paper presents a model for explicit representations of amino acid sites in the protein ontology. We handle sites that are the locations of post-translational modifications, focusing on histone proteins as our initial test case. The work explicitly represents both the entities involved (sites, residues, etc), and commonly used information about those entities such as positions relative to a reference sequence.**

I. Introduction

The Protein Ontology [1] (PRO) "... provides an ontological representation of protein-related entities by explicitly defining them and showing the relationships between them." [1] Representable entities involved in a posttranslational modification (PTM) include: the protein itself, the amino acid residue(s), the modification process, the modifying enzymes, chemical groups, and *the location at which the modification takes place*.

One type of protein-related entity that will benefit from more explicit representations than are currently used is *site*, or *location*. Examples include cleavage sites, domain binding sites, sites in secondary structures, mutation sites, and modification sites.

This work focuses on representing sites of posttranslational modifications. Specifically, we deal with sites that contain amino acids that undergo chemical modifications, e.g. phosphorylation, acetylation, and so on. The approach can be generalized to represent many types of relevant sites and their relationships to the proteins that host them, though we focus initially on modifications that involve change to a single amino acid.

In addition to terms and relations about the biological entities involved, we also represent information that is typically included in existing descriptions of these entities, such as the numeric position of a residue on a reference sequence.

II. Existing Representations of Modifications & Sites

Existing resources that represent information about protein modifications do not generally make explicit representations of all of the entities involved.

For instance, the Histone Infobase and other sources use a string of characters like "`H3T3ph`" [2] to name a modification, which is then further described in natural language text, along with PubMed IDs for provenance. The string "`H3T3ph`" is intended to succinctly express the fact that H3 histones can be modified by having the amino acid residue at position 3, which is Threonine, become phosphorylated. In this case, the information about the modification is linked to a page with information about the kinase responsible for carrying out the phosphorylation. Also in this case the modification represented has been observed to occur in some – but not all – histone H3 variants. However, the database does not appear to include a term, page, or other explicit representation of the entity that is the *Histone H3.3 variant modified in this way*.

This way of representing information about PTMs makes it difficult to use data from this source in combination with data from other sources without the intervention of a human being familiar with the domain. A further complication is that even those sources that consistently name locations using the amino acid residue present there along with a number to indicate the distances (in units of one amino acid) often have inconsistencies due to different treatment of the N-terminal initiator methionine (at position 1), which is usually removed. As an example, consider the UnitproKB[2] accession `P84243`[3], which is one of the H3 variants that undergoes the PTM above called `H3T3ph`. This modification is better treated in Uniprot's interface than in Histome, but Uniprot includes the initiator methionine in its position indices, which makes the representations used by the two resources inconsistent.

A. Example Competency Questions

The following are examples of the sort of questions that it would be useful to ask about protein modifications. Representations of modifications and the entities involved should allow these questions to be answered computationally, e.g. with the use of SPARQL queries:

- Given a protein, what PTMs does it have - what is the original residue type, modified residue type, and position with reference to uniprot sequence?
- What proteins in family X are known to have phosphorylated sites?
- Given a protein modification site, what other modifications of the site are known?
- Which types of protein have different functions conferred by being acetylated?
- What PTMs are conserved across species?

A major motivation for this work is to craft well-structured and accurate representations in OWL of the entities involved

[1] http://pir.georgetown.edu/pro/pro.shtml

[2] http://www.actrec.gov.in/histome/ptm_sp.php?ptm_sp=H3T3ph

[3] http://www.uniprot.org/uniprot/P84243

TABLE I. PR:000036802

PRO ID	PR:000036802
PRO Name	histone H3.3 acetylated and methylated 2 (mouse)
Synonyms	mH3F3B/AcMeth:2 (EXACT)PRO-short-label
Definition	A histone H3.3 that has been acetylated on the N-terminal Met and methylated on several Lys residues in mouse. UniProtKB:P84244, Met-1, MOD:00058—Lys-10/Lys-28, MOD:00083—Lys-19, MOD:00085. [TDR:PFR9332]
Comment	Category=organism-modification. Note=Top down proteomics.
Hierarchical relationship	Parent: PR:000008425 histone H3.3 Children: none only_in_taxon NCBITaxon:10090 Mus musculus

in PTMs that will facilitate inferring and retrieving the answers to such questions.

B. Existing Representations in PRO

Table I shows the information already existing in PRO for a particular modified H3.3 histone protein (http://purl.obolibrary.org/obo/PR_000036802).

By reading the name, text definition, and other fields, and by following links to Uniprot or PSI-MOD, a reader familiar with the subject matter comes to understand that the protein that this term stands for has three modifications to lysine residues, which are located at numeric positions 10, 19, and 28 relative to the N-terminus of a reference sequence. The counting scheme to produce these particular numeric positions seems to include the N-terminus Methionine, which is also described as being modified (acetylated). The linked Uniprot page for this protein's reference sequence describes the initiator Methionone as "removed." The PSI-MOD ids embedded in the definition text indicate more details about the nature of the modifications: `MOD:00085`[4] is defined as *A protein modification that effectively converts an L-lysine residue to N6-methyl-L-lysine*, while `MOD:00083`[5] converts the L-lysine residues at positions 10 and 28 to *N6,N6,N6-trimethyl-L-lysine*

III. Histone Modification Test Case

This work uses as a representation test case PTMs in histone proteins. Chromatin in the nucleus of eukaryotic cells is comprised of histones together with DNA. Posttranslational modifications to the histones change the structure of the chromatin and are hypothesized to form a "code" of downstream effects, including changes to the transcription of DNA[3].

Because histone modifications are of particular significance, there is increasing interest in discovering and cataloging the possible modifications, their combinations, and their functions. There are relatively few histone types, even including known variants. The situation is thus that there are a few proteins that play a central role in the nucleus of eukaryotic cells, and modifications to which are believed to form a complex code affecting the cell's behavior, and for which there are ongoing efforts to collect ever more data. By developing correct, precise, and computable representation schemes for these facts, and using those to represent existing knowledge about histone modifications, as well as knowledge that is generated by new assays, we aim to facilitate sharing and use of that data, and discovery of unknown facts it entails.

We have collected a preliminary histone modification data set from HIstome: The Histone Infobase [6]. This data includes information on all five human histone types and fifty five variants thereof, with one hundred and six observed modifications, as well as disease associations. The modification types included are arginine citrullination and methylation; lysine acetylation, biotinylation, methylation, ribosylation, ubiquitination; and phosphorylation of serine, threonine, and tyrosine. Histome PTM records specify the histone variant involved, the amino acid residues, and a location as the position in the amino acid sequence that makes up the primary structure of the protein. They are also linked to UniprotKB accessions for the proteins involved.

IV. Representation of Sites

This section describes in detail our approach to representing modification sites. While the basic scheme is in place, the work continues to evolve as we are now adding representations of many different histone protein modifications based on data from the Histone Infobase site, and on data gathered by top-down proteomics.

Our work in progress on representing protein, and specifically histone, modifications can be followed at: http://ctde.net/page/Protein_Modifications. The draft OWL document and related resources can be viewed from that page, or accessed directly on the **pro-ontology** Google Code repository at https://code.google.com/p/pro-ontology/source/browse/trunk/src/ontology/protein-sites/protein-site.owl

A. Site Classes

We represent PTM locations as subclasses of the Basic Formal Ontology [4][5] term `bfo:site`[7], which is elucidated as: *b is a site means: b is a three-dimensional immaterial entity that is (partially or wholly) bounded by a material entity or it is a three-dimensional immaterial part thereof. (axiom label in BFO2 Reference: [034-002]).*

We reify PTM locations using the classes

- `amino acid chain site`,
- `site of an amino acid residue in a protein`, and
- `site of post translationally modified amino acid residue`,

Figure 1 shows the subclass relations between these and their superclass, `bfo:site`.

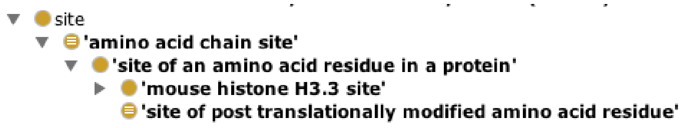

site
'amino acid chain site'
'site of an amino acid residue in a protein'
'mouse histone H3.3 site'
'site of post translationally modified amino acid residue'

Fig. 1. Types of Sites

[4] http://www.ebi.ac.uk/ontology-lookup/?termId=MOD:00085

[5] http://www.ebi.ac.uk/ontology-lookup/?termId=MOD:00083

[6] http://www.actrec.gov.in/histome/index.php

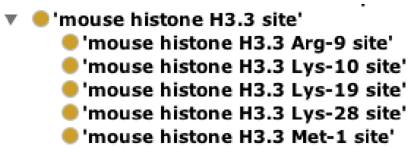
[7] http://purl.obolibrary.org/obo/BFO_0000029

A protein is made up of a chain of amino acid residues. Each amino acid residue occupies (bfo:has_location) a site. The PRO term protein[8] is a subclass of amino acid chain, defined as *A molecular entity that is a polymer of amino acids linked by peptide bonds [PRO:DAN]*

Many but not all amino acid chains are part of proteins. Every instance of site of an amino acid residue in a protein and its subclass site of post translationally modified amino acid residue is part_of some protein.

Figure 1 also shows the term mouse histone H3.3 site, which we have defined as *The site of an amino acid residue in a mouse histone H3.3 protein.*

Each instance of mouse histone H3.3 site is part_of some amino acid chain[9] and is the location of some amino acid residue[10].



▼ ● 'mouse histone H3.3 site'
● 'mouse histone H3.3 Arg-9 site'
● 'mouse histone H3.3 Lys-10 site'
● 'mouse histone H3.3 Lys-19 site'
● 'mouse histone H3.3 Lys-28 site'
● 'mouse histone H3.3 Met-1 site'

Fig. 2. Mouse Histone H3.3 Sites

The subclasses of mouse histone H3.3 site, shown in Figure 2 represent specific sites. For example, mouse histone H3.3 Lys-19 site is a specific site. The label used for the term is mnemonic to suggest to a human reader that this is a site on a mouse histone H3.3, that it contains a lysine residue, and that it is in a particular location with respect to the N-terminus. However, the label itself should not be taken as a representation of these facts.

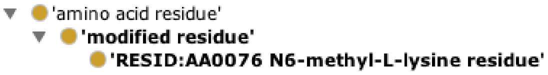
B. Residues

Recall from Table 1 and earlier discussion of PR:000036802 (histone H3.3 acetylated and methylated 2 (mouse)) that PSI-MOD IDs are currently attached to the entry for that protein, and that the PSI-MOD IDs seem to denote the *processes of modification* rather than, say, the resulting residues. MOD:00085, for instance, is defined as *A protein modification that effectively converts an L-lysine residue to N6-methyl-L-lysine*. This text definition makes reference to the residue that results from this type of modification, but MOD:00085 itself stands for the modification, not the modified residue.

We are using RESID terms for amino acid residues, as shown in Figure 3. We have defined modified residue, which is a subclass of the CHEBI[6] term amino acid residue. Subclasses of modified residue are amino acid residues that are the output of some modification process.

A MOD:00085 (process) involves the conversion of an L-lysine residue (RESID:AA0012[11]), which is located at some site of an amino acid residue in a protein, into an N6-methyl-L-lysine residue (RESID:AA0076 [12]) that is located at some site of post translationally modified amino acid residue.

These facts are currently given in PSI-MOD as part of text definitions (e.g. *A protein modification that effectively converts an L-lysine residue to N6-methyl-L-lysine* [13]) and as unstructured xref_definitions: RESID:AA0076 - the modification's output - is listed among five other xref_definitions in the PSI-MOD entry for MOD:00085.



▼ ● 'amino acid residue'
▼ ● 'modified residue'
● 'RESID:AA0076 N6-methyl-L-lysine residue'

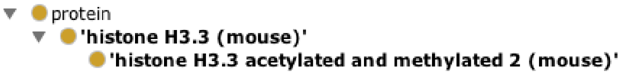
Fig. 3. Residues

Any instance of site of an amino acid residue in a protein is occupied_by some amino acid residue.

Any instance of site of post translationally modified amino acid residue is occupied_by some modified residue.

C. Modified Proteins

Figure 4 shows the term for the modified protein PR:000036802

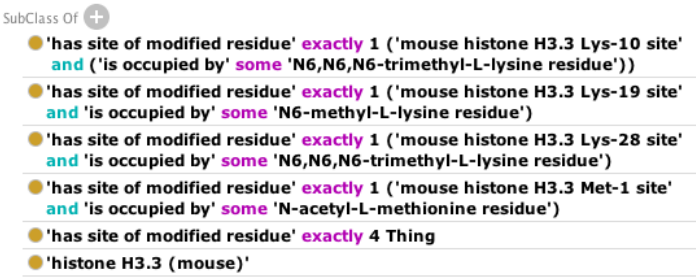


▼ ● protein
▼ ● 'histone H3.3 (mouse)'
● 'histone H3.3 acetylated and methylated 2 (mouse)'

Fig. 4. Histone H3.3 acetylated and methylated 2 (mouse)

Figure 5 shows part of the OWL definition for this term using the representation discussed above.

histone H3.3 acetylated and methylated 2 (mouse) is that particular modified protein, which is a subclass of histone (mouse) that has four sites of modified residue. One of those modified residues is a mouse histone H3.3 Lys-10 site that is occupied_by some N6,N6,N6-trimethyl-L-lysine residue [14].



SubClass Of ⊕

● 'has site of modified residue' exactly 1 ('mouse histone H3.3 Lys-10 site' and ('is occupied by' some 'N6,N6,N6-trimethyl-L-lysine residue'))

● 'has site of modified residue' exactly 1 ('mouse histone H3.3 Lys-19 site' and 'is occupied by' some 'N6-methyl-L-lysine residue')

● 'has site of modified residue' exactly 1 ('mouse histone H3.3 Lys-28 site' and 'is occupied by' some 'N6,N6,N6-trimethyl-L-lysine residue')

● 'has site of modified residue' exactly 1 ('mouse histone H3.3 Met-1 site' and 'is occupied by' some 'N-acetyl-L-methionine residue')

● 'has site of modified residue' exactly 4 Thing

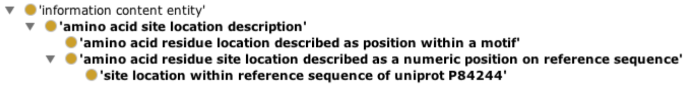
● 'histone H3.3 (mouse)'

Fig. 5. Definition of Histone H3.3 acetylated and methylated 2 (mouse)

[8] http://purl.obolibrary.org/obo/PR_000000001
[9] http://purl.obolibrary.org/obo/PR_000018263
[10] http://purl.obolibrary.org/obo/CHEBI_33708
[11] http://pir.georgetown.edu/cgi-bin/resid?id=AA0012
[12] http://pir.georgetown.edu/cgi-bin/resid?id=AA0076
[13] http://www.ebi.ac.uk/ontology-lookup/?termId=MOD:00085
[14] http://pir.georgetown.edu/cgi-bin/resid?id=AA0074

D. Position within a Reference Sequence

The position within a reference sequence of a particular amino acid residue / modification site is a key piece of information currently used to identify such sites in many different resources. We represent such descriptions as Information Artifact Ontology (IAO)[15] `information content entity`s. For instance, the Uniprot reference sequence `P84244`[16] corresponds to `histone H3.3 (mouse)`. We use the class `site location within reference sequence of uniprot P84244` for positions relative to that sequence. Members of that class are named individuals like `position 19 in reference sequence of mouse histone H3.3`, which `is about` a specific `mouse histone H3.3 site`, namely `mouse histone H3.3 Lys-19 site`.



- ▼ ● 'information content entity'
  - ▼ ● 'amino acid site location description'
    - ● 'amino acid residue location described as position within a motif'
    - ▼ ● 'amino acid residue site location described as a numeric position on reference sequence'
      - ● 'site location within reference sequence of uniprot P84244'

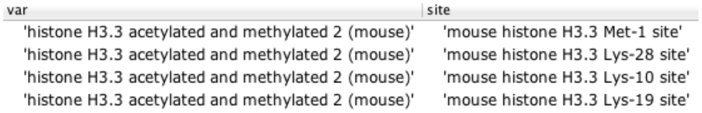
Fig. 6. Information entities about sites

V. Example Query

The following simple SPARQL query gets the subclasses of `histone H3.3 (mouse)` – `h33mouse` – and their sites that are subclasses of `site of an amino acid residue in a protein ressite:` [17] That is: it answers the question *what are the variants of* `histone H3.3 (mouse)` *and the sites of their PTMs?*

```
PREFIX h33mouse:
  <http://purl.obolibrary.org/obo/PR_P84244>
PREFIX ressite:
  <http://purl.obolibrary.org/obo/PROXXX_0001001>
SELECT ?var ?s
WHERE {
  ?var rdfs:subClassOf+ h33mouse: .
  ?var rdfs:subClassOf ?d .
  ?d owl:onClass/owl:intersectionOf/rdf:first ?s.
  ?s rdfs:subClassOf+ ressite:  .
}
```

The results of this query are shown in Figure 7. Because the ontology document queried contains only the example modified form discussed above (`histone H3.3 acetylated and methylated 2 (mouse)`), the set of results includes only its modification sites.



var	site
'histone H3.3 acetylated and methylated 2 (mouse)'	'mouse histone H3.3 Met-1 site'
'histone H3.3 acetylated and methylated 2 (mouse)'	'mouse histone H3.3 Lys-28 site'
'histone H3.3 acetylated and methylated 2 (mouse)'	'mouse histone H3.3 Lys-10 site'
'histone H3.3 acetylated and methylated 2 (mouse)'	'mouse histone H3.3 Lys-19 site'

Fig. 7. SPARL results: variants and modification sites

VI. Conclusions and Future Work

We have outlined and implemented a representation of posttranslational modifications that explicitly accounts for the biological entities involved and information about them. This includes representations of sites, residues, and modified variants as well as of information artifacts such as descriptions of sites relative to reference sequences.

This representation scheme is ready to be applied to a larger set of test data, some of which is not currently present in PRO and some of which is currently present in a less structured form. Immediate future work involves translating histone modification data from the Histone Infobase and from top-down proteomics, integrating those data with the relevant records in PRO, and creating new records where needed. Much of the work can be scripted, but some will require manual curation as well.

Other planned future work is to expand the scope of our representation of sites to include other types of modifications, mutation sites, cleavage sites, and so on. Though the basic scheme will remain the same – using terms that represent each type of site along with terms for related entities – each will present opportunities for interesting ontology work. For instance, mutations that involve the deletion of an amino acid residue raise non-trivial questions about what happens to the site when the residue that occupied it is gone.

References

[1] D. A. Natale, C. N. Arighi, W. C. Barker, J. A. Blake, C. J. Bult, M. Caudy, H. J. Drabkin, P. D'Eustachio, A. V. Evsikov, H. Huang *et al.*, "The protein ontology: a structured representation of protein forms and complexes," *Nucleic acids research*, vol. 39, no. suppl 1, pp. D539–D545, 2011.

[2] The UniProt Consortium, "The universal protein resource (uniprot)," *Nucleic Acids Research*, vol. 36, no. suppl 1, pp. D190–D195, 2008. [Online]. Available: http://nar.oxfordjournals.org/content/36/suppl_1/D190.abstract

[3] B. D. Strahl and C. D. Allis, "The language of covalent histone modifications," *Nature*, vol. 403, no. 6765, pp. 41–45, 01 2000. [Online]. Available: http://dx.doi.org/10.1038/47412

[4] P. Grenon and B. Smith, "Snap and span: Towards dynamic spatial ontology," *Spatial cognition and computation*, vol. 4, no. 1, pp. 69–104, 2004.

[5] P. Grenon, B. Smith, and L. Goldberg, "Biodynamic Ontology: Applying BFO in the Biomedical Domain," D. Pisanelli, Ed. IOS Press, 2004, vol. 102, pp. 20–38.

[6] K. Degtyarenko, P. De Matos, M. Ennis, J. Hastings, M. Zbinden, A. Mcnaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner, "Chebi: a database and ontology for chemical entities of biological interest," *Nucleic acids research*, vol. 36, no. suppl 1, pp. D344–D350, 2008.

[15] https://code.google.com/p/information-artifact-ontology/

[16] http://www.uniprot.org/uniprot/P84244

[17] The URI used here for `site of an amino acid residue in a protein` (http://purl.obolibrary.org/obo/PROXXX_0001001) is a temporary placeholder to be replaced by a real PRO ID when it is added to a PRO release