Title
ProForma: A Standard Proteoform Notation

Journal
Journal of Proteome Research, 17(3)

ISSN
1535-3893

Authors
LeDuc, Richard D
Schwämmle, Veit
Shortreed, Michael R
et al.

Publication Date
2018-03-02

DOI
10.1021/acs.jproteome.7b00851

Peer reviewed

HHS Public Access

Author manuscript

J Proteome Res. Author manuscript; available in PMC 2019 March 02.

Published in final edited form as:

J Proteome Res. 2018 March 02; 17(3): 1321–1325. doi:10.1021/acs.jproteome.7b00851.

ProForma: A Standard Proteoform Notation

Richard D. LeDuc[*,†,□,iD], **Veit Schwämmle**[‡,□], **Michael R. Shortreed**[§,□], **Anthony J. Cesnik**[§,□,iD], **Stefan K. Solntsev**[§,□,iD], **Jared B. Shaw**[‖,□,iD], **Maria J. Martin**[⊥], **Juan A. Vizcaino**[⊥], **Emanuele Alpi**[⊥,iD], **Paul Danis**[#], **Neil L. Kelleher**[†,iD], **Lloyd M. Smith**[§,▽], **Ying Ge**[§], **Jeffrey N. Agar**[○,iD], **Julia Chamot-Rooke**[◆], **Joseph A. Loo**[¶,iD], **Ljiljana Pasa-Tolic**[‖], and **Yury O. Tsybin**[+,iD]

[†]National Resource for Translational and Developmental Proteomics, Northwestern University, Evanston, Illinois 60208, United States [‡]Department of Biochemistry and Molecular Biology, University of Southern Denmark, DK-5230 Odense, Denmark [§]Department of Chemistry, University of Wisconsin, Madison, Wisconsin 53706, United States [‖]Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Washington 99354, United States [⊥]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom [#]Consortium for Top-Down Proteomics, Cambridge, Massachusetts 02142, United States [▽]Genome Center of Wisconsin, University of Wisconsin, Madison, Wisconsin 53706, United States [○]Chemistry and Chemical Biology, Northeastern University, Boston, Massachusetts 02115, United States [◆]Mass Spectrometry for Biology Unit, Institut Pasteur, CNRS USR 2000, Paris Cedex 15, France [¶]Department of Chemistry and Biochemistry and Department of Biological Chemistry, University of California, Los Angeles, Los Angeles, California 90095, United States [+]Spectroswiss, 1015 Lausanne, Switzerland

Abstract

The Consortium for Top-Down Proteomics (CTDP) proposes a standardized notation, ProForma, for writing the sequence of fully characterized proteoforms. ProForma provides a means to communicate any proteoform by writing the amino acid sequence using standard one-letter notation and specifying modifications or unidentified mass shifts within brackets following certain amino acids. The notation is unambiguous, human-readable, and can easily be parsed and written by bioinformatic tools. This system uses seven rules and supports a wide range of possible use

[*]**Corresponding Author**. Richard.leduc@northwestern.edu. Tel: +1.847.467.4362.

ORCID

Richard D. LeDuc: 0000-0002-6951-2923
Anthony J. Cesnik: 0000-0002-5326-7134
Stefan K. Solntsev: 0000-0002-1061-1476
Jared B. Shaw: 0000-0002-1130-1728
Emanuele Alpi: 0000-0003-4822-9472
Neil L. Kelleher: 0000-0002-8815-3372
Jeffrey N. Agar: 0000-0003-2645-1873
Joseph A. Loo: 0000-0001-9989-1437
Yury O. Tsybin: 0000-0001-7533-0774

□The ProForma working group was composed of R.D.L., V.S., M.R.S., A.J.C., S.K.S., and J.B.S., who contributed equally to this work. All authors have read the manuscript and support this initiative.

The authors declare the following competing financial interest(s): Some of the authors are involved in commercial software development.

cases, ensuring compatibility and reproducibility of proteoform annotations. Standardizing proteoform sequences will simplify storage, comparison, and reanalysis of proteomic studies, and the Consortium welcomes input and contributions from the research community on the continued design and maintenance of this standard.

Graphical abstract

Keywords

standard; proteoform; human readable; machine readable

INTRODUCTION

With the advent of top-down proteomics, it is increasingly possible to identify and characterize intact proteins in complex biological samples. These fully characterized proteins are known as proteoforms[1] and are defined forms of a protein with a specific set of amino acids and localized post-translational modifications (PTMs). Proteoforms are differentiated from one another by two aspects. The first is amino acid variations at known positions and includes amino acid insertions, substitutions, deletions, and alternative splicing isoforms. Such changes can lead to significant changes in the biological function of the protein.[2,3] Proteoforms may also be differentiated from one another by variations in the positioning and types of PTMs. These chemical changes play key roles in cell signaling[4] and other cellular functions, making the analysis of PTMs and PTM localizations critical for understanding biological systems.

Exchanging protein information is a common issue in all subfields of proteomics. Fortunately, exchanging unmodified protein sequences is a remarkably simple task using the IUPAC one-letter notations for amino acids.[5] For example, the FASTA format allows exchanging a protein sequence along with unstructured metadata in the header. More detailed information, such as localized PTMs and sequence variations, can be exchanged using a variety of file types (e.g., VCF[6] for DNA sequence information or the UniProt[7] XML formats). Recently, there has been interest in standardizing the description of proteoforms. One such strategy, the Protein Ontology (PRO) approach,[8] uses a single protein database (e.g., UniProt) protein accession identifiers as the foundation for describing sequence variations and PTMs. However, there has been no standardized notation for writing fully characterized proteoform sequences (Figure 1A), and we believe that establishing a standard that builds upon the simplicity and flexibility of the IUPAC one-letter notation will have a positive impact on the field.

Having a common notation promotes reproducibility and compatibility between bioinformatic tools and promotes clear understanding and interpretation. To be successful, the notation should meet five requirements: (1) it should provide an unambiguous description of the proteoform; (2) it should be human readable, that is, it should be suitable for display in written document or presentation; (3) it should be machine parsable; (4) it should contain the complete amino acid sequence of the observed proteoform; and (5) it should specify the location and type of each modification.

The Consortium for Top-Down Proteomics is a nonprofit organization that promotes the field of top-down proteomics. (More information on the CTDP can be found at http://www.topdownproteomics.org.) The Executive Committee of the CTDP formed a working group charged with developing a standard notation for exchanging proteoform information. Presented here are the results of this effort.

METHODS

The working group met weekly via conference calls and shared ideas over several months in late 2016 and early 2017. A draft of the ProForma notation was completed and socialized via GitHub. This proposal was then presented to the attendees of the 2017 ASMS Workshop on Top-Down Proteomics and the 2017 EuBIC Winter School.[11] In all cases, the public was encouraged to contribute comments and suggestions for improving the notation.

We recognize that this notation is neither perfect nor final, and we hope all interested researchers will contribute to this project. As with other standards, changes will be needed to accommodate changing technology and interests. Therefore, the notation is versioned, with subsequent versions replacing or expanding the notation as required. Version 1.0 is announced here. Future versions will be released from by the CTDP Executive Committee as needed and will be available online at https://topdownproteomics.github.io/ProteoformNomenclatureStandard/. Proposals for changes or new features can be requested via the GitHub framework or by contacting one of the authors.

RESULTS

ProForma Notation

The notation standard consists of a series of rules for writing (using ASCII characters) a proteoform sequence including modifications. The base amino acid sequence of the proteoform should be the observed sequence or represent a hypothesized sequence; this strategy intrinsically represents sequence variations. In the case of experimental proteoform observations, amino acids that were not observed are excluded from the proteoform sequence. For example, N-terminal methionine (M) cleavage is simply noted by omitting the terminal methionine in the sequence.

Protein repository accessions, such as UniProt or Refseq accessions, are explicitly avoided in lieu of providing the complete amino acid sequence of the proteoform. This is because the complete amino acid sequence is fully portable, with no need of reliance on outside organizations or databases to provide the necessary sequence details. We consider linking

proteoforms to protein accessions as an option and not a requirement. The notation is based on seven rules, which are outlined in the following text.

> **Rule 1:** The base sequence of the proteoform is written using the IUPAC capitalized single-character amino acid codes.[5] Selenocysteine is assigned to the character U, and pyrrolysine is assigned to the character O in updated standards.[9,10] Ambiguous characters, such as J, B, and Z, may be used. According to the standard: "B is assigned to aspartic acid or asparagine when these have not been distinguished. Z is assigned to glutamic acid or glutamine." ProForma is intended for writing fully characterized proteoforms, and so X is forbidden because according to the standard, it "means that the identity of an amino acid is undetermined, or that the amino acid is atypical."
>
> **Rule 2:** Tags denoted by square brackets are used to signal information regarding a modification. These tags are placed after the character representing the modified amino acid. Multiple modifications of the same amino acid are described by successive square bracket pairs.
>
> **Rule 3:** Tags contain descriptors that take the form of key–value pairs, where the key and value are separated by colons. The key indicates the type of the descriptor. To simplify the notation in several common use cases, descriptors may have implied keys that do not need to be written out, as described in Rules 5 and 6.
>
> **Rule 4:** Multiple descriptors can be placed in a single tag, provided they are separated by pipe symbols.
>
> **Rule 5:** Five types of keys are supported by the notation standard: Modification Name, Database Accession, Mass, Chemical Formula, and Additional Information. The use of each is detailed below. Some descriptors do not require a key, usually in cases where it improves readability. A key must be present in a descriptor if it is classified as mandatory, but an optional key may be omitted.

A. Modification Name—Several commonly used sources of protein modification names, such as existing controlled vocabularies or ontologies, can be used to specify modifications: Unimod,[12] UniProt,[7] RESID,[13] PSI-MOD,[14] and BRNO.[15] Modification names must come from specific fields from these databases: Unimod – Interim Name; UniProt – ID; RESID – Name; PSI-MOD – Short label; or BRNO notation. (This small set of symbols is commonly used for histone PTMs, e.g., ph, me1, ac.) In contrast with the other descriptors, the key for this type of descriptor, "mod", is optional. If it is not used, then the standard assumes Unimod Interim Names are used (http://www.unimod.org/modifications_list.php). When specifying a modification using a database other than the Unimod, the database name must be provided in parentheses following the modification name (See Figure 1B). Placing the database name after the modification name improves human readability.

B. Database Accession—Modification databases contain unique identifiers for each modification. These accessions can be used to specify modifications in proteoform sequences. This type of tag consists of an accession following the database name as a key:

Unimod, UniProt, RESID, PSI-MOD, UniCarbKB, and the PRO Ontology. The current CTDP recommendations and Web sites for these databases are in Table 1.

C. Mass—Mass differences are often characteristic of specific modifications. However, experiments are increasingly capable of revealing unidentified mass shifts.[16–18] These unidentified mass shifts can be specified in Daltons following the mandatory key "mass". Any precision may be used for these specifications (see Figure 1B). A positive mass shift can be specified either with a plus sign or without a sign. Negative shifts must be specified with a negative sign. The mass shift is assumed to be observed, neutral, and monoisotopic unless there is an "info" tag (below) explaining otherwise.

D. Chemical Formula—Chemical formulas of modifications may be specified following the mandatory key "formula". Formulas must use Unimod symbols (http://www.unimod.org/masses.html) and follow the Unimod composition rules (http://www.unimod.org/fields.html). The formula is displayed as a string of atomic symbols in any order (C, F, H, etc. are here symbols for elements within this descriptor, not one-letter codes for amino acids), and each symbol is optionally followed by the count of that atom in parentheses. The number of atoms may be negative, and if no number is specified, then the number of atoms is assumed to be 1. Isotopes are specified by the nucleon number preceding the atomic symbol (e.g., ^{13}C).

E. Additional Information—All other information can be specified using unstructured text following the mandatory key "info". The added text may not contain the pipe character. We expect this tag will commonly be used for the development of new descriptors. It is included to allow the maximum utility of this system.

> **Rule 6:** To simplify sequences that use many tags with the same key, sequences may be prefixed with a single key followed by a plus sign (see Figure 1B). This prefixed key defines every tag in the sequence. This option can only be used when there is one key in the sequence.
>
> **Rule 7:** Proteoforms may contain N- and C-terminal modifications. These modifications are specified with a tag describing the terminal modification, separated from the sequence by a dash to the left of the N-terminal amino acid or a dash to the right of the C-terminal amino acid.

Definitions

Important terms for this standard are defined in Table 2.

Best Practices

It is possible to write proteoforms following the above rules that are not easily human readable. Rather than creating rules that force sequences to be human readable, at the expense of machine parsing, best practices were adopted. These practices are not required within the ProForma standard but rather are encouraged when possible. Particular emphasis should be placed on human readability when using this notation in scientific publications.

Figure 1C provides several examples of best practices and one sequence that is problematic for human readability.

Several recommendations for writing clear sequences are as follows:

1. In a pipe-separated list, the most descriptive element should be placed first to improve human readability. Consequently, if the identity of a modification is known, it should be listed first (preferably Unimod interim names without the "mod" key). This improves the clarity over listing only masses or accessions. Example (i) of Figure 1C demonstrates this principle, with the placement of modification names before mass tags.
2. Prefix tags (see rule 6) should be used when there is only one element in the tag. The recommended use of these prefix tags is shown in example (ii) of Figure 1C. Otherwise, human readability is compromised: In the following example, the descriptors "1" and "21" inherit the Unimod key from the prefix tag, but they lack the clarity of the other key-value pairs and could cause confusion for a reader: [Unimod]+SGRGK-[mod:Acetyl|1|mass:42.010565] QGGKARGAVLLPKKT[21]-ESHHKAKGK.
3. Spacing before and after each descriptor is arbitrary and should be appropriately added to improve readability. Example (iii) of Figure 1C demonstrates this principle.
4. Unknown modifications are best described by their mass shifts and marked as unknowns, as displayed in example (iv) of Figure 1C.

DISCUSSION

This work presents a short set of rules named ProForma for researchers to write fully characterized proteoform sequences in an unambiguous manner, either by hand or through bioinformatic solutions. Proteoforms written in this way can be read by humans and parsed by software, thus simplifying the storage, retrieval, and comparisons of proteoforms revealed in proteomic studies.

The ProForma project arose from researchers at several laboratories who collectively recognize the need for this notation. The working group was careful to create a standard that is generalizable because a multitude of solutions could be presented to address this need. However, it may not address every need of the top-down proteomics community and the proteomics community in general. One such example is the need to specify modifications with ambiguous localizations. This need and others will be resolved in what we hope to be vibrant discussion on the ProForma Web site (https://topdownproteomics.github.io/ProteoformNomenclatureStandard/). In addition, researchers who find this standard does not meet the needs of specific bioinformatic tools are encouraged to provide such information. These comments and suggested changes will be considered for future versions of the standard.

Acknowledgments

A.J.C was supported by the Computation and Informatics in Biology and Medicine Training Program, T15LM007359. R.D.L. was supported in part by National Institute for General Medical Sciences under award P41 GM108569. V.S. acknowledges support from the EuBIC initiative, ELIXIR Denmark and the Danish Research Council. M.R.S., A.J.C., S.K.S., and L.M.S. acknowledge support of the National Institute of General Medical Sciences grant R01GM114292. J.A.V. acknowledges funding from ELIXIR and from EMBL core funds.

References

1. Smith LM, Kelleher NL. Proteoform: a single term describing protein complexity. Nat. Methods. 2013; 10(3):186–187. [PubMed: 23443629]
2. Pauling L, Itano Ha, Singer SJ, Wells IC. Sickle cell anemia, a molecular disease. Science. 1949; 110(2865):543. [PubMed: 15395398]
3. Yang X, Coulombe-Huntington J, Kang S, Sheynkman GM, Hao T, Richardson A, Sun S, Yang F, Shen YA, Murray RR, et al. Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. Cell. 2016; 164(4):805–817. [PubMed: 26871637]
4. Whitmarsh AJ, Davis RJ. Multisite phosphorylation by MAPK. Science. 2016; 354(6309):179–180. [PubMed: 27738159]
5. IUPAC-IUB Commission on Biochemical Nomenclature. A One-Letter Notation for Amino Acid Sequence (Definitive Rules). Pure Appl. Chem. 1972; 31(4):151–153.
6. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27(15): 2156–2158. [PubMed: 21653522]
7. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017; 45(D1):D158–D169. [PubMed: 27899622]
8. Natale DA, Arighi CN, Blake JA, Bona J, Chen C, Chen SC, Christie KR, Cowart J, D'Eustachio P, Diehl AD, et al. Protein Ontology (PRO): Enhancing and scaling up the representation of protein entities. Nucleic Acids Res. 2017; 45(D1):D339–D346. [PubMed: 27899649]
9. Liébecq C. IUPAC-IUBMB Joint Commission on Biochemical Nomenclature (JCBN) and Nomenclature Committee of IUBMB (NC-IUBMB). Biochem Mol. Biol. Int. 1997; 43(5):1151–1156. [PubMed: 9415825]
10. Cammack, R. Newsletter, 2009. Biochemical Nomenclature Committee of IUPAC and NC-IUBMB; 2009. http://www.sbcs.qmul.ac.uk/iubmb/newsletter/2009.html
11. Willems S, Bouyssié D, David M, Locard-Paulet M, Mechtler K, Schwämmle V, Uszkoreit J, Vaudel M, Dorfer V. Proceedings of the EuBIC Winter School 2017. J. Proteomics. 2017; 161:78–80. [PubMed: 28385664]
12. Creasy DM, Cottrell JS. Unimod: Protein modifications for mass spectrometry. Proteomics. 2004; 4(6):1534–1536. [PubMed: 15174123]
13. Garavelli JS. The RESID Database of Protein Modifications as a resource and annotation tool. Proteomics. 2004; 4(6):1527–1533. [PubMed: 15174122]
14. Montecchi-Palazzi L, Beavis R, Binz P-A, Chalkley RJ, Cottrell J, Creasy D, Shofstahl J, Seymour SL, Garavelli JS. The PSI-MOD community standard for representation of protein modification data. Nat. Biotechnol. 2008; 26(8):864–866. [PubMed: 18688235]
15. Turner BM. Reading signals on the nucleosome with a new nomenclature for modified histones. Nat. Struct. Mol. Biol. 2005; 12(2):110–112. [PubMed: 15702071]
16. Chick JM, Kolippakkam D, Nusinow DP, Zhai B, Rad R, Huttlin EL, Gygi SP. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. Nat. Biotechnol. 2015; 33(7):743–749. [PubMed: 26076430]
17. Li Q, Shortreed MR, Wenger CD, Frey BL, Schaffer LV, Scalf M, Smith LM. Global Post-Translational Modification Discovery. J. Proteome Res. 2017; 16(4):1383–1390. [PubMed: 28248113]

18. Shortreed MR, Frey BL, Scalf M, Knoener RA, Cesnik AJ, Smith LM. Elucidating Proteoform Families from Proteoform Intact Mass and Lysine Count Measurements. J. Proteome Res. 2016; 15:1213–1221. [PubMed: 26941048]

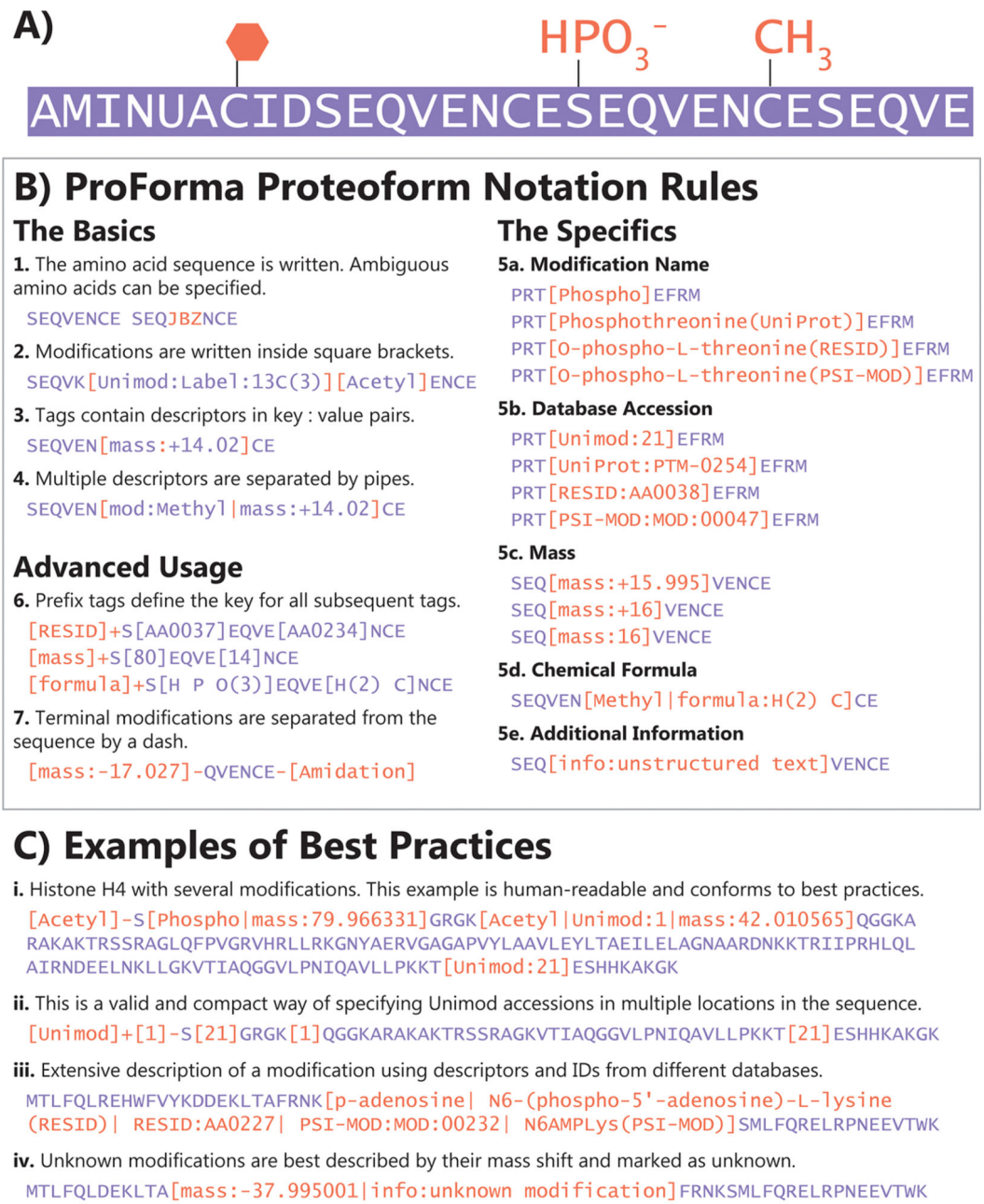

Figure 1.

Proteoform notation introduction, rules, and examples. (A) Proteoforms are composed of specific amino acid sequences with modifications at known positions along the sequence. This work presents a standard proteoform notation for writing these sequences in a flexible, human-readable way. (B) Brief examples for the seven current rules for specifying proteoform sequences. (C) Examples and explanations of best practices for writing human-readable proteoform sequences.

Table 1

Currently Supported Modification Databases

database name	CTDP recommendation	URL
Unimod	default	http://www.unimod.org/modifications_list.php
UniProt	recommended	https://www.uniprot.org/docs/ptmlist
RESID	recommended	http://pir.georgetown.edu/resid/resid.shtml
PSI-MOD	recommended	http://www.ebi.ac.uk/ols/ontologies/mod
UniCarbKB	acceptable	http://www.unicarbkb.org/
PRO Ontology/NCBI	acceptable	http://pir.georgetown.edu/pro/

Table 2

Terms Defined for This Standard

term	definition
descriptor	Member of the tag. Could be a key-value pair or a keyless entry.
human readable	A strong emphasis is placed on human readability for proteoform names. Proteoforms should be named in a manner that allows general audience members to know exactly the sequence of amino acids and the positions of any modifications, described in as accurate detail as possible.
key	An optional element of a descriptor that specifies the descriptor type. It must be followed by a colon and a value.
machine readable	Adherence to the conventions described above should facilitate the creation and utilization of generic parsers so that proteoforms can be exchanged between users using a computer interface.
modification	Includes the addition and subtraction of specific atoms, atom combinations, and/or masses at a specific residue in a proteoform.
tag	The specified way of writing a localized modification. Everything between “[” and “]” (inclusive). A collection of descriptors.
value	Contents of a descriptor, such as the mass, chemical composition, or modification name.