Published in final edited form as:

Nat Methods. 2019 October ; 16(10): 939–940. doi:10.1038/s41592-019-0573-x.

A five-level classification system for proteoform identifications

Lloyd M. Smith^{1,*}, Paul M. Thomas², Michael R. Shortreed¹, Leah V. Schaffer¹, Ryan T. Fellers², Richard D. LeDuc², Trisha Tucholski¹, Ying Ge³, Jeffrey N. Agar⁴, Lissa C. Anderson⁵, Julia Chamot-Rooke⁶, Joseph Gault⁷, Joseph A. Loo⁸, Ljiljana Paša-Toli⁹, Carol V. Robinson⁷, Hartmut Schlüter¹⁰, Yury O. Tsybin¹¹, Marta Vilaseca¹², Juan Antonio Vizcaíno¹³, Paul O. Danis¹⁴, Neil L. Kelleher^{2,*}

¹Department of Chemistry, University of Wisconsin-Madison, Madison, WI 53706 USA

²Departments of Chemistry, Molecular Biosciences and the National Resource for Translational and Developmental Proteomics, Northwestern University, Evanston, IL 60208 USA ³Department of Cell and Regenerative Biology and Human Proteomics Program, University of Wisconsin-Madison, Madison WI 53706 USA ⁴Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA 02111 USA ⁵Ion Cyclotron Resonance Program, National High Magnetic Field Laboratory, Tallahassee FL 32310 USA ⁶Mass Spectrometry for Biology Unit, Institut Pasteur, Paris, France ⁷Department of Chemistry, University of Oxford, Oxford, UK ⁸Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90095 USA

⁹Environmental Molecular Sciences Laboratory and Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA 99354 USA ¹⁰University Medical Center, Hamburg-Eppendorf, Hamburg, Germany ¹¹Spectroswiss CH, 1015 Lausanne, Switzerland ¹²Institute for Research in Biomedicine, IRB, Barcelona, Spain ¹³European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK ¹⁴Consortium for Top Down Proteomics, Cambridge, MA 02142 USA

The term proteoform, introduced in these pages in 2013¹, has rapidly gained acceptance in the proteomics community. The challenge and importance of comprehensively identifying proteoforms in complex samples has been recognized, and reports have begun to appear of new platforms towards that end^{2–5}. However, one interesting central ambiguity has emerged, namely determining precisely what is meant by a “proteoform identification”. At present, the only practical approaches for establishing the exact primary structure of a proteoform employ mass spectrometry (MS), and there is a wide range of MS results that provide “proteoform identifications⁶”. This seemingly small matter has significant impact, as the ambiguity in knowing what is meant by an “identification” makes it difficult to compare results from different laboratories and approaches. This situation hinders the ability of the community to evaluate technological progress and to efficiently expand biological knowledge.

*Lloyd Smith, Department of Chemistry, University of Wisconsin-Madison, 1101 University Avenue, Madison, WI 53706, smith@chem.wisc.edu, Phone: (608) 263-2594, Neil Kelleher, Departments of Chemistry and Molecular Biosciences and the National Resource for Translational and Developmental Proteomics, Northwestern University, 2170 Campus Drive, Evanston, IL 60208, n-kelleher@northwestern.edu, Phone: (847) 467-4362.

Competing Financial Interests

The authors declare no competing financial interests.

To address this issue, we propose a five-level system for classifying proteoform identifications. The classification scheme stems directly from a consideration of the four types of possible ambiguity possible for a proteoform identification, ranging from the most subtle (*i.e.*, precise localization of a post-translational modification, or PTM) to the most dramatic (*i.e.*, ambiguity in the gene of origin). The five classes then correspond to the level of ambiguity present in the identification, ranging from no ambiguity at all (Level 1), to ambiguity of all four types (Level 5). Details of the scheme are provided in Table 1 and Supplementary Table 1, with specific use cases and examples provided in Supplementary Figure 1.

Proteoform Level Classification System

The Proteoform Level Classification System is intended to assist researchers in representing the ambiguity within their proteoform identifications. Four classes of ambiguity are recognized as occurring in proteoform identifications:

1. **PTM Localization:** A PTM is not localized to a specific amino acid.
2. **PTM Identification:** There is ambiguity in the identity of a PTM.
3. **Amino Acid Sequence:** There is ambiguity in the amino acid sequence.
4. **Gene:** The gene of origin is unknown or ambiguous.

Level 1:

The proteoform is identified with no ambiguity and full knowledge of its gene of origin, its complete amino sequence defined, and the identities and locations of all PTMs, if present.

Level 2:

The proteoform is identified with ambiguity in just one of the classes of ambiguity described above. Further delineation of this level is shown in Table 1. Examples of this include: 2A) The amino acid sequence is completely defined with knowledge of its gene of origin and all PTMs are fully identified, but their localization is incomplete. 2B) The amino acid sequence is completely defined with knowledge of its gene of origin and the localization of PTMs is complete, but the PTM(s) are not fully identified or structurally characterized (*e.g.*, acetylation vs. trimethylation or glycoproteoforms). 2C) All PTMs are identified and localized, if present, but there exists some localized sequence ambiguity (*e.g.* the order of amino acids is unknown in a small region), yet there is still knowledge of its gene of origin. 2D) The amino acid sequence is completely defined and all PTMs are fully identified and localized, however there is ambiguity with respect to the gene of origin.

Level 3:

The proteoform is identified with ambiguity in two of the classes described above. (See Supplementary Table 1 for definitions of sub-levels.)

Level 4:

The proteoform is identified with ambiguity in three of the classes described above. (See Supplementary Table 1 for definitions of sub-levels.)

Level 5:

Insufficient information has been obtained to know from which gene the proteoform is derived, its sequence, PTMs or their localization. Only the observed mass of the proteoform is known.

We distinguish here between the definitional issue of what level of proteoform identification is being claimed, and the related issue of statistical confidence, which provides a measure of the quality of the claimed identification. The system presented here permits proteoform identifications to be classified into various types but intentionally does not address the related issue of the confidence with which such identifications have been asserted by each investigator. Ideally, every proteoform identification would be accompanied by both a classification level and a confidence metric. Early efforts to estimate proteoform characterization confidence include the C-score⁷ and the MIScore⁸ but further work is needed to develop and refine estimates such that proteoform levels can be confidently and automatically assigned.

The utility of this five-level system for classifying proteoform identifications lies in its ability to greatly reduce ambiguity in the reporting of proteoform identification results. We will use this system in future publications, breaking down proteoform identifications made into these five categories, and hope others will adopt the system as well. An important practical issue in implementation of the scheme is the development of informatic tools that assign and report these classifications automatically. We believe the resultant uniformity will help researchers publish findings of greater clarity, compare and evaluate results using different methodologies, and drive efficient progress in the field. Together with integration of different types of results (such as peptide-level data from bottom-up proteomics) and prior efforts to regularize proteoform sharing⁹, a growing infrastructure to build up proteoform databases is developing¹⁰.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by grants R35 GM126914 (LMS), and P41 GM108569 (NLK) from the NIH National Institute of General Medical Sciences, R21 LM013097 (PMT) from the National Library of Medicine and the Sherman Fairchild Foundation (NLK).

References

1. Smith LM & Kelleher NL Proteoform: a single term describing protein complexity. *Nature methods* 10, 186–187 (2013). [PubMed: 23443629]

2. Anderson LC et al. Identification and Characterization of Human Proteoforms by Top-Down LC-21 Tesla FT-ICR Mass Spectrometry. *Journal of proteome research* 16, 1087–1096 (2017). [PubMed: 27936753]
3. Lubeckyj RA et al. Single-Shot Top-Down Proteomics with Capillary Zone Electrophoresis-Electrospray Ionization-Tandem Mass Spectrometry for Identification of Nearly 600 *Escherichia coli* Proteoforms. *Analytical chemistry* 89, 12059–12067 (2017). [PubMed: 29064224]
4. Schaffer LV et al. Identification and Quantification of Murine Mitochondrial Proteoforms Using an Integrated Top-Down and Intact-Mass Strategy. *Journal of proteome research* 17, 3526–3536 (2018). [PubMed: 30180576]
5. Bush DR, Zang L, Belov AM, Ivanov AR & Karger BL High Resolution CZE-MS Quantitative Characterization of Intact Biopharmaceutical Proteins: Proteoforms of Interferon-beta1. *Analytical chemistry* 88, 1138–1146 (2016). [PubMed: 26641950]
6. Lermyte F, Tsybin YO, O'Connor PB & Loo JA Top or Middle? Up or Down? Toward a Standard Lexicon for Protein Top-Down and Allied Mass Spectrometry Approaches. *J Am Soc Mass Spectrom* (2019).
7. LeDuc RD et al. The C-score: a Bayesian framework to sharply improve proteoform scoring in high-throughput top down proteomics. *Journal of proteome research* 13, 3231–3240 (2014). [PubMed: 24922115]
8. Kou Q et al. Characterization of Proteoforms with Unknown Post-translational Modifications Using the MIScore. *Journal of proteome research* 15, 2422–2432 (2016). [PubMed: 27291504]
9. LeDuc RD et al. ProForma: A Standard Proteoform Notation. *Journal of proteome research* 17, 1321–1325 (2018). [PubMed: 29397739]
10. Natale DA et al. Protein Ontology (PRO): enhancing and scaling up the representation of protein entities. *Nucleic Acids Res* 45, D339–D346 (2017). [PubMed: 27899649]

Table 1.

Proteoform Level Classification System.

Level	PTMs Localized	PTMs Identified	Sequence Defined	Gene Identified
1	Yes	Yes	Yes	Yes
2A	No	Yes	Yes	Yes
2B	Yes	No	Yes	Yes
2C	Yes	Yes	No	Yes
2D	Yes	Yes	Yes	No
3	Two are certain and two are ambiguous			
4	One is certain and three are ambiguous			
5	All are ambiguous			