Proteomics Standards Initiative's ProForma 2.0: Unifying the Encoding of Proteoforms and Peptidoforms

Richard D. LeDuc,

National Resource for Translational and Developmental Proteomics, Northwestern University, Evanston, Illinois 60611, United States; Present Address: Children's Hospital Research Institute of Manitoba, John Buhler Research Centre, 715 McDermot Avenue, Winnipeg, Manitoba R3E 3P4, Canada

Eric W. Deutsch,

Institute for Systems Biology, Seattle, Washington 98109, United States

Pierre-Alain Binz,

Clinical Chemistry Service, Lausanne University Hospital, 1011 Lausanne, Switzerland

Ryan T. Fellers,

National Resource for Translational and Developmental Proteomics, Northwestern University, Evanston, Illinois 60611, United States

Anthony J. Cesnik,

Department of Genetics, Stanford University, Stanford, California 94305, United States; Chan Zuckerberg Biohub, San Francisco, California 94158, United States; SciLifeLab, School of Engineering Sciences in Chemistry Biotechnology and Health, KTH—Royal Institute of Technology, Solna, Stockholm, Sweden 113 51

Joshua A. Klein,

Program for Bioinformatics, Boston University, Boston, Massachusetts 02215, United States

Tim Van Den Bossche,

VIB—UGent Center for Medical Biotechnology, 9052 Ghent, Belgium; Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University, 9000 Ghent, Belgium

Ralf Gabriels,

VIB—UGent Center for Medical Biotechnology, 9052 Ghent, Belgium; Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University, 9000 Ghent, Belgium

Arshika Yalavarthi,

Corresponding Authors RLeduc@chrim.ca, juan@ebi.ac.uk.

National Resource for Translational and Developmental Proteomics, Northwestern University, Evanston, Illinois 60611, United States

Yasset Perez-Riverol,

European Molecular Biology Laboratory, EMBL-European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge CB10 1SD, United Kingdom

Jeremy Carver,

Center for Computational Mass Spectrometry, Department of Computer Science and Engineering, and Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego (UCSD), La Jolla, California 92093, United States

Wout Bittremieux,

Center for Computational Mass Spectrometry and Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego (UCSD), La Jolla, California 92093, United States

Shin Kawano,

Toyama University of International Studies, Toyama, 930-1292 Toyama, Japan; Database Center for Life Science, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, Kashiwa, Chiba 277-0871, Japan

Benjamin Pullman,

Center for Computational Mass Spectrometry and Department of Computer Science and Engineering, University of California, San Diego (UCSD), La Jolla, California 92093, United States

Nuno Bandeira,

Center for Computational Mass Spectrometry, Department of Computer Science and Engineering, and Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego (UCSD), La Jolla, California 92093, United States

Neil L. Kelleher,

National Resource for Translational and Developmental Proteomics, Northwestern University, Evanston, Illinois 60611, United States

Paul M. Thomas,

National Resource for Translational and Developmental Proteomics, Northwestern University, Evanston, Illinois 60611, United States; Present Address: AbbVie, Inc., 1401 Sheridan Road, North Chicago, IL 60064.
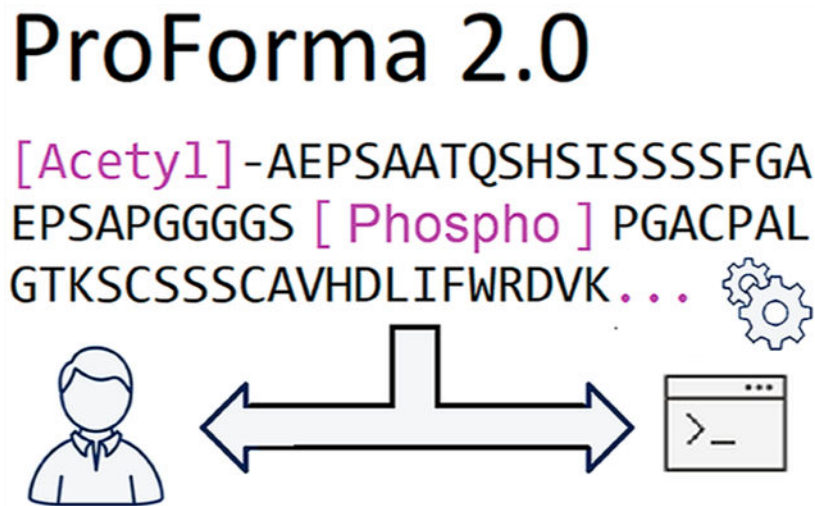
Juan Antonio Vizcaíno

European Molecular Biology Laboratory, EMBL-European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge CB10 1SD, United Kingdom

Abstract

It is important for the proteomics community to have a standardized manner to represent all possible variations of a protein or peptide primary sequence, including natural, chemically induced, and artifactual modifications. The Human Proteome Organization Proteomics Standards Initiative in collaboration with several members of the Consortium for Top-Down Proteomics

(CTDP) has developed a standard notation called ProForma 2.0, which is a substantial extension of the original ProForma notation developed by the CTDP. ProForma 2.0 aims to unify the representation of proteoforms and peptidoforms. ProForma 2.0 supports use cases needed for bottom-up and middle-/top-down proteomics approaches and allows the encoding of highly modified proteins and peptides using a human- and machinereadable string. ProForma 2.0 can be used to represent protein modifications in a specified or ambiguous location, designated by mass shifts, chemical formulas, or controlled vocabulary terms, including cross-links (natural and chemical) and atomic isotopes. Notational conventions are based on public controlled vocabularies and ontologies. The most up-to-date full specification document and information about software implementations are available at http://psidev.info/proforma.

Graphical Abstract

ProForma 2.0

[Acetyl]-AEPSAATQSHSISSSSFGA
EPSAPGGGGS [Phospho] PGACPAL
GTKSCSSSCAVHDLIFWRDVK...

Keywords

ProForma; proteoform; peptidoform; top-down proteomics; file formats; data standards; mass spectrometry; FAIR

INTRODUCTION

Protein and peptide sequences are usually represented by a string of amino acids using the well-known one-letter code that was first introduced by the International Union of Pure and Applied Chemistry (IUPAC) in 1972.[1] The linear arrangement of the amino acids is customarily written from the *N*-terminus to the *C*-terminus. However, there is no clear consensus about how to represent amino acid modifications, which can be natural [e.g., biologically relevant post-translational modifications (PTMs)], chemically induced (including, for example, reduction/alkylation and addition of tags for quantitative analysis), or artifactual as a consequence of sample preparation (such as oxidation and deamidation).

The terms "proteoform"[2] and "peptidoform"[3] are used for the specific "form" or "entity" of a given protein or peptide that results from the combination of the amino acid sequence and

modification(s) at specific amino acid positions. Multiple proteoforms can be derived from the same gene. For example, if a protein has two sites that can potentially be phosphorylated, there are four possible proteoforms: the unmodified form represented by the primary sequence and the forms with phosphorylation on the first site, the second site, and both sites. Each of these is a distinct proteoform, but only the first proteoform, the unmodified variant, can be written using the IUPAC notation. In the absence of a recognized standard notation, there is no consistency in the way modified proteins and peptides are designated. Not only can this lead to confusion in scientific publications and presentations, but it is also a major dilemma for developers of proteomics software and resources to decide what notation(s) to use for data input and output. This is applicable to widely used protein-centric database resources such as UniProtKB (UniProt Knowledge-Base),[4] ProteomeXchange proteomics resources,[5] the Protein Data Bank (PDB),[6] Reactome,[7] and IntAct,[8] among many others. This has led to the development of multiple different notational formats by various groups.

In order to make peptide and protein data more *findable, accessible, interoperable*, and *reusable* (FAIR),[9] there needs to be a single IUPAC-compatible notational standard to encode modified protein and peptide sequences. In 2018, the Consortium for Top-Down Proteomics (CTDP) introduced the ProForma notation,[10] which answered the immediate needs of the Consortium by creating a standardized method for designating a proteoform. It contained seven rules to denote both the primary structure of a proteoform and most of the commonly observed PTMs and artifactual modifications, using nomenclature from five ontologies and controlled vocabularies (CVs). In general, CVs are minimally structured lists of terms and definitions, while ontologies encode the full hierarchical relationship structure among the terms.[11]

However, this notational system was not sufficient to meet the needs of the broader proteomics community and protein data resources because some important use cases were not supported. In particular, the first ProForma version did not address issues such as ambiguity in either the order of the amino acid sequence or the modification site localization and did not support cross-links (natural or chemically induced), among many others. For proteoform and peptidoform designations to be FAIR across the broader array of protein science data resources, these and numerous other notational issues needed to be addressed. Ideally, the same notational system should be usable for both bottom-up and middle-/topdown applications.

The Proteomics Standards Initiative (PSI) of the Human Proteome Organization (HUPO) develops and ratifies community-based data standards and CVs for the field of proteomics,[12] including mzML,[13] mzIdentML,[14] mzTab,[15] PSI-MOD,[16] PEFF (PSI Extended FASTA Format),[17] and more recently the Universal Spectrum Identifier (USI)[18] and the sample metadata standard MAGE-TAB-Proteomics.[19] Each of these standards has been subjected to the PSI Document Process[20] which mandates three levels of review that must be completed before a proposed standard is ratified. In order to address the use cases needed for bottom-up and middle-/topdown approaches, members of the CTDP and HUPO-PSI worked together and devised an extended ProForma notation designed to meet the current and future needs for protein sequence data. In this article, we present an overview of the ProForma 2.0 notation, a brief description of its most salient features, and some example applications.

METHODS

Development of ProForma 2.0

The development of ProForma 2.0 started in 2019. Since then, it was an open process via conference calls in addition to discussions at the annual PSI meetings and smaller workshops. The ProForma 2.0 specification document was submitted to the PSI Document Process for review, during which time external reviewers provided their feedback. The document was also made available for comments by the public, enabling broad input on the specifications. The final version of the ProForma 2.0 specification document is provided in the Supporting Information. Potential corrections to the document, up-to-date information on software implementations, and information on future versions of ProForma are available at http://psidev.info/proforma.

The main requirements considered during the development of the standard notation were the following:

1. It must be a string of characters that is human-readable; i.e., it should be suitable for display in a written document or in a presentation.

2. It must be unambiguously parsable by software (i.e., machine-parsable).

3. It must be able to support the encoding of amino acid sequences and their modifications (including natural, chemically induced, and artifactual).

4. It must be able to support the main use cases needed by the proteomics community as a whole, including bottom-up (focused on peptides/peptidoforms) and middle-/top-down (focused on proteins/proteoforms) applications.

5. It must be flexible enough to accommodate different styles of notations that are currently in common use.

6. It must be compatible with other existing PSI file formats.

7. It must be able to accommodate ambiguity in the position of a modified site.

8. It must be able to evolve so that new use cases can be added in the future.

Requirements 1–3 were included in the original ProForma 1.0 notation.[10] The essence of the fourth requirement was in the ProForma 1.0 notation, but the current version now includes support for bottom-up proteomics-specific entities, i.e., for peptidoforms, whereas the original exclusively defined the way to designate whole proteoform sequences. Requirements 4–8 are new in ProForma 2.0.

An essential requirement of ProForma 2.0 is that it should be able to represent peptidoforms and proteoforms in a consistent and reproducible way, taking into consideration the different strategies for designating protein modifications. Moreover, it must be able to be used jointly with USIs[18] to represent peptide spectrum matches (PSMs) and proteoform spectrum matches (PrSMs).

RESULTS

Data Format Description

Here we provide a brief overview with examples of the main features of ProForma 2.0, while the full ProForma 2.0 specification document, as ratified by the PSI, provides exhaustive details on all aspects of the data format. ProForma 2.0 provides a standardized set of rules for describing the location and nature of all mass modifications on a proteoform or peptidoform. An example is shown in Figure 1. Using ProForma 2.0, there is a string of characters that linearly represents the peptidoform/proteoform primary structure, with allowance for some level of ambiguity and the possibility to link peptide chains together, such as by cross-linking. ProForma 2.0 is not intended to represent secondary or higher-order structures. ProForma 2.0 can also be used to represent the molecular interpretation of a tandem mass spectrum. It should be noted that ProForma 2.0 is designed to describe a single, specific peptidoform or proteoform and not a collection of protein sequences or a listing of all potential mass modifications that may be found on them (i.e., a protein sequence search database). Other file formats such as PEFF[17] are better suited for this purpose.

When using the ProForma 2.0 notation for peptidoforms and proteoforms, amino acids are shown as is customary from left to right, *N*- to *C*-terminus, using IUPAC single letter identifiers. Modifications of this core set of amino acids are designated by a coded string of characters enclosed in square brackets after the letter of the modified residue. The modification string is represented by CV or ontology terms. The supported CVs/ontologies in ProForma 2.0 are PSI-MOD,[16] Unimod,[21] RESID,[22] XL-MOD (cross-linking; https://github.com/HUPO-PSI/xlmod-CV), and the Glycan Naming Ontology (GNO; glycans; https://www.ebi.ac.uk/ols/ontologies/gno).

ProForma 2.0 is case insensitive. This means that the notation is agnostic with regard to the use of uppercase or lowercase characters. However, different CVs and/or ontologies generally have their own specific policies for capitalization and representation of terms. It is, therefore, recommended that the capitalization specifications for each supported CV/ontology be used. It is also important to highlight that line breaks must not be used. There is currently no limit in maximum length since ProForma 2.0 can be used to represent both peptidoforms and proteoforms. Additionally, non-ASCII (American Standard Code for Information Interchange) characters are allowed since they may be included in the supported terms in the different CVs/ontologies.

A comparison of the features of ProForma 1.0 (finished in 2018) and 2.0 is shown in Table 1. At least 18 features were either added or expanded. Examples of ProForma 2.0 notations are provided in Table 2, along with the section number in the specification document (Supporting Information) that contains the detailed description of each feature. Note that custom user-specific information may be added to ProForma 2.0 entities by means of using "Information tags." Additionally, in the 2.0 version, the use of "Information tags" is the only mechanism to add metadata for a ProForma entity.

Levels of Compliance

It is important to highlight that software that implements the ProForma 2.0 notation may not support all aspects of the specification. For example, a standard proteomics search engine that outputs the ProForma notation does not have to support the cross-linking part of the notation. We have, therefore, defined five levels of ProForma 2.0 compliance (listed below) in order to make adoption easier. Details can be found in the specification document (Supporting Information, Appendix I).

Base level (base-ProForma compliant).

Level 2 (level 2-ProForma compliant).

Top-down extensions (level 2-ProForma + top-down compliant).

Cross-linking extensions (level 2-ProForma + cross-linking compliant).

Glycan extensions (level 2-ProForma + glycans compliant).

More than one of the extensions listed above (top-down, cross-linking, and glycan) could be supported by the same software.

Software Implementations

ProForma 2.0 has already been implemented in some existing software. The CTDP has established an initial proteoform registry where experimentally verified proteoforms are assigned a unique PFR (ProteoForm Record) identifier (http://www.proteoform.org/api).[23] This identifier system is essential for enhancing interoperability between tools and databases that include proteoform data. The registry is based on an API (Application Programming Interface) that accepts ProForma 2.0 sequences, compares them to known proteoforms already stored in the registry, and returns a new PFR identifier if the proteoform is new to the system. However, if the proteoform is already stored in the registry, a PFR identifier generated previously is returned. Then, ProForma 2.0 is needed as an input to the registry so that PFR identifiers can be provided.

There are currently four implementations of parsers and writers for ProForma 2.0, including the following:

1.   A .NET version, as part of the Top-Down Software Development Kit (SDK) (https://github.com/topdownproteomics/sdk). This includes a lexer/parser with some additional proteoform validation functionality.

2.   A Java port of the .NET reader and writer (https://github.com/NRTDP/proforma-java).

3.   A Python version of a parser and writer, which is now part of the Pyteomics[24] framework (https://github.com/levitsky/pyteomics). Additional documentation is available here: https://pyteomics.readthedocs.io/en/latest/api/proforma.html.

   4.   The spectrum_utils Python package[25] includes a parser using a formal grammar
            to convert ProForma strings into abstract syntax trees (https://github.com/
            bittremieux/spectrum_utils/).

ProForma strings are also an optional part of the recently developed
USI standard for the representation of PSMs (see some examples at http://
proteomecentral.proteomexchange.org/usi/). We expect that the adoption of ProForma will
increase broadly in the field, stimulated by its inclusion in widely used bioinformatics
resources such as those created by the CTDP, ProteomeXchange,[5] and UniProtKB,[4] among
others.

DISCUSSION AND CONCLUSIONS

ProForma 2.0 is a standard notation that is capable of supporting the needs of both the
bottom-up and the top-down proteomics communities. Since peptidoforms and proteoforms
are easily encoded in the ProForma 2.0 notation, it simplifies comparing the results of
different search engines. This will greatly facilitate the reuse of experimental data. We also
anticipate that the ProForma 2.0 notation will expedite integration of bottom-up and middle-/
top-down data, which is an active field of research.[26,27] Moreover, the notation can be used
as an input for the first version of the Proteoform Registry, which generates unambiguous
PFR identifiers for proteoform entities. The use of PFR identifiers is key to facilitating
proteoform data interoperability between multiple tools and protein databases.

Proforma 2.0 has been developed as a joint effort between the PSI and the CTDP and will
be actively maintained. Both organizations expect that this version 2.0 will not change for
an extended period of time since it addresses most of the relevant use cases at the time
of writing. However, additional use cases have already been envisioned and documented
in the specification document (see Section 5, "Pending Issues-Future developments," in the
Supporting Information). We expect that these extra features can be addressed in future
versions, after the community has gained experience with the more common use cases
included in version 2.0. The current list of known open issues includes the following:
representation of cyclic peptides, representation of more complex scenarios where there is
ambiguity in the localization of different glycans attached to the same amino acid sequence,
support for rare amino acids which are not assigned to an accepted one-letter code, support
for the use of average masses in the notation, lipid modifications, support for molecular
formulas, overlapping ranges of possible protein modification localizations, ambiguous
cross-linker modification positions, representation of the distribution of different isotopes
in the sequence, and representation of sequences coming from non-MS-based proteomics
approaches (e.g., peptide nanopores and Edman-based sequencing).

PSI standards are developed via an open process in which all interested individuals and
groups are encouraged to participate. ProForma 2.0 has been developed by contributors from
both the top-down and bottom-up proteomics subfields. This fusion provides the community
with a standard that supports a diverse array of use cases and creates the potential for a
substantially higher degree of software tool interoperability within the field than in the past.
Although standards that are cooperatively developed inevitably take longer to complete than

formats proposed by a single group, the resulting standards are more broadly applicable to many more use cases than those from independent initiatives. Broad participation is, therefore, essential for the successful generation of future standards for the proteomics community. See https://www.topdownproteomics.org/ to become involved in the top-down proteomics activities of the CTDP and https://psidev.info/ for information about how to contribute to the PSI.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

ABBREVIATIONS

API	Application Programming Interface
ASCII	American Standard Code for Information Interchange
CTDP	Consortium for Top-Down Proteomics
CV	controlled vocabulary
FAIR	findable, accessible, interoperable, and reusable
GNO	Glycan Naming Ontology
HUPO	Human Proteome Organization
IUPAC	International Union of Pure and Applied Chemistry
MS	mass spectrometry
PDB	Protein Data Bank
PEFF	PSI Extended FASTA Format
PFR	Proteoform Record

PSI	Proteomics Standards Initiative
PrSM	proteoform spectrum match
PSM	peptide-spectrum match
PTM	post-translational modification
SDK	Software Development Kit
UniProtKB	UniProt Knowledge-Base
USI	Universal Spectrum Identifier

REFERENCES

(1). IUPAC-IUB Commission on Biochemical Nomenclature. A One-Letter Notation for Amino Acid Sequence (Definitive Rules). Pure Appl. Chem. 1972, 31 (4), 639–646.

(2). Smith LM; Kelleher NL Proteoform: a single term describing protein complexity. Nat. Methods 2013, 10 (3), 186–187. [PubMed: 23443629]

(3). Rosenberger G; Liu Y; Rost HL; Ludwig C; Buil A; Bensimon A; Soste M; Spector TD; Dermitzakis ET; Collins BC; et al. Inference and quantification of peptidoforms in large sample cohorts by SWATH-MS. Nat. Biotechnol. 2017, 35 (8), 781–788. [PubMed: 28604659]

(4). Bateman A; Martin M-J; Orchard S; Magrane M; Agivetova R; Ahmad S; Alpi E; Bowler-Barnett EH; Britto R; Bursteinas B; et al. UniProt: the universal protein knowledgebase in 2021. Nucleic acids research 2021, 49 (D1), D480–D489. [PubMed: 33237286]

(5). Deutsch EW; Bandeira N; Sharma V; Perez-Riverol Y; Carver JJ; Kundu DJ; Garcia-Seisdedos D; Jarnuczak AF; Hewapathirana S; Pullman BS; et al. The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics. Nucleic acids research 2019, 48 (D1), D1145–D1152.

(6). Armstrong DR; Berrisford JM; Conroy MJ; Gutmanas A; Anyango S; Choudhary P; Clark AR; Dana JM; Deshpande M; Dunlop R; et al. PDBe: improved findability of macromolecular structure data in the PDB. Nucleic acids research 2019, 48 (D1), D335–D343.

(7). Fabregat A; Sidiropoulos K; Garapati P; Gillespie M; Hausmann K; Haw R; Jassal B; Jupe S; Korninger F; McKay S; et al. The Reactome pathway Knowledgebase. Nucleic acids research 2016, 44 (D1), D481–487. [PubMed: 26656494]

(8). Kerrien S; Aranda B; Breuza L; Bridge A; Broackes-Carter F; Chen C; Duesbury M; Dumousseau M; Feuermann M; Hinz U; et al. The IntAct molecular interaction database in 2012. Nucleic acids research 2012, 40 (D1), D841–D846. [PubMed: 22121220]

(9). Wilkinson MD; Dumontier M; Aalbersberg IJ; Appleton G; Axton M; Baak A; Blomberg N; Boiten JW; da Silva Santos LB; Bourne PE; et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 2016, 3, 160018. [PubMed: 26978244]

(10). LeDuc RD; Schwammle V; Shortreed MR; Cesnik AJ; Solntsev SK; Shaw JB; Martin MJ; Vizcaino JA; Alpi E; Danis P; et al. ProForma: A Standard Proteoform Notation. J. Proteome Res. 2018, 17 (3), 1321–1325. [PubMed: 29397739]

(11). Mayer G; Jones AR; Binz PA; Deutsch EW; Orchard S; Montecchi-Palazzi L; Vizcaino JA; Hermjakob H; Oveillero D; Julian R; et al. Controlled vocabularies and ontologies in proteomics: overview, principles and practice. Biochim. Biophys. Acta 2014, 1844 (1A), 98–107. [PubMed: 23429179]

(12). Deutsch EW; Albar JP; Binz PA; Eisenacher M; Jones AR; Mayer G; Omenn GS; Orchard S; Vizcaino JA; Hermjakob H Development of data representation standards by the human proteome organization proteomics standards initiative. J. Am. Med. Inform Assoc 2015, 22 (3), 495–506. [PubMed: 25726569]

(13). Martens L; Chambers M; Sturm M; Kessner D; Levander F; Shofstahl J; Tang WH; Rompp A; Neumann S; Pizarro AD; et al. mzML–a community standard for mass spectrometry data. Mol. Cell Proteomics 2011, 10 (1), R110.000133.

(14). Vizcaino JA; Mayer G; Perkins S; Barsnes H; Vaudel M; Perez-Riverol Y; Ternent T; Uszkoreit J; Eisenacher M; Fischer L; et al. The mzIdentML Data Standard Version 1.2, Supporting Advances in Proteome Informatics. Mol. Cell Proteomics 2017, 16 (7), 1275–1285. [PubMed: 28515314]

(15). Griss J; Jones AR; Sachsenberg T; Walzer M; Gatto L; Hartler J; Thallinger GG; Salek RM; Steinbeck C; Neuhauser N; et al. The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. Mol. Cell Proteomics 2014, 13 (10), 2765–2775. [PubMed: 24980485]

(16). Montecchi-Palazzi L; Beavis R; Binz PA; 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]

(17). Binz PA; Shofstahl J; Vizcaino JA; Barsnes H; Chalkley RJ; Menschaert G; Alpi E; Clauser K; Eng JK; Lane L; et al. Proteomics Standards Initiative Extended FASTA Format. J. Proteome Res. 2019, 18 (6), 2686–2692. [PubMed: 31081335]

(18). Deutsch EW; Perez-Riverol Y; Carver J; Kawano S; Mendoza L; Van Den Bossche T; Gabriels R; Binz PA; Pullman B; Sun Z; et al. Universal Spectrum Identifier for mass spectra. Nat. Methods 2021, 18 (7), 768–770. [PubMed: 34183830]

(19). Dai C; Fullgrabe A; Pfeuffer J; Solovyeva EM; Deng J; Moreno P; Kamatchinathan S; Kundu DJ; George N; Fexova S; et al. A proteomics sample metadata representation for multiomics integration and big data analysis. Nat. Commun. 2021, 12 (1), 5854. [PubMed: 34615866]

(20). Vizcaino JA; Martens L; Hermjakob H; Julian RK; Paton NW The PSI formal document process and its implementation on the PSI website. Proteomics 2007, 7 (14), 2355–2357. [PubMed: 17570517]

(21). Creasy DM; Cottrell JS Unimod: Protein modifications for mass spectrometry. Proteomics 2004, 4 (6), 1534–1536. [PubMed: 15174123]

(22). Garavelli JS The RESID Database of Protein Modifications as a resource and annotation tool. Proteomics 2004, 4 (6), 1527–1533. [PubMed: 15174122]

(23). Hollas MAR; Robey MT; Fellers RT; LeDuc RD; Thomas PM; Kelleher NL The Human Proteoform Atlas: a FAIR community resource for experimentally derived proteoforms. Nucleic acids research 2022, 50 (D1), D526–D533. [PubMed: 34986596]

(24). Levitsky LI; Klein JA; Ivanov MV; Gorshkov MV Pyteomics 4.0: Five Years of Development of a Python Proteomics Framework. J. Proteome Res. 2019, 18 (2), 709–714. [PubMed: 30576148]

(25). Bittremieux W spectrum_utils: A Python Package for Mass Spectrometry Data Processing and Visualization. Anal. Chem. 2020, 92 (1), 659–661. [PubMed: 31809021]

(26). Lima DB; Dupre M; Duchateau M; Gianetto QG; Rey M; Matondo M; Chamot-Rooke J ProteoCombiner: integrating bottom-up with top-down proteomics data for improved proteoform assessment. Bioinformatics 2021, 37 (15), 2206–2208. [PubMed: 33165572]

(27). Schaffer LV; Millikin RJ; Shortreed MR; Scalf M; Smith LM Improving Proteoform Identifications in Complex Systems Through Integration of Bottom-Up and Top-Down Data. J. Proteome Res. 2020, 19 (8), 3510–3517. [PubMed: 32584579]
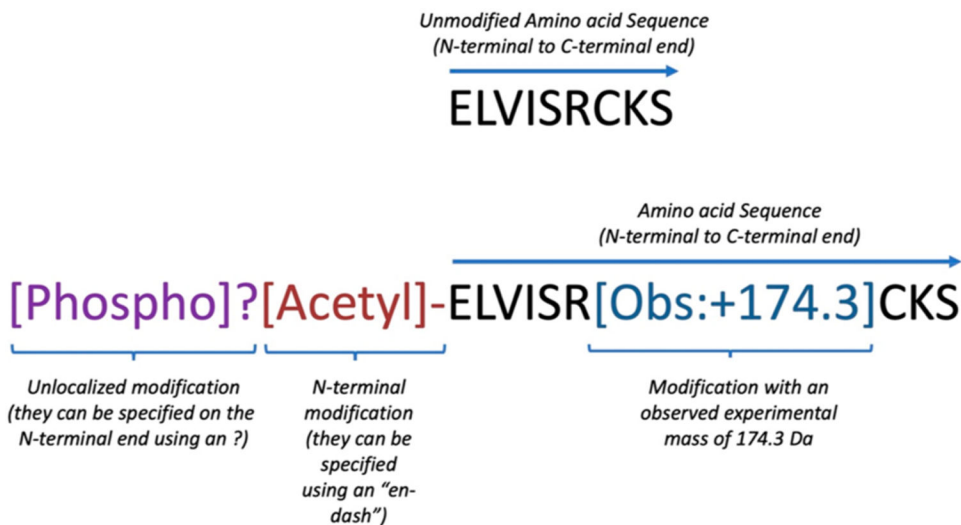
Figure 1.

Representation of the same *N*-terminal segment (sharing the same amino acid sequence) of two hypothetical proteoforms using ProForma 2.0: the unmodified proteoform (top part of the figure) and one containing different protein modifications (lower part of the figure). The text coloration is only included here to improve clarity. The purple tag encodes the existence of an unlocalized phosphorylation event somewhere on the proteoform. The keyword "Phospho" is from Unimod and can be used without additional clarification. The brown tag is a reference to an *N*-terminal modification using the term "Acetyl" from Unimod. A 174.3 Da mass shift on the arginine is also indicated.

Table 1.

Comparison of the Supported Features of ProForma 1.0 and 2.0

feature	ProForma 1.0	ProForma 2.0
protein modifications designated by CV/ontology names and accession numbers	✓	✓
representation of glycan composition	✓	✓
N-terminal and *C*-terminal modifications	✓	✓
delta mass notation for modifications	✓	✓
information tag	✓	✓
joint representation of experimental data and interpretation	✓	✓
NEW support for elemental formulas	limited	✓
NEW representation of isotopes	limited	✓
NEW cross-link notation	X	✓
NEW representation of interchain cross-links	X	✓
NEW representation of disulfide linkages	X	✓
NEW representation of glycans with GNO ontology as CV	X	✓
NEW specifying a gap of known mass	X	✓
NEW labile modifications	X	✓
NEW unknown modification position	X	✓
NEW possible set of modification positions	X	✓
NEW representation of ranges of positions for the modifications	X	✓
NEW modification position preference and localization scores	X	✓
NEW scoring for ranges of positions for a modification	X	✓
NEW fixed protein modifications	X	✓
NEW ambiguity in the order of amino acid sequences	X	✓
NEW representation of ion charges and more than one peptidoform per spectrum	X	✓
NEW representation of branched peptides	X	✓
NEW representation of ambiguity in the order of the amino acid sequence	X	✓

Table 2.

Examples of ProForma 2.0 Notations Demonstrating the Various Features of the Specification[a]

feature	example	section
CV/ontology modification names	EM[Oxidation]EVEES[Phospho]PEK	4.2.1
CV/ontology protein modification accession numbers	EM[MOD:00719]EVEES[MOD:00046]PEK	4.2.2
cross-link within the same peptide	EMEVTK[XLMOD:02001#XL1]SESPEK[#XL1]	4.2.3.1
interchain cross-links	SEK[XLMOD:02001#XL1]UENCE//EMEVTK[#XL1]SESPEK	4.2.3.2
disulfide linkages	EVTSEKC[MOD:00034#XL1]LEMSC[#XL1]EFD	4.2.3.3
branched peptides	ETFGD[MOD:00093#BRANCH]//R[#BRANCH]ATER	4.2.4
glycans using the GNO ontology as CV	NEEYN[GNO:G59626AS]K	4.2.5
delta mass notation for modifications	EM[+15.9949]EVEES[+79.9663]PEK	4.2.6
specifying a gap of known mass	RTAAX[+367.0537]WT	4.2.7
support for elemental formulas	SEQUEN[Formula:C12H20O2]CE	4.2.8
glycan composition	SEQUEN[Glycan:HexNAc1Hex2]CE	4.2.9
N-terminal and C-terminal modifications	[iTRAQ4plex]-EMEVNESPEK	4.3.1
labile modifications	{Glycan:Hex}EMEVNESPEK	4.3.2
unknown modification position	[Phospho]?EMEVTSESPEK	4.4.1
possible set of modification positions	EMEVT[#g1]S[#g1]ES[Phospho#g1]PEK	4.4.2
ranges of positions for the modifications	PROT(ESFRMS)[+19.0523]ISK	4.4.3
modification position preference and localization scores	EMEVT[#g1(0.01)]S[#g1(0.09)]ES[Phospho#g1(0.90)]PEK	4.4.4
scoring for ranges of positions for a modification	PROT(ESFRMS)[+19.0523#g1(0.01)]ISK[#g1(0.99)]	4.4.5
isotopes	<13C>ATPEILTVNSIGQLK	4.6.1
fixed protein modifications	<[MOD:01090]@C>ATPEILTCNSIGCLK	4.6.2
ambiguity in the order of the amino acid sequence	(?DQ)NGTWEMESNENFEGYMK	4.7
information tag	ELVIS[Phospho|INFO:newly discovered]K	4.8
joint representation of experimental data and interpretation	ELVIS[Phospho|Obs:+79.978]K	4.9
representation of ion charges	EMEVEESPEK/2	7.1
multiple peptidoforms assigned to chimeric spectra	EMEVEESPEK/2+ELVISLIVER/3	7.1

[a]For each feature listed in the first column, there is a representative example in the second column showing the encoding. The "section" column provides the location in the PSI specification document where the feature is explained in detail (Supporting Information).