Published in final edited form as:

Proteomics. 2019 May ; 19(10): e1800361. doi:10.1002/pmic.201800361.

Identification and Quantification of Proteoforms by Mass Spectrometry

Leah V. Schaffer[1], **Robert J. Millikin**[1], **Rachel M. Miller**[1], **Lissa C. Anderson**[2], **Ryan T. Fellers**[3], **Ying Ge**[1,4], **Neil L. Kelleher**[3,5], **Richard D. LeDuc**[3], **Xiaowen Liu**[6,7], **Samuel H. Payne**[8], **Liangliang Sun**[9], **Paul M. Thomas**[3], **Trisha Tucholski**[1], **Zhe Wang**[10], **Si Wu**[10], **Zhijie Wu**[1], **Dahang Yu**[10], **Michael R. Shortreed**[1], and **Lloyd M. Smith**[1,*]

[1.] Department of Chemistry, University of Wisconsin-Madison, Madison, Wisconsin 53706, United States

[2.] Ion Cyclotron Resonance Program, National High Magnetic Field Laboratory, Tallahassee, Florida 32310, United States

[3.] Proteomics Center of Excellence, Northwestern University, Evanston, Illinois 60208, United States

[4.] Department of Cell and Regenerative Biology and Human Proteomics Program, University of Wisconsin-Madison, Madison, Wisconsin 53706, United States

[5.] Department of Chemistry and Molecular Biosciences and the Division of Hematology-Oncology, Northwestern University, Evanston, Illinois 60208, United States

[6.] Department of BioHealth Informatics, Indiana University-Purdue University, Indianapolis, Indiana 46202, United States

[7.] Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana 46202, United States

[8.] Department of Biology, Brigham Young University, Provo, UT 84602

[9.] Department of Chemistry, Michigan State University, East Lansing, Michigan 48824, United States

[10.] Department of Chemistry and Biochemistry, University of Oklahoma, Norman, Oklahoma 73019, United States

Abstract

A proteoform is a defined form of a protein derived from a given gene with a specific amino acid sequence and localized post-translational modifications. In top-down proteomic analyses, proteoforms are identified and quantified through mass spectrometric analysis of intact proteins. Recent technological developments have enabled comprehensive proteoform analyses in complex samples, and an increasing number of laboratories are adopting top-down proteomic workflows. In

[*] Corresponding author. Lloyd M. Smith, W. L. Hubbell Professor of Chemistry. Department of Chemistry, University of Wisconsin-Madison, 1101 University Ave, Madison, Wisconsin 53706 608-263-2594 smith@chem.wisc.edu.

Conflict of Interest
The authors declare no conflict of interest.

this review, we outline some recent advances and discuss current challenges and future directions for the field.

Keywords

bioinformatics; mass spectrometry; proteoform; proteoform family; top-down proteomics

I. Introduction

Much of the biochemical diversity in cells occurs at the protein level. A proteoform consists of a defined amino acid sequence with localized modifications (Figure 1).[1] Proteoforms arise as a result of collective biological processes including amino acid variation, alternative RNA splicing, post-translation modification, and post-translational cleavage. Proteoforms derived from the same gene make up a proteoform family.[2] The biological function of different proteoforms can vary greatly, even among proteoforms belonging to the same family.[3,4] Therefore, the identification and quantification of specific proteoforms is vital to the understanding of a biological system.

Mass spectrometry (MS) has proven to be a valuable tool in the study of protein and proteoform biology.[5–7] MS-based proteomics can be divided into two broad categories: top-down and bottom-up (Figure 2). In bottom-up proteomics, proteins are digested by a protease to generate peptides that are analyzed via tandem MS.[8] Because peptides are directly detected instead of proteins, protein inference must be used to reconstruct the proteins hypothesized to exist in the sample.[9] Since proteins often contain homologous sequence regions, a number of identified peptides can reasonably emanate from more than one protein. These shared peptides significantly complicate the protein inference process.[10] The loss of information caused by the digestion of proteins, such as the relationship between the amino acid sequence and the PTMs belonging to a specific proteoform, prevents bottom-up proteomic analysis from identifying proteoforms.

Top-down proteomics directly analyzes intact proteins.[10–15] The relationship between amino acid sequence and PTMs is preserved and proteoforms can be characterized, providing a proteoform-specific understanding of biological phenomena. Recent reviews have covered sample preparation and mass spectrometry instrumentation techniques for top-down proteomic analyses.[7,11–13] Top-down proteomic experiments are subject to many analytical challenges owing to the low abundance of many proteoforms and the low signal-to-noise (S/N) ratios inherent to mass measurement of large molecules.[16] Data analysis of proteoforms is also challenging due to the complexity of intact and fragmented proteoform MS data. Co-elution of proteoforms in online LC-MS experiments further compounds these issues.

This review highlights recent advances in the field of top-down proteomics that enable the analysis of complex samples and the identification of thousands of proteoforms.[17–19] First, we define terminology used within proteomics. Next, we discuss recent developments in online and offline intact protein separation techniques, including size-based separations, reversed-phase liquid chromatography (RPLC), and capillary zone electrophoresis (CZE).

We then review techniques for proteoform identification and quantification as well as several different software tools including TDPortal[20], MSPathFinder[21], TopPIC[22], Proteoform Suite[23], MASH Suite[24], and MetaMorpheus[25]. Finally, we discuss current challenges and future directions specific to the analysis of proteoforms.

II. Proteomics Terminology

The language used in proteomics has significantly evolved over the years. Until recently, there was no established term to refer to a specific arrangement of amino acids and PTMs[1]; protein, protein form, and protein isoform were all used. In this section, we provide definitions for terms used in this review (Table 1).

A *protein* refers to a linear sequence of amino acids. PTMs may or may not be present, but a specific molecular form is not implied when the term *protein* is used. It is common to employ a one-protein-per-gene convention.[26] For example, the human genome contains ~20,000 genes, and the corresponding canonical human protein database from UniProt[27] contains ~20,000 protein sequences. Unfortunately, this is a crude approximation for the complexity of the human proteome. Genes are transcribed into pre-mRNA molecules, which are spliced into any number of different isoforms, each of which is translated into a unique protein. The UniProt definition of *protein isoform* is a member of a set of proteins from the same gene or gene family that arise from alternative splicing or variable promoter usage (https://www.uniprot.org/help/canonical_and_isoforms). Bottom-up proteomic analyses typically use a canonical protein database; including multiple isoforms per gene increases sequence redundancy in the database, complicating protein inference. Not including protein isoforms in the database, however, renders a large portion of the actual proteome invisible to identification.

In bottom-up proteomics, the presence of a protein can be confidently inferred only when a peptide unique to that protein is identified.[9] There can be high sequence homology between proteins and certainly between protein isoforms. A *protein group* is a collection of proteins that are indistinguishable based on the peptides identified.[9]

A *proteoform* is a defined sequence of amino acids with localized modifications. One gene can yield a wide variety of proteoforms.[1] The term *proteoform family* refers to all proteoforms derived from a single gene, including protein products from all mRNA splice forms, post-translational processing, and PTMs.[2] There is a desire to classify proteoforms within a gene-centric system, which creates a major need in top-down proteomics for a unified proteoform nomenclature convention. ProForma, a human-and machine-readable nomenclature for writing proteoform sequences, was created to meet this need.[28] An open-source parser/writer application called the TopDown Software Development Kit (SDK) has been developed (http://github.com/topdownproteomics/sdk) to facilitate the passing of results between software programs.

The term *proteoform spectrum match* (PrSM) is used to describe a search algorithm’s match between a proteoform identification and an observed tandem mass spectrum. This is analogous to the term *peptide spectrum match* (PSM) frequently used in bottom-up proteomics.

III. Separation of Proteoforms

Overview

Separating proteoforms is important because co-elution of proteoforms is detrimental to proteoform identification in several ways. Many high-resolution mass spectrometers have a finite ability to detect proteoforms due to limited charge capacity, so low-abundance proteoforms are often not observed without enrichment. Additionally, co-isolation of multiple proteoforms prior to fragmentation complicates data analysis and proteoform identification. MS2 spectra of intact proteins often contain many overlapping fragments that are difficult to resolve. This necessitates high resolving power as well as the averaging of considerable numbers of spectra to improve S/N. These obstacles can limit analyses to proteoforms with molecular weights below 30 kDa, rendering more than half of the human proteome inaccessible.

A wide variety of fractionation and separation techniques have been combined to reduce the complexity of samples delivered to the mass spectrometer.[11, 29] Here, we highlight such techniques (summarized in Table 2), including size-based separations, high-and low-pH reversed-phase liquid chromatography (RPLC), and capillary zone electrophoresis (CZE). Additionally, other chromatography modes, such as hydrophobic interaction chromatography (HIC)[30, 31] and ion exchange chromatography (IEX)[32], have been reported (see Chen *et al*).[11] Another related and emerging separation in top-down proteomics is ion mobility spectrometry.[33, 34]

Size-Based Separation

Top-down analysis of larger proteoforms is challenging due to the inverse relationship between S/N and proteoform mass.[16, 35] Inherently low S/N combined with the co-elution of smaller proteoforms compromises observation of large proteoforms without prior fractionation. Tran *et al.* pioneered the gel-eluted liquid fraction entrapment electrophoresis (GELFrEE) technique, which achieves high-resolution size-based separation of proteoforms.[36] GELFrEE has been coupled with other separation techniques to achieve deep coverage.[19, 37, 38] GELFrEE has also been adapted for native-state size separations to separate protein complexes[39] and preserve higher order structure.[40]

Size exclusion chromatography (SEC) is an appealing option for fractionation because of its compatibility with a variety of eluents, including those which are MS-compatible *(e.g.* 1 % formic acid, ammonium acetate).[41] Solubilizing intact proteins in MS-compatible buffers is one major challenge at the front-end of top-down proteomic workflows, especially for membrane proteins which are inherently hydrophobic.[42] Although acid-labile surfactants such as RapiGest™ (RG, also known as ALS, Waters)[43, 44], ProteaseMax™ (PM, Promega)[45], and the recently developed MS-compatible slowly-degradable surfactant (MasDeS)[46] have been proven to be highly effective for bottom-up proteomics, they have yet to be demonstrated as directly compatible with intact protein MS analysis.

Ge and co-workers recently introduced serial size exclusion chromatography (sSEC), which combines multiple matrix pore sizes for separation of complex protein mixtures in MS-compatible buffers.[47] SEC is a non-adsorptive mode of liquid chromatography that sorts molecules based on their size and differential access to matrix pore volume.[48, 49] The combination of SEC with different pore sizes provides an extension of molecular weight fractionation range and higher-resolution separation of proteins compared to conventional SEC. sSEC coupled to RPLC-quadrupole-time-of-flight mass spectrometry provided improved proteome coverage, with a 15-fold improvement in the observation of proteoforms >60 kDa compared to one-dimensional RPLC alone and enabled the observation of large proteoforms (up to 223 kDa). More recently, the Ge lab paired sSEC with FT-ICR MS analysis to enable sequence characterization of large proteoforms without RPLC separation or protein purification.[50]

Reversed-phase liquid chromatography separation

RPLC is the most prevalent approach for complex intact protein sample separation and fractionation.[12, 51–54] Various improvements such as smaller particle sizes and longer columns have been developed to increase the peak capacity for deeper proteome analysis. Smaller packing particles provide more uniform packing structure in columns, which improves peak symmetry and separation resolution.[55, 56] However, as particle size decreases, the pressure required to provide sufficient flow increases dramatically.

Longer columns and higher-pressure pumps are also used to improve peak capacity. Recent studies suggest that elution peak widths with longer columns do not increase remarkably with longer separation windows *(i.e.,* 200 minutes or longer).[57] Therefore, under ultra-high pressure (10,000 psi) and long elution gradients, column length may have a greater impact on the separation resolution than particle size. Recently, Shen *et al.* identified ~900 proteoforms from S. *oneidensis* lysate using long-column RPLC-MS analysis.[58] Similarly, Ansong and coworkers utilized a long C5 capillary column (80 cm) and identified 1,665 *Salmonella* proteoforms.[52] Longer RPLC columns and high-pressure systems can also be applied to separate proteoforms with high sequence homology. Wang and coworkers reported the long column ultrahigh pressure liquid chromatography (UPLC) top-down analysis of human serum autoantibodies that are complex with highly homologous clonal sequences. They identified 47 light chains and 16 heavy chains in the SLE patient serum, providing the first “bird’s-eye” view of the complexity of human serum autoantibodies.[59] One major limitation of long column UPLC separation for top-down proteomics is a lack of commercially-available long columns. In-house packing of long columns often requires customized equipment and expertise for uniform packing.

Wang and co-workers recently reported a two-dimensional (2D) separation using high-pH and low-pH RPLC for top-down proteomics.[57] The orthogonality between these two approaches was first demonstrated by the observation of different elution orders under different pH conditions using the same set of standard proteins. Different proteoform elution behaviors were also observed between low-pH and high-pH RPLC separation of complex samples *(e.g., E. coli* cell lysate). The orthogonality between high-pH and low-pH conditions may be related to different polarities of five amino acids (Glu, Asp, Arg, His, and

Lys) under different pH conditions.[60] Proteoform retention behaviors are less molecular weight dependent under high-pH conditions, which may be due to different distributions of charges in the proteoform primary structure.[57, 61–63] A potential drawback is that some proteins are less charged under high-pH conditions and cannot be efficiently eluted from the column.

Offline 2D-RPLC fractionation is relatively simple because fractions from the high-pH column are concentrated using vacuum drying prior to injection onto the low-pH column. When performing online 2D-RPLC, an additional dilution step must be incorporated due to the high organic solvent content of the eluent from the high-pH separation. One successful application of this 2D high-pH and low-pH approach resulted in the identification of 886 proteoforms in *E. coli* cell lysate, which is a significant improvement over the 328 proteoforms identified using one-dimensional low-pH RPLC separation.[57]

Capillary Zone Electrophoresis

CZE is a method of separating proteoforms by differences in electrophoretic mobility that does not require a stationary phase. It is orthogonal to RPLC because the separation is based on size and charge rather than hydrophobicity. CZE-MS/MS has recently attracted attention for top-down proteomics due to major improvements in the CZE-MS interface[64–66], the coating on the inner wall of separation capillaries[67], and sample stacking methods for online concentration.[68–70] In 1996, the McLafferty group detected attomole amounts of intact proteins using CZE-MS and identified carbonic anhydrase in a human red blood cell lysate using MS/MS.[71] More recently, the Yates group reported nearly 300 proteoform identifications from *Pyrococcus furiosus* lysate using RPLC-CZE-MS/MS.[72] The Yates group also showed that CZE-MS provided comparable S/N ratios to RPLC-MS for characterization of a protein complex with 100-fold less sample consumption.[73] The Dovichi group has demonstrated 600 proteoform identifications from yeast lysate using RPLC-CZE-MS/MS.[74] The Kelleher group has also demonstrated the potential of CZE-MS/MS for top-down identification of large proteins.[75]

Two major issues impede the use of CZE-MS/MS for large-scale top-down proteomics. First, the sample loading volume of CZE is low, typically 1% of the total capillary volume. For example, the typical sample loading volume of a fused silica capillary with a 50-μm inner diameter and a 1 -meter length is approximately 20 nL, which is 2 to 3 orders of magnitude lower than the volume of sample used in nanoflow RPLC. The sample loading volume of CZE must be increased to enable better characterization of low-abundance proteoforms in complex samples. Second, CZE is well known for its fast separation of various analytes and narrow separation window (typically less than 30 minutes). When analyzing complex mixtures of proteoforms, however, a longer separation window is desirable because it allows the mass spectrometer to acquire a larger number of MS and MS/MS spectra, resulting in more proteoform identifications.

The Sun group recently made progress towards increasing the sample loading volume and the separation window of CZE. They achieved a 90-min separation window and microliter-scale sample loading volume for CZE-MS/MS analysis of an *E. coli* cell lysate and identified 600 proteoforms in a single run.[76] To facilitate the wide separation window and increased sample volume, a separation capillary with a high-quality neutral coating on its inner wall and a highly efficient protein stacking method based on a dynamic pH junction principle were utilized. They also coupled SEC-RPLC to CZE-MS/MS and identified nearly 6,000 proteoforms and 850 proteoform families, which is one of the largest reported top-down proteomic datasets thus far.[18] A detailed protocol for CZE-MS/MS-based top-down proteomics has been provided to facilitate adoption of this technique.[77] More recently, the Sun lab developed an SEC-CZE-MS/MS system for native top-down proteomics which identified 23 protein complexes in *E. coli.*[78] This work showed the first example of native top-down proteomic analysis of a complex proteome using online liquid-phase separation.

There are still at least two concerns about using CZE-MS/MS for large-scale top-down proteomics. First, it is still challenging to make high-quality neutral coatings on the inner wall of the separation capillary in a reproducible manner. Second, the separation window of CZE needs to be increased further. When a 1-meter separation capillary with a high-quality neutral coating and a 30 kV voltage were employed for CZE separation of proteoforms, the separation window was about 90 min.[76] The most advanced RPLC proteoform separation with a 1-meter long column has reached an 800-min separation window.[58]

IV. Informatics Tools for Proteoform Identification

Proteoform Identification

What does it mean to *identify* a proteoform? There is not currently a unified metric for calling a proteoform *identified.* Different manuscripts utilize different levels of characterization or false discovery rates. Top-down MS analyses typically identify proteoforms by measurement of the proteoform's intact mass and the observed MS/MS fragment peaks. Search software programs select the best matching theoretical proteoform from a set of candidates, and give this match a score and a measure of statistical confidence.[7] Complete sequence coverage of the proteoform is almost never achieved with these analyses. Instead, only a subset of theoretical fragments matches to observed fragments. If PTMs are present on the proteoform, they are best localized with fragment ions containing the modified amino acid residue, which are only observed a fraction of the time. More work is needed to yield greater sequence coverage and confidence in proteoform identifications. When a count of identified proteoforms is stated in this review, readers are encouraged to consult the original source material for the authors' threshold for a proteoform to be considered identified. There has not yet been a comprehensive analysis comparing the identification thresholds of available top-down software tools and how this affects proteoform identification results. We describe below several freely available software tools for the identification of proteoforms (see summary in Table 3). Many of these tools have publicly available source code, which provides the community complete access to and understanding of the tool.

TDPortal

Although ProSightPTM versions 1[79] and 2[80] are still among the most widely cited top-down search tools, the National Resource for Translational and Developmental Proteomics at Northwestern University (NU) now also maintains the TDPortal. This service provides top-down protein and proteoform identification and label-free intensity values needed for relative proteoform quantification. Several reference proteome databases are provided. Search results include truncated and modified proteoforms. The service presents a Galaxy[81–83] frontend that allows users to queue search jobs, resulting in rapid search times. Results are returned in .tdreport files which can be viewed with the freely available TDViewer application available at http://topdownviewer.northwestern.edu. TDPortal is free to academic users and can be requested for use at http://nrtdp.northwestern.edu/tdportal-request.

TopPIC

TopPIC (Top-down mass spectrometry-based Proteoform Identification and Characterization) excels in identifying proteoforms with unknown alterations (PTMs and post-translational processing) in a blind mode.[22] TopPIC uses indexes of fragment masses[84, 85] to quickly filter candidate protein sequences in a database. TopPIC then performs spectrum alignment[86–88] to match experimental spectra to database sequences and finally uses a generating function method[89] to estimate the statistical significance for each identification. It can further match unknown mass shifts in identified PrSMs to common PTMs provided by the user and localize them using a Bayesian model.[90]

TopPIC identifies proteoforms using three search modes. The first mode is analogous to no-enzyme searches in bottom-up MS. Unknown alterations, except for terminal truncations, are excluded from the database search. The second mode is a combination of the no-enzyme search and an open database search[91], in which one unknown mass shift is allowed per identified proteoform. The third mode uses spectrum alignment to identify proteoforms with two unknown mass shifts. The algorithm for this mode utilizes internal proteoform fragments without alterations for protein sequence filtering and spectrum alignment. The internal fragment-based filtering method is used to reduce the search space, and the alignment method identifies a proteoform by connecting several unmodified protein segments, which is similar to the sequence similarity search algorithm in FASTA.[92] TopPIC is freely available at http://proteomics.informatics.iupui.edu/software/toppic/.

TopMG

TopMG[93] (Top-down mass spectrometry-based proteoform identification using Mass Graph) is a software tool for identification of ultra-modified proteoforms with variable PTMs and unknown alterations. One important challenge in identifying ultra-modified proteoforms is the combinatorial explosion of PTM patterns in a long protein. In TopMG, all proteoforms of a protein with variable PTMs are efficiently represented by a mass graph, which is aligned with a query spectrum to identify a modified proteoform that best explains the fragment masses in the spectrum.

TopMG also incorporates efficient algorithms for filtering protein sequences and estimating the statistical significance of identifications. Because a proteoform with multiple PTMs often lacks a long unmodified protein fragment, a protein sequence filtering method based on unmodified protein fragments becomes inefficient. To obtain long unmodified protein fragments, an approximate spectrum-based method is used to remove one or two PTM sites in query spectra by shifting fragment masses to cancel out mass shifts introduced by PTMs.[94] Because many proteoforms of a protein are almost the same except for PTM sites, it is a challenging problem to accurately estimate the proteoform-level statistical significance of identifications. TopMG reports protein-level E-values of identifications by combining a Markov chain Monte Carlo method and a fast algorithm for estimating similarity scores between query spectra and proteins.[95] TopMG is freely available at http://proteomics.informatics.iupui.edu/software/toppic/.

Proteoform Suite

Not all proteoforms observed in the MS1 spectra are able to be selected for fragmentation by the mass spectrometer in a typical data-dependent top-down workflow.[74, 96, 97] The MS2 resolution and ion fillings require long scan times, so only two to three proteoforms are selected for fragmentation following each MS1 spectrum. Since proteoform chromatographic peaks are on the order of a minute long, there are only a few opportunities to fragment each eluting proteoform. The problem is further exacerbated by different charge states of the same proteoform being selected for fragmentation. As a result, many proteoforms observed in the MS1 spectra are not selected for fragmentation. Additionally, many fragmentation events do not result in a successful identification.

The Smith lab recently developed Proteoform Suite[23], which identifies proteoforms observed in MS1 spectra using the observed intact mass. Proteoform masses from deconvoluted MS1 spectra are compared to a database of theoretical proteoform masses as well as to other experimental masses to form “proteoform relations”. Proteoform relations are mass differences corresponding to a known PTM, an amino acid, or combinations thereof. Proteoform families are subsequently constructed by grouping together all proteoforms connected to one another in a proteoform relation. The results are visualized as a network of proteoforms, where circles represent proteoform masses and lines connecting circles represent proteoform relations corresponding to a mass difference (Figure 3). A challenge in intact-mass analysis is that the specific arrangement of amino acids is less confidently determined than when fragmentation is utilized, and PTMs are not able to be localized to specific residues.

Recently, Proteoform Suite was augmented to enable the input of top-down MS/MS identifications. Additional identifications can be made by comparing the masses of proteoforms identified by MS/MS to unidentified proteoform masses observed in the MS1 spectra. Using intact mass to identify additional proteoforms in a top-down analysis of fractionated yeast lysate resulted in an approximately 40% increase in the number of proteoform identifications compared to MS/MS analysis alone.[98] In a subsequent study, this intact-mass approach was applied to a mammalian system of reduced complexity: mitochondrial proteins from mouse myoblasts and differentiated myotubes.[99] A similar ~40% increase in the number of proteoform identifications was observed compared to top-down MS/MS data analysis alone. Proteoform Suite also determined statistically significant proteoform abundance changes across myoblast and myotube cell types. Proteoform Suite automates identification and quantification of proteoforms observed in MS1 spectra as well as the construction and visualization of proteoform families. Proteoform Suite is open source and freely available at http://github.com/smith-chem-wisc/ProteoformSuite.

MetaMorpheus

MetaMorpheus[25] is an integrated bottom-up and top-down software program for identification of peptides and proteoforms in high-resolution data-dependent mass spectrometry experiments. The search algorithm begins by deconvoluting the isolated *m/z* window of each MS1 scan into monoisotopic masses. In addition to providing a highly accurate precursor mass for theoretical proteoform selections, this deconvolution process also provides an opportunity to reveal multiple co-fragmented precursors, so MetaMorpheus can identify multiple peptides/proteoforms from a single tandem mass spectrum. A theoretical set of proteoforms is constructed from the user-supplied protein database. PTM annotations obtained from UniProt are used by MetaMorpheus. In practice, the deconvoluted monoisotopic mass can be incorrectly estimated by several Daltons (see Deconvolution section); MetaMorpheus compensates for these monoisotopic mass errors by allowing mass differences at approximately 1 Da spacings. After determining the MS2 scan's precursor mass(es), the MS2 spectrum itself is deconvoluted. Each valid theoretical proteoform for the scan is fragmented *in silico* and matched against the spectrum's deconvoluted masses; the score of the PrSM is the number of matched fragment ions plus the fraction of the TIC belonging to the matched set of peaks *(e.g.,* a score of 74.65 means 74 fragments matched, which composed 65% of the MS2 TIC). The false-discovery rate of a dataset is estimated using the target-decoy approach.[100]

Global PTM Discovery[101] (G-PTM-D) is a strategy used by MetaMorpheus[25] to identify PTM-containing peptides in bottom-up proteomics. It has been extended to top-down analysis to identify proteoforms with unknown PTMs. Briefly, a catalog of known PTM masses is used as accepted mass differences, much like the monoisotopic mass error strategy described above. If a PrSM's theoretical mass differs from the experimental mass by a known PTM's mass, the PTM is then annotated in the database. A second-pass search with the annotated database is then performed to estimate the FDR. One advantage of MetaMorpheus is that the G-PTM-D process can also be performed on a bottom-up sample. The resulting database can then be used in a subsequent top-down search for proteoform identification.[102] This is advantageous because proteome coverage is deeper in bottom-up. Observation of PTMs in both bottom-up and top-down data from the sample provides improved confidence in PTM localization.

MetaMorpheus features include mass-calibration, open-mass searches, identification of co-isolated analytes, and automated spectrum annotation, which assists in manual evaluation of PTM assignments. MetaMorpheus is open source and freely available at http://github.com/smith-chem-wisc/MetaMorpheus.

MSPathFinder

MSPathFinder is a tool for spectrum identification and label-free quantification of top-down proteomic data.[21] It is open source and freely available at http://github.com/PNNL-Comp-Mass spec/Informed-Proteomics/. The goal of the tool was to address two significant challenges in top-down proteomics: obtaining the correct monoisotopic mass and efficient searching of modified proteoforms.

Identifying the correct monoisotopic mass is essential to an efficient and accurate identification algorithm. MSPathFinder addresses this challenge by simultaneously integrating data from all ions and across the LC time scale to boost signal intensity and improve the shape of the isotope profile. This method significantly improved monoisotopic mass determination, as evidenced by a greater number of monoisotopic masses being present in replicate data acquisitions.

The second focus of MSPathFinder is to efficiently explore the combinatorial search space of proteoforms with post-translational modifications. MSPathFinder addresses this challenge with a sequence graph. This compact graph represents all possible modification sites within a protein and takes advantage of the fact that many different proteoforms (different PTM placements) share most of their fragment ions. This approach allows the algorithm to score various fragment ions much more efficiently and reduces the search time by several orders of magnitude.

MASH Suite

The Ge research group recently developed a comprehensive and user-friendly software tool, MASH Suite Pro, with multifaceted functionality for data analysis in top-down proteomics to enable researchers to perform proteoform identification, quantification, and characterization with visual validation.[24] This software provides an intuitive interface. Users can import deconvolution and protein searching results into MASH Suite Pro for manual validation of computational outputs. This function allows users to perform manual correction of charge states and isotopic masses of fragment ions for comprehensive characterization of sequence variations and PTM sites.[103] The software also offers quantitative tools to evaluate the relative abundances of different proteoforms under various experimental conditions.[104]

Public Proteoform Repository

The Consortium for Top-Down Proteomics has taken on the task of organizing experimentally verified biological proteoforms and providing them unique identifiers. Called PFR, or Proteoform Record, the PFR is a durable identifier which uniquely identifies a proteoform. The CTDP’s Proteoform Repository can be accessed at http://atlas.topdownproteomics.org/. All laboratories with interest in top-down proteomics are encouraged to submit their experimentally verified proteoform discoveries to this repository.

V. Quantification of Proteoforms

Proteoform-level abundance changes can be determined using several different quantification strategies. Here, we describe three strategies for the large-scale measurement of the relative abundance of proteoforms. Label-free quantification (LFQ) uses measurements of MS1 chromatographic peak height or area.[105] In stable isotope labeling of amino acids in cell culture (SILAC), each sample is cultured using multiple isotopic forms of an amino acid that differ in mass.[106] The MS1 intensity ratio of the two isotopic forms provides the relative quantitative difference of the proteoform between samples. Finally, in Tandem Mass Tag (TMT), proteoforms are labelled on the N-terminus and on side-chain amino groups.[107] Each sample is labeled with a separate tag with a unique reporter ion mass. The intensities of the TMT reporter ions in fragmentation spectra provide a measure of proteoform abundance.

Because the S/N ratios of intact proteins are much lower[16] and the chromatographic separations are typically less reproducible than in bottom-up[108], the quantification of proteoforms is technically challenging. However, protein inference is a major problem in bottom-up protein quantification.[9] Because the identification of the proteins is inferred from the observed peptides, the quantification of the protein is also inferred from its constituent peptides. In top-down, protein inference is not required, which makes direct proteoform quantification possible. The information uncovered in top-down quantification – the abundance changes for the specific proteoforms in the system – is important to biological understanding.

Label-free quantification (LFQ)

Labeling reagents are not needed in LFQ, so it can be applied to samples not grown in cell culture, such as patient tissue samples. LFQ has been applied to several larger scale studies of proteoform abundance changes across two conditions.[20, 99, 105, 109–111] The Kelleher lab first demonstrated this technique using yeast mutant vs. wild-type strains and quantified over 800 proteoforms. This study implemented a hierarchical linear model to account for sources of variation when determining proteoforms with statistically significant changes across biological conditions.[110] The approach was subsequently used to quantify over 1000 human proteoforms in fibroblasts with and without induced senescence; quantitative mass targets are determined and MS2 verified identification is attempted.[112] The Smith lab has also recently implemented a label-free quantification strategy in Proteoform Suite, where the differences in abundance of observed masses are determined across two conditions.[99] Proteoform Suite was used to quantify mouse mitochondrial proteoforms in myoblasts and differentiated myotubes and determined 129 proteoforms with statistically significant abundance changes. LFQ has also been applied in targeted approaches to quantify proteoforms from a specific proteoform family across conditions.[12, 113, 114]

A major challenge in top-down LFQ is the necessity of reproducible sample handling across replicates to ensure that observed changes in intensity are biological and not artefactual. A second major challenge for LFQ of proteoforms occurs from sample fractionation. High levels of fractionation are necessary to facilitate the identification of less abundant proteoforms. However, proteoforms can be split between multiple fractions. One response to this is the summation of proteoform signal across all fractions in which it appears prior to comparison between conditions. Unfortunately, sample handling and measurement artifacts can occur across fractions or replicates.

Stable Isotope Labeling of Amino Acids in Cell Culture (SILAC)

In SILAC labeling, proteoform samples from different conditions are cultured separately but mixed at the earliest possible point in the workflow, allowing similar sample handling throughout preparation, fractionation, and MS analysis stages.[115] MS1 intensities of the two co-eluting isotopic forms are used to determine abundance differences of proteoforms between conditions. SILAC labeling results in lower quantitative variation compared to LFQ because of the uniform sample handling for each labelled form.

NeuCode labeling, a variant of SILAC, employs isotopic forms of lysine where the mass difference is only a few mDa.[116–118] NeuCode quantification has been applied to top-down analysis.[118] Measurement of the ratio of isotopic forms provides the relative abundance. The Smith lab recently used NeuCode isotopic labeling to identify proteoforms by intact mass and number of lysine residues, as the space between a proteoform's light and heavy isotopic forms is directly proportional to its number of lysines.[2] Proteoform Suite was used to identify 638 yeast proteoforms and determine that 64 experimental proteoforms exhibited statistically significant changes between normal and salt-stress conditions.[23]

There are several major challenges with using SILAC labeling in top-down proteomics. There is a decreasing likelihood of labeling the entire proteoform with the isotopic label as the number of residues in the proteoform increases. Additionally, the presence of two sets of isotopic envelopes for a given proteoform complicates the spectra and data analysis.

Tandem Mass Tag (TMT)

Isobaric chemical tags, such as isobaric tagging for relative and absolute quantification (iTRAQ)[119] and tandem mass tags (TMT)[107], have been developed to simultaneously identify and quantify proteins using tandem MS. The labeling reagents consist of an amine reactive group, a mass normalizer, and a mass reporter. The mass normalizer and the mass reporter carry isotopes in different combinations so that the mass reporters have different masses while the intact mass of reagents remains the same. Isobaric labeling techniques have been widely applied in bottom-up proteomics[120, 121], and they have been utilized to label intact proteins[122–124]; however, until recently, only a few attempts on standard proteins have been made in top-down MS.[125] Recently, Yu and co-workers developed a protein-level TMT labeling platform for intact proteoform quantification in complex protein samples *(e.g., E. coli* cell lysate) (manuscript in preparation). The HCD-based fragmentation approach was used to generate the reporter ions for quantitation and sequence fragments for identification. TopPIC was used for proteoform identification in which the TMT modification on lysine residues and N-termini is set as a fixed PTM.[22] In total, 408 intact proteoforms from 95 proteins were confidently identified and quantified from two LC-MS/MS runs after manual evaluation. Among them, 303 proteoforms were completely labeled (both at the N-terminus and at all lysine residues), while 64 proteoforms were labeled at all lysine residues with a missing label at the N-termini. The results demonstrate that the optimized proteoform-level TMT labeling platform can efficiently label and quantify intact proteoforms in complex samples.

VI. Opportunities and Challenges

While top-down proteomics offers unique advantages over bottom-up proteomics for its ability to identify proteoforms, technologies in the field are not as established or robust. In this section, we discuss several important remaining challenges in the field.

False Discovery Rate and Characterization of Proteoforms

The false-discovery rate (FDR) determination in top-down proteomics is currently understudied, and there are two concepts of proteoform identification which are frequently conflated. The first is determining the FDR associated with a proteoform or protein identification, and the second is determining the extent to which a proteoform has been characterized.

Several top-down tools have implemented proteoform FDR estimation based on target-decoy strategies *(e.g.,* TopPIC[22], Informed Proteomics[21], MetaMorpheus[25], and TDPortal[20]). The proteoform identification FDR is the fraction of a set of identifications that are expected to be incorrect. Since FDR is a multiple testing correction, proteoform identification FDR is not a property of a single observation, but rather is a property of the complete set of observations. LeDuc *et al.* have recently shown that PrSM FDRs are not sufficient to control the FDR at either the proteoform or protein level.[126] When reporting a PrSM FDR of 1%, the true proteoform FDR was several times higher; this discrepancy increases with larger study designs. The authors have provided a tool which will estimate proteoform-level FDR from the PrSM FDRs from any search tool.

The second problem is determining the extent of characterization of a proteoform given the available fragmentation data. The lack of complete overlap between detected proteoform fragment peaks and theoretical fragments requires reliance on a database to contain a faithful sequence representation of each proteoform in the sample. Some proteoforms consist of an identical set of amino acids but arranged in a different order (*e.g.,* yeast histone proteins H2A.1 and H2A.2). These proteoforms have the same intact mass and share many fragment masses. However, only observation and identification of fragments from the portion of the proteoform with the distinguishing sequence facilitates definitive identification of the appropriate proteoform. The C-score was introduced as a score that indicates the level of characterization of a proteoform identification within a fully defined search space.[127] Tools such as Informed Proteomics, MSPathFinder, MetaMorpheus, and the ProSight tools use prior knowledge of PTMs to narrow the search space for proteoform identification.

Sensitivity

Bottom-up proteomics yields higher-sensitivity for protein detection across the mass and concentration dimensions (Figure 4 and Supporting Figures S1 and S2). Proteoforms are digested into peptides, which have a uniformly low molecular weight and charge. As a result, mass spectra of peptides have fewer peaks as the ion current is distributed among fewer ion channels. Thus, the peaks have higher S/N ratios and fall well within the available scan range of most mass analyzers.[16] Peptides also generally exhibit higher chromatographic resolution and are easier to fractionate, which results in delivery of peptides to the mass spectrometer at an optimal rate.[128] The fractionation and identification of intact proteoforms is far more challenging. Intact proteoforms present solubility challenges that peptides do not possess. Membrane proteoforms, for instance, can have large regions of highly hydrophobic amino acids that span the cell membrane. These species are nearly insoluble in MS-compatible buffers, whereas peptides are usually soluble.[129] While there have been improvements in offline fractionation techniques to enhance solubility and decrease sample complexity, as mentioned earlier, the resolution of online intact proteoform separations is still often not high enough to prevent co-elution of proteoforms.

Bottom-up proteomics also exhibits higher sensitivity than top-down proteomics because S/N decreases as a function of molecular weight.[16] This translates to lower S/N and longer accumulation times in top-down analyses. Proteoforms can be hundreds of amino acids long and can fragment at multiple locations along the backbone, sometimes more than once, which generates internal fragment ions. Therefore, it is common for fragment ion signals to fall below detection limits. Techniques such as charge-reduction[130] and ion-parking[131–134] have been used to simplify spectra and enhance sensitivity. However, sensitivity for identification of proteoforms diminishes severely beyond 30 kDa.[16]

Deconvolution

Precursor and fragment mass measurements of intact proteins are often complicated by complex isotope distributions. As mass increases, the relative abundance of the monoisotopic peak of any given multiplet inevitably decreases because the likelihood of observing a proteoform containing one or more heavy isotopes increases as the number of atoms in the proteoform increases (Figure 5). Isotopic deconvolution is the process of collecting an isotopic envelope's peaks and determining its monoisotopic mass, charge, and summed intensity. Charge state deconvolution is the process of collecting all the isotopic envelopes of a proteoform in different charge states. These two processes are collectively referred to as *deconvolution*.

Proteomicists rely on deconvolution software for interpretation of mass spectra. Many deconvolution algorithms exist (*e.g.* THRASH[135], MaxEnt[136], MSDeconv[137], Promex[21], UniDec[138]), which can be divided into interpreting two categories of species: isotopically-resolved and non-isotopically-resolved. The deconvolution of isotopically resolved species determines the charge of each species from the *m/z* spacing and intensity distribution of its isotopic peaks, while the deconvolution of non-isotopically-resolved species determines the charge from the *m/z* spacing between charge states and the charge state intensity distribution. Deconvolution usually takes place without the aid of a protein database. Such "blind" interpretation of mass spectra is prone to several types of errors, a few of which are highlighted here.

The monoisotopic peak is not visible in the spectrum for many large ions. The monoisotopic mass of the ion must then be inferred, typically by fitting the isotopic distribution to a model (*e.g.,* averagine[139]). However, analytes with elemental compositions that differ greatly from the model and noise in the intensity measurements of each isotope peak result in fitting errors; *i.e.,* an incorrect mass's theoretical distribution fits better than the correct mass's theoretical distribution. Errors in the monoisotopic mass inference occur from these incorrect fits and tend to scale in frequency and magnitude with analyte mass. These mass errors occur at multiples of approximately the mass of ^{13}C minus ^{12}C, because carbon is the dominant elemental component of proteins. Compensating for these mass errors by widening precursor or fragment mass tolerances increases the search space, which causes an increase in search time and FDR. Additionally, some modifications (*e.g.,* deamidation) or combinations of modifications (*e.g.*, ammonia loss and oxidation) result in mass differences similar to monoisotopic mass errors, further exacerbating the problem.

A major difficulty for many deconvolution programs is the ability to discern electronic and chemical noise from an analytes' true signal. Deconvolution algorithms routinely report masses that do not correspond to actual species, particularly when the spectrum quality is low. ProMex uses the proteoform elution profile to inform deconvolution and decrease the likelihood of random, non-reproducible masses being reported.[21] However, the reporting of non-reproducible masses is still a major problem. This issue is especially problematic for software that identifies proteoforms from intact mass alone because false-positives are more likely without fragmentation spectra that provide additional evidence of an identification.

Proteoform Databases in Top-Down Searches

Proteomic analysis often relies on database searching to identify proteins. The completeness and accuracy of the protein database used is essential for obtaining high quality results. Canonical protein sequence databases for many organisms can be obtained from sources such as UniProt[27], Ensembl[140] or RefSeq[141]. While these databases are useful starting points for proteoform identification, they can be incomplete and can lack PTM or sequence information. Proteoforms present in a sample that lack a corresponding theoretical proteoform in the search database are challenging to identify. Experimental proteoforms are often first compared to theoretical proteoforms within a mass tolerance. If no match is found, then the search can be widened to include theoretical proteoforms with a different mass. However, one is still left with the problem of interpreting the mass difference. This can be especially difficult because the mass difference can result from a combination of sequence insertions, deletions, substitutions, truncations, PTMs, and deconvolution errors. These problems can be partially alleviated by using the spectral alignment algorithms in tools such as MSAlign+[88], or by integrating proteogenomic data and bottom-up proteomic results to generate sample-specific databases.

Nucleic acid sequencing can reveal proteoform sequence changes that are absent in the canonical protein database. This field is known as proteogenomics. A general workflow for proteogenomic database generation follows: (i) Obtain nucleotide sequencing data. (ii) Align sequences to a reference genome if available or perform *de novo* alignment. (iii) Translate sequencing data to generate protein entries.[142–149] The nucleotide sequencing data can be obtained from whole genome sequencing, exome sequencing, or RNA sequencing data. There has recently been an effort to develop software tools that facilitate the integration of genomics tools with proteomic analyses, including Spritz, ProteomeGenerator[150] and the Galaxy-P System[151].

An important factor to be considered when utilizing proteogenomic workflows is FDR. Incorporating variants into a protein database increases its size, and FDR increases with the number of proteoforms that are in the database but not in the sample. It is crucial to only annotate confidently identified variants to maintain a comprehensive but not inflated proteogenomic database.[152]

Bottom-up data provides additional insights that can be used to improve proteoform identification. Novel PTMs identified in bottom-up can be annotated in a database for subsequent top-down analyses. For example, MetaMorpheus was recently used to generate an *E. coli* protein database annotated with novel PTMs.[102] This database was used in a subsequent Proteoform Suite analysis, enabling proteoforms containing these PTMs to be identified. Additionally, the database can be filtered to contain only proteins confidently identified by bottom-up analysis. This filtering limits the search space, which decreases false-positives.

VII. A Vison for the Future

In any projection forward, it behooves one to be mindful of the admonition often attributed to Yogi Berra: *"It's hard to make predictions, especially about the future."* Nonetheless, it is possible to make some observations about likely paths.

Proteoform analysis today is limited primarily to the more abundant and lower molecular weight proteins. This offers a clear technical challenge to the community, wherein the future of proteoform analysis can evolve along two distinct axes: mass spectrometric, and "other". In the arena of mass spectrometry, the standard analytical metrics of resolution, sensitivity, and speed all need to be improved for large proteins (e.g. >50 kDa). It is noteworthy that the fundamental limits of mass spectrometry are not at issue – there is no reason in principle that the accurate mass of a single ion of a large macromolecule cannot be measured. However, improvements in ionization sources, ion transfer efficiencies, and detector sensitivity are all needed to implement this on a routine basis for complex mixtures. One can imagine that one day, individual proteoform molecules will each have their accurate masses determined in rapid succession, simply counting them to determine their abundance.

It is also apparent that the gradual accrual of knowledge about proteoforms and proteoform families, immortalized in a comprehensive database or "proteoform atlas", will allow the much more rapid and effective identification of proteoforms in the future. Proteomics can move from a "discovery" mode, involving complex data generation and interpretation, to a "scoring" mode, where proteoforms detected are matched up with members of a comprehensive proteoform database tailored for the sample under study. This strategy benefits greatly from a comprehensive approach to proteomics that integrates disparate data sources such as nucleotide sequence data, deep bottom-up data, and knowledge repositories such as UniProt and the CTDP Proteoform Repository.

Testing the biological consequences of individual proteoforms presents a remarkable challenge. Many contend that proteoforms are the ultimate biological actor. Yet, how one might introduce specific proteoforms into a biological context, or more challenging still, deliberately alter their concentration or localization for the purpose of understanding their role is still unclear. A grand challenge for the community would be to develop the power to synthesize specific proteoforms at will, fully defined with respect to PTM localization and amino acid backbone, and to introduce them into living systems. Even more challenging would be to be able to express such molecules at pre-defined locations (e.g., the cell nucleus or the mitochondria) and/or times. Such capabilities would provide a tool of unprecedented power to reveal the functions of proteoforms.

On a longer horizon, a variety of exciting new single-molecule analysis platforms are actively under development around the world, including nanoscale cantilevers[153], nanopore strategies[154], interferometric light scattering[155], cryo-electron microscopy[156], x-ray scattering[157] and others. The yet-to-be-determined and evolving strengths and limitations of such new strategies for proteoform identification will dictate how these approaches supplant or synergize with today's technologies in ways that we cannot presently foresee.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was funded by the National Institute of General Medical Sciences of the National Institutes of Health (NIH) under Award Numbers R35GM126914 and R01GM114292 (to LMS). LVS was supported by the NIH Biotechnology Training Program, T32GM008349. RJM was supported by the NIH Genomic Sciences Training Program, 5T32HG002760. RMM and TT were supported by the NIH Chemistry Biology Interface Training Program, T32GM008505. LS was supported by NIH Grant R01GM125991. XL was supported by NIH Grants R01GM118470 and R01GM125991. ZW is supported by R01 GM117058, R01 GM125085. YG would like to acknowledge NIH grants, R01HL109810, R01HL096971, R01GM117058, R01GM125085, and S100D018475. LCA was supported by NSF/DMR-1644779. RTK, NLK, RDL, and PMT acknowledge P41GM108569. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Abbreviations:

SDK	software development kit
PSM	peptide spectrum match
PrSM	proteoform spectrum match
HIC	hydrophobic interaction chromatography
IEX	ion exchange chromatography
GELFrEE	gel-eluted liquid fraction entrapment electrophoresis
sSEC	serial size exclusion chromatography
UPLC	ultrahigh pressure liquid chromatography
LFQ	label-free quantification
TMT	tandem mass tag
G-PTM-D	global PTM discovery

References

[1]. Smith LM, Kelleher NL, Consortium for Top Down Proteomics Nat Methods. 2013, 10, 186–187. [PubMed: 23443629]

[2]. Shortreed MR, Frey BL, Scalf M, Knoener RA, Cesnik AJ, Smith LM J Proteome Res. 2016, 15, 1213–1221. [PubMed: 26941048]

[3]. Santos-Rosa H, Kirmizis A, Nelson C, Bartke T, Saksouk N, Cote J, Kouzarides T Nat Struct Mol Biol. 2009, 16, 17–22. [PubMed: 19079264]

[4]. Mylona A, Theillet FX, Foster C, Cheng TM, Miralles F, Bates PA, Selenko P, Treisman R Science. 2016, 354, 233–237. [PubMed: 27738173]

[5]. Aebersold R, Mann M Nature. 2003, 422, 198–207. [PubMed: 12634793]

[6]. Han X, Aslanian A, Yates JR 3rd Curr Opin Chem Biol. 2008, 12, 483–490. [PubMed: 18718552]

[7]. Catherman AD, Skinner OS, Kelleher NL Biochem Biophys Res Commun. 2014, 445, 683–693. [PubMed: 24556311]

[8]. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM, Yates JR 3rd Nature biotechnology. 1999, 17, 676–682.

[9]. Nesvizhskii AI, Aebersold R Mol Cell Proteomics. 2005, 4, 1419–1440. [PubMed: 16009968]

[10]. Siuti N, Kelleher NL. Nat Methods 2007, 4, 817–821. [PubMed: 17901871]

[11]. Chen B, Brown KA, Lin Z, Ge Y Analytical chemistry. 2018, 90, 110–127. [PubMed: 29161012]

[12]. Cai W, Tucholski TM, Gregorich ZR, Ge Y Expert Rev Proteomics. 2016, 13, 717–730. [PubMed: 27448560]

[13]. Toby TK, Fornelli L, Kelleher NL Annu Rev Anal Chem (Palo Alto Calif). 2016, 9, 499–519. [PubMed: 27306313]

[14]. Armirotti A, Damonte G. Proteomics 2010, 10, 3566–3576. [PubMed: 20859958]

[15]. Gregorich ZR, Ge Y Proteomics. 2014, 14, 1195–1210. [PubMed: 24723472]

[16]. Compton PD, Zamdborg L, Thomas PM, Kelleher NL Analytical chemistry. 2011, 83, 6868–6874. [PubMed: 21744800]

[17]. Anderson LC, DeHart CJ, Kaiser NK, Fellers RT, Smith DF, Greer JB, LeDuc RD, Blakney GT, Thomas PM, Kelleher NL, Hendrickson CL J Proteome Res. 2017, 16, 1087–1096. [PubMed: 27936753]

[18]. McCool EN, Lubeckyj RA, Shen X, Chen D, Kou Q, Liu X, Sun L Analytical chemistry. 2018, 90, 5529–5533. [PubMed: 29620868]

[19]. Tran JC, Zamdborg L, Ahlf DR, Lee JE, Catherman AD, Durbin KR, Tipton JD, Vellaichamy A, Kellie JF, Li M, Wu C, Sweet SM, Early BP, Siuti N, LeDuc RD, Compton PD, Thomas PM, Kelleher NL Nature. 2011, 480, 254–258. [PubMed: 22037311]

[20]. Toby TK, Fornelli L, Srzentic K, DeHart CJ, Levitsky J, Friedewald J, Kelleher NL Nat Protoc. 2019, 14, 119–152. [PubMed: 30518910]

[21]. Park J, Piehowski PD, Wilkins C, Zhou M, Mendoza J, Fujimoto GM, Gibbons BC, Shaw JB, Shen Y, Shukla AK, Moore RJ, Liu T, Petyuk VA, Tolic N, Pasa-Tolic L, Smith RD, Payne SH, Kim S Nat Methods. 2017, 14, 909–914. [PubMed: 28783154]

[22]. Kou Q, Xun L, Liu X Bioinformatics. 2016, 32, 3495–3497. [PubMed: 27423895]

[23]. Cesnik AJ, Shortreed MR, Schaffer LV, Knoener RA, Frey BL, Scalf M, Solntsev SK, Dai Y, Gasch AP, Smith LM J Proteome Res. 2018, 17, 568–578. [PubMed: 29195273]

[24]. Cai W, Guner H, Gregorich ZR, Chen AJ, Ayaz-Guner S, Peng Y, Valeja SG, Liu X, Ge Y Mol Cell Proteomics. 2016, 15, 703–714. [PubMed: 26598644]

[25]. Solntsev SK, Shortreed MR, Frey BL, Smith LM J Proteome Res. 2018, 17, 1844–1851. [PubMed: 29578715]

[26]. Legrain P, Aebersold R, Archakov A, Bairoch A, Bala K, Beretta L, Bergeron J, Borchers CH, Corthals GL, Costello CE, Deutsch EW, Domon B, Hancock W, He F, Hochstrasser D, Marko-Varga G, Salekdeh GH, Sechi S, Snyder M, Srivastava S, Uhlen M, Wu CH, Yamamoto T, Paik YK, Omenn GS Mol Cell Proteomics. 2011, 10, M111 009993.

[27]. UniProt Consortium Nucleic Acids Res. 2015, 43, D204–212.

[28]. LeDuc RD, Schwammle V, Shortreed MR, Cesnik AJ, Solntsev SK, Shaw JB, Martin MJ, Vizcaino JA, Alpi E, Danis P, Kelleher NL, Smith LM, Ge Y, Agar JN, Chamot-Rooke J, Loo JA, Pasa-Tolic L, Tsybin YO J Proteome Res. 2018, 17, 1321–1325. [PubMed: 29397739]

[29]. Doucette AA, Tran JC, Wall MJ, Fitzsimmons S Expert Rev Proteomics. 2011, 8, 787–800. [PubMed: 22087661]

[30]. Valeja SG, Xiu L, Gregorich ZR, Guner H, Jin S, Ge Y Analytical chemistry. 2015, 87, 5363–5371. [PubMed: 25867201]

[31]. Chen B, Peng Y, Valeja SG, Xiu L, Alpert AJ, Ge Y Analytical chemistry. 2016, 88, 1885–1891. [PubMed: 26729044]

[32]. Muneeruddin K, Nazzaro M, Kaltashov IA Analytical chemistry. 2015, 87, 10138–10145. [PubMed: 26360183]

[33]. Nshanian M, Lantz C, Wongkongkathep P, Schrader T, Klarner FG, Blumke A, Despres C, Ehrmann M, Smet-Nocca C, Bitan G, Loo JA J Am Soc Mass Spectrom. 2019, 30, 16–23. [PubMed: 30062477]

[34]. Zinnel NF, Pai PJ, Russell DH Analytical chemistry. 2012, 84, 3390–3397. [PubMed: 22455956]

[35]. Riley NM, Mullen C, Weisbrod CR, Sharma S, Senko MW, Zabrouskov V, Westphall MS, Syka JE, Coon JJ J Am Soc Mass Spectrom. 2016, 27, 520–531. [PubMed: 26589699]

[36]. Tran JC, Doucette AA Analytical chemistry. 2008, 80, 1568–1573. [PubMed: 18229945]

[37]. Lee JE, Kellie JF, Tran JC, Tipton JD, Catherman AD, Thomas HM, Ahlf DR, Durbin KR, Vellaichamy A, Ntai I, Marshall AG, Kelleher NL J Am Soc Mass Spectrom. 2009, 20, 2183–2191. [PubMed: 19747844]

[38]. Kellie JF, Catherman AD, Durbin KR, Tran JC, Tipton JD, Norris JL, Witkowski, CE 2nd, Thomas PM, Kelleher NL Analytical chemistry. 2012, 84, 209–215. [PubMed: 22103811]

[39]. Skinner OS, Do Vale LH, Catherman AD, Havugimana PC, de Sousa MV, Compton PD, Kelleher NL Analytical chemistry. 2015, 87, 3032–3038. [PubMed: 25664979]

[40]. Skinner OS, Havugimana PC, Haverland NA, Fornelli L, Early BP, Greer JB, Fellers RT, Durbin KR, Do Vale LH, Melani RD, Seckler HS, Nelp MT, Belov ME, Horning SR, Makarov AA, LeDuc RD, Bandarian V, Compton PD, Kelleher NL Nat Methods. 2016, 13, 237–240. [PubMed: 26780093]

[41]. Chen X, Ge Y Proteomics. 2013, 13, 2563–2566. [PubMed: 23794208]

[42]. Speers AE, Wu CC Chem Rev. 2007, 107, 3687–3714. [PubMed: 17683161]

[43]. Yu YQ, Gilar M, Lee PJ, Bouvier ES, Gebler JC Analytical chemistry. 2003, 75, 6023–6028. [PubMed: 14588046]

[44]. Meng F, Cargile BJ, Patrie SM, Johnson JR, McLoughlin SM, Kelleher NL Analytical chemistry. 2002, 74, 2923–2929. [PubMed: 12141648]

[45]. Saveliev SV, Woodroofe CC, Sabat G, Adams CM, Klaubert D, Wood K, Urh M Analytical chemistry. 2013, 85, 907–914. [PubMed: 23256507]

[46]. Chang YH, Gregorich ZR, Chen AJ, Hwang L, Guner H, Yu D, Zhang J, Ge Y J Proteome Res. 2015, 14, 1587–1599. [PubMed: 25589168]

[47]. Cai W, Tucholski T, Chen B, Alpert AJ, Mcllwain S, Kohmoto T, Jin S, Ge Y Analytical chemistry. 2017, 89, 5467–5475. [PubMed: 28406609]

[48]. Hong P, Koza S, Bouvier ES J Liq Chromatogr Relat Technol. 2012, 35, 2923–2950. [PubMed: 23378719]

[49]. Alpert A, in J Size exclusion high-performance liquid chromatography of small solutes, Vol., Academic Press, San Diego, CA, 1999, pp.249–266.

[50]. Tucholski T, Knott SJ, Chen B, Pistono P, Lin Z, Ge Y Analytical chemistry. 2019, 91, 3835–3844. [PubMed: 30758949]

[51]. Zhang Z, Wu S, Stenoien DL, Pasa-Tolic L Annu Rev Anal Chem (Palo Alto Calif). 2014, 7, 427–454. [PubMed: 25014346]

[52]. Ansong C, Wu S, Meng D, Liu X, Brewer HM, Deatherage Kaiser BL, Nakayasu ES, Cort JR, Pevzner P, Smith RD, Heffron F, Adkins JN, Pasa-Tolic L Proc Natl Acad Sci U S A. 2013, 110, 10153–10158. [PubMed: 23720318]

[53]. Shishkova E, Hebert AS, Westphall MS, Coon JJ Analytical chemistry. 2018, 90, 11503–11508. [PubMed: 30179449]

[54]. Camerini S, Mauri P J Chromatogr A. 2015, 1381, 1–12. [PubMed: 25618357]

[55]. Wahab MF, Patel DC, Wimalasinghe RM, Armstrong DW Analytical chemistry. 2017, 89, 8177–8191. [PubMed: 28699732]

[56]. Kennedy RT, Jorgenson JW Analytical chemistry. 1989, 61, 1128–1135.

[57]. Wang Z, Ma H, Smith K, Wu S International Journal of Mass Spectrometry. 2018, 427, 43–51. [PubMed: 31097918]

[58]. Shen Y, Tolic N, Piehowski PD, Shukla AK, Kim S, Zhao R, Qu Y, Robinson E, Smith RD, Pasa-Tolic L J Chromatogr A. 2017, 1498, 99–110. [PubMed: 28077236]

[59]. Zhe Wang XL, Muther Jennifer, James Judith A., Smith Kenneth, Wu Si Scientific Reports. 2019, in press.

[60]. Guo D, Mant CT, Taneja AK, Hodges RS Journal of Chromatography A. 1986, 359, 519–532.

[61]. Requiao RD, Fernandes L, de Souza HJA, Rossetto S, Domitrovic T, Palhano FL PLoS Comput Biol. 2017, 13, e1005549.

[62]. Chow CC, Chow C, Raghunathan V, Huppert TJ, Kimball EB, Cavagnero S Biochemistry. 2003, 42, 7090–7099. [PubMed: 12795605]

[63]. Jha AK, Colubri A, Zaman MH, Koide S, Sosnick TR, Freed KF Biochemistry. 2005, 44, 9691–9702. [PubMed: 16008354]

[64]. Moini M Analytical chemistry. 2007, 79, 4241–4246. [PubMed: 17447730]

[65]. Wojcik R, Dada OO, Sadilek M, Dovichi NJ Rapid Commun Mass Spectrom. 2010, 24, 2554–2560. [PubMed: 20740530]

[66]. Sun L, Zhu G, Zhang Z, Mou S, Dovichi NJ J Proteome Res. 2015, 14, 2312–2321. [PubMed: 25786131]

[67]. Zhu G, Sun L, Dovichi NJ Talanta. 2016, 146, 839–843. [PubMed: 26695337]

[68]. Aebersold R, Morrison HD J Chromatogr. 1990, 516, 79–88. [PubMed: 2286630]

[69]. Britz-McKibbin P, Chen DD Analytical chemistry. 2000, 72, 1242–1252. [PubMed: 10740866]

[70]. Chen D, Shen X, Sun L Analyst. 2017, 142, 2118–2127. [PubMed: 28513658]

[71]. Valaskovic GA, Kelleher NL, McLafferty FW Science. 1996, 273, 1199–1202. [PubMed: 8703047]

[72]. Han X, Wang Y, Aslanian A, Bern M, Lavallee-Adam M, Yates JR 3rd Analytical chemistry. 2014, 86, 11006–11012. [PubMed: 25346219]

[73]. Han X, Wang Y, Aslanian A, Fonslow B, Graczyk B, Davis TN, Yates JR 3rd J Proteome Res. 2014, 13, 6078–6086. [PubMed: 25382489]

[74]. Zhao Y, Sun L, Zhu G, Dovichi NJ J Proteome Res. 2016, 15, 3679–3685. [PubMed: 27490796]

[75]. Li Y, Compton PD, Tran JC, Ntai I, Kelleher NL Proteomics. 2014, 14, 1158–1164. [PubMed: 24596178]

[76]. Lubeckyj RA, McCool EN, Shen X, Kou Q, Liu X, Sun L Analytical chemistry. 2017, 89, 12059–12067. [PubMed: 29064224]

[77]. McCool EN, Lubeckyj R, Shen X, Kou Q, Liu X, Sun L J Vis Exp. 2018.

[78]. Shen X, Kou Q, Guo R, Yang Z, Chen D, Liu X, Hong H, Sun L Analytical chemistry. 2018, 90, 10095–10099. [PubMed: 30085653]

[79]. LeDuc RD, Taylor GK, Kim YB, Januszyk TE, Bynum LH, Sola JV, Garavelli JS, Kelleher NL Nucleic Acids Res. 2004, 32, W340–345. [PubMed: 15215407]

[80]. Zamdborg L, LeDuc RD, Glowacz KJ, Kim YB, Viswanathan V, Spaulding IT, Early BP, Bluhm EJ, Babai S, Kelleher NL Nucleic Acids Res. 2007, 35, W701–706. [PubMed: 17586823]

[81]. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A Genome Res. 2005, 15, 1451–1455. [PubMed: 16169926]

[82]. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J Curr Protoc Mol Biol. 2010, Chapter 19, Unit 19 10 11–21. [PubMed: 20069539]

[83]. Goecks J, Nekrutenko A, Taylor J, Galaxy T Genome Biol. 2010, 11, R86. [PubMed: 20738864]

[84]. Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI Nat Methods. 2017, 14, 513–520. [PubMed: 28394336]

[85]. Liu X, Mammana A, Bafna Bioinformatics V.. 2012, 28, 1692–1697.

[86]. Pevzner PA, Dancik V, Tang CL J Comput Biol. 2000, 7, 777–787. [PubMed: 11382361]

[87]. Frank AM, Pesavento JJ, Mizzen CA, Kelleher NL, Pevzner PA Analytical chemistry. 2008, 80, 2499–2505. [PubMed: 18302345]

[88]. Liu X, Sirotkin Y, Shen Y, Anderson G, Tsai YS, Ting YS, Goodlett DR, Smith RD, Bafna V, Pevzner PA Mol Cell Proteomics. 2012, 11, M111 008524.

[89]. Liu X, Segar MW, Li SC, Kim S BMC Genomics. 2014, 15 Suppl 1, S9.

[90]. Kou Q, Zhu B, Wu S, Ansong C, Tolic N, Pasa-Tolic L, Liu X J Proteome Res. 2016, 15, 2422–2432. [PubMed: 27291504]

[91]. Chick JM, Kolippakkam D, Nusinow DP, Zhai B, Rad R, Huttlin EL, Gygi SP Nature biotechnology. 2015, 33, 743–749.

[92]. Lipman DJ, Pearson WR Science. 1985, 227, 1435–1441. [PubMed: 2983426]

[93]. Kou Q, Wu S, Tolic N, Pasa-Tolic L, Liu Y, Liu X Bioinformatics. 2017, 33, 1309–1316. [PubMed: 28453668]

[94]. Kou Q, Wu S, Liu X Proteomics. 2018, 18, 878–889.

[95]. Kou Q, Wang Z, Lubeckyj RA, Wu S, Sun L, Liu X J Proteome Res. 2019.

[96]. Durbin KR, Fellers RT, Ntai I, Kelleher NL, Compton PD Analytical chemistry. 2014, 86, 1485–1492. [PubMed: 24400813]

[97]. Durbin KR, Tran JC, Zamdborg L, Sweet SM, Catherman AD, Lee JE, Li M, Kellie JF, Kelleher NL Proteomics. 2010, 10, 3589–3597. [PubMed: 20848673]

[98]. Schaffer LV, Shortreed MR, Cesnik AJ, Frey BL, Solntsev SK, Scalf M, Smith LM Analytical chemistry. 2018, 90, 1325–1333. [PubMed: 29227670]

[99]. Schaffer LV, Rensvold JW, Shortreed MR, Cesnik AJ, Jochem A, Scalf M, Frey BL, Pagliarini DJ, Smith LM J Proteome Res. 2018, 17, 3526–3536. [PubMed: 30180576]

[100]. Elias JE, Gygi SP Nat Methods. 2007, 4, 207–214. [PubMed: 17327847]

[101]. Li Q, Shortreed MR, Wenger CD, Frey BL, Schaffer LV, Scalf M, Smith LM J Proteome Res. 2017, 16, 1383–1390. [PubMed: 28248113]

[102]. Dai Y, Shortreed MR, Scalf M, Frey BL, Cesnik AJ, Solntsev S, Schaffer LV, Smith LM J Proteome Res. 2017, 16, 4156–4165. [PubMed: 28968100]

[103]. Lin Z, Guo F, Gregorich ZR, Sun R, Zhang H, Hu Y, Shanmuganayagam D, Ge Y J Am Soc Mass Spectrom. 2018, 29, 1284–1294. [PubMed: 29633223]

[104]. Jin Y, Peng Y, Lin Z, Chen YC, Wei L, Hacker TA, Larsson L, Ge Y J Muscle Res Cell Motil. 2016, 37, 41–52. [PubMed: 27090236]

[105]. Ntai I, Toby TK, LeDuc RD, Kelleher NL Methods Mol Biol. 2016, 1410, 121–133. [PubMed: 26867742]

[106]. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M Mol Cell Proteomics. 2002, 1, 376–386. [PubMed: 12118079]

[107]. Thompson A, Schafer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, Neumann T, Johnstone R, Mohammed AK, Hamon C Analytical chemistry. 2003, 75, 1895–1904. [PubMed: 12713048]

[108]. Capriotti AL, Cavaliere C, Foglia P, Samperi R, Lagana A J Chromatogr A. 2011, 1218, 8760–8776. [PubMed: 21689823]

[109]. Kelleher NL, Thomas PM, Ntai I, Compton PD, LeDuc RD Expert Rev Proteomics. 2014, 11, 649–651. [PubMed: 25347991]

[110]. Ntai I, Kim K, Fellers RT, Skinner OS, Smith A. D. t., Early BP, Savaryn JP, LeDuc RD, Thomas PM, Kelleher NL Analytical chemistry. 2014, 86, 4961–4968. [PubMed: 24807621]

[111]. Davis RG, Park HM, Kim K, Greer JB, Fellers RT, LeDuc RD, Romanova EV, Rubakhin SS, Zombeck JA, Wu C, Yau PM, Gao P, van Nispen AJ, Patrie SM, Thomas PM, Sweedler JV, Rhodes JS, Kelleher NL Analytical chemistry. 2018, 90, 3802–3810. [PubMed: 29481055]

[112]. Durbin KR, Fornelli L, Fellers RT, Doubleday PF, Narita M, Kelleher NL J Proteome Res. 2016, 15, 976–982. [PubMed: 26795204]

[113]. Ntai I, Fornelli L, DeHart CJ, Hutton JE, Doubleday PF, LeDuc RD, van Nispen AJ, Fellers RT, Whiteley G, Boja ES, Rodriguez H, Kelleher NL Proc Natl Acad Sci U S A. 2018, 115, 4140–4145. [PubMed: 29610327]

[114]. Seckler HDS, Fornelli L, Mutharasan RK, Thaxton CS, Fellers R, Daviglus M, Sniderman A, Rader D, Kelleher NL, Lloyd-Jones DM, Compton PD, Wilkins JT J Proteome Res. 2018, 17, 2156–2164. [PubMed: 29649363]

[115]. Waanders LF, Hanke S, Mann M J Am Soc Mass Spectrom. 2007, 18, 2058–2064. [PubMed: 17920290]

[116]. Merrill AE, Hebert AS, MacGilvray ME, Rose CM, Bailey DJ, Bradley JC, Wood WW, El Masri M, Westphall MS, Gasch AP, Coon JJ Mol Cell Proteomics. 2014, 13, 2503–2512. [PubMed: 24938287]

[117]. Potts GK, Voigt EA, Bailey DJ, Rose CM, Westphall MS, Hebert AS, Yin J, Coon JJ Analytical chemistry. 2016, 88, 3295–3303. [PubMed: 26882330]

[118]. Rhoads TW, Rose CM, Bailey DJ, Riley NM, Molden RC, Nestler AJ, Merrill AE, Smith LM, Hebert AS, Westphall MS, Pagliarini DJ, Garcia BA, Coon JJ Analytical chemistry. 2014, 86, 2314–2319. [PubMed: 24475910]

[119]. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, Purkayastha S, Juhasz P, Martin S, Bartlet-Jones M, He F, Jacobson A, Pappin DJ Mol Cell Proteomics. 2004, 3, 1154–1169. [PubMed: 15385600]

[120]. O'Connell JD, Paulo JA, O'Brien JJ, Gygi SP J Proteome Res. 2018, 17, 1934–1942. [PubMed: 29635916]

[121]. Keshishian H, Burgess MW, Specht H, Wallace L, Clauser KR, Gillette MA, Carr SA Nat Protoc. 2017, 12, 1683–1701. [PubMed: 28749931]

[122]. Sinclair J, Timms JF Methods. 2011, 54, 361–369. [PubMed: 21397697]

[123]. Wiese S, Reidegeld KA, Meyer HE, Warscheid B Proteomics. 2007, 7, 340–350. [PubMed: 17177251]

[124]. Prudova A, auf dem Keller U, Butler GS, Overall CM Mol Cell Proteomics. 2010, 9, 894–911. [PubMed: 20305284]

[125]. Hung CW, Tholey A Analytical chemistry. 2012, 84, 161–170. [PubMed: 22103715]

[126]. LeDuc RD, Fellers RT, Early BP, Greer JB, Shams DP, Thomas P, Kelleher NL Mol Cell Proteomics. 2019.

[127]. LeDuc RD, Fellers RT, Early BP, Greer JB, Thomas PM, Kelleher NL J Proteome Res. 2014, 13, 3231–3240. [PubMed: 24922115]

[128]. Chait BT Science. 2006, 314, 65–66. [PubMed: 17023639]

[129]. Moore SM, Hess SM, Jorgenson JW J Proteome Res. 2016, 15, 1243–1252. [PubMed: 26979493]

[130]. Stephenson JL Jr., McLuckey SA Analytical chemistry. 1998, 70, 3533–3544. [PubMed: 9737205]

[131]. Chrisman PA, Pitteri SJ, McLuckey SA Analytical chemistry. 2006, 78, 310–316. [PubMed: 16383342]

[132]. Anderson LC, Karch KR, Ugrin SA, Coradin M, English AM, Sidoli S, Shabanowitz J, Garcia BA, Hunt DF Mol Cell Proteomics. 2016, 15, 975–988. [PubMed: 26785730]

[133]. Foreman DJ, Dziekonski ET, McLuckey SA J Am Soc Mass Spectrom. 2019, 30, 34–44. [PubMed: 29713964]

[134]. Holden DD, Sanders JD, Weisbrod CR, Mullen C, Schwartz JC, Brodbelt JS Analytical chemistry. 2018, 90, 8583–8591. [PubMed: 29927232]

[135]. Horn DM, Zubarev RA, McLafferty FW J Am Soc Mass Spectrom. 2000, 11, 320–332. [PubMed: 10757168]

[136]. Ferrige AG, Seddon MJ, Jarvis S, Skilling J, Aplin R Rapid Communications in Mass Spectrometry. 1991, 5, 374–377.

[137]. Liu X, Inbar Y, Dorrestein PC, Wynne C, Edwards N, Souda P, Whitelegge JP, Bafna V, Pevzner PA Mol Cell Proteomics. 2010, 9, 2772–2782. [PubMed: 20855543]

[138]. Marty MT, Baldwin AJ, Marklund EG, Hochberg GK, Benesch JL, Robinson CV Analytical chemistry. 2015, 87, 4370–4376. [PubMed: 25799115]

[139]. Senko MW, Beu SC, McLaffertycor FW J Am Soc Mass Spectrom. 1995, 6, 229–233. [PubMed: 24214167]

[140]. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigo R, Hubbard TJ Genome Res. 2012, 22, 1760–1774. [PubMed: 22955987]

[141]. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, Murphy MR, O'Leary NA, Pujar S, Rajput B, Rangwala SH, Riddick LD, Shkeda A, Sun H, Tamez P, Tully RE, Wallin C, Webb D, Weber J, Wu W, DiCuccio M, Kitts P, Maglott DR, Murphy TD, Ostell JM Nucleic Acids Res. 2014, 42, D756–763. [PubMed: 24259432]

[142]. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR Bioinformatics. 2013, 29, 15–21. [PubMed: 23104886]

[143]. Trapnell C, Pachter L, Salzberg SL Bioinformatics. 2009, 25, 1105–1111. [PubMed: 19289445]

[144]. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL Genome Biol. 2013, 14, R36. [PubMed: 23618408]

[145]. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA Genome Res. 2010, 20, 1297–1303. [PubMed: 20644199]

[146]. Langmead B, Salzberg SL Nat Methods. 2012, 9, 357–359. [PubMed: 22388286]

[147]. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL Nature biotechnology. 2015, 33, 290–295.

[148]. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L Nat Protoc. 2012, 7, 562–578. [PubMed: 22383036]

[149]. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu Front X Genet. 2012, 3, 35.

[150]. Cifani P, Dhabaria A, Chen Z, Yoshimi A, Kawaler E, Abdel-Wahab O, Poirier JT, Kentsis A J Proteome Res. 2018, 17, 3681–3692. [PubMed: 30295032]

[151]. Sheynkman GM, Johnson JE, Jagtap PD, Shortreed MR, Onsongo G, Frey BL, Griffin TJ, Smith LM BMC Genomics. 2014, 15, 703. [PubMed: 25149441]

[152]. Nesvizhskii AI J Proteomics. 2010, 73, 2092–2123. [PubMed: 20816881]

[153]. Hanay MS, Kelber S, Naik AK, Chi D, Hentz S, Bullard EC, Colinet E, Duraffourg L, Roukes ML Nat Nanotechnol. 2012, 7, 602–608. [PubMed: 22922541]

[154]. Chinappi M, Cecconi F J Phys Condens Matter. 2018, 30, 204002.

[155]. Young G, Hundt N, Cole D, Fineberg A, Andrecka J, Tyler A, Olerinyova A, Ansari A, Marklund EG, Collier MP, Chandler SA, Tkachenko O, Allen J, Crispin M, Billington N, Takagi Y, Sellers JR, Eichmann C, Selenko P, Frey L, Riek R, Galpin MR, Struwe WB, Benesch JLP, Kukura P Science. 2018, 360, 423–427. [PubMed: 29700264]

[156]. Fernandez-Leiro R, Scheres SH Nature. 2016, 537, 339–346. [PubMed: 27629640]

[157]. von Ardenne B, Mechelke M, Grubmuller H Nat Commun. 2018, 9, 2375. [PubMed: 29915244]

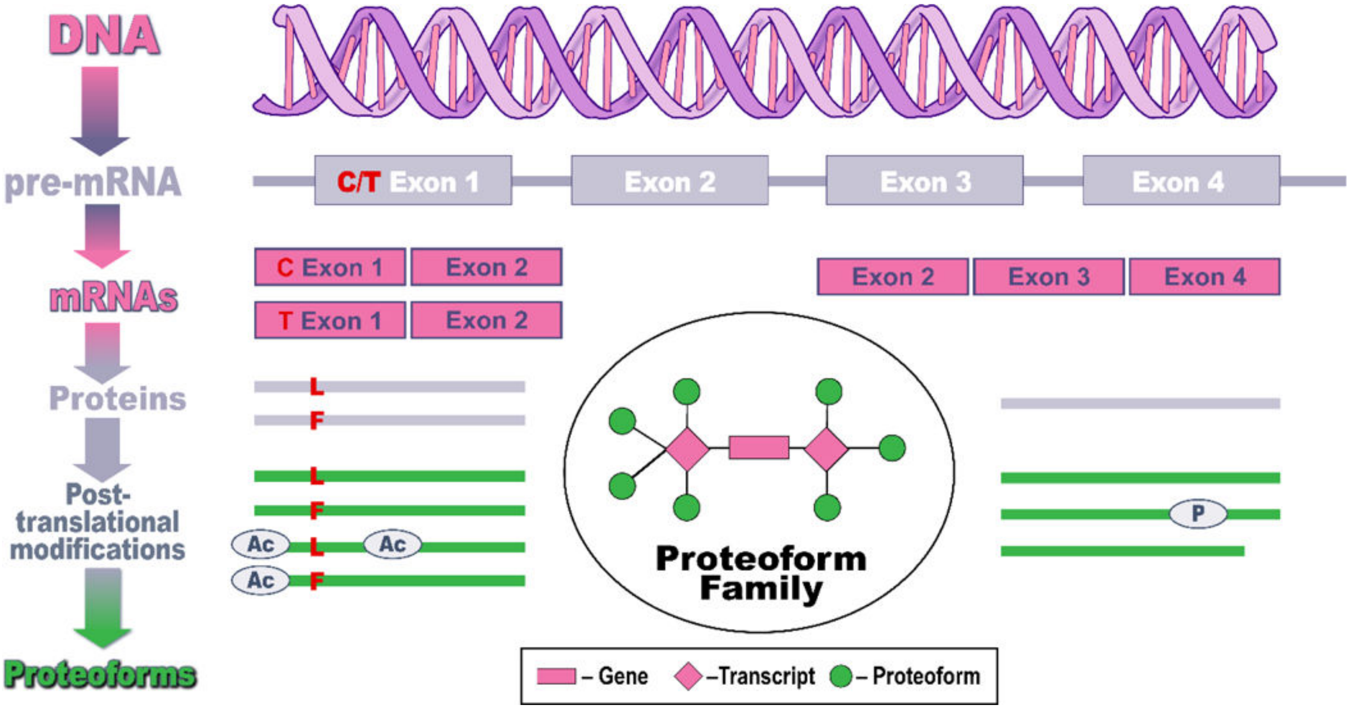

Figure 1.

The sources of variation resulting in different proteoforms. Different proteoforms arising from the same gene make up a proteoform family.

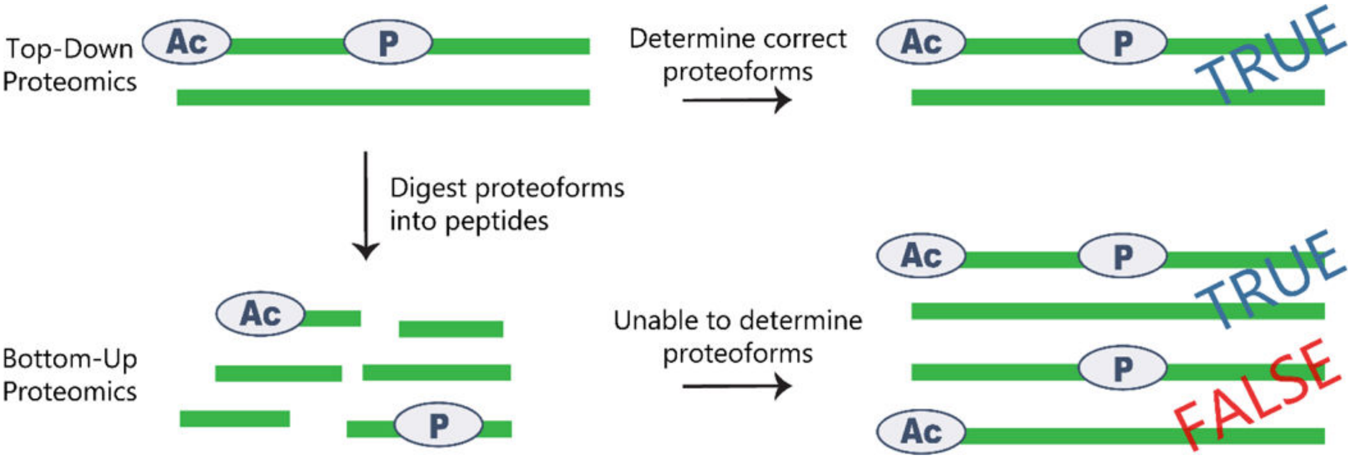

Figure 2.

Top-down and bottom-up proteomics. In bottom-up proteomics, proteoforms are digested into peptides, resulting in loss of information about the original proteoform sequence and modifications.

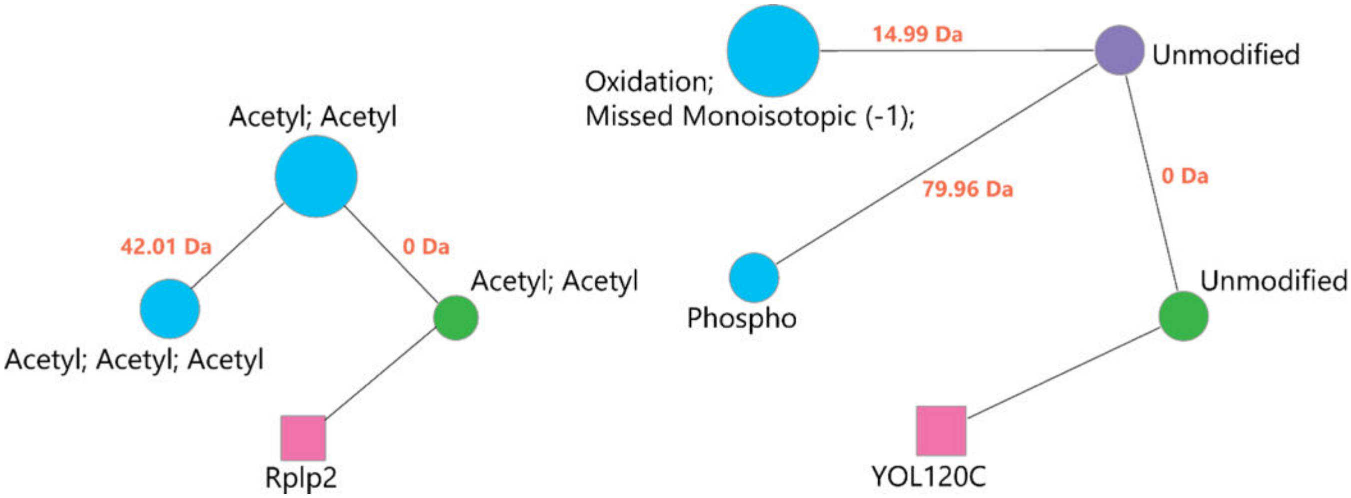

Figure 3.

The yeast genes *Rplp2* and *YOL120C* proteoform families visualized as a network of related proteoforms. Each circle represents a unique proteoform, including theoretical proteoforms (green), experimental proteoforms identified by MS/MS (purple), and experimental proteoforms observed in the MS1 spectra but unidentified by MS/MS (blue). Lines connecting circles represent mass differences corresponding to modifications. The size of blue circles is proportional to the integrated ion intensity.

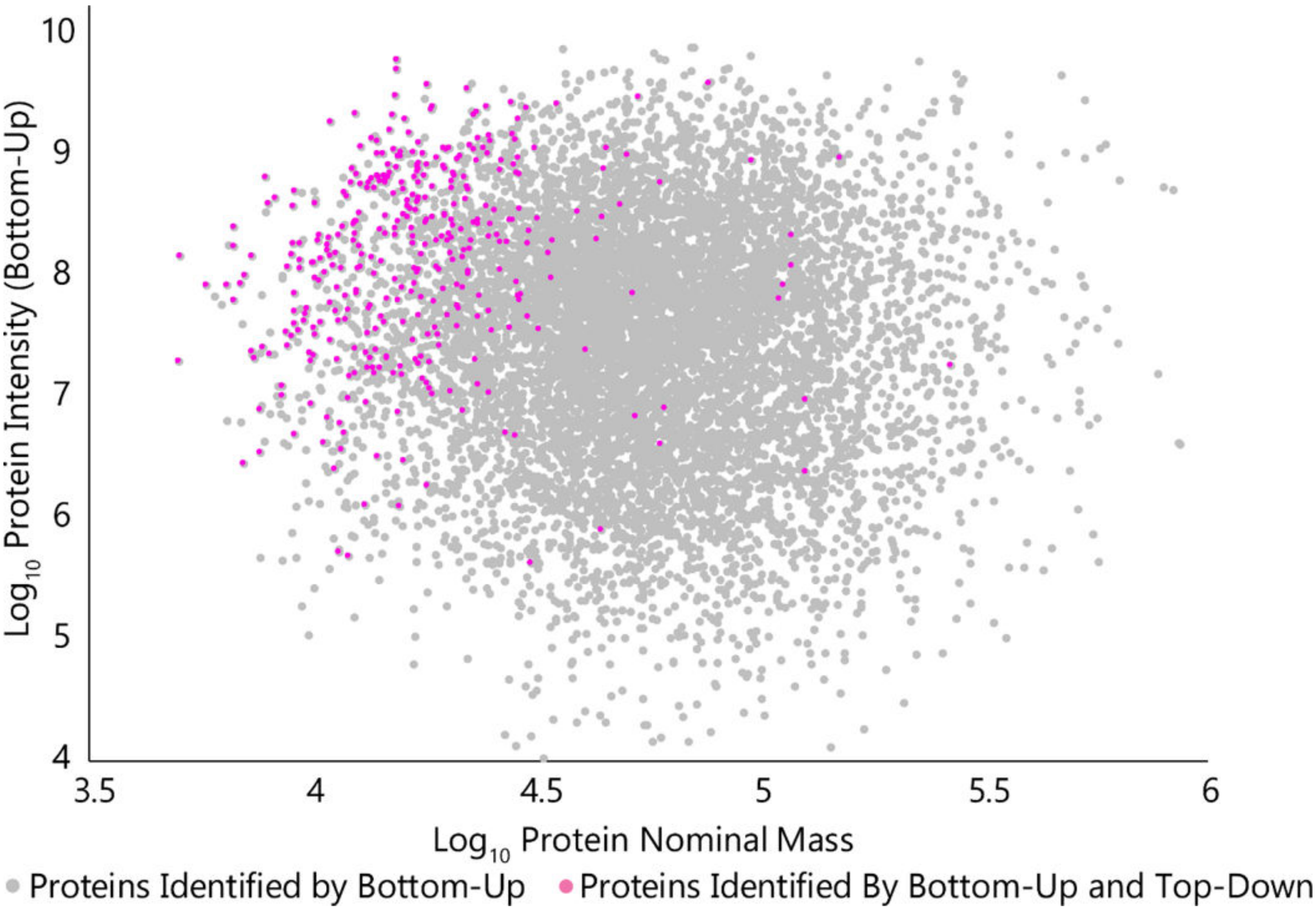

Figure 4.

A plot of $\log_{10}$ intensity vs. $\log_{10}$ nominal mass for human Jurkat proteins identified using bottom-up proteomics. Proteins also identified by top-down analysis are marked in pink. Protein “nominal mass” is the mass of the full-length unmodified protein sequence from UniProt. For bottom-up, proteins are inferred from peptide sequences by MetaMorpheus; for top-down, proteins are identified from either the full-length sequence or a subsequence with TDPortal. The subset of proteins identified by top-down corresponds to the low molecular weight and highly abundant subset of the proteome.

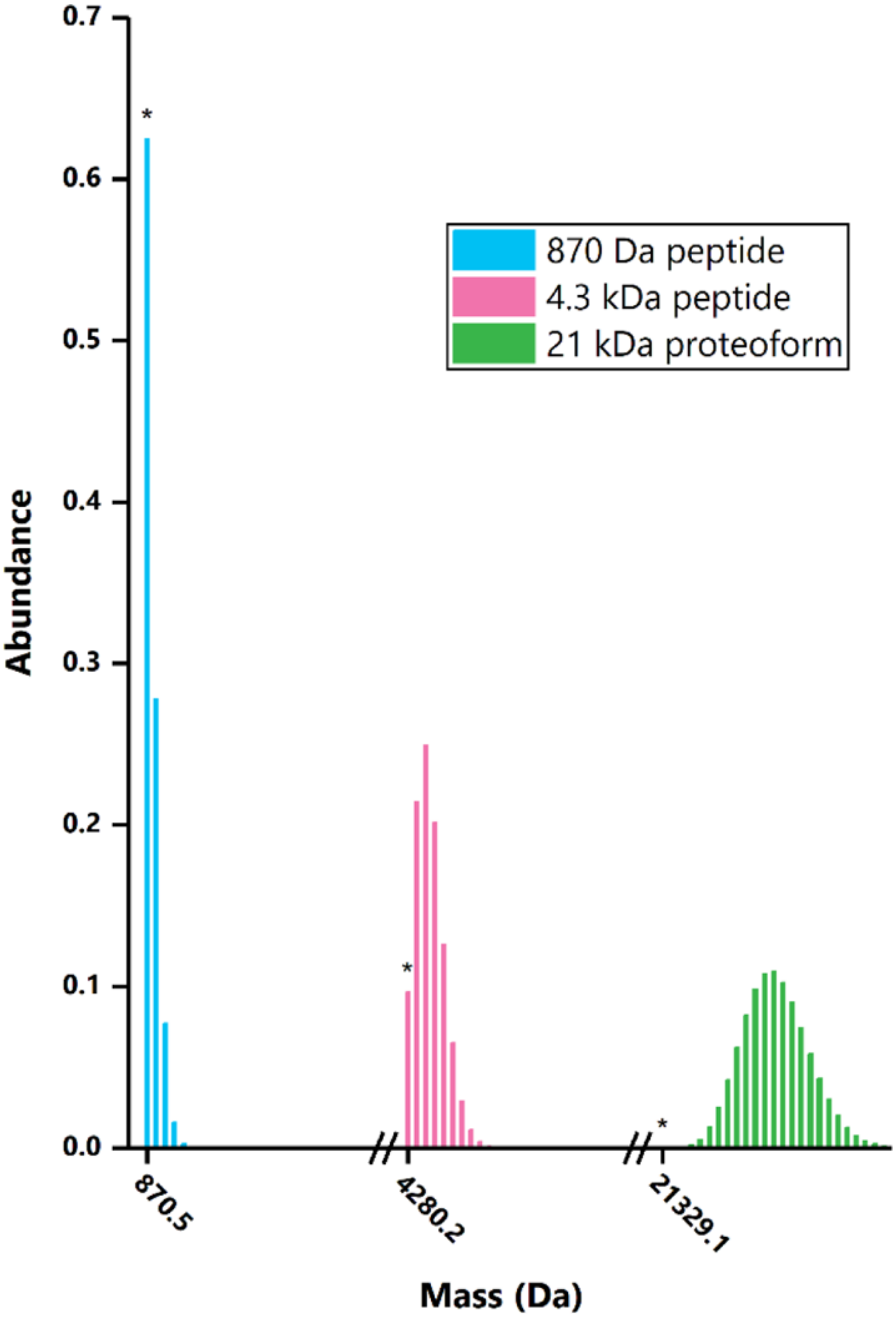

Figure 5.

Theoretical isotopic envelopes of three species of different mass (870 Da, 4.3 kDa, and 21 kDa). The monoisotopic mass of each species is annotated with an asterisk. The monoisotopic peak becomes increasingly difficult to observe as mass increases.

Table 1.

Terms and definitions relevant to proteoform analysis.

Term	**Definition**
Proteoform	Defined sequence of amino acids with localized modifications
Proteoform Family	All proteoforms derived from a single gene in a defined genome
Protein	A linear sequence of amino acids
Protein Group	A collection of proteins that are indistinguishable from each other based on the peptides identified in bottom-up analysis
Protein Isoform	A member of a set of proteins from the same gene or gene family that arise from alternative splicing or variable promoter usage
Proteoform Spectrum Match	An observed tandem mass spectrum match to a proteoform identification

Table 2.

Summary of the intact protein separation techniques described in this review.

Separation Technique	Mode of Separation	Benefits	Challenges
Gel-eluted liquid fraction entrapment electrophoresis (GELFrEE)	Size	• High-resolution • Commercially available	• MS-incompatible solvent
Serial Size Exclusion Chromatography	Size	• MS-compatible solvent	• Lower resolution • Sample load requirements
Reversed-Phase Liquid Chromatography	Hydrophobicity	• Small particle size and long columns increase peak capacity • High and low pH are orthogonal for 2D-LC	• Requires in-house equipment and expertise to pack long, small-particle columns • Some proteins difficult to elute
Capillary Zone Electrophoresis	Electrophoretic mobility	• Low sample amount needed	• Low sample loading volume • Narrow separation window • Difficult to make high-quality capillary coatings

Table 3.

Summary of proteoform identification software programs described in this review.

Software Program	Key Features
MASH Suite http://ae.crb.wisc.edu/software.html	Interface to perform MS/MS search and manually validate MS/MS identifications
MetaMorpheus httD://aithub.com/smith-chem-wisc/MetaMorDheus	MS/MS search with PTM discovery and monoisotopic mass error notch search
MSPathFinder http://aithub.com/PNNL-Comp-Massspec/Informed-Proteomics/	MS/MS search that identifies proteoforms with sequence graph and uses LC-data integration to improve monoisotopic mass determination
Proteoform Suite http://aithub.com/smith-chem-wisc/ProteoformSuite	MS1-onIy to identify proteoforms by intact-mass observations and mass differences corresponding to modifications
TDPortal http://nrtdp.northwestern.edu/tdportal-request	MS/MS search against reference databases and biomarker search for truncated proteoforms
TopMG http://proteomics.informatics.iupui.edu/software/toppic/	MS/MS tool for ultra-modified proteoforms
TopPIC http://proteomics.informatics.iupui.edu/software/toppic/	MS/MS search against database with spectral alignment to determine unknown mass shifts