Published in final edited form as:

Anal Chem. 2018 January 16; 90(2): 1325–1333. doi:10.1021/acs.analchem.7b04221.

Expanding Proteoform Identifications in Top-Down Proteomic Analyses by Constructing Proteoform Families

Leah V. Schaffer[†], Michael R. Shortreed[†], Anthony J. Cesnik[†], Brian L. Frey[†], Stefan K. Solntsev[†], Mark Scalf[†], and Lloyd M. Smith^{†,‡,*}

[†]Department of Chemistry, University of Wisconsin, 1101 University Avenue, Madison, Wisconsin 53706, United States

[‡]Genome Center of Wisconsin, University of Wisconsin, 425G Henry Mall, Room 3420, Madison, Wisconsin 53706, United States

Abstract

In top-down proteomics, intact proteins are analyzed by tandem mass spectrometry and proteoforms, which are defined forms of a protein with specific sequences of amino acids and localized post-translational modifications, are identified using precursor mass and fragmentation data. Many proteoforms that are detected in the precursor scan (MS1) are not selected for fragmentation by the instrument and therefore remain unidentified in typical top-down proteomic workflows. Our laboratory has developed the open source software program Proteoform Suite to analyze MS1-only intact proteoform data. Here, we have adapted it to provide identifications of proteoform masses in precursor MS1 spectra of top-down data, supplementing the top-down identifications obtained using the MS2 fragmentation data. Proteoform Suite performs mass calibration using high-scoring top-down identifications and identifies additional proteoforms using calibrated, accurate intact masses. Proteoform families, the set of proteoforms from a given gene, are constructed and visualized from proteoforms identified by both top-down and intact-mass

*Corresponding Author: smith@chem.wisc.edu. Phone: 608-263-2594. Fax: 608-265-6780.

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.7b04221. Materials and methods, raw data files, calibration, mass difference histogram for experiment–theoretical pairs, mass difference histogram for experiment–experiment pairs, the process for identifying experimental proteoforms in proteoform families, and mass error of top-down hits before and after calibration (PDF)

Tables of experimental proteoforms, theoretical proteoforms, experiment–theoretical pairs, experiment–experiment pairs, experiment–experiment mass difference peaks, proteoform families, mass errors of top-down hits, identified intact-mass experimental proteoforms, and top-down experimental proteoforms (ZIP) Program, T15LM007359.

ORCID

Leah V. Schaffer: 0000-0001-6339-9141

Anthony J. Cesnik: 0000-0002-5326-7134

Brian L. Frey: 0000-0002-0397-7269

Stefan K. Solntsev: 0000-0002-1061-1476

Author Contributions

L.V.S. added the calibration and top-down analyses to Proteoform Suite. A.J.C., L.V.S., and M.R.S. developed Proteoform Suite. S.K.S. developed the mass calibration algorithm and managed the software code base organization. B.L.F. and M.R.S. decided on key parameters in Proteoform Suite and Protein Deconvolution 4.0. L.V.S. prepared the yeast samples. M.S. ran the mass spectrometry experiment. L.V.S. drafted the manuscript. L.M.S. provided oversight of the work. All authors reviewed and made final edits on the manuscript.

Notes

The authors declare no competing financial interest.

analyses. Using this strategy, we constructed proteoform families and identified 1861 proteoforms in yeast lysate, yielding an approximately 40% increase over the original 1291 proteoform identifications observed using traditional top-down analysis alone.

Graphical abstract

Cells require much biochemical diversity to perform the vast array of necessary biological functions, including precise control over timing and rate of enzymatic reactions and spatial localization of proteins. This diversity is supplied in large part by the tremendous variety of forms in which proteins can exist within the cell. Proteoforms are the many different molecular forms proteins can take, including those differing by post-translational modifications (PTMs) and amino acid sequence changes due to genetic variations or alternatively spliced RNA transcripts.¹ All proteoforms that derive from a given gene are members of the same proteoform family.² Different proteoforms of the same proteoform family often exhibit different functional behavior or interaction profiles in the cell.^{3,4} A classic example of the importance of proteoforms to cellular function is found in the “histone code”, where different combinations of PTMs on histones act to regulate gene transcription.⁵ In another case, it was shown that the phosphorylations of different sites on the transcription factor Elk-1 result in a different transcriptional response.⁴ These examples illustrate how identifying the specific proteoforms present in a sample is necessary to truly understand a biological system.

The most widely used method for identifying proteins, bottom-up proteomics, is capable of identifying thousands of proteins by mass spectrometry (MS),^{6,7} but it is not capable of identifying proteoforms. Bottom-up proteomics involves digesting proteins into peptides, typically using trypsin, followed by liquid chromatography and tandem mass spectrometric analysis (LC-MS/MS).⁸ The protein digestion step results in the loss of information regarding the original amino acid sequence and combination of PTMs. For example, a proteoform with N-terminal acetylation and phosphorylation near the C-terminus will be digested into many peptides, separating these two modifications and making it impossible to know whether they are colocalized on the proteoform.

Top-down proteomics, on the other hand, can identify proteoforms by analyzing intact proteins (instead of digested peptides) by LC-MS/MS.⁹ In a typical top-down proteomic analysis, a precursor mass spectrum (MS1) of intact proteins is recorded. Then, the top few most intense peaks are each selected for fragmentation, and mass spectra (MS2) of the

resulting fragment ions are acquired. This type of analysis is referred to as data-dependent acquisition in both top-down and bottom-up proteomics because the parent ion isolation window selected for fragmentation, and thus the MS2 scan, depends on the peaks in the MS1 scan. Proteoforms are identified in most top-down software based upon the precursor monoisotopic mass and analysis of fragmentation products.^{10–14} A fully characterized proteoform will have the complete sequence determined and all PTMs localized at specific residues, but such complete characterization is not often attained. While proteoform identification typically means the protein has been successfully identified, proteoform characterization is concerned with achieving maximal fragmentation sequence coverage and modification localization for each specific proteoform.¹⁵ Statistical methods for evaluating the confidence of top-down proteoform identifications are generally less well-developed than for bottom-up peptide identifications and subsequent protein inference; this problem is being actively addressed in the field.^{15,16}

Recent advances have demonstrated that top-down analysis of complex samples is possible, such as in yeast¹⁷ and human^{18–20} cell lysates. However, top-down proteomics remains challenging for complex samples compared to the analysis of peptides in bottom-up proteomics. One reason for this is that there is an exponential decay in the signal-to-noise ratio of proteins as molecular weight increases.^{21,22} Additionally, higher resolution is required to resolve the isotopic peaks of highly charged intact proteins. As a result, longer spectral acquisition time for each MS1 and MS2 scan is required to achieve sufficient resolution and signal-to-noise to analyze intact proteins and their fragments. There is only enough time for the most intense proteoform mass peaks to be selected for fragmentation by the mass spectrometer during data-dependent acquisition (typically the top two or three peaks for each MS1, as opposed to the typical top 10 for each MS1 in bottom-up analyses), leaving many less abundant proteoforms unfragmented, and thus unidentified. For these reasons there are far more distinct intact masses observed in the MS1 spectra of top-down analyses of complex samples than are identified.²³ A strategy capable of identifying proteoform masses that were detected, but not fragmented or identified through tandem MS, would substantially increase the number of proteoform identifications in top-down studies of complex samples.

Some studies have explored identifying proteoforms using intact mass alone.¹⁴ The Kelleher lab developed a program called PTMcRAWler, which searches for protein mass shifts corresponding to PTM mass differences from proteoforms identified through tandem MS.²⁴ While this approach increases the number of identifications, a strategy that analyzes all intact masses detected throughout the run would allow more proteoforms to be identified. Such additional proteoform identifications could include those with multiple modifications or amino acid losses from proteoforms identified through top-down analysis, as well as those from families that were not identified by tandem MS at all.

Our lab recently reported a proteomic strategy for the identification of proteoforms in complex samples based upon intact mass and lysine count (determined with isotopic labeling).² This strategy differs from standard “top-down” proteomic strategies, in which proteoforms are fragmented in the gas phase and sequence and possibly PTM information is obtained from the fragmentation products (MS2 analysis). We previously developed the

software program Proteoform Suite to automate the process of identification, quantification, and visualization of proteoform families using intact-mass and lysine count determinations.^{2,25,26} In this report, we extend the capabilities of Proteoform Suite to identify proteoforms and construct proteoform families from standard top-down data sets. This new strategy allows Proteoform Suite to be integrated into current top-down workflows that use fragmentation to identify proteoforms; additionally, the cost and time of growing isotopically labeled cells (not a step in most top-down workflows) is not required. As a result, Proteoform Suite can now be used to analyze top-down data to increase the number of proteoform identifications.

We first analyzed a fractionated yeast data set using a typical top-down workflow, yielding 1291 unique proteoform identifications at 5% false discovery rate (FDR) by tandem MS. To increase the number of identifications, we developed a strategy to analyze all intact masses detected in the MS1 spectra of the top-down raw data. The strategy uses comparisons of experimental masses against those from a theoretical database and between experimental proteoforms to identify additional proteoforms by intact mass alone. Proteoform Suite integrates the top-down and intact-mass identifications into proteoform families and streamlines the visualization of these families in Cytoscape.^{27,28} Figure 1 displays an overview of the strategy. An additional 570 proteoforms were identified using intact mass in Proteoform Suite, representing an approximately 40% increase compared to a typical top-down data analysis.

EXPERIMENTAL PROCEDURES

Data Acquisition

Saccharomyces cerevisiae strain Y1788 was grown to log phase ($OD_{600} = 0.7$) in synthetic complete media, and then pelleted, flash-frozen in liquid nitrogen, and stored at -80°C until used. Cells were lysed by heat, and protein was precipitated. After resuspending the protein in 1% sodium dodecyl sulfate (SDS), a 12% GELFrEE cartridge (Expedeon) was used to perform a size-based separation²⁹ of approximately 400 μg of yeast protein into 12 fractions, following the manufacturer's recommended procedure. Top-down analysis was performed for each fraction on a QE-HF Orbitrap (Thermo Fisher Scientific), with the two most intense MS1 peaks selected for fragmentation, i.e., top-2 data-dependent acquisition. Each fraction was injected twice to produce 24 mass spectrometry raw files, which are publicly available on the MassIVE platform (<https://massive.ucsd.edu>; ID: MSV000081592). The sample preparation and mass spectrometric analysis are described in greater detail in the Materials and Methods section in the Supporting Information.

Data Analysis

All raw files can be found on the MassIVE platform, while all other files used for this analysis are available in the Vignette folder in release 0.3.0 of Proteoform Suite (<https://github.com/smith-chem-wisc/ProteoformSuite/releases/tag/0.3.0>). A tutorial document located in the Vignette folder describes how to use method files (.xml) to automatically perform end-to-end analyses that reproduce our results in the software.

Top-Down MS—We used TDPortal (from the National Resource for Translational and Developmental Proteomics, NRTDP, Northwestern University, Evanston, IL) to perform the top-down analysis of the raw files; TDPortal is available for academic collaborators at <http://nrtdp.northwestern.edu/tdportal-request/>. The analysis generated a shotgun annotated database from the May 2016 Swiss-Prot release of the yeast proteome, meaning theoretical proteoforms were generated for each protein based on the sequence, potential sequence variations, and potential modifications.^{10,30} The search used two modes as defined for ProSight PTM 2.0: tight absolute mass and biomarker searches.¹⁰ For the former, a 2.2 Da tolerance was used for MS1, and a 10 ppm tolerance was used for MS2. For the biomarker search, both MS1 and MS2 tolerances were set to 10 ppm. Search results were then analyzed in TDViewer (version 1.0; <http://topdownviewer.northwestern.edu>), which generated a list of 30 955 top-down hits, or protein spectrum matches, that met a 5% protein-level FDR cutoff. This list was saved as a Microsoft Excel file.

Deconvolution of MS1 Scans—MS1 spectra from the raw top-down data were charge-state deconvoluted and deisotoped using Thermo Protein Deconvolution 4.0 (70% fit factor, minimum S/N = 2, remainder threshold = 10%, minimum of three detected charge states, and a charge range of +5 to +30). A sliding window of 0.20 min and 34% offset was used to deconvolute the retention time range of 35–90 min, when most proteins elute. This retention time range was split into multiple smaller retention time ranges for the same raw file because the data export format (.xls) has a maximum number of rows allowed. The exported deconvolution result files contained a list of accurate monoisotopic intact masses of detectable proteoforms for each file. In total, 161 929 raw experimental components were revealed by deconvolution, where a “raw experimental component” is a monoisotopic mass and intensity of a proteoform observation.

Proteoform Suite

We developed the open-source software Proteoform Suite to automate the entire process of constructing proteoform families and identifying proteoforms by intact mass.²⁵ It is freely available at <https://github.com/smith-chem-wisc/ProteoformSuite>; all analyses in this study were performed on version 0.3.0. On a Dell Precision Tower 5810 desktop computer with a 6-core, 12-thread Xeon 3.60 GHz processor, Proteoform Suite performs calibration in 3–13 min per raw file (for 20 of the 24 total files, as discussed below), and then performs the full analysis of constructing proteoform families and identifying proteoforms in approximately 17 min.

Mass Calibration—Postspectral acquisition calibration is performed in Proteoform Suite to increase the mass accuracy of the data. The mass calibration strategy employed in this work is similar to the software lock-mass concept used in bottom-up proteomics.^{31,32} Briefly, the mass error was calculated by taking the difference between the theoretical mass and the experimentally observed mass for highly confident top-down hits (with a minimum C-score of 40, corresponding to proteoforms that are considered identified and well-characterized;¹⁵ files were only calibrated if they had a minimum of five top-down hits with this score or greater, i.e., 20 of 24 raw files). These mass errors were used to generate calibration functions that were used to shift the monoisotopic mass values for deconvoluted

intact masses from the same MS file. New Microsoft Excel files containing the calibrated deconvoluted intact-mass values were exported for subsequent use in Proteoform Suite; a new Microsoft Excel file was also created for the top-down hits containing calibrated monoisotopic mass values. These files contained a total of 30 948 calibrated top-down hits and 160 541 calibrated raw experimental components from deconvolution. For more information on the algorithm used for calibration, please see the Supporting Information.

Intact-Mass Experimental Proteoforms—First, the calibrated Excel files of deconvoluted masses were read into Proteoform Suite to generate a list of raw experimental components (each composed of a monoisotopic mass and intensity measurement). Proteoform Suite corrects for errors involving deconvolution, including (1) “missed” monoisotopic masses,³³ where the incorrect isotopic peak is reported as the monoisotopic mass, and (2) charge-state harmonics, where a multiple of the monoisotopic mass is reported. Proteoform Suite merged 19 323 raw experimental components due to missed monoisotopic masses and 1291 due to charge-state harmonics, resulting in 140 019 processed raw experimental components. Then, these components were aggregated by mass and retention time to create a list of 7801 unique experimental proteoforms, referred to as “intact-mass experimental proteoforms” in this work. A mass tolerance of 5 ppm, a retention time tolerance of 5 min, and an allowance of missed monoisotopic mass errors of up to 3 units were utilized to generate this list.

Top-Down Experimental Proteoforms—Many of the intact-mass experimental proteoforms (observed in MS1 spectra) were fragmented and identified by TDPportal; we incorporated these top-down identifications into Proteoform Suite to aid in intact-mass analysis. Therefore, in Proteoform Suite there are two types of experimental proteoforms: intact-mass (deconvoluted from MS1 spectra and unidentified by top-down) and top-down (identified by TDPportal). The calibrated top-down hits identified in TDPportal at 5% protein-level FDR were imported into Proteoform Suite. Hits with a minimum C-score of 3 were accepted for further processing, since this is considered the minimum C-score for a proteoform to be identified. Missed monoisotopic mass errors of the observed values were automatically corrected. Proteoform Suite aggregated the hits to create a list of the unique top-down experimental proteoforms, consisting of a sequence of specific amino acids with PTMs at specific residues. Top-down hits were also aggregated by retention time with a tolerance of 5 min, resulting in 1526 top-down experimental proteoforms (top-down hits with the same identification but retention times outside of this tolerance were separated). For each top-down experimental proteoform, retention time and observed mass were calculated by averaging values of the aggregated top-down hits. The 1526 top-down experimental proteoforms represent a total of 1291 unique proteoform identifications due to the same proteoform eluting at different retention times outside of the 5 min tolerance.

Top-down experimental proteoforms were added to the list of intact-mass experimental proteoforms generated from the deconvoluted intact-mass values. To create a list of unique experimental proteoform observations, top-down experimental proteoforms (which are already identified) replaced any intact-mass experimental proteoforms with the same mass and retention time; mass comparisons used tolerances of 5 ppm and 3 possible missed

monoisotopic mass units, and retention time comparisons used a 5 min tolerance. The final list of 8272 experimental proteoforms (Table S-1) therefore contained both the intact-mass experimental proteoforms and the top-down experimental proteoforms.

Theoretical Proteoforms—To help identify known proteoforms that were observed experimentally, we generated a list of theoretical proteoform masses containing PTMs known to exist in yeast. This list was generated using a UniProt XML database for *S. cerevisiae* (downloaded in February 2017), as well as a UniProt XML database with common contaminants in proteomic samples, such as human keratin protein. Proteoform Suite created theoretical proteoforms with up to two annotated modifications, limiting the combinatorial expansion while generating theoretical proteoforms with known modifications (see the Results and Discussion). Oxidation was considered a possible PTM at any methionine residue, since it is commonly observed. The theoretical proteoform database created in this way contained 29 839 entries.

Top-down analysis may reveal proteoforms that do not exist in the theoretical database. The theoretical database was supplemented with additional theoretical proteoforms originating from top-down identifications if their corresponding theoretical proteoforms were not already present. Missing theoretical proteoforms added were either subsequences of full length proteins or proteoforms with more than two PTMs, which was the maximum allowed when originally creating PTM combinations. Table S-2 shows the final theoretical database of 30 841 theoretical proteoforms.

Decoy databases of the same size were generated and used later in this analysis to assess the FDR for proteoform identifications. The amino acid sequences used in the decoy databases were created by concatenating all yeast protein sequences (including any top-down sequences added) in random order into a single continuous string, and then by selecting substrings of this sequence with lengths equal to each of the known yeast proteins. This strategy yields sequence segments with similar amino acid frequencies and arrangements as are found in the target database. PTMs from the original target sequence with that length (including top-down proteoform PTM sets) were added to the decoy sequences. Each resulting decoy proteoform sequence, set of modifications, and sequence length are therefore similar to those found in the original target database.

Construction of Proteoform Families—Proteoform Suite uses these lists of experimental and theoretical proteoforms to construct families, which allows the products from a given gene to be identified and visualized. The process for constructing proteoform families is outlined in Figure 2. This figure, as previously reported,² illustrates how top-down identifications (purple nodes) are combined with deconvoluted intact-mass results (blue nodes) to construct proteoform families. The first steps involve forming the connections between proteoforms. First, the list of experimental proteoform masses was compared to the list of theoretical proteoform masses to form experiment–theoretical pairs (ET pairs) of any comparison with a delta mass (mass difference) between ± 1 Da (Table S-3). Using the same process, experimental proteoforms were also compared to decoy theoretical proteoforms to form decoy ET pairs and assess FDR. A mass difference histogram was generated from the ET pairs with a bin size of 0.025 Da (Figure S-1). There

were 1339 ET pairs in the 0 Da peak, which were all accepted as identifications. The FDR of these identifications was 5.5%; this value was calculated as the median ratio of decoy pairs to the number of ET pairs in the same peak, i.e., within ± 0.0125 Da of the peak mass difference.

Experimental proteoforms observed may not form exact matches with theoretical proteoforms due to amino acid differences or additional PTMs. To identify additional experimental proteoforms not in the theoretical database, Proteoform Suite compares experimental proteoform masses to one another to find mass differences corresponding to known modifications. Each experimental proteoform was compared to other experimental proteoforms with a chromatographic retention time difference of less than 2.5 min to form experiment–experiment pairs (EE pairs), as related proteoforms with similar sequence lengths generally elute within this time range of one another. All EE pairs are shown in Table S-4. A mass difference histogram was plotted from the EE pairs using a bin size of 0.025 Da to determine frequent mass differences between experimental proteoforms (Figure S-2). The mass difference peaks with a count of at least 75 EE pairs were manually inspected and marked as accepted if corresponding to a common and probable set of modifications. The EE pairs in accepted peaks were also accepted. The EE mass difference peaks are shown in Table S-5, displaying which peaks were accepted as well as possible peak assignments (the modification sets that could correspond to that peak's mass difference).

For EE comparisons, there is no decoy database to use for FDR assessment. Instead, we created decoy EE pairs, where experimental masses were compared to other experimental proteoforms outside the retention time tolerance of 5 min. The number of these false EE pairs is much larger than the number of EE pairs eluting within 2.5 min. Therefore, a random subset of decoy pairs was selected that was equal in count to the number of target EE pairs. The FDR for each EE histogram peak was the median ratio of decoy pairs with a mass difference within ± 0.0125 Da of the peak to target EE pairs in the peak.

Finally, accepted ET pairs and EE pairs (target pairs) were joined together into proteoform families (Figure 2). Proteoform families are visualized as a network of circles representing proteoforms, connected by lines representing mass differences that correspond to modifications or amino acid differences.² The visualization of theoretical proteoforms (green nodes) makes it clear which proteoforms were identified by ET pairs and which were identified through EE connections. Proteoform Suite outputs a script that allows the user to quickly and easily visualize proteoform families in the network visualization software Cytoscape.^{27,28} Intact-mass experimental proteoforms are represented as blue circles, top-down experimental proteoforms are represented as purple circles, and theoretical proteoforms are represented as green circles. Each green theoretical circle is connected to a pink gene rectangle, bringing all proteoforms of the same gene together into one proteoform family. The size of each circle representing intact-mass experimental proteoforms is proportional to the integrated ion intensity. Table S-6 lists all proteoform families. A Cytoscape file of the visualized proteoform families can be found in the Vignette folder in release 0.3.0 of Proteoform Suite.

Proteoform Identification—Proteoform Suite automatically identifies proteoforms within each proteoform family.²⁵ Starting with a theoretical node, the connections representing exact matches or characteristic PTM mass differences provide the initial identification of experimental proteoforms; all accepted ET pairs are considered identifications in this way. EE pairs can reveal daisy chains of additional identifications starting from original ET identifications. Multiple modifications are often possible for a given mass difference (e.g., phosphorylation and sulfonation), and so we applied heuristics to choose the most likely assignment based on the protein sequence, the modifications known for that protein, and the frequency of the modification in the theoretical database. This limits the FDR by preventing false identifications when identifying proteoforms using EE pairs, such as assigning amino acid losses that cannot occur on a given protein sequence. An example of how Proteoform Suite performs this identification process by following pairs in proteoform families is shown in Figure S-3.

The same identification process was performed in decoy proteoform families to assess the global FDR for proteoform identification. Decoy proteoform families were constructed for each of the 10 sets of accepted decoy proteoform pairs (i.e., within ± 0.0125 Da of an accepted ET or EE peak mass difference) in the same way as ET and EE pairs (target proteoform pairs). The FDR was then calculated as the ratio of the experimental proteoforms identified in target proteoform families to the average number of experimental proteoforms identified in the 10 sets of decoy proteoform families.

RESULTS AND DISCUSSION

Top-down proteomics is a powerful and useful proteomic strategy because of its ability to fully characterize proteoforms, which are the major effectors in most biological systems. However, the number of proteoforms observed in MS1 spectra that can be identified by this strategy is limited by the substantial instrument time required for the tandem MS analysis of intact proteins. Therefore, many proteoforms remain unidentified in top-down proteomic analyses of complex samples. We identified 1291 unique proteoforms in yeast lysate at 5% FDR by top-down analysis, but deconvolution of all MS1 spectra revealed that 7801 unique proteoforms were represented in the spectra. This much larger count of proteoform observations illustrates that many more proteoforms are present in complex proteomic samples than are identified by typical top-down workflows. In the present work, we extended the capabilities of the software program Proteoform Suite to identify some of these unidentified proteoform observations.

To identify proteoforms by intact mass alone, mass accuracy is of utmost importance. A strategy for postacquisition spectral calibration in bottom-up proteomics has been developed that uses a two-dimensional minimization of peptide identification mass errors to calibrate masses;³¹ a similar approach was implemented in Proteoform Suite. The mass error of high-scoring top-down hits (identified in TDPportal) was minimized globally for each file as a function of precursor mass-to-charge ratio (m/z) and retention time. These functions were used to calibrate all intact masses deconvoluted from the MS1 spectra, as well as the top-down precursor masses, thus minimizing mass errors. Before calibration, the average mass error of top-down hits was 1.1 ppm, and the standard deviation of the mass error was 3.2

ppm. Proteoform Suite mass calibration improved the average mass error to 0.5 ppm with a standard deviation of 3.0 ppm. Figure S-4 shows a histogram of the precursor mass errors of all top-down hits before and after mass calibration, and Table S-7 shows the mass error for each top-down hit before and after calibration in Proteoform Suite. An analysis of the uncalibrated data resulted in fewer intact-mass identifications with a higher FDR compared with the calibrated analysis, as described in the Supporting Information. Mass calibration in Proteoform Suite improved mass accuracy and the ability to identify proteoforms by intact mass; future versions of Proteoform Suite will perform retention time calibration to account for elution differences of proteoforms across different MS runs to enhance aggregation and EE comparisons.

The use of Proteoform Suite to analyze all proteoform intact masses observed in MS1 spectra yielded an additional 570 unique proteoform identifications at 4.5% FDR, yielding a total of 1861 unique proteoform identifications and increasing the number of identifications by almost 40% over the original 1291 identifications identified by top-down analysis at 5% FDR (Figure 3). Additionally, the number of unique proteins identified increased from 375 proteins to 443 proteins, meaning entire proteoform families that were unidentified in the top-down analysis were revealed by Proteoform Suite. Proteoforms that were newly identified by Proteoform Suite, i.e., identified by intact mass alone, are listed in Table S8. We note that many of these identifications are potential artifacts of sample handling or chemical modifications (e.g., oxidation, ammonia loss). However, the identification of these proteoforms is important because (1) these potential artifacts could indeed be biological modifications and (2) it is important to identify artifacts, including SDS adducts, so that they are not misidentified as a different proteoform. There were 65 intact-mass experimental proteoform identifications that represented duplicate identifications of ones that had been identified previously by TDPortal. These 65 intact-mass proteoforms were not counted as additional Proteoform Suite identifications, and they resulted from instances where the intact-mass experimental proteoforms and the corresponding top-down experimental proteoforms were not merged because they had retention time differences of greater than 5 min. There were 102 experimental proteoforms that resulted from an adduct observation of hydrogen dodecyl sulfate (sodium dodecyl sulfate was used in sample preparation), which were visualized in families but not counted as additional identifications. There were three intact-mass experimental proteoforms with ambiguous identifications, where the proteoform had equal numbers of connections from theoretical proteoforms of different genes. These ambiguous identifications make up a small number of intact-mass identifications and are examples of where subsequent top-down fragmentation would be necessary to disambiguate these proteoforms.

Proteoform Suite also formed pairs and constructed families with the top-down experimental proteoforms, which can be compared with the identifications assigned by TDPortal. Table S-9 shows all top-down experimental proteoforms and, if applicable, the identity assigned by Proteoform Suite. Of the 1526 top-down proteoforms (counting separately proteoforms with more than a 5 min difference in retention time), 1090 were correctly identified in Proteoform Suite, meaning they were formed into an ET pair with the same theoretical proteoform as assigned by TDPortal. There were only 56 top-down experimental proteoforms where the identification did not match between Proteoform Suite and TDPortal, which is around 3% of

the total number of proteoform identifications. Due to the 5% FDR in both TDPortal identifications and the final Proteoform Suite identifications, either software may be correct in these disagreeing identifications. There were 380 top-down proteoforms that did not match with any theoretical proteoform because the mass error of these top-down proteoforms was too large, and so the ET proteoform pairs formed were not included in the mass difference histogram peak at 0 Da (Figure S-1). To reduce the FDR in Proteoform Suite for identifying proteoforms by intact mass alone, stringent mass error tolerances are necessary; inevitably, some proteoforms with a mass error outside of the allowed tolerance of ± 0.0125 Da will only be identifiable by fragmentation data. Our results illustrate how accurate precursor monoisotopic mass values can allow additional proteoforms to be identified from top-down data.

One limitation of the intact-mass strategy for proteoform identification is that it does not provide localization of PTMs to specific residues. PTMs can in principle be localized on such proteoforms through subsequent targeted top-down analysis, although sequence coverage of regions containing the PTMs is not guaranteed. Fragmentation in top-down proteomics allows PTMs to be localized in well-characterized proteoforms (C-score greater than 40).¹⁵ However, many proteoforms identified by top-down analysis do not have sequence coverage over the range of the protein sequence containing the PTM, compromising localization. Comprehensive and accurate PTM localization thus remains an outstanding challenge in proteomics.

The theoretical database generated in Proteoform Suite included PTMs annotated in UniProt, as well as methionine oxidations. A major challenge in top-down proteomics is the generation of a theoretical database of proteoforms that includes enough modified proteins to identify proteoforms in the sample but does not lead to a combinatorial explosion that increases the FDR.³⁴ In bottom-up proteomics, unmodified peptides of a heavily modified proteoform can still be identified, whereas in top-down proteomics, modifications in the theoretical database are necessary to identify the modified intact protein. The major trade-off observed in Proteoform Suite, where proteoform identification is based on intact mass alone, is thus including enough modification combinations in the database without increasing the database size to an extent that significantly impacts the FDR. In the present work, we found that including combinations of up to two modifications yielded the most identifications but kept the FDR below 5%. (This is a parameter in the software that can be set depending on the data set and original database size employed.) A major advantage of Proteoform Suite is that additional modified proteoforms can be identified through the EE mass difference comparisons, as long as one proteoform in the family is contained in the theoretical database, including those identified by top-down analysis. Additionally, unexpected mass differences present promise for identifying novel modifications. A novel PTM could potentially be revealed by either the ET or EE mass difference histograms if an unknown mass is observed at high frequency; however, the in-depth analysis of uncommon or novel PTMs likely requires fragmentation and is therefore more suited for a bottom-up or top-down analysis.

Proteoform Suite constructs proteoform families from both intact-mass and top-down proteoforms. Figure 4A depicts all proteoform families constructed from this top-down yeast

data set, including identifications from TDPortal (top-down experimental identifications) and Proteoform Suite (intact-mass experimental identifications). In total, 1022 proteoform families were formed, consisting of 3903 experimental proteoforms. There were 660 unidentified families, meaning no theoretical node was paired with any of the experimental proteoforms in these families. A total of 4369 experimental proteoforms were not paired with another proteoform, referred to as “orphans.” (Orphans are not displayed in Figure 4.) A future targeted top-down analysis or an improved theoretical database could be used to identify these unknown families and orphan experimental proteoforms.

Proteoform families can be visualized as a network of distinct, related masses. The visualization of proteoform families allows the user to view all products of a given gene in a single graphic, clearly portraying PTM combinations and abundance differences between proteoforms. Intact-mass experimental node sizes are proportional to the MS intensity sum reported by the deconvolution software. MS intensity has been used for determining relative abundance of different proteoforms from the same family in other studies, since, unlike small peptides in bottom-up proteomics, modifications on intact proteins have a less significant effect on ionization efficiency.³⁵ Future versions of Proteoform Suite will implement quantification within proteoform families based on these relative proteoform intensities.

Proteoform Suite combines both intact-mass observations (deconvoluted from MS1 spectra) and top-down identifications (from TDPortal) into proteoform families, increasing the overall number of identifications and allowing the visualization of proteoforms derived from the same gene. The proteoform family in Figure 4B was identified by top-down analysis in TDPortal, but not in Proteoform Suite. The mass error between each top-down proteoform and the corresponding theoretical proteoforms in the database was greater than 0.0125 Da, and so fragmentation was necessary to identify this proteoform family. The family in Figure 4C is an example of a proteoform family that was only identified in Proteoform Suite; no proteoforms from the yeast gene *STF2* were identified by top-down analysis. Figure 4D depicts a proteoform family with a top-down identification and also additional intact-mass experimental proteoforms identified by Proteoform Suite. In this family, EE comparisons were leveraged to extend top-down or ET identifications and identify additional proteoforms from the same gene and proteoform family.

CONCLUSIONS

Identifying proteoforms using intact-mass measurements complements top-down proteomic analysis and increases the total number of identifications. Using a traditional top-down analysis, 1291 proteoforms were identified in yeast lysate; our strategy identified an additional 570 proteoforms, a 40% increase. Proteoform Suite software provides a powerful, automated strategy that can be implemented into current top-down workflows to increase the number of proteoforms identified and also to construct readily visualized proteoform families.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Ryan Fellers, Richard LeDuc, Joseph Greer, Bryan Early, and AJ van Nispen at the NRTDP, who developed TDPortal and continuously provided guidance in using the software. This work was supported by Grant R01GM114292 from the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health (NIH). L.V.S. was supported by the NIGMS Biotechnology Training Program, T32GM008349. A.J.C was supported by the Computation and Informatics in Biology and Medicine Training Program, T15LM007359.

References

1. Smith LM, Kelleher NL, et al. 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. Yang X, Coulombe-Huntington J, Kang S, Sheynkman GM, Hao T, Richardson A, Sun S, Yang F, Shen YA, Murray RR, Spirohn K, Begg BE, Duran-Frigola M, MacWilliams A, Pevzner SJ, Zhong Q, Trigg SA, Tam S, Ghamsari L, Sahni N, Yi S, Rodriguez MD, Balcha D, Tan G, Costanzo M, Andrews B, Boone C, Zhou XJ, Salehi-Ashtiani K, Charlotiaux B, Chen AA, Calderwood MA, Aloy P, Roth FP, Hill DE, Iakoucheva LM, Xia Y, Vidal M. Cell. 2016; 164:805–817. [PubMed: 26871637]
4. Mylona A, Theillet F-X, Foster C, Cheng TM, Miralles F, Bates PA, Selenko P, Treisman R. Science. 2016; 354:233–237. [PubMed: 27738173]
5. Jenuwein T, Allis CD. Science. 2001; 293:1074–1080. [PubMed: 11498575]
6. Kim M-S, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, Thomas JK, Muthusamy B, Leal-Rojas P, Kumar P, Sahasrabudhe NA, Balakrishnan L, Advani J, George B, Renuse S, Selvan LDN, Patil AH, Nanjappa V, Radhakrishnan A, Prasad S, Subbannayya T, Raju R, Kumar M, Sreenivasamurthy SK, Marimuthu A, Sathe GJ, Chavan S, Datta KK, Subbannayya Y, Sahu A, Yelamanchi SD, Jayaram S, Rajagopalan P, Sharma J, Murthy KR, Syed N, Goel R, Khan AA, Ahmad S, Dey G, Mudgal K, Chatterjee A, Huang T-C, Zhong J, Wu X, Shaw PG, Freed D, Zahari MS, Mukherjee KK, Shankar S, Mahadevan A, Lam H, Mitchell CJ, Shankar SK, Satishchandra P, Schroeder JT, Sirdeshmukh R, Maitra A, Leach SD, Drake CG, Halushka MK, Prasad TSK, Hruban RH, Kerr CL, Bader GD, Iacobuzio-Donahue CA, Gowda H, Pandey A. Nature. 2014; 509:575–581. [PubMed: 24870542]
7. Hebert AS, Richards AL, Bailey DJ, Ulbrich A, Coughlin EE, Westphall MS, Coon JJ. Mol Cell Proteomics. 2014; 13:339–347. [PubMed: 24143002]
8. Zhang Y, Fonslow BR, Shan B, Baek M, Yates JR. Chem Rev. 2013; 113:2343–2394. [PubMed: 23438204]
9. Catherman AD, Skinner OS, Kelleher NL. Biochem Biophys Res Commun. 2014; 445:683–693. [PubMed: 24556311]
10. Zamdborg L, LeDuc RD, Glowacz KJ, Kim Y, Viswanathan V, Spaulding IT, Early BP, Bluhm EJ, Babai S, Kelleher NL. Nucleic Acids Res. 2007; 35:W701–W706. [PubMed: 17586823]
11. Kou Q, Xun L, Liu X. Bioinformatics. 2016; 32:3495–3497. [PubMed: 27423895]
12. Sun RX, Luo L, Wu L, Wang RM, Zeng WF, Chi H, Liu C, He SM. Anal Chem. 2016; 88:3082–3090. [PubMed: 26844380]
13. Frank AM, Pesavento JJ, Mizzen CA, Kelleher NL, Pevzner PA. Anal Chem. 2008; 80:2499–2505. [PubMed: 18302345]
14. Karabacak NM, Li L, Tiwari A, Hayward LJ, Hong P, Easterling ML, Agar JN. Mol Cell Proteomics. 2009; 8:846–856. [PubMed: 19074999]
15. Leduc RD, Fellers RT, Early BP, Greer JB, Thomas PM, Kelleher NL. J Proteome Res. 2014; 13:3231–3240. [PubMed: 24922115]

16. Kou Q, Zhu B, Wu S, Ansong C, Toli N, Paša-Toli L, Liu X. *J Proteome Res.* 2016; 15:2422–2432. [PubMed: 27291504]
17. Kellie JF, Catherman AD, Durbin KR, Tran JC, Tipton JD, Norris JL, Witkowski CE, Thomas PM, Kelleher NL. *Anal Chem.* 2012; 84:209–215. [PubMed: 22103811]
18. Catherman AD, Durbin KR, Ahlf DR, Early BP, Fellers RT, Tran JC, Thomas PM, Kelleher NL. *Mol Cell Proteomics.* 2013; 12:3465–3473. [PubMed: 24023390]
19. 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]
20. Cleland TP, DeHart CJ, Fellers RT, VanNispen AJ, Greer JB, LeDuc RD, Parker WR, Thomas PM, Kelleher NL, Brodbelt JS. *J Proteome Res.* 2017; 16:2072–2079. [PubMed: 28412815]
21. Compton PD, Zamdborg L, Thomas PM, Kelleher NL. *Anal Chem.* 2011; 83:6868–6874. [PubMed: 21744800]
22. Riley NM, Mullen C, Weisbrod CR, Sharma S, Senko MW, Zabrouskov V, Westphall MS, Syka JEP, Coon JJ. *J Am Soc Mass Spectrom.* 2016; 27:520–531. [PubMed: 26589699]
23. Zhao Y, Sun L, Zhu G, Dovichi NJ. *J Proteome Res.* 2016; 15:3679–3685. [PubMed: 27490796]
24. Durbin KR, Tran JC, Zamdborg L, Sweet SMM, Catherman AD, Lee JE, Li M, Kellie JF, Kelleher NL. *Proteomics.* 2010; 10:3589–3597. [PubMed: 20848673]
25. Cesnik AC, Shortreed MR, Schaffer LV, Knoener RA, Frey BL, Scalf M, Solntsev SK, Dai Y, Gasch AP, Smith LM. *J Proteome Res.* 2017; doi: 10.1021/acs.jproteome.7b00685
26. 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]
27. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. *Genome Res.* 2003; 13:2498–2504. [PubMed: 14597658]
28. Smoot ME, Ono K, Rucseinski J, Wang P, Ideker T. *Bioinformatics.* 2011; 27:431–432. [PubMed: 21149340]
29. Tran JC, Doucette AA. *Anal Chem.* 2008; 80:1568–1573. [PubMed: 18229945]
30. LeDuc RD, Taylor GK, Kim Y-B, Januszyk TE, Bynum LH, Sola JV, Garavelli JS, Kelleher NL. *Nucleic Acids Res.* 2004; 32:340–345.
31. Cox J, Michalski A, Mann M. *J Am Soc Mass Spectrom.* 2011; 22:1373–1380. [PubMed: 21953191]
32. Solntsev SK, Shortreed MR, Frey BL, Smith LM. University of Wisconsin, Madison, WI. Unpublished work. 2017
33. Senko MW, Beu SC, McLaffertycor FW. *J Am Soc Mass Spectrom.* 1995; 6:229–233. [PubMed: 24214167]
34. Kou Q, Wu S, Toli N, Paša-Toli L, Liu Y, Liu X. *Bioinformatics.* 2017; 33:1309–1316. [PubMed: 28453668]
35. Gregorich ZR, Ge Y. *Proteomics.* 2014; 14:1195–1210. [PubMed: 24723472]

Figure 1.

Overview of the strategy for constructing proteoform families in Proteoform Suite with top-down proteomic data. Proteoform Suite contains a workflow consisting of four steps. Top-down proteomic data is first analyzed by TD Portal to produce a list of top-down proteoform identifications (step 1), and MS1 spectra from the same MS files are deconvoluted to produce a list of observed intact-mass experimental proteoforms (step 2). Then, a protein database is used to create a list of theoretical proteoforms (step 3). Finally, the masses of these three types of proteoforms are compared (step 4). Proteoform Suite outputs a list of identified proteoforms and visualized proteoform families, proteoforms derived from the same gene.

Figure 2.

Proteoform pairs are formed between experimental proteoform masses and theoretical proteoform masses. Experimental proteoforms are composed of top-down measurements (purple circles, identifications from TDPportal) and intact-mass measurements (blue circles, observed by deconvolution, not identified by TDPportal). Theoretical nodes are generated from a UniProt database (green circles). Experimental proteoform masses are compared to theoretical proteoform masses (ET pairs, lines between green and blue or purple circles), as well as to one another (EE pairs, lines between blue or purple circles). The lines representing the EE and ET pairs are labeled in orange with the mass difference between the two proteoforms connected. Proteoform pairs that correspond to a known set of modifications are accepted and joined to form proteoform families.

Figure 3.

Proteoform and protein identification results. The top graphic displays how Proteoform Suite increased the number of proteoform identifications by 40% (570 new identifications) using intact-mass determinations from a top-down (MS2) data set. The bottom graphic displays how the number of unique protein IDs (each corresponding to a particular gene) increased by 18% (68 newly identified proteoform families).

Figure 4.

(A) Visualization of the 1022 proteoform families, composed of 3903 proteoforms from the integration of top-down and intact-mass experimental proteoforms. The visualization of proteoform families allows all identified proteoforms from a given gene to be viewed in a single graphic, illustrating the combinations of PTMs and/or cleavage products present in the family. In this figure, proteoform families are arranged with a gene at the bottom; moving counterclockwise, any theoretical proteoforms are arranged by decreasing mass. Continuing counterclockwise, any experimental proteoforms are arranged by increasing mass. (B) The proteoform family for yeast gene *RPL11A* was identified by top-down analysis in TDPortal but not in Proteoform Suite due to mass error of the precursor proteoform masses. Therefore, the top-down proteoforms were not formed into accepted ET pairs. (C) The proteoform family for yeast gene *STF2* was missed by top-down analysis but was identified by intact-mass analysis. The lines between visualized theoretical and experimental proteoforms allow the user to know which proteoforms were identified by ET pairs (were present in the theoretical database) and which proteoforms were identified by the EE comparison (a mass shift from previously identified experimental proteoforms). This demonstrates how intact-mass analysis of proteoforms observed in MS1 spectra can identify additional proteoforms missed by top-down analysis. (D) The proteoform family for yeast gene *LSM5* shows how a proteoform identified by top-down analysis can be leveraged to identify additional experimental proteoforms by their intact masses alone. In this case, the top-down analysis identified the acetylated form of *LSM5* and comparison to other experimental proteoform masses revealed an acetylated form with a cleaved C-terminal leucine residue.