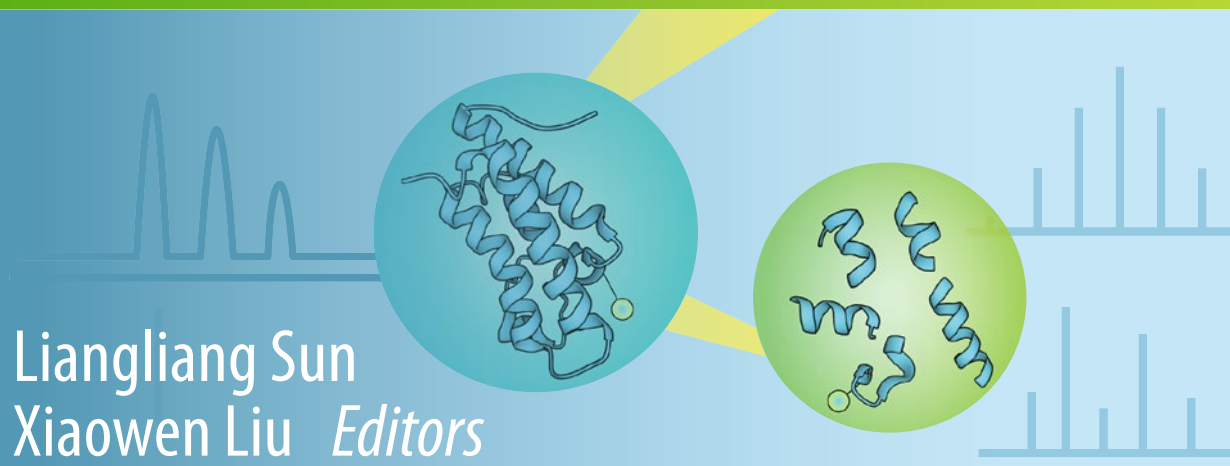
Methods in Molecular Biology 2500

Springer Protocols



Liangliang Sun
Xiaowen Liu *Editors*

Proteoform Identification

Methods and Protocols

Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor
John M. Walker
School of Life and Medical Sciences
University of Hertfordshire
Hatfield, Hertfordshire, UK

For further volumes:
http://www.springer.com/series/7651

For over 35 years, biological scientists have come to rely on the research protocols and methodologies in the critically acclaimed *Methods in Molecular Biology* series. The series was the first to introduce the step-by-step protocols approach that has become the standard in all biomedical protocol publishing. Each protocol is provided in readily-reproducible step-by-step fashion, opening with an introductory overview, a list of the materials and reagents needed to complete the experiment, and followed by a detailed procedure that is supported with a helpful notes section offering tips and tricks of the trade as well as troubleshooting advice. These hallmark features were introduced by series editor Dr. John Walker and constitute the key ingredient in each and every volume of the *Methods in Molecular Biology* series. Tested and trusted, comprehensive and reliable, all protocols from the series are indexed in PubMed.

Proteoform Identification

Methods and Protocols

Edited by

Liangliang Sun

Department of Chemistry, Michigan State University, East Lansing, MI, USA

Xiaowen Liu

Deming Department of Medicine, School of Medicine, Tulane University, New Orleans, LA, USA

Humana Press

Editors
Liangliang Sun
Department of Chemistry
Michigan State University
East Lansing,
MI, USA

Xiaowen Liu
Deming Department of Medicine
School of Medicine
Tulane University
New Orleans, LA, USA

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-0716-2324-4 ISBN 978-1-0716-2325-1 (eBook)
https://doi.org/10.1007/978-1-0716-2325-1

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Science+Business Media, LLC, part of Springer Nature 2022
This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Humana imprint is published by the registered company Springer Science+Business Media, LLC, part of Springer Nature.
The registered company address is: 1 New York Plaza, New York, NY 10004, U.S.A.

Preface

Mass spectrometry (MS)-based proteoform identification and characterization are extremely important for pursuing an accurate understanding of protein functions in diseases and development. The book aims to be a useful resource on various MS-based technologies for proteoform identification, characterization, and quantification, including sample preparation, proteoform separation, proteoform gas-phase fragmentation, bioinformatics tools for MS data analysis, and some important applications in the field. The book will benefit researchers in both academia and the biopharmaceutical industry who are interested in protein analysis using MS.

East Lansing, MI, USA — *Liangliang Sun*
New Orleans, LA, USA — *Xiaowen Liu*

Contents

Contributors

Joseph S. Beckman • *Department of Biochemistry and Biophysics, Oregon State University, Corvallis, OR, USA; e-MSion Inc., Corvallis, OR, USA*

Gili Ben-Nissan • *Department of Biomolecular Sciences, Weizmann Institute of Science, Rehovot, Israel*

Hao Chi • *Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China*

In Kwon Choi • *Deming Department of Medicine, School of Medicine, Tulane University, New Orleans, LA, USA*

Kellye A. Cupp-Sutton • *Department of Chemistry and Biochemistry, University of Oklahoma, Norman, OK, USA*

Jun Dai • *NovaBioAssays LLC, Woburn, MA, USA*

Rachel Franklin • *Department of Biochemistry and Biophysics, Oregon State University, Corvallis, OR, USA*

Ying Ge • *Department of Chemistry, University of Wisconsin – Madison, Madison, WI, USA; Department of Cell and Regenerative Biology, University of Wisconsin – Madison, Madison, WI, USA; Human Proteomics Program, University of Wisconsin – Madison, Madison, WI, USA*

Michael Hare • *e-MSion Inc., Corvallis, OR, USA*

Sophie R. Harvey • *Department of Chemistry and Biochemistry and Resource for Native Mass Spectrometry Guided Structural Biology, The Ohio State University, Columbus, OH, USA*

Si-Min He • *Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China*

Kyowon Jeong • *Department of Computer Science, University of Tübingen, Tübingen, Germany*

Chengjie Ji • *NovaBioAssays LLC, Woburn, MA, USA*

Jihyung Kim • *Department of Computer Science, University of Tübingen, Tübingen, Germany*

Oliver Kohlbacher • *Department of Computer Science, University of Tübingen, Tübingen, Germany*

Marius M. Kostelic • *Department of Chemistry and Biochemistry, University of Arizona, Tucson, AZ, USA*

Yu Liang • *CAS Key Lab of Separation Sciences for Analytical Chemistry, National Chromatographic Research and Analysis Center, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian, China*

Chao Liu • *Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China*

Xiaowen Liu • *Deming Department of Medicine, School of Medicine, Tulane University, New Orleans, LA, USA*

Lan Luo • *Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China*

Michael T. Marty • *Department of Chemistry and Biochemistry, University of Arizona, Tucson, AZ, USA; Bio5 Institute, University of Arizona, Tucson, AZ, USA*

Jake A. Melby • *Department of Chemistry, University of Wisconsin – Madison, Madison, WI, USA*
Suideng Qin • *School of Chemical Science & Engineering and Shanghai Key Laboratory of Chemical Assessment and Sustainability, Tongji University, Shanghai, China*
Leah V. Schaffer • *Department of Chemistry, University of Wisconsin-Madison, Madison, WI, USA*
Michal Sharon • *Department of Biomolecular Sciences, Weizmann Institute of Science, Rehovot, Israel*
Michael R. Shortreed • *Department of Chemistry, University of Wisconsin-Madison, Madison, WI, USA*
Lloyd M. Smith • *Department of Chemistry, University of Wisconsin-Madison, Madison, WI, USA*
Liangliang Sun • *Department of Chemistry, Michigan State University, East Lansing, MI, USA*
Rui-Xiang Sun • *National Institute of Biological Sciences, Beijing, China; Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China*
Timothy N. Tiambeng • *Department of Chemistry, University of Wisconsin – Madison, Madison, WI, USA*
Zhixin Tian • *School of Chemical Science & Engineering and Shanghai Key Laboratory of Chemical Assessment and Sustainability, Tongji University, Shanghai, China*
Rui-Min Wang • *Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China*
Zhe Wang • *Department of Chemistry and Biochemistry, University of Oklahoma, Norman, OK, USA*
Jesse W. Wilson • *Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA, USA*
Si Wu • *Department of Chemistry and Biochemistry, University of Oklahoma, Norman, OK, USA*
Zhijie Wu • *Department of Chemistry, University of Wisconsin – Madison, Madison, WI, USA*
Vicki H. Wysocki • *Department of Chemistry and Biochemistry and Resource for Native Mass Spectrometry Guided Structural Biology, The Ohio State University, Columbus, OH, USA*
Qiangwei Xia • *CMP Scientific Corp, Brooklyn, NY, USA*
Zhichang Yang • *Department of Chemistry, Michigan State University, East Lansing, MI, USA*
Dahang Yu • *Department of Chemistry and Biochemistry, University of Oklahoma, Norman, OK, USA*
Wen-Feng Zeng • *Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China*
Lihua Zhang • *CAS Key Lab of Separation Sciences for Analytical Chemistry, National Chromatographic Research and Analysis Center, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian, China*
Yukui Zhang • *CAS Key Lab of Separation Sciences for Analytical Chemistry, National Chromatographic Research and Analysis Center, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian, China*
Mowei Zhou • *Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA, USA*

Chapter 1

Proteoforms and Proteoform Families: Past, Present, and Future

Lloyd M. Smith

Abstract

The Human Proteoform Project is an ambitious international effort to accelerate the development of technologies for proteoform analysis and to establish comprehensive atlases of proteoforms for humans and model organisms. Proteoforms are the ultimate molecular effectors of function in biology and are thus central to understanding that function. Proteoform analysis as it is practiced today is almost exclusively accomplished by mass spectrometry (MS) and is rapidly advancing in its capabilities. This volume presents a beautiful snapshot of emerging technologies at the exciting frontier of MS-based proteoform analysis.

Key words Proteoform, Proteoform Family, Human Proteoform Project, Mass spectrometry

1 Introduction

The term "proteoform" was first introduced in 2013 and was quickly adopted by the research community [1]. A proteoform is a defined form of a protein from a given gene with a specific amino acid sequence and localized posttranslational modifications. Proteoforms are thus the collection of diverse protein molecules (in many forms) that derive from each gene in the genome. The importance of proteoforms in biology and medicine reflects the fact that understanding of biological systems relies upon knowledge of their elements. Proteoforms are the ultimate molecular effectors of function in biology and as such are central to understanding that function.

Proteomics as it is practiced today is almost exclusively accomplished by mass spectrometry. It comes in three main flavors: "bottom-up," "top-down," and "middle-down." Bottom-up, in which proteins are digested into peptides, is by far the most well-developed and widely used approach; it provides deeper proteome coverage than top-down or middle-down, but at the cost of loss of molecular context— you cannot determine what proteoforms are

Liangliang Sun and Xiaowen Liu (eds.), *Proteoform Identification: Methods and Protocols*,
Methods in Molecular Biology, vol. 2500, https://doi.org/10.1007/978-1-0716-2325-1_1,
© The Author(s), under exclusive license to Springer Science+Business Media, LLC, part of Springer Nature 2022

present from their component peptides. Top-down, in contrast, skips the digestion step and seeks to identify the proteoforms in the mass spectrometer directly. It does allow complete characterization of the intact proteoform, but at the cost of decreased sensitivity, which translates into lower proteome coverage. Middle-down is between the two, seeking to reduce protein sizes to manageable levels by means of controlled and limited digestion, but this approach remains a bit artisanal in nature and is not well suited for comprehensive proteome analysis.

One powerful way to conceptualize proteomics at the proteoform level is through "proteoform families" (Fig. 1) [2, 3]. A proteoform family is the set of proteoforms derived from a given gene. For the ~20,000 genes in human, for example, there would thus be ~20,000 proteoform families, and a comprehensive proteoform-level analysis in human would reveal the members and their abundances detected in the sample for each family. There is much to be done to actualize this vision. Advances in mass spectrometry and other emerging platforms such as nanopore sequencing and cryo-EM are in active development worldwide, and the technological progress is breathtaking.

A concept that may help to drive progress concerns the distinction and transition between "discovery" technologies and "scoring" technologies. A prominent example of this occurred in the Human Genome Project, where the early large-scale sequencing efforts (discovery technology) revealed millions of single-nucleotide polymorphisms that were compiled in public databases such as dbSNP (https://www.ncbi.nlm.nih.gov/snp/); once compiled, however, simpler and less expensive approaches such as hybridization arrays (scoring technology) could be used to screen large numbers of samples, enabling deep analysis across populations in genome-wide association studies (GWAS). This kind of possibility motivates the Human Proteoform Project, an ambitious international effort to accelerate proteoform technology development and to establish comprehensive atlases of proteoforms for humans and model organisms [4]. Such atlases can enable the transition to a scoring technology, by, for example, allowing a relatively simple proteoform mass measurement to serve as a proteoform identifier [2, 3]. Sample-specific databases built on a foundation of multiple data types, such as genomic and transcriptomic nucleic acid sequence data and deep bottom-up proteomic data, will play an important role in establishing and employing proteoform atlases, as they will allow proteoform databases to be built that correctly reflect the particular individual and tissue under study [5, 6].

This volume presents a compendium of cutting-edge emerging tools for proteoform analysis, in a form designed to allow others to be able to practice them. Chapters describe emerging approaches to proteoform separation, data handling and interpretation, and important application areas.

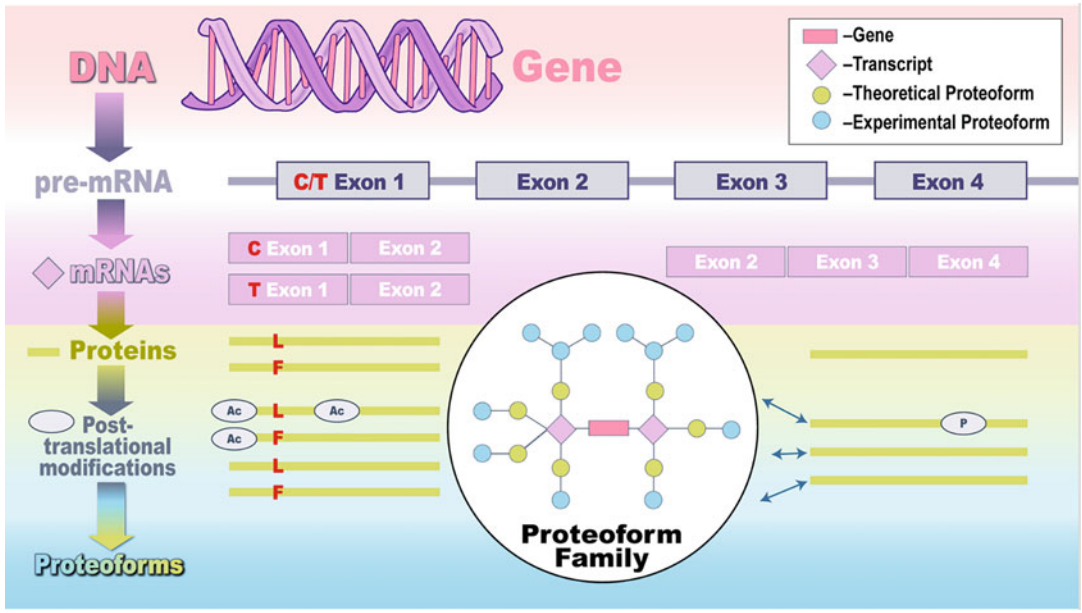


Fig. 1 Illustration of a proteoform family, encompassing sources of protein variation at genomic, transcriptomic, and posttranslational levels

2 Separations

Separations are central to proteomics analyses of complex mixtures due to the fundamental mismatch between the complexity of a proteome and the much more limited ability of a mass spectrometer to resolve mixtures. While a proteome-wide study might seek to reveal thousands or tens of thousands of peptides or proteins, a decipherable high-quality mass spectrum of peptide or protein mixtures will generally have far fewer components, arguably in the range from 1 to 100. While separation strategies for comprehensive peptide analysis are fairly mature and widely practiced, separations of proteins and their proteoforms are much less developed. New strategies and materials for proteoform separations by capillary electrophoresis, size exclusion chromatography, reverse-phase chromatography, and isoelectric focusing are described.

3 Bioinformatic Tools

Bioinformatic tools are similarly critical to proteoform analysis. The complexity of the information required for proteoform-level analysis can be daunting. Complete sequence coverage of even a single proteoform relies upon the acquisition of very complex data; for example, if a proteoform had 500 amino acids, corresponding to a molecular weight of around 50 kDa, uniform cleavage at every

peptide bond would produce 1000 b and y ions, each of which would have only 1/1000th of the ion intensity of the parent ion. In reality, cleavage is not uniform, and the spectrum will be further complicated by internal ions, missing ions, possible co-eluting interferences, and so on. In the case of a proteome-wide analysis, complex data of this sort needs to be obtained for thousands or tens of thousands of components. Powerful bioinformatic tools are clearly needed to rapidly and reliably process these massive data streams, providing the needed information on identities and abundances of the proteoforms present. Development of new algorithms and software tools to this end is a rich and vibrant area of current research, and several such state-of-the art capabilities are presented in this volume.

4 Applications

Applications comprise the third and final set of chapters presented here. In many ways, such applications are "where the rubber hits the road" in state-of-the-art proteoform analysis. Tools are only as good as the results they are able to provide when faced with real-world problems. Three concluding chapters present the application of top-down proteomics for analysis of histone proteoforms for the discovery of new post-translational modifications, for the characterization of metalloproteins, and for the delineation of protein complexes.

Taken together, this volume presents a beautiful snapshot of the status of emerging technologies in the exciting frontier of proteoform analysis. Readers interested in learning about and applying the latest approaches will find it well worth their time.

References

1. Smith LM, Kelleher NL, Consortium for Top-down Proteomics (2013) Proteoform: a single term describing protein complexity. Nat Methods 10:186–187
2. Shortreed MR, Frey BL, Scalf M, Knoener RA, Cesnik A, Smith LM (2016) Elucidating proteoform families from proteoform intact mass and lysine count measurements. J Proteome Res 15: 1213–1221
3. Cesnik A, Shortreed M, Schaffer L, Knoener R, Frey B, Scalf M, Solntsev S, Dai Y, Gasch A, Smith LM (2018) Proteoform Suite: software for constructing, quantifying and visualizing proteoform families. J Proteome Res 17:568–578
4. Smith LM, Agar JN, Chamot-Rooke J, Danis PO, Ge Y, Loo JA, Paša-Tolić L, Tsybin YO, Kelleher NL; Consortium for Top-Down Proteomics (2021) The Human Proteoform Project: Defining the human proteome. Sci Adv 7(46): eabk0734
5. Dai Y, Shortreed M, Scalf M, Frey B, Cesnik A, Solntsev S, Schaffer L, Smith L (2017) Elucidating E. coli proteoform families using intact-mass proteomics and a global PTM discovery database. J Proteome Res 16:4156–4165
6. Cesnik AJ, Miller RM, Ibrahim K, Lu L, Millikin RJ, Shortreed MR, Frey BL, Smith LM (2021) Spritz: a proteogenomic database engine. J Proteome Res 20:1826–1834

Chapter 2

Membrane Ultrafiltration-Based Sample Preparation Method and Sheath-Flow CZE-MS/MS for Top-Down Proteomics

Zhichang Yang and Liangliang Sun

Abstract

Mass spectrometry (MS)-based denaturing top-down proteomics (dTDP) identify proteoforms without pretreatment of enzyme proteolysis. A universal sample preparation method that can efficiently extract protein, reduce sample loss, maintain protein solubility, and be compatible with following up liquid-phase separation, MS, and tandem MS (MS/MS) is vital for large-scale proteoform characterization. Membrane ultrafiltration (MU) was employed here for buffer exchange to efficiently remove the sodium dodecyl sulfate (SDS) detergent in protein samples used for protein extraction and solubilization, followed by capillary zone electrophoresis (CZE)-MS/MS analysis. The MU method showed good protein recovery, minimum protein bias, and nice compatibility with CZE-MS/MS. Single-shot CZE-MS/MS analysis of an *Escherichia coli* sample prepared by the MU method identified over 800 proteoforms.

Key words Denaturing top-down proteomics, Mass spectrometry, Membrane ultrafiltration, Capillary zone electrophoresis-mass spectrometry, Proteoform, Sodium dodecyl sulfate

1 Introduction

Denaturing top-down proteomics (dTDP) aims to characterize proteoforms in cells with high throughput [1–3]. It is becoming an important tool for better understanding of protein structure, post translational modifications (PTMs), and function in biological system. Mass spectrometry (MS)-based dTDP has achieved great advance due to tremendous efforts in the development of proteoform liquid-phase separation [4–14], MS instrumentation [7, 15, 16], and new bioinformatics tools for proteoform identifications (IDs) through database search [17–19], leading to thousands of proteoform IDs from a complex proteome.

A comprehensive proteoform characterization with high throughput cannot do without an efficient and comprehensive extraction of proteins with high recovery, good reproducibility,

Liangliang Sun and Xiaowen Liu (eds.), *Proteoform Identification: Methods and Protocols*,
Methods in Molecular Biology, vol. 2500, https://doi.org/10.1007/978-1-0716-2325-1_2,
© The Author(s), under exclusive license to Springer Science+Business Media, LLC, part of Springer Nature 2022

minimum bias, and free of MS incompatible salts, chaotropes, and detergents [20]. Protein extraction is normally implemented with assistance of additives such as detergents and chaotropic reagents for a thorough protein extraction and denaturing. However, these detergents and chaotropic reagents need to be removed before MS analysis as they can cause significant ion suppression. Multiple strategies were developed for protein cleanup including membrane ultrafiltration (MU) [21], chloroform–methanol precipitation (CMP) [22], and single-spot solid-phase sample preparation using magnetic beads (SP3) [23, 24]. Membrane ultrafiltration (MU) has been widely used by the bottom-up proteomics community for filter-aided sample preparation (FASP) method to remove sodium dodecyl sulfate (SDS) before enzymatic digestion of proteins [21]. Basically, a protein sample in 1–5% (w/v) SDS solution is loaded onto a commercialized membrane filter unit with a 10- to 30-kDa molecular weight cutoff (MWCO), followed by washing with a 8 M urea solution to remove SDS, which is based on the fact that 8 M urea can destroy the hydrophobic interaction between SDS and proteins. The cleaned protein can then be recovered with designated buffer for dTDP analysis or be subjected to enzyme digestion for bottom-up analysis. We systematically compared MU, SP3, and CMP methods for protein clean-up for dTDP analysis through capillary zone electrophoresis mass spectrometry (CZE-MS) and concluded that MU method can be a universal strategy for dTDP sample preparation, due to its nature of high efficiency, high protein recovery, comprehensiveness, and compatibility with downstream MS analysis. A workflow of MU sample processing for dTDP is shown in Fig. 1 [25]. With sample prepared from MU strategy, we applied dynamic pH junction-based CZE-MS/MS on *Escherichia coli* proteoform analysis and achieved over 800 proteoform IDs in a single shot. In this chapter, we have provided a detailed description on the MU strategy and highlighted some critical steps for high protein recovery.

Fig. 1 Workflow of the dTDP sample preparation using membrane ultrafiltration. (Reproduced from ref. 25 with permission from American Chemical Society, copyright (2020))

2 Materials

All reagents were purchased from Sigma-Aldrich (St. Louis, MO) unless stated otherwise. All solvents were prepared with water and reagents at LC-MS grade. All processing buffers/solutions need to be prepared fresh.

2.1 Prepare Protein for MU Cleanup

1. LB (Luria-Bertani) medium for *E. coli* culture.
2. Sterile phosphate-buffered saline (PBS) without calcium and magnesium.
3. 50 mL Falcon tube.
4. 500 mL conical flask.
5. *E. coli* (strain K-12 substrain MG1655).
6. Cell lysis buffer: 1%(w/w) SDS, 100 mM NH_4HCO_3 (pH 8.0), 5 mg/mL protease inhibitor, 5 mg/mL phosphatase inhibitor. Store at 4 °C.
7. Centrifuge (compatible with 1.5 mL Eppendorf tube with maximum speed of 14,800 rpm; e.g., Thermo Scientific Legend Micro 21).
8. Ultrasonication (capable of inducing cell disruption, homogenization, and emulsification through cavitation; e.g., Branson Sonifier 250).

2.2 Membrane Ultrafiltration

1. Urea buffer: 8 M urea, 100 mM NH_4HCO_3 (pH 8.0).
2. NH_4HCO_3 buffer: 100 mM NH_4HCO_3 (pH 8.0).
3. 30 kDa MWCO membrane unit (Millipore: MRCF0R030).

2.3 Prepare LPA (Linear Polyacrylamide)-Coated Capillary

1. Capillary electrophoresis (CE) separation capillary: 360/50 (OD/ID) fused silica capillary.
2. 1 M NaOH.
3. 1 M HCl.
4. γ-MAPS solution: 50%(v/v) 3-(trimethoxysilyl)propyl-methacrylate (in methanol (MeOH)).
5. Acrylamide solution: 40 mg/mL acrylamide solution.
6. APS buffer: 5% w/v ammonium persulfate buffer (APS).
7. Coating solution: mix 3.5 μL APS and 500 μL acrylamide solution.
8. Hydrofluoric acid: 40% hydrofluoric acid.
9. 10 M NaOH solution.
10. 600 μL Eppendorf tubes.

2.4 CZE-MS

1. Background electrolyte (BGE): 20% acetic acid.
2. Sheath buffer: 10% MeOH, 0.2% formic acid.
3. CE autosampler: CMP Scientific autosampler.
4. CE interface: electrokinetically pumped sheath flow CE-MS interface (EMASS II, CMP Scientific, Brooklyn, NY).
5. Mass spectrometer: Q-Exactive HF mass spectrometer (Thermo Fisher Scientific).

3 Method

3.1 Prepare Protein for MU Cleanup

1. Culture the *E. coli* (strain K-12 substrain MG1655) in the LB (Luria-Bertani) medium at 37 °C until OD600 reached 0.7.
2. Harvest the *E. coli* cells by centrifugation at 4000 rpm for 10 min. Wash the cell pellets with PBS three times to remove the leftover culture medium.
3. Add 400 μL cell lysis buffer into the cell pellet. Pipette up and down a few times to lysis the cell. For thorough lysis, immerse sample tube into ultrasonication equipment filled with ice water and sonicate for 10 min (*see* **Note 1**).
4. Spin down the sample tube at a speed of 14,000 g for 5 min. After spinning down, protein is in supernatant. Take the supernatant out carefully and measure the protein concentration of the supernatant by BCA protein assay. Aliquot the protein into a 600 μL Eppendorf tube with 100 μg in each tube and store the protein at −80 °C before use.

3.2 Ultrafiltration Buffer Exchange

1. Rinse the ultrafiltration membrane with 200 μL 100 mM NH_4HCO_3 buffer for membrane pretreatment.
2. Dilute the protein solution with 8 M urea buffer to a concentration of up to 1 μg/μL if the protein original concentration is too high (*see* **Note 2**).
3. Load the protein solution onto membrane and spin down the ultrafiltration unit with high speed until all solution go through the membrane or the volume of the solution doesn't reduce with more spinning (*see* **Note 3**).
4. Wash the membrane with 100 μL 8 M urea buffer at least two times and 100 μL 100 mM NH_4HCO_3 buffer three times through high speed (<14,000 g) spin-down (*see* **Note 4**).
5. Add 50–100 μL of 100 mM NH_4HCO_3 onto the membrane, and pipette up and down a couple times to resuspend the protein left on the membrane. Additional resuspension can be performed through 5 min of vortex.
6. Flip the membrane unit onto a new collection tube and collect the protein solution through a quick spin down.

7. Measure the protein concentration with BCA protein assay to estimate protein recovery and to adjust the amount of protein to be loaded for follow-up MS identification (*see* **Note 5**).

3.3 Prepare Capillary for CE-MS

1. Capillary pretreatment: Flush the capillary with following buffer in order: 1 M NaOH for more than 30 min, water for 30 min, 1 M HCl for 30 min, water for 30 min, MeOH for 30 min.
2. Dry the capillary by flushing with nitrogen gas for 10 min.
3. Flush the capillary with 50% 3-(trimethoxysilyl)propyl-methacrylate for 10 min and seal both ends with rubber. Incubate the capillary at room temperature for at least 24 h.
4. Rinse the capillary with MeOH and dry the capillary with nitrogen gas.
5. Capillary coating: Remove oxygen from the LPA coating solution by introducing pure nitrogen into the solution. Introduce LPA coating solution into the capillary and seal both ends with rubber. Incubate the capillary in a 50 °C water bath for 50 min. *See* **Note 6.**
6. Flush the capillary with water and store the capillary at room temperature, if not use right away. *See* **Note 7.**
7. Etch capillary: To accommodate the capillary into interface emitter with minimum sample dilution. One end of capillary needs to be etched with hydrofluoric (HF) acid to reduce the outer diameter. Use lighter fire to burn out the outer coating of the capillary, in the middle part that is about 1 cm away from the end, to create a 1–2 cm etching segment. Wipe out the coating residue with wet paper towel. Install the capillary into a 600 μL Eppendorf tube through a hole pierced at the bottom of the tube so that the etching segment is in the Eppendorf tube and the pierced hole is clogged tightly by the capillary (as shown in Fig. 2). Add 60–80 μL HF acid into the tube so that the HF acid can cover part of the etching segment. Leave the capillary in HF acid for 90 min. Carefully remove the HF acid in the tube and neutralize the HF acid using 10 M NaOH buffer. Rinse the etched end with water. Cut off the etching segment and leave a desired length of etched capillary (~5 mm) (*see* **Note 8**).

3.4 CE-MS

1. Install the etched tip of the capillary into the glass emitter of the interface. The glass emitter (orifice controlled at 20–40 μm) needs to be filled with sheath buffer. Control the distance of the etched capillary tip to the emitter orifice less than 500 μm and the distance of the emitter orifice to the MS entrance around 2 mm.

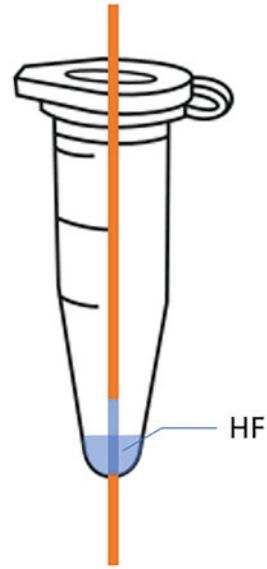


Fig. 2 Setup of capillary etching

2. Perform the CE separation with the autosampler. Set the electrospray ionization (ESI) voltage at 2 kV. Flush the capillary with BGE at 15 psi for 10 min. For 500 nL sample injection, immerse the distal end of the capillary into sample solution and apply 5 psi for 90 s (*see* **Note 9**). For separation, immerse the distal end of the capillary back into BGE buffer and apply 30 kV voltage for 115 min. Flush the capillary at 15 psi for 10 min for capillary cleanup and equilibrate after separation is finished. The example electropherograms of the prepared *E. coli* sample are shown in Fig. 3. *See* **Note 10**.
3. During the separation, MS instrument is operated in data-dependent mode. Acquire MS1 data with full scan range of 600–2000 m/z. Set the MS1 resolution at 120,000 FWHM (at 200 m/z); set the AGC 1E6 and set maximum injection time as 50 ms. Allow up to three precursor ions for higher energy collisional dissociation (HCD) fragmentation in tandem mass spectra. Acquire MS2 spectra at 60,000 resolution (at 200 m/z), set AGC target value of 1E5 and max injection time of 200 ms. Set precursor ion isolation width to 4 m/z. Set the dynamic exclusion as 30 s.
4. Data analysis: Perform database search with TopPIC Suite [17]. Convert RAW files into mzML file using msconvert tool [26]. Process the mzML file with TopFD software for spectral deconvolution. Process the msalign files that resulted from spectral deconvolution with TopPIC software for database search. Specify the designated database, uniport *E. coli* in this case. Specify the maximum shift value, 1 in this case.

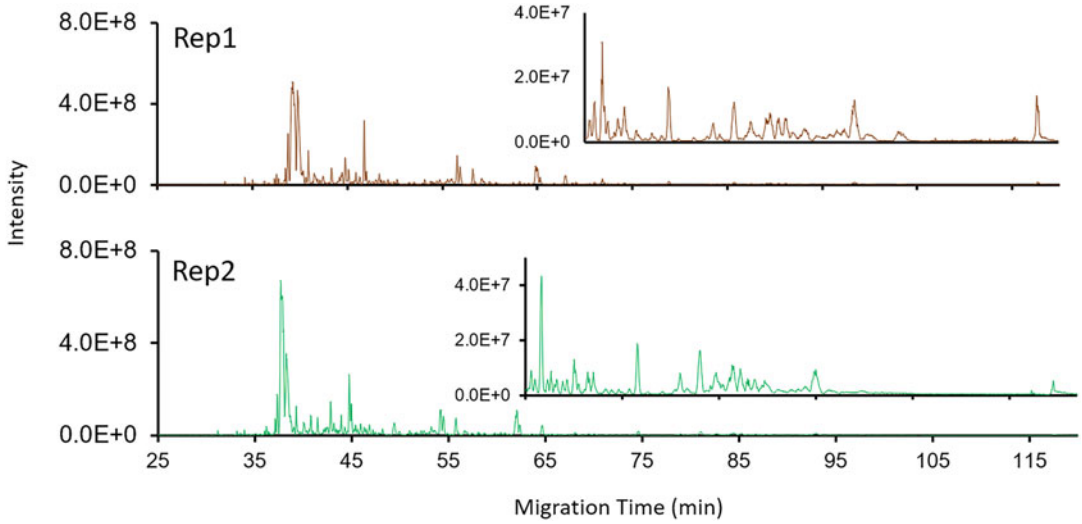


Fig. 3 Electropherograms of dTDP CE-MS analysis on two replicates of *E. coli* proteome processed by membrane ultrafiltration. (Reproduced from ref. 25 with permission from American Chemical Society, copyright (2020))

Specify the approach for proteome false discovery rate (FDR) evaluation, target-decoy in this case. Set the PrSM-level FDR to 1% and proteoform level to 5%. *See* **Note 11**.

4 Notes

1. For thorough cell lysis, sonication was set at a power output of 6. To avoid accumulated heating during long session of sonication, sonication session can be divided into multiple sessions with break of 30 s.
2. Total protein amount loaded onto the ultrafiltration membrane should not exceed 100 μg to avoid protein precipitation. It is recommended to dilute the protein with urea first to disturb the interaction of protein and SDS and facilitate protein dissolution.
3. During through centrifugation, try not to exceed 14,000 g for centrifugation speed as too high of speed might cause membrane clogging with protein molecules and protein loss. A spin-down on the protein solution to remove any precipitate is also recommended before ultrafiltration.
4. Make sure that minimum solution is left above the membrane after each cycle of wash.
5. An SDS-PAGE analysis can be performed to check the comprehensiveness of sample preparation (whether specific MW proteins are missing compared to pre-processed sample).

6. When preparing LPA coating solution, the degas step is crucial as oxygen will influent the polymerization process. A continuous bubbling from the degassing tube that is immersed in the coating solution should be maintained for an efficient degas. Avoid disturbing the coating solution after the degas step and introduce the coat solution into the capillary with vacuum pump. The incubation time should be at least 50 min. Insufficient incubation could result in incomplete reaction and unstable coating. Incubating for too long (for example 1.5 h) could lead to capillary clogging.
7. When the reaction time is too long, it is hard to push the polymer in the capillary out. HPLC pump can be used to flush the capillary in this case.
8. Hydrofluoric acid is very dangerous and needs to be processed with care in the hood. Follow appropriate safety procedures. All the tubes and tips involved in the etching steps need to be placed in proper trash container.
9. The total loading amount of protein should be adjusted. The loading amount in Fig. 3 was 400 ng. Too high of loading amount can cause protein precipitation during CE separation and odd current flow chart. Dilute the protein with 100 mM NH_4HCO_3 if the concentration is too high. A loading amount of 100–500 ng should be applicable.
10. It is proved that proteins extracted with SDS as additive tend to be more hydrophobic than those extracted with urea as additive [25]. Hydrophobic protein is more insoluble in aqueous sample buffer such as NH_4HCO_3 and can precipitate out during CE separation. To facilitate the dissolution, we applied 20% acetic acid as BGE as opposed to 5% that was regularly used in our previous studies.
11. 1% proteoform FDR can also be used for more strict setting. A total of 800 proteoforms can be identified typically using the CZE-MS/MS system from the *E. coli* sample in a single run with 5% FDR at proteoform level.

Acknowledgments

We thank Prof. Heedeok Hong's group at the Department of Chemistry of Michigan State University for kindly providing the *E. coli* cells for this project. We thank the support from the National Institute of General Medical Sciences (NIGMS) through Grant R01GM125991 and the National Science Foundation through Grant DBI1846913 (CAREER Award).

References

1. Smith LM, Kelleher NL (2013) Consortium for Top Down, P., Proteoform: a single term describing protein complexity. Nat Methods 10:186–187
2. Toby TK, Fornelli L, Kelleher NL (2016) Progress in top-down proteomics and the analysis of proteoforms. Annu Rev Anal Chem 9: 499–519
3. Smith LM, Kelleher NL (2018) Proteoforms as the next proteomics currency. Science 359: 1106–1107
4. 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 (2011) Mapping intact protein isoforms in discovery mode using top-down proteomics. Nature 480:254–258
5. Catherman AD, Durbin KR, Ahlf DR, Early BP, Fellers RT, Tran JC, Thomas PM, Kelleher NL (2013) Large-scale top-down proteomics of the human proteome: membrane proteins, mitochondria, and senescence. Mol Cell Proteomics 12:3465–3473
6. Ahrne E, Martinez-Segura A, Syed AP, Vina-Vilaseca A, Gruber AJ, Marguerat S, Schmidt A (2015) Exploiting the multiplexing capabilities of tandem mass tags for high-throughput estimation of cellular protein abundances by mass spectrometry. Methods 85:100–107
7. Anderson LC, DeHart CJ, Kaiser NK, Fellers RT, Smith DF, Greer JB, LeDuc RD, Blakney GT, Thomas PM, Kelleher NL, Hendrickson CL (2017) Identification and characterization of human proteoforms by top-down LC-21 tesla FT-ICR mass spectrometry. J Proteome Res 16:1087–1096
8. Cai W, Tucholski T, Chen B, Alpert AJ, McIlwain S, Kohmoto T, Jin S, Ge Y (2017) Top-down proteomics of large proteins up to 223 kDa enabled by serial size exclusion chromatography strategy. Anal Chem 89: 5467–5475
9. Liang Y, Jin Y, Wu Z, Tucholski T, Brown KA, Zhang L, Zhang Y, Ge Y (2019) Bridged hybrid monolithic column coupled to high-resolution mass spectrometry for top-down proteomics. Anal Chem 91:1743–1747
10. McCool EN, Lubeckyj RA, Shen X, Chen D, Kou Q, Liu X, Sun L (2018) Deep top-down proteomics using capillary zone electrophoresis-tandem mass spectrometry: identification of 5700 proteoforms from the Escherichia coli proteome. Anal Chem 90: 5529–5533
11. Yu D, Wang Z, Cupp-Sutton KA, Liu X, Wu S (2019) Deep intact proteoform characterization in human cell lysate using high-pH and low-pH reversed-phase liquid chromatography. J Am Soc Mass Spectrom 30:2502–2513
12. 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 (2013) Top-down proteomics reveals a unique protein S-thiolation switch in Salmonella Typhimurium in response to infection-like conditions. Proc Natl Acad Sci U S A 110:10153–10158
13. Lubeckyj RA, Basharat AR, Shen X, Liu X, Sun L (2019) Large-scale qualitative and quantitative top-down proteomics using capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry with nanograms of proteome samples. J Am Soc Mass Spectrom 30:1435–1445
14. Liu Z, Wang R, Liu J, Sun R, Wang F (2019) Global quantification of intact proteins via chemical isotope labeling and mass spectrometry. J Proteome Res 18:2185–2194
15. Riley NM, Westphall MS, Coon JJ (2017) Activated ion-electron transfer dissociation enables comprehensive top-down protein fragmentation. J Proteome Res 16:2653–2659
16. Shaw JB, Li W, Holden DD, Zhang Y, Griep-Raming J, Fellers RT, Early BP, Thomas PM, Kelleher NL, Brodbelt JS (2013) Complete protein characterization using top-down mass spectrometry and ultraviolet photodissociation. J Am Chem Soc 135:12646–12651
17. Kou Q, Xun L, Liu X (2016) TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. Bioinformatics 32:3495–3497
18. Cai W, Guner H, Gregorich ZR, Chen AJ, Ayaz-Guner S, Peng Y, Valeja SG, Liu X, Ge Y (2016) MASH Suite Pro: a comprehensive software tool for top-down proteomics. Mol Cell Proteomics 15:703–714
19. Sun RX, Luo L, Wu L, Wang RM, Zeng WF, Chi H, Liu C, He SM (2016) pTop 1.0: a high-accuracy and high-efficiency search engine for intact protein identification. Anal Chem 88: 3082–3090
20. Donnelly DP, Rawlins CM, DeHart CJ, Fornelli L, Schachner LF, Lin Z, Lippens JL, Aluri KC, Sarin R, Chen B, Lantz C, Jung W, Johnson KR, Koller A, Wolff JJ, Campuzano IDG, Auclair JR, Ivanov AR, Whitelegge JP,

Pasa-Tolic L, Chamot-Rooke J, Danis PO, Smith LM, Tsybin YO, Loo JA, Ge Y, Kelleher NL, Agar JN (2019) Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. Nat Methods 16:587–594

21. Wisniewski JR, Zougman A, Nagaraj N, Mann M (2009) Universal sample preparation method for proteome analysis. Nat Methods 6:359–362
22. Wessel D, Flugge UI (1984) A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. Anal Biochem 138:141–143
23. Hughes CS, Foehr S, Garfield DA, Furlong EE, Steinmetz LM, Krijgsveld J (2014) Ultrasensitive proteome analysis using paramagnetic bead technology. Mol Syst Biol 10: 757
24. Hughes CS, Moggridge S, Muller T, Sorensen PH, Morin GB, Krijgsveld J (2019) Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. Nat Protoc 14:68–85
25. Yang Z, Shen X, Chen D, Sun L (2020) Toward a universal sample preparation method for denaturing top-down proteomics of complex proteomes. J Proteome Res 19: 3315–3325
26. Kessner D, Chambers M, Burke R, Agus D, Mallick P (2008) ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics 24:2534–2536

Chapter 3

Size Exclusion Chromatography Strategies and MASH Explorer for Large Proteoform Characterization

Timothy N. Tiambeng, Zhijie Wu, Jake A. Melby, and Ying Ge

Abstract

Top-down mass spectrometry (MS)-based analysis of larger proteoforms (>50 kDa) is typically challenging due to an exponential decay in the signal-to-noise ratio with increasing protein molecular weight (MW) and coelution with low-MW proteoforms. Size exclusion chromatography (SEC) fractionates proteins based on their size, separating larger proteoforms from those of smaller size in the proteome. In this protocol, we initially describe the use of SEC to fractionate high-MW proteoforms from low-MW proteoforms. Subsequently, the SEC fractions containing the proteoforms of interest are subjected to reverse-phase liquid chromatography (RPLC) coupled online with high-resolution MS. Finally, proteoforms are characterized using MASH Explorer, a user-friendly software environment for in-depth proteoform characterization.

Key words Size exclusion chromatography, Proteoforms, Data analysis, Top-down proteomics

1 Introduction

Top-down mass spectrometry (MS)-based proteomics is a powerful technology because of its capability to characterize proteoforms arising from sequence variations, alternative splicing, and post-translational modifications (PTMs) [1–8]. Despite the capability of top-down proteomics, high molecular weight (MW) proteoforms are often under-represented in the MS analysis of the proteome [9–11]. This challenge arises from both the high dynamic range of the proteome, in which protein expression can vary in orders of magnitude, and the exponential decay in the MS signal-to-noise ratios of large proteoforms, resulting mainly from the increased number of charge states observed with electrospray ionization of proteoforms, the greater contribution of heavy isotopes at higher precursor mass, and the detrimental influence of adducting and interfering species [12]. MS identification and characterization of large proteoforms are particularly challenging if they are coeluted with smaller proteoforms [9, 10]. Moreover,

Liangliang Sun and Xiaowen Liu (eds.), *Proteoform Identification: Methods and Protocols*,
Methods in Molecular Biology, vol. 2500, https://doi.org/10.1007/978-1-0716-2325-1_3,
© The Author(s), under exclusive license to Springer Science+Business Media, LLC, part of Springer Nature 2022

user-friendly software tools remain underdeveloped for proteoform identification and comprehensive characterization of large proteoforms.

Size exclusion chromatography (SEC) addresses a critical challenge in top-down MS-based analysis of large, high-MW proteoforms [9, 10, 13, 14]. This chromatographic technique fractionates the proteome by separating proteoforms based on their size or hydrodynamic volumes, while reducing sample loss due to minimal interactions of proteins with the SEC stationary phase [13, 14]. SEC is usually performed using hydrophilic stationary phases with well-defined pore diameters [15]. Typically, smaller proteoforms more readily diffuse into the pores, while relatively larger proteoforms are less likely to enter the pores. Thus, proteoforms are separated based on their size as they pass through the column and are eluted in order of decreasing molecular weight [16]. SEC can be used as an effective and versatile separation method that is highly compatible with top-down MS-based proteomics and is orthogonal to other chromatographic methods. SEC experiments can be performed in both denatured and native modes, using mobile phases such as formic acid in water [10] and ammonium acetate solution [17], respectively. The resulting SEC fractions containing proteoforms eluting at specific chromatographic elution times can then be subjected to high-resolution MS with simple sample processing. As a notable example, Cai and Tucholski et al. developed serial size exclusion chromatography (sSEC) by connecting together SEC columns in series with different stationary phase pore diameters to achieve efficient size-based separation over a broad MW range [9]. When sSEC fractions were further separated by RPLC coupled online with high-resolution MS, it enabled a 15-fold improvement in the observation of proteoforms greater than 60 kDa compared to one-dimensional RPLC-MS alone. Additionally, Tucholski et al. utilized sSEC fractionation directly with Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS) analysis to enable sequence characterization of large proteoforms without RPLC separation or extensive protein purification [10].

For characterizing the primary sequence and localizing the PTMs of proteoforms, MASH Explorer is a universal, user-friendly, and freely available software environment for top-down proteomics [18]. MASH Explorer is built upon the previous successes of MASH Suite [19] and MASH Suite Pro [20] to bolster the continued growth of the top-down proteomics community. In comparison to MASH Suite Pro, MASH Explorer can process MS, tandem MS (MS/MS), and liquid chromatography tandem MS (LC-MS/MS) across multiple vendor-specific-formats, with automated database searching for protein identification and tools for proteoform characterization and data validation. MASH Explorer incorporates various deconvolution and database-

searching algorithms to assist in fragment ion identification and large proteoform characterization. MASH Explorer provides an effective and comprehensive solution to process many types of MS data and can be used in combination with other freely available software tools, such as Proteoform Suite [21, 22], to assist in the identification of large proteoforms [11] that are not monoisotopically resolved.

Here in this protocol, we demonstrate that SEC can be generally used to fractionate high-MW proteoforms from low-MW proteoforms. SEC fractions are then subjected to online LC-MS and LC-MS/MS analysis using a quadrupole time-of-flight (QTOF) instrument. Subsequently, MASH Explorer [18] is used to process and analyze MS data to characterize the sequence and PTMs of proteoforms.

2 Materials

All reagents were purchased from MilliporeSigma Inc. (St. Louis, MO, USA) unless otherwise noted. HPLC-grade solvents such as water, acetonitrile, ethanol, isopropanol, and formic acid were purchased from Fisher Scientific (Fair Lawn, NJ, USA) (*see* **Note 1**).

2.1 Size Exclusion Chromatography

1. Waters ACQUITY H-Class UPLC system equipped with a UV detector and an automatic fraction collector (Waters Corporation, Milford, MA, USA).
2. PolyHYDROXYETHYL A (PolyHEA) columns (PolyLC Inc., Columbia, MD, USA) (*see* **Note 2**).
3. Mobile phase, such as 1% formic acid in water (v/v) (*see* **Note 3**).
4. Ultra-centrifugal 10 kDa molecular weight cutoff (MWCO) filters (0.5 mL) (Fisher Scientific, Fair Lawn, NJ, USA).
5. 1/16″ OD (outer diameter) PEEK tubing with 100 μm ID (inner diameter), 2 cm (VICI Valco instruments, Houston, TX, USA).
6. Supelco® 1/16″ OD PEEK ferrules and fittings (MilliporeSigma Inc., St. Louis, MO, USA).

2.2 Online Reverse-Phase Liquid Chromatography Mass Spectrometry

1. maXis II quadrupole time-of-fight (QTOF) mass spectrometer (Bruker Daltonics, Bremen, Germany).
2. Waters nanoACQUITY UPLC system (Waters Corporation, Milford, MA, USA).
3. Home-packed PLRP column, (200 × 0.5 mm ID with PLRP-S bulk media (10 μm, 1000 Å, Agilent Technologies, Santa Clara, CA, USA) (*see* **Note 4**).

4. 1/16″ OD PEEK tubing with 100 μm ID, 20 cm (VICI Valco instruments, Houston, TX, USA).
5. Supelco® 1/16″ OD PEEK ferrules and fittings (MilliporeSigma Inc., St. Louis, MO, USA).
6. Mobile phase A: 0.1% formic acid in water (v/v).
7. Mobile phase B: 0.1% formic acid in 50:50 acetonitrile/ethanol (v/v).
8. Absorbance microplate reader, such as the BioTek ELx808 Absorbance Reader (Winooski, VT, USA).
9. Bradford protein assay reagents, such as the Quick Start™ Bradford Protein Assay Kit (Hercules, CA, USA).

2.3 MASH Explorer Data Analysis

1. Personal computer meeting the minimum hardware requirements (*see* **Note 5**).
2. MASH Explorer software, version 2.1.1. (https://labs.wisc.edu/gelab/software.html).
3. MASH Explorer User Installation Guide: (https://labs.wisc.edu/gelab/MASH_Explorer/doc/User%20Installation%20Guide.pdf) (*see* **Note 6**).

3 Methods

3.1 Size Exclusion Chromatography

1. For new and previously unused PolyHEA columns, flush the column with 15 column volumes of water to remove methanol.
2. Condition the column(s) by flushing with the intended mobile phase overnight at room temperature to ensure reproducible retention times.
3. For serial SEC, connect the desired number of SEC columns together using 2 cm segments of 1/16″ OD PEEK tubing (100 μm ID) and 1/16″ OD PEEK ferrules and fittings.
4. Ensure that an appropriately sized sample loop is installed on the UPLC system for SEC or serial SEC applications (*see* **Note 7**).
5. Set the flow rate to 0.5 mL/min and perform three sample injections using water at the same flow rate and isocratic gradient as the intended sample to ensure stable baseline and column backpressure. Ensure that the sample manager is kept at 4 °C to reduce artifactual temperature-induced protein modification such as oxidation (*see* **Note 8**).
6. As standards for reproducibility, inject 5 μg of protein solution that is buffer exchanged with the starting SEC mobile phase (*see* **Note 9**).

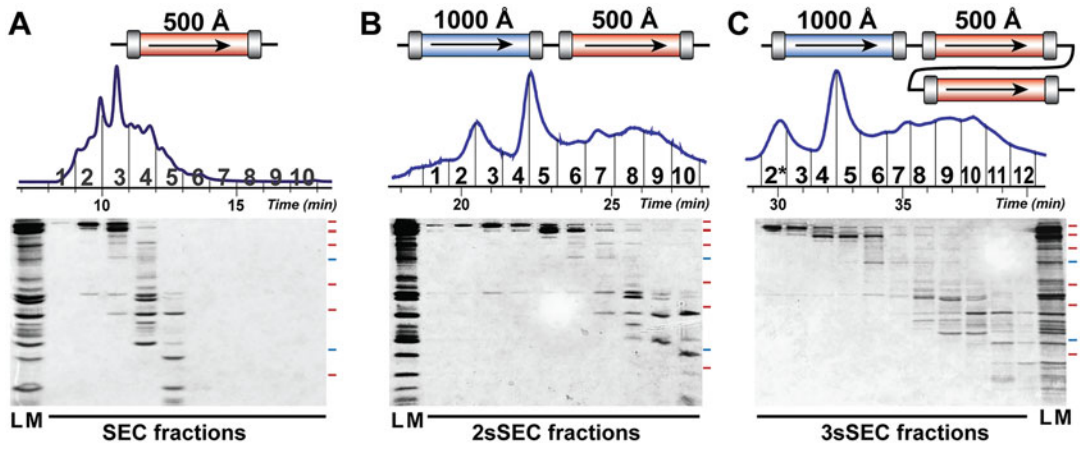


Fig. 1 Example of an SEC separation of a complex protein mixture. Comparison of (**a**) single column SEC (500 Å), (**b**) SEC with two columns connected serially (1000 Å–500 Å), and (**c**) SEC with three columns connected serially (1000 Å–500 Å–500 Å) for the fractionation of the same protein loading mixture (LM). The top panel illustrates the UV chromatogram for the corresponding SEC experiment, annotated with numbers corresponding to the collected SEC fractions. The bottom panel illustrates the SDS-PAGE analysis corresponding to the collected and annotated SEC fractions. (Adapted and reprinted with permission from ref. 9. Copyright 2017 American Chemical Society)

7. Monitor protein elution from the SEC column by recording the change in UV absorbance at 280 nm as a function of retention time (*see* **Note 10**).
8. For fraction collection, begin collecting fractions when the absorbance intensity increases about 10 times the signal-to-noise ratio, or at the approximate elution time of the protein of interest based on the elution times of the standard proteins (*see* **Note 11**). Ensure that the fraction manager is kept at 4 °C to reduce artifactual temperature-induced protein modification.
9. Concentrate the desired protein-containing SEC fractions using 10 kDa MWCO (*see* **Note 12**) for SDS-PAGE (*see* **Note 13**) and LC-MS/MS analysis (Fig. 1).

3.2 Online Reverse-Phase Chromatography and Top-Down MS Analysis

1. Prepare LC-MS grade 0.1% formic acid in water (v/v) and LC-MS grade 0.1% formic acid in 50:50 acetonitrile:ethanol (v/v).
2. Dilute the protein fractions to a total protein concentration of 100 ng/μL total protein with 0.1% formic acid in water containing 2 mM TCEP (*see* **Note 14**).
3. Load 500 ng of total protein for a 0.5 mm ID PLRP-S column. Separate the proteins chromatographically using a nanoAcquity UPLC system equipped with a PLRP-S column. Connect the PLRP-S column to the electrospray ionization source using

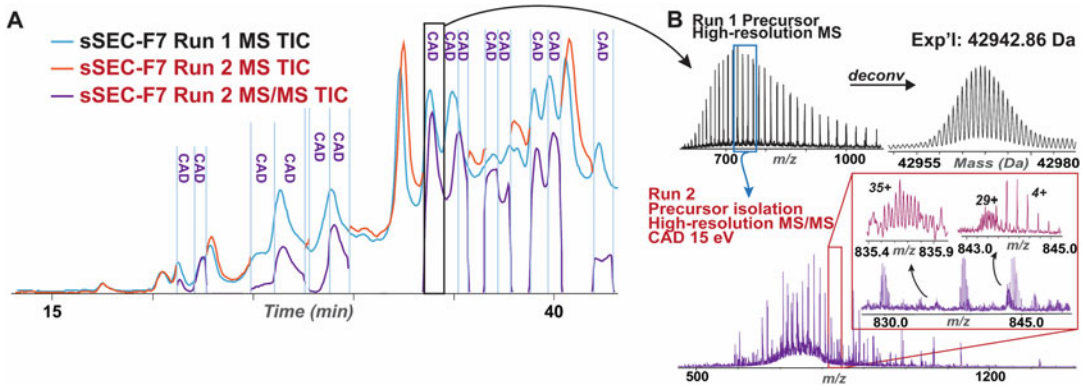


Fig. 2 Top-down targeted MS/MS for protein identification in a selected sSEC fraction. (**a**) CAD-based MS/MS experiments are performed on selected LC retention time windows containing proteoforms of interest. (**b**) High-resolution MS demonstrating representative precursor ion isolation and effective CAD-based fragmentation of a 42.9 kDa protein, yielding high-resolution tandem mass spectra. (Adapted and reprinted with permission from ref. 9 Copyright 2017 American Chemical Society)

1/16″ OD PEEK tubing with 100 μm ID and 1/16″ OD PEEK ferrules and fittings.

4. Elute proteins using a mobile-phase gradient going from 5% B to 95% B over 45 min at a flow rate of 8 μL/min (mobile phase A: 0.1% formic acid in water, mobile phase B: 0.1% formic acid in 50:50 acetonitrile:ethanol) (*see* **Note 15**).
5. For the electrospray ionization source on the QTOF mass spectrometer, set the "End Plate Offset" at 500 V, the "Capillary" at 4500 V, the "Nebulizer" at 0.5 bar, the "Dry Gas" at 4.0 L/min, and the "Dry Temp" at 220 °C. Collect mass spectra at a scan rate of 0.5 Hz over 500–2000 *m/z* range.
6. Complete an initial LC-MS run on an SEC protein fraction of interest and perform data analysis to determine the charge state ions, molecular weight, and retention time of each target protein.
7. For targeted protein analysis, complete a second LC-MS run using online collisionally activated dissociation (CAD) by inputting the precursor ion(s), ion isolation range (*m/z*), CAD energy, and the chromatographic elution interval. The precursor ion(s) will be isolated and fragmented in the targeted MS/MS experiment for protein identification (Fig. 2) (*see* **Note 16**).

3.3 Data Analysis Using MASH Explorer's Targeted Mode

1. Import raw MS/MS data files into the MASH Explorer software under Targeted Mode (Fig. 3) (*see* **Note 17**).
2. Perform spectral deconvolution using algorithms such as THRASH [23] or TopFD [24] (*see* **Note 18**). After selecting a deconvolution method, proceed to the Advanced tab to

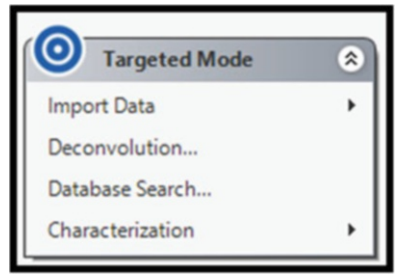


Fig. 3 MASH Explorer's Targeted Mode interface for MS/MS data import

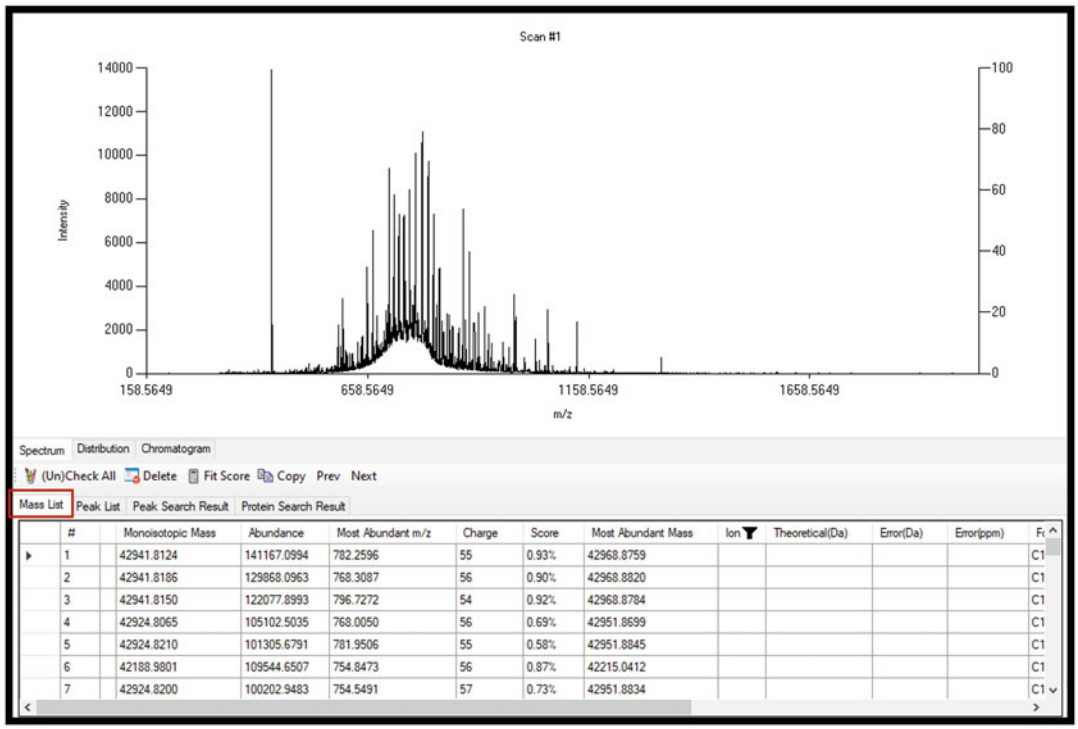


Fig. 4 MS/MS spectral deconvolution and the calculated mass list imported by MASH Explorer software from the selected deconvolution algorithm, displayed in MASH Explorer's Mass List panel (highlighted in red)

change settings such as the precursor ion *m/z* and charge in the General subheading, as well as specific algorithm parameters such as Max Charge and Max Mass in the Deconvolution subheading. A list of deconvoluted fragment ions will be generated by the selected algorithm and automatically imported back to the MASH Explorer software interface (Fig. 4).

3. Verify the deconvoluted fragment ions in the Mass List by adjusting their charge states and monoisotopic mass. A more precise list of fragment ions will allow for shorter processing time and more accurate proteoform identifications in the database search.

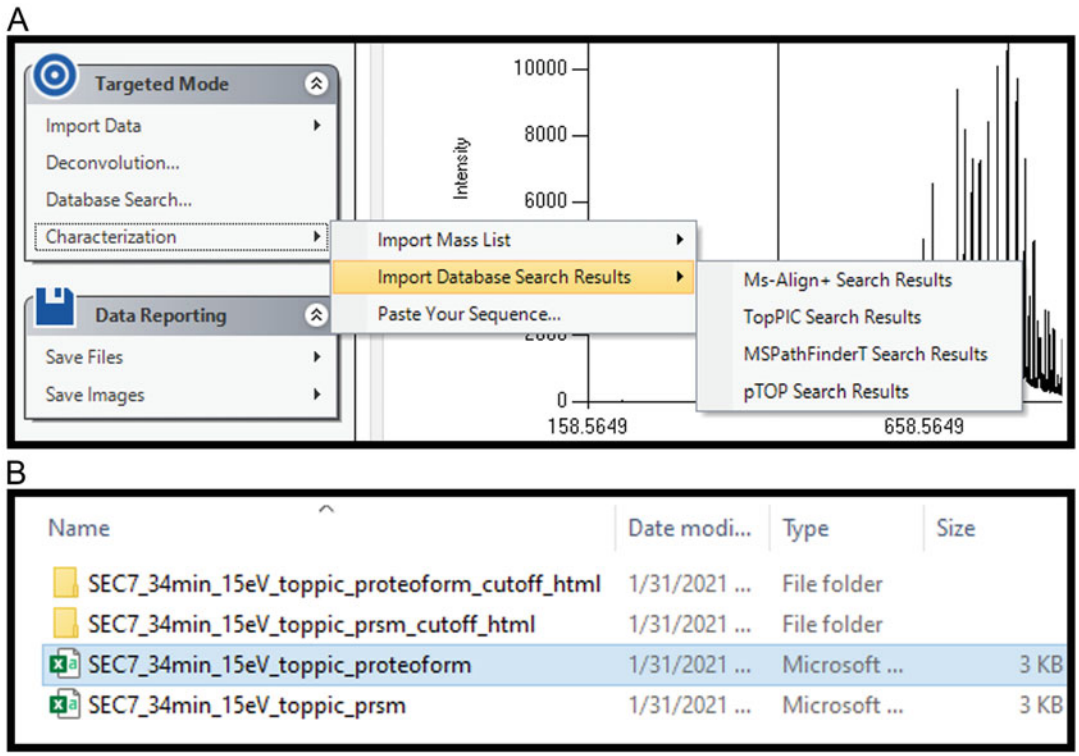


Fig. 5 MASH Explorer enables import of database search results. (**a**) MASH Explorer supports multiple database search algorithms including TopPIC [24], MS-Align+ [25], pTop [26], and Informed-Proteomics [27] workflows. (**b**) Highlighted .CSV file containing database search results from the "MASH_Workflows" folder

4. Using the verified fragment ion list, perform database search using algorithms such as MS-Align+ [25] (*see* **Note 19**). The precursor ion information obtained during data analysis discussed in Subheading 3.2, such as the precursor ion *m/z* and charge, will assist the database search algorithm to identify targeted proteoforms more accurately.
5. Import the identified proteoform identification after completing the database search. Under the Characterization heading, select Import Database Search Results, and select the database algorithm used for database search (e.g., TopPIC Search Results). Select the folder labeled "MASH_Workflows" and select the appropriate MassList folder labeled with the database search algorithm (e.g., . . .MassList_TopPIC_1_1) (Fig. 5). The folder contains a .CSV file with the Protein Search results. After opening the .CSV file, the protein sequence will be uploaded to MASH Explorer's Sequence Table.
6. Using MASH Explorer software, visualize the verified fragment ion list against the imported proteoform sequence in the Sequence Table for proteoform characterization (Fig. 6). A

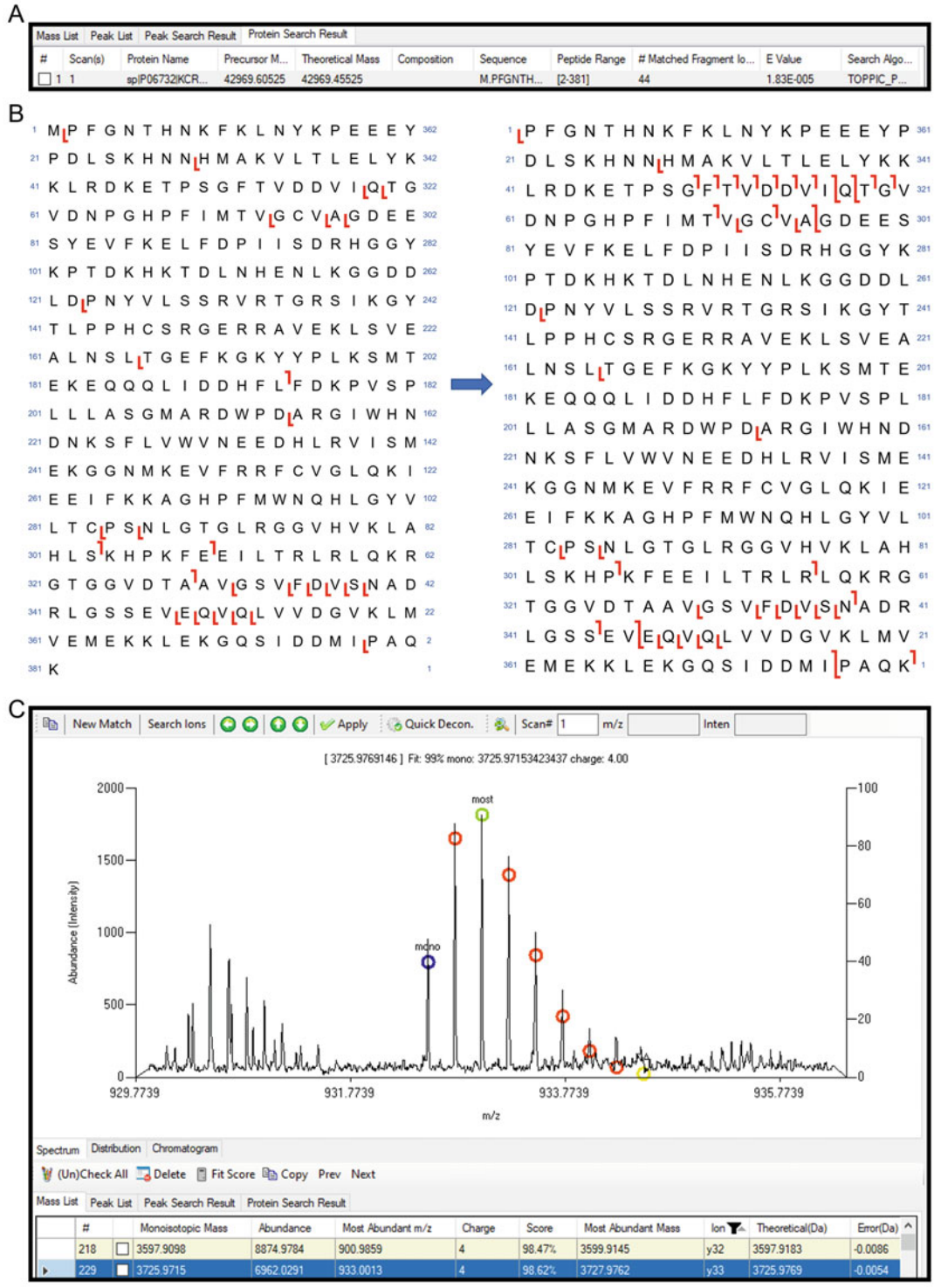


Fig. 6 Visualizing and verifying fragment ions against an imported protein sequence in MASH Explorer. (**a**) Imported database search results can be found under the Protein Search Result tab. (**b**) MASH Explorer's "Protein Sequence Table" allows users to visualize fragment ion mapping for different MS/MS techniques to characterize the protein sequence and PTMs. The protein amino acid sequence corresponding to the identified

warning message will appear from MASH Explorer reminding the user to correct the protein sequence to obtain the correct fragmentation list.

7. Characterize the protein's PTMs. Users can obtain this information by intact mass analysis discussed in Subheading 3.2 or by database search algorithm results (*see* **Note 20**).
8. Use database search algorithms in MASH Explorer to identify proteoforms (*see* **Note 21**).

4 Notes

1. Ensure that all mobile phases are prepared in thoroughly cleaned glassware intended for HPLC solvents. Always clean glassware with the same mobile phase/solvent used for the chromatography; never wash glassware with detergent as this can introduce contaminants during LC-MS analysis.
2. A variety of column dimensions can be used for SEC and sSEC experiments. Typical column dimensions include 200 mm × 9.4 mm, 3 μm particles with pore sizes of 500 Å, and 1000 Å. Ideally for SEC, the use of small particles, wider ID columns, and shorter column lengths can result in columns where wall effects are negligible and protein band spreading decreases [28]. Strategies such as minimizing sample injection volume and increasing initial protein loading can assist in reducing significant dilution and loss of UV signal during SEC.
3. A variety of MS-compatible mobile phase compositions can be used for the fractionation of different protein classes. Previously, we have demonstrated that 1% formic acid is effective for acid-soluble and water-soluble proteins [9, 10]. For membrane protein fractionation, our lab has shown that addition of 40% isopropanol (1% formic acid, 40% isopropanol, 59% water v/v) can facilitate membrane protein solubility in the absence of surfactants [29]. For proteins that strongly interact with each other, even under denaturing conditions, the addition of hexafluoroisopropanol can be helpful in disrupting protein–protein interactions and can be added to a final concentration of 100 mM. Samples with hexafluoroisopropanol can be centrifuged in MWCO filters at 15,000 × *g* for 30 min at 4 °C prior to SEC.

Fig. 6 (continued) protein is directly imported from the database search. Correcting the protein sequence according to the database search (N-terminal methionine excision) results in significantly improved ion fragmentation lists. (**c**) MASH Explorer enables manual verification of MS/MS fragment ion data and adjustment of charge state and monoisotopic mass

4. The maximum backpressure of this PLRP-S material is approximately 3000 psi; exceeding this pressure will reduce the performance and lifetime of the column.
5. MASH Explorer requires Microsoft Windows operating system with .NET framework installation. It is recommended that the personal computer should have at least central processing units with two cores and two threads, and four gigabytes of random-access memory for operation. For deconvolution and database search algorithms, it is suggested that a personal computer with a minimal computing power of a central processing unit with four cores and four threads, and eight gigabytes of random-access memory is used.
6. The MASH Explorer installation package does not contain any deconvolution or database search algorithms except for the THRASH deconvolution algorithm. To access other supported deconvolution and database search algorithms, they must be downloaded from the individual research groups which developed them. MASH Explorer allows for convenient import of these deconvolution and database search algorithms by accessing the Tools menu and selecting the Configuration dialog.
7. For single column SEC (200 × 4.6 mm, 5 μm, 500 Å PolyHEA), the protein-loading capacity is approximately 500 μg. A sample loop volume of 50–250 μL can be used for sample injection. Using smaller sample loop volumes when possible helps to reduce protein band broadening. For single column SEC, we typically use a 50 μL sample loop and 100 μg protein loading at 2 μg/μL sample concentration; for two SEC columns connected serially, we use a 250 μL sample loop and 200 μg protein loading; for three SEC columns connected serially, we use a 250 μL sample loop and 300 μg protein loading. Serial SEC can improve protein separation and fractionation range but requires a larger sample injection and greater protein loading to avoid diluting protein signal.
8. Using a single SEC column (200 × 9.6 mm, 3 μm, 500 Å PolyHEA), the backpressure is approximately 1000 psi using a flow rate of 0.5 mL/min with 1% formic acid in water (v/v) as mobile phase. For two SEC columns connected serially, the pressure is approximately 1500 psi; for three SEC columns connected serially, the pressure is approximately 2000 psi. Use a KimWipe™ to detect solvent leaks at the junction of the fittings and the columns.
9. A panel of standard proteins are often used in our lab to evaluate the performance of SEC columns. Depending on native or denaturing SEC applications, these standard proteins include thyroglobulin (669 kDa), apoferritin (443 kDa), β-amylase (200 kDa), alcohol dehydrogenase (150 kDa),

bovine serum albumin (66 kDa), and myoglobin (17 kDa). For denaturing SEC, β-amylase, bovine serum albumin, and myoglobin are typical standards used to evaluate the performance of SEC columns.

10. Proteins usually show UV absorption maxima between 275 and 280 nm, which are caused by absorbance of the two aromatic amino acids tryptophan (Trp) and tyrosine (Tyr) and, to a small extent, by the absorbance of cystine (i.e., disulfide bonds) [30].
11. We normally collect SEC fractions using 1-min intervals (~0.5 mL elution fractions), but longer time intervals (e.g., 1.5- or 2-min intervals) can be collected and pooled together.
12. MWCO filters are pre-rinsed by centrifuging with the mobile phase used in the SEC separation to reduce protein loss resulting from concentration of the SEC fractions. A total of 2 mM TCEP is added to reduce protein oxidation and is active under acidic conditions such as 1% formic acid in water (v/v).
13. For typical SDS-PAGE experiments, commercially available products such as 8–16% Mini-PROTEAN® 12.5% Tris–Glycine eXtended (TGX) gels or Novex™ 8–16% Tris–Glycine Plus Midi Gels can be used according to the manufacturer's instructions (e.g., 125 V until electrophoresis is complete). For protein loading between 100 and 1000 ng per lane, it is recommended to use a highly sensitive staining method such as SYPRO Ruby or Silver Staining. For protein loading greater than 1 μg per lane, we normally perform Coomassie brilliant blue staining.
14. To quantify and normalize total protein loading for equal protein loading in SDS-PAGE and LC-MS experiments, we typically use the Bradford microplate assay for protein samples that are prepared without detergents and measure the absorbance of standards and samples at 595 nm using an absorbance microplate reader. For samples which contain detergents, it is recommended to use the bicinchoninic acid (BCA) microplate assay, which is compatible with most ionic and non-ionic detergents, and measure the absorbance of standards and samples at 562 nm using an absorbance microplate reader.
15. While LC mobile-phase gradients will require optimization and vary depending on the protein mixture, a typical LC gradient used for the separation of cardiac proteins commonly performed in our lab is 0–5 min 20% B, 5–17 min 20–40% B, 17–25 min 40–50% B, 25–32 min 50–65%, 32–43 min 65–95% B, 43–48 min 95% B, 48–52 min 95–5% B, and 52–60 min 5% B.
16. To select the appropriate fragmentation energy for targeted MS/MS experiments using collisionally activated dissociation (CAD), the precursor ion in the resulting MS/MS spectrum

should still be present at 50–70% base peak intensity to avoid over-fragmentation, which may result in internal fragments and lower identified fragments. For larger MW proteins, they may require lower fragmentation energy than that required for smaller MW proteins. A good starting range for CAD energy is 10–20 eV for large MW proteins.

17. To support data import, MASH Explorer uses both ProteoWizard [31] and vendor-specific software (Thermo. RAW, Bruker. BAF, Bruker .ascii, Waters. RAW,. MzXML,. MGF,. MzML).
18. MASH Explorer supports several deconvolution algorithms including THRASH, TopFD [24], pParseTD [26], and MS-Deconv [32]. The results from THRASH, MS-Deconv, and TopFD are most often used for fragment ion verification.
19. MASH Explorer supports multiple database search algorithms including TopPIC [24], MS-Align+ [25], pTop [26], and Informed-Proteomics [27] workflows.
20. PTMs are supported directly in MASH Explorer software. Modifications such as acetylation, trimethylation, and phosphorylation can be directly added on the target amino acid in the protein sequence. Other modifications can be added using custom modification. Database search algorithms are useful tools in identifying PTMs such as N-terminal acetylation and methylation. However, for phosphorylation, manual proteoform characterization is often needed using the verified fragment ion list [33–40].
21. Identification of larger MW proteoforms may not be confidently and accurately identified by database search algorithms. Using the verified fragment ion list, a series of fragment ions in a protein sequence may be used to derive a three or four consecutive amino acid sequence, called sequence tag. This sequence tag can be used as an alternative method to identify target proteoform matching the intact protein mass. The amino acid sequence for the sequence tag can be derived by mass differences of several fragment ions. The mass differences can be matched to one or the sum of multiple amino acid residue masses (https://proteomicsresource.washington.edu/mascot/help/aa_help.html). The obtained sequence tag can be searched against the database using the guide provided by UniProt (https://web.expasy.org/tagident/).

Acknowledgments

We would like to thank Trisha Tucholski for helpful discussions. This work was supported by NIH R01 GM117058 and NIH R01

GM125085 (to YG). YG also would like to acknowledge R01 HL096971 and S10 OD018475. TNT would like to acknowledge support from the NIH Chemistry-Biology Interface Training Program, T32 GM008505. JAM would like to acknowledge support from the Training Program in Translational Cardiovascular Science, T32 HL007936-20.

References

1. Smith LM, Kelleher NL (2018) Proteoforms as the next proteomics currency. Science 359(6380):1106
2. Smith LM, Thomas PM, Shortreed MR, Schaffer LV, Fellers RT, LeDuc RD, Tucholski T, Ge Y, Agar JN, Anderson LC, Chamot-Rooke J, Gault J, Loo JA, Pasa-Tolic L, Robinson CV, Schluter H, Tsybin YO, Vilaseca M, Vizcaino JA, Danis PO, Kelleher NL (2019) A five-level classification system for proteoform identifications. Nat Methods 16:939
3. Aebersold R, Agar JN, Amster IJ, Baker MS, Bertozzi CR, Boja ES, Costello CE, Cravatt BF, Fenselau C, Garcia BA, Ge Y, Gunawardena J, Hendrickson RC, Hergenrother PJ, Huber CG, Ivanov AR, Jensen ON, Jewett MC, Kelleher NL, Kiessling LL, Krogan NJ, Larsen MR, Loo JA, Loo RRO, Lundberg E, MacCoss MJ, Mallick P, Mootha VK, Mrksich M, Muir TW, Patrie SM, Pesavento JJ, Pitteri SJ, Rodriguez H, Saghatelian A, Sandoval W, Schluter H, Sechi S, Slavoff SA, Smith LM, Snyder MP, Thomas PM, Uhlen M, Van Eyk JE, Vidal M, Walt DR, White FM, Williams ER, Wohlschlager T, Wysocki VH, Yates NA, Young NL, Zhang B (2018) How many human proteoforms are there? Nat Chem Biol 14(3):206
4. Brown KA, Chen BF, Guardado-Alvarez TM, Lin ZQ, Hwang L, Ayaz-Guner S, Jin S, Ge Y (2019) A photocleavable surfactant for top-down proteomics. Nat Methods 16(5):417
5. Tiambeng TN, Roberts DS, Brown KA, Zhu Y, Chen B, Wu Z, Mitchell SD, Guardado-Alvarez TM, Jin S, Ge Y (2020) Nanoproteomics enables proteoform-resolved analysis of low-abundance proteins in human serum. Nat Commun 11(1):3903. https://doi.org/10.1038/s41467-020-17643-1
6. Chen B, Brown KA, Lin Z, Ge Y (2018) Top-down proteomics: ready for prime time? Anal Chem 90(1):110–127. https://doi.org/10.1021/acs.analchem.7b04747
7. Toby TK, Fornelli L, Kelleher NL (2016) Progress in top-down proteomics and the analysis of proteoforms. Annu Rev Anal Chem 9(1):499–519. https://doi.org/10.1146/annurev-anchem-071015-041550
8. Brown KA, Melby JA, Roberts DS, Ge Y (2020) Top-down proteomics: challenges, innovations, and applications in basic and clinical research. Expert Rev Proteomics 17(10):719–733. https://doi.org/10.1080/14789450.2020.1855982
9. Cai W, Tucholski T, Chen B, Alpert AJ, McIlwain S, Kohmoto T, Jin S, Ge Y (2017) Top-down proteomics of large proteins up to 223 kDa enabled by serial size exclusion chromatography strategy. Anal Chem. https://doi.org/10.1021/acs.analchem.7b00380
10. Tucholski T, Knott SJ, Chen B, Pistono P, Lin Z, Ge Y (2019) A top-down proteomics platform coupling serial size exclusion chromatography and Fourier transform ion cyclotron resonance mass spectrometry. Anal Chem 91(6):3835–3844. https://doi.org/10.1021/acs.analchem.8b04082
11. Schaffer LV, Tucholski T, Shortreed MR, Ge Y, Smith LM (2019) Intact-mass analysis facilitating the identification of large human heart proteoforms. Anal Chem 91(17):10937–10942. https://doi.org/10.1021/acs.analchem.9b02343
12. Compton PD, Zamdborg L, Thomas PM, Kelleher NL (2011) On the scalability and requirements of whole protein mass spectrometry. Anal Chem 83(17):6868–6874. https://doi.org/10.1021/ac2010795
13. Doucette AA, Tran JC, Wall MJ, Fitzsimmons S (2011) Intact proteome fractionation strategies compatible with mass spectrometry. Expert Rev Proteomics 8(6):787–800. https://doi.org/10.1586/epr.11.67
14. Chen X, Ge Y (2013) Ultrahigh pressure fast size exclusion chromatography for top-down proteomics. Proteomics 13(17):2563–2566. https://doi.org/10.1002/pmic.201200594
15. Alpert AJ (2016) Protein fractionation and enrichment prior to proteomics sample preparation. In: Mirzaei H, Carrasco M (eds) Modern proteomics - sample preparation,

analysis and practical applications, vol 919. Advances in Experimental Medicine and Biology, pp 23–41. https://doi.org/10.1007/978-3-319-41448-5_2

16. Hong P, Koza S, Bouvier ESP (2012) Size-exclusion chromatography for the analysis of protein biotherapeutics and their aggregates. J Liq Chromatogr Relat Technol 35(20): 2923–2950. https://doi.org/10.1080/10826076.2012.743724
17. Hengel SM, Sanderson R, Valliere-Douglass J, Nicholas N, Leiske C, Alley SC (2014) Measurement of in vivo drug load distribution of cysteine-linked antibody–drug conjugates using microscale liquid chromatography mass spectrometry. Anal Chem 86(7):3420–3425. https://doi.org/10.1021/ac403860c
18. Wu Z, Roberts DS, Melby JA, Wenger K, Wetzel M, Gu Y, Ramanathan SG, Bayne EF, Liu X, Sun R, Ong IM, McIlwain SJ, Ge Y (2020) MASH explorer: a universal software environment for top-down proteomics. J Proteome Res 19(9):3867–3876. https://doi.org/10.1021/acs.jproteome.0c00469
19. Guner H, Close PL, Cai W, Zhang H, Peng Y, Gregorich ZR, Ge Y (2014) MASH Suite: a user-friendly and versatile software interface for high-resolution mass spectrometry data interpretation and visualization. J Am Soc Mass Spectrom 25(3):464
20. Cai W, Guner H, Gregorich ZR, Chen AJ, Ayaz-Guner S, Peng Y, Valeja SG, Liu X, Ge Y (2016) MASH Suite Pro: a comprehensive software tool for top-down proteomics. Mol Cell Proteomics 15(2):703–714. https://doi.org/10.1074/mcp.o115.054387
21. Cesnik AJ, Shortreed MR, Schaffer LV, Knoener RA, Frey BL, Scalf M, Solntsev SK, Dai Y, Gasch AP, Smith LM (2018) Proteoform Suite: software for constructing, quantifying, and visualizing proteoform families. J Proteome Res 17(1):568–578. https://doi.org/10.1021/acs.jproteome.7b00685
22. Schaffer LV, Shortreed MR, Cesnik AJ, Frey BL, Solntsev SK, Scalf M, Smith LM (2018) Expanding proteoform identifications in top-down proteomic analyses by constructing proteoform families. Anal Chem 90(2): 1325–1333. https://doi.org/10.1021/acs.analchem.7b04221
23. Horn DM, Zubarev RA, McLafferty FW (2000) Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. J Am Soc Mass Spectrom 11(4):320
24. Kou Q, Xun L, Liu X (2016) TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. Bioinformatics 32(22):3495
25. Liu X, Sirotkin Y, Shen Y, Anderson G, Tsai YS, Ting YS, Goodlett DR, Smith RD, Bafna V, Pevzner PA (2012) Protein identification using top-down spectra. Mol Cell Proteomics 11(6):M111.008524
26. Sun RX, Luo L, Wu L, Wang RM, Zeng WF, Chi H, Liu C, He SM (2016) pTop 1.0: a high-accuracy and high-efficiency search engine for intact protein identification. Anal Chem 88(6): 3082
27. 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 (2017) Informed-proteomics: open-source software package for top-down proteomics. Nat Methods 14(9):909
28. Regnier FE, Gooding KM (1980) High-performance liquid chromatography of proteins. Anal Biochem 103(1):1–25. https://doi.org/10.1016/0003-2697(80)90229-8
29. Brown KA, Tucholski T, Alpert AJ, Eken C, Wesemann L, Kyrvasilis A, Jin S, Ge Y (2020) Top-down proteomics of endogenous membrane proteins enabled by cloud point enrichment and multidimensional liquid chromatography–mass spectrometry. Anal Chem 92(24):15726–15735. https://doi.org/10.1021/acs.analchem.0c02533
30. Schmid FX (2001) Biological macromolecules: UV-visible spectrophotometry. e LS
31. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, Kessner D, Tasman N, Shulman N, Frewen B, Baker TA, Brusniak M-Y, Paulse C, Creasy D, Flashner L, Kani K, Moulding C, Seymour SL, Nuwaysir LM, Lefebvre B, Kuhlmann F, Roark J, Rainer P, Detlev S, Hemenway T, Huhmer A, Langridge J, Connolly B, Chadick T, Holly K, Eckels J, Deutsch EW, Moritz RL, Katz JE, Agus DB, MacCoss M, Tabb DL, Mallick P (2012) A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol 30(10): 918–920. https://doi.org/10.1038/nbt.2377
32. Liu X, Inbar Y, Dorrestein PC, Wynne C, Edwards N, Souda P, Whitelegge JP, Bafna V, Pevzner PA (2010) Deconvolution and database search of complex tandem mass spectra of intact proteins: a combinatorial approach. Mol Cell Proteomics 9(12):2772
33. Jin Y, Wei L, Cai W, Lin Z, Wu Z, Peng Y, Kohmoto T, Moss RL, Ge Y (2017) Complete

characterization of cardiac myosin heavy chain (223 kDa) enabled by size-exclusion chromatography and middle-down mass spectrometry. Anal Chem 89(9):4922–4930. https://doi.org/10.1021/acs.analchem.7b00113

34. Wu Z, Jin Y, Chen B, Gugger MK, Wilkinson-Johnson CL, Tiambeng TN, Jin S, Ge Y (2019) Comprehensive characterization of the recombinant catalytic subunit of cAMP-dependent protein kinase by top-down mass spectrometry. J Am Soc Mass Spectrom 30(12):2561–2570. https://doi.org/10.1021/jasms.8b06294
35. Jin Y, Lin Z, Xu Q, Fu C, Zhang Z, Zhang Q, Pritts WA, Ge Y (2019) Comprehensive characterization of monoclonal antibody by Fourier transform ion cyclotron resonance mass spectrometry. MAbs 11(1):106–115. https://doi.org/10.1080/19420862.2018.1525253
36. Lin Z, Guo F, Gregorich ZR, Sun R, Zhang H, Hu Y, Shanmuganayagam D, Ge Y (2018) Comprehensive characterization of swine cardiac troponin T proteoforms by top-down mass spectrometry. J Am Soc Mass Spectrom 29(6):1284–1294. https://doi.org/10.1021/jasms.8b05834
37. Tucholski T, Cai W, Gregorich ZR, Bayne EF, Mitchell SD, McIlwain SJ, de Lange WJ, Wrobbel M, Karp H, Hite Z, Vikhorev PG, Marston SB, Lal S, Li A, dos Remedios C, Kohmoto T, Hermsen J, Ralphe JC, Kamp TJ, Moss RL, Ge Y (2020) Distinct hypertrophic cardiomyopathy genotypes result in convergent sarcomeric proteoform profiles revealed by top-down proteomics. Proc Natl Acad Sci 117(40):24691. https://doi.org/10.1073/pnas.2006764117
38. Chen B, Guo X, Tucholski T, Lin Z, McIlwain S, Ge Y (2017) The impact of phosphorylation on electron capture dissociation of proteins: a top-down perspective. J Am Soc Mass Spectrom 28(9):1805–1814. https://doi.org/10.1007/s13361-017-1710-3
39. Jin Y, Diffee GM, Colman RJ, Anderson RM, Ge Y (2019) Top-down mass spectrometry of sarcomeric protein post-translational modifications from non-human primate skeletal muscle. J Am Soc Mass Spectrom 30(12):2460
40. Peng Y, Gregorich ZR, Valeja SG, Zhang H, Cai W, Chen YC, Guner H, Chen AJ, Schwahn DJ, Hacker TA, Liu X, Ge Y (2014) Top-down proteomics reveals concerted reductions in myofilament and Z-disc protein phosphorylation after acute myocardial infarction. Mol Cell Proteomics 13(10):2752–2764. https://doi.org/10.1074/mcp.M114.040675

Chapter 4

RPLC-RPLC-MS/MS for Proteoform Identification

Kellye A. Cupp-Sutton, Zhe Wang, Dahang Yu, and Si Wu

Abstract

Top-down proteomics methods have a distinct advantage over bottom-up methods in that they analyze intact proteins rather than digested peptides which can result in loss of information regarding the intact protein. However, the analysis of intact proteins using top-down proteomics methods has been impeded by the low resolution of typical separation approaches applied in bottom-up proteomics studies. To increase the coverage of intact proteomes, orthogonal, two-dimensional separation techniques have been developed to improve the separation efficiency; in this chapter, we describe a two-dimensional HPLC separation technique that utilizes a high-pH mobile phase in the first dimension followed by a low-pH mobile phase in the second dimension. This two-dimensional pH-based HPLC approach demonstrates increased separation efficiency of intact proteins and increased proteome coverage when compared to one-dimensional HPLC in the analysis of larger and lower abundance proteoforms.

Key words Top-down MS, Proteomics, 2D separation, RPLC, Intact proteoforms

1 Introduction

Top-down mass spectrometry (MS)-based proteomics has become an increasingly popular technique as the technologies regarding instrumentation and data acquisition and processing have improved [1, 2]. These methods have been successfully used to study the dynamics of protein posttranslational modifications (PTMs) in biological pathways and disease mechanisms. High separation efficiency of intact proteoforms is required for high-throughput top-down proteomics analysis of complex mixtures; currently, reversed-phase liquid chromatography (RPLC) is the most common separation approach for top-down proteomics [2–6]. Efforts to improve RPLC separation of complex samples of intact proteoforms including increasing the column pressure and length [7] and decreasing the particle size [8] have been successful, but improvements are still needed to optimize coverage of cellular proteomes. Orthogonal, two-dimensional separation techniques have been developed to improve the separation efficiency and increase the number of

Liangliang Sun and Xiaowen Liu (eds.), *Proteoform Identification: Methods and Protocols*,
Methods in Molecular Biology, vol. 2500, https://doi.org/10.1007/978-1-0716-2325-1_4,
© The Author(s), under exclusive license to Springer Science+Business Media, LLC, part of Springer Nature 2022

identified intact proteins. Additionally, separation techniques including size exclusion chromatography (SEC) [9], gel-eluted liquid fraction entrapment electrophoresis (GELFrEE) [10], hydrophilic and hydrophobic interaction chromatography (HILIC/HIC, respectively) [11], and capillary zone electrophoresis (CZE) [12] have all been paired with RPLC as orthogonal separation methods prior to MS analysis of complex intact protein samples.

Here, we describe a 2D pH RP/RPLC-MS/MS method that utilizes a high-pH RPLC (pH = 10) as the first-dimension separation followed by a low-pH (pH = 2) RPLC in the second-dimension separation [13–15]. The high orthogonality between high-pH RPLC and low-pH RPLC relies on the change in hydrophobicity at the pH extremes, and coupling these RPLC methods has improved the identification of intact proteoforms. The orthogonality of the high-pH and low-pH RPLC separations can be visualized in a heatmap (Fig. 1) [14]. The separation methods in both dimensions are "salt-free" and utilize only volatile buffers compatible with MS analysis. Limited sample handling that does not require desalting or buffer exchange increases the reproducibility and limits the sample loss between separations. Furthermore, we have developed an online ultra-high-pressure nano-LC system to improve the throughput and sensitivity of the offline 2D pH RP/RPLC-MS/MS approach.

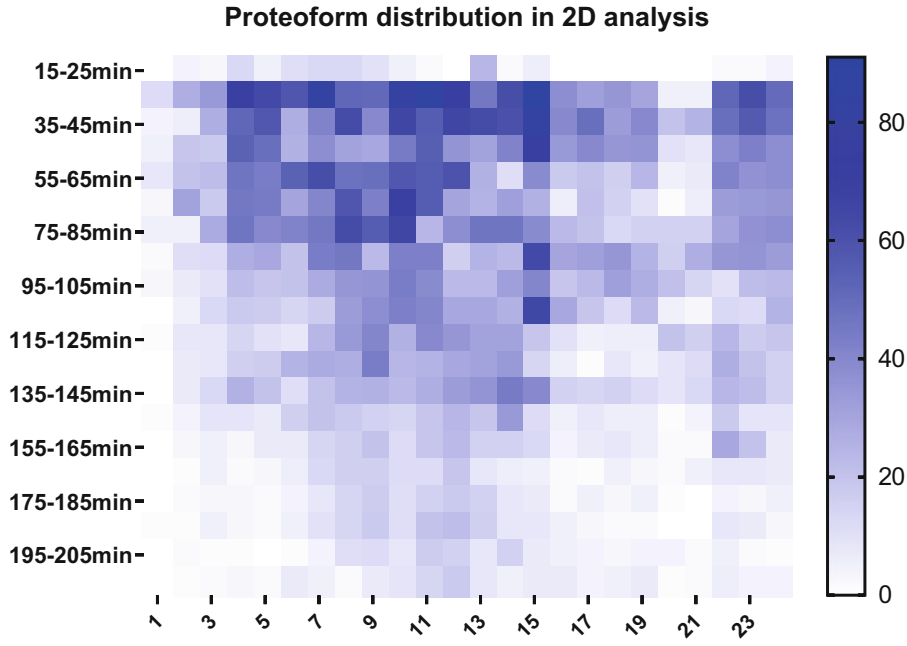


Fig. 1 Identification heatmap demonstrating the orthogonality of the low-pH/high-pH RPLC separations in the first and second dimensions, respectively. Each low-pH bin in the heatmap represents the number of deconvoluted mass features (>2.5 kDa) in a 10-min portion of the elution window

2 Materials

2.1 High-pH Mobile Phases (See Notes 1 and 2)

1. Ammonium formate high concentration stock solution: 200 mM ammonium formate, pH 10.0. Add 6.3 g ammonium formate to a 500 mL media storage bottle with 450 mL LC-MS grade water. Adjust the pH of the mixture to 10.0 by adding ammonium hydroxide dropwise while stirring. After pH adjustment, dilute the solution to 500 mL using LC-MS grade water.
2. High-pH mobile phase A (high-pH MPA): 20 mM ammonium formate, pH 10.0. Add 50 mL of the ammonium formate high concentration stock solution prepared in **step 1** to a 500 mL media storage bottle and dilute to 500 mL using LC/MS grade water. Measure the pH and adjust to pH 10.0 if needed.
3. High-pH mobile phase B (high-pH MPB): 20 mM ammonium formate in 90% acetonitrile (ACN), pH 10.0. Add 50 mL ammonium formate high concentration stock solution prepared in **step 1** to a 500 mL media storage bottle and dilute to 500 mL using LC/MS grade ACN.

2.2 Low-pH Mobile Phases (See Notes 1 and 2)

1. Low-pH mobile phase A (low-pH MPA): 0.01% trifluoracetic acid, 0.585% acetic acid, 2.5% isopropanol (2-propanol), and 5% ACN in water. Add 50 mL ACN, 25 mL 2-propanol, and 920 mL LC/MS grade water to a 1000 mL media storage bottle and mix well. Then, add 5.85 mL acetic acid and 0.1 mL trifluoroacetic acid using a glass pipette (*see* **Note 3**). Store at room temperature.
2. Low-pH mobile phase B (low-pH MPB): 0.01% trifluoracetic acid, 0.585% acetic acid, 45% 2-propanol, and 45% ACN in water. Add 450 mL ACN, 450 mL 2-propanol, and 95 mL LC/MS grade water to a 1000 mL media storage bottle and mix well. Then, add 5.85 mL acetic acid and 0.1 mL trifluoroacetic acid to the solution using a glass pipette. Store at room temperature.

2.3 HeLa Cell Culture and Lysis Reagents (See Note 4)

1. *HeLa* medium: 1% penicillin–streptomycin (pen-strep), 9% fetal bovine serum (FBS), 90% Dulbecco's Modified Eagle's medium–high glucose (DMEM). Add 5 mL pen-strep solution and 50 mL FBS to a new 500 mL bottle of high glucose DMEM. This solution should be stored at 4 °C.
2. *HeLa* lysis buffer: 20 mM sodium fluoride (NaF), 1× Halt™ protease, and phosphatase inhibitor cocktail in 1× phosphate-buffered saline (PBS). Add 4 mL ice cold 10× concentrated PBS, 4 mL 200 mM sodium fluoride (NaF), and 400 μL 100× Halt™ protease and phosphatase inhibitor cocktail to a 50 mL conical centrifuge tube. Dilute this solution to 40 mL with LC/MS grade water.

2.4 E. coli *Cell Culture and Lysis Reagents* (See *Note 5*)

1. Lysogeny broth (LB) for the initial *E. coli* inoculation: 2% w/v lysogeny broth. Dissolve 2 g of lysogeny broth powder in 100 mL of Milli-Q water in a 250 mL Erlenmeyer flask.
2. Lysogeny broth for the second *E. coli* inoculation: 2% w/v lysogeny broth. Dissolve 20 g of lysogeny broth powder into 1 L of Milli-Q water in a 2 L Erlenmeyer flask (two flasks required).
3. Ammonium bicarbonate buffer stock solution: 1 M ammonium bicarbonate, pH 7.4. Dissolve 79 g ammonium bicarbonate powder into 1 L LC/MS grade water in a 1000 mL media bottle. Adjust the pH to 7.4 by dropwise addition of ammonium hydroxide while stirring. Monitor the pH of the solution using a pH meter.
4. Ammonium bicarbonate (ABC) buffer: 25 mM ammonium bicarbonate, pH 7.4. Dilute 100 mL of the ammonium bicarbonate buffer stock solution **(item 3)** to 4 L with LC/MS grade water in a 4 L bottle. Test the pH using a pH meter to ensure a pH of 7.4, adjust the pH using ammonium hydroxide if necessary.

3 Methods

3.1 *HeLa Cell Culture and Lysis* (See *Note 4*)

1. To recover frozen *HeLa* cells from storage at −80 °C, warm the cells to 37 °C in a water bath. Add 8 mL of the *HeLa* medium (*see* Subheading 2.3, **step 1**) to a cell culture plate. Pipette the recovered cells to the cell culture plate and swirl gently. Incubate the plate for 24 h at 37 °C in a humidified atmosphere containing 5% carbon dioxide (CO_2). After 24 h, remove the *HeLa* medium from the plate and replace with 10 mL fresh *HeLa* medium.
2. Passage the cells when the cell density is approximately 80–90%. To passage the cells, remove the medium from the plates. Wash the cells by adding 5 mL of pH 7.4 phosphate-buffered saline (PBS) to each plate and swirl gently. Remove the PBS using a pipette and add 2.5 mL TrypLE Express to each plate to dissociate the adherent cells from the plate surface. Incubate the plates at 37 °C for 5 min.
3. After incubation, add 6.5 mL of the *HeLa* medium solution to each plate to quench the dissociation. Prepare new plates for cell passaging by adding 3 mL of cells and 7 mL of the *HeLa* medium solution to each new plate and swirl gently. Passage each plate of cells to three plates and record the passage number. Incubate the plates and exchange the *HeLa* medium solution with 10 mL fresh *HeLa* medium solution every 48 h until the cell density is approximately 80–90%. At this point, either passage the cells again by repeating **steps 2** and **3** or harvest the cells as discussed in **step 4**.

4. To harvest the *HeLa* cells, remove the *HeLa* medium and wash the cells with 5 mL of ice-cold PBS buffer at pH 7.4. Swirl the plate gently and remove the buffer (repeat washing once more for a total of two washes). Then, add 2 mL of ice-cold PBS to each plate and scratch the cells off the plates. Transfer the buffer containing the cells to a 50 mL conical centrifuge tube (approximately 20 plates per tube).
5. Centrifuge the tubes at 4000 rpm and 4 °C for 20 min. Wash the cell pellet by adding 20 mL of ice-cold PBS and mix gently. Centrifuge the cells again at 4000 rpm and 4 °C for 20 min and repeat the wash step for a total of 2 washes. Then, centrifuge the cells at 2000 rpm for 5 min and remove the supernatant. The cell pellets may be frozen at −80 °C until lysis.
6. To lyse the *HeLa* cell pellet produced in **step 5**, thaw the pellet on ice for 1 h and resuspend in 10 mL of the *HeLa* lysis buffer (*see* Subheading 2.3, **step 2**) (*see* **Note 6**). Aliquot the solution into six 2 mL Eppendorf tubes, 1.5 mL per tube. Then, wash the 50 mL conical centrifuge tubes with an additional 2 mL of lysis buffer and aliquot the remainder of the solution into two additional 2 mL Eppendorf tubes, 1.5 mL per tube. Sonicate the tubes on ice for 5 minutes and allow the tubes to rest for 5 minutes (repeat sonication 3 times for a total of 4 rounds of sonication).
7. Centrifuge the tubes at 13,000 rpm and 4 °C for 30 min. Remove 1 mL of the supernatant from each tube and combine into a 15 mL conical centrifuge tube (*see* **Note 7**). Filter the lysate using a 0.2 um syringe filter (*see* **Note 8**). Aliquot the filtered lysate into 12 Eppendorf tubes (1 mL per tube) and centrifuge at 13,000 rpm and 4 °C for 30 min. Finally, move 900 μL of the supernatant to fresh Eppendorf tubes without disturbing the bottom 100 μL of sample or any pellet that may have formed. The lysate may be frozen at −80 °C until use.

3.2 E. coli *Cell Culture and Lysis*

1. For the initial *E. coli* inoculation, add 5 μL of *E. coli* origin media to the pre-prepared lysogeny broth (*see* Subheading 3.4, **step 1**) and shake the solution at 250 rpm and 37 °C for 8 h.
2. Add 50 mL of the initial inoculation solution described in **step 1** to each pre-prepared LB flask for the second inoculation (*see* Subheading 3.4, **step 2**). Shake the flasks at 250 rpm and 37 °C for 12 h.
3. Centrifuge the solutions from the second inoculation (**step 2**) at 7800 rpm for 15 min at 4 °C to pellet the *E. coli* cells. Remove the supernatant and resuspend the pellets in 25 mM ABC buffer (*see* Subheading 2.4, **step 4**) at a ratio of 1 g of cell pellet to 5 mL of ABC. Add PMSF to a final concentration of 0.1% (v/v) to inhibit proteases.

4. Lyse the *E. coli* cells using a high-pressure homogenizer (Avestin C3 EmulsiFlex homogenizer) according to the manufacturer's instructions using 25 mM ABC buffer (*see* Subheading 2.4, **step 4**). Aliquot the solutions into 1.5 mL Eppendorf tubes.
5. Centrifuge the tubes at 13,000 rpm for 60 min at 4 °C to remove the cell debris. Move the supernatant to fresh 1.5 mL Eppendorf tubes and freeze at −80 °C until needed.

3.3 Offline First-Dimension High-pH RPLC Separation and Fractionation [13, 14]

1. The offline first-dimension high-pH RPLC separation and fractionation utilizes a Waters XBridge Protein BEH C4 column (300 Å, 3.5 μm, 2.1 mm × 250 mm), a Thermo Scientific Accela LC system with UV detection, and a Triversa NanoMate.
2. For first-dimension protein fractionation, inject 1 mg of the desired protein lysate (*HeLa* or *E. coli*) onto the first-dimension column. Utilizing a flow rate of 150 μL/min, load the sample for 5 min using 97% high-pH MPA. Elute the proteins using a 60-min gradient from 10% high-pH MPB to 70% high-pH MPB. Collect 24, 2.5-min fractions into a 96-well plate using a Triversa NanoMate.
3. Begin fraction collection when the first proteins are eluted as observed using UV-Vis detection, Fig. 2a (*see* **Note 9**). SpeedVac the fractions to 20–30 μL at low temperature (*see* **Note 10**).
4. After fraction collection, flush the column with 90% MPB for 10 min and then equilibrate back to 97% high-pH MPA for 30 min at a flow rate of 75 μL/min.

3.4 Offline Second-Dimension Low-pH RPLC Separation and Top-Down MS/MS Analysis

1. Connect a nano-LC pump to a trapping column (150 μm ID, 10 cm length, 5 μm diameter, and 300 Å pore size Jupiter particles) and a nano-flow C5 capillary separation column (75 μm ID, 70 cm length, 5 μm diameter, and 300 Å pore size Jupiter particles). Connect the separation column directly to an etched capillary ESI tip for sample introduction to the mass spectrometer [16].
2. Equilibrate the trapping column using 97% low-pH MPA and 3% low-pH MPB at a flow rate of 5 μL/min for 30 min and equilibrate the separation column using the same buffer at a flow rate of 0.4 μL/min for 30 min.
3. Reconstitute the concentrated fractions (*see* Subheading 3.3, **step 3**) to 100 μL using 25 mM ABC buffer (*see* Subheading 2.4, **step 4**). Load 25 μL of each reconstituted fraction individually onto the trapping column using 97% low-pH MPA at a flow rate of 5 μL/min over 15 min. Elute proteins from the trapping column and separate using a 200-min gradient from

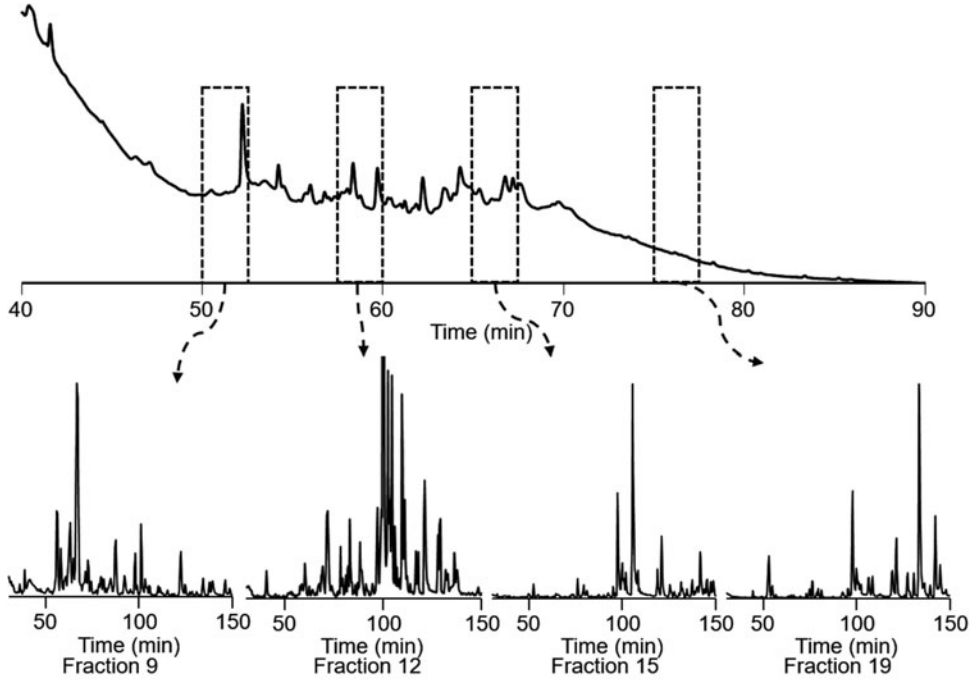


Fig. 2 First-dimension high-pH UV-Vis spectrum and representative LC-MS of fractions. (Reprinted from ref. 13, Copyright (2018), with permission from Elsevier)

10 to 65% low-pH MPB at a flow rate of 0.4 μL/min. Introduce the sample to the LTQ Orbitrap Velos Pro Mass Spectrometer using a nano-ESI interface for data collection, Fig. 2b. Set the temperature of the inlet capillary to 275 °C and the spray voltage to 2.6 kV.

4. Collect the full MS scans at a resolving power of 100,000 (at 400 m/z). Use data dependent analysis (DDA) to select the top 6 most abundant precursor ions in each full MS for MS/MS fragmentation with a 3.0 m/z isolation window. Use collision-induced dissociation (CID) with a normalized energy of 35 eV to fragment selected ions. Set the resolving power of the MS/MS scans to 60,000 (at 400 m/z) with two micro scans. Set the AGC target to 5×10^5 for full MS scans and 3×10^5 for MS/MS scans. Collect MS data with Xcalibur 3.0.

3.5 Online High-pH/ Low-pH RPLC Separation and Top-Down MS/MS Analysis

1. The components of the online 2D nano-LC system consist of a C_4 capillary column (Waters BEH300 particles, 3.5 μm particle size and 300 Å pore size, 10 cm length, and 75 μm ID) for first-dimension high-pH separation, a C_5 nano-flow capillary column (Jupiter particles, 5 μm particle size and 300 Å pore size, 100 cm length, and 75 μm ID) for second-dimension separation, and a trapping column (Waters BEH300 particles, 3.5 μm particle size and 300 Å pore size, 20 mm length, and 150 μm

ID). Additionally, two Thermo Accela LC pumps (operational pressure 10,000 psi and operational flow rate 10–1000 μL/min), two high-pressure six-port switching valves, and two splitting columns are utilized. Assemble these components according to Fig. 3 to complete the setup of the online 2D nano LC system [15].

2. LC configurations for online sample injection and first-dimension separation are demonstrated in Fig. 3a, b, respectively. With switching valve #1 at position B, inject 5 μg of protein into the sample loop. Move switching valve #1 to position A to connect the sample loop to the first-dimension high-pH C4 column and load the sample onto the column using 100% high-pH MPA at a flow rate of 6 μL/min for 15 min.
3. During sample loading, set switching valve #2 to position A to connect the first-dimension high-pH column to the micro-trap column to collect the proteins that do not bind with the first-dimension column (flow-through proteins) on the SPE micro-trap column. Dilute the flow-through approximately 1:10 with low-pH MPA from pump 2 prior to the eluent entering the micro-trap column (*see* **Note 11**). To produce 1:10 dilution, regulate the flow rate through the first-dimension column to 3 μL/min and split the flow from pump 2 to produce a flow rate of 25 μL/min through the micro-trap column. The flow-through was then eluted from the micro-trap column and separated using the second-dimension low-pH column according to **step 4**.
4. To elute proteins from the first-dimension high-pH column, set switching valve #1 to position B to bypass the sample loop. Set switching valve #2 to position A to connect the first-dimension column to the micro-trap column. Elute proteins from the first-dimension column using 9-step gradients with steps from 0–40%, 40–45%, 45–50%, 50–55%, 55–60%, 60–65%, 65–70%, 70–75%, and 75–100% low-pH MPB for 10 min at 3-5 μL/min for each step of the gradient. Dilute the eluent from the first-dimension separation column approximately 1:10 with low-pH MPA from pump 2 prior to the eluent entering the micro-trap as discussed in **step 3**.
5. After each step of the gradient, the proteins were eluted from the micro-trap and second-dimension columns according to **step 6**. While the proteins are separated and eluted from the second-dimension high-pH column, equilibrate the first-dimension column back to 100% high-pH MPA and flow 100% high-pH MPA through the first-dimension column at 5 μL/min until the second-dimension separation for that fraction is complete. Upon completion of the second-dimension elution for each fraction, perform the next step of the first-dimension step gradient.

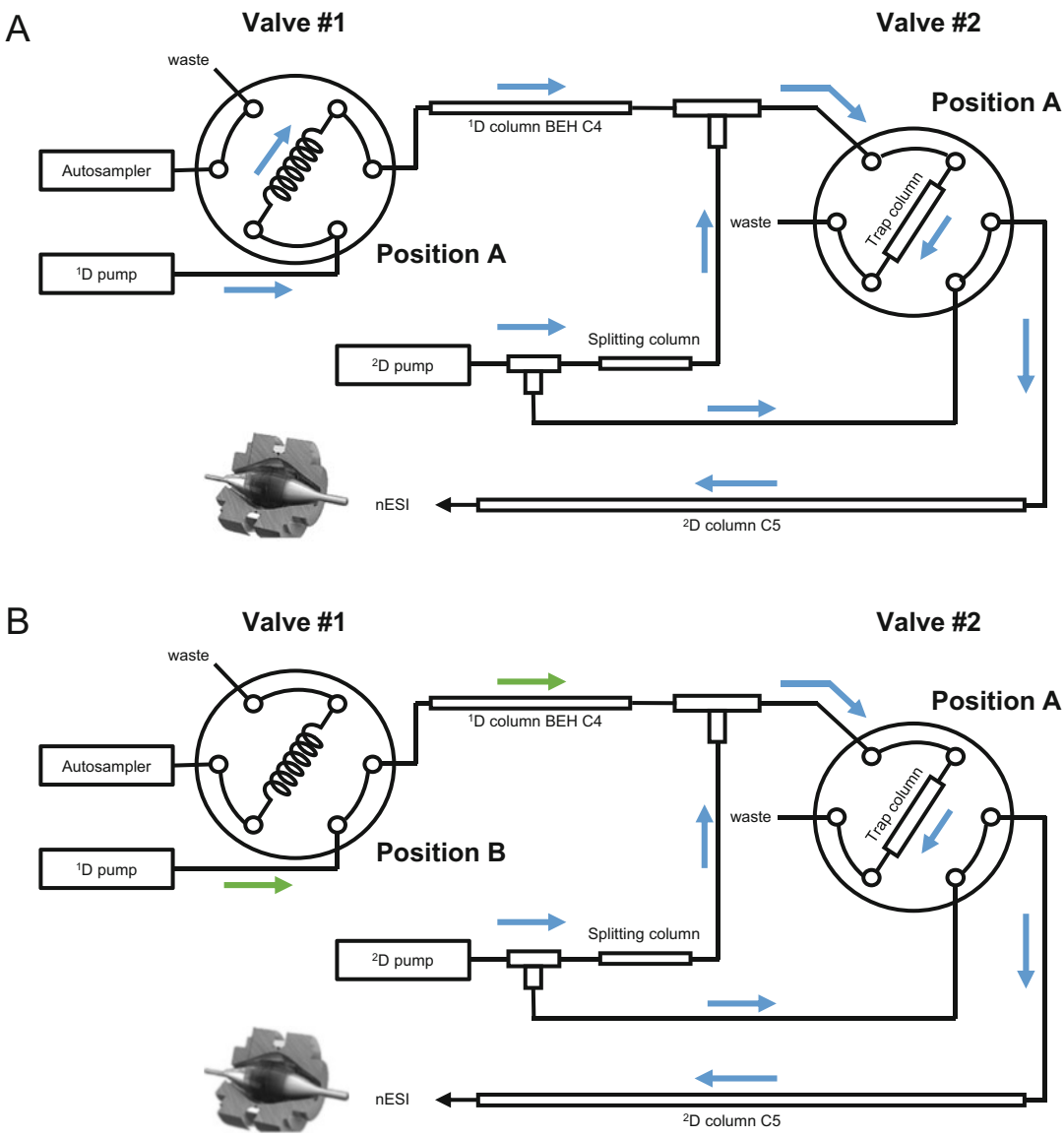


Fig. 3 Stepwise schematic of online 2D high-pH/low-pH nano-LC system. (Reprinted with permission from ref. 15. Copyright (2020) American Chemical Society)

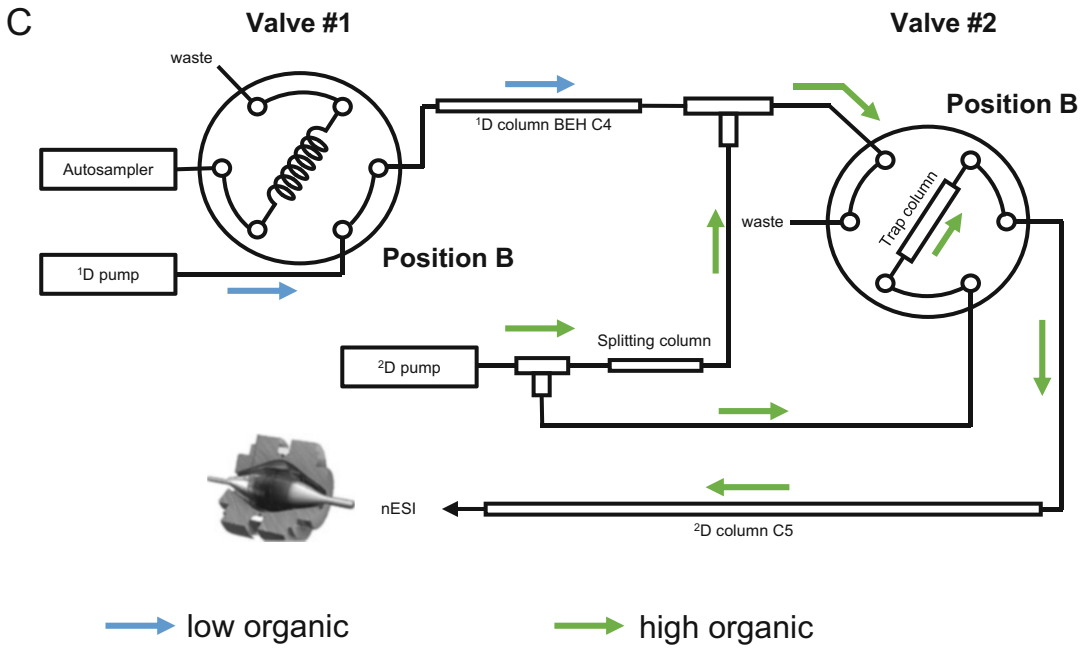


Fig. 3 (continued)

6. LC configuration for second-dimension sample separation is demonstrated in Fig. 3c. To elute the proteins from the micro-trap column and separate using the second-dimension low-pH column, set valve #2 to position B so that the micro-trap column is connected to the second-dimension low-pH column. Elute the proteins using a gradient from 10% to 70% low-pH MPB over 200 min at 400 µL/min. After the elution gradient, flush the second-dimension low-pH column with 90% low-pH MPB at 200 µL/min for 5 min and equilibrate with 10% low-pH MPB at 400 µL/min for 25 min. After the column is equilibrated, set valve #2 to position A to collect the next first-dimension fraction.
7. Repeat **steps 4–6** moving through the first-dimension step gradient until all protein have been eluted.
8. Couple the second-dimension low-pH column to an LTQ Orbitrap Elite mass spectrometer using a nano-ESI interface to perform top-down MS/MS analysis. Set the electrospray voltage to 2.6 kV and the inlet capillary temperature to 300 °C.
9. Set the MS resolving setting to 120,000 (m/z 400) with three micro scans and the maximum injection time to 1000 ms. Select the top six most abundant precursor ions in the MS scans for data-dependent MS/MS acquisition using an isolation window of 6.0. Do not select ions with charge states less than +4 for MS/MS fragmentation. Utilize dynamic exclusion

with repeat count 1, 90 s exclusion duration, and exclusion list size of 500. Fragment the selected precursor ions using CID with a normalized energy of 35%. Collect MS/MS spectra at a resolving power setting of 60,000 (m/z 400) with three micro scans and maximum injection time of 500 ms. Set the AGC target to 1×10^6 for MS and 3×10^5 for MS/MS. Collect the data using Xcalibur 3.0 software.

3.6 Protein and Proteoform Identification

1. Convert MS .raw files to .mzML using MSConvert (ProteoWizard) with default parameters.
2. Deconvolute the spectra using TopFD (Top-Down Mass Spectrometry Based Proteoform Identification and Characterization (TopPic) Suite) with a maximum mass of 50,000 Da and all other parameters set to default.
3. Search the deconvoluted .msalign data files produced in **step 2** against the annotated Swiss-Prot reviewed protein database for the appropriate cell type (*E. coli* or *HeLa*) using TopPic (TopPic Suite). Use an error tolerance of 15 ppm, maximum unexpected mass shift of ±500 Da, maximum number of unexpected modifications of 2, and spectrum level and proteoform level cutoff of EValue <0.01. Remove redundant proteoforms with mass differences ≤3.7 Da.

4 Notes

1. All reagents and buffers should be prepared with LC/MS grade solvents.
2. All LC/MS buffers solutions should be made air-free by sonicating for 20 min before use.
3. Trifluoroacetic acid (TFA) may react with oxygen-containing plastics and should only be transferred using glass pipettes.
4. *HeLa* cells must be cultured in a sterile environment in a biosafety level 2 hood and all reagents should be warmed to 37 °C before use.
5. *E. coli* cells must be cultured in a sterile environment and all lysogeny broth solutions, glassware, and pipette tips must be autoclaved before use.
6. Make sure the pellet is completely dissolved in the lysis buffer.
7. When removing the supernatant, do not disturb the debris pellet.
8. Wash the syringe filter using 1× PBS three times before using to filter the lysate.
9. For our system, the first protein peaks were observed approximately 16 min after sample injection.

10. If the fractions are not immediately used for second-dimension low-pH separation and analysis, they may be stored at −20 °C until needed.
11. The 1:10 dilution with low-pH MPA of the eluent from the first-dimension column is designed to reduce the percentage of organic and allow the proteins to bind to the micro-trap column.

References

1. Cupp-Sutton KA, Wu S (2020) High-throughput quantitative top-down proteomics. Molecular Omics 16:91–99
2. Zhang Z et al (2014) High-throughput proteomics. Annu Rev Anal Chem 7(1):427–454
3. Anderson LC et al (2017) Identification and characterization of human proteoforms by top-down LC-21 tesla FT-ICR mass spectrometry. J Proteome Res 16(2):1087–1096
4. Gregorich ZR, Ge Y (2014) Top-down proteomics in health and disease: challenges and opportunities. Proteomics 14(10):1195–1210
5. Camerini S, Mauri P (2015) The role of protein and peptide separation before mass spectrometry analysis in clinical proteomics. J Chromatogr A 1381:1–12
6. Ansong C et al (2013) Top-down proteomics reveals a unique protein S-thiolation switch in salmonella typhimurium in response to infection-like conditions. Proc Natl Acad Sci U S A 110(25):10153–10158
7. Shen Y et al (2017) High-resolution ultrahigh-pressure long column reversed-phase liquid chromatography for top-down proteomics. J Chromatogr A 1498:99–110
8. Jorgenson JW (2010) Capillary liquid chromatography at ultrahigh pressures. Annu Rev Anal Chem 3(1):129–150
9. Vellaichamy A et al (2010) Size-sorting combined with improved nanocapillary liquid chromatography-mass spectrometry for identification of intact proteins up to 80 kDa. Anal Chem 82(4):1234–1244
10. Lee JE et al (2009) A robust two-dimensional separation for top-down tandem mass spectrometry of the low-mass proteome. J Am Soc Mass Spectrom 20(12):2183–2191
11. Xiu L et al (2014) Effective protein separation by coupling hydrophobic interaction and reverse phase chromatography for top-down proteomics. Anal Chem 86(15):7899–7906
12. McCool EN et al (2018) Deep top-down proteomics using capillary zone electrophoresis-tandem mass spectrometry: identification of 5700 proteoforms from the Escherichia coli proteome. Anal Chem 90(9):5529–5533
13. Wang Z et al (2018) Two-dimensional separation using high-pH and low-pH reversed phase liquid chromatography for top-down proteomics. Int J Mass Spectrom 427:43–51
14. Yu D et al (2019) Deep intact proteoform characterization in human cell lysate using high-pH and low-pH reversed-phase liquid chromatography. J Am Soc Mass Spectrom 30(12):2502–2513
15. Wang Z et al (2020) Development of an online 2D ultrahigh-pressure nano-LC system for high-pH and low-pH reversed phase separation in top-down proteomics. Anal Chem 92(19): 12774–12777
16. Wang Z et al (2019) Top-down mass spectrometry analysis of human serum autoantibody antigen-binding fragments. Sci Rep 9(1):2345–2345

Chapter 5

Monolithic Materials-Based RPLC-MS for Proteoform Separation and Identification

Yu Liang, Lihua Zhang, and Yukui Zhang

Abstract

High-performance separation of proteoforms plays an important role in top-down proteomic ananlysis due to high complexity of the proteome. To this end, the functionalized ethylene-bridged hybrid monolithic materials have been developed for reversed-phase liquid chromatographic separation of proteoforms followed by online combination with high-resolution mass spectrometry (MS) for top-down proteomic analysis. Such monoliths have advantages of homogenously distributed functional groups in the framework, good chemical stability, and high permeability and, thus, show high resolution, good reproducibility, and low backpressure for proteoform separation. This chapter describes in detail the preparation of such monoliths and online combination with high-resolution MS for proteoform separation and identification.

Key words Monolithic materials, Bridged hybrid monolith, Proteoform separation, Top-down proteomic analysis

1 Introduction

Efficient separation of proteoforms prior to top-down mass spectrometry (MS) identification is very important due to high complexity of the proteome [1, 2]. Otherwise, the co-elution of proteoforms would make MS identification and data analysis difficult [3]. Reversed phase liquid chromatography (RPLC) is one of the most powerful methods for online separating proteoforms [4]. To reduce the co-elution of proteoforms, the columns packed with small particles and long columns have been developed [5, 6]. However, ultrahigh backpressure from small particles (9 k psi) and meter-long particle-packed column (14 k psi) affect the retention behavior of proteins [7, 8] and need the state-of-the-art UHPLC instrumentation, limiting its application.

The alternative stationary phases are monolithic materials [9–11], which have advantages of fast mass transfer, high permeability, and low backpressure. To achieve high-resolution protein

Liangliang Sun and Xiaowen Liu (eds.), *Proteoform Identification: Methods and Protocols*,
Methods in Molecular Biology, vol. 2500, https://doi.org/10.1007/978-1-0716-2325-1_5,
© The Author(s), under exclusive license to Springer Science+Business Media, LLC, part of Springer Nature 2022

separation, the monolithic materials should have homogeneous pore structure and uniform highly-loaded functional groups [12, 13]. In addition, chemical and mechanical stability is necessary to ensure good separation reproducibility. Periodic mesoporous organosilicas (PMOs) monoliths [12, 14–16], with organic groups directly integrated within the framework by sol-gel reaction of bridged organoalkoxysilane precursors ($(R'O)_3Si–R–Si(OR')_3$), have attracted much more attention in the field of chromatographic separation due to their unique features of narrow pore size distribution, high loading capacity of organic content, homogeneously distributed functional groups, and good chemical stability with Si-C bonds in the framework of structures.

In our previous work, the bridged hybrid monolithic matrix with 100% vinyl groups in the framework was prepared in the capillary by the one-step "sol-gel" reaction of bis(triethoxysilyl)-ethylene (BTSEY) [15]. With vinyl groups highly loaded and homogeneously distributed throughout the structure, the monolithic matrix can be facilely functionalized for chromatographic separation by effective "thiol-ene" click chemistry. Then C8 functionalized BTSEY monoliths were prepared and combined with maXis II Q-TOF MS for proteoform separation and identification [17]. With advantages of homogenously distributed C8 groups in the framework, good chemical stability, and high permeability, such monoliths showed high resolution, good reproducibility, and low backpressure for protein separation. In the analysis of the sample extracted from swine cardiac tissue, ~300 proteoforms with wide mass range of 3–104 kDa were separated in a single 80-min run. The effective protein separation allowed unambiguous identification of ~100 proteoforms with MW up to 104 kDa. Furthermore, such monoliths were also combined with Orbitrap Fusion™ Lumos™ for top-down proteomic analysis, and ~200 proteoforms were unambiguously identified from purified *E. coli* 70S ribosome in a single 72-min run. These results show great promising of such monoliths for proteoform analysis. Herein, we provide detailed procedures for the preparation of such monolithic columns and combination with high-resolution MS is also described.

2 Materials

2.1 Fused Silica Capillary Pretreatment

1. Fused silica capillary: ID 100 μm, OD 365 μm, length 3 m.
2. Water: deionized by Mill-Q (18 MΩ/cm at 25 °C).
3. 1 M HCl: dilute 8.32 mL of 37% (w/v) HCl to a final volume of 100 mL with water (*see* **Note 1**).
4. 1 M NaOH: dissolve 4.00 g of NaOH in 100 mL water.

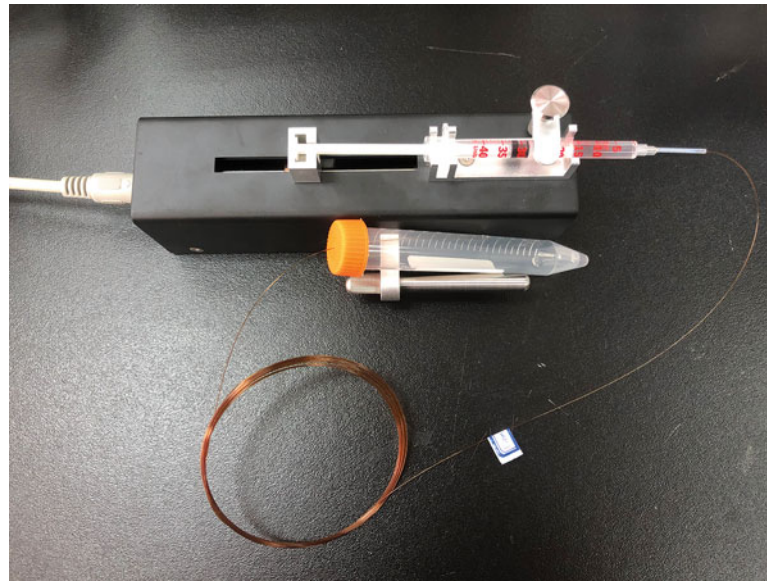


Fig. 1 Flushing of the solution into the capillary via the syringe pump and disposable syringe

5. 4% HF: dilute 100 μL of 40% (w/v) HF to a final volume of 1 mL with water (*see* **Note 2**).
6. Devices: the syringe pump and disposable syringe for flushing the solution into the capillary, shown in Fig. 1 (*see* **Note 3**).

2.2 Monolithic Column Preparation

1. Silane reagent: bis(triethoxysilyl)ethylene (BTSEY, 95%) (*see* **Note 4**).
2. Templates: F-127, hexadecyltrimethylammonium chloride (CTAC).
3. Solvents: methanol and water with HPLC grade.
4. Catalyst: triethylamine (TEA).
5. Devices: vortex mixer, water bath.

2.3 Monolithic Column Functionalization

1. Alkyl thiols: 1-octanethiol (*see* **Note 5**).
2. AIBN (2,2-azobisisobutyronitrile): recrystallized before use (*see* **Note 6**). Store at 4 °C.
3. Solvent: anhydrous ethanol.
4. Devices: vortex mixer, water bath.

2.4 Proteoform Separation and Identification

1. Mobile phase A: 2%(v/v) acetonitrile (ACN) containing 0.05% (v/v) trifluoroacetic acid (TFA) and 0.05% (v/v) formic acid (FA).
2. Mobile phase B: 80%(v/v) ACN containing 0.05%(v/v) TFA and 0.05% (v/v) FA.

3. Separation column: C8 functionalized bridged hybrid monolithic column with ID of 100 μm and length of 50 cm.
4. Instruments: nanoLC system (*see* **Note 7**) and high-resolution mass spectrometry.
5. Software: TopPIC Suite for proteoform identification.

3 Methods

3.1 Pretreatment of the Fused Silica Capillary

To make monolithic materials attached to the inner wall of the capillary, the capillary should be pretreated and the maximum number of silanol groups at the inner wall should be exposed.

1. The capillary is flushed continuously with 1 M HCl for 2 h at room temperature to remove impurities on the inner wall.
2. The capillary is flushed with water until the pH of washing solution is neutral.
3. The capillary is filled with 4% HF and then both ends were sealed by rubbers (*see* **Note 2**). Etching is performed at 35 °C for 3 h to increase the surface area of the inner wall (*see* **Note 8**).
4. After removing the rubbers and cutting the sealed ends, the capillary is flushed with water again until the pH of washing solution is neutral.
5. The capillary is flushed continuously with 1 M NaOH for 2 h at room temperature to regenerate silanol groups at the inner wall.
6. The capillary is rinsed with water until the pH of washing solution is neutral, and then rinsed with methanol.
7. The capillary is dried at 120 °C with nitrogen for about 12 h (*see* **Note 9**).

3.2 Synthesis of Ethylene-Bridged Hybrid Monolith in the Capillary

To make sure the good reproducibility of preparation, the room temperature is strictly controlled at 20 °C. Carry out all procedures at room temperature unless otherwise specified.

1. Weigh 10 mg of F127 and 32 mg of CTAC in 1.5 mL centrifugal tube, and resolve them in 160 μL of methanol and 20 μL of water. Add 120 μL of BTSEY and mix them with vortex mixer. Add 5 μL of TEA and mixed them again (*see* **Note 10**).
2. The prepared mixture is filled into the pretreated capillary using a disposable syringe as soon as possible (*see* **Note 11**).
3. With both ends sealed by rubber, the capillary is incubated at 40.0 °C for 12 h (*see* **Note 12**). By the sol-gel reaction (Fig. 2), the ethylene-bridged hybrid monolith is obtained.

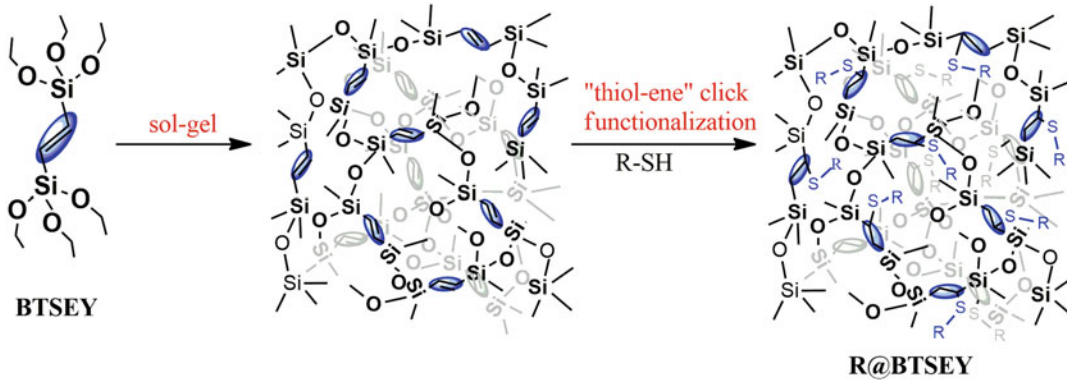


Fig. 2 Functionalized ethylene-bridged hybrid monolith

4. After removing the rubbers and cutting the sealed ends, the ethylene-bridged hybrid monolith is washed with methanol by HPLC pump to remove the templates and unreacted reagents, and then washed with ethanol for following modification (*see* **Note 13**).

3.3 "Thiol-ene" Click Modification of Ethylene-Bridged Hybrid Monolith

1. Add 173 μL of 1-octanethiol in 827 μL of ethanol (0.5 M), and then add 3 mg AIBN (3 mg/mL), followed by mixing them with vortex mixer to obtain a homogeneous solution (*see* **Note 5**).
2. The prepared mixture is degassed with nitrogen for 30 s to remove residual oxygen.
3. The prepared mixture is pumped into the ethylene-bridged hybrid monolith at room temperature by nitrogen (as shown in Fig. 3) until the whole column is filled with such solution (*see* **Note 14**). Both ends of the column are sealed by rubber, and the "thiol-ene" click reaction is carried out in a water bath at 65 °C for 4 h.
4. After removing the rubbers and cutting the sealed ends, the ethylene-bridged hybrid monolith is rinsed with ethanol by HPLC pump to remove unreacted regents.
5. Repeat **steps 1–4** twice to achieve maximum modification (*see* **Note 15**).
6. The functionalized column is sequentially washed with ethanol and acetonitrile before usage (*see* **Note 13**).

3.4 Functionalized Ethylene-Bridged Hybrid Monolithic Column Coupling to MS for "Top-Down" Proteomics

1. As shown in Fig. 4, one end of the C8 functionalized ethylene-bridged hybrid monolithic column (ID 100 μm × 25 ~ 50 cm) is combined with nanoLC system using ID 20 μm × 50 cm capillary and the MicroTee, and another end of such column is combined with MS using the nanoelectrospray emitter (*see* **Note 16**).

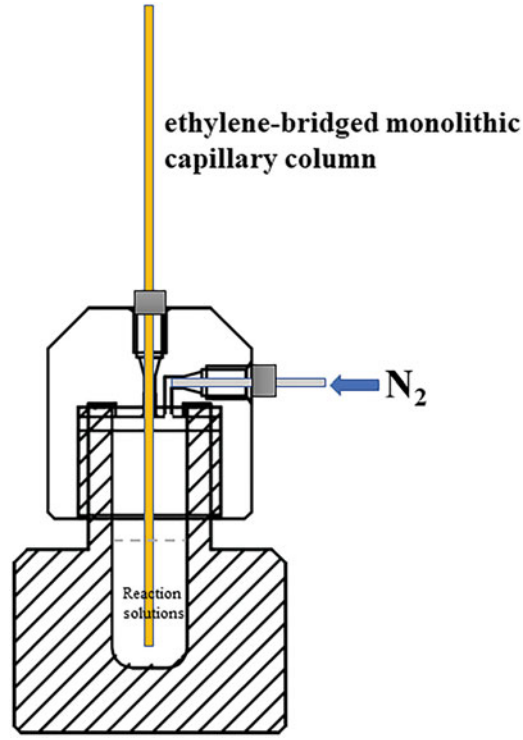


Fig. 3 Pumping of the reaction solution by N2

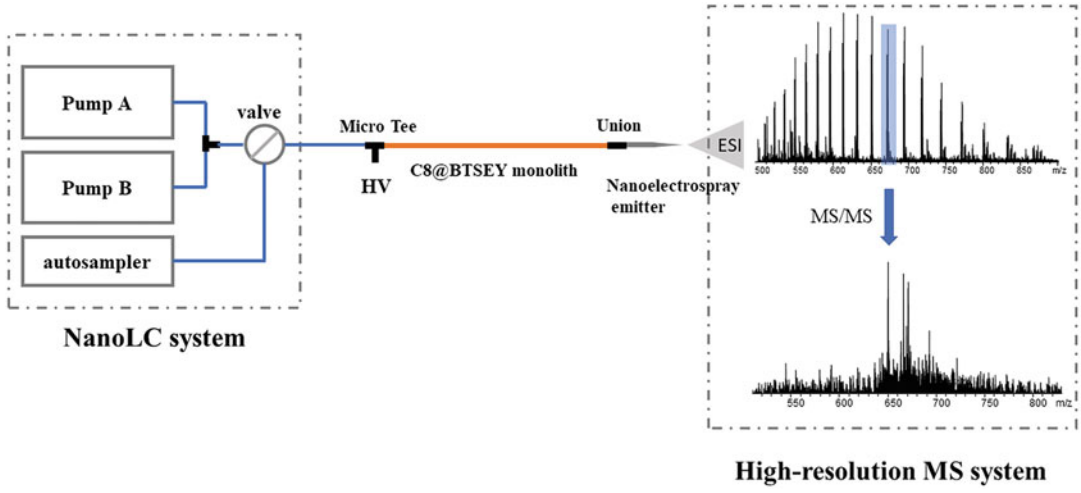


Fig. 4 C8-functionalized ethylene-bridged hybrid monolith coupling to MS for LC-MS analysis of proteoforms

2. Equilibrate the column with 10 μL mobile phase A at constant pressure or constant flow rate.
3. Sample is loaded onto such column with mobile phase A followed by separation and identification (*see* **Note 17**). The separation parameters are shown in Table 1 (*see* **Note 18**) and

Table 1
Parameters of proteoform separation with C8-functionalized ethylene-bridged hybrid monolithic column

Mobile phases Mobile phase A	2%(v/v) ACN + 0.05%(v/v) TFA + 0.05% (v/v) FA	
Mobile phase B	80%(v/v) ACN + 0.05%(v/v) TFA + 0.05% (v/v) FA	
Separation column	C8@BTSEY monolith (ID 100 μm × 25 ~ 50 cm)	
Sample loading		
Volume (μL)	7	
Constant pressure or constant flow rate	200 bar or 500 nL/min	
Injection volume (μL)	1	
Column equilibration		
Volume (μL)	10	
Constant pressure or constant flow rate	200 bar or 500 nL/min	
Separation gradient		
Time (mm:ss)	Flow (nl/min)	Mixture (%B)
00:00	500	0
01:00	500	10
61:00	500	70
62:00	500	95
72:00	500	95

the MS parameters could be set up based on the type of MS instrument according to some references. The example of the obtained data is shown in Fig. 5.

3.5 Data Analysis

The raw data were converted into .mzXML files with Msconvert. Then, spectral deconvolution was performed with TopFD to generate .msalign files followed by database search using TopPIC software.

4 Notes

1. Measure the volume of 37% (w/v) HCl in a fume hood because concentrated HCl is volatile liquid.
2. Extreme care must be taken to prevent exposure to HF liquid or vapor, since HF is extremely hazardous and corrosive. HF solutions should be used in a fume hood, and appropriate protective gloves should be worn.

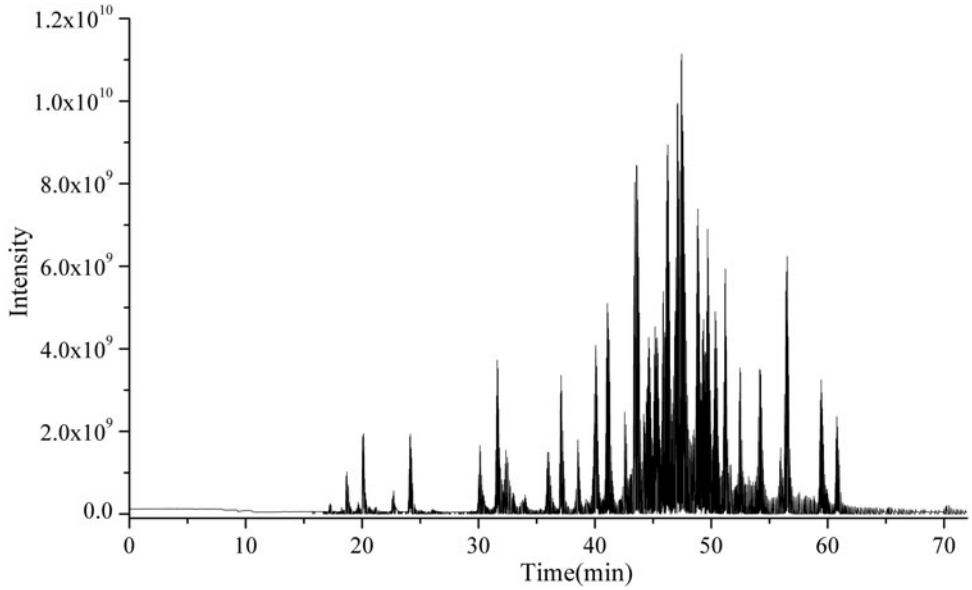


Fig. 5 Total ion chromatography of nanoLC-MS analysis of proteoforms from *E. coli* 70S ribosome using C8@BTSEY monolithic column and Orbitrap Fusion™ Lumos™. The separation conditions are shown in Table 1. The MS parameters are shown as follows. Data-dependent acquisition (DDA) is utilized with intact protein mode. Both the precursor and fragment ions are acquired with orbitrap at 120 K resolving power. The type of fragmentation is Higher-Energy Collision Dissociation (HCD) with Normalized Collision Energy (NCE) of 20%

3. The syringe-capillary interface is prepared by the Teflon tube with the length of about 2 cm. One end of the Teflon tube is tightly connected with the syringe and another end is tightly connected with the capillary.
4. BTSEY should be stored in vacuum drier at 4 °C since it is sensitive to hydrolysis.
5. Alkyl thiols for functionalization are not limited to 1-octanethiol. Other alkyl thiols, e.g., 1-butanethiol and 1-hexanethiol, can also be used for functionalization. Besides, use alkyl thiols in a fume hood since they are harmful.
6. For recrystallization, add 500 mL ethanol in a glass beaker and heat to 45 °C. Then, add 30.0 g AIBN while stirring quickly. After filtration, the filtrate is placed at room temperature for 4 h followed by placing at 4 °C for 12 h. The recrystallized AIBN is filtrated and dried in vacuum at room temperature and then stored at 4 °C.
7. NanoLC system should have accuracy flow rate from 200 to 800 nL/min and small dead volume to ensure good separation performance at the nano scale flow rate.
8. Etching by HF could make the monolithic materials attached onto the inner wall of the capillary and thus make sure good separation performance. The etching time and etching temperature should be well controlled.

9. The inner wall of the capillary should be completely dried in order to ensure that the following sol-gel reaction could be successfully performed.
10. Measure each components as accurately as possible, since subtle changes in the reaction mixture can drastically affect the properties of monolithic column [15]. Besides, the mixture proportion is just suitable to prepare the monolith in the ID-100-μm capillary. For the column with other ID, the mount of F127 and CTAC can be adjusted to ensure homogenous structure and good permeability.
11. The prepared mixture should be immediately filled into the pretreated capillary since the sol-gel reaction can occur even at the room temperature. Besides, avoid bubbles in the capillary during filling.
12. The temperature of the water bath should be exactly controlled by thermometer with division value of 0.1 °C. It is crucial to exactly control the temperature of the water bath since different sol-gel reaction temperatures affect the porous properties of the monolith [18].
13. The prepared monolithic materials should be homogenous and attached onto the inner wall of the capillary, which could be roughly characterized by optical microscopy or precisely characterized by SEM.
14. To make sure the whole column is completely filled with the reaction solution, the effluent volume should be twice the column volume.
15. The reaction solution should be prepared fresh each time, since thiol group easily gets oxygenized in the air.
16. The ID of nanoelectrospray emitter is preferred to be 10 μm to avoid possible clogging and ensure good stability of nanoelectrospray. Such emitter can be commercially available or prepared by gravity-assisted etching method [19].
17. The sample amount is often controlled below 1 μg to ensure good separation performance of the column.
18. TFA added in the mobile phases in RPLC as an ion-pairing reagent can sharpen peak shape. Nevertheless, the ion suppression from TFA reduces the signal intensity of proteins in MS analysis. To balance the peak shape of RPLC and the sensitivity of mass spectrometry, the amount of TFA in mobile phases is optimized as 0.05% (v/v) [17]. The flow rate is preferred to be 500 nL/min according to the ID of the column (100 μm), but higher flow rate up to 800 nL/min can also be used. The gradient can be changed according to the properties of the sample.

Acknowledgments

The authors are grateful for the financial support from National Natural Science Foundation (32088101, 21725506) and National Key Research and Development Program of China (2017YFA0505003).

References

1. Aebersold R, Agar JN, Amster IJ, Baker MS, Bertozzi CR, Boja ES, Costello CE, Cravatt BF, Fenselau C, Garcia BA, Ge Y, Gunawardena J, Hendrickson RC, Hergenrother PJ, Huber CG, Ivanov AR, Jensen ON, Jewett MC, Kelleher NL, Kiessling LL, Krogan NJ, Larsen MR, Loo JA, Ogorzalek Loo RR, Lundberg E, MacCoss MJ, Mallick P, Mootha VK, Mrksich M, Muir TW, Patrie SM, Pesavento JJ, Pitteri SJ, Rodriguez H, Saghatelian A, Sandoval W, Schlüter H, Sechi S, Slavoff SA, Smith LM, Snyder MP, Thomas PM, Uhlén M, Van Eyk JE, Vidal M, Walt DR, White FM, Williams ER, Wohlschlager T, Wysocki VH, Yates NA, Young NL, Zhang B (2018) How many human proteoforms are there? Nat Chem Biol 14:206–214
2. Regnier FE, Kim J (2018) Proteins and proteoforms: new separation challenges. Anal Chem 90:361–373
3. Schaffer LV, Millikin RJ, Miller RM, Anderson LC, Fellers RT, Ge Y, Kelleher NL, LeDuc RD, Liu X, Payne SH, Sun L, Thomas PM, Tucholski T, Wang Z, Wu S, Wu Z, Yu D, Shortreed MR, Smith LM (2019) Identification and quantification of proteoforms by mass spectrometry. Proteomics 19:e1800361
4. Chen B, Brown KA, Lin Z, Ge Y (2018) Top-down proteomics: ready for prime time? Anal Chem 90:110–127
5. Wu Z, Wei B, Zhang X, Wirth MJ (2014) Efficient separations of intact proteins using slip-flow with nano-liquid chromatography–mass spectrometry. Anal Chem 86:1592–1598
6. Shen YF, Tolic N, Piehowski PD, Shukla AK, Kim S, Zhao R, Qu Y, Robinson E, Smith RD, Pasa-Tolic L (2017) High-resolution ultrahigh-pressure long column reversed-phase liquid chromatography for top-down proteomics. J Chromatogr A 1498:99–110
7. Tanaka N, McCalley DV (2016) Core–Shell, ultrasmall particles, monoliths, and other support materials in high-performance liquid chromatography. Anal Chem 88:279–298
8. Fekete S, Guillarme D (2015) Estimation of pressure-, temperature- and frictional heating-related effects on proteins' retention under ultra-high-pressure liquid chromatographic conditions. J Chromatogr A 1393:73–80
9. Ma S, Li Y, Ma C, Wang Y, Ou J, Ye M (2019) Challenges and advances in the fabrication of monolithic bioseparation materials and their applications in proteomics research. Adv Mater 31:1902023
10. Lynch KB, Ren J, Beckner MA, He C, Liu S (2019) Monolith columns for liquid chromatographic separations of intact proteins: a review of recent advances and applications. Anal Chim Acta 1046:48–68
11. Toby TK, Fornelli L, Srzentić K, DeHart CJ, Levitsky J, Friedewald J, Kelleher NL (2019) A comprehensive pipeline for translational top-down proteomics from a single blood draw. Nat Protoc 14:119–152
12. Liang Y, Zhang L, Zhang Y (2019) Well-defined materials for high-performance chromatographic separation. Annu Rev Anal Chem 12:451–473
13. Dores-Sousa JL, Fernández-Pumarega A, De Vos J, Lämmerhofer M, Desmet G, Eeltink S (2019) Guidelines for tuning the macropore structure of monolithic columns for high-performance liquid chromatography. J Sep Sci 42:522–533
14. Mizoshita N, Tani T, Inagaki S (2011) Syntheses, properties and applications of periodic mesoporous organosilicas prepared from bridged organosilane precursors. Chem Soc Rev 40:789–800
15. Wu C, Liang Y, Yang K, Min Y, Liang Z, Zhang L, Zhang Y (2016) Clickable periodic mesoporous organosilica monolith for highly efficient capillary chromatographic separation. Anal Chem 88:1521–1525
16. Wu C, Liang Y, Zhu X, Zhao Q, Fang F, Zhang X, Liang Z, Zhang L, Zhang Y (2018) Macro-mesoporous organosilica monoliths with bridged-ethylene and terminal-vinyl: high-density click functionalization for

chromatographic separation. Anal Chim Acta 1038:198–205

17. Liang Y, Jin Y, Wu Z, Tucholski T, Brown KA, Zhang L, Zhang Y, Ge Y (2019) Bridged hybrid monolithic column coupled to high-resolution mass spectrometry for top-down proteomics. Anal Chem 91:1743–1747
18. Meinusch R, Ellinghaus R, Hormann K, Tallarek U, Smarsly BM (2017) On the underestimated impact of the gelation temperature on macro- and mesoporosity in monolithic silica. Phys Chem Chem Phys 19:14821–14834
19. Zhu XD, Liang Y, Weng YJ, Chen YB, Jiang H, Zhang LH, Liang Z, Zhang YK (2016) Gold-coated nanoelectrospray emitters fabricated by gravity-assisted etching self-termination and electroless deposition. Anal Chem 88: 11347–11351

Chapter 6

Capillary Isoelectric Focusing: Mass Spectrometry Method for the Separation and Online Characterization of Monoclonal Antibody Charge Variants at Intact and Subunit Levels

Jun Dai, Qiangwei Xia, and Chengjie Ji

Abstract

Monoclonal antibodies (mAbs) are one of the most widely used types of protein therapeutics. Charge variants are important quality attributes for evaluating developability, activity, and safety for mAb therapeutics. Here, we report a novel online capillary isoelectric focusing-mass spectrometry (CIEF-MS) method for mAb charge variant analysis using an electrokinetically pumped sheath-flow nanospray ion source on a time-of-flight (TOF) MS with a pressure-assisted chemical mobilization. Key factors that enable online CIEF-MS include effective capillary electrophoresis-MS (CE-MS) interface with enhanced sensitivity, utilization of MS-friendly electrolytes, beneficial effects of glycerol that reduces non-CIEF electrophoretic mobility and limits band broadening, appropriate ampholyte type and concentration selection for balanced separation resolution and MS detection sensitivity, optimized sheath liquid composition to realize high-resolution CIEF separation and effective MS electrospray ionization, as well as judiciously selected CIEF running parameters. The fundamental premise of CIEF has been verified by the linear correlation between isoelectric point (pI) values and migration time using a mixture of pI markers. By achieving high separation resolutions that are similar as those obtained from imaged CIEF (iCIEF), this method successfully provides highly sensitive MS identification for intact mAb charge variants. Furthermore, a middle-up sample treatment workflow can be adopted to provide in-depth charge variant analysis at subunit level for mAbs with complex charge heterogeneity. The mAb subunit CIEF-MS reveals the source of charge variant with enhanced resolution on both CIEF separation and MS spectra. This novel CIEF-MS method is a valuable tool with distinct advantage for objective and accurate assessment of charge heterogeneity of protein therapeutics.

Key words Monoclonal antibody, Charge variant, CIEF-MS, iCIEF, Electrokinetically pumped sheath-flow nanospray, Time-of-flight MS, Middle-up, Subunit

1 Introduction

Charge heterogeneity is an important quality attribute of protein therapeutics [1–3]. Because of its unequivocal solely pI-based separation mechanisms and very high resolution, CIEF, or imaged CIEF (iCIEF) developed later, has become a critical benchmark

Liangliang Sun and Xiaowen Liu (eds.), *Proteoform Identification: Methods and Protocols*,
Methods in Molecular Biology, vol. 2500, https://doi.org/10.1007/978-1-0716-2325-1_6,
© The Author(s), under exclusive license to Springer Science+Business Media, LLC, part of Springer Nature 2022

technique for routine charge variant analysis [4–6]. The commonly used optical detection has the advantages of easy hyphenation and provides universal response for quantitative analysis. However, optical detection is incapable of providing information for source of charge variants. Understanding the source of charge variance has become increasingly critical to help gain the insights of its implication, not only for protein molecular engineering optimization, developability assessment, and process control but also for evaluation of efficacy and safety as required by regulatory agency [7, 8]. As a powerful tool, MS provides unparalleled rich information for characterization and identification of therapeutic mAbs [9]. Combining the high-resolution capability of CIEF and the characterization power of MS would yield a highly desirable hyphenated analytical technique, with amplified merits for both techniques, for deciphering mAb charge heterogeneity.

Due to a number of reasons [10], online coupling of CIEF separation and electrospray ionization MS (ESI-MS) detection has been technically challenging. Effective online hyphenation of CIEF to ESI-MS has been hindered by several factors [11, 10]: (1) lack of effective interference between CE electrical circuit and ESI [12], (2) challenges in automation of the two-step CIEF process (i.e., focusing followed by mobilization with different catholytes at each step) [13], and (3) requirement to overcome the MS incompatibility of the key CIEF reagents [13–16].

Here, we report a reliable, high-resolution, fully automated CIEF-MS method for challenging mAb charge variant analysis using a common capillary-based CE instrument and a time-of-flight (TOF) MS system. In order to address the sensitivity issue of traditional high sheath flow CE-MS interfaces, we adopt a commercial EMASS-II CE-MS ion source with nano sheath flow that is based on the electrokinetically pumped sheath liquid technology [17–19]. Instead of switching the catholyte between focusing and chemical mobilization, the basic catholyte solution plug is first injected inside the capillary. Then, the sample mixture with low concentration of ampholyte is injected. After that, a voltage is applied to the separation capillary for focusing. Meanwhile, the acidic sheath solution at the catholyte side starts to titrate the catholyte. Pressure-assisted chemical mobilization of the focused sample starts when basic catholyte is consumed. Fully automated CIEF-MS is enabled by basic catholyte injection, acidic sheath solution titration, and pressure-assisted chemical mobilization.

In order to attain high CIEF resolution and to enable sensitive MS detection of mAb charge variants, a number of critical method parameters have been extensively optimized. It is found that the combination of utilizing 15–20% glycerol (in both electrolytes and sample buffer) with a sheath liquid containing acetic acid and acetonitrile yields high MS sensitivity and robust mobilization, and retains the high-resolution CIEF separation. We also

determined that 1.5% or less (v/v) Pharmalyte® 3–10 in the sample buffer offers good balance for CIEF resolution and MS sensitivity. For samples with high salt formulations, desalting is performed for effective CIEF focusing. In addition, a urea-based capillary rinsing procedure is implemented to favorably extend the life of coated capillaries.

The developed automated CIEF-MS method offers charge variants separation that correlates well with iCIEF-UV profiles, along with sensitive online TOF-MS detection and characterization of the individual charge variant [20].

For variants with relatively small mass differences or complex mAb molecules with heavy glycosylation, a middle-up approach can effectively reduce sample complexity through limited enzymatic cleavage and chemical reduction. For example, the intact mAb can be fragmented to subunits of light chain LC, antigen binding F $(ab')_2$ domain, single chain scFc, and Fd. The CIEF-MS method for the intact charge variant analysis can then be readily applied for fragmented mAb charge variants analysis [21].

2 Materials

Freshly prepare all solutions daily and store at 4 °C.

1. Agilent 7100 CE (Agilent Technologies, Santa Clara, CA).
2. Agilent 6200 series TOF (Agilent Technologies, Santa Clara, CA).
3. EMASS-II CE-MS ion source (CMP Scientific Corp, Brooklyn, NY).
4. Capillary: neutral coating PS1 capillary at 75 cm length with 360 μm OD and 50 μm ID (CMP Scientific Corp, Brooklyn, NY).
5. Electrospray emitter: 1.0 mm OD, 0.75 mm ID, and 20–30 μm tip size (CMP Scientific Corp, Brooklyn, NY).
6. Carrier ampholyte: Pharmalyte® 3–10 (GE Healthcare).
7. Catholyte: 0.2 N ammonium hydroxide aqueous solution with 15% glycerol.
8. Anolyte: 1% formic acid aqueous solution with 15% glycerol.
9. Sheath solution: acetic acid, acetonitrile, and water at volume ratio of 20:25:55.
10. CIEF solution: 3% Pharmalyte® 3–10 aqueous solution with 40% glycerol.
11. Buffer exchange solution: 10 mM ammonium acetate at pH 6.5.

12. pI marker solutions at 4.1, 5.5, 7.0, and 9.5 (Sciex, Framingham, MA).
13. Filter for buffer exchange: Amicon® Ultra 0.5 mL 10 K centrifugal filter units (EMD Millipore, Billerica, MA).
14. Protease: IdeS protease (Promega Corp., Madison, WI).
15. Reduction solution: 1M dithiothreitol (DTT) aqueous solution.

3 Methods

3.1 Online CIEF-MS

1. The Agilent 6200 series TOF mass spectrometer is coupled with an Agilent 7100 CE, using a CMP Scientific EMASS-II CE-MS ion source. The regular ESI ion source on the TOF is modified to accommodate nanospray by replacing the ESI spray shield with an Agilent nanospray shield, which comes with a gas diverter (*see* **Note 1**).
2. Install the capillary to CE instrument and EMASS-II CE-MS ion source by following the instructions provided by the manufacturers.
3. Rinse the capillary with water at 100 mbar for 10 min.
4. Rinse the capillary with anolyte at 100 mbar for 20 min (*see* **Note 2**).
5. Install the spray emitter by following the instructions provided by CMP Scientific.
6. Adjust the emitter and capillary with the help of a microscope camera: Move the spray emitter toward the mass spectrometer and maintain a distance of 2–4 mm between the emitter tip and the spray shield. Adjust the capillary outlet tip to a ~1 mm distance from the spray emitter tip.
7. Set TOF MS conditions: (1) set capillary voltage on TOF at 0 V; (2) set drying gas at 350 °C with a flow at 6 L/min; (3) set skimmer voltage at 65 V and OCT 1RF Vpp at 750 V; (4) set the fragmentor voltage (FV) at 380–400 V for intact mAb, 200–380 V for fragmented mAb, and 125 V for pI marker and small proteins (*see* **Note 3**).
8. Turn on the external power supply that comes with the EMASS-II CE-MS ion source and set at 2.0–2.4 kV (*see* **Note 4**).
9. Prepare sample solution: mix equal volume buffer exchanged sample (at 0.2–2.0 mg/mL in 10 mM ammonium acetate) with CIEF solution (3% Pharmalyte® 3–10 with 40% glycerol) to achieve a final composition of 1.5% Pharmalyte® 3–10 with 20% glycerol (*see* **Note 5**).

10. CIEF-MS analysis: (1) inject catholyte solution under 950 mbar for 10 s, (2) inject sample solution under 950 mbar for 75 s, (3) apply a voltage with a field strength of 250 V/cm and a pressure of 10 mbar for CIEF separation. (4) After complete elution of the analyte peak (~75 min), increase the pressure to 100 mbar and flush out residual ampholyte to get the capillary ready for next sample analysis. The total run time is about 100 min (*see* **Note 6**).

3.2 Protein Sample Treatment

1. Add deionized water to the lyophilized IdeS Protease to make a 50 unit/μL solution.
2. Add 60 μg of sample to PBS to get the sample solution at ~1 mg/mL.
3. Add 1 unit of IdeS protease to 1 ug mAb sample.
4. Incubate the sample at 37 °C for 30 min. Split the sample into two aliquots. Take one aliquot as the IdeS-digested sample.
5. Take the second aliquot of IdeS-digested sample and spike in 1M DTT solution to reach 50 mM DTT concentration.
6. Incubate the sample with 50 mM DTT at 37 °C for 30 min to get the IdeS- plus reduction-treated sample.
7. Desalt and buffer exchange (*see* Subheading 3.3) the treated samples prior to CIEF-MS analysis.

3.3 Protein Desalting and Buffer Exchange Procedure

Salt-free protein standards can be used without desalting. The formulated or treated samples are desalted and buffer exchanged using EMD Millipore (Billerica, MA) Amicon centrifugal filter units.

1. Add 400 μL 10 mM ammonium acetate (pH 6.5) and 30 uL sample to an Amicon® Ultra 0.5 mL10 K filter.
2. Centrifuge the sample at 14,000 × *g* 10 °C for 12 min (*see* **Note 7**).
3. Recover the desalted solute by reversing the filter and spin at 1000 × *g* for 2 min.

3.4 Capillary Cleaning Procedure

Rinse the PS1 capillary at the end of day for cleaning.

1. Remove the emitter: Move the spray emitter away from the mass spectrometer and remove the emitter from the capillary by following the instructions provided by CMP Scientific.
2. Clean the capillary: Rinse the capillary at 950 mbar with solutions in the order of water (5 min), 4.5 M urea (5 min), water (5 min), followed by anolyte buffer (10 min).

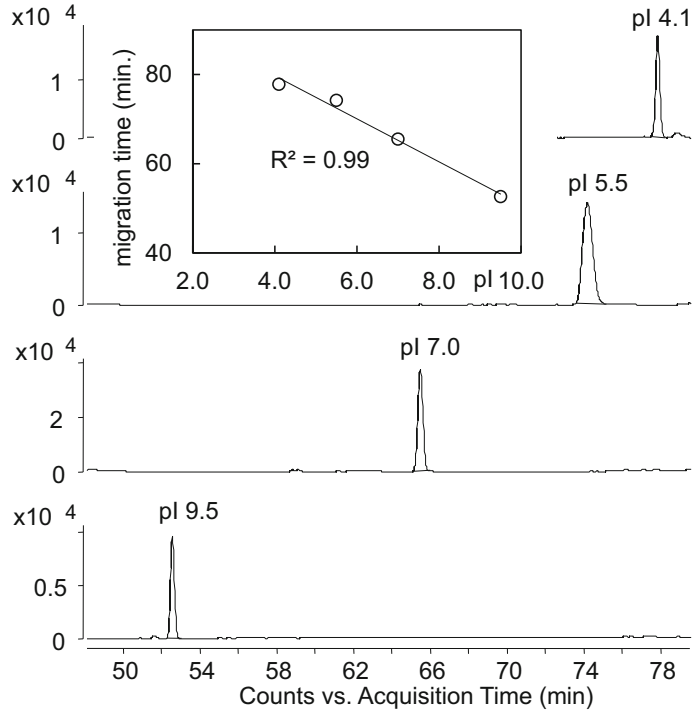


Fig. 1 CIEF-MS separation of four pI markers. From top to bottom: the extracted ion electropherograms of pI 4.1 (m/z 591.25), pI 5.5 (m/z 471.19), pI 7.0 (m/z 627.29), and pI 9.5 (m/z 950.47). Inset: Linear regression fitting of pI values and corresponding migration time. (Figure adapted with permission from ACS https://pubs.acs.org/doi/full/10.1021/acs.analchem.7b04608 ref. 20. Further permissions related to the material excerpted should be directed to the ACS)

3.5 Example Results

3.5.1 Linearity of Established pH Gradient

Figure 1 shows the separation of four pI markers and the linear regression between the pI values and corresponding migration time [20]. This linearity demonstrates the established pH gradient of the CIEF-MS method.

3.5.2 CIEF-MS Intact Charge Variant Characterization of mAbs

The charge variant profiles revealed by the CIEF-MS analysis are highly consistent with those in iCIEF-UV experiments as shown in Figs. 2 and 3 for bevacizumab and infliximab (*see* **Note 8** for iCIEF-UV conditions used) [20].

Figure 2 illustrates the comparison of bevacizumab CIEF-MS and iCIEF-UV results [20]. Good MS spectra enable reliable deconvolution for intact mass. Based on the deconvoluted intact mass (Fig. 2c), main peak M has a mass of 149,202 Da. Basic peak B1 matches the common C-terminal lysine variant with a mass difference from the main species (Δm) of 128 Da (+1K), and basic peak B2 has a Δm of −17 Da. Acidic peak A1 has a Δm of +1 Da, possibly a deamidation species. No reliable intact mass information has been derived for peak A2 because of the weak signal.

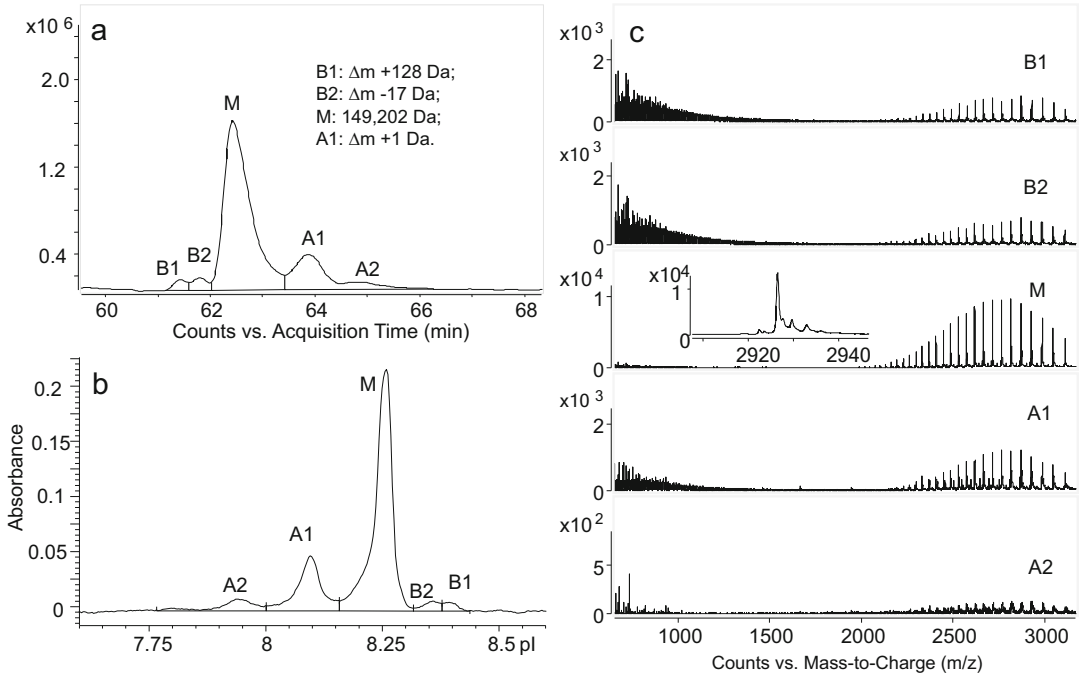


Fig. 2 Bevacizumab CIEF-MS analysis in comparison with iCIEF-UV. (**a**) Extracted ion electropherogram (m/z 2960–3200 Da) showing basic variants B1 and B2, main peak M, and acidic variants A1 and A2; (**b**) iCIEF-UV electropherogram; (**c**) mass spectra of major variants. Mass of each variant shown was based on the most abundant deconvoluted peak. (Figure adapted with permission from ACS https:/pubs.acs.org/doi/full/10.1021/acs.analchem.7b04608 ref. 20. Further permissions related to the material excerpted should be directed to the ACS)

Figure 3 gives the charge variant separation and identification of infliximab [20]. Deconvoluted mass confirms that the two basic species are the C-terminal lysine (K) variants [22–24] with B1 as +2K (Δm +258 Da) and B2 as +1K (Δm +129 Da) variants. Deconvoluted mass for acidic variant revealed intact MS at Δm of +5 Da, indicating the possibility of deamidation species.

3.5.3 Middle-Up CIEF-MS Charge Variant Characterization of mAb Subunits

Charge variant analysis of mAbs with significant microheterogeneity (e.g., cetuximab, with complex glycosylation at both Fc and Fab regions [25]) can be greatly benefited from middle-up CIEF-MS approach. Figure 4 shows the comparison of CIEF-MS and iCIEF-UV profiles of cetuximab at intact and subunit levels [21]. Nine charge variants are observed at the intact level. Eight charge variants are detected after IdeS digestion, and 11 charge variants are separated after IdeS plus reduction treatment. Peak assignments for all variants are achieved based on the MS spectra [21].

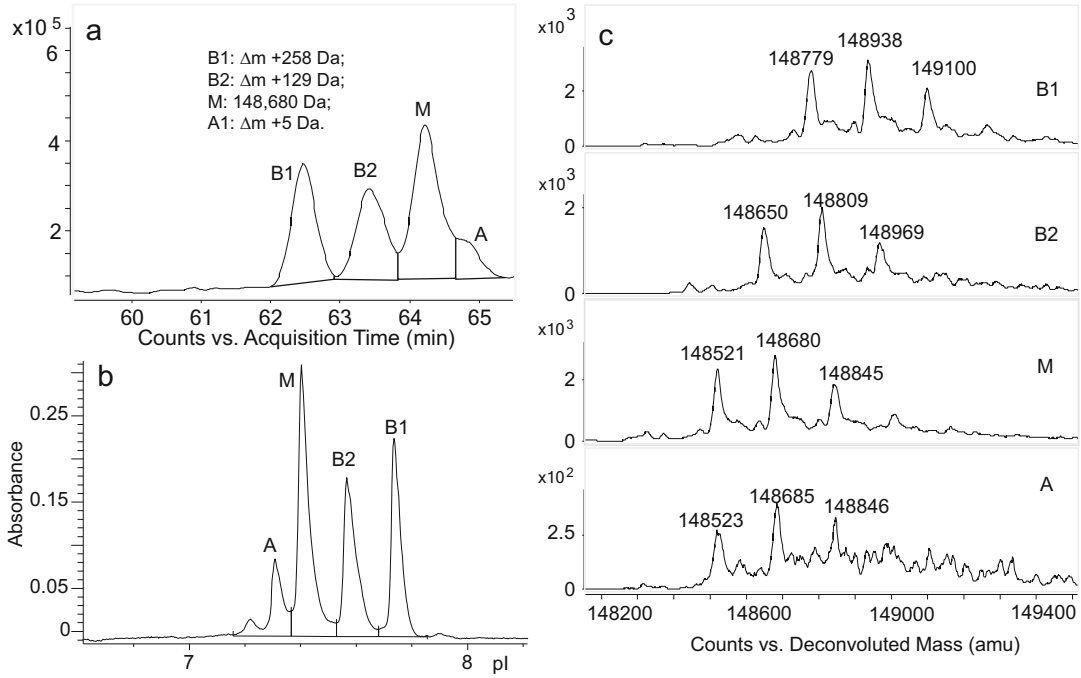


Fig. 3 Infliximab CIEF-MS analysis in comparison with iCIEF-UV. (**a**) Extracted ion electropherogram (m/z 2960–3200 Da): basic variants B1 and B2, main peak M, and acidic variant A; (**b**) iCIEF-UV electropherogram; (**c**) deconvoluted mass spectra. (Figure adapted with permission from ACS https:/pubs.acs.org/doi/full/10.1021/acs.analchem.7b04608 ref. 20. Further permissions related to the material excerpted should be directed to the ACS)

4 Notes

1. The regular ESI source spray shield on the 6200 series TOF MS is not ideal for nanospray generated by EMASS-II ion source. To circumvent the adverse effects of high volume of drying gas, a single bore inline spray shield with a radial gas diverter is used for gentle gas delivery for effective introduction of analytes into MS.
2. Use different anolyte solution vials for washing and CIEF separation. Use a new set of electrolyte solution vials for each sample analysis.
3. For samples that range with different molecular weights, two fragmentor voltages on the TOF mass spectrometer that are optimal for different sizes of molecule can be set within the same acquisition experiment (e.g., 200 V for ~25 kDa scFc and 380 V for~100 kDa F(ab')2).
4. Use real-time monitoring on TOF instrument to check the electrospray before injecting sample. Fine-tune the distance between the spray emitter tip and the spray shield if needed.

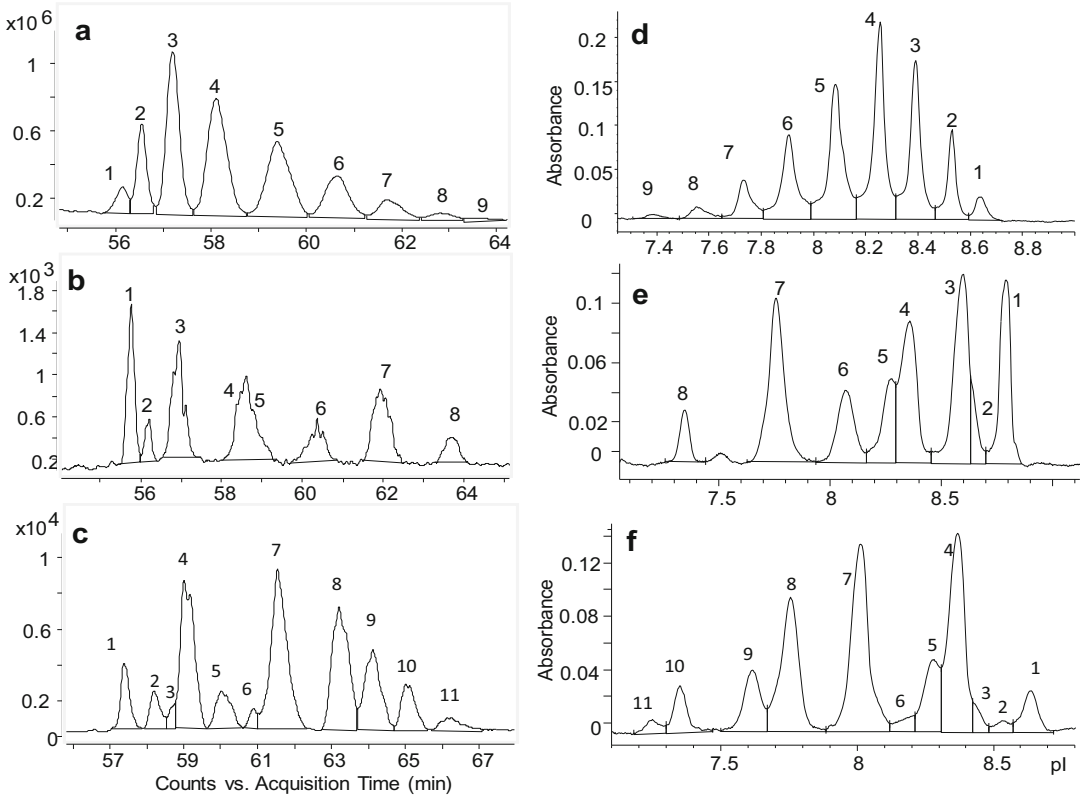


Fig. 4 Intact and fragmented cetuximab CIEF-MS analysis in comparison with iCIEF-UV. The CIEF-MS results are as follows: (**a**) extracted ion electropherogram of intact mAb at m/z 2960–3200, (**b**) extracted base peak electropherogram of IdeS-treated mAb at m/z 1500–2600, (**c**) extracted base peak electropherogram of IdeS-plus reduction-treated mAb at m/z 1000–2000. The iCIEF-UV results are as follows: (**d**) intact, (**e**) IdeS treated, (**f**) IdeS plus reduction treated. (Figure adapted with permission from ref. 21. Copyright (2018) American Chemical Society)

5. Minimum sample volume required is ~10 μL using Agilent CE sample vial (Agilent PN 9301-0978).
6. After sample injection, the initial current should be at ~2.5–1 μA. The current normally drops below 1 μA in ~10 min and reaches its lowest plateau (~0.5 μA) at ~30 min. The current remains at its lowest plateau till the elution of the analyte peaks. After increasing the pressure from 10 to 100 mbar, current gradually increases. Once the current reaches its maximum plateau (7–9 μA), the system becomes ready for next sample analysis.
7. For samples that have very high salt formulations, the buffer exchange process can be performed twice for effective desalting.
8. Imaged CIEF-UV conditions used for example results: The iCIEF-UV separations were performed on iCE3

(ProteinSimple, San Jose, CA). Fluorocarbon-coated capillary cartridge (5 cm × 100 μm ID) is used with 80 mM phosphoric acid as the anolyte and 100 mM sodium hydroxide as the catholyte, with 0.1% methyl cellulose in both electrolytes. Sample buffer contains 0.35% methyl cellulose, 4% Pharmalyte® 3–10, and 2 M urea. Focusing is conducted at 1.5 kV for 1 min, followed by 3.0 kV for 8 min.

References

1. Du Y, Walsh A, Ehrick R, Xu W, May K, Liu H (2012) Chromatographic analysis of the acidic and basic species of recombinant monoclonal antibodies. MAbs 5:578–585. Taylor & Francis
2. Liu H, Ponniah G, Zhang H-M, Nowak C, Neill A, Gonzalez-Lopez N, Patel R, Cheng G, Kita AZ, Andrien B (2014) In vitro and in vivo modifications of recombinant and human IgG antibodies. MAbs 5:1145–1154. Taylor & Francis
3. Gangopadhyay A, Petrick AT, Thomas P (1996) Modification of antibody isoelectric point affects biodistribution of 111-indium-labeled antibody. Nucl Med Biol 23(3): 257–261
4. Wu J, Pawliszyn J (1992) Capillary isoelectric focusing with a universal concentration gradient imaging system using a charge-coupled photodiode array. Anal Chem 64(23): 2934–2941
5. Sosic Z, Houde D, Blum A, Carlage T, Lyubarskaya Y (2008) Application of imaging capillary IEF for characterization and quantitative analysis of recombinant protein charge heterogeneity. Electrophoresis 29(21):4368–4376. https://doi.org/10.1002/elps.200800157
6. Michels DA, Salas-Solano O, Felten C (2011) Imaged capillary isoelectric focusing for charge-variant analysis of biopharmaceuticals. BioProcess Int 9(10):48–54
7. Wu G, Yu C, Wang W, Wang L (2018) Inter-laboratory method validation of icIEF methodology for analysis of monoclonal antibodies. Electrophoresis 39(16):2091–2098
8. Dakshinamurthy P, Mukunda P, Kodaganti BP, Shenoy BR, Natarajan B, Maliwalave A, Halan V, Murugesan S, Maity S (2017) Charge variant analysis of proposed biosimilar to Trastuzumab. Biologicals 46:46–56
9. Beck A, Sanglier-Cianférani S, Van Dorsselaer A (2012) Biosimilar, biobetter, and next generation antibody characterization by mass spectrometry. Anal Chem 84(11):4637–4646. https://doi.org/10.1021/ac3002885
10. Hühner J, Lämmerhofer M, Neusüß C (2015) Capillary isoelectric focusing-mass spectrometry: coupling strategies and applications. Electrophoresis 36(21–22):2670–2686
11. Zhong X, Maxwell EJ, Ratnayake C, Mack S, Chen DD (2011) Flow-through microvial facilitating interface of capillary isoelectric focusing and electrospray ionization mass spectrometry. Anal Chem 83(22):8748–8755
12. Týčová A, Ledvina V, Klepárník K (2017) Recent advances in CE-MS coupling: instrumentation, methodology, and applications. Electrophoresis 38(1):115–134. https://doi.org/10.1002/elps.201600366
13. Tang Q, Harrata AK, Lee CS (1995) Capillary isoelectric focusing-electrospray mass spectrometry for protein analysis. Anal Chem 67(19):3515–3519
14. Kristl T, Stutz H, Wenz C, Rozing G (2014) Principles and applications of capillary isoelectric focusing. Agilent Technologies, Inc, Primer
15. Cruzado-Park ID, Mack S, Ratnayake CK (2008) Identification of system parameters critical for high-performance cIEF. AB SCIEX, Application Notes A-11634A
16. Mack S, Cruzado-Park I, Chapman J, Ratnayake C, Vigh G (2009) A systematic study in CIEF: defining and optimizing experimental parameters critical to method reproducibility and robustness. Electrophoresis 30(23): 4049–4058
17. Wojcik R, Dada OO, Sadilek M, Dovichi NJ (2010) Simplified capillary electrophoresis nanospray sheath-flow interface for high efficiency and sensitive peptide analysis. Rapid Commun Mass Spectrom 24(17):2554–2560. https://doi.org/10.1002/rcm.4672
18. Sun L, Zhu G, Zhao Y, Yan X, Mou S, Dovichi NJ (2013) Ultrasensitive and fast bottom-up analysis of femtogram amounts of complex proteome digests. Angew Chem Int Ed 52(51):13661–13664. https://doi.org/10.1002/anie.201308139

19. Peuchen EH, Zhu G, Sun L, Dovichi NJ (2017) Evaluation of a commercial electrokinetically pumped sheath-flow nanospray interface coupled to an automated capillary zone electrophoresis system. Anal Bioanal Chem 409(7):1789–1795. https://doi.org/10.1007/s00216-016-0122-8
20. Dai J, Lamp J, Xia Q, Zhang Y (2018) Capillary isoelectric focusing-mass spectrometry method for the separation and online characterization of intact monoclonal antibody charge variants. Anal Chem 90(3):2246–2254
21. Dai J, Zhang Y (2018) A middle-up approach with online capillary isoelectric focusing/mass spectrometry for in-depth characterization of Cetuximab charge heterogeneity. Anal Chem 90(24):14527–14534
22. Shion H, Chen W (2016) Comparison of infliximab and a biosimilar via subunit analysis using the waters biopharmaceutical platform with UNIFI. Waters Corporation, Application Note 720004796EN
23. Redman EA, Batz NG, Mellors JS, Ramsey JM (2015) Integrated microfluidic capillary electrophoresis-electrospray ionization devices with online MS detection for the separation and characterization of intact monoclonal antibody variants. Anal Chem 87(4):2264–2272. https://doi.org/10.1021/ac503964j
24. Jung SK, Lee KH, Jeon JW, Lee JW, Kwon BO, Kim YJ, Bae JS, Kim D-I, Lee SY, Chang SJ (2014) Physicochemical characterization of Remsima®. MAbs 6:1163–1177. Taylor & Francis.
25. Ayoub D, Jabs W, Resemann A, Evers W, Evans C, Main L, Baessmann C, Wagner-Rousset E, Suckau D, Beck A (2013) Correct primary structure assessment and extensive glyco-profiling of cetuximab by a combination of intact, middle-up, middle-down and bottom-up ESI and MALDI mass spectrometry techniques. MAbs 5(5):699–710

Chapter 7

Proteoform Analysis and Construction of Proteoform Families in Proteoform Suite

Leah V. Schaffer, Michael R. Shortreed, and Lloyd M. Smith

Abstract

Proteoform Suite is an interactive software program for the identification and quantification of intact proteoforms from mass spectrometry data. Proteoform Suite identifies proteoforms observed by intact-mass (MS1) analysis. In intact-mass analysis, unfragmented experimental proteoforms are compared to a database of known proteoform sequences and to one another, searching for mass differences corresponding to well-known post-translational modifications or amino acids. Intact-mass analysis enables proteoforms observed in the MS1 data without MS/MS (MS2) fragmentation to be identified. Proteoform Suite further facilitates the construction and visualization of proteoform families, which are the sets of proteoforms derived from individual genes. Bottom-up peptide identifications and top-down (MS2) proteoform identifications can be integrated into the Proteoform Suite analysis to increase the sensitivity and accuracy of the analysis. Proteoform Suite is open source and freely available at https://github.com/smith-chem-wisc/proteoform-suite.

Key words Proteoform, Proteoform family, Top-down proteomics, Mass spectrometry, Post-translational modification

1 Introduction

Much of the biological complexity in cells is due to variations at the protein level. Proteoforms are the different forms of a protein that arise due to biological processes in the cell, such as alternative splicing, amino acid variation, and post-translational modifications (PTMs) [1]. The set of proteoforms derived from the same gene is a proteoform family [2]. Proteoform identification is typically performed by top-down mass spectrometry, where intact proteins are analyzed by liquid chromatography–tandem mass spectrometry (LC-MS/MS), which includes by definition fragmentation of the precursor [3–5]. However, due to sensitivity limitations, many proteoforms observed at the MS1-level are either not selected for MS2 fragmentation or do not have high enough quality MS2 fragmentation data for proteoform identification.

Liangliang Sun and Xiaowen Liu (eds.), *Proteoform Identification: Methods and Protocols*,
Methods in Molecular Biology, vol. 2500, https://doi.org/10.1007/978-1-0716-2325-1_7,
© The Author(s), under exclusive license to Springer Science+Business Media, LLC, part of Springer Nature 2022

We developed the software program Proteoform Suite (https://github.com/smith-chem-wisc/proteoform-suite) to construct proteoform families and identify proteoforms based on analysis of MS1-level intact-mass data [6]. A list of unique experimental proteoforms from a sample is generated by aggregating results from top-down MS/MS search results (identified proteoforms) and/or MS1 deconvolution results (observed but unidentified proteoforms) [7]. Aggregated experimental proteoform masses are compared to both a database of theoretical proteoform masses (experiment–theoretical comparisons) and to one another (experiment–experiment comparisons) by Proteoform Suite. Proteoform families are then constructed by grouping together experimental and theoretical proteoforms with mass differences corresponding to known PTMs and amino acids. In each proteoform family, the experiment–theoretical and experiment–experiment comparisons are used to identify additional experimental proteoforms by intact mass. Proteoform families are visualized as a network of nodes (proteoform masses) and edges (mass differences corresponding to modifications or amino acid differences) using the software program Cytoscape [8, 9].

This chapter describes the workflow used to perform intact-mass analysis and construct proteoform families in Proteoform Suite. We first describe preprocessing steps performed on the raw mass spectrometry files to generate results for Proteoform Suite input. We then describe how to perform intact-mass analysis in Proteoform Suite and the visualization of proteoform families. Finally, we discuss remaining challenges in intact-mass analysis and caveats to be aware of when using Proteoform Suite.

Proteoform analysis in Proteoform Suite requires mass spectrometry data from an intact protein sample. This may include intact-only data (MS1) or tandem MS/MS data, which contains both MS1 and MS2 spectra. See previous work for information on sample preparation, fractionation, and mass spectrometry settings [10, 11]. To observe a greater number of proteoforms at the MS1 level, we recommend acquisition of both MS1-only and tandem MS/MS data from the sample.

A release of Proteoform Suite can be downloaded from https://github.com/smith-chem-wisc/proteoform-suite. At least 8 GB of RAM is recommended, with more required for larger human databases. A 64-bit operating system should be utilized with .NET Core 3.1 installed. Release version 0.4.0 contains a user manual with detailed descriptions of the parameters and results on each page of Proteoform Suite.

2 Data Preprocessing

A typical Proteoform Suite analysis requires MS1 deconvolution results, top-down search results, and a protein database (Fig. 1). The Proteoform Suite graphical user interface (GUI) opens on the Load Results page (Fig. 2). On this page, results are loaded in for the analyses performed on subsequent pages (select Standard under Choose Analysis, labeled 1 in Fig. 2). If necessary, this page can also be used to generate deconvolution and top-down results and to calibrate results. Subheadings 2.1 through 2.5 describe how to obtain results to input on the Load Results page using either external software programs or other analysis options on the Load Results page.

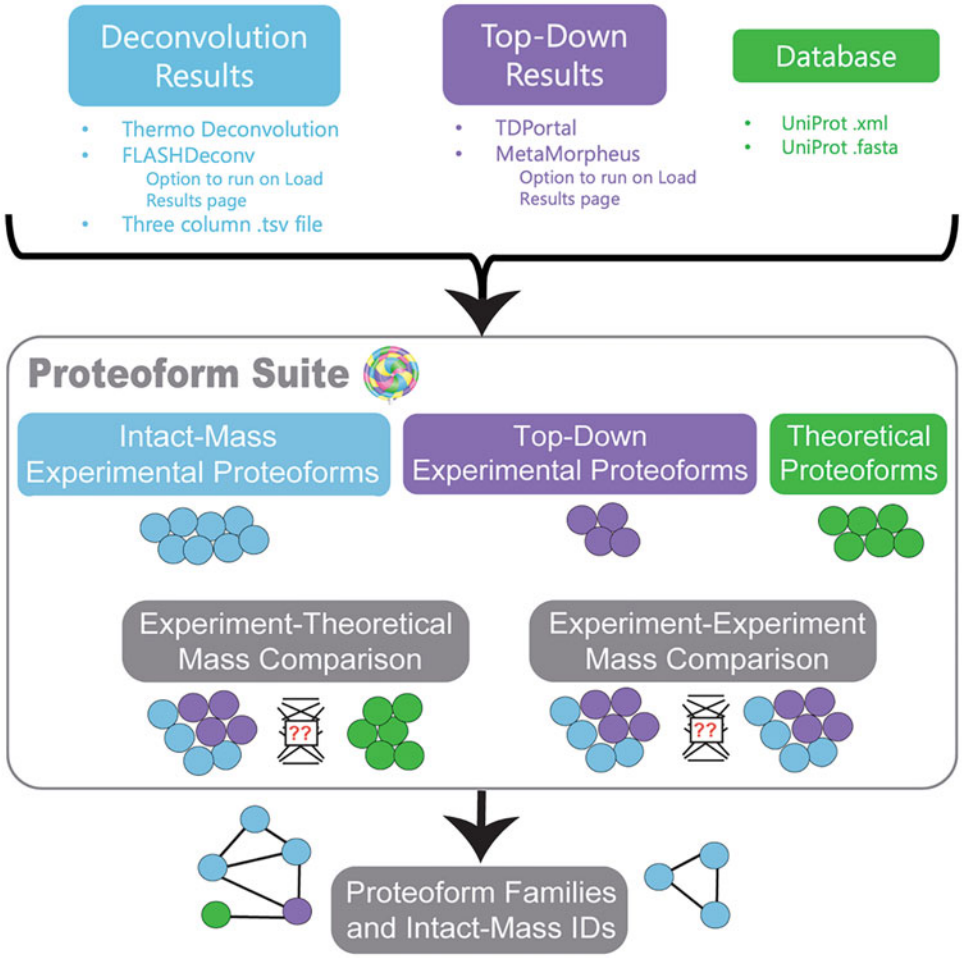


Fig. 1 Overview of inputs and outputs for Proteoform Suite analysis

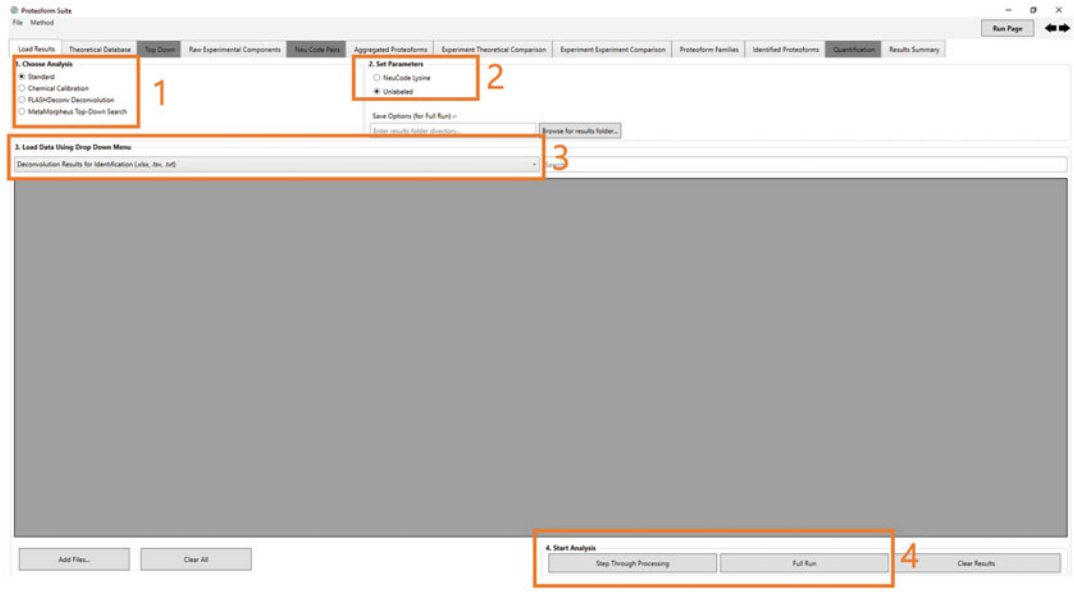


Fig. 2 Load Results page in Proteoform Suite

2.1 Deconvolution Results

Deconvolution results input into Proteoform Suite provide a list of observed experimental proteoform masses that are used for proteoform family construction and intact-mass identification of proteoforms (*see* **Note 1**). On the Load Results page under Standard analysis, set the drop-down menu (labeled 3 in Fig. 2) to Deconvolution Results for Identification and add deconvolution results. There are three options to produce deconvolution results for Proteoform Suite:

1. Thermo Deconvolution 4.0 (see Thermo Fisher website for a quote and user guide). Run the Xtract algorithm for high-resolution data, then open the result for each .raw file and export the results table. Save each file as an .xlsx in Microsoft Excel. The resulting .xlsx file is loaded under Deconvolution Results for Identification.
2. FLASHDeconv. FLASHDeconv is an ultrafast deconvolution algorithm developed by the OpenMS team for high-resolution mass spectrometry data [12]. On the Load Results page of Proteoform Suite, select FLASHDeconv Deconvolution under Choose Analysis (labeled 1 in Fig. 2). Input .mzML spectra files into the table. Set the desired parameters under Set Parameters (labeled 2 in Fig. 2). Click the Deconvolute button (labeled 4 in Fig. 2). For more advanced parameter options, a user can also externally run the command-line version of FLASHDeconv, available at https://www.openms.de/comp/FLASHDeconv/. The resulting .tsv file is loaded under Deconvolution Results for Identification.

3. Results from any external deconvolution software can also be input. Create a three-column tab-separated .tsv or .txt file: monoisotopic mass, intensity, retention time. The resulting .tsv or .txt file is loaded under Deconvolution Results for Identification.

For proteoform quantification, there is the option to label the Biological Replicate, Fraction, Technical Replicate, and Condition for each file. To change one of these labels for a single file, click the appropriate cell in the table. To change the label for more than one file or cell, select the cells, right click, enter a label, click Okay.

2.2 Top-Down Results

Top-down results input into Proteoform Suite provide a list of proteoforms identified by MS/MS analysis. These results are useful for improving intact-mass analysis by identifying additional proteoform families and proteoforms that might not be identified by intact-mass alone. On the Load Results page under Standard analysis, set the drop-down menu (labeled 3 in Fig. 2) to Top-Down Hit Results and add top-down results. There are two options to produce top-down results for Proteoform Suite:

1. TDPortal. TDPortal is a high-throughput global proteome analysis software for top-down data [10] available through the National Resource for Translational and Developmental Proteomics. Request access to TDPortal at http://nrtdp.northwestern.edu/tdportal-request/. Under the Reports tab in TDViewer, export the Hit Report. The resulting .xlsx file is loaded under Top-Down Hit Results.
2. MetaMorpheus. MetaMorpheus is an MS/MS search software program for both bottom-up and top-down high-resolution MS data. On the Load Results page of Proteoform Suite, select MetaMorpheus Top-Down Search under Choose Analysis (labeled 1 in Fig. 2). Set the desired parameters under Set Parameters (labeled 2 in Fig. 2). Set the drop-down menu (labeled 3 in Fig. 2) to Spectra Files and add .raw or .mzML files. Set the drop-down menu to Protein Databases and add an .xml or .fasta database. Click the MetaMorpheus Top-Down Search button (labeled 4 in Fig. 2). For more advanced parameter options, externally run the GUI or command line versions of MetaMorpheus available at https://github.com/smith-chem-wisc/metamorpheus. The resulting AllPSMs.tsv file is loaded under Top-Down Hit Results.

2.3 Database

Download a protein database from UniProt (https://www.uniprot.org/proteomes/). Typically, only reviewed entries are employed but unreviewed entries can be included at the user's discretion. A database from MetaMorpheus generated from a bottom-up search and the global post-translational modification discovery strategy

(G-PTM-D) can also be utilized, as previously described [13–15]. On the Load Results page under Standard analysis, set the drop-down menu (labeled 3 in Fig. 2) to Protein Databases and add at least one database. There are two options for protein databases:

1. .xml or .xml.gz: contains annotated PTM information and subsequences.
2. .fasta: optionally includes protein isoforms.

2.4 Bottom-Up Results

Bottom-up results input into Proteoform Suite provide a list of peptides identified by MS/MS analysis. To run a bottom-up search, download a release of MetaMorpheus and run a bottom-up search (https://github.com/smith-chem-wisc/metamorpheus). On the Load Results page under Standard analysis, set the drop-down menu (labeled 3 in Fig. 2) to MetaMorpheus Bottom-Up Unique Peptides and add the AllPeptides.psmtsv generated by the MetaMorpheus bottom-up search.

2.5 Mass and Retention Time Calibration

Mass and retention time calibration is an optional preprocessing step that is recommended when performing intact-mass analysis with deconvolution results [7]. Mass and retention time across runs can be calibrated for deconvolution and top-down results to improve mass accuracy and correct run-to-run variation, respectively. Calibrated files are output in the same file location as the input files; the calibrated results files can then be loaded on the Load Results page under Deconvolution Results for Identification and Top-Down Hit Results in the Standard analysis. For more details on how to perform calibration in Proteoform Suite, please see the user manual in Proteoform Suite release version 0.4.0.

3 Data Analysis

Proteoform Suite is an interactive software program for intact-mass analysis and proteoform family construction; each page is run in sequence, from left to right. To navigate to the next page, either click the right arrow in the top right of the GUI window or click the page tab name at the top of the GUI window. The user manual in release version 0.4.0 contains detailed information about each parameter and table column. The default parameters are set such that they will provide an acceptable starting point for most standard analyses.

1. Load Results. On this page, the user loads results files for analysis, as described in the Subheading 2.
   (a) Select Standard under Choose Analysis (labeled 1 in Fig. 2).

(b) Under Load Data Using Drop Down Menu (labeled 3 in Fig. 2), use the drop-down menu to select the result type to be loaded in. Then use the Add button or drag-and-drop to add files of the selected result type to the table.

(c) Navigate to the Theoretical Proteoform Database page by clicking that page tab at the top of the GUI window or by clicking the right arrow at the top right of the GUI window.

2. Theoretical Proteoform Database. On this page, theoretical proteoforms are created using the file(s) input under Protein Databases on the Load Results page. The theoretical proteoform database includes theoretical proteoforms with combinations of annotated PTMs and subsequences. Theoretical proteoforms are used in the subsequent pages for identifying proteoforms by intact-mass analysis.

(a) Set the parameters as necessary. Parameters to note are the Average Mass checkbox (check if deconvolution results are reported as average masses instead of monoisotopic masses), the Carbamidomethylation checkbox (check if sample was carbamidomethylated), and the Max Modifications per Proteoform (increase or decrease to generate theoretical proteoforms with combinations of more or fewer annotated PTMs). Higher values for Max Modifications per Proteoform will include more theoretical proteoforms in the database for potential identification (see Experiment-Theoretical comparison) but may also result in an increased false discovery rate for intact-mass identifications (*see* **Note 2**).

(b) Click Run Page button at the top right of the GUI window.

(c) The main table will fill with a list of the theoretical proteoforms.

3. Top-Down. On this page, top-down results are read from the file(s) input under Top-Down Hit Results on the Load Results page. A top-down hit is a proteoform spectral match. Top-down hits are loaded, and then hits are aggregated into unique top-down proteoforms using both identification information and retention time. After aggregation, the theoretical database is supplemented with top-down proteoform identifications not already present in the database (i.e., those with unannotated PTMs, unannotated subsequences, amino acid variations, or with a greater number of PTMs than the Max Modifications per Proteoform number on the Theoretical Proteoform Database page).

(a) Set the parameters as necessary. Parameters to note are the Min. C-Score (the minimum proteoform characterization

score [16] for TDPortal results) and the Ret. Time Tolerance (retention time used for merging top-down hits of the same proteoform identification). Higher values for the minimum C-score will exclude less characterized and potentially false top-down hits but may also exclude uncharacterized but true top-down identifications.

(b) Click Run Page button at the top right of the GUI window.

(c) The main table will fill with a list of the top-down proteoforms. Click a top-down proteoform to fill the bottom box with the sequence and PTM information. If bottom-up results were provided on the Load Results page, the second table will fill with bottom-up peptides that correspond to the selected top-down proteoform.

4. Raw Experimental Components. On this page, raw experimental components are read from the file(s) loaded under Deconvolution Results for Identification on the Load Results page. A raw experimental component is an individual deconvoluted proteoform observation at the MS1 level, as reported in the deconvolution result file(s). Deconvolution artifacts (including missed monoisotopic mass errors and charge state harmonics) are corrected using the mass tolerance set on the page.

   (a) Set the parameters as necessary. Parameters to note are the cosine threshold between per-charge-intensities and fitted Gaussian distribution and the cosine threshold between averagine and observed isotope pattern [12]. These parameters determine the thresholds for reported FLASHDeconv results to be included as raw experimental components. Higher values will exclude more raw experimental components that correspond to experimental artifacts but may also exclude raw experimental components that correspond to real observed experimental proteoforms.

   (b) Click the Run Page button at the top right of the GUI window.

   (c) The main table will fill with a list of raw experimental components. Click a raw experimental component to fill the bottom table with the observed charge state information.

5. Aggregated Proteoforms. On this page, experimental proteoforms are created by aggregating raw experimental components. These experimental proteoforms are analyzed and identified by intact-mass analysis and used to construct proteoform families. The Add Top-Down Proteoforms checkbox provides the option to supplement this list with top-down identified experimental proteoforms from the Top-Down page.

(a) Set the parameters as necessary. Parameters to note are the Mass Tolerance and Ret. Time Tolerance, which are used to determine the mass and retention time tolerances used to merge raw experimental components into an experimental proteoform. Additionally, the Minimum Required Observations setting can be used to determine the minimum number of biological and/or technical replicates an experimental proteoform must have been observed in (determined by the raw experimental component files input and optionally labeled on the Load Results page). Higher values will exclude more raw experimental components that correspond to experimental artifacts but may also exclude raw experimental components that correspond to real observed experimental proteoforms.

(b) Click the Run Page button at the top right of the GUI window.

(c) The main table will fill with a list of aggregated experimental proteoforms. Click an experimental proteoform to fill the bottom table with raw experimental components aggregated into the selected proteoform. If the selected proteoform is a top-down proteoform, the bottom table will fill with the top-down hits aggregated into the selected proteoform.

6. Experiment–Theoretical Comparison. On this page, masses of experimental proteoforms created on the Aggregated Proteoforms page are compared to masses of theoretical proteoforms created on the Theoretical Proteoforms page, generating a list of experiment–theoretical pairs. Each experiment–theoretical pair has a mass difference (delta mass) between the experimental proteoform and the theoretical proteoform in the pair; pairs are generated for mass differences corresponding to a known single modification or set of modifications. Proteoform Suite creates a histogram of the mass differences for all generated experiment–theoretical pairs and a corresponding list of delta mass peaks observed in the histogram. Each experiment–theoretical pair with a delta mass in a delta mass peak accepted by the user is utilized to construct proteoform families (see Proteoform Families below).

   (a) Set the parameters as necessary. There are several parameters of note on this page. The Peak Width Base determines the width of peaks when generating the delta mass histogram; a larger value will enable more experiment–theoretical pairs to be included in each peak but will also increase the false discovery rate of each peak. The checkbox Notch Search provides the option to perform a notch search [17]. If checked, a tolerance parameter will become

visible, and experiment–theoretical pairs will only be generated if within tolerance from an accepted notch (corresponding to an exact match at 0 Da, a known PTM, a known PTM combination or an amino acid mass). The larger the tolerance, the more the experiment–theoretical pairs that will be included at each notch, and also potentially larger the false discovery rate will be. The checkbox Add Top-Down Theoretical determines if theoretical proteoforms that were added to the database on the Top-Down page (not present in the original database) are included in the analysis.

(b) Click the Run Page button at the top right of the GUI window.

(c) The top right table will fill with all experiment–theoretical pairs. The bottom graph will display the delta mass histogram created from the experiment–theoretical pairs. The top left table will fill with the list of delta mass peaks from the histogram.

(d) Browse the list of delta mass peaks (top left table). Accept peaks that have an acceptable false discovery rate and that correspond to common/likely modifications or amino acid differences (*see* **Note 3**). The Peak Threshold determines the minimum number of experiment–theoretical pairs that must be grouped in a delta mass peak in order for it to be accepted. This value can be changed to accept/unaccept peaks, or individual peaks can be accepted/unaccepted by checking/unchecking the accepted checkbox for the peak. Typically, only the delta mass peak closest to 0 Da (exact match experiment–theoretical pairs) is accepted. A higher number of accepted peaks results in more proteoform family connections and intact-mass identifications but also a larger false discovery rate.

7. Experiment–Experiment Comparisons. On this page, masses of experimental proteoforms created on the Aggregated Proteoforms page are compared to one another within a set retention time difference tolerance, generating a list of experiment–experiment pairs. Each experiment–experiment pair has a mass difference (delta mass) between the two experimental proteoforms in the pair. Proteoform Suite creates a histogram of the mass differences for all generated experiment–experiment pairs and a corresponding list of delta mass peaks observed in the histogram. Each experiment–experiment pair with a delta mass in a delta mass peak accepted by the user is utilized to construct proteoform families (*see* Proteoform Families below).

(a) Set the parameters as necessary. There are several parameters of note on this page. The Peak Width Base

determines the width of peaks when generating the delta mass histogram; a larger value will enable more experiment–experiment pairs to be included in each peak but will also increase the false discovery rate of each peak. The checkbox Notch Search provides the option to perform a notch search [17]. If checked, a tolerance parameter will become visible, and experiment–experiment pairs will only be generated if within tolerance from an accepted notch (corresponding to a known PTM or a known PTM set). Larger tolerance values result in more experiment–experiment pairs included at each notch, but also increase the false discovery rate. The Max Retention Time Difference determines the maximum allowed difference in retention time between two experimental proteoforms to be eligible to form an experiment–experiment pair. Larger values result in more experiment–experiment pairs, but also increase the false discovery rate.

(b) Click the Run Page button at the top right of the GUI window.

(c) The top right table will fill with all experiment–experiment pairs. The bottom graph will display the delta mass histogram created from the experiment–experiment pairs. The top left table will fill with the list of delta mass peaks from the histogram.

(d) Browse the list of delta mass peaks (top left table). Accept peaks that have an acceptable false discovery rate and that correspond to common/likely modifications or amino acid differences (*see* **Note 3**). The Peak Threshold determines the minimum number of experiment–experiment pairs that must be grouped in a delta mass peak in order for it to be accepted. This value can be changed to accept/unaccept peaks, or individual peaks can be accepted/unaccepted by checking/unchecking the accepted checkbox for the peak. Typically, only the most abundant five to ten delta mass peaks corresponding to known and common PTMs and amino acid differences are accepted. A higher number of accepted peaks results in more proteoform family connections and intact-mass identifications and can also result in a larger false discovery rate.

8. Proteoform Families. On this page, accepted experiment–theoretical and experiment–experiment pairs are used to construct proteoform families. Experimental proteoforms are identified by intact-mass analysis; beginning with each theoretical proteoform in each family, connections between proteoforms are used to identify proteoforms first from experiment–theoretical pairs and then from subsequent experiment–experiment pairs. A false discovery rate is calculated for intact-mass identifications

(*see* **Note 4**). A script for Cytoscape [8, 9] can be exported to visualize proteoform families as a network of nodes (proteoforms) and edges (mass differences between proteoforms corresponding to modifications or amino acid differences).

(a) Set the parameters as necessary. There are several parameters to note on this page. If the checkbox Only Assign Common/Known Mods is checked, a modification will only be considered for each proteoform identification if annotated for that protein in the theoretical database or if it is a common modification (e.g., acetylation, methylation, phosphorylation). The Use ppm tolerance checkbox can be checked to require a proteoform identification to be within a certain ppm mass error tolerance.

(b) Click the Run Page button at the top right of the GUI window.

(c) The main table will fill with a list of proteoform families. Click a row in the table and the bottom table will fill with either a list of experimental proteoforms, theoretical proteoforms, or proteoform relations depending on which column is selected.

(d) To build proteoform families on this page, click the Browse button on the right and select a file path. Set the visualization parameters as desired (top right of the screen). Select either Build All Families in Cytoscape (writes a script to visualize all proteoform families) or Build Selected Families in Cytoscape (writes a script to visualize proteoform families selected in the main table).

9. Results

(a) Identified Proteoforms. Navigate to the Identified Proteoforms page to view all identified experimental proteoforms. This page displays two tables with intact-mass identifications (from deconvolution results and proteoform family construction) and top-down identifications (from input top-down results). If bottom-up results were provided on the Load Results page, click either an intact-mass or a top-down proteoform to fill the bottom table with bottom-up peptides that correspond to the selected proteoform identification.

(b) Results Summary. Navigate to the Results Summary page to view a summary of results from Proteoform Suite. Select a file path for the Results Folder and click the Save All button. A summary of the results text file, a method .xml file for subsequent runs, Cytoscape scripts, and various results tables will export. See the user manual in Proteoform Suite release 0.4.0 for a list and description of exported results.

(c) Visualize Proteoform Families. Scripts for Cytoscape are exported on either the Proteoform Families page or the Results Summary page. To visualize proteoform families in Cytoscape, install Cytoscape version 3.5.0 at https://cytoscape.org. In Cytoscape, select Execute Command File under Tools, then select the cytoscape_script_timestamp.tsv file output by Proteoform Suite.

4 Notes

1. Intact-mass analysis is highly dependent on the quality of the results of deconvolution. Missed monoisotopic mass errors, where the reported mass is one or more isotopic units from the monoisotopic mass, are a challenging issue for intact-mass analysis. Additionally, artifactual masses reported by deconvolution also increase the false discovery rate. The user should do their best to ensure high-quality MS1 data and deconvolution results.
2. It is important for the user to consider the database size when performing an intact-mass analysis. There is a tradeoff between expanding the database to include more PTM combinations and amino acid variants and limiting increases to the FDR. Typically, our studies have used two or three PTM combinations per theoretical proteoform; this is because proteoforms with more modifications are able to be identified via the experiment–experiment comparisons if a less modified proteoform is also identified from the same proteoform family. Proteoforms that only exist in ultra-modified forms are thus best identified by top-down MS/MS analysis.
3. When accepting and unaccepting peaks in the experiment–theoretical and experiment–experiment comparisons, we have found it is best to be conservative and accept the most abundant peaks closest to the delta mass of known, common, and expected post-translational modifications. If a large number of high-FDR and high mass error experiment–theoretical and experiment–experiment peaks are employed, the effect will be an increase in FDR and proteoform families may be contaminated with incorrect proteoform members. At this time, less common modifications are better identified either through top-down MS/MS analysis or by using bottom-up analysis to add the modification to the protein sequence, enabling it to be included in the theoretical proteoform database and thus identifiable through an exact match in the experiment–theoretical comparison.
4. In intact-mass analysis with Proteoform Suite, a remaining challenge is identifying previously unidentified, observed

proteoforms by intact mass alone without exceeding a reasonable false discovery rate. Proteoform Suite calculates a global false discovery rate by performing decoy experiment–theoretical and experiment–experiment comparisons to construct decoy proteoform families, described in detail previously [6, 7]. This globally calculated false discovery rate is dependent on several aspects of the analysis, including the size of the theoretical database, the deconvolution results, the top-down results, and the settings utilized when performing the experiment–theoretical and experiment–experiment comparisons.

Acknowledgments

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health (NIH) under Award Number R35GM126914. LVS was supported by the NIH Biotechnology Training Program, T32GM008349.

References

1. Smith LM, Kelleher NL, Consortium for Top Down Proteomics (2013) Proteoform: a single term describing protein complexity. Nat Methods 10(3):186–187. https://doi.org/10.1038/nmeth.2369
2. Shortreed MR, Frey BL, Scalf M, Knoener RA, Cesnik AJ, Smith LM (2016) Elucidating proteoform families from proteoform intact-mass and lysine-count measurements. J Proteome Res 15(4):1213–1221. https://doi.org/10.1021/acs.jproteome.5b01090
3. Cai W, Tucholski TM, Gregorich ZR, Ge Y (2016) Top-down proteomics: technology advancements and applications to heart diseases. Expert Rev Proteomics 13(8):717–730. https://doi.org/10.1080/14789450.2016.1209414
4. Chen B, Brown KA, Lin Z, Ge Y (2018) Top-down proteomics: ready for prime time? Anal Chem 90(1):110–127. https://doi.org/10.1021/acs.analchem.7b04747
5. Toby TK, Fornelli L, Kelleher NL (2016) Progress in top-down proteomics and the analysis of proteoforms. Annu Rev Anal Chem (Palo Alto, Calif) 9(1):499–519. https://doi.org/10.1146/annurev-anchem-071015-041550
6. Cesnik AJ, Shortreed MR, Schaffer LV, Knoener RA, Frey BL, Scalf M, Solntsev SK, Dai Y, Gasch AP, Smith LM (2018) Proteoform Suite: software for constructing, quantifying, and visualizing proteoform families. J Proteome Res 17(1):568–578. https://doi.org/10.1021/acs.jproteome.7b00685
7. Schaffer LV, Shortreed MR, Cesnik AJ, Frey BL, Solntsev SK, Scalf M, Smith LM (2018) Expanding proteoform identifications in top-down proteomic analyses by constructing proteoform families. Anal Chem 90(2):1325–1333. https://doi.org/10.1021/acs.analchem.7b04221
8. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504. https://doi.org/10.1101/gr.1239303
9. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27(3):431–432. https://doi.org/10.1093/bioinformatics/btq675
10. Toby TK, Fornelli L, Kristina S, DeHart CJ, Fellers RT (2019) A comprehensive pipeline for translational top-down proteomics from a single blood draw. Nat Protoc 14:119–152. https://doi.org/10.1038/s41596-018-0085-7
11. Donnelly DP, Rawlins CM, DeHart CJ, Fornelli L, Schachner LF, Lin Z, Lippens JL, Aluri KC, Sarin R, Chen B, Lantz C, Jung W, Johnson KR, Koller A, Wolff JJ, Campuzano IDG, Auclair JR, Ivanov AR, Whitelegge JP, Pasa-Tolic L, Chamot-Rooke J, Danis PO,

Smith LM, Tsybin YO, Loo JA, Ge Y, Kelleher NL, Agar JN (2019) Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. Nat Methods 16(7): 587–594. https://doi.org/10.1038/s41592-019-0457-0

12. Jeong K, Kim J, Gaikwad M, Hidayah SN, Heikaus L, Schluter H, Kohlbacher O (2020) FLASHDeconv: ultrafast, high-quality feature deconvolution for top-down proteomics. Cell Syst 10(2):213–218 e216. https://doi.org/10.1016/j.cels.2020.01.003
13. Dai Y, Shortreed MR, Scalf M, Frey BL, Cesnik AJ, Solntsev S, Schaffer LV, Smith LM (2017) Elucidating Escherichia coli proteoform families using intact-mass proteomics and a global PTM discovery database. J Proteome Res 16(11):4156–4165. https://doi.org/10.1021/acs.jproteome.7b00516
14. Dai Y, Buxton KE, Schaffer LV, Miller RM, Millikin RJ, Scalf M, Frey BL, Shortreed MR, Smith LM (2019) Constructing human proteoform families using intact-mass and top-down proteomics with a multi-protease global post-translational modification discovery database. J Proteome Res 18(10): 3671–3680. https://doi.org/10.1021/acs.jproteome.9b00339
15. Li Q, Shortreed MR, Wenger CD, Frey BL, Schaffer LV, Scalf M, Smith LM (2017) Global post-translational modification discovery. J Proteome Res 16(4):1383–1390. https://doi.org/10.1021/acs.jproteome.6b00034
16. LeDuc RD, Fellers RT, Early BP, Greer JB, Thomas PM, Kelleher NL (2014) The C-score: a Bayesian framework to sharply improve proteoform scoring in high-throughput top down proteomics. J Proteome Res 13(7):3231–3240. https://doi.org/10.1021/pr401277r
17. Solntsev SK, Shortreed MR, Frey BL, Smith LM (2018) Enhanced global post-translational modification discovery with MetaMorpheus. J Proteome Res 17(5):1844–1851. https://doi.org/10.1021/acs.jproteome.7b00873

Chapter 8

Top-Down Mass Spectrometry Data Analysis Using TopPIC Suite

In Kwon Choi and Xiaowen Liu

Abstract

With the advances of mass spectrometry (MS) techniques, top-down MS-based proteomics has gained increasing attention because of its advantages over bottom-up MS in studying complex proteoforms. TopPIC Suite is a widely used software package for top-down MS-based proteoform identification and quantification. Here, we present the methods for top-down MS data analysis using TopPIC Suite.

Key words Top-down mass spectrometry, Proteomics, Spectral deconvolution, Proteoform identification, Proteoform quantification

1 Introduction

Top-down mass spectrometry (MS)-based proteomics studies intact proteoforms in biological samples [1, 2]. Compared with bottom-up MS, which analyzes short peptides, top-down MS enables the analysis of whole proteoforms and the combinatorial patterns of various alterations on proteoforms [3]. In recent years, rapid developments of MS techniques and the availability of high-resolution mass spectrometers facilitated the applications of top-down MS in many research areas, such as proteoform biomarker discovery [4].

It is a challenging computational problem to analyze top-down MS data [5]. The first challenge is the complexity of top-down mass spectra [6]. Because a proteoform has various isotopomers, a top-down mass spectrum contains a group of isotopic peaks for a proteoform, increasing the number of peaks. High-charge state ions are used to measure proteoforms with large molecular masses, increasing the complexity of mass spectra. As a result, a top-down mass spectrum may contain tens of thousands of peaks. Another challenge is that intact proteoforms often have multiple alterations, including sequence mutations and posttranslational modifications

Liangliang Sun and Xiaowen Liu (eds.), *Proteoform Identification: Methods and Protocols*,
Methods in Molecular Biology, vol. 2500, https://doi.org/10.1007/978-1-0716-2325-1_8,
© The Author(s), under exclusive license to Springer Science+Business Media, LLC, part of Springer Nature 2022

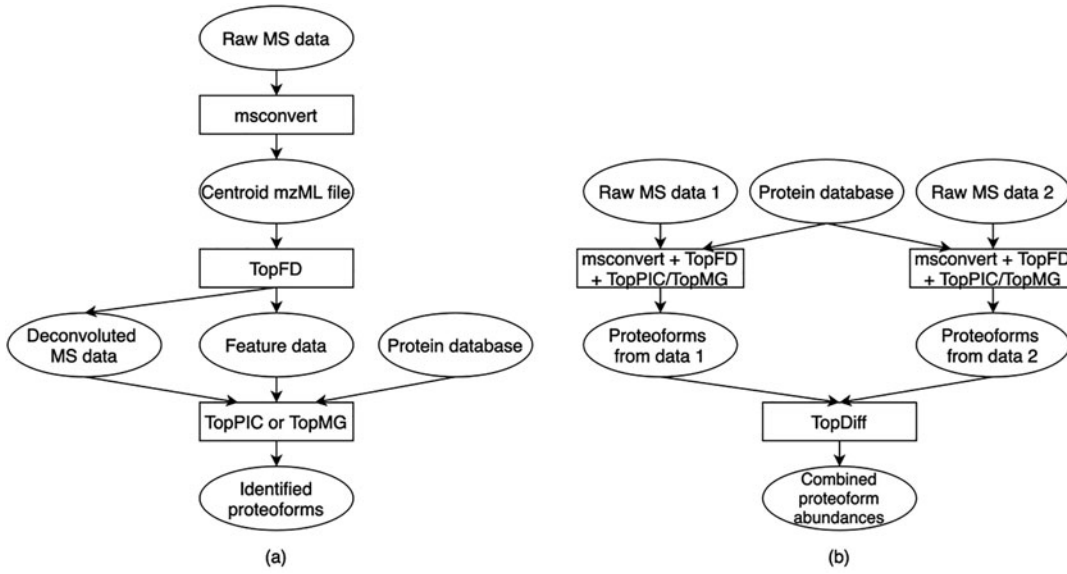


Fig. 1 Data analysis steps for top-down MS-based proteoform identification and quantification using TopPIC Suite. (**a**) First, a raw mass spectrum data file is converted into an mzML file using msconvert in ProteoWizard (Subheading 3.1). Second, TopFD is used for data processing that generates a list of LC-MS features and converts each MS/MS spectrum into a list of deconvoluted monoisotopic masses (Subheading 3.2). Third, TopPIC or TopMG is employed to identify proteoforms by searching deconvoluted mass spectra against a protein sequence database (Subheadings 3.3 and 3.4). TopMG is capable of identifying complex proteoforms with multiple variable PTMs. (**b**) To compare proteoform abundances in top-down MS data sets of multiple biological samples, msconvert, TopFD, and TopPIC/TopMG are used to identify proteoforms from each data set separately. Then TopDiff is used to combine and compare the abundances of proteoforms identified in multiple biological samples (Subheading 3.5)

(PTMs). Database search is the method of choice for proteoform identification by top-down MS. But these alterations are often unexpected in database search, so open search [3] or spectral alignment methods [7] are needed for identifying proteoforms with alterations.

TopPIC Suite is a commonly used software package for proteoform identification and quantification by top-down MS (Fig. 1). It consists of four software tools: TopFD, TopPIC [8], TopMG [9], and TopDiff. TopFD is a tool for top-down spectral deconvolution, which converts complex isotopic patterns of proteoforms or proteoform fragments into their monoisotopic masses. TopPIC and TopMG identify modified proteoforms using spectral alignment and graph alignment between mass spectra and database protein sequences. TopDiff compares the abundances of proteoforms identified from multiple samples. Here we guide you through the process of top-down MS data analysis using TopPIC Suite.

2 Material

A desktop computer with a Windows 10 operating system is needed for top-down MS data analysis. In addition, ProteoWizard [10] and TopPIC Suite will be installed on the desktop, and the tools will be used to analyze two top-down tandem mass spectrometry (MS/MS) data files of *Escherichia coli*.

2.1 Installation of ProteoWizard

1. Download ProteoWizard from http://proteowizard.sourceforge.net/download.html.
2. Choose the platform "Windows 64-bit installer" for end-users.
3. Check the license agreement checkbox and click the "Download" button to download ProteoWizard.
4. Double click the downloaded file pwiz-setup-3.0-×86–64.msi to install it.

2.2 Installation of TopPIC Suite

1. Create a new folder named toppic_tutorial on the C: drive of your computer.
2. Create a new folder named toppic inside the toppic_tutorial folder. The resulting file structure is shown in the screenshot in Fig. 2.
3. Download TopPIC Suite from https://www.toppic.org/software/toppic/register.html.
4. Choose the download type "Windows 64-bit zip file" and fill out the registration form.

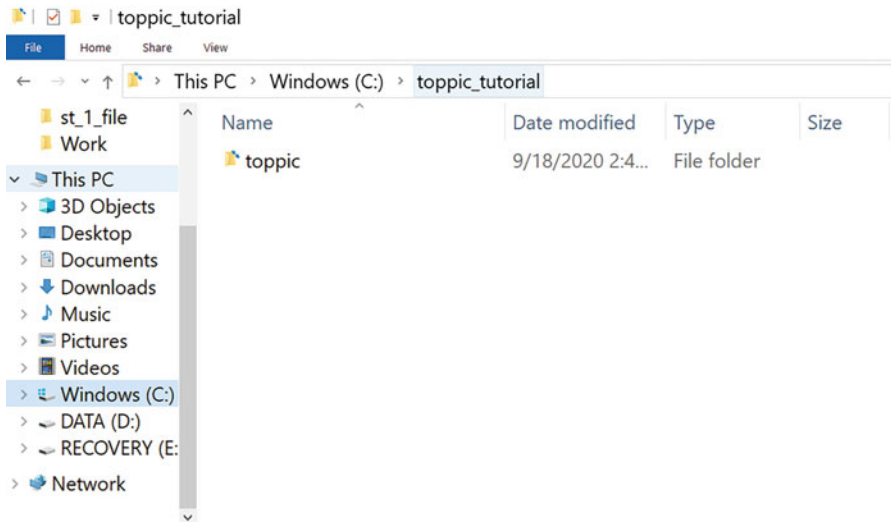


Fig. 2 The file folder for TopPIC Suite

5. Click "I accept the license agreement and download TopPIC suite" to download TopPIC Suite.
6. Unzip the toppic-suite zip file into the folder C:\toppic_tutorial\toppic\.

2.3 Preparation of Data Files

1. Download two top-down MS data files of *E. coli* from https://www.toppic.org/protocoldata/data-1-raw.zip, save it into the folder C:\toppic_tutorial, and unzip the file. (*See* **Note 1** for more information about the data set.)
2. Download the Swiss-Prot proteome database of *E. coli* from https://www.toppic.org/protocoldata/uniprot-ecoli.fasta and save it into the folder C:\toppic_tutorial. The protein database contains 4271 proteins.
3. Download an example file for variable PTMs from https://www.toppic.org/protocoldata/variable_mods.txt and save it in the folder C:\toppic_tutorial.

3 Methods

3.1 Convert Raw MS Files to mzML Files Using msconvert

1. Search "msconvert" in the search box on the taskbar of Windows 10 and run the desktop app "MSConvert."
2. In MSConvert window (Fig. 3), click the "Browse" button (Label A in Fig. 3) and add C:\toppic_tutorial\ECOLI_1.raw and C:\toppic_tutorial\ECOLI_2.raw as input files.
3. Click the "Add" button (Label B in Fig. 3) after the files are selected.
4. Select the filter "Peak Picking" from the dropdown list for filters (Label C in Fig. 3) and click the "Add" button (Label D in Fig. 3). "peakPeaking vendor msLevel=1-" will be displayed in the filter window (Label E in Fig. 3). The Peak Picking filter will convert profile mode MS data to centroid MS data. Remove any other filters in the filter window.
5. Click the "Start" button to start file format conversion. MSConvert outputs two mzML files containing centroided peaks.

3.2 Spectral Deconvolution Using TopFD

1. In the folder C:\toppic_tutorial\toppic, double click the executable file "topfd_gui.exe." (*See* **Note 2** for parameters of TopFD.)
2. Add the file C:\toppic_tutorial\ECOLI_1.mzML and C:\toppic_tutorial\ECOLI_2.mzML as input files (Label A in Fig. 4).

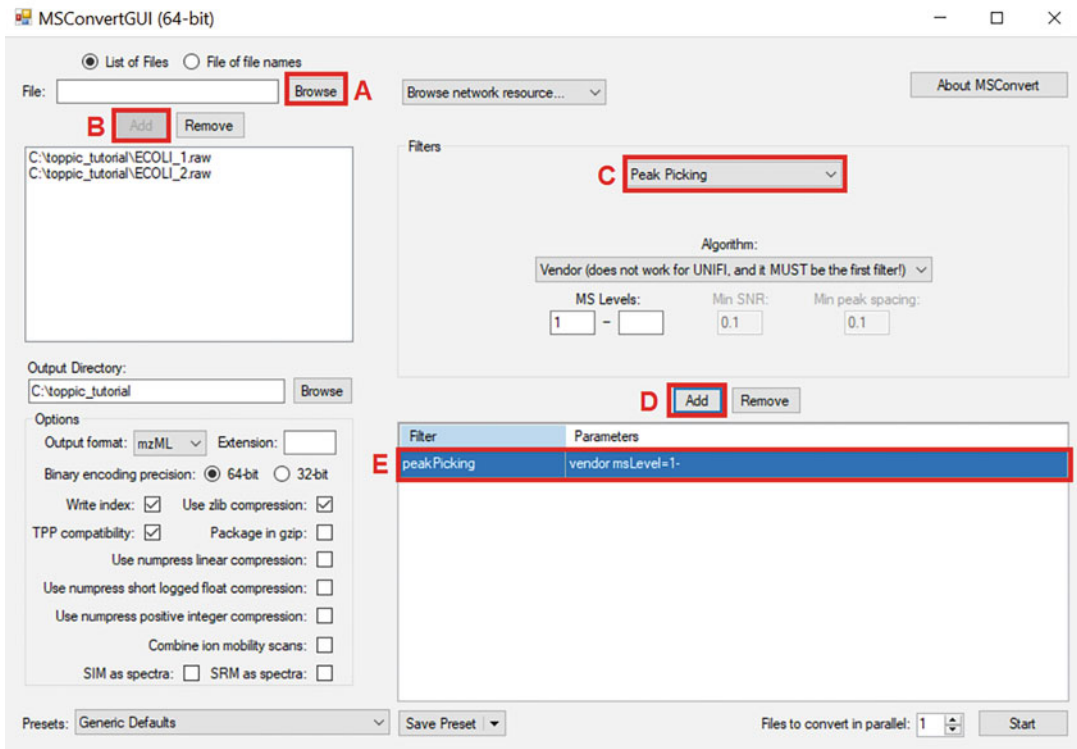


Fig. 3 The window of msconvert for MS data file format conversion

3. Click the button "Start" to deconvolute the two files (Label B in Fig. 4). TopFD outputs two LC-MS feature text files with a file extension "feature" and one LC-MS feature file with a file extension "xml." It also reports a deconvoluted mass spectral data file for MS/MS spectra in the msalign format with a file extension "msalign," which is similar to the MGF file format. TopFD creates a folder containing deconvoluted MS1 spectra in the msalign format and annotations of LC-MS data with a file extension "mzrt.csv," and another folder containing JavaScript files of MS1 and MS/MS data for spectral visualization.

3.3 Proteoform Identification Using TopPIC

1. Double click the executable file "toppic_gui.exe" in the folder C:\toppic_tutorial\toppic (Figs. 5, 6, and 7). (*See* **Note 3** for the parameters of TopPIC.)
2. Click "Select" and select C:\toppic_tutorial\uniprot-ecoli.fasta (Label A in Fig. 5).

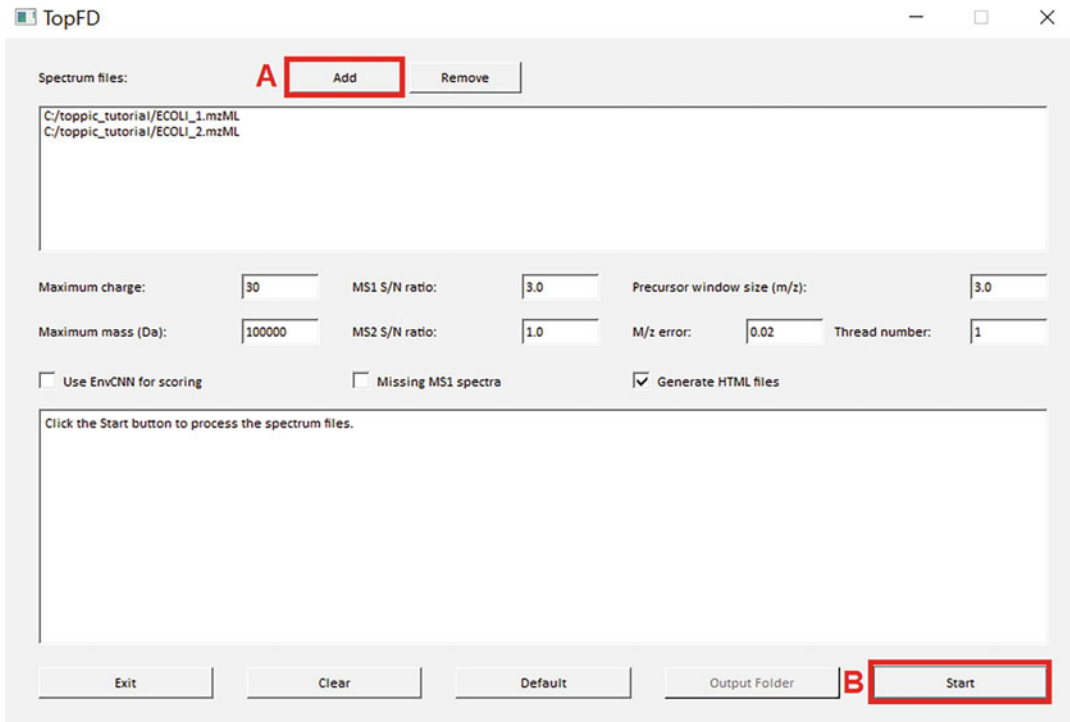


Fig. 4 The window of TopFD for top-down MS spectral deconvolution

3. Click "Add" and add C:\toppic_tutorial\ECOLI_1_ms2.msalign and C:\toppic_tutorial\ECOLI_2_ms2.msalign (Label B in Fig. 5).
4. Check the checkbox Decoy database (Label C in Fig. 6).
5. Select FDR as the spectrum-level cutoff type (Label D in Fig. 6).
6. Select FDR as the proteoform-level cutoff type (Label E in Fig. 6).
7. Click the button "Start" (Label F in Fig. 6). TopPIC outputs two tab-separated value (TSV) files containing identified proteoforms and proteoform spectrum-matches (PrSMs) with an E-value or false discovery rate (FDR) cutoff and an XML file containing identified proteoforms with the E-value or proteoform-level FDR cutoff. (*See* **Note 4** for a detailed description of the columns in the resulting TSV files.) It also reports a folder containing JavaScript files of identified PrSMs for visualization.

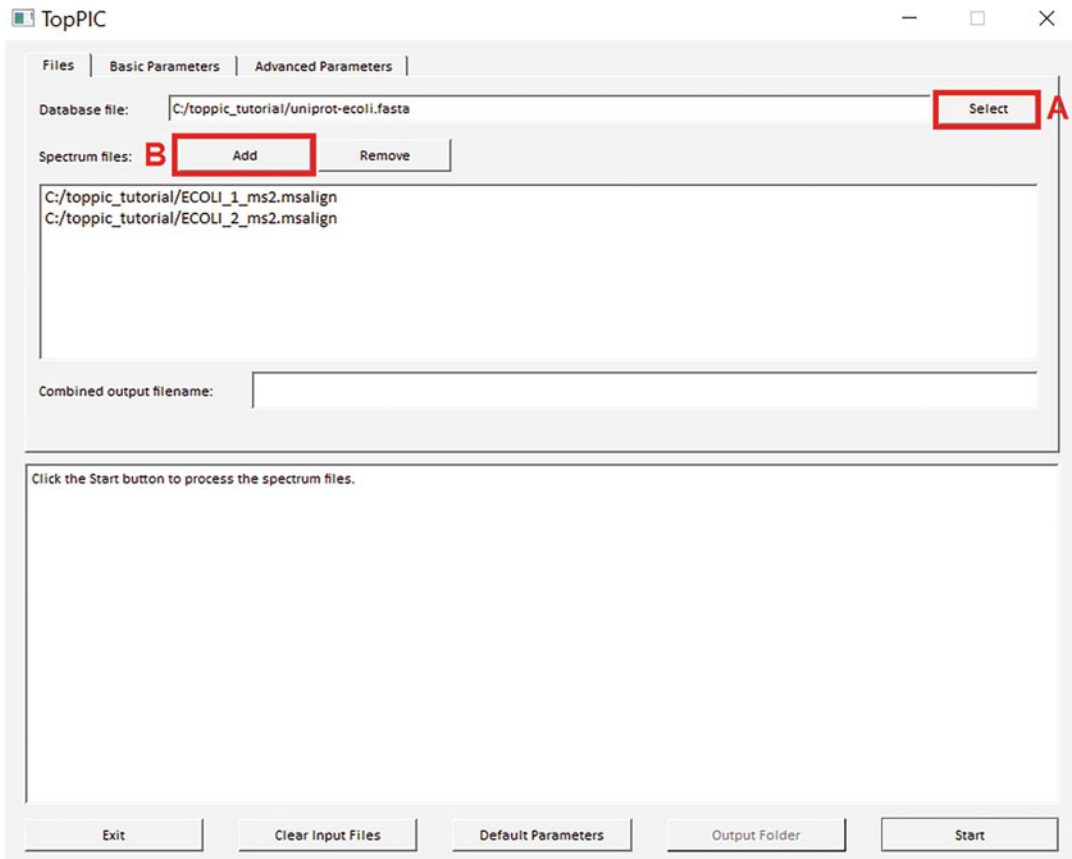


Fig. 5 The Files panel of TopPIC for proteoform identification by database search

3.4 Proteoform Identification Using TopMG

TopMG is used to identify highly modified proteoforms with expected PTMs. If the study is unrelated to highly modified proteoforms, the steps in this subsection can be skipped.

1. Double click the executable file "topmg_gui.exe" in the folder C:\toppic_tutorial\toppic (Figs. 8, 9, and 10). (*See* **Note 5** for parameters of TopMG.)
2. Click "Select" and select C:\toppic_tutorial\uniprot-ecoli.fasta (Label A in Fig. 8).
3. Click "Add" and add C:\toppic_tutorial\ECOLI_2_ms2.msalign (Label B in Fig. 8).
4. Click "Select" to add C:\toppic_tutorial\variable_mods.txt (Label C in Fig. 8).
5. Check the checkbox Decoy database (Label D in Fig. 9).
6. Select FDR as the spectrum-level cutoff type (Label E in Fig. 9).
7. Set the spectrum-level FDR cutoff to 0.05 (Label F in Fig. 9).

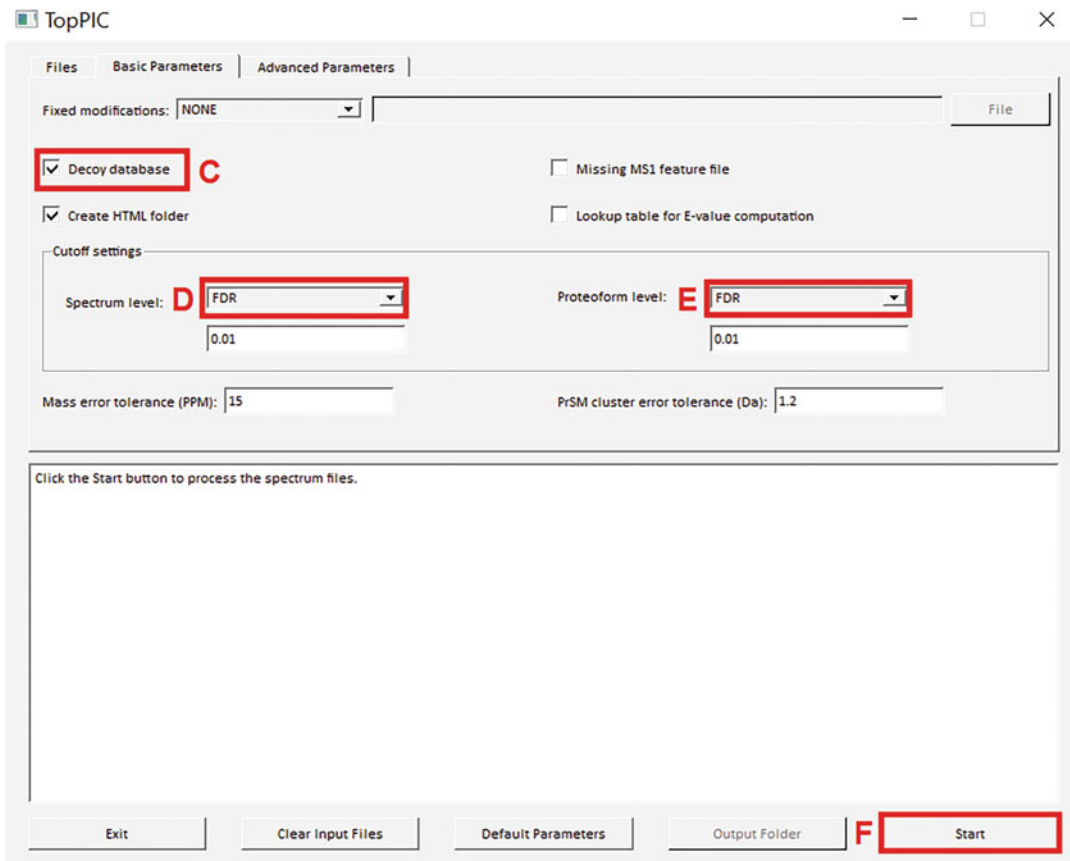


Fig. 6 The Basic Parameters panel of TopPIC for proteoform identification by database search

8. Select FDR as the proteoform-level cutoff type (Label G in Fig. 9).
9. Set the proteoform-level FDR cutoff to 0.05 (Label H in Fig. 9).
10. Click the button "Start" (Label I in Fig. 9). TopMG outputs two TSV files containing identified proteoforms and PrSMs with an E-value or FDR cutoff and an XML file containing identified proteoforms. (*See* **Note 4** for a detailed description of the columns in the resulting TSV files.) It also reports a folder containing JavaScript files of identified PrSMs for visualization.

3.5 Proteoform Abundance Comparison Using TopDiff

1. Double click the executable file "topdiff_gui.exe" in C:\toppic_tutorial\toppic (Fig. 11). (*See* **Note 6** for parameters of TopDiff.)
2. Click "Select" and select C:\toppic_tutorial\uniprot-ecoli.fasta (Label A in Fig. 11).

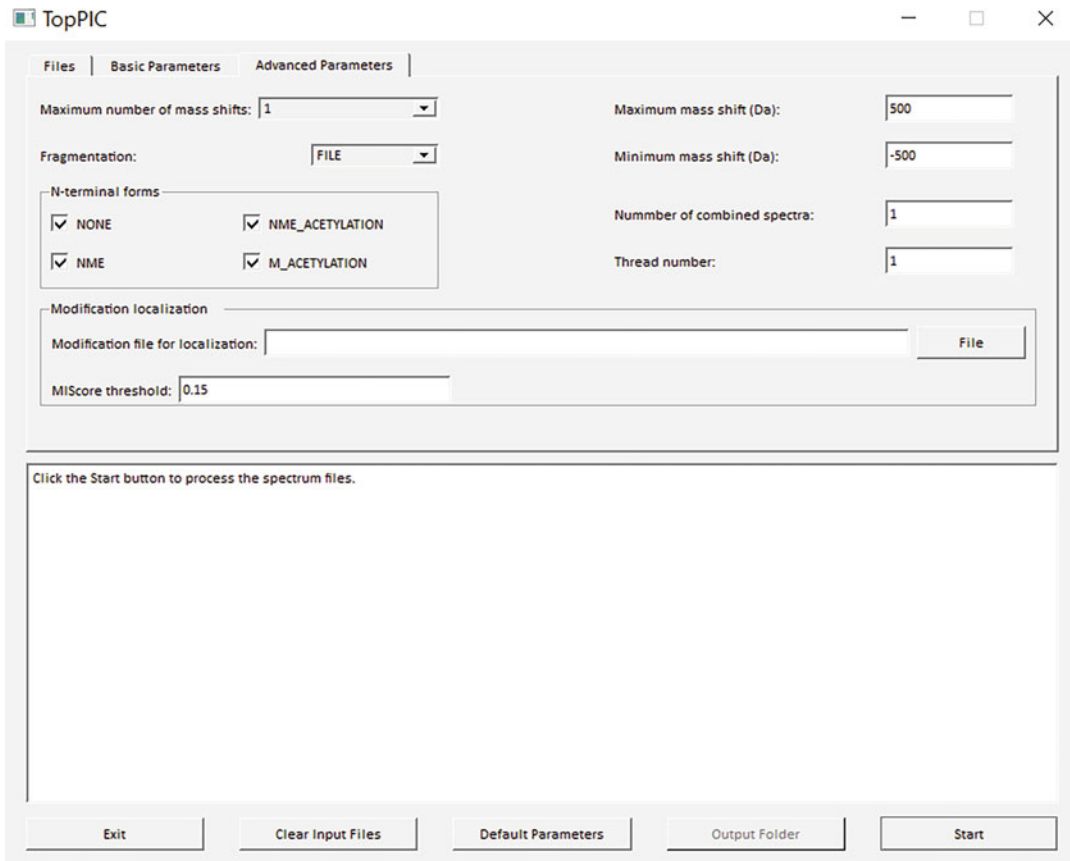


Fig. 7 The Advanced Parameters panel of TopPIC for proteoform identification by database search

3. Click "Add" and add C:\toppic_tutorial\ECOLI_1_ms2.msalign and C:\toppic_tutorial\ECOLI_2_ms2.msalign (Label B in Fig. 11).
4. Click the button "Start" (Label C in Fig. 11). TopDiff generates a TSV file containing proteoform identifications and their abundances in multiple mass spectra data files. (*See* **Note 7** for a detailed description of the columns in the resulting TSV file.)

3.6 Result Visualization Using TopMSV

(Attention: Currently, TopMSV only supports the Google Chrome browser. Please use the Google Chrome browser for the steps below.)

1. In C:\toppic_tutorial, click a folder "ECOLI_1.html" (if the user used different data from the example data, the folder name would be "input-mzML-file-name.html").
2. In ECOLI_1.html folder, click a folder "topmsv."
3. Click index.html to visualize PrSM identifications.

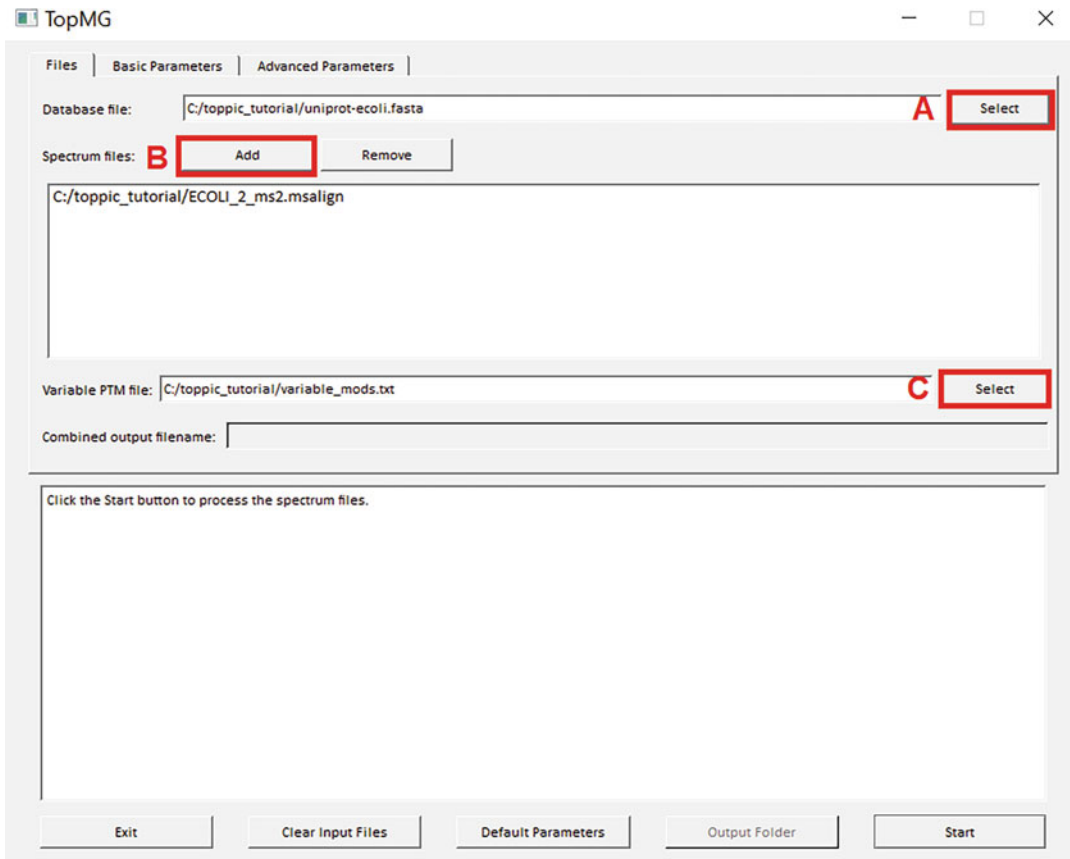


Fig. 8 The Files panel of TopMG for proteoform identification by database search

4 Troubleshooting

1. TopPIC, TopMG, or TopDiff quits after clicking the "Start" button.

   Possible reason: Required files are missing.

   Solution: The files ECOLI_1_ms2.feature and ECOLI_2_ms2.feature are required to be in the same folder as the spectrum files ECOLI_1_ms2.msalign and ECOLI_2_m2.msalign. If you have deleted or moved feature files from the folder after spectral deconvolution using TopFD, then rerun TopFD to generate the files.

2. In TopPIC and TopMG, the following error is reported: "LOG ERROR: TopFD feature file does not exist! Please use -x option in command line or select 'Missing MS1 feature file in GUI'."

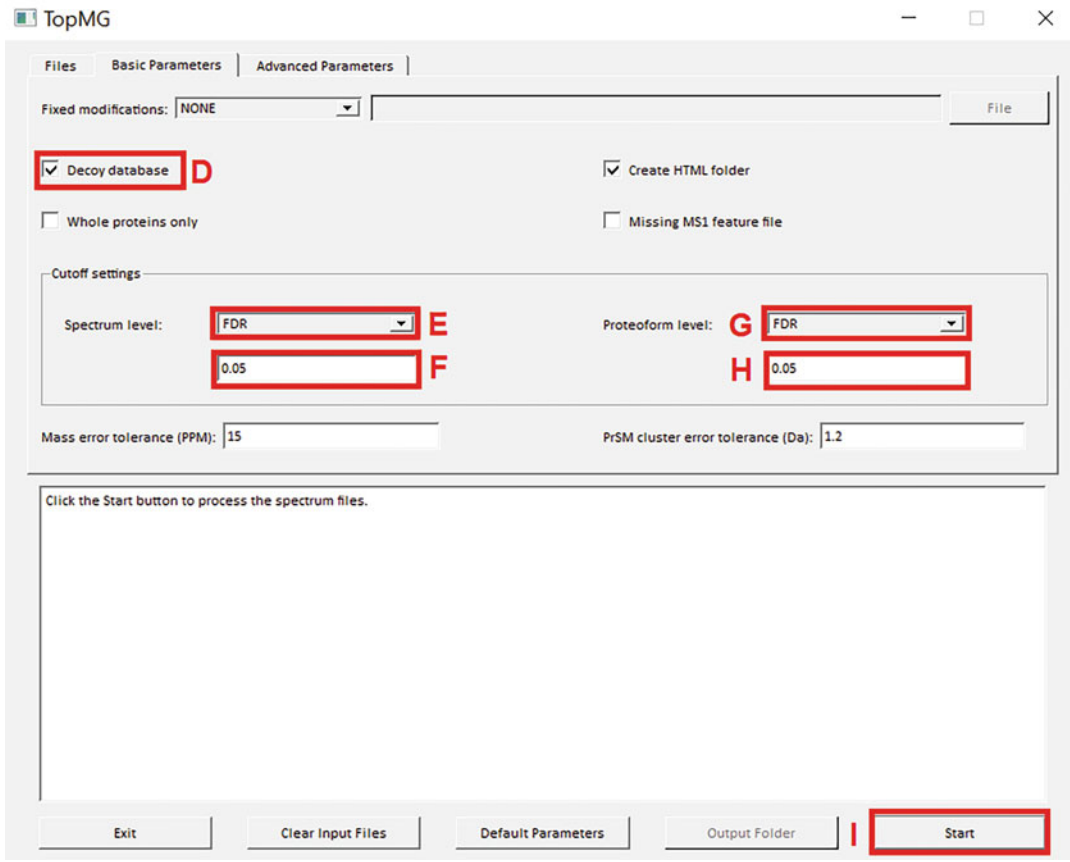


Fig. 9 The Basic Parameters panel of TopMG for proteoform identification by database search

Possible reason: TopFD feature files are missing.

Solution: Go to the Basic parameters tab and check "missing MS1 feature" checkbox. If you removed feature files from the folder by mistake, copy them back to the folder or rerun TopFD to generate the feature files.

3. In TopDiff, the following warning is reported: "LOG WARN: feature_sample_merge.cpp: The file *.xml does not contain any PrSM identifications! Proteoform number: 0 matched number: 0"

   Possible reason: Required files are missing. XML files reported by TopPIC or TopMG are needed as the input of TopDiff.

   Solution: If you have deleted or moved the XML files in the folder, rerun proteoform identification by TopPIC or TopMG.

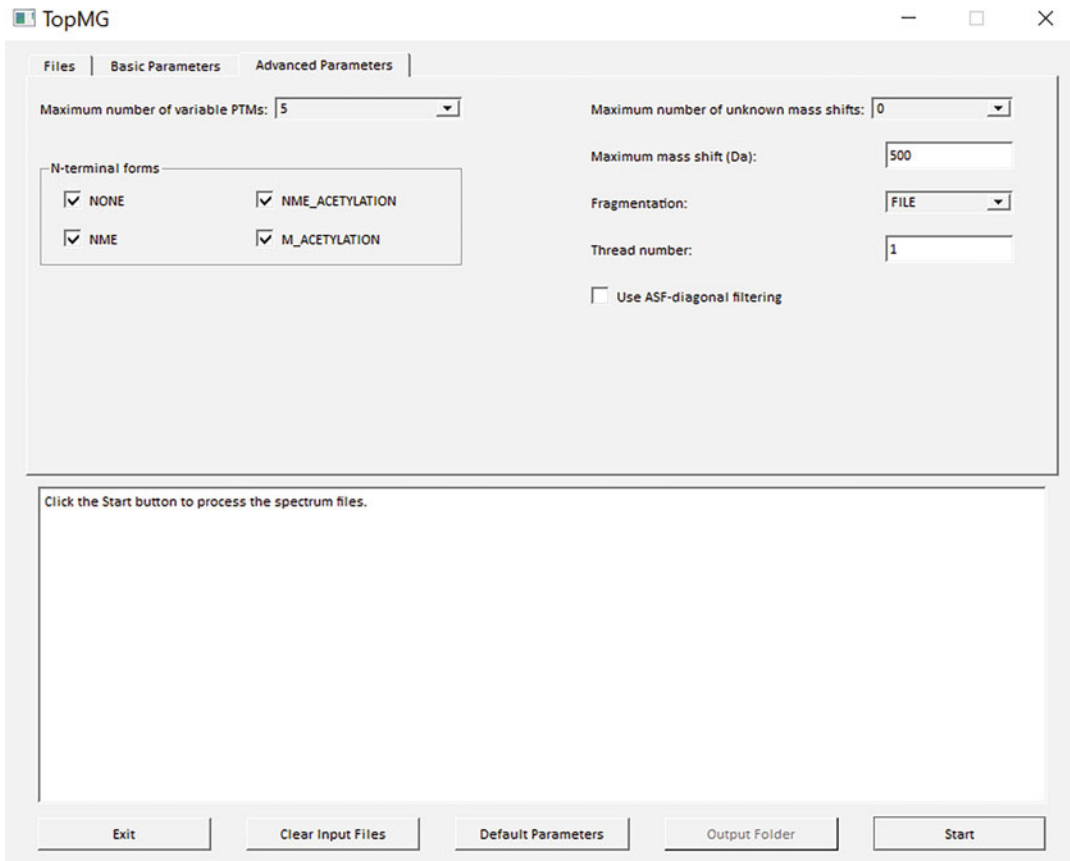


Fig. 10 The Advanced Parameters panel of TopMG for proteoform identification by database search

5 Notes

1. The data set was generated from a top-down LC-MS/MS experiment of *E. coli*. The protein extract of *E. coli* was separated by capillary zone electrophoresis and analyzed by an LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific). Technical duplicates were generated for testing proteoform quantification in two runs of the same sample.
2. TopFD parameters
   (a) Maximum charge (a positive integer): Set the maximum charge state of precursor and fragment ions. Default value is 30.
   (b) Maximum mass (a positive number): Set the maximum monoisotopic mass of precursor and fragment ions. Default value is 100,000 Dalton.
   (c) MS1 S/N ratio (a positive number): Set the signal/noise ratio for MS1 spectra. Default value is 3.

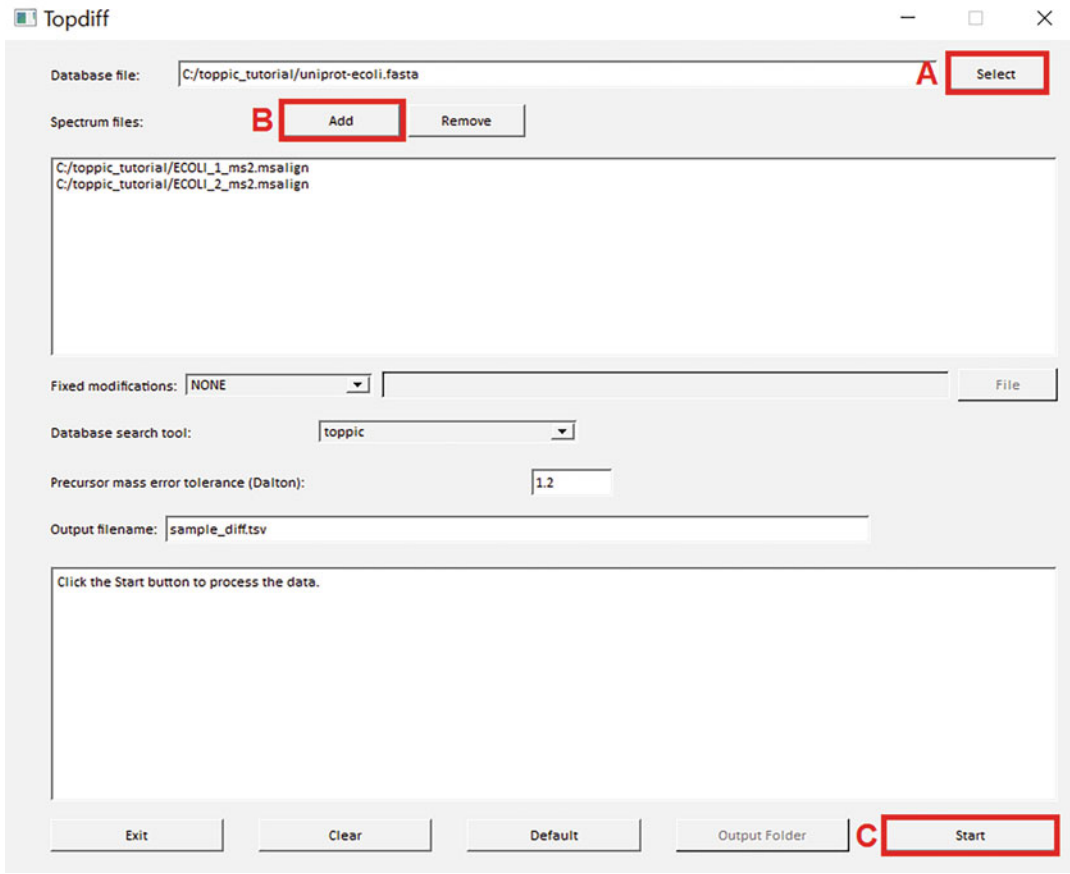


Fig. 11 The window of TopDiff for comparing proteoform abundances in multiple files

(d) MS2 S/N ratio (a positive number): Set the signal/noise ratio for MS/MS spectra. Default value is 1.

(e) Precursor window size (a positive number): Set the precursor isolation window size. Default value is 3.0 m/z.

(f) M/z error (a positive number): Set the error tolerance of m/z values of spectral peaks. Default value is 0.02 m/z.

(g) Thread number (a positive number): Set the number of threads used in the computation. Default value is 1.

(h) Using EnvCNN score: Use EnvCNN as the scoring function of isotopic envelopes. A detailed description of EnvCNN can be found in [11].

(i) Missing MS1 spectra: Specify that the input file does not contain MS1 spectra.

(j) Generate HTML files: Generate JavaScript files for spectral visualization.

3. TopPIC parameters
   (a) Combined output filename: When an MS data set contains data files from multiple fractions, input a file name in "Combined output filename" field. Then the proteoform identifications of the MS files are combined into a single file with the specified output filename.
   (b) Fixed modifications: Set fixed modifications. There are four available options: None, C57, C58, or the name of a text file containing fixed modifications. When None is selected, there is no fixed modification. When C57 is selected, carbamidomethylation on cysteine is the only fixed modification. When C58 is selected, carboxymethylation on cysteine is the only fixed modification. Default value: None.
   (c) Decoy database: Use a shuffled decoy protein database to estimate spectrum and proteoform-level FDRs. When the check box is selected, a shuffled decoy database is automatically generated and appended to the target database before database search. FDRs of identifications are estimated using the target-decoy approach.
   (d) Create HTML folder: Generate JavaScript files for spectral visualization.
   (e) Missing MS1 feature file: Specify that there are no TopFD feature files for proteoform identification.
   (f) Lookup table for E-value computation: Use a lookup table method for computing E-values of identification. It is faster than the default generating function approach, but it may reduce the number of identifications.
   (g) Spectrum-level cutoff type (EVALUE or FDR): Set the spectrum-level cutoff type for filtering proteoform spectrum matches (PrSMs). Default value: EVALUE.
   (h) Spectrum-level cutoff value (a positive number): Set the spectrum-level cutoff value for filtering PrSMs. Default value: 0.01.
   (i) Proteoform-level cutoff type (EVALUE or FDR): Set the proteoform-level cutoff type for filtering proteoforms and PrSMs. Default value: EVALUE.
   (j) Proteoform-level cutoff value (a positive number): Set the proteoform-level cutoff value for filtering proteoforms and PrSMs. Default value: 0.01.
   (k) Mass error tolerance (a positive integer): Set the error tolerance for precursor and fragment masses in part per million (ppm). Default value: 15 ppm. When the lookup table approach is selected for E-value estimation, valid error tolerance values are 5, 10, and 15 ppm.

(l) PrSM cluster error tolerance (a positive number): Set the error tolerance for identifying PrSM clusters (in Dalton). Default value: 1.2 Dalton.

(m) Maximum number of mass shifts (0, 1, or 2): Set the maximum number of unexpected mass shifts in a PrSM. Default value: 1.

(n) Fragmentation (CID, HCD, ETD, UVPD or FILE): Set the fragmentation method(s) of MS/MS spectra. When "FILE" is selected, the fragmentation methods of spectra are given in the input spectrum data file. Default value: FILE.

(o) N-terminal forms: Set N-terminal forms of proteins. Four N-terminal forms can be selected: NONE, NME, NME_ACETYLATION, and M_ACETYLATION. NONE stands for no modifications, NME for N-terminal methionine excision, NME_ACETYLATION for N-terminal acetylation after the initiator methionine is removed, and M_ACETYLATION for N-terminal methionine acetylation.

(p) Maximum mass shift (a number): Set the maximum value for unexpected mass shifts (in Dalton). Default value: 500.

(q) Minimum mass shift (a number): Set the minimum value for unexpected mass shifts (in Dalton). Default value: −500.

(r) Number of combined spectra (a positive integer): Set the number of combined spectra. The parameter is set to 2 (or 3) for combining spectral pairs (or triplets) generated by the alternating fragmentation mode. Default value: 1.

(s) Thread number (a positive number): Set the number of threads used in the computation. Default value: 1. The maximum number of threads is determined by the CPU and memory of the computer used for computation. About 4 GB of memory is required for each thread. If the computer has 16 GB memory and a CPU with 8 cores, the maximum number of threads is 4 because about 16 GB memory is needed for 4 threads.

(t) Modification file for localization: Specify a text file containing a list of common PTMs for proteoform characterization. The PTMs are used to identify and localize PTMs in reported PrSMs with unknown mass shifts.

(u) MIScore threshold (a number between 0 and 1): Set the score threshold (MIScore) for filtering results of PTM characterization. Default value: 0.45.

4. Columns in output TSV files of TopPIC and TopMG
   (a) Data file name: Name of the MS file. Example: ECOLI_1_ms2.msalign.
   (b) PrSM ID: Proteoform spectrum match ID. Example: 1.
   (c) Spectrum ID: Spectrum ID. Example: 10.
   (d) Fragmentation: Fragmentation method for MS/MS spectra. Example: HCD.
   (e) Scan(s): Scan of the MS/MS spectrum for proteoform identification. Example: 502.
   (f) Retention time: Retention time of the MS/MS spectrum for proteoform identification: Example: 965.29 s.
   (g) # peaks: Number of deconvoluted monoisotopic masses in the MS/MS spectrum. Example: 19.
   (h) Charge: Charge state of the precursor ion of the MS/MS spectrum. Example: 7.
   (i) Precursor mass: Monoisotopic neutral mass of the precursor ion. Example: 11922.896 Dalton.
   (j) Adjusted precursor mass: Monoisotopic neutral mass of the precursor ion adjusted based on the molecular mass of the matched proteoform. Example: 10923.251 Dalton.
   (k) Proteoform ID: Proteoform ID assigned by TopPIC or TopMG. Example: 170.
   (l) Feature intensity: The sum of peak intensities of the identified proteoform in LC-MS data. All peaks of various retention times, charge states, and isotopic patterns of the proteoform are included. Example: 2.369E6.
   (m) Feature score: A score for evaluating the quality of the LC-MS feature of the identified proteoform. Example: 15.05.
   (n) Protein accession: Protein accession ID. Example: sp|P76402|YEGP_ECOLI.
   (o) Protein description: A description of the identified protein. Example: UPF0339 protein YegP OS = Escherichia coli (strain K12) GN = yegP PE = 1 SV = 2
   (p) First residue: The position of the first residue of the identified proteoform in the database protein sequence. Example: 2.
   (q) Last residue: The position of the last residue of the identified proteoform in the database protein sequence: Example: 110.

(r) Proteoform: In the proteoform sequence, PTMs and mass shifts are enclosed in squared brackets. For each pair of brackets with PTMs or mass shifts, the candidate sites of the PTMs or mass shifts are annotated using a pair of parentheses right before the pair of brackets. The "." symbol represents the start and end positions of the proteoform. Example: M.AGWFELSKSSDNQFRFVLK AGNGETILTSEL(YTSKTSAEKGIASVRSNSPQEERY EKKTASNGKFYFNLKAANHQIIGSSQMYATAQSRE TGIASVKANGTSQTVKDNT)[37.3184].

(s) # unexpected modifications: The number of unexpected modifications (mass shifts) in the identified proteoform. Example: 1.

(t) MIScore: Confidence scores for PTM characterization and localization. For each identified PTM, a list of candidate sites and their confidence scores are reported. Example: Methyl[K3:88.3%].

(u) # variable PTMs. The number of variable PTMs in the identified proteoform. Only TopMG reports proteoforms with variable PTMs. Example: 0.

(v) # matched peaks: The number of matched deconvoluted masses in the MS/MS spectrum. Example: 5.

(w) # matched fragment ions: The number of matched fragments in the identified proteoform. Example: 5.

(x) E-value: E-value of the PrSM. Example: 0.00058.

(y) Spectrum-level Q-value: Spectrum-level Q-value of the PrSM. Example: 0.

(z) Proteoform-level Q-value: Proteoform-level Q-value of the PrSM. Example: 0.

5. Parameters of TopMG

(a) Combined output filename: When an MS data set contains data files from multiple fractions, input a file name in "Combined output filename" field. Then the proteoform identifications of the MS files are combined into a single file with the specified output filename.

(b) Fixed modifications: Set fixed modifications. There are four available options: None, C57, C58, or the name of a text file containing the information of fixed modifications. When None is selected, there is no fixed modification. When C57 is selected, carbamidomethylation on cysteine is the only fixed modification. When C58 is selected, carboxymethylation on cysteine is the only fixed modification. Default value: None.

(c) Decoy database: Use a shuffled decoy protein database to estimate spectrum and proteoform-level FDRs. When the checkbox is selected, a shuffled decoy database is automatically generated and appended to the target database before database search, and FDRs are estimated using the target-decoy approach.

(d) Whole proteins only: Report only proteoforms of whole protein sequences.

(e) Create HTML folder: Generate JavaScript files for spectral visualization.

(f) Missing MS1 feature file: Specify that there are no TopFD feature files for proteoform identification.

(g) Spectrum-level cutoff type (EVALUE or FDR): Set the spectrum-level cutoff type for filtering PrSMs. Default value: EVALUE.

(h) Spectrum-level cutoff value (a positive number): Set the spectrum-level cutoff value for filtering PrSMs. Default value: 0.01.

(i) Proteoform-level cutoff type (EVALUE or FDR): Set the proteoform-level cutoff type for filtering proteoforms and PrSMs. Default value: EVALUE.

(j) Proteoform-level cutoff value (a positive number): Set the proteoform-level cutoff value for filtering proteoforms and PrSMs. Default value: 0.01.

(k) Mass error tolerance (a positive integer): Set the error tolerance for precursor and fragment masses in ppm. Default value: 15.

(l) PrSM cluster error tolerance (a positive number): Set the error tolerance for identifying PrSM clusters (in Dalton). Default value: 1.2 Dalton.

(m) Maximum number of variable PTMs (a positive number): Set the maximum number of variable PTM sites in a proteoform. Default value: 5.

(n) N-terminal forms: Set N-terminal forms of proteins. Four N-terminal forms can be selected: NONE, NME, NME_ACETYLATION, and M_ACETYLATION. NONE stands for no modifications, NME for N-terminal methionine excision, NME_ACETYLATION for N-terminal acetylation after the initiator methionine is removed, and M_ACETYLATION for N-terminal methionine acetylation. Default value: NONE, M_ACETYLATION, NME, and NME_ACETYLATION.

(o) Maximum number of unknown mass shifts (0, 1, or 2): Set the maximum number of unexpected mass shifts in a proteoform. Default value: 0.

(p) Maximum mass shift (a number): Set the maximum absolute value for unexpected mass shifts (in Dalton). Default value: 500.

(q) Fragmentation (CID, HCD, ETD, UVPD, or FILE): Fragmentation method of tandem mass spectra. When FILE is used, fragmentation methods of spectra are given in the input spectral data file. Default value: FILE.

(r) Thread number (a positive number): Set the number of threads used in the computation. Default value: 1. The maximum number of threads is determined by the CPU and memory of the computer. About 4 GB of memory is required for each thread. If the computer has 16 GB memory and a CPU with 8 cores, the maximum number of threads is 4 because 16 GB memory is required for 4 threads.

(s) Use ASF-diagonal filtering: Use the ASF-DIAGONAL method for protein sequence filtering. The default filtering method is ASF-RESTRICT. When the checkbox is selected, both ASF-RESTRICT and ASF-DIAGONAL will be used. The combined approach may identify more PrSMs, but it is much slower than using ASF-RESTRICT only.

6. Parameters of TopDiff

(a) Fixed modifications: Set fixed modifications. There are four available options: None, C57, C58, or the name of a text file containing fixed modifications. When None is selected, there is no fixed modification. When C57 is selected, carbamidomethylation on cysteine is the only fixed modification. When C58 is selected, carboxymethylation on cysteine is the only fixed modification. Default value: None.

(b) Database search tools (TopPIC or TopMG): Set the identification results used for comparison. When TopPIC is selected, the identifications reported by TopPIC are compared. Otherwise, the identifications reported by TopMG are compared. Default value: TopPIC.

(c) Precursor mass error tolerance. Set the error tolerance for matching proteoforms identified in multiple samples (in Dalton). Default value: 1.2 Dalton.

(d) Output file name: Set the output file name. Default value: sample_diff.tsv.

7. Columns in output files of TopDiff
   (a) Protein accession: Protein accession ID. Example: sp|P0ACF8|HNS_ECOLI.
   (b) Protein description: A description of the identified protein. Example: DNA-binding protein H-NS OS = Escherichia coli (strain K12) GN = hns PE = 1 SV = 2.
   (c) First residue: The position of the first residue of the identified proteoform in the database protein sequence. Example: 2.
   (d) Last residue: The position of the last residue of the identified proteoform in the database protein sequence: Example: 89.
   (e) Proteoform: In the proteoform sequence, PTMs and mass shifts are enclosed in squared brackets. For each pair of brackets with PTMs or mass shifts, the candidate sites of the PTMs or mass shifts are annotated using a pair of parentheses right before the pair of brackets. The "." symbol represents the start and end position of the proteoform. Example: M.SEALKILNNIRTLRAQAR(E)[305.1361]CTLETLEEMLEKLEVVVNERREEESAAAAEVEERTRKLQQYREMLIADGIDPNELLNSLAAVKSGTKAK.R
   (f) Precursor mass. The monoisotopic neutral mass of the proteoform. Example: 10268.322 Dalton.
   (g) Match: If matched LC-MS features are found in all data files, it is "Match." Otherwise, "No Match."
   (h) Spectrum file 1 abundance: Proteoform abundance in the first MS file. Example: 7.03E8.
   (i) Spectrum file 1 spectrum ID: The spectrum ID of the representative MS/MS spectrum of the proteoform in the first spectrum MS file. Example: 396.
   (j) Spectrum file 1 retention time begin: The beginning retention time of the proteoform feature in LC-MS data in the first spectrum MS file. Example: 1315.22 s.
   (k) Spectrum file 1 retention time end: The ending retention time of the proteoform feature in LC-MS data in the first spectrum MS file. Example: 1562.90 s.
   (l) When there are multiple MS files, the information of columns h-l is also provided for the second and other files.

Acknowledgments

The research was supported by National Institutes of Health (NIH) through Grants R01GM118470, R01GM125991, R01AI141625, and R01CA247863.

References

1. Catherman AD, Skinner OS, Kelleher NL (2014) Top down proteomics: facts and perspectives. Biochem Biophys Res Commun 445(4):683–693. https://doi.org/10.1016/j.bbrc.2014.02.041
2. Gregorich ZR, Ge Y (2014) Top-down proteomics in health and disease: challenges and opportunities. Proteomics 14(10): 1195–1210. https://doi.org/10.1002/pmic.201300432
3. Schaffer LV, Millikin RJ, Miller RM, Anderson LC, Fellers RT, Ge Y, Kelleher NL, LeDuc RD, Liu X, Payne SH, Sun L, Thomas PM, Tucholski T, Wang Z, Wu S, Wu Z, Yu D, Shortreed MR, Smith LM (2019) Identification and quantification of proteoforms by mass spectrometry. Proteomics 19(10):e1800361. https://doi.org/10.1002/pmic.201800361
4. Cho WC (2014) Proteomics in translational cancer research: biomarker discovery for clinical applications. Expert Rev Proteomics 11(2): 131–133. https://doi.org/10.1586/14789450.2014.899908
5. Chen C, Hou J, Tanner JJ, Cheng J (2020) Bioinformatics methods for mass spectrometry-based proteomics data analysis. Int J Mol Sci 21(8). https://doi.org/10.3390/ijms21082873
6. Liu X, Inbar Y, Dorrestein PC, Wynne C, Edwards N, Souda P, Whitelegge JP, Bafna V, Pevzner PA (2010) Deconvolution and database search of complex tandem mass spectra of intact proteins: a combinatorial approach. Mol Cell Proteomics 9(12):2772–2782. https://doi.org/10.1074/mcp.M110.002766
7. Liu X, Hengel S, Wu S, Tolic N, Pasa-Tolic L, Pevzner PA (2013) Identification of ultramodified proteins using top-down tandem mass spectra. J Proteome Res 12(12):5830–5838. https://doi.org/10.1021/pr400849y
8. Kou Q, Xun L, Liu X (2016) TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. Bioinformatics 32(22):3495–3497. https://doi.org/10.1093/bioinformatics/btw398
9. Kou Q, Wu S, Tolic N, Pasa-Tolic L, Liu Y, Liu X (2017) A mass graph-based approach for the identification of modified proteoforms using top-down tandem mass spectra. Bioinformatics 33(9):1309–1316. https://doi.org/10.1093/bioinformatics/btw806
10. Kessner D, Chambers M, Burke R, Agus D, Mallick P (2008) ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics 24(21):2534–2536. https://doi.org/10.1093/bioinformatics/btn323
11. Basharat AR, Ning X, Liu X (2020) EnvCNN: a convolutional neural network model for evaluating isotopic envelopes in top-down mass-spectral deconvolution. Anal Chem 92(11): 7778–7785. https://doi.org/10.1021/acs.analchem.0c00903

Chapter 9

Accurate Proteoform Identification and Quantitation Using pTop 2.0

Rui-Xiang Sun, Rui-Min Wang, Lan Luo, Chao Liu, Hao Chi, Wen-Feng Zeng, and Si-Min He

Abstract

The remarkable advancement of top-down proteomics in the past decade is driven by the technological development in separation, mass spectrometry (MS) instrumentation, novel fragmentation, and bioinformatics. However, the accurate identification and quantification of proteoforms, all clearly-defined molecular forms of protein products from a single gene, remain a challenging computational task. This is in part due to the complicated mass spectra from intact proteoforms when compared to those from the digested peptides. Herein, pTop 2.0 is developed to fill in the gap between the large-scale complex top-down MS data and the shortage of high-accuracy bioinformatic tools. Compared with pTop 1.0, the first version, pTop 2.0 concentrates mainly on the identification of the proteoforms with unexpected modifications or a terminal truncation. The quantitation based on isotopic labeling is also a new function, which can be carried out by the convenient and user-friendly "one-key operation," integrated together with the qualitative identifications. The accuracy and running speed of pTop 2.0 is significantly improved on the test data sets. This chapter will introduce the main features, step-by-step running operations, and algorithmic developments of pTop 2.0 in order to push the identification and quantitation of intact proteoforms to a higher-accuracy level in top-down proteomics.

Key words Top-down proteomics, Tandem mass spectrometry, Proteoform identification and quantitation, Search engine, Semi-supervised learning

1 Introduction

The past decade has witnessed the increasing technological and application-oriented progresses in top-down proteomics (TDP), which is the study of all proteins from a cell or an organism on the proteoform level [1–14]. The term "proteoform" designates all of the different molecular forms in which the protein product of a single gene can be found, including all forms of genetic variations, alternative splicing of RNA transcripts, and posttranslational modifications (PTMs) [15]. The latest research reveals that the number of all human proteoforms is estimated to be much larger than the

Liangliang Sun and Xiaowen Liu (eds.), *Proteoform Identification: Methods and Protocols*,
Methods in Molecular Biology, vol. 2500, https://doi.org/10.1007/978-1-0716-2325-1_9,
© The Author(s), under exclusive license to Springer Science+Business Media, LLC, part of Springer Nature 2022

number of the protein-coding genes due to the high-level complexity of the biological processes from the genes to the proteins [1, 2]. The development of sensitive and high-throughput analytical techniques has been the current hotspot to achieve identifying and quantifying all human proteoforms. So far, intact proteoforms are not measured individually one-by-one, but in mixtures by TDP, utilizing a combination of separations such as high-performance liquid chromatography (LC) and high-resolution mass spectrometry (MS). The resulting MS and MS/MS data are then analyzed by bioinformatics tools for intact protein identification and quantitation. Therefore, the analysis of proteoforms in TDP necessitates the use of unique approaches for separation, high-resolution mass spectrometers, and novel data analysis algorithms and tools [14]. The great advances in protein separation and mass spectrometry in the past decade make the software tools dealing with large-scale mass spectra become an urgent need.

Large-scale MS-based proteomic experiments generate a large amount of very complex data, and how to identify and quantify proteoforms from these mass spectra is what bioinformatics needs to address. Because top-down mass spectra are of high complexity, they are often firstly simplified by several deconvolution algorithms such as THRASH [16], DeconMSn [17], MS-Deconv [18], and pParseTD [19], all of which can convert fragment ion peak clusters into a list of single-charged or neutral monoisotopic masses. The subsequent deconvoluted data then serve as the input of top-down search engines such as ProSight [20, 21], MS-Align+ [22], MS-Align-E [23], pTop [19], TopPIC [24], and Informed-Proteomics [25], which deal with the qualitative and quantitative analysis of intact proteins. The community of TDP has contributed more on sample preparation, MS analysis, and designation of proteoforms in recent years [6–8]. However, there is currently no such a similar standard protocol published on how to process the mass spectral data from large-scale intact proteoforms.

Database searching is currently the dominant approach to proteoform identification in TDP. Most search engines could be divided into two categories according to the databases they search. One is those searching the shotgun-annotated protein database (e.g., ProSightPC [20, 21]) and the other is those searching protein sequence database (e.g., TopPIC [24] and pTop [19]). As the first top-down software, ProSightPC searches spectra against a shotgun-annotated protein database, which is generated by considering all potential PTMs and variations. However, this kind of shotgun-annotated protein database is much larger in size than the original protein database. This will be essentially difficult to scale up. ProSightPC can also identify some (but not all) proteins with unexpected PTMs using the optional "biomarker" search mode, where a big mass tolerance (e.g., ±2000Da) is used for precursor mass tolerance, but it is not originally designed to

identify truncated proteins with PTMs that are not represented in the shotgun-annotated database.

MS-Align+ is another popular tool which searches the protein sequence database. Based on a spectral alignment algorithm from MS-TopDown, MS-Align+ could identify proteoforms with at most two unknown mass shifts. It treats all primary structure alterations (PSAs) including genetic variations, alternative splicing, and PTMs, as unknown mass shifts except for fixed PTMs and N-terminal PTMs. MS-Align-E enables to identify proteoforms with variable PTMs, but not those with terminal truncations. TopPIC is an improved version of MS-Align+ with reduced memory and running time. The recently published TopPG is capable of identifying proteoforms with genetic alterations and alternative splicing events [12]. pTop 1.0 supports the identification of proteoforms with both fixed and variable PTMs, but not those with terminal truncations and unexpected PTMs. Moreover, all the above-mentioned software tools do not contain a comprehensive workflow including quantitative analysis for intact proteins. Although several programs exist to do various pieces of top-down data analysis, there is a dearth of software that is capable of handling deconvolution, identification, and quantification within one program.

In this chapter, we present an updated version of pTop 1.0, namely, pTop 2.0, providing a suite of comprehensive analysis tools for TDP, with deconvolution, identification, quantitation modules integrated together. Based on the tag-index process of pTop 1.0, we propose a novel algorithm to identify terminal truncations using the mass deviations of the flanking mass of sequence tags. pTop 2.0 enables the search with variable modifications as well as one unknown modification through a mass index of common modifications. An ion-index process is designed in pTop 2.0 to act as a secondary-search module to further improve the identification rate of MS/MS data. The semi-supervised machine learning approach is introduced to re-rank the search results of pTop 2.0, thus improving the sensitivity of the engine significantly. Tests on both simple and complex sample data sets show that pTop 2.0 could identify much more spectra than pTop 1.0 and other software tools. With multithreading to accelerate the searching process, pTop 2.0 is now a more accurate and efficient software tool for both identification and quantitation of high-throughput intact proteins on the proteoform level.

2 Materials

2.1 General Workflow of pTop 2.0

The top-down MS/MS data analysis pipeline of pTop 2.0 is summarized in Fig. 1, which includes four main procedures.

1. Preprocessing the raw or extracted data.
2. Searching and scoring for each MS2 spectrum.
3. Re-ranking all proteoform-spectrum matches (PrSMs) and filtering to generate the result report.
4. Quantifying identified proteins based on the MS data if needed.

pTop 2.0 integrates pQuant [26], pXtract [27], and pParseTD [19] as its data preprocessing module. pXtract converts RAW data files from Thermo's instruments to *.ms1* and *.ms2* files with centroided peaks. Then pParseTD is employed to detect potential precursors and extract a list of monoisotopic masses with a single charge for each top-down MS/MS spectrum. These two-step processing will greatly simplify the subsequent matching process between an MS/MS spectrum and its candidate proteins. With the incorporation of pQuant, pTop 2.0 can automatically search multiplex-labeled data to achieve accurate quantification. The PrSMs output by pTop 2.0 can be viewed by pLabel, one of our in-house-developed visualization programs.

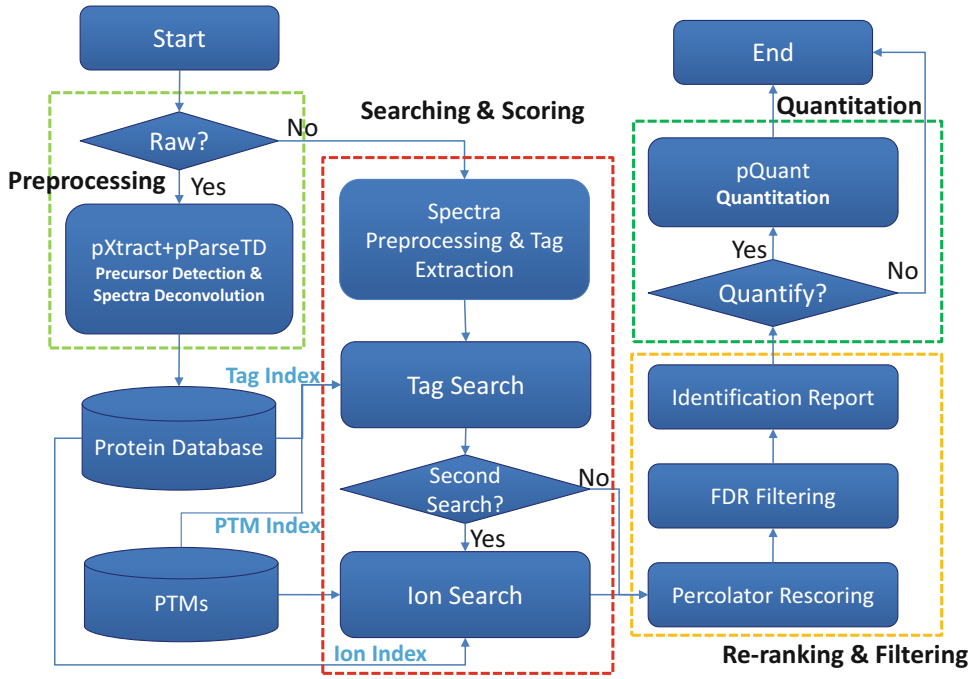


Fig. 1 The workflow of pTop 2.0

2.2 Step-by-Step Running Operations of pTop 2.0

In this section, we will briefly introduce the installation, activation, interface, configuration of PTMs, searching parameters, and results files of pTop 2.0.

2.2.1 Installation and Activation of pTop 2.0

Hardware Requirements
4GB or higher recommended memory.

Software Requirements
Windows 7 or above, microsoft. NET Framework 4.5 or above, MSFileReader 2.2(From Thermo Scientific) or higher version.

The Windows setup package of pTop 2.0 can be downloaded from the website http://pfind.ict.ac.cn/software/pTop/index.html.

After downloading pTop 2.0, the installation step can be done according to the guidance as shown in Fig. 2. All users are required to go through a software activation process to use pTop2.0. A license wizard will guide users for the activation process when the first time pTop2.0 is launched.

2.2.2 The Main Interface of pTop 2.0

There are four main sub-pages of pTop 2.0: MS data, Identification, Quantification, and Summary (Fig. 3). The format of MS data includes Thermo's .RAW and generic .mgf files. MS fragmentation includes HCD, CID, ETD, EThcD, ETciD, and UVPD (Fig. 3a).

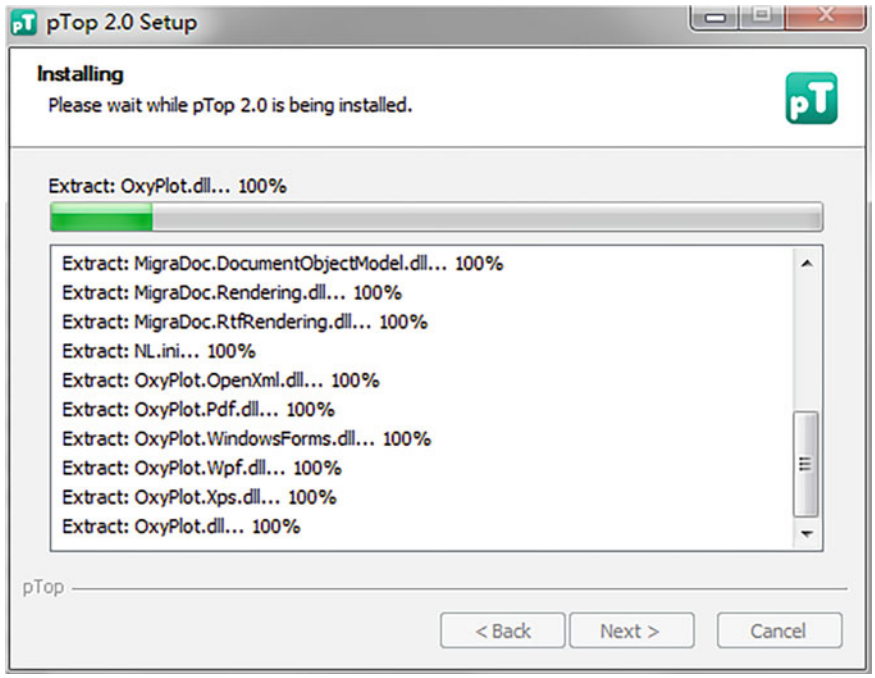
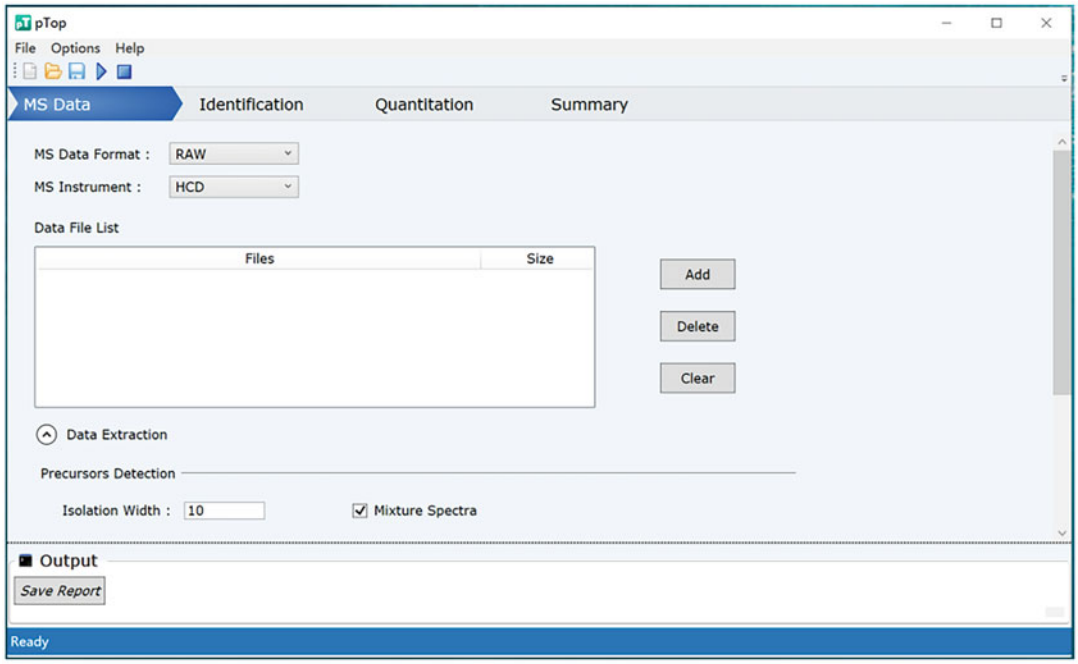
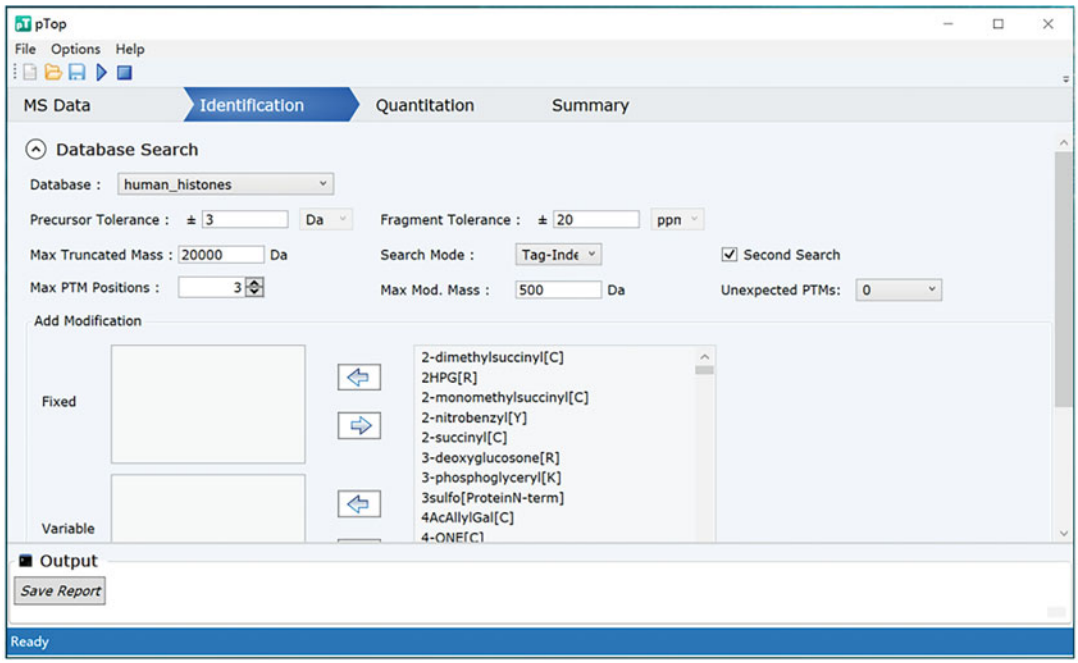


Fig. 2 The installation of pTop 2.0.

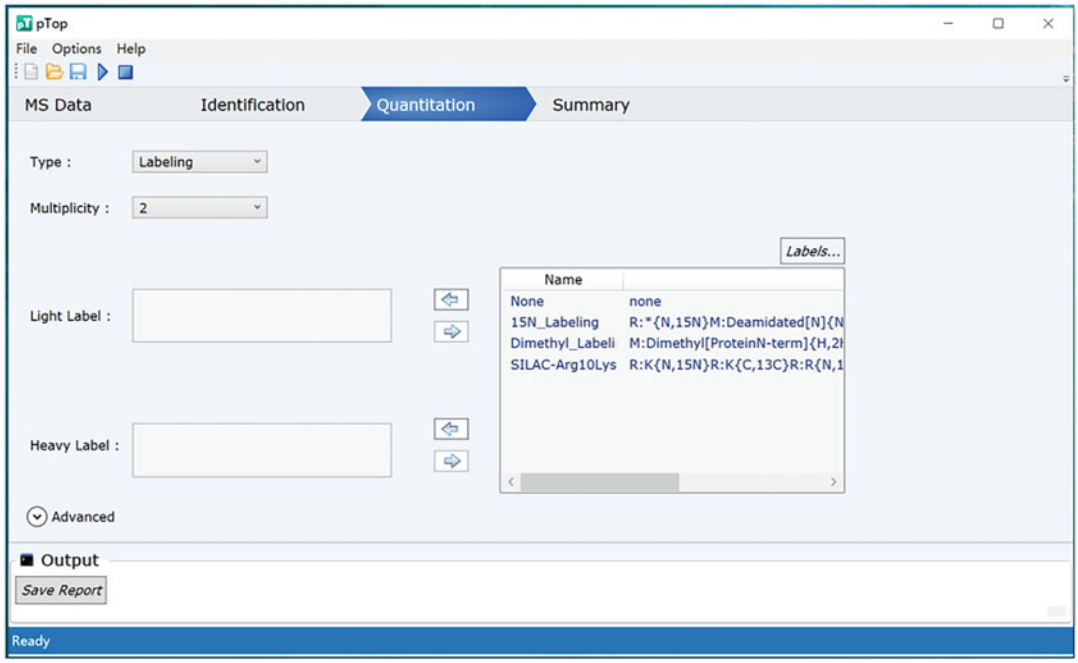


(a) MS data selection

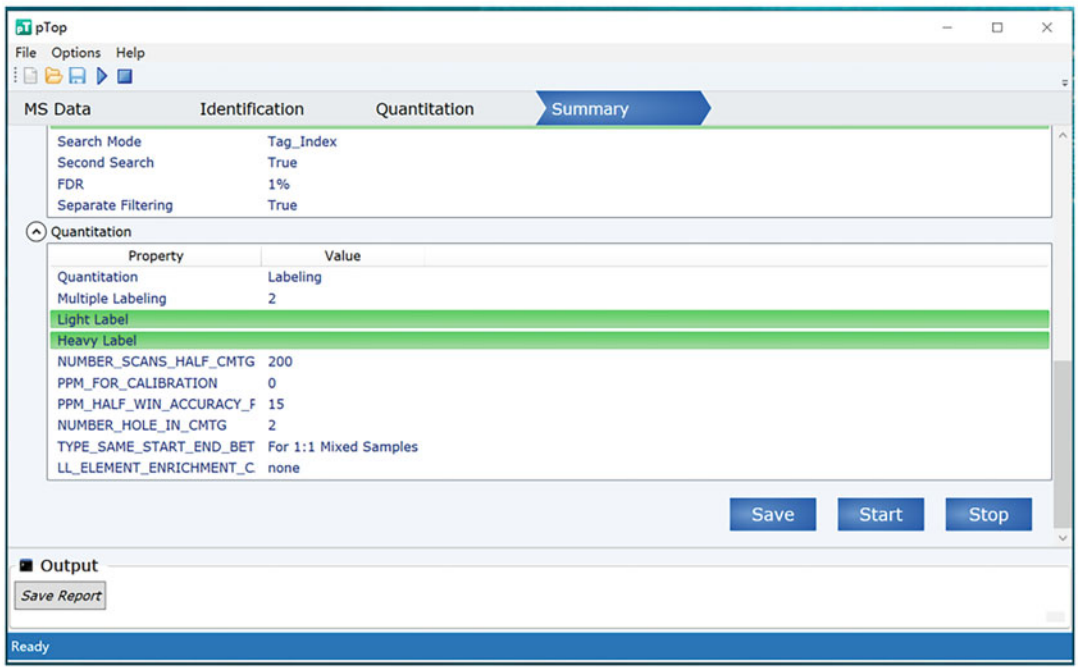


(b) Identification parameters

Fig. 3 The interface of pTop 2.0



(c) Quantification page



(d)Summary of all parameters

Fig. 3 (continued)

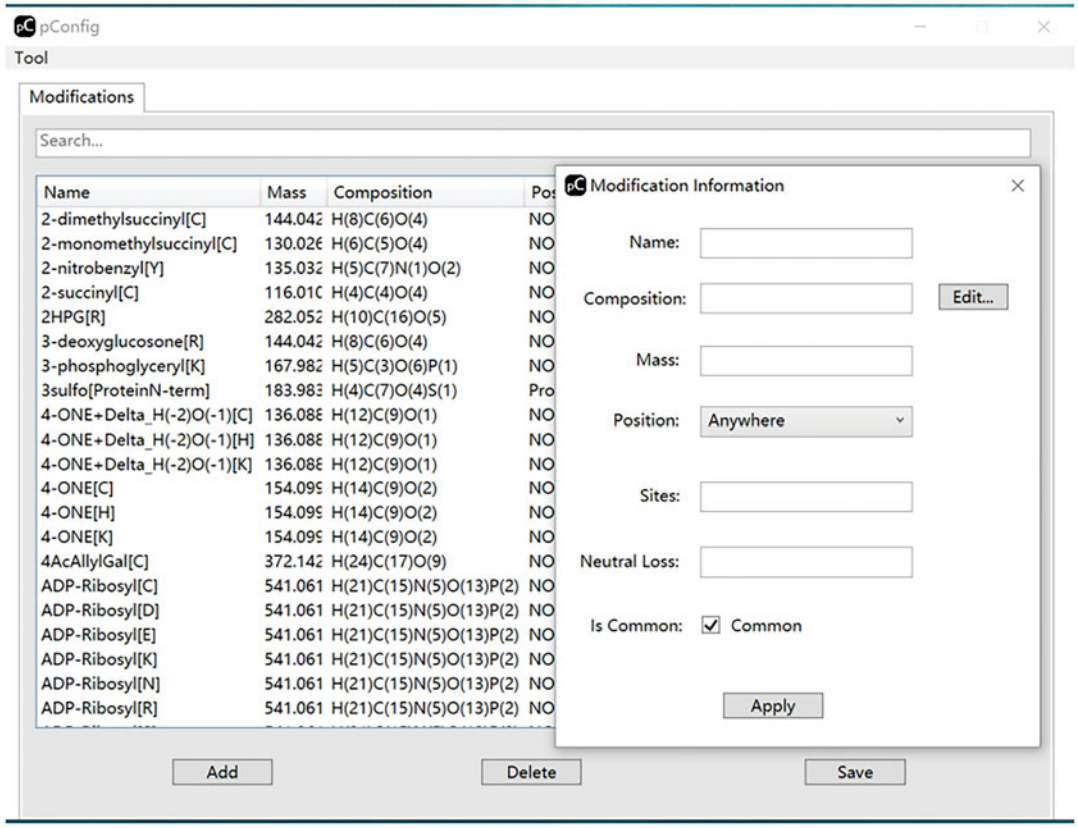


Fig. 4 Edit modification information

The database and searching parameter settings are shown on Identification page (Fig. 3b), which includes the precursor and fragment mass tolerances, max truncated mass, max PTM positions, number of unexpected PTMs, and the fixed and variable PTM selection. Figure 3c is the quantification page, which supports ^{15}N, dimethyl, and SILAC methods. All settings of pTop 2.0 are summarized in Fig. 3d.

2.2.3 Configuration of New Modifications

When searching the protein database with modifications, if a new modification which is not contained in the list is required to be set, pTop 2.0 in Fig. 4 provides the function to define new modifications or edit the old ones. After completing the edition, press the save button to save the information into the modification.ini file.

2.2.4 Parameters in Identification

The detailed illustration for each parameter in identification is listed in Table 1.

2.2.5 Searching Results of pTop 2.0

In the output path, you can find your task folder containing all the results (Fig. 5). The "pTopTask" folder with time stamp is the symbol of a pTop new task. The "param" folder contains the

Table 1
Parameters in identification

Parameter	Description
Database	Protein sequence database to be searched, required
Precursor tolerance	Error tolerance for precursor mass in Dalton. The default value is 5.2 Da.
Fragment tolerance	Error tolerance for fragment ions in ppm. The default value is 15
Max truncated mass	Max mass allowed to be truncated on the N/C terminus. The default value is 20,000
Search mode	The two search modes in pTop 2.0 are tag-index mode and ion-index mode. Tag-index mode gets candidate proteins through tag-index, while ion-index mode acquires candidate proteins through ion-index. When ion-index mode is used, the precursor tolerance can be set as the most, e.g., 50,000
Second search	Once tag-index mode is selected, a second search switch could be turned on. Second search flow use ion index to search those spectra missed by tag index, which may take a little longer time
Max PTM positions	The maximum modification sites (including variable and unexpected) allowed on each protein. The default value is 3
Max Mod. Mass	Maximum absolute value of the mass shift (in Dalton) of a modification. Default value: 500
Unexpected PTMs	Maximum number of unexpected modifications in a proteoform. Default value: 0
Fixed Mods.	Fixed modifications which are certain to occur on the proteins
Variable Mods.	Variable modifications which may occur on some proteins
FDR	The threshold of false discovery rate (FDR). The default value is 0.01
Separate filtering	Whether to calculate FDR and filter the search results for each input file individually. If the switch is turned off, the search results of all the input files will be merged; then, estimate FDR and filter out the results above the FDR threshold

parameter files of this task. There is a folder for each input data file. In the "summary.txt" file, you can find all results about the number of total MS/MS, the identification rate for each input file. In the "out.log" file, you can find the running log of pTop. The ".cfg" file is also a copy of the search parameters. The "pTop.spectra" file contains all search results and the "pTop_filtered.spectra" file contains all identification results above the FDR threshold set previously.

If quantification analysis is done, there will be more files including 1.aa/2.aa and 1.mod/2.mod containing the information of amino acids and modifications under different labels. "pQuant.

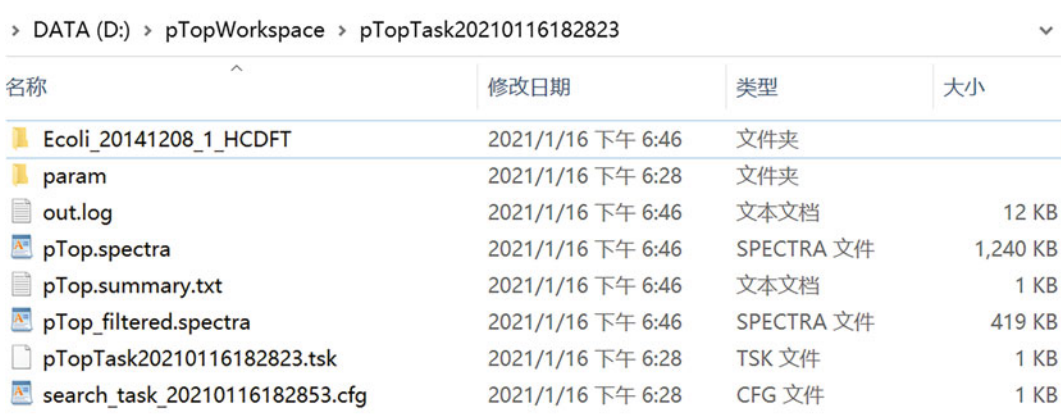


Fig. 5 Output files of pTop 2.0

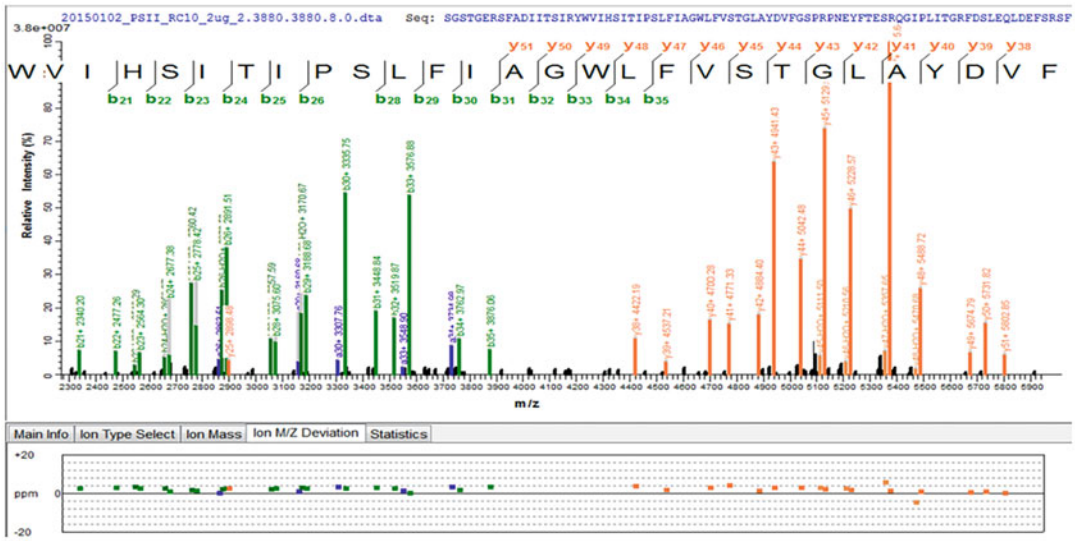


Fig. 6 An identified MS/MS annotated by pLabel program (below is the mass deviations of matched peaks, unit: ppm)

cfg" is a copy of pQuant's parameter file. The "pQuant.protein" file and the "pQuant.protein.list" file contain information of quantified proteins, while the "pQuant.spectra" file and the "pQuant.spectra.list" file contain information of quantified PrSMs.

All identified MS/MS spectra can be visualized by pLabel. One example is shown in Fig. 6.

3 Methods

3.1 Method of Tag-Based Identification of Truncated Proteins

In the main search flow (Fig. 1), pTop 2.0 employs a tag-index method [19] to speed up the retrieval of candidate proteins. For each deconvoluted mass spectrum, we first extract its sequence tags and search these tags against the protein sequence database through the pre-established index of sequence tags. As each tag has a score related to its matching errors and matched peaks intensity, the score of a candidate protein is the sum of scores of the sequence tags it contains. Then all the candidate proteins are sorted by their scores and the top 100 remain for further checking.

With the help of the tag information, the identification of proteoforms with terminal truncations becomes feasible. During the extraction of sequence tags from a spectrum, we record not only the sequences and scores of the tag, but also its two flanking masses. Let $[p_1, p_2, p_3, p_4, p_5]$ be the five peaks that generate the sequence tag *HYTK*, then the left flanking mass of *HYTK* is $mass(p_1)$ and the right flanking mass of *HYTK* is $Precursor_{Mass} - Water\_Mass - mass(p_5)$. For each candidate protein, we sort its tags by their start positions in the protein sequence, and then we could use the left flanking mass difference of the left-most tag to check the N-terminal truncation and the right flanking mass difference of the right-most tag to check the C-terminal truncation. For example, the protein in Fig. 7 has four tags: HYTK, TKRV, AVRL, and VRLL. The left flanking mass of the left-most tag HYTK in the

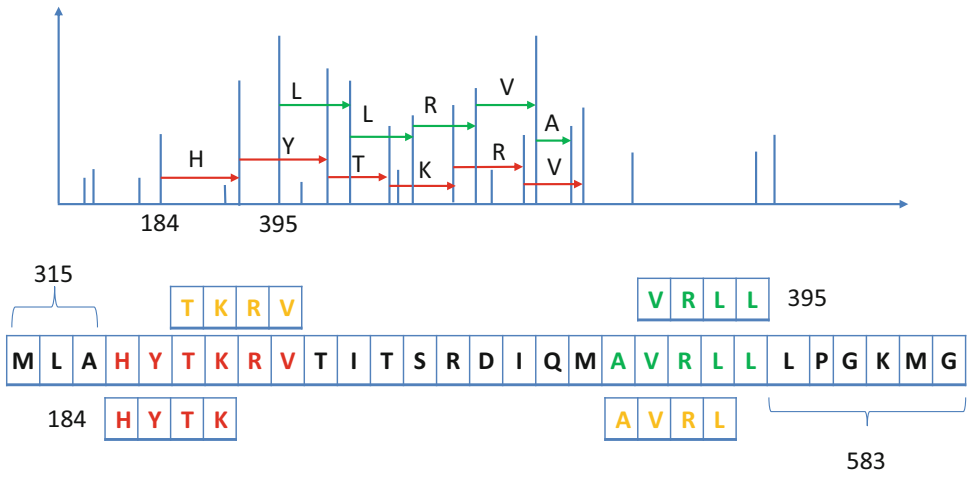


Fig. 7 Identification of terminal truncations. The protein has four tags: HYTK, TKRV, AVRL, and VRLL. The left flanking mass of the left-most tag HYTK in the spectrum is 184, but the left flanking mass of the tag in the protein sequence is 315(MLA), and the mass difference is the mass of "M" ($184 - 315 = -131$). The right flanking mass of the right-most tag VRLL in the spectrum is 395, but the right flanking mass of the tag in the protein sequence is 583(LPGKMG), and the mass difference is the mass of "MG" ($395 - 583 = -188$). Therefore, the truncated protein sequence is LAHYTK. . .VRLLLPGK

spectrum is 184, but the left flanking mass of the tag in the protein sequence is 315 (the mass of "MLA"), and the mass difference is exactly for the mass of "M" (184 − 315 = − 131). The right flanking mass of the right-most tag VRLL in the spectrum is 395, but the right flanking mass of the tag in the protein sequence is 583 (the mass of "LPGKMG"), and the mass difference is exactly for the mass of "MG" (395 − 583 = − 188).

Not all spectra contain sequence tags whose lengths are above a fixed number, such as 4 or 5 (pTop 2.0 uses 4-tag and 5-tag to create index). For those spectra without adequate sequence tags, a large mass tolerance, such as ±500 Da, will be used for searching the candidate proteins. If the mass difference between the precursor ion and the candidate protein is negative, we cut several amino acids from the N-terminus of the protein sequence to make it positive and add the N-terminus-truncated protein sequence to the candidate set. In the similar fashion, we can also add the C-terminus-truncated protein sequence to the candidate set. Finally, we may get hundreds of candidate protein sequences for a specific spectrum. pTop 2.0 employs a rough score to filter the candidate protein sequences and remains the top 100 for later identification of proteoforms.

3.2 Identification and Characterization of Proteoforms

Given the spectrum S and the candidate protein P, we could use the mass difference between the precursor mass of S and the mass of P to retrieve the candidate modifications through a pre-established PTMs index. There are two components in the PTMs index: one is mass index for user-specified variable modifications (variable PTMs index) and the other is mass index for modification combinations with at least one unexpected modification (unexpected PTMs index). We select 460 common modifications from Unimod [28] as candidate unexpected modifications. To identify proteoforms with both variable modifications and unexpected modifications, we take a two-step search strategy. We first retrieve the variable PTMs index to get variable PTM combinations. For each combination, we use a graph-based method [19] to locate the modifications and get the candidate proteoforms. Then each proteoform is scored by the following formula:

$$\text{Score} = \sum_{i=1}^{n} \left(\frac{(1+b)I_i}{b+I_i} \right) \left(0.5 + \cos\left(\frac{|\,\text{Tol}_i\,|\,\pi}{2\text{FragTol}} \right) \right) \qquad (1)$$

In Eq. 1, I_i is the absolute intensity of a matched peak, the parameter b is well-tuned as 0.03. FragTol refers to the user-defined matching mass tolerance of fragment ions, e.g., 20 ppm. Tol_i refers to the mass tolerance of a matched fragment ion ranging from –FragTol to +FragTol.

For the spectrum *S*, we sort its candidate proteoforms by their scores. If the top score is below a threshold, e.g. 18, we then retrieve the unexpected PTMs index to get unexpected PTM combinations and then use the same method to locate the PTMs and score the candidate proteoforms. The mass index of PTMs and the two-step search strategy has substantially improved the search efficiency of pTop 2.0.

3.3 Re-ranking and Filtering

In addition to the matching error and intensity which are considered during the refined scoring, we need another normalized score to re-rank all the PrSMs and ensure that a large fraction of the true positive identifications are retained. We implemented the re-ranking module based on Percolator's semi-supervised learning algorithm, and thus improved significantly pTop 2.0's sensitivity.

To establish a machine learning model, the feature extraction and training samples selection should be of prime importance. We extract 11 features, including nine intra-spectra features and two inter-spectra features. The intra-spectra features are mainly extracted from the matching process between the spectra and proteoforms while the inter-spectra features are from the statistical information of the samples. The SVM score of the last round is one of the intra-spectra features. To select proper training samples, we employ the target-decoy strategy to estimate FDR, then we choose the target identifications within FDR $\leq$ 0.5% as positive samples. All decoy identifications are negative samples. We use these 11 features to iteratively train an SVM model and predict the SVM score for 10 times and yield a better normalized score for each spectrum. The workflow of Percolator in pTop 2.0 is shown in Fig. 8. In each iteration, training samples would be reselected and features would be updated.

The SVM package we used is LibLinear [29], a widely used library for large linear classification. To improve the precision of the model and speed up the convergence of gradient descent, we perform feature scaling on both training set and testing set with min-max normalization. For the convenience of comparison, we normalize the predicted score with function $y = 1/(1 + e^{10x})$ as the final score in place of *E*-value. The test results have shown that sensitivities are improved after re-ranking on two complex sample data sets compared with ranking with *E*-value.

3.4 Protein Quantitation

We integrate pQuant [26] into pTop 2.0 as the quantitation module, which makes it possible to accurately quantify intact proteoforms based on MS data. pQuant detects interference signals and identifies for each peptide/protein a pair of least interfered isotopic chromatograms—one for the light isotope-labeled and the other for the heavy isotope-labeled peptide/protein. Based on the isotopic pairs, pQuant calculates the relative heavy/light peptide/protein ratios. pTop 2.0 could search both the light isotope-labeled

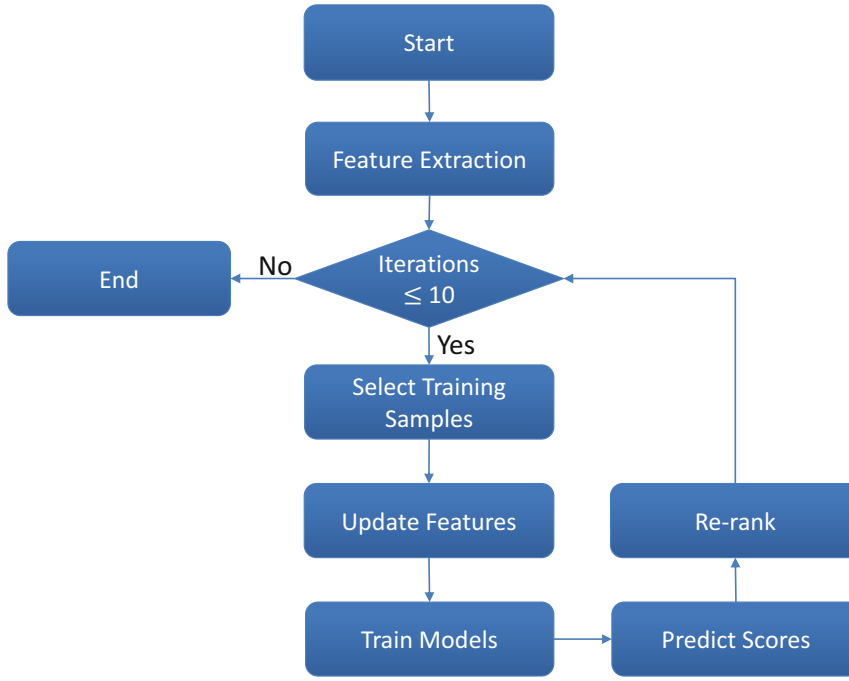


Fig. 8 Workflow of Percolator in pTop2.0

and the heavy isotope-labeled tandem mass spectrometry data, merge the result together, and finally use pQuant to quantify each identified PrSM and estimate the ratios of proteoforms. For those PrSMs, whose light or heavy signals are missing or very low in MS scans, pQuant outputs "NaN" (Not a Number) ratios. The proportion of NaN in all identified PrSMs can be used to evaluate the accuracy of the search engine's FDR.

4 Results of pTop 2.0 on Three Test Data Sets

4.1 Data Sets

The performance of pTop 2.0 was tested on three data sets. The first data set containing 11,378 CID spectra was generated from the human histone H2A, H2B, H3, and H4 proteins by LTQ Orbitrap Velos mass spectrometer [30]. The second data set is a much more complex one from *E. coli* intact proteome generated on Q Exactive with three technical replicates [31]. The total numbers of HCD spectra were 16,573, 16,533, and 16,591, respectively. The third data set is a mixture of ^{14}N-labeled fission yeast (*Schizosaccharomyces pombe*) and ^{15}N-labeled yeast with a mixing ratio of 1:1 generated on Q Exactive HF (unpublished). The fission yeast sample was fractionated into discrete molecular weight (MW) ranges by use of GELFrEE. GELFrEE fractions C-G were collected from three technical replicates with a resolution of 120,000 for MS spectra and 60,000 for MS/MS spectra (at mass 200 Da), and a total of 23,837 HCD spectra were acquired.

The first data set was searched against a small database containing only target proteins. The second data set was searched against UniProt *E. coli* database and the third data set was searched against Uniprot yeast database. We first tested the performance of pTop 2.0 on the first two data sets and compared it with pTop 1.2 and TopPIC in terms of sensitivity and accuracy. Here, the sensitivity of the search engine is mainly measured by the identification rate of the spectra and the number of identified PrSMs while filtered with FDR $\leq$ 1% at the spectra level. The accuracy of the search engine is measured by the ratio/count of matched fragment ions and the ratio of matched peaks intensity of the identified PrSMs. To compare fairly, all the top-down MS/MS spectra were deconvoluted by pParseTD and searched against the target-decoy concatenated database. Then, we tested the quantitation performance of pTop 2.0 on the ^{15}N-labeled yeast data set. We proposed here, for the first time, the idea of validating the FDR of large-scale identification results by using quantification information of ^{15}N-labeled data in top-down proteomics. We also assessed the accuracy and stability of FDR estimates on the ^{15}N-labeled yeast data set by using a sample and entrapment database.

4.2 Proteoform Identifications on the Human Histone Data Set

The human histone data set was searched by pTop 1.2, TopPIC, and pTop 2.0, respectively. Two different searching modes of pTop 2.0 were tested in this study. The first mode is searching without unexpected modifications (referred to as pTop 2.0 (uPTM = 0)) which has the same parameters as pTop 1.2. The second mode is searching with one unexpected modification as well as user-defined variable modifications. TopPIC was tested with similar parameters, such as two unexpected modifications.

The identification rates as well as the number of identified PrSMs and proteoforms of pTop 1.2, pTop 2.0, and TopPIC are shown in Table 2. pParseTD may export more than one precursor for some MS/MS scan, and that is the reason why the number of PrSMs is larger than the number of MS/MS scans. When searching without unexpected modifications, pTop 2.0 yielded an identification rate of 51.9%, which is 6.4% higher than that of pTop 1.2 (45.5%) and 11.9% higher than that of TopPIC (40%). pTop 2.0 (uPTM = 0) identified 2025 more PrSMs with at least 20 matched ions and 890 more proteoforms than pTop 1.2. When searching with one unexpected modification, the identification rate was increased by about 3% and 631 more PrSMs with at least 20 matched ions were identified comparing with "uPTM = 0" mode. As modification sites cannot be accurately determined with deficient fragment ions, they are not considered when calculating the number of proteoforms. That is to say, we count two proteoforms as one when both their protein sequences and modification combinations are same. pTop 2.0 could identify more proteoforms than pTop 1.2 (Table 2).

Table 2
Identification results of pTop 1.2, pTop 2.0, and TopPIC

Software	ID Rate[a]	# ID PrSMs[b]	# ID PrSMs with at least 20 matched ions	# Proteoforms[c]
pTop 1.2	**45.5%**	21,096	14,546	792
pTop 2.0(uPTM = 0)	**51.9%**	25,052	16,571	1682
pTop 2.0(uPTM = 1)	**54.8%**	27,202	17,202	2454
TopPIC	**40%**	17,322	14,496	*

[a]ID Rate = identified scan number/total scan number
[b]As pParseTD could export co-eluted spectra, one scan may correspond to several spectra and each spectrum corresponds to a PrSM
[c]Modification sites are not considered when counting proteoforms here, and the number of proteoforms identified by TopPIC cannot be counted exactly because it does not provide full modification information for some unknown mass shifts(*)

The consistency between pTop 2.0 and pTop 1.2 is shown in Fig. 9a. pTop 2.0 could cover 67.7% of PrSMs identified by pTop 1.2, and the percentages of results individually identified by pTop 2.0 and pTop 1.2 were 43.0% and 32.3%, respectively (two PrSMs are considered consistent if and only if the protein sequences are identical in spite of the modification types for the same spectrum). To assess the accuracy and reliability of the identifications, the ratios of matched fragment ions (referred to as Matched_Ion_Ratio, defined as the proportion of matched fragment ions in all the theoretical fragment ions) and the ratios of matched peaks intensity (referred to as Matched_Intensity_Ratio, defined as the proportion of matched peaks' intensity in all the peaks' intensity) are analyzed. The statistical analysis of matched intensity ratio and matched ion ratio in the consensus and individual parts is shown in Fig. 9b. The ratios of matched intensity and matched ions in pTop 2.0's individual identifications tend to be larger than those in pTop 1.2, indicating that the reliability of individual identifications of pTop 2.0 is higher than that of pTop 1.2 to some extent. Among the 10,780 PrSMs identified individually by pTop 2.0, ~68% were truncated either in the N terminal or C terminal. Figure 9c shows the top-score proteoform with terminal truncations.

The consistency between pTop 2.0 and TopPIC and the matched fragment ions of the two engines were also analyzed. pTop 2.0 could cover 99.5% of spectra identified by TopPIC and identify 57% more spectra than TopPIC. For all 17,228 spectra identified by both pTop 2.0 and TopPIC or the individual identifications of the two engines, proteoforms reported by pTop 2.0 tend to match more fragment ions than those reported by TopPIC (data not shown).

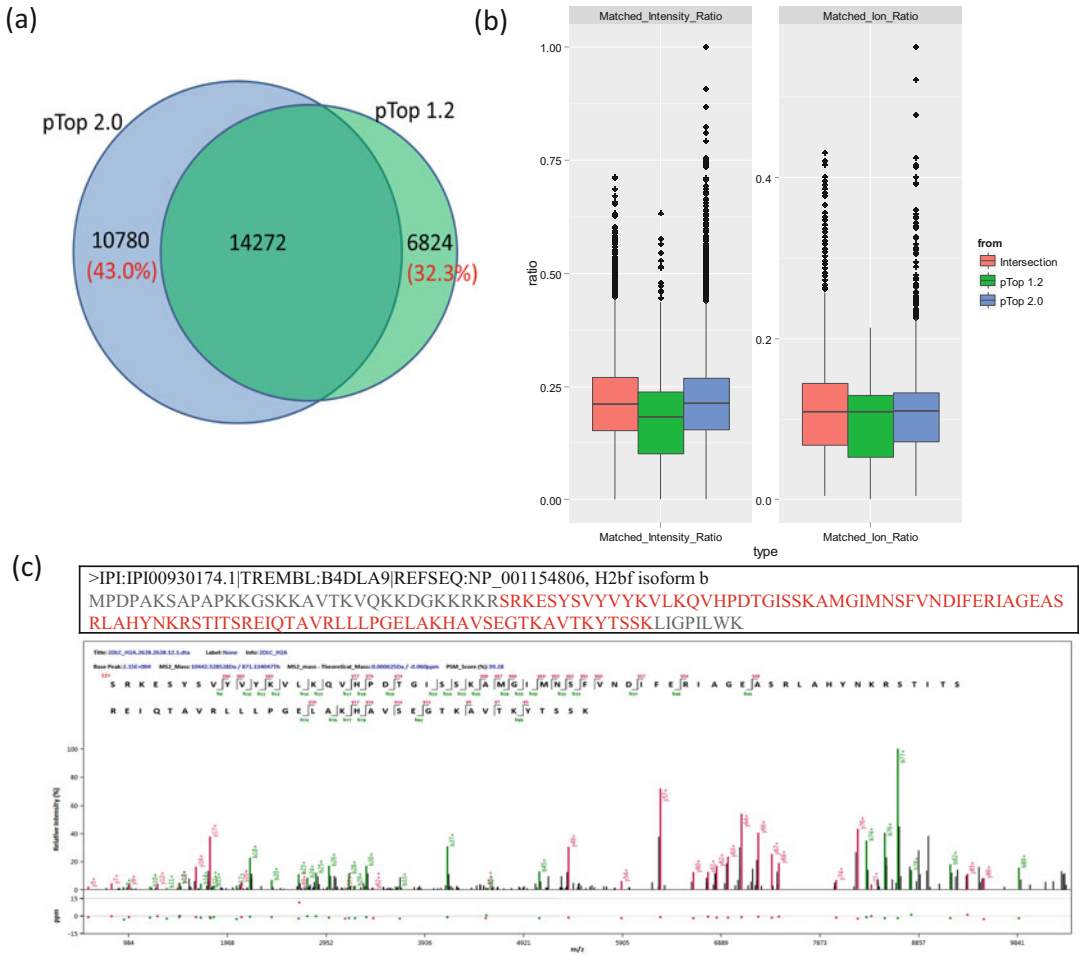


Fig. 9 Comparisons of results between pTop 1.2 and pTop 2.0(uPTM = 0) on human histone data set. (**a**) Venn diagram showing the PrSMs identified by pTop 1.2 and pTop 2.0. 14,272 PrSMs were identified by both pTop 1.2 and pTop 2.0 indicating that the protein sequences are identical for the same spectrum. (**b**) A comparison of matched intensity ratio and matched ion ratio between the intersection and subtraction. "Intersection" means the same spectrum with the same protein sequence, the 14,272 PrSMs identified by both pTop 1.2 and pTop 2.0; "pTop 1.2" represents the 6824 PrSMs identified only by pTop 1.2; "pTop 2.0" represents the 10,780 PrSMs identified only by pTop 2.0. (**c**) A proteoform with N-terminal truncation (32 amino acids are removed) and C-terminal truncation (8 amino acids are removed) identified by pTop 2.0

4.3 Proteoform Identifications on the E. coli *Data Set*

To test the performance of pTop 2.0 over complex sample data sets, the *E. coli* intact proteome data including three technical replicates were searched by pTop 1.2 and pTop 2.0. Similar analysis and comparison were performed over the two engines. Table 3 shows the identification rates as well as the number of proteins and proteoforms. On all of the three data sets, pTop 2.0 yielded an identification rate of 37.4–41.4%, which is about 8% on average higher than that of pTop 1.2. On the overall data set, pTop 2.0 identified

Table 3
Identification results of pTop 1.2 and pTop 2.0 on *E. coli* data set

	pTop 1.2			pTop 2.0		
Raw	**ID Rate**	**# Proteins**	**# Proteoforms**	**ID Rate**	**# Proteins**	**# Proteoforms**
E.coli_rep1	**28.73%**	152	980	**37.38%**	213	1285
E.coli_rep2	**32.59%**	166	1122	**38.53%**	209	1716
E.coli_rep3	**31.55%**	155	1081	**41.41%**	233	1520
Overall	**30.96%**	242	2075	**39.11%**	401	3124

401 proteins and 3124 proteoforms, which is 159 more proteins and 1049 more proteoforms than pTop 1.2.

The consistency of PrSMs identified by pTop 2.0 and pTop 1.2 is shown in Fig. 10. About 75% of the results reported by pTop 1.2 can also be obtained by pTop 2.0 and the individual part of pTop 2.0 takes 45% of its own results. As shown in Fig. 9b, the matched intensity ratios and matched ion ratios of the consensus part tend to be higher than those of the individual parts. In terms of the individual identification results, PrSMs identified only by pTop 2.0 tend to have higher matched intensity ratio and matched ion ratio than PrSMs identified only by pTop 1.2. Among the 18,302 PrSMs identified individually by pTop 2.0, ~66% were those with terminal truncations, and the top-score one was shown in Fig. 10c.

As shown above, pTop 2.0 could identify more spectra and proteoforms on both simple data and complex data. What is more, the PrSMs identified by pTop 2.0 tend to match more fragment ions and high-intensity peaks than those identified by pTop 1.2 and TopPIC.

4.4 Quantitative Analysis on the Yeast Data Set

pTop 2.0 completed the identification and quantitation of the ^{15}N-labeled yeast proteome data set in the user-friendly "one-key operation." A total of 14,121 scans (59%), 460 proteins, and 2391 proteoforms (without consideration of modification sites) were quantified. The ratios of the spectra are concentrated around 1:1 whose log base 2 is 0. The mean, median, and standard deviation of spectra ratios are 1.02, 1.01, and 1.03, respectively. We took the median of the spectra ratios as the ratio of the proteoform on condition that the spectra identified the same proteoform. The mean, median, and standard deviation of spectra ratios are 1.06, 1.03, and 2.04, respectively. Thus the ratios of proteoforms are more diverse than those of spectra but still concentrated around 1:1 (data not shown).

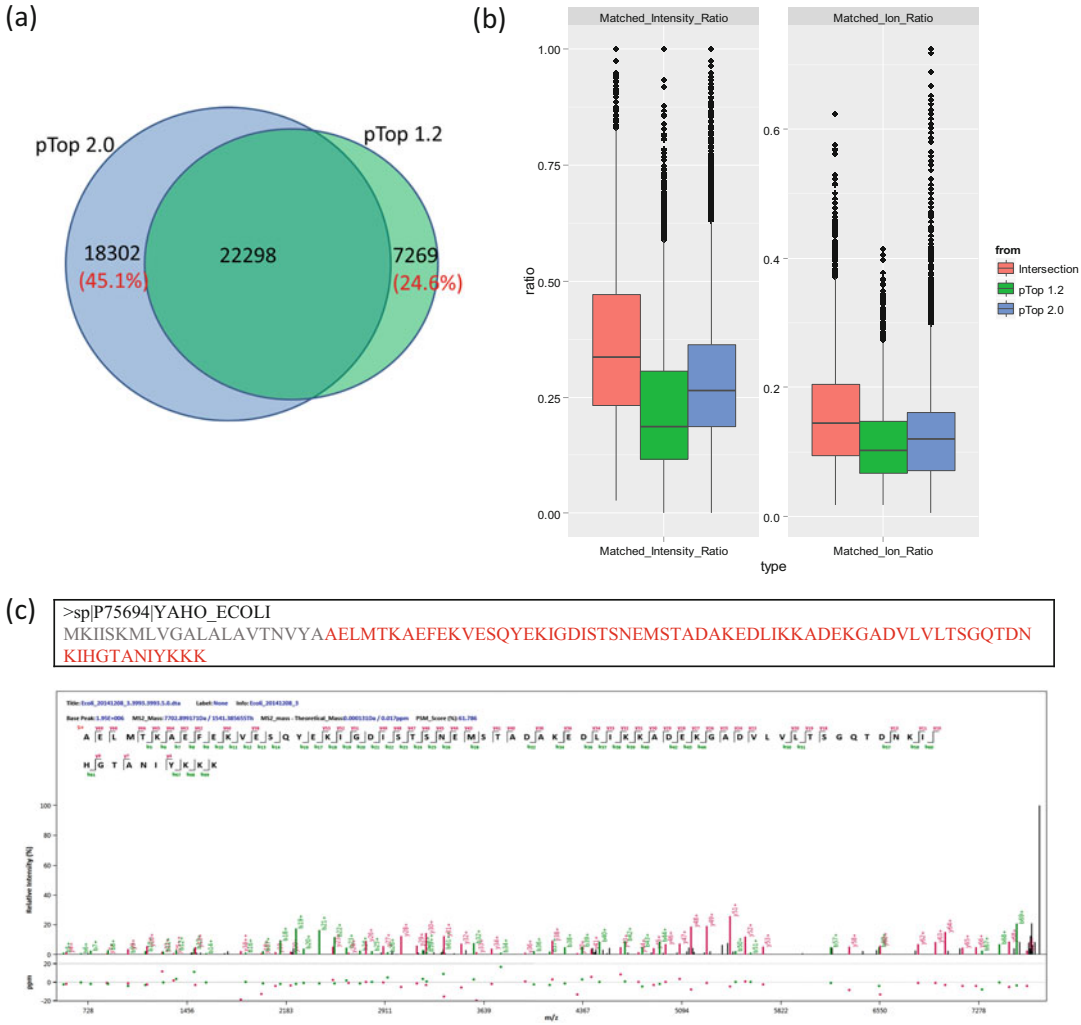


Fig. 10 Comparison of results between pTop 1.2 and pTop 2.0 on *E. coli* data set. (**a**) Venn diagram showing the PrSMs identified by pTop 1.2 and pTop 2.0. 22,298 PrSMs were identified by both pTop 1.2 and pTop 2.0 indicating that the protein sequence and modification types are identical for the same spectrum. (**b**) A comparison of matched intensity ratio and matched ion ratio between the intersection and subtraction. "Intersection" means same spectrum with same protein sequence, that is the 22,298 PrSMs identified by both pTop 1.2 and pTop 2.0; "pTop 1.2" represents the 7269 PrSMs identified only by pTop 1.2; "pTop 2.0" represents the 18,302 PrSMs identified only by pTop 2.0. (**c**) A proteoform with N-terminal truncation (21 amino acids are removed) identified by pTop 2.0

4.5 Validation of FDR Estimates

The most commonly used target-decoy approach (TDA) is employed in pTop to estimate the false discovery rate (FDR). We then used pQuant to quantify the PrSMs filtered by FDR $\leq$ 1% at spectra level. The "NaN" ratios reported by pQuant on the above ^{15}N-labeled yeast data set were used to validate the FDR estimates of pTop 2.0. Those PrSMs with "NaN" ratios, whose light or heavy signals are missing or very low in MS spectra, are most possibly the

Table 4
The number of "NaN" ratios of pTop 1.2 and pTop 2.0 on ^{15}N-labeled yeast data sets

	pTop 1.2		pTop 2.0	
Data set	**# ID spectra**	**# NaN (Ratio)**	**# ID spectra**	**# NaN (ratio)**
U+15N	4369	28 (0.64%)	6721	36 (0.54%)
U+15N_rep1	3197	17 (0.53%)	4831	20 (0.41%)
U+15N_rep2	4376	42 (0.96%)	6564	47 (0.72%)
Overall	11,942	87 (0.73%)	18,116	103 (0.57%)

wrong identifications. Therefore, the proportion of "NaN" ratios in all the identified PrSMs can be used for an independent merit to estimate the error rates on the isotopic labeled datasets. The number of identified spectra and "NaN" ratios of the three technical replicates are shown in Table 4. Both pTop 1.2 and pTop 2.0 have less than 1% "NaN" ratios on the three technical replicates, i.e., most of their results can be assigned with expected quantitation ratios. What is more, pTop 2.0 reported less "NaN" ratios than pTop 1.2 which indicates that the error rate of pTop 2.0 is lower than that of pTop 1.2.

To further assess the accuracy and stability of FDR estimates of pTop 2.0, we searched the three technical replicates of yeast proteome against a big target-decoy concatenated database. The target database was a concatenation of the sample database containing 5149 target proteins and the entrapment database containing 91,464 human proteins. The decoy database was constructed by connecting the reversed sequence of the first half and the second half of the protein sequences in the target database. We still used the target-decoy strategy to estimate the FDR and report the identification results at 1% spectra-level FDR. Assuming that any match to the sample database is a true positive and any match to the entrapment database is a false positive, the percentage of identifications from the entrapment database can reflect relatively the error rate and also act as another validation of the FDR estimates. The number of identified spectra and those from the entrapment database are shown in Table 5. Except for the error rate of the first technical replicate being 1.06%, both the other two replicates have an error rate of less than 1%, and the overall error rate under the entrapment strategy is 0.75%, indicating that the 1% spectra-level FDR of pTop 2.0 is of high accuracy.

4.6 Analysis of the Running Time

The running time of pTop 1.2, pTop 2.0, and TopPIC on human histone H2A data set is listed in Table 6. While pTop 1.2 does not support multi-threading, its single-threaded speed is the fastest, even faster than pTop 2.0 with three threads and TopPIC with six

Table 5
The number of identified spectra from entrapment database on ^{15}N-labeled yeast data sets

Data set	# ID spectra	# ID spectra of entrapment (ratio)
U+15N	5823	62 (1.06%)
U+15N_rep1	4116	18 (0.44%)
U+15N_rep2	5837	39 (0.67%)
Overall	15,776	119 (0.75%)

Table 6
Running time of pTop 1.2, pTop 2.0, and TopPIC on H2A (unit: min)

	pTop 1.2	pTop 2.0 (uPTM = 0)	pTop 2.0 (uPTM = 1)	TopPIC
Threads = 1	17.5	68.5	73.1	421.3
Threads = 3	*	25.1	26.6	372.4
Threads = 6	*	14.2	16.5	259.9

All of the running time tests were performed on the same PC (AMD Opteron(tm) Processor 6320, 2.80GHz, 128GB Memory).
*Multiple threads function was not supported in pTop 1.2.

threads. The main reason is that it cannot identify truncated proteins and unknown modifications. Thus, its search space is smaller than pTop 2.0 and TopPIC. As shown in Table 6, even if one unexpected modification is allowed, the search time of pTop 2.0 does not increase too much, especially with three or more threads, consuming only 2 more minutes.

We also tested the running time of pTop 2.0 and TopPIC with 8, 12, 16, and 20 threads on H2A data set. The running time and speedup ratio of pTop 2.0 and TopPIC with multi-threading are shown in Fig. 11. For single thread, the speed of pTop 2.0 searching with ten variable modifications and one unexpected modification is about 5 times faster than TopPIC searching with two unexpected modifications. For eight and twenty threads, the speed of pTop 2.0 is about 20 and 42 times faster than TopPIC, respectively. When the number of threads increases from 3 to 20, the speedup of pTop 2.0 varies from 2.7 to 11.9 with its efficiency decreasing from 91% to 60%, and the speedup of TopPIC varies from 1.1 to 1.7 with its efficiency decreasing from 38% to 8%. In summary, the speed of pTop 2.0 with six threads is comparable to that of pTop 1.2 with single thread. In multi-threading mode, the speedup as well as the efficiency of pTop 2.0 is higher than that of TopPIC.

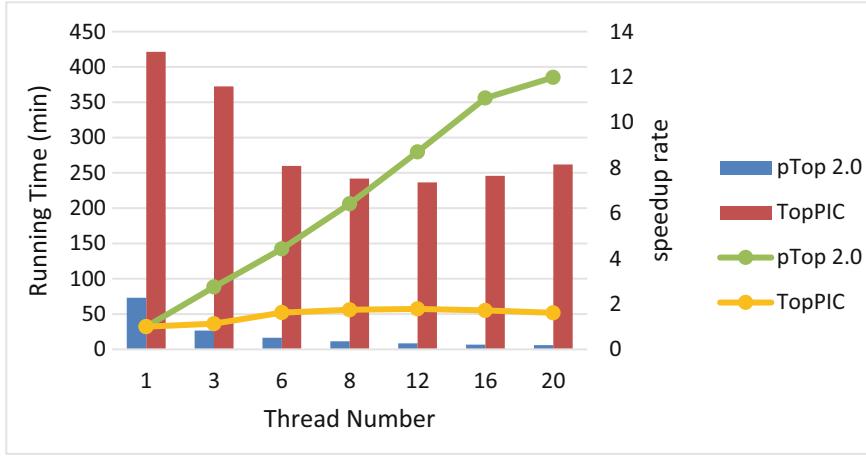


Fig. 11 Running time and speed-up ratio of pTop 2.0 and TopPIC with multi-threading on H2A data

5 Conclusion and Discussions

In this study, we presented a novel software tool pTop 2.0, which can more accurately identify and quantify complex proteoforms with primary structure alterations such as amino acid mutations, terminal truncations, and posttranslational modifications. The semi-supervised machine learning approach is introduced in TDP to re-rank the search results of pTop 2.0, thus improving the sensitivity of the engine significantly. Tests on both simple and complex sample data sets showed that pTop 2.0 can identify more spectra and proteoforms with more matched fragment ions and higher matched peaks. Compared with pTop 1.2, pTop 2.0 could identify truncated proteins and one unknown modification besides variable modifications, which makes its search space much larger. And for spectra which cannot be identified by tag-based process, pTop 2.0 would search them a second time with ion-index process. TopPIC could search with at most two unknown modifications whose advantage is to identify unknown modifications that are not in Unimod.

pTop 2.0 integrated pQuant as the quantitation module, which enables it to accurately quantify intact proteoforms based on the MS data. Test on ^{15}N-labeled yeast data set showed that pTop2.0 could automatically search multiplex-labeled data and call pQuant to compute relative quantitation ratios.

The isotopic labeling strategy and entrapment strategy, which both indicated that the quality control of pTop 2.0 was relatively strict and more accurate, thus have provided a new insight on how to assess and validate the spectra-level FDR of a search engine. However, the accurate estimation and validation of FDR of identified proteoforms and modifications is still a challenging open

problem. As for proteoform-level FDR estimation, a variety of PSAs may have happened on the identified proteoforms compared with the database protein sequences, making it hard to generate decoy proteoforms.

To improve the performance of pTop 2.0, we also investigated the fragmentation patterns of large-scale proteoforms using the data set from ref. [32]. Our result shows that the quality of PrSMs depends heavily on the sequence characteristics [33]. We will in future incorporate these new features into pTop. For the identification of truncated proteins, it has attracted more and more attentions in the pharmaceutical industry [34–36]. We will improve the performance of pTop 2.0 to identify more truncations, such as more complex species in its future versions.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No. 31670837). The authors thank Hao Yang, Wen-Jing Zhou, Xiao-Jin Zhang, and Zhao-Wei Wang from the Institute of Computing Technology; Zhe-Yi Liu and Fang-Jun Wang from Dalian Institute of Chemical Physics, Chinese Academy of Sciences; and Yong Cao and Meng-Qiu Dong from National Institute of Biological Sciences, Beijing, for MS Data acquisition and valuable discussions to improve this chapter.

References

1. Aebersold R, Agar JN, Jonathan Amster I, Baker MS, Bertozzi CR, Boja ES, Costello CE, Cravatt BF, Fenselau C, Garcia BA, Ge Y, Gunawardena J, Hendrickson RC, Hergenrother PJ, Huber CG, Ivanov AR, Jensen ON, Jewett MC, Kelleher NL, Kiessling LL, Krogan NJ, Larsen MR, Loo JA, Ogorzalek Loo RR, Lundberg E, MacCoss MJ, Mallick P, Mootha VK, Mrksich M, Muir TW, Patrie SM, Pesavento JJ, Pitteri SJ, Rodriguez H, Saghatelian A, Sandoval W, Schlüter H, Sechi S, Slavoff SA, Smith LM, Snyder MP, Thomas PM, Uhlén M, Van Eyk JE, Vidal M, Walt DR, White FM, Williams ER, Wohlschlager T, Wysocki VH, Yates NA, Young NL, Zhang B (2018) How many human proteoforms are there? Nat Chem Biol 14:206–214
2. Smith LM, Kelleher NL (2018) Proteoforms as the next proteomics currency. Science 359(6380):1106–1107
3. Naryzhny S (2019) Inventory of proteoforms as a current challenge of proteomics: some technical aspects. J Proteomics 191:22–28
4. Schaffer LV, Shortreed MR, Cesnik AJ, Frey BL, Solntsev SK, Scalf M, Smith LM (2018) Expanding proteoform identifications in top-down proteomic analyses by constructing proteoform families. Anal Chem 90: 1325–1333
5. Schaffer LV, Anderson LC, Butcher DS, Shortreed MR, Miller RM, Pavelec C, Smith LM (2021) Construction of human proteoform families from 21 tesla Fourier transform ion cyclotron resonance mass spectrometry top-down proteomic data. J Proteome Res 20:317–325
6. LeDuc RD, Schwämmle 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 (2018) ProForma: a standard proteoform notation. J Proteome Res 17:1321–1325
7. Smith LM, Thomas PM, Shortreed MR, Schaffer LV, Fellers RT, LeDuc RD, Tucholski T, Ge Y, Agar JN, Anderson LC, Chamot-Rooke-J, Gault J, Loo JA, Paša-Tolić L, Robinson CV,

Schlüter H, Tsybin YO, Vilaseca M, Vizcaíno JA, Danis PO, Kelleher NL (2019) A five-level classification system for proteoform identifications. Nat Methods 16:939–940

8. Donnelly DP, Rawlins CM, DeHart CJ, Fornelli L, Schachner LF, Lin Z, Lippens JL, Aluri KC, Sarin R, Chen B, Lantz C, Jung W, Johnson KR, Koller A, Wolff JJ, Campuzano IDG, Auclair JR, Ivanov AR, Whitelegge JP, Paša-Tolić L, Chamot-Rooke J, Danis PO, Smith LM, Tsybin YO, Loo JA, Ge Y, Kelleher NL, Agar JN (2019) Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. Nat Methods 16:587–594
9. McCool EN, Lubeckyj RA, Shen X, Chen D, Kou Q, Liu X, Sun L (2018) Deep top-down proteomics using capillary zone electrophoresis-tandem mass spectrometry: identification of 5700 proteoforms from the *Escherichia coli* proteome. Anal Chem 90: 5529–5533
10. Shen X, Yang Z, McCool EN, Lubeckyj RA, Chen D, Sun L (2019) Capillary zone electrophoresis-mass spectrometry for top-down proteomics. Trends Analyt Chem 120. pii: 115644
11. Xu T, Shen X, Yang Z, Chen D, Lubeckyj RA, McCool EN, Sun L (2020) Automated capillary isoelectric focusing-tandem mass spectrometry for qualitative and quantitative top-down proteomics. Anal Chem 92: 15890–15898
12. Chen W, Liu X (2021) Proteoform identification by combining RNA-Seq and top-down mass spectrometry. J Proteome Res 20: 261–269
13. Wu Z, Roberts DS, Melby JA, Wenger K, Wetzel M, Gu Y, Ramanathan SG, Bayne EF, Liu X, Sun R, Ong IM, McIlwain SJ, Ge Y (2020) MASH explorer: a universal software environment for top-down proteomics. J Proteome Res 19:3867–3876
14. Brown KA, Melby JA, Roberts DS, Ge Y (2020) Top-down proteomics: challenges, innovations, and applications in basic and clinical research. Expert Rev Proteomics 17(10): 719–733
15. Smith LM, Kelleher NL, Consortium for Top Down Proteomics (2013) Proteoform: a single term describing protein complexity. Nat Methods 10(3):186–187
16. Horn DM, Zubarev RA, McLafferty FW (2000) Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. J Am Soc Mass Spectrom 11(4):320–332
17. Mayampurath AM, Jaitly N, Purvine SO, Monroe ME, Auberry KJ, Adkins JN, Smith RD (2008) DeconMSn: a software tool for accurate parent ion monoisotopic mass determination for tandem mass spectra. Bioinformatics 24(7):1021–1023
18. Liu X, Inbar Y, Dorrestein PC, Wynne C, Edwards N, Souda P, Whitelegge JP, Bafna V, Pevzner PA (2010) Deconvolution and database search of complex tandem mass spectra of intact proteins: a combinatorial approach. Mol Cell Proteomics 9(12):2772–2782
19. Sun RX, Luo L, Wu L, Wang RM, Zeng WF, Chi H, Liu C, He SM (2016) pTop 1.0: a high-accuracy and high-efficiency search engine for intact protein identification. Anal Chem 88(6): 3082–3090
20. LeDuc RD, Taylor GK, Kim YB, Januszyk TE, Bynum LH, Sola JV, Garavelli JS, Kelleher NL (2004) ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry. Nucleic Acids Res 32(Web Server issue): W340–W345
21. Zamdborg L, LeDuc RD, Glowacz KJ, Kim YB, Viswanathan V, Spaulding IT, Early BP, Bluhm EJ, Babai S, Kelleher NL (2007) ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. Nucleic Acids Res 35(Web Server issue):W701–W706
22. Liu X, Sirotkin Y, Shen Y, Anderson G, Tsai YS, Ting YS, Goodlett DR, Smith RD, Bafna V, Pevzner PA (2012) Protein identification using top-down spectra. Mol Cell Proteomics 11(6):M111 008524
23. Liu X, Hengel S, Wu S, Tolic N, Pasa-Tolic L, Pevzner PA (2013) Identification of ultramodified proteins using top-down tandem mass spectra. J Proteome Res 12(12):5830–5838
24. Kou Q, Xun L, Liu X (2016) TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. Bioinformatics 32(22):3495–3497
25. Park J, Piehowski PD, Wilkins C et al (2017) Informed-proteomics: open-source software package for top-down proteomics. Nat Methods 14:909–914
26. Liu C, Song CQ, Yuan ZF, Fu Y, Chi H, Wang LH, Fan SB, Zhang K, Zeng WF, He SM, Dong MQ, Sun RX (2014) pQuant improves quantitation by keeping out interfering signals and evaluating the accuracy of calculated ratios. Anal Chem 86(11):5286–5294
27. Yuan ZF, Liu C, Wang HP, Sun RX, Fu Y, Zhang JF, Wang LH, Chi H, Li Y, Xiu LY,

Wang WP, He SM (2012) pParse: a method for accurate determination of monoisotopic peaks in high-resolution mass spectra. Proteomics 12(2):226–235

28. Creasy DM, Cottrell JS (2004) Unimod: protein modifications for mass spectrometry. Proteomics 4(6):1534–1536
29. LibLinear. https://www.csie.ntu.edu.tw/~cjlin/liblinear/
30. Tian Z, Tolic N, Zhao R, Moore RJ, Hengel SM, Robinson EW, Stenoien DL, Wu S, Smith RD, Pasa-Tolic L (2012) Enhanced top-down characterization of histone post-translational modifications. Genome Biol 13(10):R86
31. Xiao K, Yu F, Fang H, Xue B, Liu Y, Tian Z (2015) Accurate and efficient resolution of overlapping isotopic envelopes in protein tandem mass spectra. Sci Rep 5:14755
32. Anderson LC, DeHart CJ, Kaiser NK, Fellers RT, Smith DF, Greer JB, LeDuc RD, Blakney GT, Thomas PM, Kelleher NL, Hendrickson CL (2017) Identification and characterization of human proteoforms by top-down LC-21 tesla FT-ICR mass spectrometry. J Proteome Res 16(2):1087–1096
33. Sun R-X, Wang R-M, Chi H, Liu C, He S-M, Ge Y (2017) Statistical fragmentation pattern discovery of intact proteins based on their large-scale top-down MS/MS spectra, the 65th ASMS conference, 4–8 June 2017, Indianapolis
34. Liu Z, Wang R, Liu J, Sun R, Wang F (2019) Global quantification of intact proteins via chemical isotope labeling and mass spectrometry. J Proteome Res 18:2185–2194
35. Chen D, Geis-Asteggiante L, Gomes FP, Ostrand-Rosenberg S, Fenselau C (2019) Top-down proteomic characterization of truncated proteoforms. J Proteome Res 18: 4013–4019
36. Zhang Z, Hug C,Tao Y, Bitsch F, Yang Y (2021) Solving complex biologics truncation problems by top-down mass spectrometry. J Am Soc Mass Spectrom 32(8):1928–1935

Chapter 10

Proteoform Identification and Quantification Using Intact Protein Database Search Engine ProteinGoggle

Suideng Qin and Zhixin Tian

Abstract

Proteomics studies the proteome of organisms, especially proteins that are differentially expressed under certain physiological or pathological conditions; qualitative identification of protein sequences and post-translational modifications (PTMs) and their positions can help us systematically understand the structure and function of proteoforms. With the development and relative popularity of soft ionization technology (such as electrospray ionization technology) and high mass measurement accuracy and high-resolution mass spectrometers (such as orbitrap), the mass spectrometry (MS) characterization of complete proteins (the so-called top-down proteomics) has become possible and has gradually become popular. Corresponding database search engines and protein identification bioinformatics tools have also been greatly developed. This chapter provides a brief overview of intact protein database search algorithm "isotopic mass-to-charge ratio and envelope fingerprinting" and search engine ProteinGoggle.

Key words Proteoform, Identification, Quantification, Posttranslational modification, Cross-talk, Localization

1 Introduction

The proteome, first proposed by Wilkins et al. [1] in 1994, refers to a collection of proteins expressed by biological cells, tissues, or organs; it is used to describe the protein counterpart of the genome. Proteomics is the study of the proteome, which determines the state of all proteins in a specific organism, including expression, quantification, modification, mutation, etc. The identification of protein sequences and the determination of protein PTMs are of great significance for the systematic understanding of protein structure and function.

Proteomics is widely used in the research and discovery of biomarkers to find characteristic proteins for disease prevention and diagnosis. The systematic and precise quantification of differentially expressed proteins in physiological and pathological states is the major goal of MS-based proteomics.

Liangliang Sun and Xiaowen Liu (eds.), *Proteoform Identification: Methods and Protocols*,
Methods in Molecular Biology, vol. 2500, https://doi.org/10.1007/978-1-0716-2325-1_10,
© The Author(s), under exclusive license to Springer Science+Business Media, LLC, part of Springer Nature 2022

The bottom-up proteomics method [2–4] cuts proteins into peptides and uses MS to analyze the peptides to identify the proteins. It has been very mature and has achieved great success in the identification of various proteins and proteomes. There are some limitations in protein characterization through unique peptide identification. For example, if a peptide is shared by multiple proteins, it cannot be used to quantify any one of the proteins. Top-down proteomics [5, 6] directly performs MS analysis on intact proteins, which can achieve 100% sequence coverage and complete protein characterization. The relative popularity of high mass measurement accuracy and high-resolution orbitrap MS has greatly promoted the development of top-down proteomics; corresponding database search engines and protein identification bioinformatics tools have also been largely developed. This chapter reports the "isotopic mass-to-charge ratio and envelope fingerprinting" search algorithm and protein database search engine ProteinGoggle.

1.1 New Features of ESI Mass Spectra of Intact Proteins

In the late 1980s, two new soft ionization techniques appeared, namely electrospray ionization (ESI) invented by Fenn et al. [7] and matrix-assisted laser desorption ionization (MALDI) invented by Karas et al. [8]. These two technologies solve the problems of ionization of highly polar, thermally unstable proteins and peptide molecules and MS analysis of biological macromolecules.

The characteristic of the electrospray ion source is to generate multi-charged molecular ions, which reduces the mass-to-charge ratio to the range that most mass analyzers can detect, thus greatly expanding the analysis range of molecular weight. The true molecular mass together with the charge state of the ion can alse be calculated based on the mass-to-charge ratio. With the advantage that it can be easily combined with a variety of separation techniques, electrospray MS is more commonly used. The emergence of electrospray makes MS analysis and qualitative identification of intact proteins possible.

The MS analysis of the intact protein is called the top-down method, that is, the protein is directly detected by the MS without predigestion in sample preparation. Protein ions pass through different dissociation channels to obtain different dissociation fragments, thereby enabling qualitative identification of amino acid sequences (also covering PTMs, varied amino acids, and their positions). Therefore, the top-down method directly uses intact proteins as the research object, which can display more accurate and richer biological information of intact proteins, especially for the identification of PTMs. Compared with the EI or CI mass spectra of small inorganic molecules, small organic molecules, peptides, and metabolites, protein ESI MS has two new features: (1) proteins appear in high valence states and (2) monoisotopic peaks are often no longer directly observed experimentally.

1.2 The Deconvolution (Deisotoping) Approach

To accommodate the new features of protein mass spectra, especially the fact that monoisotopic peaks are normally not observed any more, one solution is to deconvolute experimental isotopic envelopes using model molecules, such as averagine which stands for average amino acid. Based on fingerprinting of deconvoluted monoisotopic masses, lots of top-down search engines are developed such as MS-Top down [9], MS-Align+ [10], MS-Align-E [11], Big Mascot [12, 13], PIITA [14], and recent pTop [15]. In addition, SEQUEST [16] and OMMSA [17] have also been adapted for identification of intact proteins. These search engines have made large-scale identification of proteoforms possible and greatly pushed the field forward.

For the overall protein database search, there is only one commercial search engine so far, namely ProSightPC of Thermo [18–21]. ProSightPC uses the THRASH deisotope method to approximate the isotope profile data directly measured by MS to the monoisotopic mass before database search.

ProSightPTM is the first search engine for protein top-down identification [20]. It uses a method called "Shotgun Annotation" [20] to generate a theoretical protein variant database. In fact, it is based on the existing annotation information in the database to combine the variant forms of the protein in the database in an enumerated manner to generate all possible proteoforms including amino acid mutations, alternative splicing, and PTMs. This method greatly reduces the burden of matching and scoring protein variants with their mass spectra.

ProSightPTM comes with three search modes [20]: absolute mass search, sequence tag search (called biomarker search in ProSightPC), and a combination of the two modes. ProSightPTM uses a Poisson distribution probability scoring model [22] to evaluate the credibility of the matching results. ProSightPC is the commercialized version of ProSightPTM. In addition, the newly developed ProSight Lite [18] as a single protein search tool can be used for the study of unknown protein modifications.

ProSightPC adopts the method of isotopic removal. The isotope profile data directly measured by the MS is approximated to the monoisotopic mass before database search. The subsequent quality fingerprint comparison is only a comparison between two numbers and is very efficient. As such, top-down search engines utilizing the deisotoping and monoisotopic masses have been widely used in proteoform identification of various biological systems. Yet, besides often missing monoisotopic peaks, the other issues associated with the new features of protein mass spectra have not been fully addressed by the deconvolution approach and need alternative approaches.

Due to the sensitivity of the isotope profile to the composition of amino acids, the deisotoping method cannot accurately deisotope proteins with a relative molecular mass of more than 10,000

(equivalent to 100 amino acids) [23]. First of all, the error rate of the monoisotopic mass obtained by the deisotopic method based on the data preprocessing of the model molecule is often high, reaching more than 40%; at the same time, due to the different models in the different preprocessing methods, the results of the respective analysis are not comparable and sharing [24, 25]. Secondly, due to the large protein molecules, the number of fragment ions during dissociation and the different valences of the same ion are very large, which directly leads to the crowding of the MS/MS spectrum, and also the density of overlapping data is very large for the same reason; the deisotoping method is not known due to the unknown pretreatment processes, so these overlapping data cannot be well separated and resolved [24, 25]. Thirdly, the complete contour information is completely lost in the process of isotopic removal, so it is impossible to effectively distinguish ideal high-confidence data from non-ideal low-confidence data, as well as to guarantee the reproducibility and reliability of protein identification [24, 25].

Altogether, for the two new features of electrospray protein MS, new analytical methods must be developed to overcome the bottlenecks and many problems that have not been fully addressed by the deconvolution approach.

2 The IMEF Search Algorithm and Workflow

2.1 Isotopic Mass-to-Charge Ratio and Envelope Fingerprinting [26]

There are many problems with the traditional method of deisotopic processing on the isotope profile directly measured by MS. ProteinGoggle abandons the process of isotopic removal and directly analyzes the mass spectrum in situ based on the original mass spectrum data. ProteinGoggle uses (develops) a new algorithm, isotopic mass-to-charge ratio, and envelope fingerprinting (iMEF).

The principle of isotopic envelope fingerprinting is shown in Fig. 1. The vertical line in the figure is the experimental value, and the circle is the theoretical value. The numbers annotated above each isotope peak in the figure are the *m/z* deviation and relative intensity deviation of each experimental isotope peak. IPACO (Isotopic Peak Abundance Cutoff) is the isotope peak intensity threshold. Therefore, if all experimental isotope peaks of an ion above the threshold are observed, and the isotopic *m/z* deviation (IPMD) and isotopic peak abundance deviation (IPAD) are less than or equal to the specified search parameters, this ion is a matched ion; if any of the above criteria is not met, the ion is a non-matched ion.

The advantages of this algorithm are as follows: (1) the process of removing isotope of the original MS data and the problems or errors introduced during the process are eliminated, and the in situ analysis based on the original data has a higher credibility; (2) each ion is no longer a (mere) comparison of the mass of a single isotope,

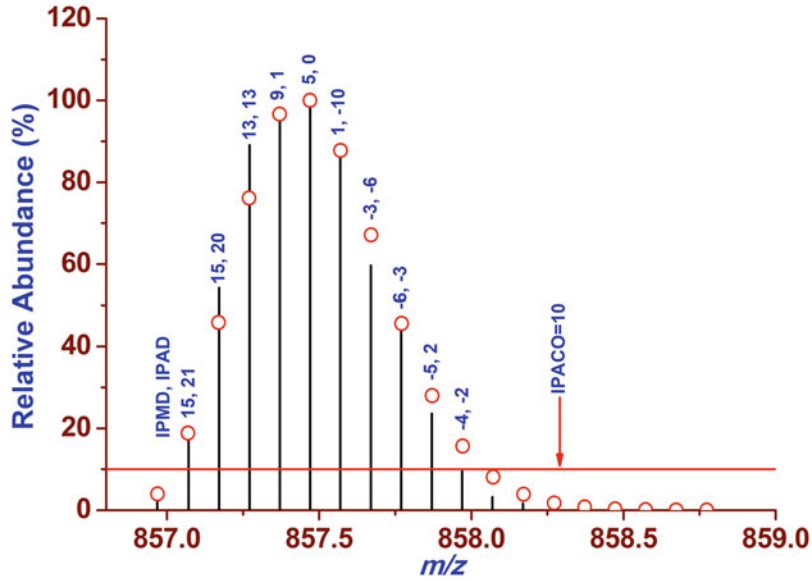


Fig. 1 Intepretation and identification of ubiquitin precursor ion ($z = 10$) using isotopic envelope fingerprinting. Bar, experimental data; circle, theoretical

but a comparison of the entire profile, and the obtained result is more reliable; (3) the isotope profile comparison map of each ion is outputted so that the user can perform manual verification more conveniently.

2.2 Comprehensive Coverage of Dissociation Methods and Pathways

ProteinGoggle supports common collision-based, electron-based, and photon-based dissociation methods for secondary MS data analysis (the graphical user interfaces are shown in Figs. 2, 3, and 4). Collision-based dissociation methods mainly include collision-induced dissociation and high-energy collision-induced dissociation. Apart from the most common b- and y-type ions, a-, b- and y-type ions with 1–2 neutral molecules lost, internal dissociation yb-, ya-type ions and corresponding yb-, ya-type ions with neutral loss are taken into consideration to fully cover all the dissociation channels of the protein. Electron-based dissociation methods mainly include electron transfer dissociation and electron capture dissociation, generating 4 major types of fragment ions including c-type, z-type, c-type with neutral loss and z-type with neutral loss. The photon-based dissociation methods mainly include ultraviolet multiphoton dissociation, etc. The ion types are more complex and diverse, mainly including a, b, c, x, y, z-type ions, and a, b, c, x, y, z-type ions appear as ion types in which 1–2 neutral molecules are lost. In addition, there are specific types of ions such as a + 1, x + 1, and y − 1.

Through the effective treatment of all the above dissociation methods and dissociation ion types, ProteinGoggle achieves a comprehensive coverage of protein dissociation channels.

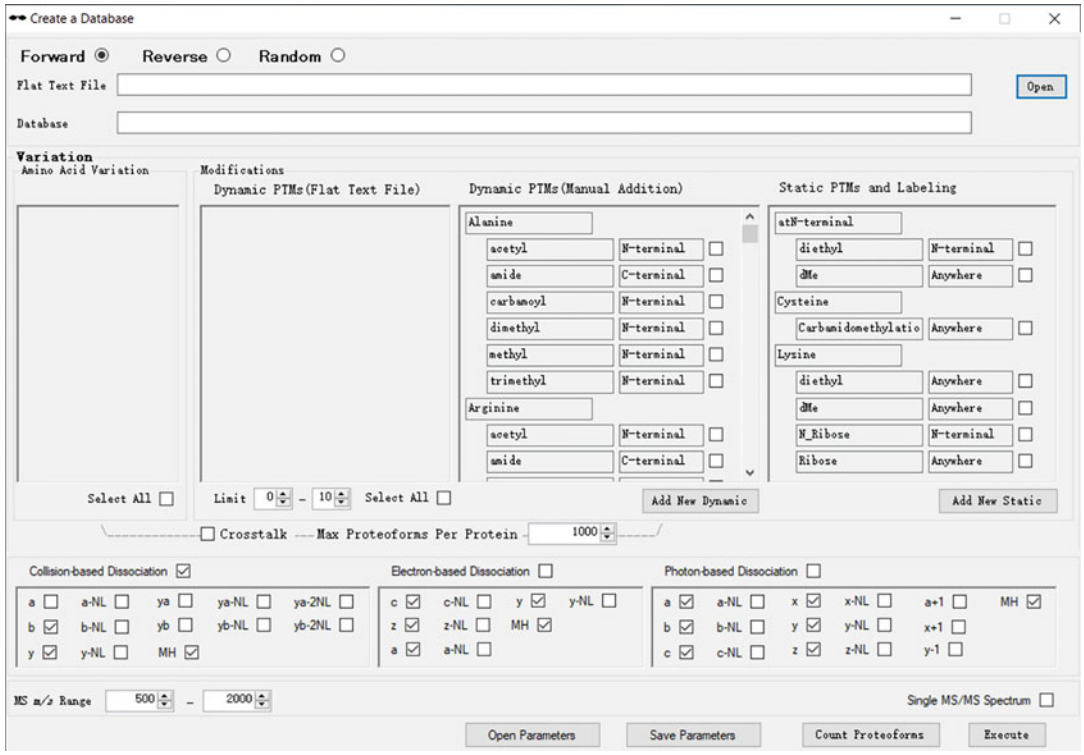


Fig. 2 Creation of an intact protein database in ProteinGoggle

2.3 Overlapping Isotopic Envelopes

For overall protein MS, the problem of ion overlap becomes more prominent and important due to the generally higher ion valence, larger number of ions, and increasing complexity of the spectrum. Either partial overlap of ions or complete overlap of ions will seriously affect the matching of ions because the overlap of ions will cause serious deformation of the isotope profile, and the result is that the isotope profile cannot be correctly matched.

ProteinGoggle uses the overlap analysis algorithm of OIE_-CARE [27] (Fig. 5) to use the highest peak of the theoretical isotope corresponding to the overlapped isotope profile as a reference to realize the assignment of the peak intensity and the accurate discrimination of the isotope profile. The corrected isotope profile is resolved and can get a good match.

2.4 Neutral Loss and Internal Fragment Ions

Neutral loss and internal product ions are common during proteoform fragmentation. The utility of these two types of fragment ions for intact protein identification is investigated at both the proteome level using RPLC-MS/MS analysis of *E. coli* proteome and the protein level using selected *E. coli* proteins [28] and a public dataset of *Pseudomonas aeruginosa* PAO1. Because of the significant random match probability, both neutral loss and internal product ions

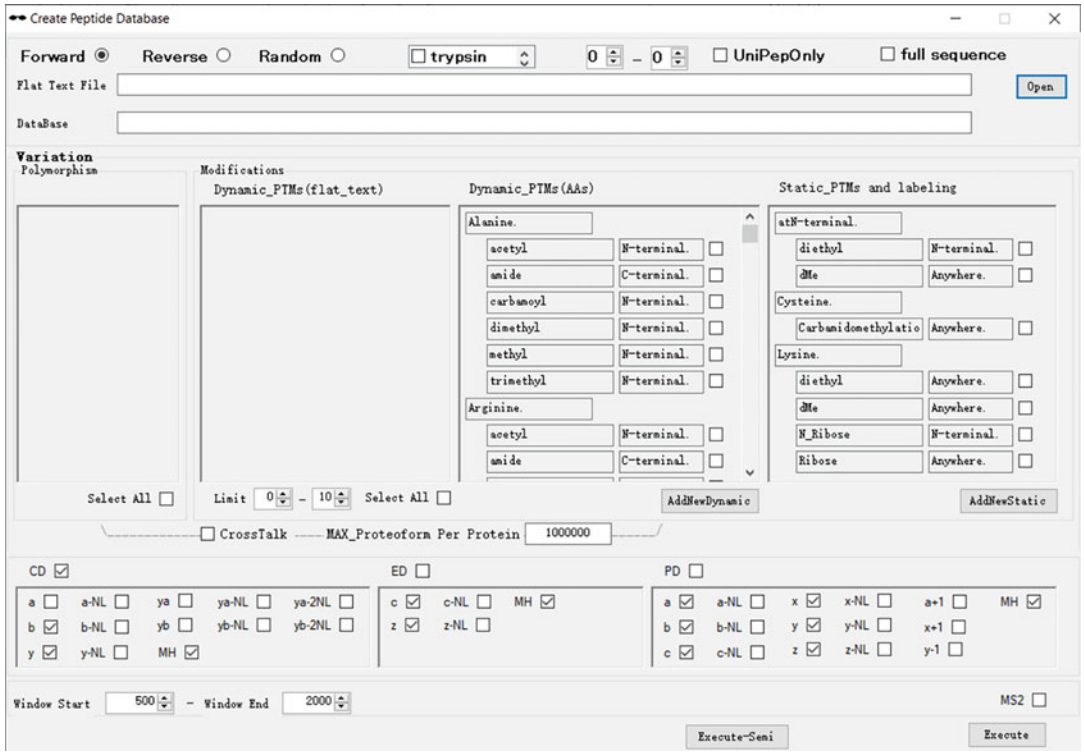


Fig. 3 Creation of a peptide database in ProteinGoggle

are found to have no direct utility for protein identification with target-decoy searches for false discovery rate control. However, the indirect utility of these types of product ions is not negligible because of their contribution to more matching canonical product ions after resolution of overlapping product ions. Every matched product ion contributes to protein identification and potentially PTM localization.

2.5 The Intact Protein Database Search Engine ProteinGoggle [26, 29]

The ProteinGoggle search engine based on the iMEF algorithm first selects candidate protein sequences with MS spectra and then compares the highest isotopic peak mass-to-charge ratio of the precursor ion corresponding to each MS/MS spectra with the corresponding theoretical database (in which theoretical values are stored). If the comparison is successful, the entire experimental and theoretical isotope profile of the ion is compared to determine the candidate protein. Simply put, isotope profile comparison is to compare the mass-to-charge ratio and relative abundance of each isotope peak of two ions. Each ion in the experimental MS/MS spectra is compared with the theoretical fragment ion generated by each candidate protein sequence using an isotope profile fingerprint comparison algorithm to select the best matching protein from the candidate proteins, which then becomes the protein ID.

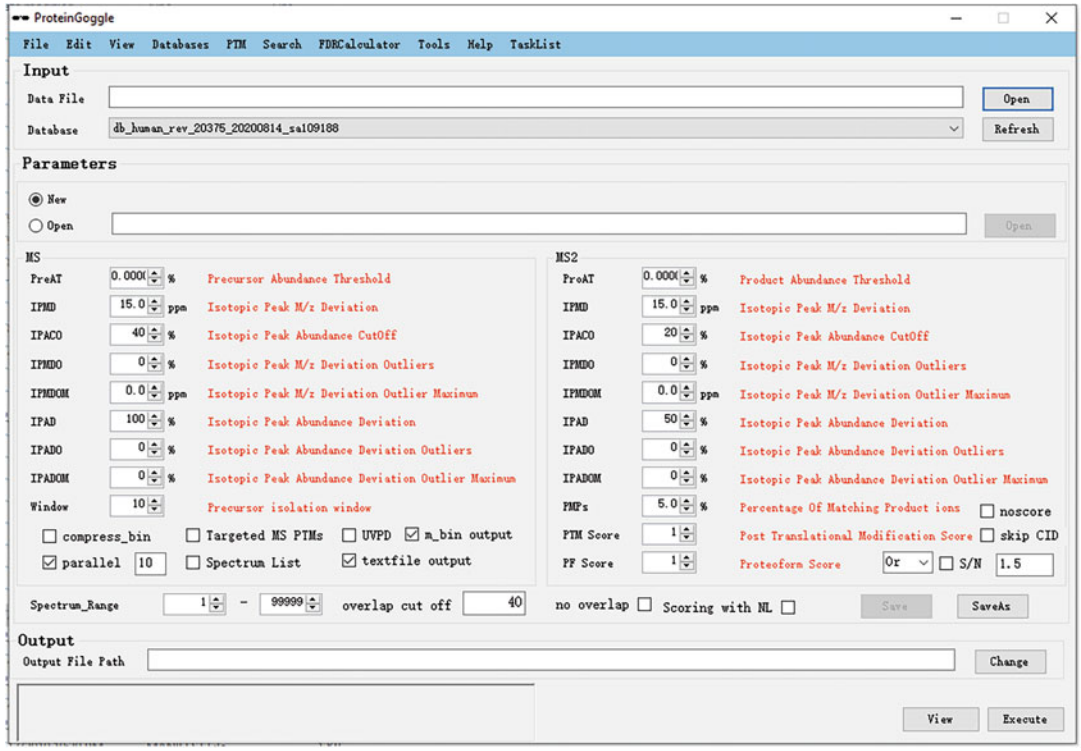


Fig. 4 Run-search graphical user interface in ProteinGoggle

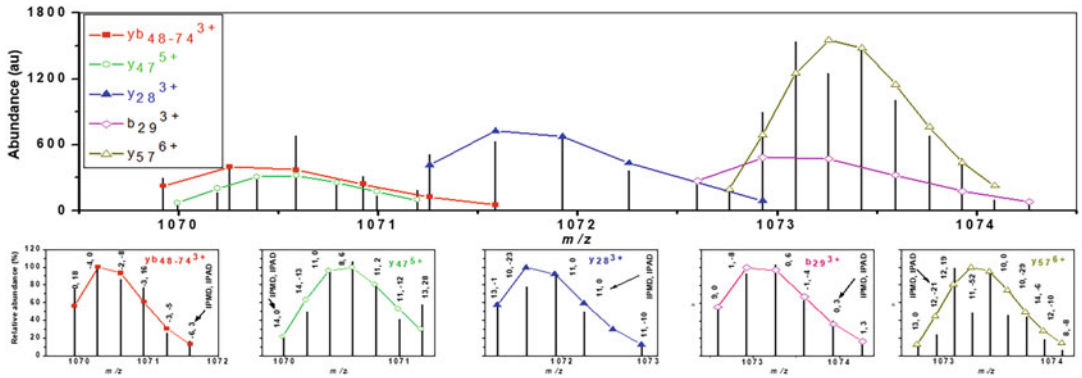


Fig. 5 The main search graphical user interface of ProteinGoggle

3 Application of ProteinGoggle in Proteoform Identification

One strength of ProteinGoogle with iMEF search algorithm is that there is no limit of protein size, because experimental isotopic envelopes are interpreted as they are measured. For example, HCD of enolase 1 ($z = 47$) was well searched by ProteinGoggle to give the isotopic envelope fingerprinting map of the precursor ion (Fig. 6, top panel), the annotated MS/MS spectra (Fig. 6,

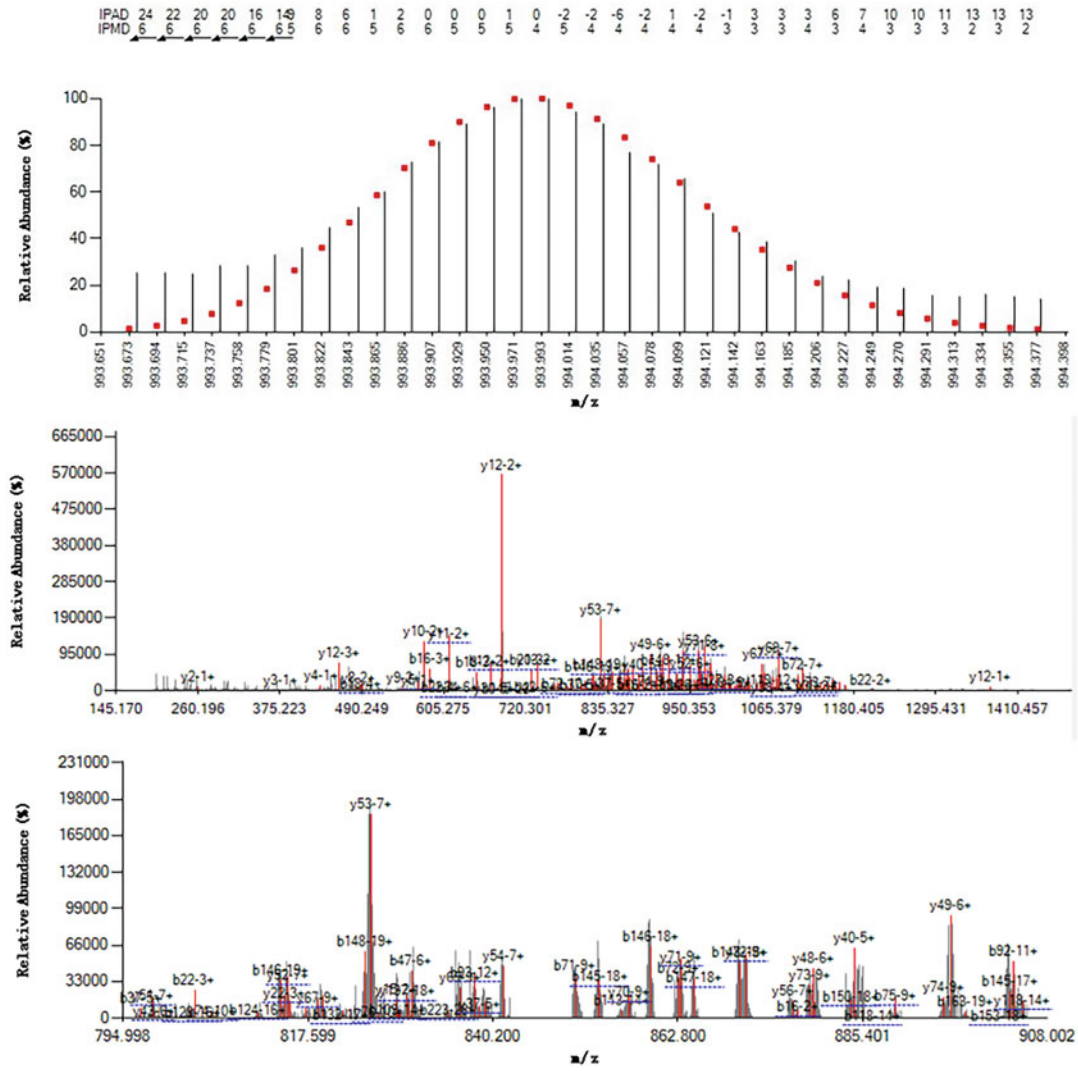


Fig. 6 The annotated MS/MS spectra (middle panel) and pop-up m/z range of 794–908 (bottom panel) from HCD of enolase 1 ($z = 47$, top panel)

middle and bottom panels) with the matched fragment ions, and the graphical fragmentation map (Fig. 7). The sequence length is 436, and 200 matched b/y ions were observed; the percentages of interpreted isotopic peaks and overall abundance are 97.5% and 93.6%, respectively.

Histones are key chromatin structure proteins and are often heavily and densely modified. The need for characterization at the intact protein level to catch the cross-talk of different PTMs on a single proteoform as well as the median sizes make the top-down approach an ideal choice. With its comprehensive handling of dynamic PTMs and creation of proteoforms with all possible combinatorial PTMs, as well as its capability of PTM localization using

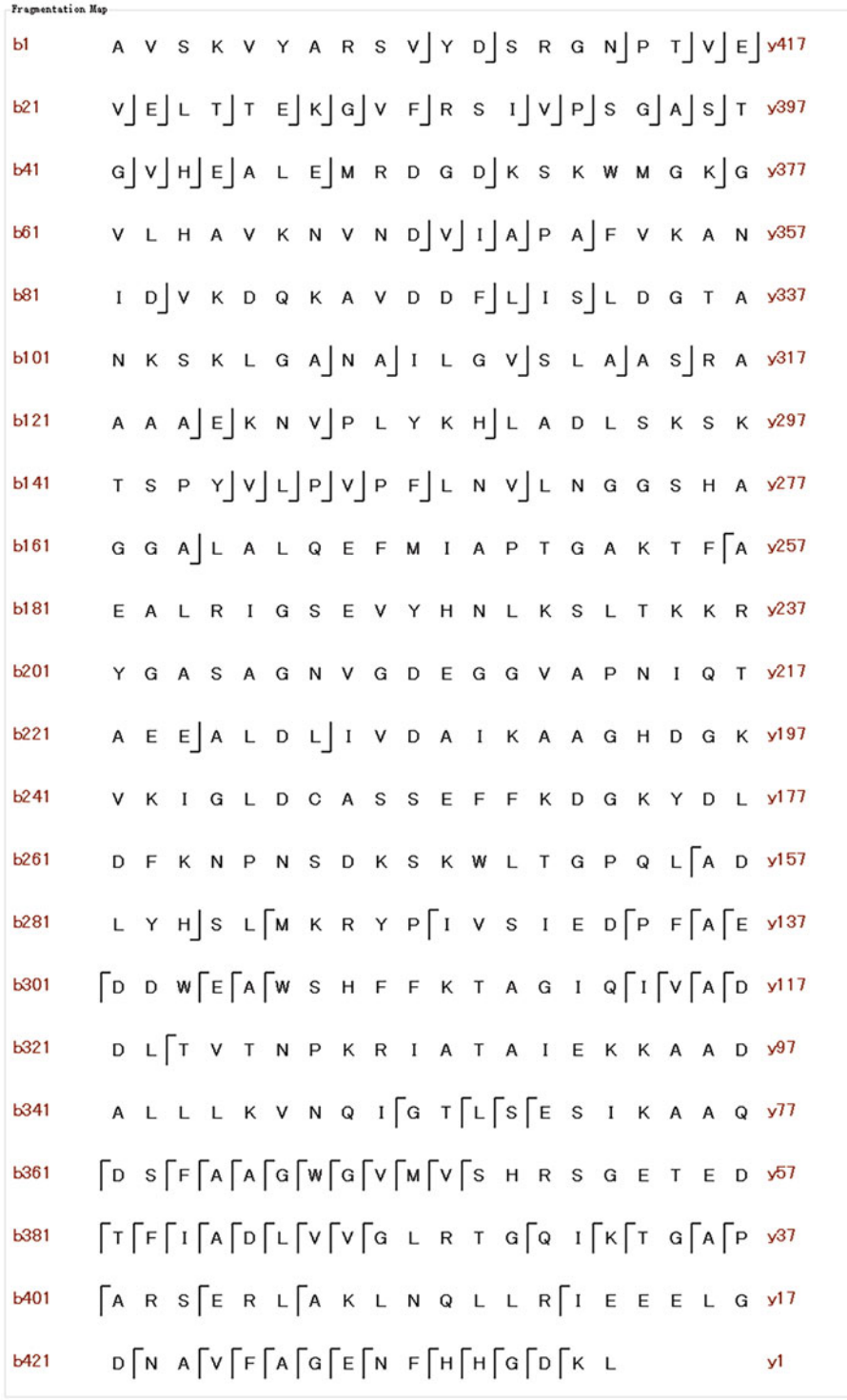


Fig. 7 The graphical fragmentation map of enolase 1 with the matched b and y fragment ions from HCD

site-determining fragment ions, ProteinGoggle is ideal for database search of top-down data of histones. From the WCX-MS/MS (alternative CID and ETD) dataset of HeLa histone H4 family which was pre-fractionated [30] from the core histone mixture by RPLC, 426 proteoforms were confidently identified with a spectrum-level false discovery rate of less than 1%, which represented the most comprehensive H4 proteoforms reported so far.

In RPLC-MS/MS (HCD) analysis of chicken blood core histones [31], 58 proteoforms were identified for the core histone families of H4, H2B, H2A, and H3; PTMs in 33 of these proteoforms were uniquely localized.

Mouse are common model animal. With nanoRPLC-MS/MS analysis and ProteinGoggle database search, 547 protein species were identified with spectrum-level FDR $\leq$1%, where PTMs in 51 protein species were unambiguously localized with PTM scores $\geq$1 [32].

Besides common proteins, ProteinGoggle has also been successfully applied to search for intact N-glycoprotein RNase B as a benchmark study [33], where selective fragmentation of the protein amino acid backbone and N-glycan moiety was achieved; the former was observed in both HCD and EThcD at low collisional energies, and the latter was observed in all the five dissociation methods (CID, ETD, HCD, ETciD, and EThcD) available on the Orbitrap Fusion Lumos Tribrid MS.

4 Application of ProteinGoggle in Proteoform Quantification [34]

Besides qualitative identification, ProteinGoggle also supports quantitative search with the ProteinGoggleQuan module. In top-down quantitation of differentially expressed proteins in hepatocellular carcinoma HepG2 cells (relative to LO2) using RPLC-MS/MS (HCD) and pseudo-isobaric dimethyl labeling, 33 proteoforms were quantified within 10% RSD (Fig. 8).

5 Conclusions and Perspectives

The most basic function of the overall protein database search engine is to analyze the original experimental data provided by the user to obtain a qualitative identification list of the protein's amino acid sequence, PTM, and its position. However, in order to obtain accurate and comprehensive overall protein qualitative and quantitative results, the qualitative and quantitative analyses of electrospray MS still face the following problems that need to be improved or solved: (1) Whether relying on the decoy database for false-positive control can guarantee identification (the accuracy of which is still a controversial issue). The number of IDs obtained

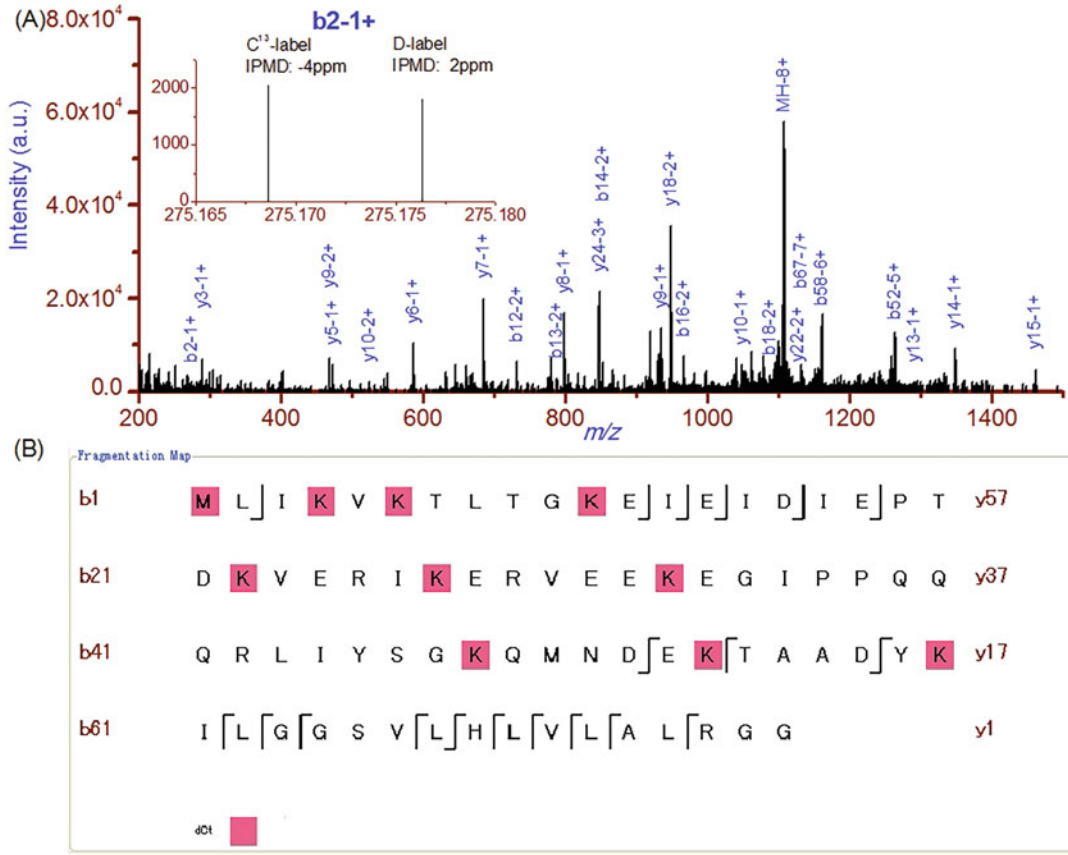


Fig. 8 The annotated MS/MS spectrum (**a**), the pseudo-isobaric b2-1+ ion pair (**a**, inset), and the graphical fragmentation map (**b**) for the quantified protein NEDD8_HUMAN from HepG2 and LO2 cells. Dimethyl-labeled amino acids with amino groups are highlighted with solid red squares. (Adopted from reference [34])

after FDR control seems to be related to the output conditions (tolerance) of the initial PrSMs. How to define it, how to choose the appropriate decoy database, whether it is an anti-database, a random library, or other libraries, and how to choose a suitable scoring system, all these are problems that need to be solved. (2) For proteins with many PTMs, the exhaustive list of proteoforms seems to be necessary because co-elution always exists. But using the exhaustive list will make the number of proteoforms very large, which may heavily reduce the search efficiency. The balance mechanism between the two remains to be studied. (3) Ensuring the resolution rate of the secondary mass spectrum is key to achieving comprehensive identification. How to explain the problem of low resolution, how to search for hydrolyzed protein efficiently and accurately, and whether it can be solved by sequence tag search remain unknown. (4) How to realize the quantitative analysis of the whole protein is still blank in each TD search engine.

References

1. Wilkins MR, Sanchez J-C, Gooley AA, Appel RD, Humphery-Smith I, Hochstrasser DF, Williams KL (1996) Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. Biotechnol Genet Eng Rev 13:19–50
2. Peng JM, Elias JE, Thoreen CC, Licklider LJ, Gygi SP (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. J Proteome Res 2:43–50
3. Chait BT (2006) Mass spectrometry: bottom-up or top-down? Science 314:65–66
4. Foster MS, Edwards MS, Reed DC, Schiel DR, Zimmerman RC (2006) Top-down vs. bottom-up effects in kelp forests. Science 313:1737–1738
5. Lakshmanan R, Wolff JJ, Alvarado R, Loo JA (2014) Top-down protein identification of proteasome proteins with nanoLC-FT-ICR-MS employing data- independent fragmentation methods. Proteomics 14:1271–1282
6. Moradian A, Kalli A, Sweredoski MJ, Hess S (2014) The top-down, middle-down, and bottom-up mass spectrometry approaches for characterization of histone variants and their post-translational modifications. Proteomics 14:489–497
7. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM (1989) Electrospray ionization for mass spectrometry of large biomolecules. Science 246:64–71
8. Karas M, Hillenkamp F (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Anal Chem 60:2299–2301
9. Frank AM, Pesavento JJ, Mizzen CA, Kelleher NL, Pevzner PA (2008) Interpreting top-down mass spectra using spectral alignment. Anal Chem 80:2499–2505
10. Liu XW, Sirotkin Y, Shen YF, Anderson G, Tsai YS, Ting YS, Goodlett DR, Smith RD, Bafna V, Pevzner PA (2012) Protein identification using top-down. Mol Cell Proteomics 11: M111.008524
11. Liu XW, Hengel S, Wu S, Tolic N, Pasa-Tolic L, Pevzner PA (2013) Identification of ultramodified proteins using top-down tandem mass spectra. J Proteome Res 12:5830–5838
12. Kaplan DA, Hartmer R, Speir JP, Stoermer C, Gumerov D, Easterling ML, Brekenfeld A, Kim T, Laukien F, Park MA (2008) Electron transfer dissociation in the hexapole collision cell of a hybrid quadrupole-hexapole Fourier transform ion cyclotron resonance mass spectrometer. Rapid Commun Mass Spectrom 22: 271–278
13. Karabacak NM, Li L, Tiwari A, Hayward LJ, Hong PY, Easterling ML, Agar JN (2009) Sensitive and specific identification of wild type and variant proteins from 8 to 669 kDa using top-down mass spectrometry. Mol Cell Proteomics 8:846–856
14. Tsai YS, Scherl A, Shaw JL, Mackay CL, Shaffer SA, Langridge-Smith PRR, Goodlett DR (2009) Precursor ion independent algorithm for top-down shotgun proteomics. J Am Soc Mass Spectrom 20:2154–2166
15. Sun RX, Luo L, Wu L, Wang RM, Zeng WF, Chi H, Liu C, He SM (2016) pTop 1.0: a high-accuracy and high-efficiency search engine for intact protein identification. Anal Chem 88: 3082–3090
16. Karabacak NM, Li L, Tiwari A, Hayward LJ, Hong P, Easterling ML, Agar JN (2009) Sensitive and specific identification of wild type and variant proteins from 8 to 669 kDa using top-down mass spectrometry. Mol Cell Proteomics 8:846–856
17. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH (2004) Open mass spectrometry search algorithm. J Proteome Res 3:958–964
18. Fellers RT, Greer JB, Early BP, Yu X, Leduc RD, Kelleher NL, Thomas PM (2015) ProSight Lite: graphical software to analyze top-down mass spectrometry data. Proteomics 15:1235–1238
19. Zamdborg L, Leduc RD, Glowacz KJ, Kim YB, Viswanathan V, Spaulding IT, Early BP, Bluhm EJ, Babai S, Kelleher NL (2007) ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. Nucleic Acids Res 35:W701–W706
20. Leduc RD, Taylor GK, Kim YB, Januszyk TE, Bynum LH, Sola JV, Garavelli JS, Kelleher NL (2004) ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry. Nucleic Acids Res 32:W340–W345
21. Wenger CD, Boyne MT, Ferguson JT, Robinson DE, Kelleher NL (2008) Versatile online-offline engine for automated acquisition of high-resolution tandem mass spectra. Anal Chem 80:8055–8063

22. Meng FY, Cargile BJ, Miller LM, Forbes AJ, Johnson JR, Kelleher NL (2001) Informatics and multiplexing of intact protein identification in bacteria and the archaea. Nat Biotechnol 19:952–957
23. Zubarev R (2006) Protein primary structure using orthogonal fragmentation techniques in Fourier transform mass spectrometry. Expert Rev Proteomics 3:251–261
24. Park K, Yoon JY, Lee S, Paek E, Park H, Jung HJ, Lee SW (2008) Isotopic peak intensity ratio based algorithm for determination of isotopic clusters and monoisotopic masses of polypeptides from high-resolution mass spectrometric data. Anal Chem 80:7294–7303
25. Shaw JB, Li WZ, Holden DD, Zhang Y, Griep-Raming J, Fellers RT, Early BP, Thomas PM, Kelleher NL, Brodbelt JS (2013) Complete protein characterization using top-down mass spectrometry and ultraviolet photodissociation. J Am Chem Soc 135:12646–12651
26. Li L, Tian ZX (2013) Interpreting raw biological mass spectra using isotopic mass-to-charge ratio and envelope fingerprinting. Rapid Commun Mass Spectrom 27:1267–1277
27. Xiao KJ, Yu F, Fang HQ, Xue BB, Liu Y, Tian ZX (2015) Accurate and efficient resolution of overlapping isotopic envelopes in protein tandem mass spectra. Sci Rep 5:14755
28. Kaijie-Xiao, Yu F, Fang HQ, Xue BB, Liu Y, Li YH, Tian ZX (2017) Are neutral loss and internal product ions useful for top-down protein identification? J Proteome 160:21–27
29. Xiao KJ, Yu F, Tian ZX (2017) Top-down protein identification using isotopic envelope fingerprinting. J Proteome 152:41–47
30. Xiao KJ, Tian ZX (2016) Top-down characterization of histone H4 proteoforms with ProteinGoggle 2.0. Chin J Chromatogr 34: 1255–1263
31. Wu H, Xiao KJ, Tian ZX (2018) Top-down characterization of chicken core histones. J Proteome 184:34–38
32. Xue BB, Xiao KJ, Tian ZX (2019) Top-down characterization of mouse core histones. J Mass Spectrom 54:258–265
33. Li SS, Zhou Y, Xiao KJ, Li J, Tian ZX (2018) Selective fragmentation of the N-glycan moiety and protein backbone of ribonuclease B on an Orbitrap fusion Lumos Tribrid mass spectrometer. Rapid Commun Mass Spectrom 32: 2031–2039
34. Fang HQ, Xiao KJ, Li YH, Yu F, Liu Y, Xue BB, Tian ZX (2016) Intact protein quantitation using pseudoisobaric dimethyl labeling. Anal Chem 88:7198–7205

Chapter 11

Mass Deconvolution of Top-Down Mass Spectrometry Datasets by FLASHDeconv

Kyowon Jeong, Jihyung Kim, and Oliver Kohlbacher

Abstract

Mass deconvolution, the determination of proteoform precursor and fragment masses, is crucial for top-down proteomics data analysis. Here we describe the detailed procedure to run FLASHDeconv, an ultrafast, high-quality mass deconvolution tool. Both spectrum- and feature-level deconvolution results are obtainable in various output formats by FLASHDeconv. FLASHDeconv is runnable in different environments such as the command line and OpenMS workflows.

Key words Top-down proteomics, Mass deconvolution, Feature deconvolution, Intact protein analysis, Protein characterization

1 Introduction

In mass spectrometry (MS)-based top-down proteomics approach, mass deconvolution, the determination of proteoform precursor and fragment masses, is crucial for downstream data analysis such as proteoform identification, characterization, and quantification [1–4]. However, due to its complex signal structure that stems from the large masses of analytes, fast and accurate mass deconvolution is challenging [5]. Many tools and algorithms have been introduced such as TopFD [6], ProMex [7], Xtract [8], and ReSpect (Positive Probability Ltd), but they often have both runtime and quality issues [9]. Recently Jeong et al. presented FLASHDeconv [9], a deconvolution tool for high-quality mass deconvolution two orders of magnitude faster than existing tools.

Here we describe the detailed procedure to run FLASHDeconv both for MS1 and MS2 spectra. FLASHDeconv performs deconvolution in less than a few milliseconds per spectrum and reports both deconvoluted masses in each spectrum (called *spectrum masses* from here on) and those of MS features (called *feature masses* from here on). As a part of OpenMS software [10], FLASHDeconv runs

Liangliang Sun and Xiaowen Liu (eds.), *Proteoform Identification: Methods and Protocols*,
Methods in Molecular Biology, vol. 2500, https://doi.org/10.1007/978-1-0716-2325-1_11,
© The Author(s), under exclusive license to Springer Science+Business Media, LLC, part of Springer Nature 2022

in different operating systems including Windows, macOS, and Linux and in different environments including the command line and workflows like KNIME, Galaxy, nextflow, and TOPPAS. The protocol described here assumes running FLASHDeconv in the command line. Running FLASHDeconv in other workflows or software is described shortly at the end of this chapter (*see* Subheading 3.6).

2 Materials

1. FLASHDeconv download and supported operating systems
   FLASHDeconv is a part of OpenMS software [10] which is a free software available under the three clause BSD license and runs under Windows, macOS, and Linux. The installation files and source code of FLASHDeconv are found in https://www.openms.de.comp.flashdeconv..
2. Supported input file formats
   FLASHDeconv takes the spectrum files in the mzML format [11]. ProteoWizard [12] msconvert is primarily used to generate mzML files from raw data files. It is recommended to generate the raw file in profile mode from MS instruments since the profile mode spectra generate more sensitive results for MS1 spectra by FLASHDeconv. The conversion method is described in Subheading 3.2.
3. System requirement
   To run FLASHDeconv, a 64-bit computer with at least 4 GB memory is required. 8GB memory space is strongly recommended.

3 Methods

3.1 Installation

Installation of FLASHDeconv is done by running the installation file. After downloading the installation file from https://www.openms.de.comp.flashdeconv., double-click the download to start installing. More information on the installation of FLASHDeconv (or of OpenMS) is found in https://www.openms.de.tutorials..

3.2 Input mzML File Generation

1. ProteoWizard [12] msconvert tool is used to generate input mzML files from raw spectrum files. Except for the peak picking options, the default parameters are recommended.
2. If the spectra are generated in profile mode, either profile or centroid (i.e., peak picked) spectra can be used for MS1 spectra. For MS2 spectra, centroid spectra should be used. To generate centroid spectra both for MS1 and MS2 levels, run the command

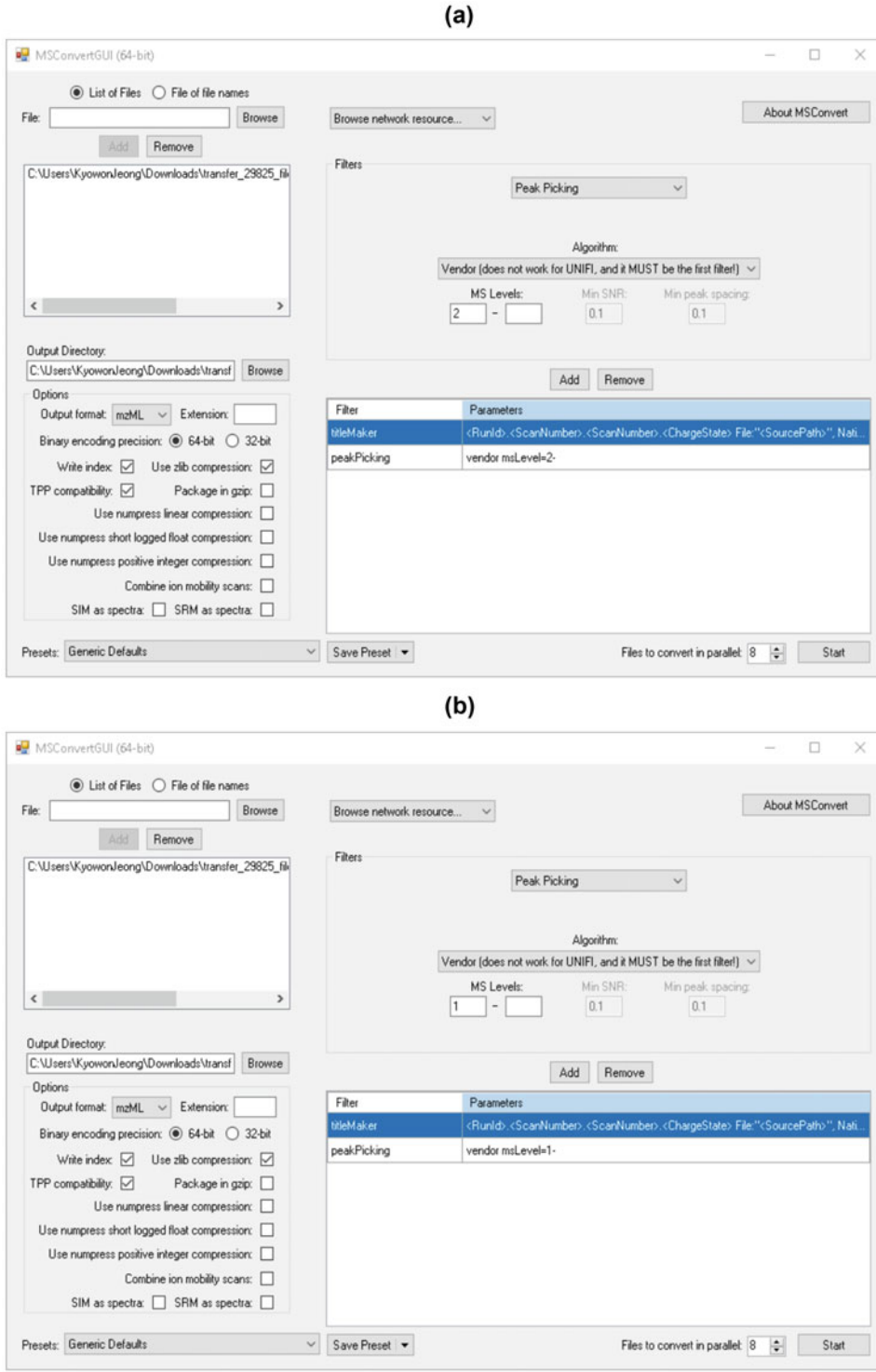


Fig. 1 ProteoWizard msconvert GUI screenshots to convert raw files into mzML files for (**a**) the case in which only MS2 spectra are centroid spectra and (**b**) the case in which both MS1 and MS2 spectra are centroid spectra

```
Msconvert.exe data.RAW --filter "peakPicking true 1-"
```

in a command prompt. To generate centroid spectra only for MS2, run

```
msconvert.exe data.RAW --filter "peakPicking true 2-"
```

in a command prompt (*see* Fig. 1 for screenshots of msconvert GUI usage for the both cases).

3.3 Performing Basic Mass Deconvolution

FLASHDeconv firstly performs the mass deconvolution for each input spectrum, generating spectrum masses. Then, by tracing the spectrum masses along the retention time direction (by a mass trace detection algorithm [13]), the masses of eluted proteoform masses (i.e., feature masses) are found. The feature mass list is the default output of FLASHDeconv. The spectrum masses can be reported in user-specified file formats. The list of all options for FLASHDeconv is shown in Tables 1 (basic options) and 2 (advanced options).

1. After installing FLASHDeconv, go to the directory containing FLASHDeconv binary file (*see* **Note 1**). To see the basic options for FLASHDeconv, type

   ```
   FLASHDeconv.exe --help
   ```

   in a Windows command prompt. To run FLASHDeconv in macOS or Linux, type

   ```
   ./FLASHDeconv --help
   ```

   in a terminal. To see all the options (including advanced ones), type

   ```
   FLASHDeconv.exe --helphelp
   ```

   in a Windows command prompt. To run FLASHDeconv in macOS or Linux, type

   ```
   ./FLASHDeconv --helphelp
   ```

   in a terminal. From here on, we assume that we run FLASHDeconv in a macOS or Linux terminal.

2. To perform the deconvolution, the input and output files should be specified. The input file is the mzML file (*see* Subheading 3.2), and the output file is a tsv file containing the deconvoluted feature masses and relevant information thereof (*see* Table 3). Assume the input mzML file name is `spectrum.mzml`, and the desired output tsv file name is `mass.tsv`. Then, type

   ```
   ./FLASHDeconv -in spectrum.mzml -out mass.tsv
   ```

   in a terminal to run FLASHDeconv with the default parameters. The generated `mass.tsv` contains the tabular data with the column names shown in Table 3. `in` and `out` options are mandatory.

3. In addition to the default output tsv file, the feature masses can be reported in the ProMex [7] output format (ms1ft file) by

Table 1
All basic options of FLASHDeconv. Only `in` and `out` options are mandatory in most cases (*see* Note 4)

Name	Format	Default	Description
`-in`	*.mzML file	–	Input mzML file
`-out`	*.tsv file	–	Feature mass output tsv file
`-out_spec`	*.tsv files	–	Spectrum mass output tsv files (per MS level)
`-out_mzml`	*.mzML file	–	Spectrum mass output mzML file
`-out_promex`	*.ms1ft file	–	Feature mass in ProMex output format
`-out_topFD`	*.msalign files	–	Spectrum mass in TopFD output format (per MS level)
`-mzml_mass_charge`	Integer	0	Charge status of deconvoluted masses in mzML output file
`-write_detail`	0 or 1	0	To write detailed information in spectrum mass tsv files
`-use_ensemble_spectrum`	0 or 1	0	To use ensemble spectrum
`-Algorithm:tol`	ppm values	[10.0 10.0]	Mass ppm tolerance (per MS level)
`-Algorithm:min_mass`	Value	50.0	Minimum mass in Da
`-Algorithm:max_mass`	Value	100000.0	Maximum mass in Da
`-Algorithm:min_charge`	Integer	1	Minimum charge
`-Algorithm:max_charge`	Integer	100	Maximum charge
`-Algorithm: min_isotope_cosine`	Values	[0.75 0.75]	Isotope cosine similarity threshold (per MS level)
`-FeatureTracing: mass_error_da`	Value	1.5	Mass tolerance in Dalton for mass trace detection algorithm
`-FeatureTracing: min_trace_length`	Second	10.0	Minimum feature trace length

specifying `out_promex` option. For example, to generate `mass.ms1ft`, type

`./FLASHDeconv -in spectrum.mzml -out mass.tsv -out_promex mass.ms1ft`. The generated ms1ft file can be used as an input file of MSPathFinder [7] for the proteoform identification (*see* **Note 2**).

4. The deconvoluted spectrum masses can be reported in three formats: tsv, mzML, and TopFD output file format, msalign file (or, TopPIC input format). `out_spec`, `out_mzml`, and

Table 2
All advanced options of FLASHDeconv. The advanced options are displayed with `--helphelp` option

Name	Format	Default	Description
`-max_MS_level`	Integer	2	Maximum MS level
`-use_RNA_averagine`	0 or 1	0	To use RNA averagine model
`-Algorithm:min_mz`	Value	−1.0	Minimum m/z
`-Algorithm:max_mz`	Value	−1.0	Maximum m/z
`-Algorithm:min_RT`	Second	−1.0	Minimum retention time
`-Algorithm:max_RT`	Second	−1.0	Maximum retention time
`-Algorithm:min_peaks`	Integers	[3 1]	Minimum supporting peak count (per MS level)
`-Algorithm:min_intensity`	Value	0	Minimum peak intensity
`-Algorithm:RT_window`	Second	20.0	Retention time window size
`-FeatureTracing: quant_method`	"Area," "median," or "max_height"	'area'	Mass tracing quantification method
`-FeatureTracing: max_trace_length`	Second	−1.0	Maximum feature trace length
`-FeatureTracing: min_isotope_cosine`	Value	0.75	Isotope cosine similarity threshold for features

`out_topFD` options are used to specify different spectrum mass output formats (see below).

5. `out_spec` option specifies the spectrum mass tsv file names. The tsv file name should be specified per MS level. For example,

   `./FLASHDeconv -in spectrum.mzml -out mass.tsv -out_spec specmass_ms1.tsv specmass_ms2.tsv`

   will generate `specmass_ms1.tsv` containing the results only for MS1 spectra and `specmass_ms2.tsv` containing the results only for MS2 spectra. `out_spec` option is not mandatory, but should be used in the case in which `use_ensemble_spectrum` option is used (*see* **step 1** in Subheading 3.4). Table 4 shows the columns in the spectrum mass tsv file.
6. If `write_detail` option is set, more detailed information is reported in the spectrum mass tsv file. The additional columns are also given in Table 4.
7. `out_mzml` option specifies the mzML file name for the deconvoluted spectra. The mzML output file contains the results for all MS levels. `out_mzml` option can be coupled with `mzml_mass_charge` option that determines the charges of the deconvoluted masses.

Table 3
The title of columns in the feature mass tsv file, which is specified by `out` option

Name	Description
FeatureIndex	Index for features
FileName	The input mzML file name
MonoisotopicMass	Monoisotopic feature mass
AverageMass	Average feature mass
MassCount	Number of the spectrum masses in a feature
StartRetentionTime	Start retention time in seconds
EndRetentionTime	End retention time in seconds
RetentionTimeDuration	Retention time duration in seconds
ApexRetentionTime	Apex retention time in seconds
SumIntensity	Summed intensity of all masses in a feature
MaxIntensity	Maximum intensity out of all mass intensities
FeatureQuantity	Quantity of features (calculated with the method chosen by `FeatureTracing:quant_method` option)
MinCharge	Minimum charge
MaxCharge	Maximum charge
ChargeCount	The number of distinct charges
IsotopeCosine	Cosine similarity between observed and avergine isotope patterns
PerChargeIntensity	Intensities per distinct charges (separated by semicolon)
PerIsotopeIntensity	Intensities per distinct isotope indices (separated by semicolon)

8. `out_topFD` option specifies the msalign file names for the deconvoluted spectra. The msalign file name should be specified per MS level. For example,

   ```
   ./FLASHDeconv -in spectrum.mzml -out mass.tsv -out_topFD specmass_ms1.msalign specmass_ms2.msalign
   ```

   will generate `specmass_ms1.msalign` containing the results only for MS1 spectra and `specmass_ms2.msalign` containing the results only for MS2 spectra (*see* **Note 3**).

3.4 Adjusting Basic Parameters

The remaining options of FLASHDeconv control the parameters for the deconvolution and the mass tracing. The options starting with "`Algorithm:`" and "`FeatureTracing:`" determine the parameter values for the deconvolution and for the mass tracing, respectively. The options without any prefix affect both the deconvolution and the tracing. In this section, only the parameters from

Table 4
The title of columns in the spectrum mass tsv file, which is specified by `out_spec` option

Name	Description	Property
FileName	Input file (*.tsv) name	
ScanNum	Scan number	
RetentionTIme	Retention time in seconds	
MassCountInSpec	The number of spectrum masses in a spectrum	
AverageMass	Average mass	
MonoisotopicMass	Monoisotopic mass	
SumIntensity	Sum of peak intensities	
MinCharge	Minimum charge	
MaxCharge	Maximum charge	
PeakCount	The number of peaks belonging to the mass	
IsotopeCosine	Cosine similarity between observed and avergine isotope patterns	
ChargeScore	Charge distribution score (between 0 and 1, the higher the better)	
MassSNR	SNR of the mass	
RepresentativeCharge	Representative charge	
RepresentativeMzStart	m/z start of the isotopomer envelope for the representative charge	
RepresentativeMzEnd	m/z end of the isotopomer envelope for the representative charge	
QScore	QScore to measure the quality of the signal (*see* **Note 11**)	
PerChargeIntensity	Intensities per distinct charges (separated by semicolon)	
PerIsotopeIntensity	Intensities per distinct isotope indices (separated by semicolon)	
PrecursorScanNum	Precursor scan number	MS2
PrecursorMz	Precursor m/z	MS2
PrecursorIntensity	Precursor intensity	MS2
PrecursorCharge	Precursor charge (determined by FLASHDeconv)	MS2
PrecursorMonoisotopicMass	Precursor monoisotopic mass (determined by FLASHDeconv)	MS2
PrecursorQScore	Precursor QScore (*see* **Note 11**)	MS2
PeakMZs	Peak m/zs (of the peaks belonging to the mass; separated by semicolon)	Detail
PeakIntensities	Peak intensities (separated by semicolon)	Detail
PeakCharges	Peak charges (separated by semicolon)	Detail

(continued)

Table 4
(continued)

Name	Description	Property
PeakMasses	Peak masses (masses before deisotoping; separated by semicolon)	Detail
PeakIsotopeIndices	Peak isotope indices (separated by semicolon)	Detail
PeakPPMErrors	ppm mass error between peak masses and theoretical masses (before deisotoping; separated by semicolon)	Detail

The columns with "MS2" property are the ones appearing only for MS2 spectrum output file. The columns with "Detail" property are the ones appearing with `write_detail` option activated

the basic options (shown by typing "`./FLASHDeconv --help`" in a terminal) are explained. The list of all basic options is given in Table 1.

1. If `use_ensemble_spectrum` option is set, a single ensemble spectrum is generated per MS level. An ensemble spectrum of an MS level is generated by aggregating the spectra of the MS level across the retention time direction. Then FLASHDeconv performs the deconvolution for each ensemble spectrum (*see* **Note 4**).
2. `Algorithm:tol` option sets the ppm tolerance values for the deconvolution and the mass tracing. The tolerance is set per MS level. For example, `Algorithm:tol 10.0 20.0` specifies 10 ppm and 20 ppm mass tolerances for MS1 and MS2, respectively. The tolerance for the mass tracing is the same as MS1 tolerance, which is 10 ppm in this example.
3. `Algorithm:min_mass` and `Algorithm:max_mass` options specify the mass range of the deconvoluted masses. These parameters are useful for targeted studies where target proteoform mass ranges are known.
4. `Algorithm:min_charge` and `Algorithm:max_charge` options specify the charge range of the deconvoluted masses for MS1 spectra (*see* **Note 5**). The charge values can be negative for negative-mode MS experiments.
5. `Algorithm:min_isotope_cosine` option sets the threshold of the cosine similarity between the observed and the averagine isotope distributions, per MS level (*see* **Note 6**). False-positive masses can be effectively filtered out by raising this cosine threshold. However, using high-cosine threshold value (e.g., 0.9 or higher) results in the loss of large proteoform masses (e.g., larger than 50 kDa).

6. `FeatureTracing:mass_error_da` option determines the Dalton tolerance of the mass tracing algorithm. A value larger than 1.5 Da is recommended due to frequent isotope index error.
7. `FeatureTracing:min_trace_length` option determines the minimum feature retention time length (in seconds).

3.5 Adjusting Advanced Parameters

The advanced options allow for more detailed control of FLASHDeconv (*see* **Note 7**). The list of all advanced options is given in Table 2.

1. `max_MS_level` option determines the maximum MS-level subject to the analysis.
2. If `use_RNA_averagine` option is set, the averagine isotope patterns are generated with RNA nucleotides instead of protein amino acids.
3. `Algorithm:min_mz` and `Algorithm:max_mz` options specify the minimum and maximum m/z values of the spectrum subject to the analysis. If set as negative values, these options are deactivated.
4. `Algorithm:min_RT` and `Algorithm:max_RT` options specify the retention time range (in seconds) subject to the analysis. If set as negative values, these options are deactivated.
5. `Algorithm:min_peaks` option sets the number of the peaks that support each mass (*see* **Note 8**). The number is set per MS level. For example,

   ```
   ./FLASHDeconv -in spectrum.mzml -out mass.tsv
   -Algorithm:min_peaks 4 3
   ```

   sets the numbers of supporting peaks for MS1 and MS2 to be 4 and 3, respectively.
6. `Algorithm:min_intensity` option sets the intensity threshold of the peak intensity.
7. `Algorithm:RT_window` option sets the size (in seconds) of the retention time window in which spectra are considered altogether for sensitive deconvolution (*see* **Note 9**).
8. `FeatureTracing:quant_method` option determines the method of quantification for mass traces. For LC (liquid chromatography) data, `area` is recommended while `median` is recommeneded for direct injection data. `max_height` simply uses the most intense peak in the trace. The resulting quantification value is written in the FeatureQuantity column of the feature mass tsv output file (specified by `out` option).
9. `FeatureTracing:max_trace_length` option determines the maximum feature retention time length (in seconds). If set as a negative value, this option is deactivated.

10. `FeatureTracing:min_isotope_cosine` option sets the threshold of the cosine similarity between the observed and the averagine isotope distributions for features (*see* **Note 10**). Similar to `Algorithm:min_isotope_cosine`, false-positive features can be effectively filtered out by raising this cosine threshold. However, using high-cosine threshold value (e.g., 0.9 or higher) results in the loss of most proteoform features larger than 50 kDa.

3.6 Running FLASHDeconv in Other Environments

1. Running FLASHDeconv in workflows: OpenMS tools are integrated into workflow engines KNIME, Galaxy, nextflow, and TOPPAS. FLASHDeconv is part of OpenMS and thus can be used in pipelines for different workflow systems. The instruction and information on running OpenMS in workflows are given in https://www.openms.de.getting.started.creating.workflows..
2. Running FLASHDeconv in MASH Suite Pro [5] software: FLASHDeconv is integrated in MASH Suite Pro software and can be used for spectral deconvolution. The installation instruction and the user manual are found in https://labs.wisc.edu.gelab.MASH.Explorer.UserResources.php.

4 Notes

1. It is `[OpenMS folder]/bin` folder.
2. The maximum charge for ProMex is 60. Thus, the charge range for FLASHDeconv should be adjusted accordingly (from 1 to 60), using `Algorithm:min_charge` and `Algorithm:max_charge` options (*see* **step 4** in Subheading 3.4).
3. To be able to be recognized by TopPIC, the msalign file name for MS level *n* should be `*msn.msalign` (as shown in the example).
4. If `use_ensemble_spectrum` option is used, the feature mass tsv file specified by `-out` option reports no feature mass. And `out_spec` option should be used to specify the tsv files for ensemble spectra.
5. Reducing the charge range as much as possible reduces possible harmonic mass artifacts. For an MS2 spectrum, the charge range is set to be from 1 to the charge state of its precursor.
6. The Averagine model is determined by `use_RNA_averagine` option (*see* **step 2** in Subheading 3.5).
7. FLASHDeconv is optimized for the default values of the advanced parameters; it is not recommended to change these parameters without thorough understanding of the algorithm.

8. A mass is reported by FLASHDeconv only if it has a sufficient number of peaks that explain the mass (called the supporting peaks). The exact definition of the supporting peaks is beyond the scope of this book chapter. But briefly, the supporting peaks of a mass are the peaks corresponding to the ions of the mass or its neutral losses. For example, the peaks of the mass of distinct charge states are possible supporting peaks. Also the peaks corresponding to the water loss of the mass can be supporting peaks. For higher specificity, the intensities and inter-location of such peaks are considered to define the final supporting peak list.
9. When performing the deconvolution for a spectrum, FLASHDeconv uses information from spectra close in retention time (similar to Xtract and ReSpect), as described in [9]. The value given by `RT_window` option sets the size of the retention time window in which the spectra are considered for the deconvolution.
10. If `FeatureTracing:min_isotope_cosine` option is not used, the MS1 cosine threshold set by `Algorithm:min_isotope_cosine` is used.
11. QScore is a score that measures the quality of the signals representing a single molecule. The exact definition of QScore is beyond the score of this chapter. But briefly, QScore is obtained by a logistic regression method using multiple features that are extracted from raw spectrum signals representing the same molecule. The features include the cosine similarity between observed and averagine isotope patterns, signal-to-noise ratio (SNR), charge distribution score, etc.

References

1. Chen B, Brown KA, Lin Z, Ge Y (2018) Top-down proteomics: ready for prime time? Anal Chem 90(1):110–127. https://doi.org/10.1021/acs.analchem.7b04747
2. Li H, Nguyen HH, Ogorzalek Loo RR, Campuzano IDG, Loo JA (2018) An integrated native mass spectrometry and top-down proteomics method that connects sequence to structure and function of macromolecular complexes. Nat Chem 10(2):139–148. https://doi.org/10.1038/nchem.2908
3. Schaffer LV, Millikin RJ, Miller RM, Anderson LC, Fellers RT, Ge Y, Kelleher NL, LeDuc RD, Liu X, Payne SH, Sun L, Thomas PM, Tucholski T, Wang Z, Wu S, Wu Z, Yu D, Shortreed MR, Smith LM (2019) Identification and quantification of Proteoforms by mass spectrometry. Proteomics 19(10):e1800361. https://doi.org/10.1002/pmic.201800361
4. Donnelly DP, Rawlins CM, DeHart CJ, Fornelli L, Schachner LF, Lin Z, Lippens JL, Aluri KC, Sarin R, Chen B, Lantz C, Jung W, Johnson KR, Koller A, Wolff JJ, Campuzano IDG, Auclair JR, Ivanov AR, Whitelegge JP, Pasa-Tolic L, Chamot-Rooke J, Danis PO, Smith LM, Tsybin YO, Loo JA, Ge Y, Kelleher NL, Agar JN (2019) Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. Nat Methods 16(7):587–594. https://doi.org/10.1038/s41592-019-0457-0
5. Cai W, Guner H, Gregorich ZR, Chen AJ, Ayaz-Guner S, Peng Y, Valeja SG, Liu X, Ge Y (2016) MASH suite pro: a comprehensive software tool for top-down proteomics. Mol Cell Proteomics 15(2):703–714. https://doi.org/10.1074/mcp.O115.054387

6. Kou Q, Xun L, Liu X (2016) TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. Bioinformatics 32(22):3495–3497. https://doi.org/10.1093/bioinformatics/btw398
7. 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 (2017) Informed-proteomics: open-source software package for top-down proteomics. Nat Methods 14(9):909–914. https://doi.org/10.1038/nmeth.4388
8. Zabrouskov V, Senko MW, Du Y, Leduc RD, Kelleher NL (2005) New and automated MSn approaches for top-down identification of modified proteins. J Am Soc Mass Spectrom 16(12):2027–2038. https://doi.org/10.1016/j.jasms.2005.08.004
9. Jeong K, Kim J, Gaikwad M, Hidayah SN, Heikaus L, Schluter H, Kohlbacher O (2020) FLASHDeconv: ultrafast, high-quality feature deconvolution for top-down proteomics. Cell Syst 10(2):213–218. e216. https://doi.org/10.1016/j.cels.2020.01.003
10. Rost HL, Sachsenberg T, Aiche S, Bielow C, Weisser H, Aicheler F, Andreotti S, Ehrlich HC, Gutenbrunner P, Kenar E, Liang X, Nahnsen S, Nilse L, Pfeuffer J, Rosenberger G, Rurik M, Schmitt U, Veit J, Walzer M, Wojnar D, Wolski WE, Schilling O, Choudhary JS, Malmstrom L, Aebersold R, Reinert K, Kohlbacher O (2016) OpenMS: a flexible open-source software platform for mass spectrometry data analysis. Nat Methods 13(9):741–748. https://doi.org/10.1038/nmeth.3959
11. Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, Tang WH, Rompp A, Neumann S, Pizarro AD, Montecchi-Palazzi L, Tasman N, Coleman M, Reisinger F, Souda P, Hermjakob H, Binz PA, Deutsch EW (2011) mzML – a community standard for mass spectrometry data. Mol Cell Proteomics 10(1):R110 000133. https://doi.org/10.1074/mcp.R110.000133
12. Kessner D, Chambers M, Burke R, Agus D, Mallick P (2008) ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics 24(21):2534–2536. https://doi.org/10.1093/bioinformatics/btn323
13. Kenar E, Franken H, Forcisi S, Wormann K, Haring HU, Lehmann R, Schmitt-Kopplin P, Zell A, Kohlbacher O (2014) Automated label-free quantification of metabolites from liquid chromatography-mass spectrometry data. Mol Cell Proteomics 13(1):348–359. https://doi.org/10.1074/mcp.M113.031278

Chapter 12

Deconvolving Native and Intact Protein Mass Spectra with UniDec

Marius M. Kostelic and Michael T. Marty

Abstract

Intact protein, top-down, and native mass spectrometry (MS) generally requires the deconvolution of electrospray ionization (ESI) mass spectra to assign the mass of components from their charge state distribution. For small, well-resolved proteins, the charge can usually be assigned based on the isotope distribution. However, it can be challenging to determine charge states with larger proteins that lack isotopic resolution, in complex mass spectra with overlapping charge states, and in native spectra that show adduction. To overcome these challenges, UniDec uses Bayesian deconvolution to assign charge states and to create a zero-charge mass distribution. UniDec is fast, user-friendly, and includes a range of advanced tools to assist in intact protein, top-down, and native MS data analysis. This chapter provides a step-by-step protocol and an in-depth explanation of the UniDec algorithm, and highlights the parameters that affect the deconvolution. It also covers advanced data analysis tools, such as macromolecular mass defect analysis and tools for assigning potential PTMs and bound ligands. Overall, this chapter provides users with a deeper understanding of UniDec, which will enhance the quality of deconvolutions and allow for more intricate MS experiments.

Key words Mass spectrometry data analysis, Native mass spectrometry, Top-down proteomics, Bayesian deconvolution, Electrospray ionization, Nanodiscs, Intact protein analysis

1 Introduction

One of the breakthroughs of electrospray ionization (ESI) is that it enabled high-resolution analysis of intact proteins. Intact protein analysis has since become useful in the routine characterization of biopharmaceuticals, which enables rapid profiling of the mass distribution and covalent protein modifications [1]. Building on intact protein analysis, top-down proteomics uses gas-phase fragmentation to gain further insights into the protein sequence and sites of covalent modifications, which provides an indispensable tool for characterizing proteoforms [2–6]. A related tool, native MS uses nondenaturing ionization conditions to observe not only covalent modifications but also noncovalent interactions such as bound

Liangliang Sun and Xiaowen Liu (eds.), *Proteoform Identification: Methods and Protocols*,
Methods in Molecular Biology, vol. 2500, https://doi.org/10.1007/978-1-0716-2325-1_12,
© The Author(s), under exclusive license to Springer Science+Business Media, LLC, part of Springer Nature 2022

ligands or complex formation [7–11]. The combination of native MS with top-down provides a powerful tool for the simultaneous characterization of covalent and noncovalent protein interactions [12–14]. For all of these techniques, the initial spectrum of the intact protein has multiple charge states that must be assigned to determine the mass. Charge state assignment can either be done by comparing the spacing between peaks with different charge states or by comparing the spacing of the peaks within an isotope series when high-resolution analysis is available.

However, as intact protein MS analysis moves toward more complex experiments and larger proteins, data analysis becomes challenging due to adduction, loss of isotopic resolution, and overlapping charge state peaks from multiple charge state distributions. These challenges can make it difficult to assign charge states for different m/z peaks. Furthermore, there is an increasing demand for automated and higher throughput data analysis approaches that are capable of handling these more complex mass spectra. To overcome these challenges, multiple deconvolution software programs have been created to help assign charge states and deconvolve intact MS data [15–25]. A recent review has covered a wide variety of deconvolution software [26]. UniDec (Universal Deconvolution) software uses Bayesian deconvolution to quickly assign the charge states of m/z peaks and thus determine the mass [27, 28]. UniDec can deconvolve a wide variety of complex spectra such as native ion-mobility (IM), intact protein spectra, native membrane proteins, nanodiscs, antibodies, and DNA–protein complexes [10, 16, 29–33].

Although the UniDec algorithm is open source and has been published [27, 28, 30, 34, 35], there is no written step-by-step protocol for using the UniDec software. Moreover, because the default parameters work well for many applications, many users may lack a clear understanding of the algorithm and the advanced features. UniDec is not magic. It will create the most likely deconvolved mass spectrum based on the input parameters and can produce a distorted deconvolution if used incorrectly. Conversely, a deeper understanding of the available parameters provides powerful tools for deconvolving complex spectra. This chapter provides a step-by-step protocol for using UniDec, including comprehensive descriptions of the most common deconvolution parameters and data analysis tools. By understanding these crucial parameters, users will be able to deconvolve a wide variety of complex native and intact protein MS samples.

2 How Does UniDec Work?

Before outlining a step-by-step protocol, we will first provide an overview of the underlying UniDec algorithm, focusing on a typical 1D mass spectrum (Fig. 1). The goal of UniDec deconvolution is to separate the charge and mass components from an *m/z* spectrum. To accomplish this, UniDec goes through multiple iterations of three steps to create a 2D charge vs *m/z* matrix that contains the most likely charge state distribution for each *m/z* data point in the spectrum. The 2D charge vs *m/z* plot is shown directly below the processed *m/z* spectrum after deconvolution in the software.

Initially, UniDec assumes that all charge states are equally likely for any *m/z* point (Fig. 1b). UniDec can determine the most probable charge states in three ways: by using the charge state distribution, the mass distribution, or the isotope distribution. These options are not mutually exclusive and can be used as complimentary filters. Without at least one of these filters, UniDec will not be able to assign a charge state. Here, we will focus on the most common case for native and intact protein spectra, which assumes a smooth charge state distribution common in ESI of proteins.

The first step is to filter the charge state distribution (Fig. 1c). This means that UniDec will adjust the intensity of potential charge states in the 2D charge vs. *m/z* matrix based on their neighboring charge states. For a point at *z* and *m/z* in the matrix, UniDec will look at the intensities at neighboring pairs at $z + 1$ with $m/(z + 1)$ and $z - 1$ with $m/(z - 1)$. If the neighboring charge states have strong intensities, UniDec will increase the intensity of the point at *z* and *m/z*. If not, it will decrease the intensity assigned to that point in the matrix.

Second, UniDec will sum the intensities of all charge assignments in the 2D charge vs *m/z* matrix into a simulated $1 - D$ *m/z* spectrum that has the same dimensions as the data (Fig. 1d). The simulated spectrum can be convolved with a known peak shape—based on the full width at half maximum (FWHM) and peak shape parameters—to define what the minimum peak size should be. Setting a minimum peak width can help to avoid deconvolution artifacts by telling the algorithm that points within the peak width should have the same charge state.

Third, the simulated mass spectrum intensities are compared to the intensities in the actual mass spectrum. If an *m/z* peak in the simulated spectrum has a higher intensity than its actual counterpart in the original data, UniDec will lower the intensity of all charge states at that *m/z*. If the simulation is lower, UniDec will increase the intensity of all potential charge states at that *m/z* (Fig. 1e). After this, the loop restarts by adjusting the charge state distributions and continues.

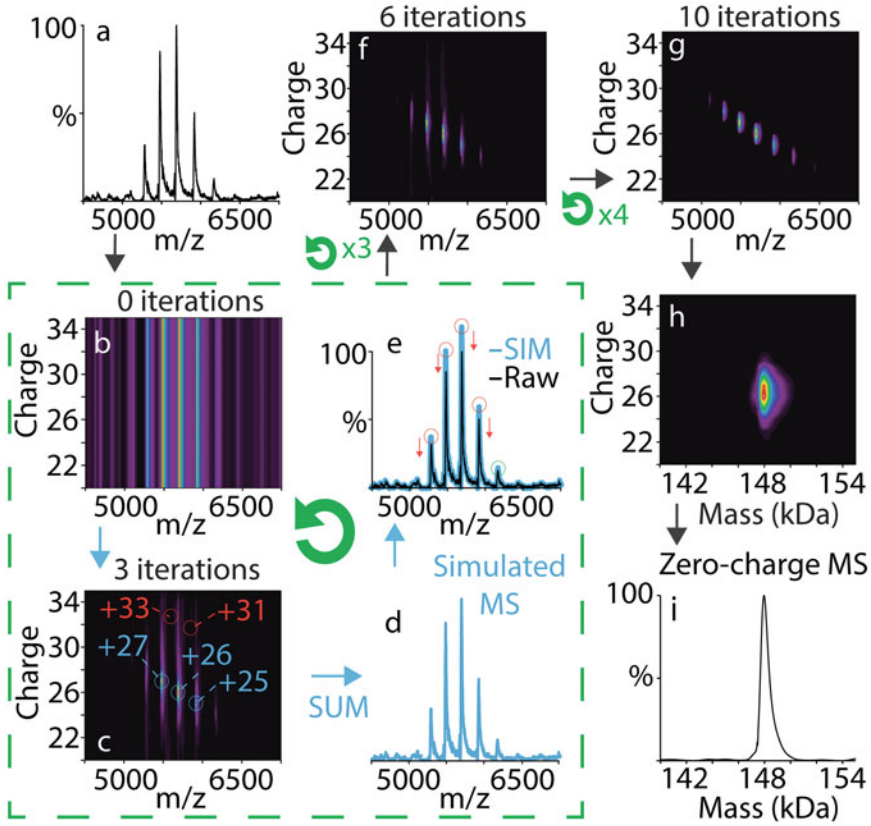


Fig. 1 Schematic of the UniDec algorithm for a mass spectrum of alcohol dehydrogenase (ADH). (**a**) Raw mass spectrum of ADH where the charge states and intact mass are unknown. (**b**) Initial charge vs *m/z* matrix where all charge states are equally likely. (**c**) The charge vs *m/z* plot after 3 iterations, which also conceptually shows the first step of the algorithm filtering the charge state distribution. The +26 charge state is more likely than the +32 charge state because it has support from the neighboring +27 and +25 charge states. (**d**) Step 2 of the algorithm, where the charge vs *m/z* plot is summed into a 1D simulated mass spectrum. (**e**) Step 3 of the algorithm, where the simulated mass spectrum is compared to the raw data. Steps (**c**–**e**) show one iteration of the algorithm and are symbolized by the green loop. The charge vs *m/z* matrix is then shown at (**f**) 6 and (**g**) 10 iterations before transformation into the (**h**) charge vs mass (kDa) matrix, which is summed to yield the (**i**) zero-charge mass spectrum after completing all the iterations

The loop ends when the intensities in the simulated mass spectrum match those in the original data or have converged. Finally, the 2D charge vs. *m/z* matrix is converted into a 2D charge vs. mass matrix (Fig. 1h). Summing the charge vs. mass matrix across all charge states produces the zero-charge mass spectrum (Fig. 1i). Armed with a basic understanding of the algorithm, we will now outline how to use the software and describe specific parameters.

3 Downloading and Installing UniDec

The latest version of UniDec can be downloaded from GitHub at (https://github.com/michaelmarty/UniDec/releases). This protocol will cover version 4.3.0 but applies to other versions as well. Unzip and extract the folder into a convenient location such as the desktop. No further installation is necessary because UniDec runs as a standalone, portable program. Then, run the GUI_UniDec.exe file. This compiled version is only available for Windows, but Mac and Linux versions can be built from the source code. UniDec can also be run from the Python source code, which allows command line usage and scripting. Further information on installing the source code can be found here: https://github.com/michaelmarty/UniDec/wiki/.

4 UniDec Data Processing and Deconvolution

4.1 Opening Mass Spectra

The first step in analyzing mass spectra with UniDec is opening the data file. UniDec can read a variety of mass spectra file formats. Where applicable, UniDec will average all spectra in the file into a single mass spectrum. Chromatography data can be parsed externally (see **step 3** below) or opened with MetaUniDec [28] or UniChrom to parse mass spectra from specific time sections.

1. Text files, mzML, or Thermo RAW files can be opened in UniDec by going to *File* > *Open File (text, mzML, or Thermo RAW)*. These files can also be opened by dragging and dropping them into the main window. Text files should be in the form of where each row is a data point with the first column as the *m/z* value and the second column as the intensity. For Bruker data, the spectrum needs to be exported as a "simple x-y ASCII," which can then be opened as a text file in UniDec.
2. Waters and Agilent data are stored in directories rather than single files and can be opened using *File* > *Open Waters or Agilent File*. These files can also be opened with drag and drop.
3. For quick analysis, data can be copied to the clipboard from Thermo QualBrowser or Waters MassLynx. For the Spectrum window in MassLynx, *Edit* > *Copy Spectrum List* will copy the selected spectrum to the clipboard. For Thermo QualBrowser, right-clicking the mass spectrum and selecting *Export* > *Clipboard* (*Exact Mass*) will copy the spectrum to the clipboard. To import data from the clipboard into UniDec, go to *File* > *Get Spectrum from Clipboard* or use the Ctrl+G shortcut. Opening data from the clipboard is a quick way to deconvolve a selected mass spectrum with UniDec, especially during data collection, but it will generally be saved as a temporary file. To save the

data in a permanent location, data from the clipboard can be pasted into a text file, saved, and opened as described in step 1. We recommend saving permanent files so that data analysis can be traced.

4.2 UniDec Presets

After opening a spectrum, a preset can be selected to better tailor the data processing and deconvolution settings. There are multiple fixed presets in *File* > *Presets*. *High Resolution Native* is a good starting preset for the native protein MS data. For denatured intact protein analysis, the maximum charge state may need to be increased, but this preset should otherwise be a good starting point. Settings are saved as <*filename*>*_conf.dat* files in the <*filename*>*_unidecfiles* folder. Past deconvolution settings can be opened using *File* > *Load External Config File* and selecting any *_conf.dat* file. Adding a config file to the Presents folder in the UniDec directory will make this file available as a custom preset in the Presets menu when the program is reopened. UniDec will automatically reload settings from previously opened files as long as the <*filename*>*_unidecfiles* folder and <*filename*>*_conf.dat* files are still in the same directory as the top file.

4.3 Data Processing

Before deconvolution, the data needs to be processed. In the simplest case, this will simply copy over the raw data and clean up any issues like negative intensities or duplicate *m/z* values. However, additional data processing can be applied to improve and speed up the deconvolution. Data processing parameters can be found in the blue tabs on the control panel to the right of the main UniDec window (Fig. 2a, b).

1. In the data processing tab, the *m/z* range can be set by entering the minimum and maximum values in the corresponding boxes. Another option to set an *m/z* range is to click and hold the left-click and drag across the *m/z* area of interest in the plot to zoom in and then right-click on the plot to automatically save the new minimum and maximum from the selected range. Clicking the "*Full*" button, next to the *m/z* range, will reset the *m/z* range to the lowest and highest *m/z* values in the spectrum.
2. Below the *m/z* range, clicking the "*Use Background Subtraction*" checkbox will subtract a curved baseline from the spectrum based on smoothing the local minima throughout the spectrum. The type (flat, linear, or curved) and the degree of background subtraction can be further adjusted in the Additional Data Processing Parameters tab (*see* **Note 1**).
3. There are multiple additional parameters within the *Additional Data Processing Parameters* tab that are useful for improving the deconvolution. For an in-depth description of each parameter in the *Additional Data Processing Parameters* tab, *see* **Note 2**.

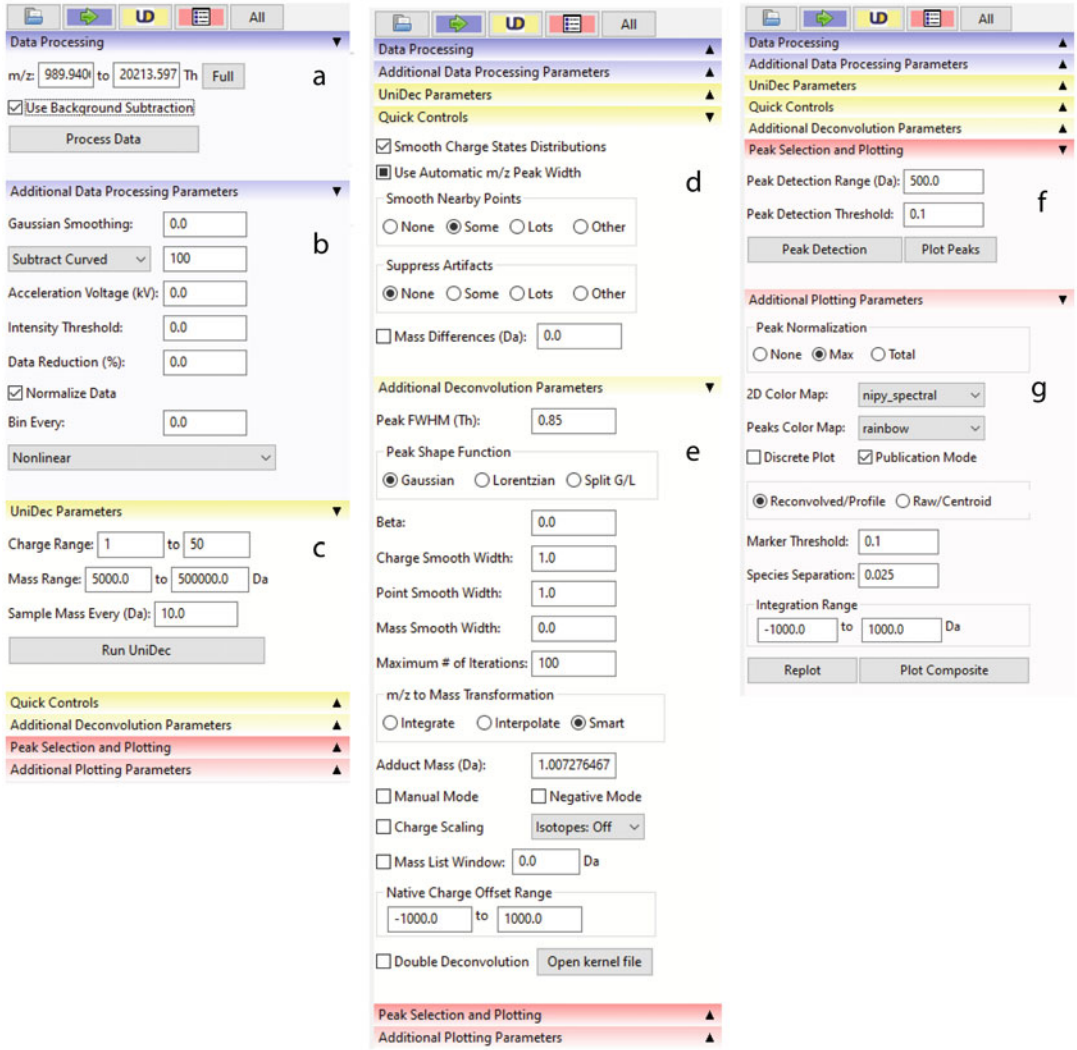


Fig. 2 The UniDec control panel: (**a**) data processing parameters (*blue*), (**b**) advanced data processing parameters (*blue*), (**c**) UniDec parameters tab (*yellow*), (**d**) quick control parameters (*yellow*), (**e**) advanced deconvolution parameters (*yellow*), (**f**) peak selection and plotting parameters (*red*), and (**g**) additional plotting parameters (*red*)

4. Clicking "*Process Data*" will apply the parameters within the data processing tabs and write the processed data into the unidecfiles folder (<*filename*>*_unidecfiles*) in the same directory as the file or embedded in the raw directory for Waters and Agilent data. This folder initially contains the <*filename*>*_conf.dat*, <*filename*>*_rawdata.txt*, and <*filename*>*_input.dat* files. The *_input.dat* file contains the processed data with *m/z* values in the first column and the corresponding intensities in the second column.

4.4 Deconvolution Parameters and Quick Controls

After the data has been processed, the next step is to deconvolve the data. A few simple deconvolution parameters set the range of potential masses and charge states as well as how precisely to sample the deconvolved mass data. The quick controls then allow the deconvolution to be subtly guided by assumptions about the data (Fig. 2c, d).

1. In the *UniDec Parameters* tab, the *Charge Range* and *Mass Range* set the allowed charges and mass values for the deconvolution. The charge range will define the rows in the 2D charge vs. *m/z* matrix, and the mass range will forbid points in the matrix outside the mass range. Setting a *Charge Range* or *Mass Range* that is too narrow will cause artifacts in the deconvolution. Thus, it is better to start with a broader range and then narrow the charge and mass range to the area of interest. Narrowing the ranges is a useful tool to eliminate artifacts from harmonics at double or half the charge and to speed up the algorithm.
2. *Sample Mass Every* sets the sample rate for the deconvolved mass spectrum. It is important to set a sample rate with adequate mass resolution. If the sample rate is set to 10 Da, then each mass value will fall on an even 10 Da. Thus, it may be more beneficial to set the sample rate to 1 or 0.1 Da for a better mass accuracy at the expense of speed.
3. Within the *Quick Controls* tab, the first option is to tell UniDec to expect a Smooth Charge State Distribution by clicking the "*Smooth Charge State Distributions"* check box, which sets the *Charge Smooth Width* in the *Additional Deconvolution Parameters* tab to the default of 1 (*see* **Note 3**). As described above in the first step of the algorithm (Subheading 2), charge smoothing tells UniDec that the probability of a potential charge state, *z*, for an *m/z* data point in the charge vs. *m/z* matrix is correlated to the intensity of neighboring charge states at $z - 1$ and $z + 1$ at data points $m/(z-1)$ and $m/(z + 1)$, respectively.
4. Checking the "*Use Automatic m/z Peak Width*" box will tell UniDec to automatically determine the peak width and peak shape from the most abundant peak in the processed mass spectrum and apply this for the deconvolution. The peak width and peak shape affect the second step of the algorithm. In this step, the data is convolved with a specified peak shape, which means that each mass peak will have the same minimum width and shape. This convolution can be ignored by unchecking the box or by setting the peak width to 0. The peak width and peak shape parameters can be set manually in the *Additional Deconvolution Parameters* tab. Manually increasing the peak width is useful when the deconvolution is overfitting the

data and presenting artifacts. Decreasing the peak width is useful when the deconvolution is underfitting the data.

5. Setting the *Smooth Nearby Points* option will tell UniDec that nearby data points should have the same charge state [30]. This will smooth nearby data points into the same charge state to reduce artifacts in the deconvolved mass spectrum, which is usually caused by misassigned charge states. The four options in the *Smooth Nearby Points* affect the *Point Smooth Width* parameter in the *Additional Deconvolution Parameters* tab. The "*some*" option works well for most mass spectra, but some mass spectra may require a higher setting (*see* **Note 4**).
6. The *Suppress Artifacts* feature tells UniDec that each *m/z* data point should have a single charge state, which helps eliminate artifacts from misassigned charge states. It uses a SoftMax function to push more intense charge states higher for each *m/z*, and the degree to which it pushes toward a single charge state is specified by the *Beta* parameter in the *Additional Deconvolution Parameters* tab. The *Suppress Artifacts* feature is useful when each peak should have a single charge state assignment but harmonics or satellites artifacts are present (*see* **Note 5**) [35].
7. *Mass Differences* helps assign charge states based on an expected mass distribution. Mass distribution smoothing is applied in the first step of the algorithm alongside charge state smoothing (see Subheading 2). It tells UniDec that each *m/z* peak should have neighboring peaks that are $(m - nM)/z$ and $(m + nM)/z$, where M is a repeating mass in the spectrum that can be set in the *Mass Differences* box. The repeating mass could be a repeating monomer unit in a polymer/oligomer or a repeating lipid mass in a nanodisc [36]. The n value is set by the *Mass Smooth Width* parameter in the *Additional Deconvolution Parameters* tab, which tells UniDec how many masses to consider on either side of the initial *m/z* value. Clicking the *Mass Differences* checkbox will set the *Mass Smooth Width* to 1.

4.5 Advanced Deconvolution Parameters

In addition to the basic deconvolution parameters and the quick controls, there are a number of advanced parameters that are useful for different applications (Fig. 2e). Although the default values are sufficient for most cases, we will highlight a few parameters that users should consider. Other parameters, such as *Charge Scaling*, *Mass List Window*, and *Native Charge Offset Range*, are described in **Note 6**.

1. Clicking the *"Negative Mode"* checkbox will switch the *Adduct Mass* to the negative mass of a proton (*see* **Note 6**). Unchecking the box will return it to the default for positive mode, the positive mass of a proton. These account for the charge carriers

typically present in ESI and ensure that UniDec reports the neutral mass of the protein without charge carriers.

2. *Manual Mode* allows the user to manually specify the charge state for specific regions of the spectrum and is useful when charge state assignments are difficult. For example, *Manual Mode* is essential for MS/MS where a single charge state is isolated because the lack of a charge state distribution makes this isolated peak impossible to identify without isotopic resolution. Clicking the "*Manual Mode*" checkbox tells UniDec to use manual assignments and will automatically open the *Tools* > *Manual Assignment* window if no assignments are already set. The *Manual Assignment* window allows the user to select specific regions of the spectrum—defined by a central *m/z* value and an *m/z* range around this value—and set the charge state of this region.

5 Analyzing Deconvolution Data

5.1 Peak Selection and Plotting

After the deconvolution, UniDec has a wide variety of tools to analyze and visualize the data. We will first discuss tools for peak selection and plotting (Fig. 2f, g).

1. The parameters in the *Peak Selection and Plotting* tab allow the user to automatically detect peaks within the deconvolved mass spectrum. The *Peak Detection Range* tells UniDec to select peaks in the deconvolved mass spectrum that are local maxima within a certain window. Thus, to detect two peaks that are 500 Da apart, a *Peak Detection Range* of less than 500 would be required.
2. The *Peak Detection Threshold* tells UniDec to only pick peaks that are above a set intensity threshold. The threshold is the decimal value entered in the box multiplied by the intensity of the most abundant peak. For example, a value of 0.1 will set a peak threshold at 10% of the maximum intensity.
3. Clicking "*Peak Detection*" will detect peaks in the deconvolved mass spectrum, label them with symbols on both the deconvolved mass and the processed data below, and display the *Mass*, *Intensity*, *DScore*, and *Name* for each peak directly to the right of the deconvolved mass spectrum. The *DScore* is part of the UniDec scoring function and a user would usually want a DScore above 60. For a further explanation of the *DScore* function, *see* **Note 7**.
4. Clicking "*Plot Peaks*" next to the *Peak Detection* button will plot each mass species charge state distribution independently. Note that each peak will have the same simulated minimum peak width and shape and may not reflect the peak shapes real

data. *Species Separation*, in the *Additional Plotting Parameters*, can be used to increase the distance between the mass species within the same plot (*see* **Note 8**).

5. UniDec can integrate the selected peaks in the deconvolved mass spectrum by clicking *Analysis* > *Integrate Peaks*. The range of the integration will be highlighted for each selected peak in the deconvolved mass spectrum and is determined by the *Integration Range* in the *Additional Plotting Parameters* (*see* **Note 8**). An asymmetric integration range can be set by adjusting these parameters. The area calculated from the integration for each peak is shown in the peak list in the new *Area* column.
6. Right-clicking a peak in the list will provide a drop-down menu with options on how to isolate, display, and export the data.
7. To label the charge states of a specific mass species, right-click the species in the peak list and select "*Label Charge States*." This will plot all possible charge states for that mass, which helps visualize and confirm the charge state assignment for a deconvolved mass peak.
8. A useful method to display the mass differences between mass peaks is to right-click any mass species and select "*Display Differences*." With *Display Differences* selected, the mass peaks will be labeled as their mass difference from the selected peak. This plotting is useful for quickly identifying posttranslational modifications (PTMs), oligomers, and bound ligands in the mass spectrum.

5.2 Mass and Oligomer Tools for Peak Assignment

The *Mass and Oligomer Tools* is useful for quickly determining possible PTMs or bound ligands based on the masses of each peak. The overall workflow is to first define potential mass species in the *Oligomer Maker* and then match them to measured peaks (Fig. 3).

1. After selecting peaks in the main UniDec window, go to *Tools* > *Mass and Oligomer Tools* or use the shortcut Ctrl+T. To the left is the *Mass List*, which is discussed in **Note 6**. The *Oligomer Maker* is in the center of the window (Fig. 3b).
2. UniDec defines potential species as oligomers with five parameters: a *Base Offset* (Da), *Monomer Mass* (Da), *Min # of Oligomers*, *Max # of Oligomers*, and *Name*. Potential masses will be calculated as *Base Offset* + *Monomer Mass* * *n*, where *n* is an integer from *Min# of Oligomers* to *Max# of Oligomers*. Thus, a potential monomeric protein can be input using a base mass of 0, the protein mass for the *Monomer Mass*, a minimum of 0, and a maximum of 1. Setting the minimum to 1 will force all potential species to have that protein mass as a fixed component.

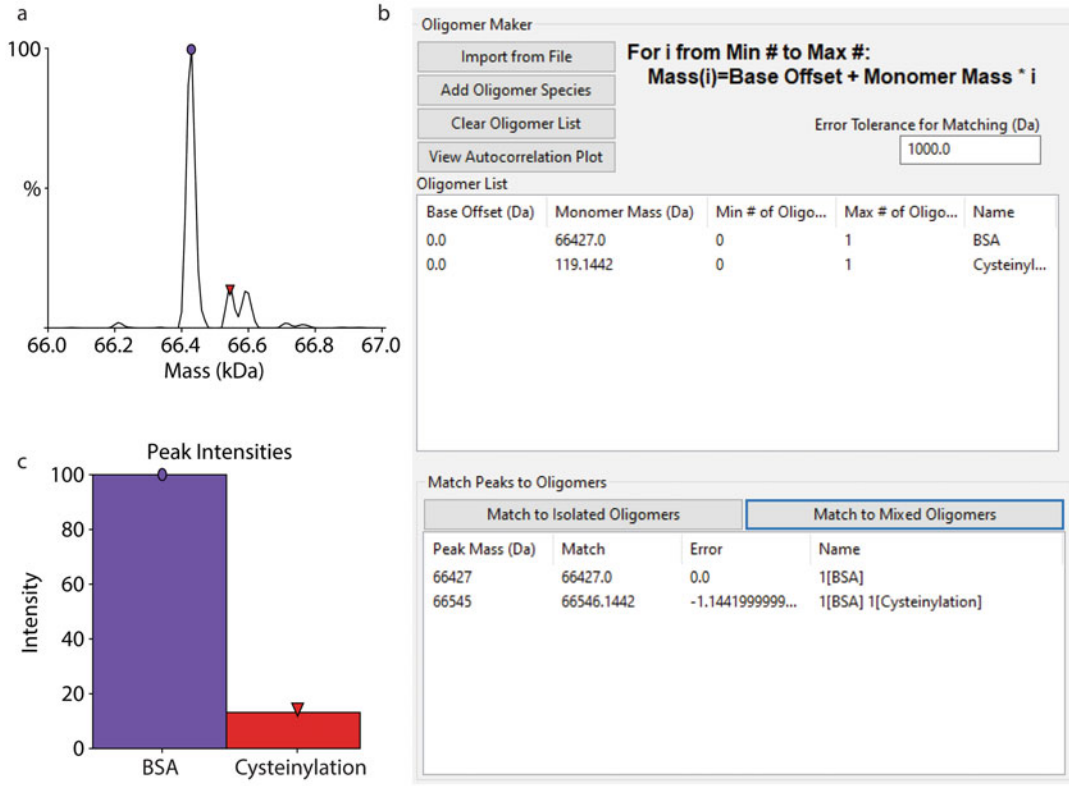


Fig. 3 The *Mass and Oligomer Tools* used to identify a cysteinylation of BSA [37]. (**a**) Deconvolved mass spectrum of BSA. (**b**) *Mass and Oligomer Tools* window, the mass of BSA was manually entered, and the mass of the cysteinylation was imported from the *Common Masses List*. (**c**) Relative peak intensities for BSA and BSA with a cysteinylation created in the main panel

3. Within the *Oligomer Maker*, potential species can be manually added by selecting "*Add Oligomer Species.*" Once a new oligomer species has been added, each parameter can be edited. UniDec will automatically export the oligomers to a file in the unidecfiles folder called <*filename*>*_ofile.dat*. These files can be imported for other analysis and are simply text files with columns and rows that match the list in the window. Masses can also be calculated from the sequence of a protein by right clicking and opening the Biopolymer Calculator window.
4. The *Common Masses List* presents a variety of common PTMs, lipids, ions, and detergents that may be added or removed from the analyte. To add an item from the *Common Mass List* into the *Oligomer Maker*, right-click an item and select "*Add to Oligomer Builder.*" The *Common Mass List* is automatically imported from a CSV file and can be edited to incorporate additional species.

5. After specifying potential oligomers, the *Mass and Oligomer Tools* can match measured peaks to potential oligomer masses by selecting either "*Match to Isolated Oligomers*" or "*Match to Mixed Oligomers*" in the *Match Peaks to Oligomers* section. The *Match Peaks to Oligomer* section presents each match with the measured *Peak Mass* (Da), the theoretical mass of the *Match*, the *Error* (Da) between the two, and the *Name*. The *Error* is the mass difference, positive or negative, between the measured mass and the theoretical mass. The *Name* is taken from the name in the *Oligomer List*; a monomer match would be labeled [*Name*] and a dimer would be labeled *2*[*Name*]. A match will only be displayed if the absolute value of the error is lower than the value set in the *Error Tolerance for Matching* box.
6. "*Match to Isolated Oligomers*" will only allow isolated rows when calculating potential masses, which is useful for matching multiple protein species (each row) that should not combine to form complexes.
7. Selecting "*Match to Mixed Oligomers*" will allow all possible combinations of rows, which is useful for matching potential combinations of PTMs on a protein or for matching combinations of oligomers.
8. Clicking *Ok* will save the matches and label each peak in the main panel with the matched name.

5.3 Macromolecular Mass Defect Analysis

Macromolecular mass defect analysis can be highly useful for visualizing and quantifying complex mass spectra, especially when there are repeating mass units. The principle behind macromolecular mass defect analysis is that small shifts in mass relative to a repeating multiple of a reference mass can be used to assign features of the spectrum. Kendrick mass defect analysis was originally used in hydrocarbon analysis, where the reference mass of CH_2 was used to differentiate a wide diversity of hydrocarbons [38, 39]. Recently, mass defect has been used to determine the stoichiometry of membrane proteins and antimicrobial peptides in lipid nanodiscs [10, 29, 31, 40]. Nanodiscs have a repeating pattern of lipids, and subtle shifts in this pattern can reveal the contents of the nanodiscs even when the overall mass distribution does not shift appreciably, as shown in Fig. 4. The key advantage of mass defect analysis is that polydisperse spectra that differ only in multiples of a reference mass will have the same mass defect, so common features in complex spectra can be visualized.

The mass defect is calculated by dividing the measured masses by the reference mass. The remainder of this division is the mass defect. It can be expressed as a normalized mass defect with a decimal value between 0 and 1 or as an absolute mass defect in units of Da between 0 and the reference mass. Combining mass defects follows modular arithmetic such that a normalized mass

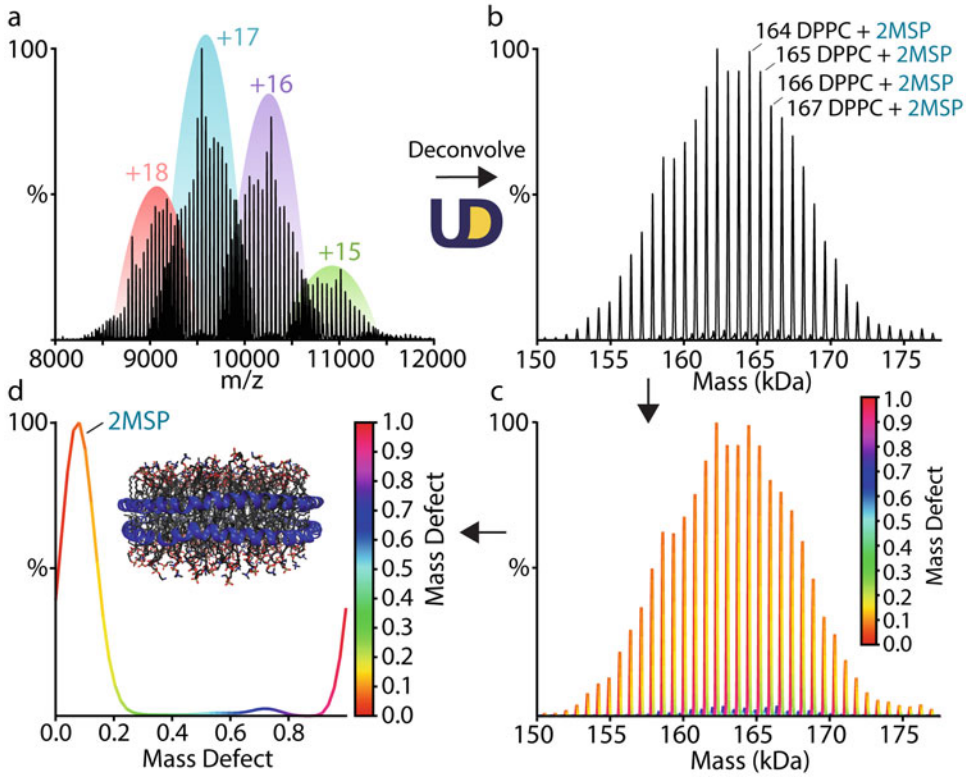


Fig. 4 Native mass spectrum of DPPC nanodiscs with the corresponding mass defect analysis. (**a**) Mass spectrum of DPPC nanodiscs with charge states highlighted. (**b**) Deconvolved mass distribution of DPPC nanodiscs with the number of lipids annotated for select peaks. (**c**) The deconvolved mass distribution colored by mass defect. (**d**) Summed mass defects from the deconvolved mass distribution. The peak around 0.1 corresponds to a nanodisc with 2 MSPs

defect larger than 1 would simply drop the integer part and retain the decimal place. For example, an exact multiple of the reference mass would have a mass defect of 0, which is also equivalent to 1. A component that is 0.4 times the reference mass would have a mass defect of 0.4. Two such components would cause a mass defect of 0.8, and three would cause a mass defect of 0.2. Importantly, these mass defect values are independent of the absolute mass and are the same across any number of repeating reference mass units.

1. After deconvolving the spectrum, go to *Analysis* and select "*Mass Defect Tools*" or use the shortcut Ctrl+K. The reference mass is automatically imported from the *Mass Difference* in the main panel. If no *Mass Difference* value is set, a reference mass will need to be manually entered in the *Mass Defect Tools* window and replotted.
2. Within the *Mass Defect* window, the zero-charge mass spectrum in the top left is colored with mass defect, which shows the mass defect values for each point on the zero-charge mass

spectrum (Fig. 4c). Below the zero-charge mass spectrum, a 2D plot shows mass defect on the *y*-axis and mass (kDa) on the *x*-axis. This plot is useful for also visualizing the mass defect of different mass distributions in the spectrum. To the right, the two plots show the total relative intensity summed across all masses on the *y*-axis and mass defect on the *x*-axis (Fig. 4d). This plot is useful for identifying different mass defect species in the zero-charge mass spectrum. The apex of each peak can be detected by clicking *Plot* > *Label Peaks*.

3. For higher resolution data, increasing the number of mass defect bins may be necessary to capture all relevant features. Clicking "*Replot*" will reset the plots with the new parameters. Additional options allow the user to specify the range and whether to use normalized or absolute mass defect units.

6 Exporting Plots and Mass Spectra

Now that the data has been analyzed and plotted, UniDec has multiple tools to export any plot or mass spectra as figures for publication. Many windows have additional save figure tools in the file menu, but we will cover basic exporting tools that will be useful for any user.

1. All plots in the main UniDec window can be automatically exported as a range of file types, including PDF or PNG, by selecting the *File* > *Save Figures As* menu or selecting one of the *File* > *Save Figure* presets.
2. Additionally, users can middle-click any plot within UniDec, including any analysis window, to open a save figure dialog and specify a save location. Ending the file name with .png or .pdf will set the file type, and any file types supported by MatPlotLib can be used [41].

7 Conclusion

Here, we covered how the UniDec algorithm works, the deconvolution parameters, mass and oligomer tools, and mass defect analysis. Although this chapter provides an explanation of the key features of UniDec, we omitted some tools that may be useful for different types of analysis, including K_D fitting, native charge state analysis tools, and Fourier analysis. Additionally, MetaUniDec is useful for analyzing collections of MS data sets [28], and UniChrom adapts UniDec to time-resolved MS data, such as from LC/MS. Additional information on some of these tools, future features, and more can be found at https://github.com/

michaelmarty/UniDec/wiki. Overall, with the deeper understanding of UniDec and associated analysis tools, we hope users will achieve the best quality deconvolution and analysis for intact and native MS spectra.

8 Notes

1. There are three ways to subtract the background in UniDec: *minimum*, *line*, and *curved*. *Subtract minimum* will find the lowest intensity data point and subtract the whole spectrum by that intensity. The number in the box, next to the subtract setting, does not matter for *Subtract minimum*, any number other than zero will turn it on. *Subtract minimum* is useful for a constant raised baseline. *Subtract line* will create a sloped line from the first and last *m/z* data points in the spectrum. The resulting line will be subtracted from the whole spectrum. The number in the box, *n*, sets how many data points to average at the front and end of the spectrum. *Subtract line* is most useful for mass spectra with linearly sloping baselines. *Subtract curved* will create a curved baseline based on local minima throughout the spectrum. The number in the box, *n*, will set how many data points are considered while creating each local minimum. A smaller *n* will create a rough baseline based on many local minima, and a larger *n* will create a smoother baseline based on fewer local minima. The default of *Subtract Curved* 100 works well for most spectra.
2. This note will describe some of the advanced data processing parameters. *Gaussian Smoothing* can be useful for smoothing noisy data but may broaden the peaks in the mass spectrum. UniDec will smooth the data with a Gaussian that has a width, in data points, set to the number in the box. However, nonlinear resampling (described below) is often a more effective way to smooth data while also speeding up the algorithm.

   Acceleration Voltage is used to correct for the detector efficiency between big ions and small ions in time-of-flight (ToF) mass spectrometers with multichannel plate (MCP) detectors [42].

   Intensity Threshold is useful for removing data points that are under a fixed relative intensity. For example, if the threshold is set at 0.1, all data points below 10% of the most abundant peak will be removed from the spectrum.

   Data Reduction, like *Intensity Threshold*, is useful for removing less abundant data points from the spectrum. However, *Data Reduction* ranks all the data points from least abundant to most abundant and removes a fixed percentage of the least intense based on the percentage entered in the box.

To normalize the data, click the box labeled "*Normalize Data.*" Normalizing the data will set the *y*-axis of all mass and deconvolved mass spectra in relative intensity units such that the maximum is 1. Turning normalization off will leave the *y*-axis in absolute signal intensity.

At the bottom of the Additional Data Processing Parameters tab, a dropdown menu presents options on how to resample the data: *Linear*, *Linear Resolution*, *Nonlinear*, *Linear Interpolated*, and *Linear Resolution Interpolated*. Entering a nonzero value in the Bin Every option will turn the resampling on.

Linear will linearize the spectrum such that the intensities of all data points within a fixed *m/z* range will be summed together into one data point. The value entered in the *Bin Every* option specifies the spacing between data points in the linearized data. For example, a bin value of 5 will create a data point every 5 *m/z*, and all intensities will be summed into the nearest 5 *m/z* bin.

Linear Resolution works similarly to *Linear*, but it linearizes based on a constant resolution rather than a constant *m/z*. If the bin value is set to 5, then the first two data points would be separated by 5 *m/z*, but the next two would be separated by a slightly larger value to keep a constant resolution.

Nonlinear resamples the data differently than linear. *Nonlinear* will tell UniDec to average every *n* number of data points together across the mass spectrum. Both the *m/z* and intensity values are averaged. The value, *n,* is set by the *Bin Every* parameter, and this value specifies the number of data points averaged instead of the *m/z* spacing.

Linear Interpolated works similarly to *Linear*. However, *Linear Interpolated* interpolates the intensity at each linearized data point rather than simply summing everything into fixed bins. *Linear Interpolated* should only be used for oversampling the data.

Linear Resolution Interpolated works similarly to *Linear Interpolated*, but it linearizes with a constant resolution rather than a constant *m/z*.

The *Nonlinear* option is usually the best for reducing the data. It has the least potential for artifacts, resists mass peak broadening, and matches the nonlinear nature of most mass spectra. It is an effective way to smooth the data and speed up the algorithm without introducing artifacts.

3. With a *Charge Smooth Width* of 1, UniDec will use a log mean filter to increase the intensity of potential charge states with neighboring charge states and decrease the intensity of those without neighboring charge states. The *Charge Smooth Width*

value tells UniDec how many neighboring charge states to expect on either side. The default value of 1 is usually optimal.

4. The *Smooth Nearby Points* feature will smooth nearby points within a charge state using a mean filter to reduce artifacts in the deconvolved mass spectrum with a width set by the value in the *Point Smooth Width* box. There are four options for this function: *None*, *Some*, *Lots*, and *Other*. Clicking "*Some*" will set the *Point Smooth Width* to 1, which means a mean filter of width ±1 will be used to smooth each row of the charge vs. *m/z* matrix. Clicking "*Lots*" will set the point smooth filter width to 10 and may broaden the mass peaks. Inputting any value other than 1 or 10 in the *Point Smooth Width*, will set the *Smooth Nearby Points* to *Other*.
5. There are four options for *Suppress Artifacts*: *None*, *Some*, *Lots*, and *Other*. Clicking "*Some*" will set the *Beta* parameter to 50, which is a medium setting. Clicking "*Lots*" will set the *Beta* parameter to 500, which is a high setting that will reduce lots of artifacts; this setting may distort the deconvolved mass spectrum, especially with mass spectra that have overlapping mass peaks. In cases with many overlapping peaks, a *Beta* value of 5–15 may be more appropriate. Manually setting a *Beta* value other than 50 or 500 will switch the *Suppress Artifacts* setting to *Other*.
6. This note describes several advanced deconvolution parameters. *Charge Scaling* divides the intensity of each ion by its charge state to correct for different signal responses for different charge states, which is usually the case in FTICR and Orbitrap mass spectra. For example, an ion with a +2 charge state would have double the signal intensity of the same ion with a +1 charge state because it induces double the current in an FTICR or Orbitrap detector.

   Checking the *Mass List Window* option enforces a list of allowed masses that are entered in the *Mass List* section of the *Mass and Oligomer Tools*, which can be opened by going to *Tools* > *Mass and Oligomer Tools* or by using the shortcut Ctrl +T. The value in the box specifies the range around each entered mass value, in Da, that is allowed. For example, setting the value to 1000 would allow a ±1000 Da window around each mass entered in the *Mass List* but would exclude any masses not within that window from one of the defined masses.

   The *Native Charge Offset Range* sets how far an allowed charge state can be from the predicted native charge state of a globular protein of the same mass [43]. This feature is useful for clipping off high or low charge states that are outside of the expected range for a given mass without having to limit the global charge range.

In the *Additional Deconvolution Parameters* tab, there is a *Negative Mode* check box that will automatically switch the adduct mass to a negative value when turned on. By default, the *Adduct Mass* is set to the mass of a proton. If the charge carrier for the ions were sodium, then the mass of sodium would be used as the *Adduct Mass*. However, if negative ionization mode is being used, then the *Adduct Mass* will have to be switched to a negative value because the charge is caused by the loss of a proton.

7. The *DScore* value is a part of the UniScore tool within UniDec. The UniScore rates the quality of the deconvolution with a range of 0–100, with 0 or less being completely wrong and 100 being perfect. The *DScore* scores each mass peak on the same scale. Briefly, the *DScore* is based on the presence of unique charge state peaks for a mass species, the fit of the peak shape across all charge states, the smoothness of the charge state distribution, and the asymmetry of the mass peak shape. Normally, a *DScore* above 60 indicates a confident deconvolution, and peaks with lower *DScore* values should be checked. Lowering the FWHM parameter can help increase the *DScore*. To learn more about the UniScore function see [34].
8. Other parameters in the *Additional Plotting Parameters* tab include *Peak Normalization*, *Color Maps*, and *Publication Mode*. *Peak Normalization* will tell UniDec how to normalize the intensity of each mass peak. Setting this parameter to "*None*" will turn off the normalization. "*Max*" will normalize the intensities to a percent relative intensity with the most abundant peak having an intensity of 100%. *"Total"* will set the normalization such that the intensity of all peaks added together will be 100%.

   The *Integration Range* parameter is important for integrating the mass peaks in the deconvolved mass spectrum. UniDec will integrate the area under the selected mass peak with a range specified in the minimum and maximum boxes of the *Integration Range*. For example, the range can be asymmetric such that a range of −500 to 1000 Da could be used.

   Clicking the "*Publication Mode*" checkbox will plot the raw *m/z* and deconvolved mass spectra in a simpler plot that is more amenable for publication.

Acknowledgments

The authors thank James Keener for providing the ADH example spectrum. This work was funded by the National Science Foundation (CHE-1845230). The authors thank the broad and diverse community of UniDec users for their helpful suggestions and continued support.

References

1. Beck A, Wagner-Rousset E, Ayoub D, Van Dorsselaer A, Sanglier-Cianferani S (2013) Characterization of therapeutic antibodies and related products. Anal Chem 85(2):715–736. https://doi.org/10.1021/ac3032355
2. Gregorich ZR, Ge Y (2014) Top-down proteomics in health and disease: challenges and opportunities. Proteomics 14(10): 1195–1210. https://doi.org/10.1002/pmic.201300432
3. Cai W, Tucholski TM, Gregorich ZR, Ge Y (2016) Top-down proteomics: technology advancements and applications to heart diseases. Expert Rev Proteomics 13(8):717–730. https://doi.org/10.1080/14789450.2016.1209414
4. Toby TK, Fornelli L, Kelleher NL (2016) Progress in top-down proteomics and the analysis of proteoforms. Annu Rev Anal Chem (Palo Alto, Calif) 9(1):499–519. https://doi.org/10.1146/annurev-anchem-071015-041550
5. Smith LM, Kelleher NL, Consortium for Top Down P (2013) Proteoform: a single term describing protein complexity. Nat Methods 10(3):186–187. https://doi.org/10.1038/nmeth.2369
6. Schaffer LV, Millikin RJ, Miller RM, Anderson LC, Fellers RT, Ge Y, Kelleher NL, LeDuc RD, Liu X, Payne SH, Sun L, Thomas PM, Tucholski T, Wang Z, Wu S, Wu Z, Yu D, Shortreed MR, Smith LM (2019) Identification and quantification of proteoforms by mass spectrometry. Proteomics 19(10):e1800361. https://doi.org/10.1002/pmic.201800361
7. Gault J, Liko I, Landreh M, Shutin D, Bolla JR, Jefferies D, Agasid M, Yen HY, Ladds M, Lane DP, Khalid S, Mullen C, Remes PM, Huguet R, McAlister G, Goodwin M, Viner R, Syka JEP, Robinson CV (2020) Combining native and 'omics' mass spectrometry to identify endogenous ligands bound to membrane proteins. Nat Methods 17(5):505–508. https://doi.org/10.1038/s41592-020-0821-0
8. Bolla JR, Agasid MT, Mehmood S, Robinson CV (2019) Membrane protein-lipid interactions probed using mass spectrometry. Annu Rev Biochem 88:85–111. https://doi.org/10.1146/annurev-biochem-013118-111508
9. Calabrese AN, Radford SE (2018) Mass spectrometry-enabled structural biology of membrane proteins. Methods 147:187–205. https://doi.org/10.1016/j.ymeth.2018.02.020
10. Keener JE, Zambrano DE, Zhang G, Zak CK, Reid DJ, Deodhar BS, Pemberton JE, Prell JS, Marty MT (2019) Chemical additives enable native mass spectrometry measurement of membrane protein oligomeric state within intact nanodiscs. J Am Chem Soc 141(2): 1054–1061. https://doi.org/10.1021/jacs.8b11529
11. Eschweiler JD, Kerr R, Rabuck-Gibbons J, Ruotolo BT (2017) Sizing up protein-ligand complexes: the rise of structural mass spectrometry approaches in the pharmaceutical sciences. Annu Rev Anal Chem (Palo Alto, Calif) 10(1):25–44. https://doi.org/10.1146/annurev-anchem-061516-045414
12. Zhou M, Lantz C, Brown KA, Ge Y, Paša-Tolić L, Loo JA, Lermyte F (2020) Higher-order structural characterisation of native proteins and complexes by top-down mass spectrometry. Chem Sci. https://doi.org/10.1039/d0sc04392c
13. Peetz O, Hellwig N, Henrich E, Mezhyrova J, Dotsch V, Bernhard F, Morgner N (2019) LILBID and nESI: different native mass spectrometry techniques as tools in structural biology. J Am Soc Mass Spectrom 30(1):181–191. https://doi.org/10.1007/s13361-018-2061-4
14. Allison TM, Reading E, Liko I, Baldwin AJ, Laganowsky A, Robinson CV (2015) Quantifying the stabilizing effects of protein-ligand interactions in the gas phase. Nat Commun 6(1):8551. https://doi.org/10.1038/ncomms9551
15. Cleary SP, Prell JS (2019) Liberating native mass spectrometry from dependence on volatile salt buffers by use of Gabor transform. ChemPhysChem 20(4):519–523. https://doi.org/10.1002/cphc.201900022
16. Campuzano IDG, Robinson JH, Hui JO, Shi SD, Netirojjanakul C, Nshanian M, Egea PF, Lippens JL, Bagal D, Loo JA, Bern M (2019) Native and denaturing MS protein deconvolution for biopharma: monoclonal antibodies and antibody-drug conjugates to polydisperse membrane proteins and beyond. Anal Chem 91(15):9472–9480. https://doi.org/10.1021/acs.analchem.9b00062
17. Cleary SP, Li H, Bagal D, Loo JA, Campuzano IDG, Prell JS (2018) Extracting charge and mass information from highly congested mass spectra using Fourier-domain harmonics. J Am Soc Mass Spectrom 29(10):2067–2080. https://doi.org/10.1007/s13361-018-2018-7

18. Bern M, Caval T, Kil YJ, Tang W, Becker C, Carlson E, Kletter D, Sen KI, Galy N, Hagemans D, Franc V, Heck AJR (2018) Parsimonious charge deconvolution for native mass spectrometry. J Proteome Res 17(3): 1216–1226. https://doi.org/10.1021/acs.jproteome.7b00839
19. Cleary SP, Thompson AM, Prell JS (2016) Fourier analysis method for analyzing highly congested mass spectra of ion populations with repeated subunits. Anal Chem 88(12): 6205–6213. https://doi.org/10.1021/acs.analchem.6b01088
20. Tseng YH, Uetrecht C, Yang SC, Barendregt A, Heck AJ, Peng WP (2013) Game-theory-based search engine to automate the mass assignment in complex native electrospray mass spectra. Anal Chem 85(23): 11275–11283. https://doi.org/10.1021/ac401940e
21. Sivalingam GN, Yan J, Sahota H, Thalassinos K (2013) Amphitrite: a program for processing travelling wave ion mobility mass spectrometry data. Int J Mass Spectrom 345-347:54–62. https://doi.org/10.1016/j.ijms.2012.09.005
22. Morgner N, Robinson CV (2012) Massign: an assignment strategy for maximizing information from the mass spectra of heterogeneous protein assemblies. Anal Chem 84(6): 2939–2948. https://doi.org/10.1021/ac300056a
23. Ferrige AG, Seddon MJ, Green BN, Jarvis SA, Skilling J (1992) Disentangling electrospray spectra with maximum-entropy. Rapid Commun Mass Spectrom 6(11):707–711. https://doi.org/10.1002/rcm.1290061115
24. Lu J, Trnka MJ, Roh SH, Robinson PJ, Shiau C, Fujimori DG, Chiu W, Burlingame AL, Guan S (2015) Improved peak detection and deconvolution of native electrospray mass spectra from large protein complexes. J Am Soc Mass Spectrom 26(12):2141–2151. https://doi.org/10.1007/s13361-015-1235-6
25. Silvia MM, Sara C, Florian F, Jennifer S, Paul G, Ken C, Kai S, Jonathan B (2020) Optimisation of the use of sliding window deconvolution for comprehensive characterisation of trastuzumab and adalimumab charge variants by native high resolution mass spectrometry. Eur J Pharm Biopharm 158:83–95. https://doi.org/10.1016/j.ejpb.2020.11.006
26. Allison TM, Barran P, Benesch JLP, Cianferani S, Degiacomi MT, Gabelica V, Grandori R, Marklund EG, Menneteau T, Migas LG, Politis A, Sharon M, Sobott F, Thalassinos K (2020) Software requirements for the analysis and interpretation of native ion mobility mass spectrometry data. Anal Chem 92(16):10881–10890. https://doi.org/10.1021/acs.analchem.9b05792
27. Marty MT, Baldwin AJ, Marklund EG, Hochberg GK, Benesch JL, Robinson CV (2015) Bayesian deconvolution of mass and ion mobility spectra: from binary interactions to polydisperse ensembles. Anal Chem 87(8): 4370–4376. https://doi.org/10.1021/acs.analchem.5b00140
28. Reid DJ, Diesing JM, Miller MA, Perry SM, Wales JA, Montfort WR, Marty MT (2019) MetaUniDec: high-throughput deconvolution of native mass spectra. J Am Soc Mass Spectrom 30(1):118–127. https://doi.org/10.1007/s13361-018-1951-9
29. Walker LR, Marty MT (2020) Revealing the specificity of a range of antimicrobial peptides in lipid Nanodiscs by native mass spectrometry. Biochemistry 59(23):2135–2142. https://doi.org/10.1021/acs.biochem.0c00335
30. Kostelic MM, Ryan AM, Reid DJ, Noun JM, Marty MT (2019) Expanding the types of lipids amenable to native mass spectrometry of lipoprotein complexes. J Am Soc Mass Spectrom 30(8):1416–1425. https://doi.org/10.1007/s13361-019-02174-x
31. Walker LR, Marzluff EM, Townsend JA, Resager WC, Marty MT (2019) Native mass spectrometry of antimicrobial peptides in lipid nanodiscs elucidates complex assembly. Anal Chem 91(14):9284–9291. https://doi.org/10.1021/acs.analchem.9b02261
32. Campuzano ID, Li H, Bagal D, Lippens JL, Svitel J, Kurzeja RJ, Xu H, Schnier PD, Loo JA (2016) Native MS analysis of bacteriorhodopsin and an empty nanodisc by orthogonal acceleration time-of-flight, orbitrap and ion cyclotron resonance. Anal Chem 88(24): 12427–12436. https://doi.org/10.1021/acs.analchem.6b03762
33. Ahdash Z, Lau AM, Martens C, Politis A (2018) Analyzing protein architectures and protein-ligand complexes by integrative structural mass spectrometry. J Vis Exp 140: e57966. https://doi.org/10.3791/57966
34. Marty MT (2020) A universal score for deconvolution of intact protein and native electrospray mass spectra. Anal Chem 92(6): 4395–4401. https://doi.org/10.1021/acs.analchem.9b05272
35. Marty MT (2019) Eliminating artifacts in electrospray deconvolution with a SoftMax function. J Am Soc Mass Spectrom 30(10): 2174–2177. https://doi.org/10.1007/s13361-019-02286-4
36. Marty MT, Hoi KK, Robinson CV (2016) Interfacing membrane mimetics with mass

spectrometry. Acc Chem Res 49(11): 2459–2467. https://doi.org/10.1021/acs.accounts.6b00379

37. Pratesi A, Cirri D, Ciofi L, Messori L (2018) Reactions of auranofin and its pseudohalide derivatives with serum albumin investigated through ESI-Q-TOF MS. Inorg Chem 57(17):10507–10510. https://doi.org/10.1021/acs.inorgchem.8b02177
38. Kendrick E (1963) A mass scale based on CH2= 14.0000 for high resolution mass spectrometry of organic compounds. Anal Chem 35(13):2146–2154. https://doi.org/10.1021/ac60206a048
39. Marshall AG, Rodgers RP (2004) Petroleomics: the next grand challenge for chemical analysis. Acc Chem Res 37(1):53–59. https://doi.org/10.1021/ar020177t
40. Marty MT, Hoi KK, Gault J, Robinson CV (2016) Probing the lipid annular belt by gas-phase dissociation of membrane proteins in nanodiscs. Angew Chem Int Ed Engl 55(2):550–554. https://doi.org/10.1002/anie.201508289
41. Hunter JD (2007) Matplotlib: a 2D graphics environment. Comput Sci Eng 9(3):90–95. https://doi.org/10.1109/mcse.2007.55
42. Fraser GW (2002) The ion detection efficiency of microchannel plates (MCPs). Int J Mass Spectrom 215(1–3):13–30. https://doi.org/10.1016/s1387-3806(01)00553-x
43. Stengel F, Baldwin AJ, Painter AJ, Jaya N, Basha E, Kay LE, Vierling E, Robinson CV, Benesch JL (2010) Quaternary dynamics and plasticity underlie small heat shock protein chaperone function. Proc Natl Acad Sci U S A 107(5):2007–2012. https://doi.org/10.1073/pnas.0910126107

Chapter 13

Discovery of Unknown Posttranslational Modifications by Top-Down Mass Spectrometry

Jesse W. Wilson and Mowei Zhou

Abstract

Protein encoding genes can undergo modifications posttranscriptionally and posttranslationally, yielding many different "proteoforms." The chemical diversity of such modifications is known to be important biomarkers of function within biological systems but is not completely understood. Top-down mass spectrometry is a valuable tool for the characterization of proteoforms, especially for histones that have complex combinations of posttranslational modifications (PTMs). In this chapter, we present a top-down liquid chromatography-mass spectrometry experimental and data analysis workflow for the identification of novel, unexpected modifications on histones. Proteoforms of interest are first discovered using the "open" modification search in TopPIC. Then target proteoforms are manually confirmed using the data visualization tool—LcMsSpectator, part of the Informed-Proteomics package. The workflow can be very helpful in targeted PTM analysis and can be expanded to other types of proteins for discovery of unknown PTMs.

Key words Proteoform, Posttranslational modification, Histone, Liquid chromatography, Mass spectrometry, Top-down

1 Introduction

The advance in genome analysis has significantly improved our understanding of biology. However, the genomic information does not fully capture the chemical diversity of proteins, the primary effectors of biological functions. A single gene may produce many different forms of proteins, known as "proteoforms" that arise from posttranslational modifications (PTMs), sequence variance, and proteolysis [1, 2]. As an analytical technique, mass spectrometry (MS) is uniquely suited for the identification of proteoforms within complex mixtures [1, 3, 4]. In a typical "bottom-up" approach, proteins are digested with proteases (e.g., trypsin) and then separated by liquid chromatography coupled with MS detection [4]. This analytical scheme reduces the complexity for protein identification and structural analysis by dividing the protein into peptide fragments. With bottom-up, modifications are

Liangliang Sun and Xiaowen Liu (eds.), *Proteoform Identification: Methods and Protocols*,
Methods in Molecular Biology, vol. 2500, https://doi.org/10.1007/978-1-0716-2325-1_13,
© The Author(s), under exclusive license to Springer Science+Business Media, LLC, part of Springer Nature 2022

identified by measuring mass shifts on peptides relative to the sequence mass. Then, with tandem mass spectrometry (tandem-MS), peptides are further fragmented in the gas phase and sequenced to localize and identify PTMs. In comparison, a "top-down" approach to proteoform identification skips the protease digestion step and transfers intact proteins into the mass spectrometer for mass analysis and tandem-MS sequencing. The key advantage with the top-down approach is that the connectivity between PTMs and/or sequence variants, which would otherwise be lost due to enzymatic digestion, is maintained. This allows for the combination of PTMs on a particular proteoform to be precisely defined [3, 5].

Histones belong to a class of proteins (in all eukaryotes and most archaea) that form the protein core of nucleosomes for chromatin organization and are crucially involved in the dynamic regulation of gene expression [6, 7]. To fulfill their regulatory role, histones are heavily modified proteins where the modifications directly influence gene silencing or activation [7–9]. Changes to these modifications on one or more histones can therefore directly affect gene expression and play a role in disease states such as cancer. Multiple MS methods, including top-down, bottom-up, and middle-down (e.g., enzymatic digestion with less frequent cutters such as Asp-N or Glu-C that generate larger peptide fragments), have been utilized for histone proteoform profiling [8]. Each method has advantages and integration of the resulting data can be complementary [10, 11].

Currently, most proteomics software packages only consider PTMs that are specified in the analysis. Development of "open" modification algorithms have been used for both bottom-up [12–15] and top-down [16, 17], which report potential PTMs as mass shifts to the unmodified peptides or proteins. Accurate mass of the putative PTMs can then be used to assign their identities. Such analysis can potentially reveal "dark matter" in proteomics data—unexpected PTMs that are overlooked by regular data analysis methods. But novel PTMs still require manual confirmation to match fragmentation spectra and ensure a quality fit of the data.

Herein, we used two top-down histone datasets from published studies (one from mouse brain [18], one from sorghum leaf [19]) to illustrate the workflow of identifying unknown PTMs. The "open" search (i.e., not specifying PTMs) by TopPIC [16] was first used for discovering mass shifts on proteins that may correspond to novel proteoforms. Selected targets were then manually examined with the data visualization tool LcMsSpectator from the Informed Proteomics package [20]. Most PTMs induce mild or negligible change in retention time on the intact proteins; therefore, proteoforms that originated from the same protein/gene typically show up at similar retention times (except significant sequence truncations). The "feature map" in LcMsSpectator,

which displays all detected proteoforms by their intact mass and retention time window (MS^1 data), can assist in rapid identification of groups of related proteoforms that differ slightly in PTMs (known as "proteoform family" [21]). Once at least one proteoform in the group is confidently identified, other related proteoforms can be also assigned based on their mass differences with high mass accuracy. When MS^2 fragmentation data are available, these assignments can be further supported. Unique spectral features (e.g., isotope pattern, neutral losses) specific to PTMs provide additional evidence, especially for unusual PTMs. Using this workflow, we identified novel histone proteoforms with bromine and pyruvic acid PTMs, which were also confirmed by orthogonal bottom-up analysis at the tryptic peptide level. These novel PTMs may have been easily confused with other common PTMs or PTM combinations because of similar masses. The examples highlight the benefit of manual analysis, assisted by software tools, for the identification of proteoforms with novel PTMs. Future software development integrating the knowledge from such manual analysis would make the workflow more automated for high-throughput characterization of "dark matters" at the proteoform level.

2 Materials

2.1 Online Reverse-Phase Liquid Chromatography (RPLC) Coupled to Mass Spectrometry

1. A Waters NanoAquity or similar liquid chromatography system equipped with two binary nanoflow pumps (one for online trapping/desalting at 3 μL/min, the other for gradient at 0.3 μL/min). Follow manufacturer's instructions for the configurations.
2. RPLC separation (*see* **Note 1** for non-histone samples):
   (a) Trapping column: In-house packed short RPLC C18 column (Phenomenex Aeris wide pore 3.6 μm, column inner diameter 150 μm, outer diameter 360 μm, length 5 cm) or a C2 column (3 μm 300 Å, ID 150 μm, OD 360 μm, length 5 cm).
   (b) Analytical column: In-house packed short RPLC C18 column (Phenomenex 3 μm 300 Å, column inner diameter 75 μm, outer diameter 360 μm, length 70 cm).
   (c) RPLC mobile phase A composition (MPA): 0.1% formic acid (FA) in water; mobile phase B (MPB): 0.1% FA in acetonitrile (all solvents should be of MS grade).

2.2 Mass Spectrometry

1. A Thermo Scientific Orbitrap Fusion Lumos or Eclipse (or similar) equipped with higher-energy collisional dissociation (HCD) and electron transfer dissociation (ETD) to acquire high-resolution MS^1 and MS^2 spectra.

2. The raw data discussed here can be downloaded from the following addresses.
   (a) Mouse histone datafile (MZ20160222DS_histone48.raw) is at MendeleyData (https://doi.org/10.17632/k72vc6vxyv.1).
   (b) Sorghum histone datafile (Sorghum-Histone0810162L01.raw) is at the ProteomeXchange Consortium via the PRIDE partner repository [22] with dataset identifier PXD014660 (http://www.ebi.ac.uk/pride).

2.3 Data Analysis Software

1. MSConvert (http://proteowizard.sourceforge.net/tools.shtml).
2. Informed-Proteomics package(https://github.com/PNNL-Comp-Mass-Spec/Informed-Proteomics) [20].
3. TopPIC Suite (http://proteomics.informatics.iupui.edu/software/toppic/) [16].
4. (Optional) Molecular Weight Calculator (https://omics.pnl.gov/software/molecular-weight-calculator) for calculating theoretical isotope distributions.

3 Methods

3.1 Online RPLC Coupled with Mass Spectrometry Detection

1. Inject 1–2 μg of purified histones (*see* **Note 2** for histone sample information) onto the dual pump nanoflow LC (e.g., Waters NanoAcquity) coupled to the MS instrument. *See* **Note 3** for further information about sample loading.
2. Load the injected samples onto the trapping column using 1% MPB for 5–10 min at 3 μL/min to flush away the remaining salts from injected histones.
3. Switch the eluent flow from the trapping column to the analytical C18 column. Then start the gradient and MS acquisition (*see* **Note 4**). With the RPLC condition described here, histone proteins typically elute between 30% and 50% MPB. To optimize separation of core histone proteins use a shallow gradient over at least 100 min between 30% and 50% MPB (*see* **Note 5**).
4. Set up a data-dependent acquisition method, alternating fragmentation for the same precursor with both HCD and ETD (*see* **Note 6**).
5. Acquire MS^1 precursor ion mass spectra of intact histones with recommended settings: a resolution of 120k, averaged over 4 microscans, an AGC target of 1E6, and a max injection time of 50 ms (*see* **Note 7**).
6. Acquire data-dependent MS^2 fragment ion mass spectra with recommended settings: a resolution of 120k with 1 microscan,

AGC target of 1E6, isolation window of 0.6 Da (*see* **Note 8**). MS^2 scans should alternate ETD (reaction time ~20 ms, max injection time 500 ms) and HCD (28% normalized collision energy with ±5% stepped energy, max injection time 100 ms). Set dynamic exclusion to 120 s with a mass tolerance of ±0.7 Da. Exclude charge states lower than 5 and undetermined charge states (*see* **Note 9**). Activate MS^2 on single charge state per precursor only.

3.2 LC-MS Data Processing

1. Figure 1 displays a general workflow for using TopPIC and LcMsSpectator for LC-MS data processing and proteoform identification. Begin by obtaining the protein database in FASTA format for the corresponding sample. For the examples given here, the *Mus musculus* database and the *Sorghum bicolor* database from UniProt were used.
2. Convert MS raw data files to mzML format using MSConvert.
3. Download the TopPIC suite and run the program from the graphical interface or the command line mode following the tutorial on the TopPIC website (*see* **Note 10**).
4. Use TopFD from the TopPIC suite to deconvolute the mzML-formatted spectra from **step 2**. Default parameters can be used except that the "precursor window" (−w) needs to be reduced to 1 *m/z* due to the narrow isolation window.
5. Use TopPIC from the TopPIC suite to identify proteoforms. Default parameters can be used. Set the spectrum and proteoform cutoff type to FDR (false discovery rate) and FDR cutoff value to 0.01 or as desired. The "proteoform error tolerance" can be set to 5 Da (*see* **Note 11**). Load the protein database FASTA file from **step 1** and the "_ms2.msalign" file from the deconvolution process in **step 4** to start the search.
6. To view the identified proteoforms, open the "*****_proteoform.csv" file (after combining proteoform identifications with similar masses) or "*****_prsm.csv" file (all proteoform-spectrum matches). Visual display of proteoform sequence coverage can be seen in the Topview module found in the output folder under the "*****_html" file.
7. The generated proteoforms list from TopPIC annotates PTMs on proteins as mass shifts in square brackets. The residue (or range of residues) where the mass shift was identified is annotated with parenthesis. Identify interesting targets for manual confirmation in the following steps, for example, a proteoform with low FDR (or low *E*-value, *P*-value) and a mass shift that cannot be explained by common PTMs (*see* **Note 12**).
8. Prepare the data for visualization and manual analysis in LcMsSpectator. Perform the following steps (optional but

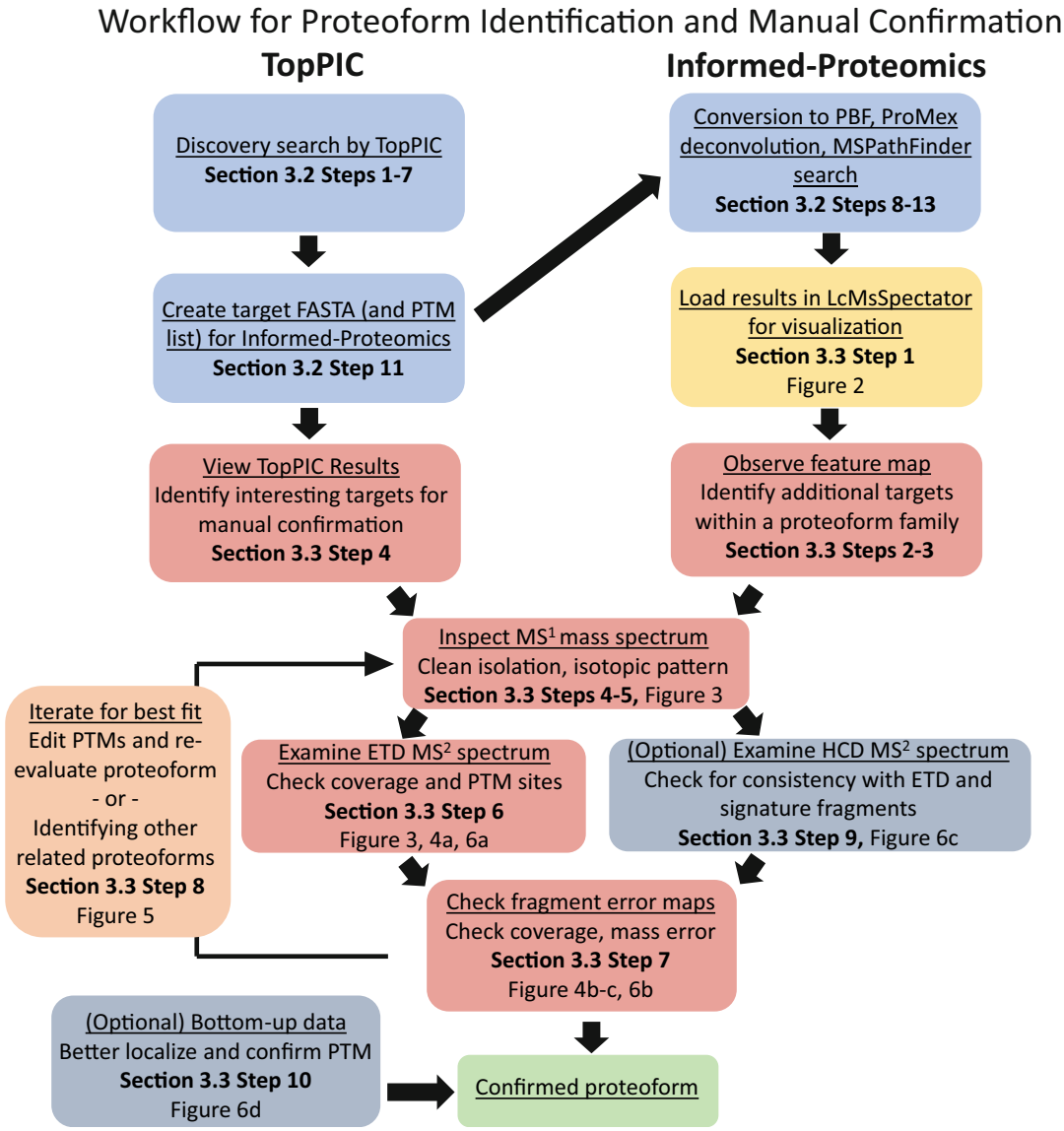


Fig. 1 Workflow scheme for proteoform discovery by TopPIC and manual confirmation by LcMsSpectator. Method steps and related figures are noted. Mouse brain histone data were used in Figs. 2, 3, 4, and 5 for demonstrating identification of brominated histone H2A.X and H4 proteoforms. Sorghum leaf histone data were used in Fig. 6 as an example to highlight the complementarity of HCD data to ETD data for identifying unexpected PTMs with unique neutral losses

recommended) to process the data using the Informed-Proteomics package (*see* **Note 13**).

9. Convert the raw/mzML file to a PBF file. This can be performed from LcMsSpectator simply with the "Open Raw/MzML" button on the home screen. Then click on the "XIC View" tab and press the "Create PBF now" button (*see* **Note 14**).

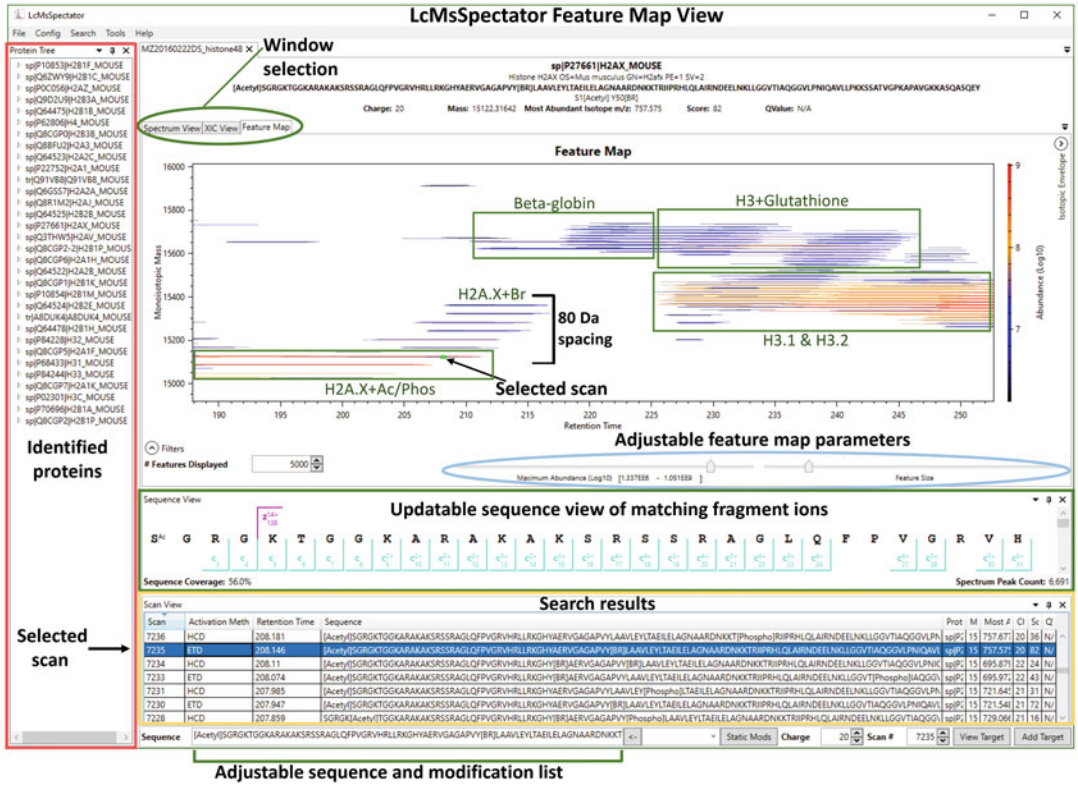


Fig. 2 Representative LC-MS feature map view using LcMsSpectator. The horizontal axis is the LC retention time and the vertical axis is the deconvoluted monoisotopic mass. The selected region highlights a series of unusual histone H2A.X proteoforms from mouse brain with characteristic ~80 Da spacing between each feature that forms a "ladder." The selected scan (number 7235) is highlighted by the green rectangle in the feature map. Sequence, PTM, and scan information are shown above the feature map. Sequence coverage view is shown below the feature map. The proteoform sequence at the bottom of the window can be edited to manually adjust fit. All the displays will update in real time

10. Use the ProMex module in command line to deconvolute the MS[1] data and output a ms1ft file, which includes the feature list (each feature is a unique combination of mass and retention time).
11. Using the identified protein list from TopPIC in **step 6** creates a focused FASTA file to reduce the search space for MSPathFinder in the Informed-Proteomic package.
12. Create a targeted modification list for searching histone PTMs using the example file provided as template. For histones, the common PTMs are as follows: lysine acetylation, lysine monomethylation, lysine dimethylation, lysine trimethylation, serine/threonine/tyrosine phosphorylation, N-terminal acetylation, methionine/cystine oxidation. In the case of plant/algae histones, N-terminal mono-, di-, and trimethylation should be added. Also add any putative PTMs identified from manual

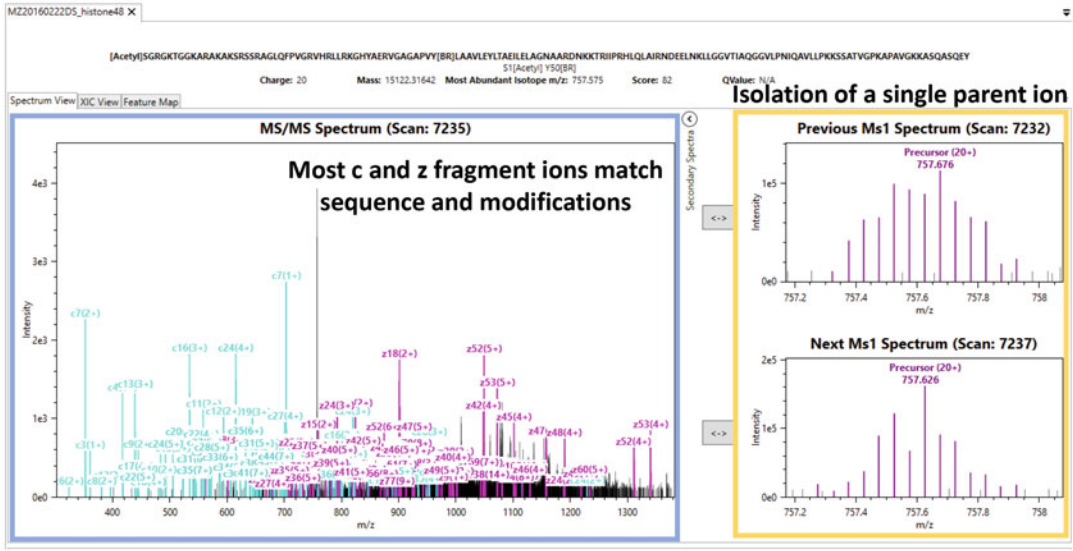


Fig. 3 Spectrum view in LcMsSpectator of H2A.X histone assigned with N-terminal acetylation and bromine modification. The left panel shows the MS^2 fragment matches. The right panel shows the precursor ion within the isolation window. Purple peaks indicate the experimental data matched to the theoretical isotope distribution of the assigned proteoform. The x-axes of the MS^1 spectra on the right default to be the isolated *m/z* window for MS^2. In this example, the target ions were in the middle of the isolation window and free of significant interference

analysis from **step 7**. Minimize the number of PTMs by removing uncommon PTMs in the sample to reduce search time (*see* **Note 15**).

13. Run the MSPathFinder module to identify proteoforms from the created PBF file in **step 9**, the ms1ft file in **step 10**, the targeted FASTA file from **step 11**, and the modification list from **step 12**. Default parameters can be used.

3.3 Identification and Confirmation of Unknown Proteoforms

1. Load all the required files into LcMsSpectator: These include the search parameter file (*.param) and results (*.tsv) from MSPathFinder (Subheading 3.2, **step 13**), the LC-MS data in PBF format (Subheading 3.2, **step 9**), the focused FASTA file (Subheading 3.2, **step 11**), and the deconvolved Promex ms1ft file (Subheading 3.2, **step 10**).
2. Observe feature map - display options: Navigate to the "Feature Map" in LcMsSpectator by clicking on the window selection tab (the three tabs near top left, the default tab upon file loading is "Spectrum View"). Identify regions of interest based on groupings of proteoforms. Figure 2 displays the feature map with example data from mouse histones (zoomed into H2A and H3 regions) and scan 7235 selected. Common PTMs can be initially identified by analyzing mass shifts between features from the feature map. For example, methylation causes a mass

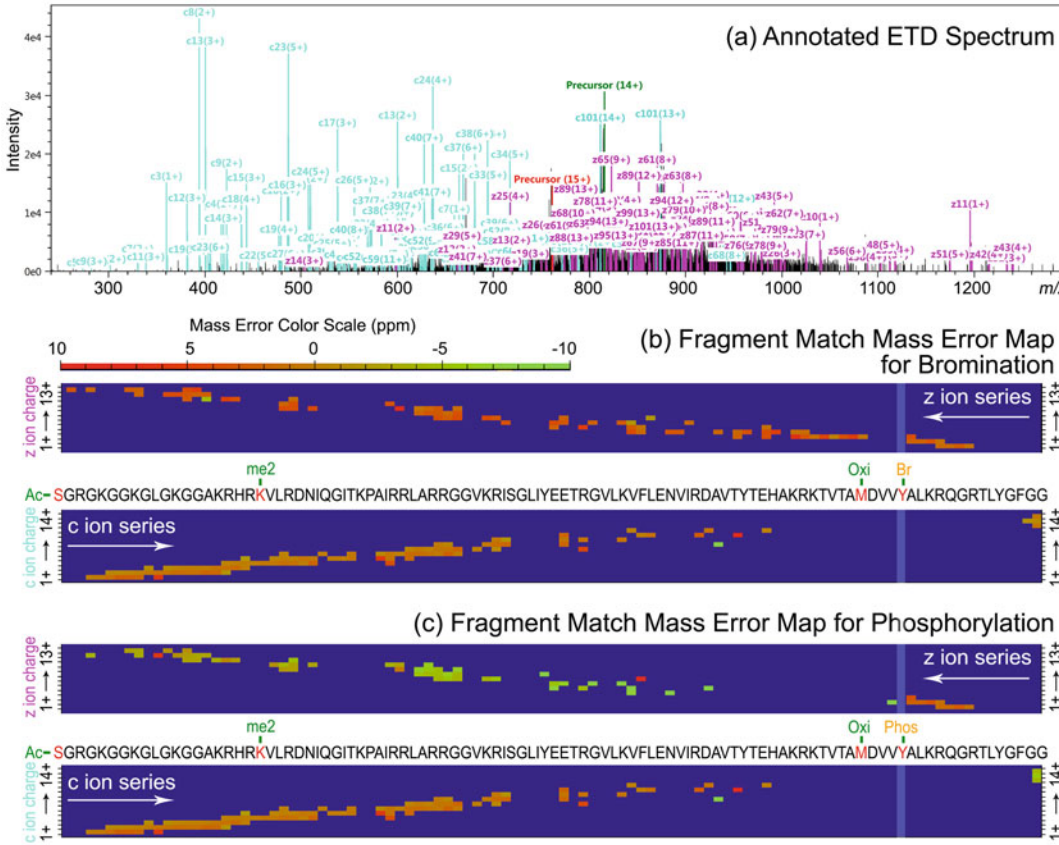


Fig. 4 The ETD fragment mass spectrum and error maps confirm tyrosine 88 bromination over phosphorylation for a histone H4 proteoform in mouse brain sample. (**a**) Annotated ETD spectrum for a brominated histone H4 proteoform S1AcK20me2M84OxidationY88Br. (**b**) Fragment mass error map for the matching H4 proteoform with Y88Br in comparison with phosphorylation in (**c**). The error map has x axis as the protein sequence, y axis as the fragment charge, and color represents the mass error of matched fragments in ppm. Higher levels of mass error are seen in the z-ion series than with bromination. (Reprinted with permission from [18]. Copyright 2017 American Chemical Society)

shift of 14.01 Da, giving rise to the dense, parallel features from H3 with different numbers of methylation. In addition, the features above the identified H2A.X proteoforms showed an unusual "ladder" pattern, with each feature spaced by ~80 Da. Because these unusual features are closely associated with major H2A.X proteoforms, they can be assigned tentatively as part of the H2A.X "proteoform family." Scroll the middle mouse button to zoom in/out of the feature map. Right click and check "show manual adjustment" to adjust the displayed axis ranges with typable values.

3. Observe feature map - identify proteoform family: Use the feature map to help identify other related proteoforms. Right click on the feature map and select "Show Identified MS/MS

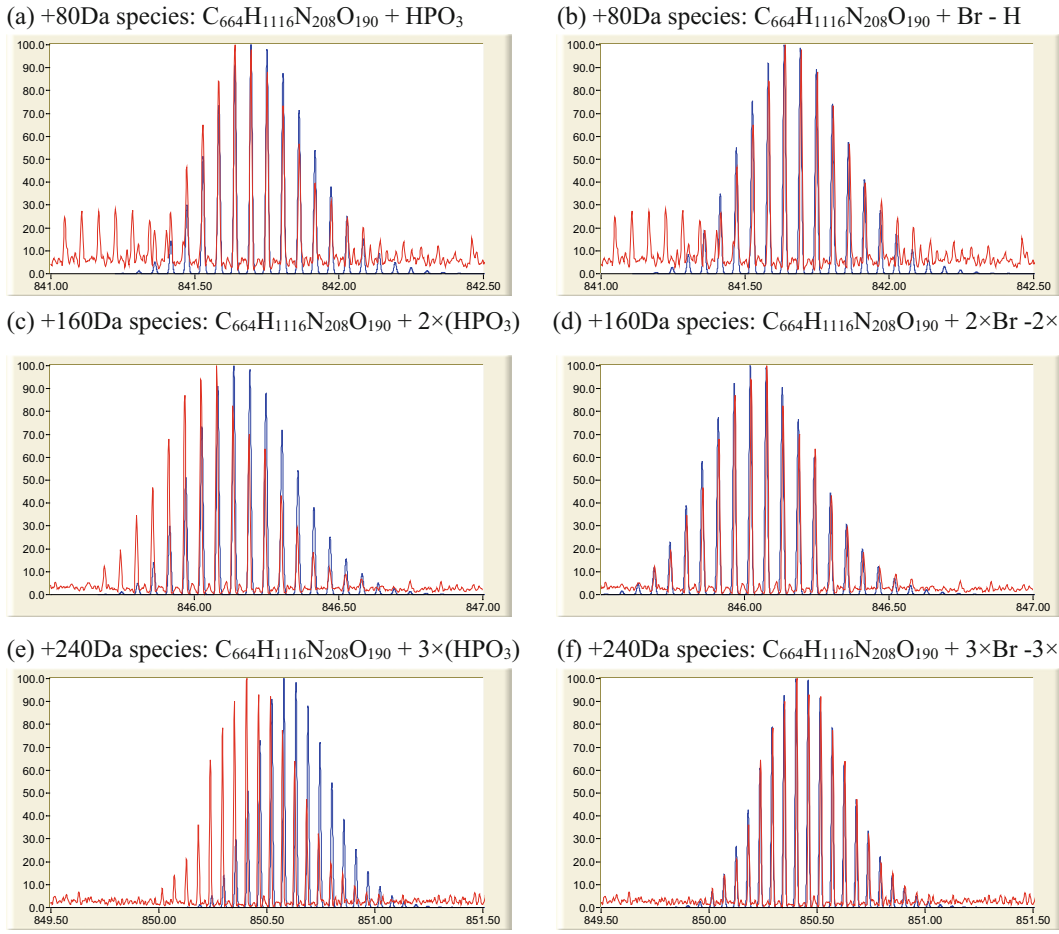


Fig. 5 MS^1 spectra of the major H2A.X species containing one acetylation and a single (**a**, **b**), double (**c**, **d**), or triple (**e**, **f**) ~80 Da PTM. The red traces display the experimental isotope distribution while the theoretical distribution is shown in blue. The theoretical distribution with bromination of the H2A.X histone (**d**, **f**) matches more closely with the experimental isotope distribution than phosphorylation (**c**, **e**). For the single ~80 Da PTM, the bromination (**b**) only matched slightly better than the phosphorylation (**a**), where the theoretical distribution is slightly shifted to the right. Typically, this small shift is not significant in automatic analysis, given the experimental fluctuation of isotope distributions. Manual analysis of the proteoform family in the high-resolution MS^1 spectra suggests the ~80 Da species is bromine instead of phosphorylation. Figures were produced using the Molecular Weight Calculator software. (Reprinted with permission from [18]. Copyright 2017 American Chemical Society)

scans." This will draw "X" over features with MS^2 spectra and proteoform identifications by Informed-Proteomics (*see* **Note 16**). Similarly, selecting "Show UnIdentified MS/MS scans" will draw dots over features with MS^2 spectra but not with proteoform identifications. This is a useful tool for quickly looking for MS^2 spectra of target features for manual confirmation of tentative assignments by MS^1. For example, activate

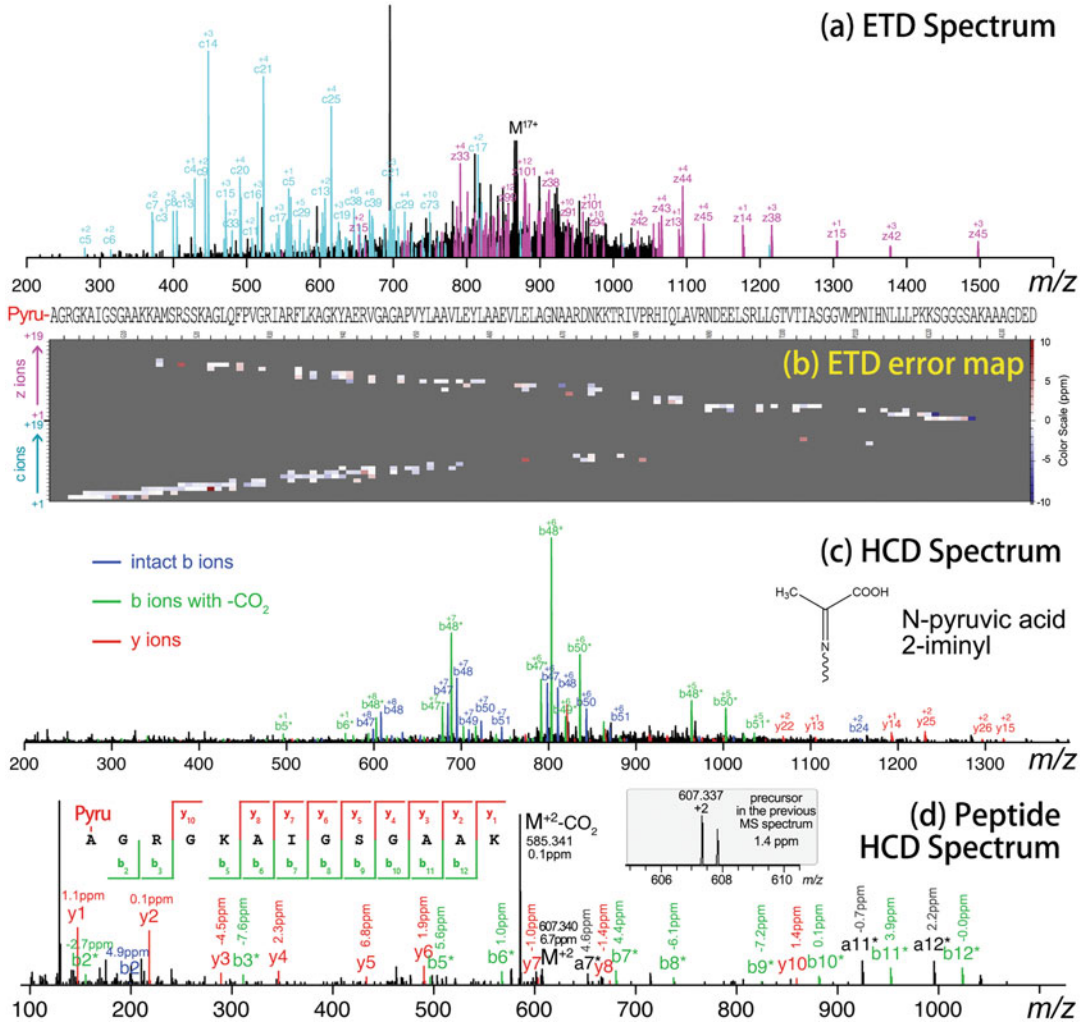


Fig. 6 Confirmation of N-pyruvic acid 2-iminyl modification on 14 kDa H2A (UniProt accession: C5XAT9) from sorghum. (**a**) Top-down ETD fragmentation spectrum with matching fragment ions. (**b**) Corresponding error map for the ETD spectrum shown in (**a**) exhibiting good, consistent mass accuracy. (**c**) HCD spectrum of the same precursor as in (**a**). The major b-ions are mass shifted by the neutral loss of CO_2 shown in green, while the fragment ions in blue maintain the intact modification. (**d**) HCD spectrum of the tryptic peptide from the same proteoform. The inset on the left shows the coverage map along the peptide sequence, and the inset on the right shows the zoom-in view of the precursor ion from the previous MS^1 spectrum. The mass errors (ppm) for each manually assigned ion are labeled on the top of each peak. The peptide spectrum showed neutral loss peaks (b-ions in green) consistent with the assigned PTM by the top-down data. (Figure reprinted from [19]. Copyright 2020 with permission from Elsevier)

the display and look for other spectra for H2A.X proteoforms near scan 7235 (e.g., scans 7244 and 7249).

4. View TopPIC results: To manually examine the H2A.X proteoforms, first look for H2A.X proteoform matches in the TopPIC results (e.g., *_prsm.csv file). Look up identification by scan

numbers (such as scan 7235, *see* **Note 17**). Click on the proteoforms or spectra of interest in the "Scan View" window where the search results are displayed (Fig. 2, yellow box). Check for consistency between TopPIC and Informed-Proteomics identifications. They should report the same protein and terminal truncations (if any) for high confidence matches. The following steps will correlate the mass shift in TopPIC to the PTM assignments in Informed-Proteomics (*see* **Note 18**).

5. Inspect MS^1 mass spectrum: Check the quality of precursor ion match in the MS^1 spectra and look for potential co-isolation of other species (*see* **Note 19**). Figure 3 displays the "Spectrum View" window in LcMsSpectator for a H2A.X proteoform in scan 7235. Here, the previous MS^1 spectrum and the next MS^1 spectrum are consistent (Fig. 3, yellow box), ensuring that the isolated MS^2 spectrum corresponds to the same species. The default view in the MS^1 spectra is displayed at the width of the isolation window in the experiment. An ideal match should have the full precursor isotope distribution in the center of the window.
6. Examine ETD MS^2 spectrum: A good match will account for most of the major fragment ions (Fig. 3, blue box). If the match is not ideal, the sequence and assigned PTMs at the bottom of the LcMsSpectator window (Fig. 2) can be manually changed. All displays (e.g., spectral match, theoretical mass, feature map) will update in real time to reflect changes in sequence and PTMs (*see* **Note 20**). Iterate this process to find the proteoform that best fits to the data. The zoom in/out control is the same as the feature map, described in **step 2**.
7. Check fragment error maps: Access the "Error map" by right-clicking and selecting "Open Error/Coverage Map" in the fragmentation spectrum on the "Spectrum View" (*see* **Note 21**). When the correct proteoform is assigned, a consistent level of mass error across the sequence should be seen with each identified fragment ion (*see* **Note 22**). For example, the H4 proteoform in scan 1164 of the mouse brain data (Fig. 4) was assigned as N-terminal acetyl, K20 dimethyl, M84 oxidation, and a ~78 Da PTM. By manually fitting the PTMs as described in **step 6** above, the accurate mass of the PTM was determined to be 77.9 Da at Y88, which showed good coverage and mass accuracy (Fig. 4a, b). However, this mass shift was matched to bromination in UniMod, which is quite unusual. A more common PTM, phosphorylation (79.97 Da), may also explain the data considering potential deisotoping error (initially assigned to phosphorylation by automated searches with a common PTM list). However, the fragment error map of phosphorylated proteoform showed that the matched z-ions

have significantly lower mass accuracy than c-ions. Some smaller z-ions after Y88 were also missing (Fig. 4c).

8. Identifying other related proteoforms: Error map analysis (same process as described in **step 7** above) on the H2A.X proteoform (scan 7235 in Fig. 2) suggested N-terminal acetyl and ~78 Da PTM near residue 50. Due to lack of coverage around residue 50 in the ETD data, both phosphorylation and bromination can fit the MS^2 spectrum. Other ETD data with related proteoforms (e.g., scan 7244 and 7249) also had limited coverage. To differentiate the two potential PTMs without sufficient MS^2 data, examine the precursor isotope distributions. Figure 5 shows the experimental high-resolution MS^1 spectra for the 18+ and ~80 Da singly mass shifted (same proteoform as scan 7235), doubly shifted (~160 Da, scan 7244), and triply shifted (~240 Da, scan 7249) proteoforms overlaid with the theoretical distributions corresponding to the addition of phosphorylation or bromination. Multiple addition of bromination (Fig. 5b, d, f) matches much more closely with the experimental MS^1 spectra than the addition of phosphorylation series (Fig. 5a, c, e), and is thus assigned as the major PTM on the H2A.X proteoform family in this region.
9. (Optional) Examine HCD MS^2 spectrum: The HCD data of the same precursor (*see* **Note 23**) should be consistent with the ETD data (*see* **Note 24**). In addition, HCD spectrum can provide crucial support for novel PTM assignments. As an example, Fig. 6a–c displays the identification of a novel N-pyruvic acid 2-iminyl modification on 14 kDa histone H2A from sorghum (ETD scan 2920 and HCD scan 2922 from the sorghum top-down dataset). HCD caused the neutral loss of CO_2, suggesting that the PTM likely has carboxylic acid group and further confirms the assignment (also *see* **Note 25**).
10. (Optional) Perform bottom-up peptide LC-MS analysis to further confirm the PTM: For example, the N-pyruvic acid 2-iminyl modification on 14 kDa histone H2A in sorghum also showed the signature neutral loss on the tryptic peptide (Fig. 6d).

4 Notes

1. Histones are in the mass range of 11–16 kDa, which can be well separated by C18 RPLC. The method here can be applied to other small proteins. However, for proteins larger than 20–30 kDa, or more hydrophobic ones, shorter chain length chromatography material (e.g., C2) should be considered. The same MPA and MPB can still be used. We have applied the similar top-down workflow with C2 column to characterize

intact chromophore PTMs on phycobilisome proteins (17 ~ 20 kDa) in cyanobacterium [23].

2. Histone purification protocols generally consist of a nuclei isolation step and a histone extraction step [24]. Histone purification from nuclei often relies on the high positive charge of histones, so generic protocols (e.g., acid extraction, ion exchange chromatography) can be applied. Nuclei isolation protocols can vary significantly between different sample types. For example, detergent concentration to selectively lyse cells while keeping nuclei intact needs to be optimized. For brevity, we did not list the detailed steps here specific to mouse brain and sorghum leaf samples. Many commercial kits are available for nuclei isolation, and some are available specifically for histone extraction. Notably, some PTMs are acid labile and thus protocols involving strong acid should be avoided if these PTMs are the targets of study [25]. For method development, commercial histone standards (e.g., HeLa histone from ActiveMotif). For the mouse brain sample, we used the commercial histone purification kit from ActiveMotif, following the manufacturer's instructions [18]. For the sorghum plant leaf samples, we used sucrose density gradient to enrich nuclei from other tissue components [19, 26]. Then a generic weak cation exchange (WCX) method purifies histones from isolated nuclei. The WCX method is generic for histone purification and can also be used to remove polymer contamination (commonly seen in histone samples) and improve top-down analysis.
3. The optimum sample load depends on the size of the analytical column and the sensitivity of the instrument. To prevent carryover, it is important to not overload the column and to perform blank LC runs between samples. For single protein (or highly purified proteins), the loading amount can be reduced by 5–10 times.
4. The dead volume in the setup described here corresponds to ~20 min of delay of protein elution at the MS. It is recommended to set up a delay in starting MS data acquisition after LC gradient start.
5. A long gradient (e.g., 400 min) increases the number of MS^2 spectra that can be collected leading to higher proteoform identification, especially for instruments with low speed/sensitivity that require longer signal averages.
6. ETD generally yields high sequence coverage at the termini of proteins. The majority of the PTMs on histones are on the N-terminus, making ETD highly effective. HCD tends to give sequence-specific fragments, which provides complementary datasets to ETD. For some labile PTMs, HCD can generate

diagnostic fragments (e.g., neutral losses) that help confirm the chemical identify of PTMs.

7. More microscans help improve the signal-to-noise ratio of low abundance species. However, too many microscans slow down the acquisition and may reduce coverage.
8. A narrow isolation window can be especially useful for isolating methylated forms of histones, which are only 14 Da apart (which translates to m/z difference <1 for most detected peaks).
9. Most histones can be readily resolved isotopically and assigned with correct charge "on the fly" for MS^2. However, bigger proteins or low-intensity proteins may have bad isotopic distribution, resulting in unassigned charge in single MS^1 spectra. If targeting high mass or very low abundance proteins, exclusion of undetermined charge states can be turned off to increase the chance of selecting them for MS^2.
10. Alternatively, MashExplorer [27] can be used as a graphical interface for spectral deconvolution and proteoform identification using TopPIC.
11. This setting will combine proteoforms with similar masses (e.g., within 5 Da) and reduce redundant proteoforms that are solely results of deisotoping error. But proteoforms with PTMs at different sites (isomers) will also be combined. Use this function with caution.
12. UniMod (http://www.unimod.org) compiles a comprehensive database of PTMs reported in the literature. Login as guest to view the list of PTMs sorted by mass. Look for the mass shift identified by TopPIC for potential assignments.
13. For manual validation of single spectrum, other software tools such as ProSightLite (http://prosightlite.northwestern.edu/) and TDValidator (https://proteinaceous.net/) can be used. LcMsSpectator can also directly load MS raw data, and then you can manually type in sequence for specific spectra for confirming selected identifications. However, it is recommended to at least create the feature file using ProMex to enable the "feature map" and visualize all detected proteoforms in the full LC run. The feature map can help identify proteoforms even without MS^2 data. Note that LcMsSpectator is a Windows program. PSpecteR (https://github.com/EMSL-Computing/PSpecteR) is a similar program built in R that can be used for non-Windows users.
14. PBF files are required to visualize extracted ion chromatogram (XIC) and feature map, but not required to load raw data. Alternatively, the conversion can be performed using command line tool "PbfGen.exe." All instructions for using the

software (PbfGen, ProMex, and MSPathFInder) in command line mode are on the Informed-Proteomics GitHub page.

15. Minimizing FASTA size and PTM list are necessary to successfully perform the MSPathFinder analysis. If the search becomes too computationally intensive, omit rare PTMs. Confirmation of rare PTMs can still be performed manually by examining target spectra. The MSPathfinder search here is mostly for identifying major, common proteoforms for display during manual analysis.
16. The protein filter in Scan View (*see* **Note 17**) can be activated here to only display the identified MS^2 spectra for selected proteins. Multiple proteins can be selected at a time in the filter.
17. The spectral matches in LcMsSpectator can also be sorted by each column by clicking on the column headers. Right click on the column header to open filter settings. This will help find the target proteins, proteoforms, or scan numbers.
18. The "open" search in TopPIC annotates PTMs as mass shifts on a given identified sequence, these results are crucial for finding unknown/unexpected PTMs in the form of uncommon mass shifts. Default search only allows 1 unexpected mass shift (max allowed is 2); therefore, summed mass of multiple PTMs will be annotated in the report. Informed-Proteomics has the tendency to force match unknown PTMs to a combination of known PTMs and terminal truncations. So cross-referencing the two can help improve confidence in identification. TopPIC also has a function to specify target PTMs using a PTM list file. When it is activated, TopPIC will fit the mass shift with known PTMs in the final report. But this may also result in forced match to known (combinations of) PTMs.
19. Sometimes the precursor ions do not have good signal and may be misassigned in the deconvolution. Another potential issue of co-isolating multiple species can be common, especially when a wide isolation window is used. Some software tools have options (e.g., ProSightPC) to report multiple proteoforms per spectrum to account for such "chimera" fragmentation spectra. LcMsSpectator and TopPIC currently only output the best matched proteoform per spectrum.
20. This allows the user to guess and check modifications to see how well they match the fragmentation spectrum. Protein sequences are represented by single letter amino acid notations. PTMs are placed next to the resides with the PTM names in square brackets (e.g., [Oxidation]). The names must be predefined to be recognized. The PTM list can be found in "Config"—"manage modifications" in the top menu. Custom modifications can be added. PTMs can also be represented simply with numbers next to the residues (e.g., +15.99,

−18.01 for oxidation and water loss, respectively). No bracket is needed. Ensure that PTM annotation format is consistent. They must be annotated all with PTM names in square brackets, or all with numbers. Changing the sequence in LcMsSpectator won't change and save the identification results (only for real-time display).

21. The default error map combines fragments of all charge states. Uncheck the "Combine charge states" box in the left bottom corner to display fragments of all charge states, same as the ones shown in Figs. 4 and 6. The "Table" tab has the information of the heatmap including all theoretical fragments and matched ions. Because the charge generally scales with fragment mass, real fragments should be on a slope in the error map with individual charge states displayed (not combined). Sometimes, random match to noise peaks can occur, but with unusually high or low charge. Such random matches will not be on the correct "slope" of real fragments in the error map.
22. Error maps can be used to differentiate PTMs with small mass differences such as trimethylation and acetylation, which only vary by 36 mDa. Abrupt changes in mass error along a protein sequence are suggestive of a misidentification of a closely fitting PTM [28].
23. Newer generation Orbitraps do not necessarily alternate ETD and HCD spectra in a fixed order (scan numbers are not next to each other). Find the HCD scan number with the same precursor *m/z* as ETD in the same cycle.
24. ETD generally does not cleave labile PTMs and provides high-terminal sequence coverage. The accurate mass of the PTM can be determined and localized. HCD usually provides lower sequence coverage and induces fragmentation of labile PTMs (e.g., phosphorylation, glycosylation). Some PTMs will produce signature fragments in HCD and further support the assignment. In the brominated H2A.X data, no significant neutral loss was seen, suggesting the proteoforms are unlikely phosphorylated. Some of the y-ions for brominated H4 containing the bromine PTM showed the unique isotopes for bromine, confirming the assignment [18]. However, the brominated H2A.X proteoforms did not produce enough resolved fragments to confirm the exact site and unique isotope pattern. The assignment was largely based on the MS^1 data and the pattern of proteoform family in the feature map.
25. The N-pyruvic acid 2-iminyl modification was identified by manually cross-referencing TopPIC and LcMsSpectator results using the same workflow in **steps 4–7**. The Informed-Proteomics search results contained an unusual combination of acetylation, methylation, and oxidation on histone H2A.

TopPIC and manual analysis identified a mass shift of 70.005 Da near the N-terminus for the intact protein.

Acknowledgments

We acknowledge the support by the Intramural program at EMSL (grid.436923.9), a DOE Office of Science User Facility sponsored by the Biological and Environmental Research program and operated under Contract No. DE-AC05-76RL01830.

References

1. Smith LM, Kelleher NL, Consortium for Top Down P (2013) Proteoform: a single term describing protein complexity. Nat Methods 10(3):186–187. https://doi.org/10.1038/nmeth.2369
2. Smith L, Agar J, Chamot-Rooke J, Danis P, Ge Y, Loo J, Pasa-Tolic L, Tsybin Y, Kelleher N (2020) The human proteoform project: a plan to define the human proteome. Preprints 2020100368. https://doi.org/10.20944/preprints202010.0368.v1
3. Schaffer LV, Millikin RJ, Miller RM, Anderson LC, Fellers RT, Ge Y, Kelleher NL, LeDuc RD, Liu X, Payne SH, Sun L, Thomas PM, Tucholski T, Wang Z, Wu S, Wu Z, Yu D, Shortreed MR, Smith LM (2019) Identification and quantification of proteoforms by mass spectrometry. Proteomics 19(10):e1800361. https://doi.org/10.1002/pmic.201800361
4. Han X, Aslanian A, Yates JR 3rd (2008) Mass spectrometry for proteomics. Curr Opin Chem Biol 12(5):483–490. https://doi.org/10.1016/j.cbpa.2008.07.024
5. Donnelly DP, Rawlins CM, DeHart CJ, Fornelli L, Schachner LF, Lin Z, Lippens JL, Aluri KC, Sarin R, Chen B, Lantz C, Jung W, Johnson KR, Koller A, Wolff JJ, Campuzano IDG, Auclair JR, Ivanov AR, Whitelegge JP, Pasa-Tolic L, Chamot-Rooke J, Danis PO, Smith LM, Tsybin YO, Loo JA, Ge Y, Kelleher NL, Agar JN (2019) Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. Nat Methods 16(7):587–594. https://doi.org/10.1038/s41592-019-0457-0
6. Bannister AJ, Kouzarides T (2011) Regulation of chromatin by histone modifications. Cell Res 21(3):381–395. https://doi.org/10.1038/cr.2011.22
7. Venkatesh S, Workman JL (2015) Histone exchange, chromatin structure and the regulation of transcription. Nat Rev Mol Cell Biol 16(3):178–189. https://doi.org/10.1038/nrm3941
8. Moradian A, Kalli A, Sweredoski MJ, Hess S (2014) The top-down, middle-down, and bottom-up mass spectrometry approaches for characterization of histone variants and their post-translational modifications. Proteomics 14(4–5):489–497. https://doi.org/10.1002/pmic.201300256
9. Tropberger P, Schneider R (2013) Scratching the (lateral) surface of chromatin regulation by histone modifications. Nat Struct Mol Biol 20(6):657–661. https://doi.org/10.1038/nsmb.2581
10. Liu X, Dekker LJ, Wu S, Vanduijn MM, Luider TM, Tolic N, Kou Q, Dvorkin M, Alexandrova S, Vyatkina K, Pasa-Tolic L, Pevzner PA (2014) De novo protein sequencing by combining top-down and bottom-up tandem mass spectra. J Proteome Res 13(7):3241–3248. https://doi.org/10.1021/pr401300m
11. Schaffer LV, Anderson LC, Butcher DS, Shortreed MR, Miller RM, Pavelec C, Smith LM (2020) Construction of human Proteoform families from 21 tesla Fourier transform ion cyclotron resonance mass spectrometry top-down proteomic data. J Proteome Res. https://doi.org/10.1021/acs.jproteome.0c00403
12. Chick JM, Kolippakkam D, Nusinow DP, Zhai B, Rad R, Huttlin EL, Gygi SP (2015) A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. Nat Biotechnol 33(7):743–749. https://doi.org/10.1038/nbt.3267
13. Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI (2017) MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based

proteomics. Nat Methods 14(5):513–520. https://doi.org/10.1038/nmeth.4256

14. Na S, Bandeira N, Paek E (2012) Fast multi-blind modification search through tandem mass spectrometry. Mol Cell Proteomics 11(4):M111 010199. https://doi.org/10.1074/mcp.M111.010199
15. Ahmad Izaham AR, Scott NE (2020) Open database searching enables the identification and comparison of bacterial glycoproteomes without defining glycan compositions prior to searching. Mol Cell Proteomics 19(9): 1561–1574. https://doi.org/10.1074/mcp.TIR120.002100
16. Kou Q, Xun L, Liu X (2016) TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. Bioinformatics 32(22):3495–3497. https://doi.org/10.1093/bioinformatics/btw398
17. LeDuc RD, Fellers RT, Early BP, Greer JB, Thomas PM, Kelleher NL (2014) The C-score: a Bayesian framework to sharply improve proteoform scoring in high-throughput top down proteomics. J Proteome Res 13(7):3231–3240. https://doi.org/10.1021/pr401277r
18. Zhou M, Pasa-Tolic L, Stenoien DL (2017) Profiling of histone post-translational modifications in mouse brain with high-resolution top-down mass spectrometry. J Proteome Res 16(2):599–608. https://doi.org/10.1021/acs.jproteome.6b00694
19. Zhou M, Malhan N, Ahkami AH, Engbrecht K, Myers G, Dahlberg J, Hollingsworth J, Sievert JA, Hutmacher R, Madera M, Lemaux PG, Hixson KK, Jansson C, Pasa-Tolic L (2020) Top-down mass spectrometry of histone modifications in sorghum reveals potential epigenetic markers for drought acclimation. Methods 184:29–39. https://doi.org/10.1016/j.ymeth.2019.10.007
20. 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 (2017) Informed-proteomics: open-source software package for top-down proteomics. Nat Methods 14(9):909–914. https://doi.org/10.1038/nmeth.4388
21. Cesnik AJ, Shortreed MR, Schaffer LV, Knoener RA, Frey BL, Scalf M, Solntsev SK, Dai Y, Gasch AP, Smith LM (2018) Proteoform suite: software for constructing, quantifying, and visualizing proteoform families. J Proteome Res 17(1):568–578. https://doi.org/10.1021/acs.jproteome.7b00685
22. Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, Inuganti A, Griss J, Mayer G, Eisenacher M, Perez E, Uszkoreit J, Pfeuffer J, Sachsenberg T, Yilmaz S, Tiwary S, Cox J, Audain E, Walzer M, Jarnuczak AF, Ternent T, Brazma A, Vizcaino JA (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res 47(D1):D442–D450. https://doi.org/10.1093/nar/gky1106
23. Nagarajan A, Zhou M, Nguyen AY, Liberton M, Kedia K, Shi T, Piehowski P, Shukla A, Fillmore TL, Nicora C, Smith RD, Koppenaal DW, Jacobs JM, Pakrasi HB (2019) Proteomic insights into phycobilisome degradation, a selective and tightly controlled process in the fast-growing cyanobacterium synechococcus elongatus UTEX 2973. Biomol Ther 9(8):374
24. Shechter D, Dormann HL, Allis CD, Hake SB (2007) Extraction, purification and analysis of histones. Nat Protoc 2(6):1445–1457. https://doi.org/10.1038/nprot.2007.202
25. Rodriguez-Collazo P, Leuba SH, Zlatanova J (2009) Robust methods for purification of histones from cultured mammalian cells with the preservation of their native modifications. Nucleic Acids Res 37(11):e81. https://doi.org/10.1093/nar/gkp273
26. Zhou M, Abdali SH, Dilworth D, Liu L, Cole B, Malhan N, Ahkami AH, Winkler TE, Hollingsworth J, Sievert J, Dahlberg J, Hutmacher R, Madera M, Owiti JA, Hixson KK, Lemaux PG, Jansson C, Paša-Tolić L Isolation of histone from sorghum leaf tissue for top down mass spectrometry profiling of potential epigenetic markers. JoVE:e61707. https://doi.org/10.3791/61707
27. Wu Z, Roberts DS, Melby JA, Wenger K, Wetzel M, Gu Y, Ramanathan SG, Bayne EF, Liu X, Sun R, Ong IM, McIlwain SJ, Ge Y (2020) MASH explorer: a universal software environment for top-down proteomics. J Proteome Res 19(9):3867–3876. https://doi.org/10.1021/acs.jproteome.0c00469
28. Zhou M, Wu S, Stenoien DL, Zhang Z, Connolly L, Freitag M, Paša-Tolić L (2017) Profiling changes in histone post-translational modifications by top-down mass spectrometry. In: Wajapeyee N, Gupta R (eds) Eukaryotic transcriptional and post-transcriptional gene expression regulation. Springer New York, New York, pp 153–168. https://doi.org/10.1007/978-1-4939-6518-2_12

Chapter 14

Determining Copper and Zinc Content in Superoxide Dismutase Using Electron Capture Dissociation Under Native Spray Conditions

Rachel Franklin, Michael Hare, and Joseph S. Beckman

Abstract

Localizing metal binding to specific sites in proteins remains a challenging analytical problem in vitro and in vivo. Although metal binding can be maintained by "native" electrospray ionization with intact proteins for quantitation by mass spectrometry, subsequent fragmentation of proteins with slow-heating methods like collision-induced dissociation (CID) can scramble and detach metals. In contrast, electron capture dissociation (ECD) fragmentation produces highly localized bond cleavage that is well known to preserve posttranslational modifications. We show how a newly available ECD tool that can be retrofitted on standard QTOF mass spectrometers allows the sites of copper and zinc binding to be localized in the antioxidant enzyme Cu, Zn superoxide dismutase (SOD1). The loss of zinc from Cu, Zn SOD1 has been shown to induce motor neuron death and could have a causal role in the fatal neurodegenerative disease, amyotrophic lateral sclerosis (ALS). The methods described enable copper loss to be distinguished from zinc using distinct ECD fragments of SOD1 and are broadly applicable to other metalloproteins.

Key words Native mass spectrometry, Metal coordination, Metalloprotein, Superoxide dismutase, Amyotrophic lateral sclerosis

1 Introduction

Many key enzymes and a third of all proteins bind metals that are essential for their function and structure [1]. However, site occupancy by metals in proteins remains a major analytical challenge to quantify. This is particularly problematic for proteins that bind more than one metal such as Cu, Zn superoxide dismutase (SOD1) (Fig. 1). Not only can SOD1 exist with only copper or zinc bound, but copper can also migrate to the zinc site and zinc will readily bind to the copper site [2]. In addition, SOD1 is a homodimer in its native state and the dimerization affects metal-binding substantially [3]. Mass spectrometry is the only technique

Liangliang Sun and Xiaowen Liu (eds.), *Proteoform Identification: Methods and Protocols*,
Methods in Molecular Biology, vol. 2500, https://doi.org/10.1007/978-1-0716-2325-1_14,
© The Author(s), under exclusive license to Springer Science+Business Media, LLC, part of Springer Nature 2022

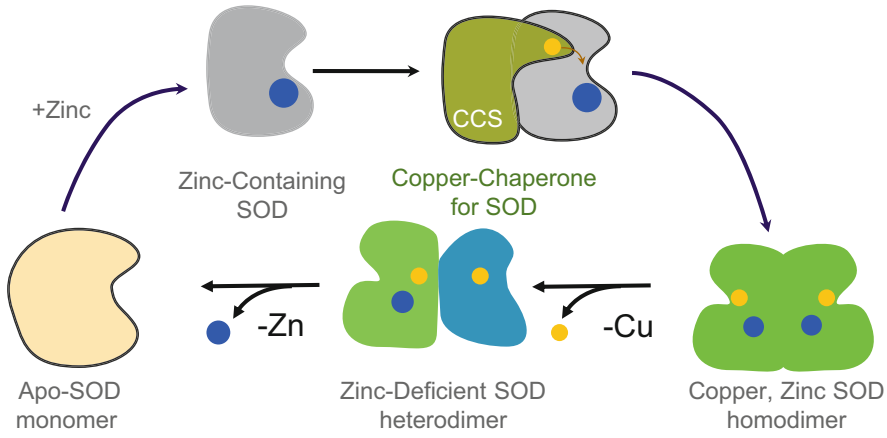


Fig. 1 Metals and maturation of SOD1. Zinc is bound first into newly synthesized SOD1 protein (apo), which mostly remains monomeric. The copper chaperone for SOD1 (CCS) binds to zinc-containing SOD1 and inserts copper. The binding of Cu and Zn to SOD1 promotes dimerization and forms the remarkably stable homodimer. Zinc is prone to fall out of monomeric Cu, Zn SOD to form zinc-deficient SOD1 that is highly toxic to motor neurons

that has the potential to fully characterize the heterogeneous combination of metal-bound states in SOD1.

Over 180 different missense mutations to SOD1 have been shown to cause amyotrophic lateral sclerosis (ALS), a fatal neurodegenerative disease in humans [4]. In our prior work to determine how mutations to SOD1 could cause ALS, we found that certain mutations to SOD1 reduce zinc affinity and activate oxidative chemistry in the remaining copper which can selectively target motor neurons [5, 6]. Our work has shown that determining metal binding to SOD1 is critical to understand how mutations can cause the progressive death of motor neurons in ALS [7, 8]. Measuring the intact protein spectra of metals bound to SOD1 proved invaluable for developing a promising copper-containing therapeutic called CuATSM to treat ALS. [9, 10] However, because copper and zinc are neighbors in the periodic chart, their isotopic mass distributions overlap substantially, making it difficult to distinguish which metal is bound in singly-metallated SOD1 using intact mass spectrometry. [11]

Our challenge has been to determine how copper and zinc binding is perturbed in ALS-associated SOD1, particularly in motor neurons in the spinal cords of transgenic animals. To adequately distinguish whether copper or zinc is missing in the SOD1 fraction containing only one metal, we turned to native protein mass spectrometry, which can quantify the amounts of copper and zinc bound to SOD1 directly in spinal cords [12]. Using a newly developed electron capture dissociation (ECD) tool, we identify metal binding in SOD1 using fragmentation spectra. A major advantage of ECD for the localization of posttranslational

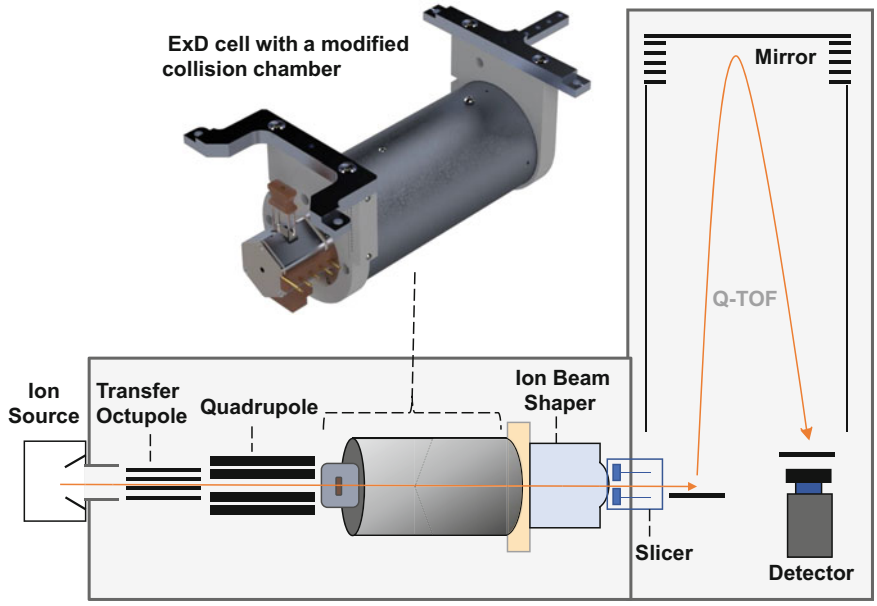


Fig. 2 Placement of the ExD cell in the Agilent 6545 XT Q-TOF mass spectrometer. The ExD cell is located after the quadrupole mass filter and before the collision cell. The collision cell was shortened by 18% to accommodate for ExD cell

modifications (PTMs) and metal binding over collision-induced dissociation (CID) is that specific bonds are broken at the electron capture, whereas CID heats the entire peptide or protein, breaking the weakest bonds first [13]. ECD has recently been used to characterize a variety of metal cofactors that bind to different sites in amyloid ß, a 43 amino acid found in Alzheimer's disease [14].

ECD has historically been limited to use in specialized FT-ICR mass spectrometers, which hindered its widespread adoption. Recently, a 30-mm-long electromagnetostatic cell (ExD cell) has been developed which enables ECD fragmentation in commercially available mass spectrometers (Fig. 2) [15]. The ExD cell uses carefully optimized electrostatic and magnetic fields to confine low-energy electrons. Charged analytes are guided through the cell, where electron capture causes dissociation between the N-C_α bond, resulting in c-type and z-type ions [15–18]. Small amounts of supplemental collision energy can be applied to increase the dissociation of protein fragments that can be held together by residual non-covalent interactions [19].

Here, we demonstrate the use of an ExD cell in a Q-TOF mass spectrometer to successfully identify copper and zinc binding in SOD1. The methodology described here for ECD fragmentation of SOD1 may be combined with ion mobility separations in the future to reveal new information about how copper and zinc binding affects protein stability and dimerization [20, 21]. This new method makes ECD practical in common mass spectrometers and offers better ways to analyze metal binding and other post-translational modifications in vitro and in vivo [16].

2 Materials

2.1 Reagents and Supplies

1. Lyophilized bovine Cu, Zn superoxide dismutase, Sigma Aldrich (CAS RN 9054-89-1, lot # 5LBC0305) stored at 4 °C.
2. Sub-micron filtered HPLC-grade water (Fisher Chemicals).
3. 15 mL conical tube and 1.5 mL microcentrifuge tube.
4. Pipettes and pipette tips.

2.2 Nanoelectrospray Ionization

1. Borosilicate glass capillaries (Sutter Instrument Co.) OD: 1 mm; ID: 0.78 mm: 10 cm in length.
2. Tip puller (Sutter Instrument Co., Model P-87).
3. A nanospray ionization source was assembled for these experiments using custom parts for compatibility with the Agilent HPLC-chip source recognition sensor. The source consists of a mount, a capillary holder, a thin stainless-steel wire connected to ground potential, and a camera system to assess emitter placement. A spray shield was used that replaced the capillary cap over the inlet into the mass spectrometer (*see* **Note 1**).
4. Gel-loading pipette tips.
5. Ceramic glass cutting tool.
6. Capillary holding forceps.

2.3 Mass Spectrometer and ExD Cell

1. Agilent 6545 Q-ToF mass spectrometer retrofit with an AQ-250 series ExD cell (e-MSion Inc.). The ExD cell is located between the quadrupole mass filter and a shortened collision cell.

3 Methods

3.1 Protein Preparation

1. Prior to analysis, thaw proteins at room temperature and then dissolve in HPLC-grade water to a final concentration of 1–10 μM (*see* Fig. 3).

3.2 Nanospray Ionization Conditions

1. Cut a clean glass capillary emitter to approximately 2 inches with a ceramic cutting tool (*see* **Note 1**).
2. Hold the capillary emitter with forceps and load 5–7 microliters of protein solution into the open end of the emitter using a long, gel-loading pipette tip (*see* **Note 2**).
3. Load the capillary into the capillary holder of the source. Feed the grounded wire into the open end of the capillary until the wire is submerged in the solution (*see* **Note 2**).
4. Position the capillary tip approximately 1–5 mm away from the source spray shield. Adjust the alignment of the emitter to ensure the tip is positioned directly in front of the inlet (*see* **Note 3**).

Fig. 3 Tune window for ExD cell controller. ECD profiles are optimized through systematic increases and decreases in lens voltages implemented through +/− controls on the left side of the window

5. After the emitter is positioned, set the initial source voltage below 1000 V and slowly increase the voltage until a stable spray is obtained. Bovine Cu, Zn superoxide dismutase was ionized with the following source parameters: gas temp. = 150 °C, gas flow = 5 L/min, capillary voltage = 1500 V and 1250 V for 10 μM and 1 μM, respectively (*see* **Notes 5** and **6**).

3.3 Mass Spectrometry

1. Before analysis, calibrate the Agilent 6546XT Q-TOF mass spectrometer for the appropriate masses using the Agilent LC/MS Calibration Standard (*see* **Note 7**).
2. After mass calibration, tune the ExD cell before introducing the SOD1 sample. Tune for ECD fragmentation using the substance P doubly charged precursor at m/z 674. The intensity of the prominent c5 fragment at m/z 624 can be optimized through the adjustment of ExD cell voltages (*see* Fig. 4). Start tuning the ExD cell by systematically increasing and decreasing the inner lenses (LM3-LM5) of the cell until the highest intensity of 624 m/z is reached. Then, increase/ decrease the filament bias (FB) independently to further improve ECD efficiency. Finally, optimize the voltages of L2 and L6 as a pair before L1 and L7 (*see* **Note 8**). This process may be repeated to fine-tune the optimal ExD profile voltages (*see* **Note 9**).
3. Set the ExD cell filament current to 2.38 A for all profiles. The value of 2.38 A is specifically used for looped filaments.
4. In the "Optics 1" Tune Tab, set the fragmentor to 300 V and the skimmer to 200 V. Set Lens 1 to 28 V, Lens 2 to 18.5 V, and Oct 1 DC to 29.5 V.
5. Set the Quad atomic mass unit (AMU) to 2000 and perform ECD/CID fragmentation in MS1 mode without isolation of a precursor. The nominal quad AMU setting of 2000 results in

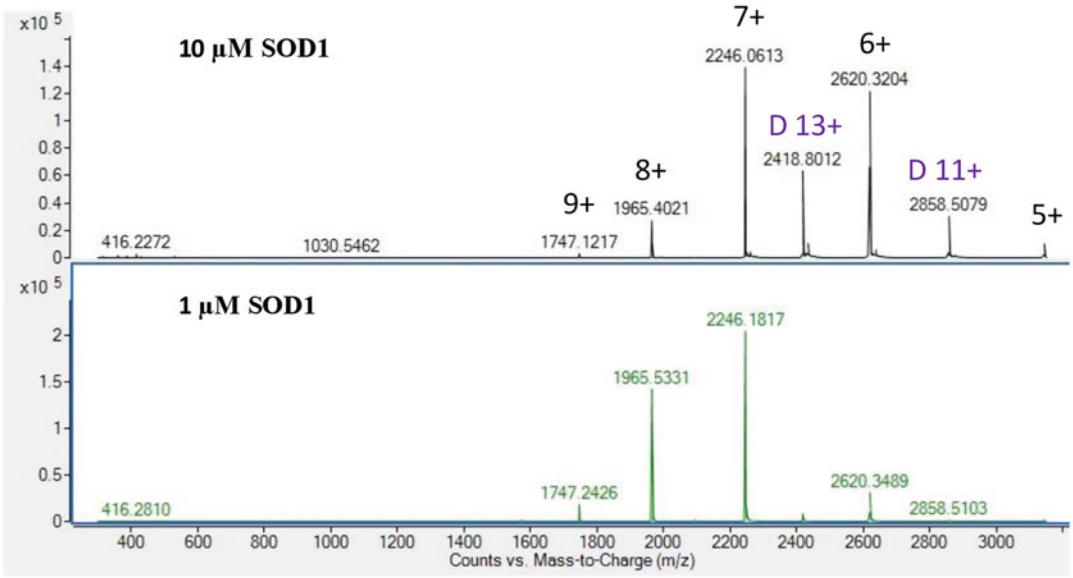


Fig. 4 10 µM and 1 µM holo-SOD1 MS1 spectra. Observed monomer charge states were +5 through +9 and the SOD homodimer was visible in its +11 and +13 state. The acquisition time for both spectra was 2 min

an effective low-mass cutoff of approximately 1600. Use alternative ExD voltage profiles for MS1 transmission and MS1 fragmentation experiments (*see* **Note 10**).

6. Add supplemental energy by increasing collision energy to improve ECD fragment detection (*see* Fig. 5) (*see* **Note 11**).
7. Increase the inner lenses and filament bias of the ExD cell (LM3, L4, FB, LM5) proportionally to match increases in collision energy. ExD profiles used for SOD1 analysis are listed in Table 1.
8. Collect spectra with Agilent MassHunter and use Agilent BioConfirm for mass deconvolution.

3.4 ECD Fragment Ion Analysis

1. ExD Viewer (e-MSion Inc.) is used for data analysis. One can also use LCMS Spectator available from the Pacific Northwest National Laboratories.
2. Gather the protein sequence from Uniprot and load into the "Targets Editor" in the ExD viewer interface.

1. Bovine SOD1 sequence: ATKAVCVLKGDGPVQGTIH FEAKGDTVVVTGSITGL TEGDHGFHVHQFGDNTQGCT SAGPHFNPLSKKHGGPKDEERHVGDLGNVTADKNG VAIVDIVDPLISLSGEYSIIGRTMVVHEKPDDLGRGG NEESTKTGNAGSRLACGVIGIAK.

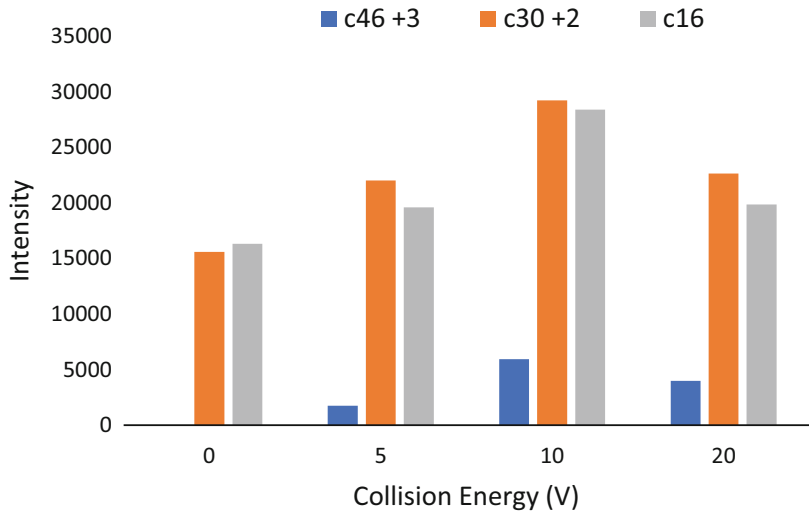


Fig. 5 A comparison of representative c-ions intensities with the addition of 0, 5, 10, and 20 volts of supplemental collision energy. These fragments were detected using a 1 μM SOD1 sample

Table 1
ExD voltage profiles for the ExD Controller used for SOD analysis

	L1	L2	LM3	L4	FB	LM5	L6	L7
Transmission	12	15	21	17	16.5	21	12	0
ECD (MS1)	−10	26.5	31.5	36.5	26.5	30.5	16	0

Voltage across LM3-LM5 was increased proportionally to increases in CE. All of the voltage settings were specifically for MS1 and would need to be separately optimized for MS2 operation

2. For bovine SOD1, include acetylation at the N-terminal, copper (II) coordination on H46, zinc (II) coordination on D81, and dehydrations on C55 and C144 for the intact disulfide bond. Include H60 and H118 dehydration modifications for Cu and Zn to account for the second metal coordination site [17].
3. Apply the Top Hat baseline filter to spectra before peak identification (*see* **Note 12**) [22].
4. Set peak finding parameters to require a minimum of 3 peaks, an error limit of 20 ppm, and a Pearson cutoff score of 0.85–0.90.
5. Use the theoretical fragment masses calculated from the reference sequence to manually confirm ExD Viewer identified peaks.
6. Use the described hybrid ECD/CID method to identify SOD1 fragments containing copper or zinc. Our experiments determined the site of copper binding in SOD1 down to a concentration of 1 μM (*see* Figs. 6 and 7).

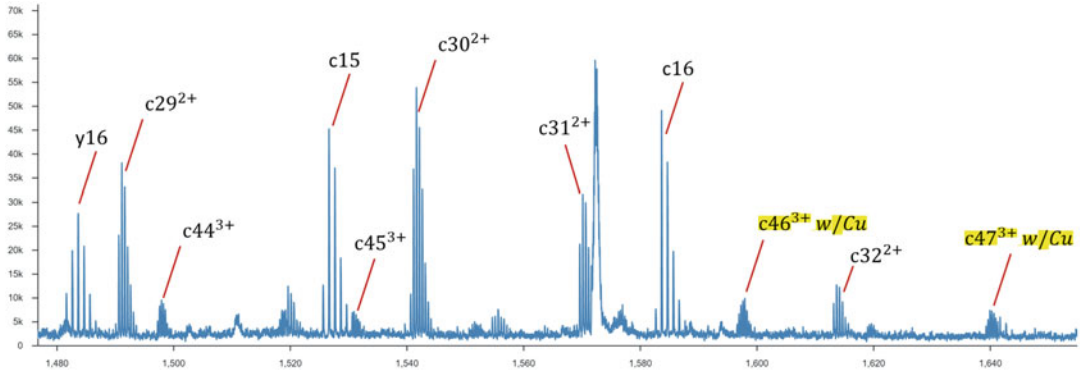


Fig. 6 A manually curated representative spectrum of 10 μM SOD1 added 10 V of collision energy. Ions indicative of copper on H46 are highlighted in yellow

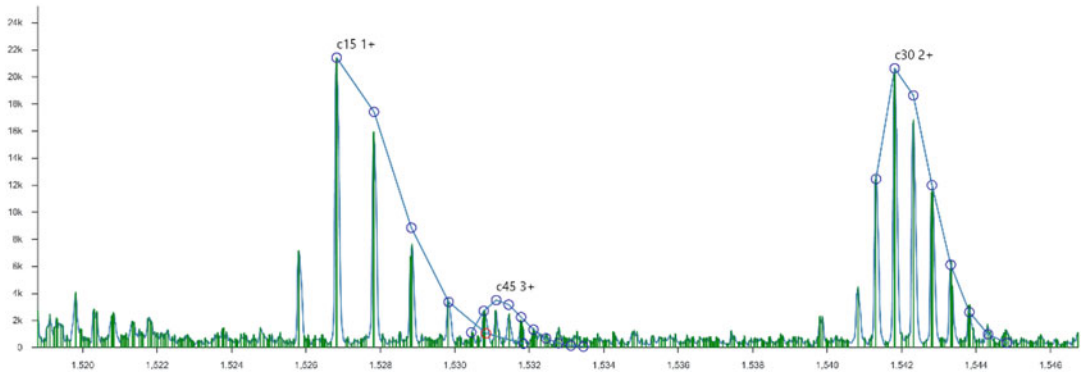


Fig. 7 Representative ExD Viewer assignments of ECD fragments

4 Notes

1. For our experiments, borosilicate glass capillary emitters with an original ID of 0.78 mm were pulled using a Sutter Model P-87 tip puller to a final ID of 1.5 μM.
2. Take care not to damage the tip of the capillary emitter while loading the sample or placing ground wire too far inside the capillary tube. A microscope is a useful tool to investigate tip integrity.
3. CAUTION: Ambient air ionization is an open design that exposes the high voltage spray shield of mass spectrometer inlet.
4. The sample should not run out quickly. A sample volume of 8–10 μL should be enough for hours of stable ionization. If the sample drains quickly, your needle may be damaged or the ionization voltage too high.

5. We typically do not exceed a capillary voltage of 1600 V to obtain suitable ionization of most proteins. Larger proteins may require higher voltages.
6. Using these source parameters, we typically observe the 5+ through 8+ states of holo-SOD1 with peaks corresponding to the SOD1 homodimer with +11 and +13 charges.
7. All spectra were collected in extended dynamic range (2 GHz) mode with a scan range of m/z 300–3200. The slicer was set to high resolution.
8. Tuning of the inner lenses and filament bias has the largest influence on ECD fragmentation.
9. An auto-tuning feature is being developed for this technology which will streamline the tuning process for future applications.
10. Ionization of SOD1 in water using nanoelectrospray resulted in a narrow charge state distribution which made ECD fragmentation without precursor isolation practical for analyzing SOD1.
11. Supplemental collision energy of 10 V seemed to give optimal ECD peak intensities in our experiments with bovine SOD1.
12. The ExD Viewer uses the PeakPickerHiRes centroiding algorithm.

References

1. Shi W, Chance M (2011) Metalloproteomics: forward and reverse approaches in metalloprotein structural and functional characterization. Curr Opin Chem Biol 15:144–148
2. Valentine JS, Pantoliano MW (1981) Protein-metal ion interactions in cuprozinc protein (superoxide dismutase). In: Spiro TG (ed) Copper proteins: metal ions in biology, vol 3. John Wiley and Sons, New York, pp 291–358
3. Kayatekin C, Zitzewitz JA, Matthews CR (2008) Zinc binding modulates the entire folding free energy surface of human Cu,Zn superoxide dismutase. J Mol Biol 384(2):540–555
4. Beckman JS, Estevez AG, Crow JP, Barbeito L (2001) Superoxide dismutase and the death of motoneurons in ALS. Trends Neurosci 24(11 Suppl):S15–S20
5. Crow JP, Sampson JB, Zhuang Y, Thompson JA, Beckman JS (1997) Decreased zinc affinity of amyotrophic lateral sclerosis-associated superoxide dismutase mutants leads to enhanced catalysis of tyrosine nitration by peroxynitrite. J Neurochem 69(5):1936–1944
6. Crow JP, Ye YZ, Strong M, Kirk M, Barnes S, Beckman JS (1997) Superoxide dismutase catalyzes nitration of tyrosines by peroxynitrite in the rod and head domains of neurofilament-L. J Neurochem 69(5):1945–1953
7. Estévez AG, Crow JP, Sampson JB et al (1999) Induction of nitric oxide-dependent apoptosis in motor neurons by zinc-deficient superoxide dismutase. Science 286:2498–2500
8. Sahawneh MA, Ricart KC, Roberts BR et al (2010) Cu,Zn superoxide dismutase (SOD) increases toxicity of mutant and Zn-deficient superoxide dismutase by enhancing protein stability. J Biol Chem 285(44):33885–33897
9. Williams JR, Trias E, Beilby PR, Lopez NI et al (2016) Copper delivery to the CNS by CuATSM effectively treats motor neuron disease in SODG93A mice co-expressing the Copper-Chaperone-for-SOD. Neurobiol Dis 89:0969–9961
10. Roberts BR, Lim NK, McAllum NK et al (2014) Oral treatment with CuII(atsm) increases mutant SOD1 in vivo but protects motor neurons and improves the phenotype

of a transgenic mouse model of amyotrophic lateral sclerosis. J Neurosci 34(23):8021–8031

11. Rhoads T, Williams J, Lopez N et al (2013) Using theoretical protein isotopic distributions to parse small-mass-difference post-translational modifications via mass spectrometry. J Am Soc Mass Spectrom 24:115–124
12. Rhoads TW, Lopez NI, Zollinger DR et al (2011) Measuring copper and zinc superoxide dismutase from spinal cord tissue using electrospray mass spectrometry. Anal Biochem 415(1):52–58
13. Kelleher N, Zubarev R, Bush K et al (1999) Localization of labile posttranslational modifications by electron capture dissociation: the case of γ-carboxyglutamic acid. Anal Chem 71:4250–4253
14. Lermyte F, Everett J, Lam Y et al (2019) Metal ion binding to the amyloid β monomer studied by native top-down FTICR mass spectrometry. J Am Soc Mass Spectrom 30:2123–2134
15. Fort K, Cramer C, Voinov V et al (2018) Exploring ECD on a benchtop Q exactive orbitrap mass spectrometer. J Proteome Res 17: 926–933
16. Beckman JS, Voinov VG, Hare M et al (2021) Improved protein and PTM characterization with a practical electron-based fragmentation on Q-TOF instruments. J Am Soc Mass Spectrom 32(8):2081–2091
17. Syrstad E, Turecček F (2005) Toward a general mechanism of electron capture dissociation. J Am Soc Mass Spectrom 16:208–224
18. Zubarev R, Kelleher N, McLafferty F (1998) Electron capture dissociation of multiply charged protein cations. A nonergodic process. J Am Chem Soc 120:3265–3266
19. Voinov V, Beckman J, Deinzer M, Barofsky D (2009) Electron-capture dissociation (ECD), collision-induced dissociation (CID) and ECD/CID in a linear radio-frequency-free magnetic cell. Rapid Commun Mass Spectrom 23:3028–3030
20. Williams J, Morrison L, Brown J et al (2020) Top-down characterization of denatured proteins and native protein complexes using electron capture dissociation implemented within a modified ion mobility-mass spectrometer. Anal Chem 92:3674–3681
21. Butler KE, Takinami Y, Rainczuk A, Baker ES, Roberts BR (2021) Utilizing ion mobility-mass spectrometry to investigate the unfolding pathway of Cu/Zn superoxide dismutase. Front Chem 9:614595
22. Stanford T, Bagley C, Solomon P (2016) Informed baseline subtraction of proteomic mass spectrometry data aided by a novel sliding window algorithm. Proteome Sci 7:14–19

Chapter 15

Surface-Induced Dissociation for Protein Complex Characterization

Sophie R. Harvey, Gili Ben-Nissan, Michal Sharon, and Vicki H. Wysocki

Abstract

Native mass spectrometry (nMS) enables intact non-covalent complexes to be studied in the gas phase. nMS can provide information on composition, stoichiometry, topology, and, when coupled with surface-induced dissociation (SID), subunit connectivity. Here we describe the characterization of protein complexes by nMS and SID. Substructural information obtained using this method is consistent with the solved complex structure, when a structure exists. This provides confidence that the method can also be used to obtain substructural information for unknowns, providing insight into subunit connectivity and arrangements. High-energy SID can also provide information on proteoforms present. Previously SID has been limited to a few in-house modified instruments and here we focus on SID implemented within an in-house-modified Q Exactive UHMR. However, SID is currently commercially available within the Waters Select Series Cyclic IMS instrument. Projects are underway that involve the NIH-funded native MS resource (nativems.osu.edu), instrument vendors, and third-party vendors, with the hope of bringing the technology to more platforms and labs in the near future. Currently, nMS resource staff can perform SID experiments for interested research groups.

Key words Native mass spectrometry, Proteoform identification, Protein complex, Surface-induced dissociation, High-resolution mass spectrometry, Protein mass spectrometry

1 Introduction

Native mass spectrometry (nMS) has emerged as a powerful tool in the structural biologist's toolbox, enabling intact non-covalent complexes to be transferred and studied in the gas phase [1–3]. nMS typically utilizes nano-electrospray ionization (nESI), non-denaturing solution conditions (most commonly ammonium acetate), and soft instrumental conditions, which have enabled protein–protein, protein–ligand, and protein–RNA/–DNA complexes to be studied [3–6]. nMS is advantageous in such studies as the sample requirements are low, typically microliters of sample at micromolar or lower concentrations. In addition, nMS can handle

Liangliang Sun and Xiaowen Liu (eds.), *Proteoform Identification: Methods and Protocols*,
Methods in Molecular Biology, vol. 2500, https://doi.org/10.1007/978-1-0716-2325-1_15,
© The Author(s), under exclusive license to Springer Science+Business Media, LLC, part of Springer Nature 2022

heterogeneous samples and therefore is advantageous in the study of samples with multiple proteoforms [7–9].

The first stage of an nMS experiment typically involves introducing the intact non-covalent complex into the gas phase and measuring its mass. However, in order to confirm composition and stoichiometry, determine which proteoforms are present, or obtain substructural information, it is often necessary to dissociate the complex. The most commonly used method of dissociation in nMS is collision-induced dissociation (CID), which is incorporated into almost all tandem mass spectrometers. In CID the analyte ion is accelerated into a neutral collision gas (typically nitrogen or argon), and the ion undergoes multiple low-energy collisions with the gas molecules resulting in a step-wise buildup of internal energy which results in dissociation. For protein complexes, CID typically causes unfolding/restructuring and results in a highly charged monomer being ejected from the complex [10–12]. Hence CID can provide information on the monomeric molecular weights, and often proteoforms, but provides limited information on connectivity. An alternative method of dissociation that has been increasingly applied to study protein complexes is surface-induced dissociation (SID) [13]. In SID, the analytes are accelerated toward, and collide against, a surface. This is a very rapid energy deposition process and dissociation can occur without unfolding. SID has been shown to be advantageous in protein structural studies as SID typically cleaves the weakest interfaces in protein complexes first, providing substructural information consistent with the solved structures, when available [14–16]. We demonstrate here the use of nMS and SID for protein complex, and proteoform characterization. The discussion in this chapter focuses on the use of the Thermo Scientific Q Exactive UHMR for these studies; while different instrument platforms can be used for SID studies, the high resolution, and sensitivity, of this instrument is beneficial in proteoform characterization studies [17].

2 Equipment and Materials

2.1 Self-Assembled Monolayer (SAM) Surface Preparation

1. Suitable gold-coated surface—we use 17-mm × 13-mm × 0.5-mm gold surface slide [the glass surface is coated with a 50 Å layer of Ti and then a 1000 Å layer of Au] (custom ordered from EMF corp).
2. 1 mM perflourothiol (FC_{12}) in ethanol, synthesis of which has been previously described [14].
3. 200 proof ethanol.
4. 20 mL scintillation vial.

5. Soft-tipped tweezers or stainless-steel tweezers with the ends wrapped in Teflon tape.
6. UV cleaner.
7. Sonicator.

2.2 Offline Sample Preparation

1. Ammonium acetate (≥99.99%).
2. Ultrapure water.
3. Either of the following:
   (a) Spin column (e.g., Biorad micro bio-spin p6).
   (b) Dialysis cassette (e.g., Pierce 96-well microdialysis plates).
   (c) Centrifugal concentrator (e.g., Amicon Ultra 0.5 centrifugal filters).
4. Centrifuge, capable of going to 14,000 × *g* if using a centrifugal concentrator.
5. Shaker plate for dialysis.

2.3 Pulling and Loading nESI Tips

1. Glass capillaries (we use glass capillaries with a 1.0 mm OD and 0.78 mm ID with a glass filament).
2. Tip puller (we use a P97 puller from Sutter instrument Co. with a 3.0 mm box filament installed).
3. Platinum wire 0.25 mm diameter.
4. Gel loader tips or glass syringe.

2.4 Calibration of Q Exactive UHMR

1. Cesium iodide.
2. Isopropanol.
3. Ultrapure water.

2.5 SID on a Q Exactive UHMR [18, 19]

1. SID device [18, 19].
2. Flange for back of HCD cell with Fisher connector feedthrough.
3. Switch to convert between using onboard power supplies and external power supply to apply the C-trap offset.
4. Two 10-channel power supplies and appropriate cables to connect one to Fisher connector for SID device and one to control the C-trap offset.
5. Access to control C-trap offset entrance and exit inject voltages.

3 Methods

3.1 How Is SID Implemented in Different Mass Spectrometers?

SID was first developed by Cooks and coworkers and since has been implemented by several groups coupled with multiple different mass analyzers [20–27] initially for small molecule and peptide fragmentation. In recent years, SID has been incorporated by Wysocki and coworkers into several different types of mass spectrometer and applied to the study of protein complexes. These include quadrupole time-of-flight (QTOF) [28, 29], ion mobility Q-TOFs (Q-IM-TOF) [30], Fourier transform ion cyclotron resonance (FTICR) [31–33], and Orbitrap mass spectrometers [18, 34]. In all cases, the instruments were modified in-house to incorporate an SID device, the design of which depends on the instrument in question. In its simplest form, the SID device consists of a surface, optics to steer the ion beam toward the surface, and optics to collect the fragments after surface collision. In all cases, the SID device can be operated either in "transmission mode" where the ion beam passes through the device or "SID" mode where the ion beam is directed toward and collides against the surface.

A schematic representation of an SID device designed by Zhou et al. [30]. and installed into a Waters Synapt G2 instrument is shown in Fig. 1. In this instrument, the SID device can be placed before or after the mobility cell by either truncating the trap or transferring CID cell to make space for the SID device [30, 35]. The truncated CID cells can still be used to perform CID and hence this functionality is not lost through the incorporation of SID. Incorporation of SID before IM allows SID products to be separated by mobility, which is advantageous as species may overlap in m/z (e.g., a singly charged monomer and doubly charged dimer would have the same m/z but would be separated in the IM dimension). However, incorporation after IM allows mobility-separated species to be individually interrogated. Furthermore, SID can be incorporated in both locations in SID-IM-SID studies enabling multistage dissociation experiments to be performed and giving increased information on connectivity [36]. Similar SID devices have been installed in a Bruker 15T FTICR [31, 32], Thermo Exactive plus EMR [18], and Q Exactive plus UHMR [34]. On the Bruker 15T FTICR, the standard collision cell was replaced with a custom designed cell which incorporates both an SID device and a CID cell [31, 32]. In these designs and the device incorporated into the Q Exactive UHMR, discussed below, a fluorocarbon self-assembled monolayer (SAM) was used as the surface, and directions for preparing the surface are given in the following section.

More recently, a very simple split lens SID device has been reported by Snyder et al. [33], reducing the number of electrodes

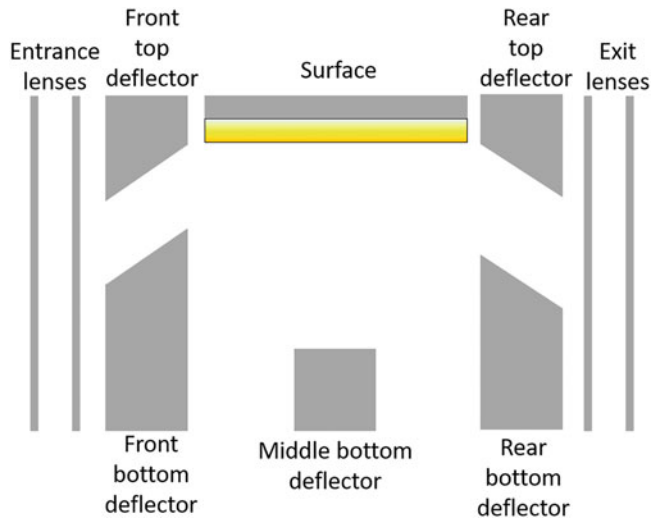


Fig. 1 Schematic representation of an SID device incorporated into a Waters Synapt G2 [30]

from 10–12 to three, and the device size to several millimeters. This split lens device has been incorporated in-house into multiple mass spectrometers, replacing the entrance lens on the Bruker FT-ICR collision cell, the dynamic range enhancement lens on a Waters Synapt, or the exit lens of a transfer multipole in a Thermo Scientific Extended Mass Range Orbitrap. In this design, the surface is a stainless-steel electrode, as opposed to a SAM, and therefore users do not need to prepare a surface in the traditional way. This device is now commercially available in the Waters Select Series Cyclic IMS, enabling researchers to perform SID without having to modify their instrument in-house.

3.2 Incorporation of SID into a Thermo Scientific Q Exactive UHMR

In the Q Exactive UHMR, a 4 cm SID device is fabricated and placed between the quadrupole and the C-trap, replacing the transport multipole, as shown in Fig. 2. A similar design was previously incorporated in the same location into an Exactive EMR [18], which was modified to include a quadrupole. To supply the voltages to the SID device, a modified flange is installed behind the HCD cell to provide the necessary feedthrough connections. Voltages are supplied with an external power supply (Ardara Technologies, Ardara, PA). When operating in transmission mode (no surface collision), voltages are applied to the entrance and exit lenses of the device to help with ion focusing and transmission. When operating in SID mode, the ion beam is directed toward the surface for collision, and the products are then collected and guided toward the C-trap with the rear electrodes. The SID voltage is defined as the voltage difference between the last optic before the SID device or the SID entrance lens and the surface. In order to increase the

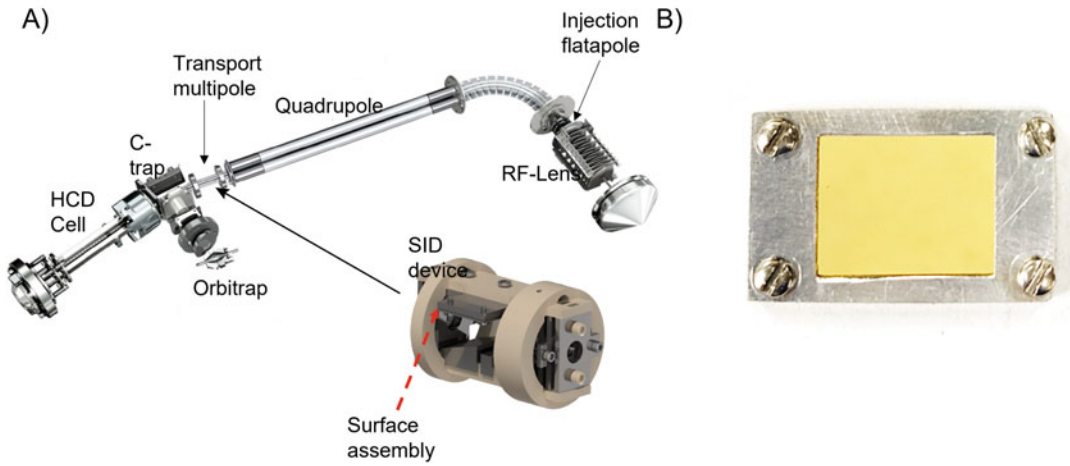


Fig. 2 (**a**) Schematic representation of a modified Q Exactive UHMR with SID installed in place of the transport multiple. (Reproduced from Ref [34] with permission from the Royal Society of Chemistry) and (**b**) photograph of surface assembly

SID voltage, the post-surface optics in the SID device are made more negative along with the C-trap DC offset during injection. Higher SID energies are achieved by modifying the ion optics board, to enable the C-trap DC offset during the injection step to be applied by an external power supply. This is achieved through the use of a circuit with a switch that allows a user to switch between supplying the C-trap offset from the onboard power-supply (typical for MS experiments when no SID is required) to the external power supply (when SID is desired) [18]. More details on tuning the SID device for transmission and SID modes can be found in the following sections.

3.3 Preparing a Self-Assembled Monolayer Surface

The majority of protein complex SID studies have been performed using a perfluorothiol (FC_{12}) SAM assembled on a 17-mm × 13-mm × 0.5-mm gold surface slide (1000 Å of Au on 50 Å of Ti on glass). While preparing a surface care must be taken to avoid scratching or damaging the gold surface, this is best done by handling the surface only by the edges using soft-tipped tweezers or tweezers wrapped in Teflon tape.

1. Perfluorothiol can be synthesized following a previously established protocol [14].
2. Prepare a 1 mM solution of perfluorothiol in ethanol. The FC_{12} stock solution should be stored at 4 °C wrapped in aluminum foil to prevent exposure to light; for long-term storage, it is recommended to store under nitrogen to reduce oxidation of the thiol.
3. Clean the surface with ethanol and then dry with nitrogen.

4. The surface should then be UV cleaned (gold side up) for 15 min.
5. Place the surface into a scintillation vial, gold surface side up, and cover with ~2–3 mL of 1 mM FC_{12}. Incubate overnight at room temperature in the dark, best achieved by wrapping the vial in aluminum foil.
6. Remove the FC_{12} solution and sonicate in ethanol for 1 min, repeat the wash step 6 times. If using immediately dry with nitrogen after the final wash step, alternatively the surface can be stored for future use in ethanol.
7. Once dry, the surface can be installed in the surface holder (shown in Fig. 2b).

The FC_{12} SAM surface should be replaced every time the instrument is vented. Venting can allow residual oil from turbomolecular pumps to deposit on the surface, changing the surface properties; hence, a fresh surface should be installed after every instrument vent.

3.4 Sample Preparation for Native MS

For native mass spectrometry, samples should be prepared in an MS-compatible electrolyte solution, typically ammonium acetate (*see* **Note 1**), with or without triethyl ammonium acetate for charge reduction, or ethylenediamine diacetate [37, 38]. Prior to analysis, samples can be buffer exchanged into the MS-compatible electrolyte solutions and non-volatile components are removed (e.g., salts, additives, electrolytes); however, low concentrations of some components can be retained in solution if required for activity or stability [39, 40]. Buffer exchange can be performed either offline or online to the MS. Offline buffer exchange typically uses either buffer exchange spin column, dialysis, or diafiltration. Alternatively, during the last stages of protein purification, ammonium acetate (AmAc) can be used as the electrolyte solution, avoiding the need for buffer exchange later on.

3.4.1 Buffer Exchange Spin Columns

The quickest offline buffer exchange method is to use a buffer exchange spin column. We typically use the Biorad Micro P6 spin columns. Buffer exchange can be achieved following the manufacturer's protocol (*see* **Notes 2–4**).

3.4.2 Concentration and Diafiltration

An alternative method of buffer exchange is diafiltration or ultrafiltration that utilizes a molecular weight cutoff (MWCO) filter. This method can be used to concentrate your sample at the same time if required. We typically use Amicon-ultra 0.5 mL concentrators which have a range of MWCOs (*see* **Note 5**). The appropriate MWCO is chosen based on the protein of interest. The MWCO is dependent on the hydrodynamic radii rather than the true molecular weight and, therefore, care should be taken with proteins which

have a small hydrodynamic radii. The MWCO chosen should typically be at most half of the molecular weight of the smallest protein of interest. To perform diafiltration, follow the recommended manufacturer's protocol (*see* **Note 6**).

3.4.3 Dialysis

Dialysis is a more time-consuming method of buffer exchange but is more efficient at removing high concentrations of salts and non-volatiles. There are multiple different dialysis cassettes that can be used, the choice of which will depend both on the MWCO required and on the sample volumes. We are typically working with sample volumes of 10–100 μL and hence use either the Thermo Slide-A-Lyzer MINI (*see* **Note 7**) or the Pierce 96-well microdialysis cassettes (*see* **Note 8**). For the Pierce 96-well microdialysis cassettes, a single cassette can be used and placed in a microcentrifuge tube instead of the 96-well format if desired. In both cases, the manufacturer's protocol should be followed.

3.4.4 Online Buffer Exchange

An alternative to the offline buffer exchange methods summarized here is to perform the exchange online to the MS. Online buffer exchange (OBE) utilizes liquid chromatography, with a non-denaturing mobile phase such as ammonium acetate, directly coupled to the MS [41]. In this case, samples in MS incompatible buffers are injected onto a short-size exclusion column which separates the protein from the small salt molecules, which can be diverted away from the MS to waste. The proteins are eluted in the non-denaturing mobile phase and directly analyzed with MS. A detailed protocol describing how to perform online buffer exchange has been previously reported [41]. More recently, OBE has been coupled with affinity purification online to an MS allowing the rapid screening of his-tagged protein overexpression [42]. In this approach, cells can be lysed, clarified, and then injected onto the LC first being loaded onto a Ni-NTA column to which the his-tagged proteins will bind. These proteins are then eluted with imidazole and loaded onto the second column, a short SEC column for OBE, before being introduced into the MS. This rapid screening approach is particularly useful when screening overexpression conditions. Finally, when looking at abundant proteins produced from overexpression samples, the samples can be lysed with ammonium acetate and then directly analyzed, removing the requirement to exchange into AmAc [43–45].

3.5 Pulling Nanospray Needles

The majority of nMS experiments are performed via static nanospray ionization (nESI), using glass or quartz capillaries that have been pulled into a small tip, typically of ~1 μm inner diameter, although for some applications much smaller tip sizes are used [46, 47]. In order to apply the spray, potential users can either use capillaries which are coated with a conductive material or simply insert a platinum wire into the solution, ensuring the platinum wire

Table 1
Typical puller program values for a 3.0 mm box filament and 1.0 mm outer diameter and 0.78 mm inner diameter borosilicate capillaries containing a glass filament

Heat	Pull	Velocity	Time
Ramp test value+5	15–40	15–40	200

is in contact with the sample holder to which the potential is applied (see details below). Coated capillaries are commercially available from multiple vendors including New Objective and Thermo Scientific. With a suitable tip-puller, however, capillaries can be made in-house, which can be more cost-efficient. A detailed protocol has been previously reported for how to pull, sputter coat in gold, and then clip very small capillaries suitable for use with backing pressure (applying low gas flows or pressure to the back of the capillary to help start or maintain spray) [48]. Here we present a protocol for pulling open tips which do not need to be clipped and are used without sputter coating. We use a P97 Sutter Instrument Co. puller with a 3.0 × 3.0 mm box filament puller. While the exact parameters will need to be reoptimized when changing filaments, and between tip pullers, Table 1 below gives some standard settings that can be used as a starting point. It is also important to note that settings will vary for different types of capillaries—we use filaments with a 1.0 mm outer diameter and 0.78 mm inner diameter and a glass filament (Sutter Instrument Co.) (*see* **Note 9**).

1. Loosen the knobs on both puller bars. Place a glass capillary in the groove on one of the puller bars and push glass slowly through the clamp until 1–2 cm of the capillary protrudes through the clamp. Carefully tighten the clamp so that the capillary is secure, but be sure to not over-tighten and risk damaging the glass.
2. Push the puller release on both puller bars and carefully move the puller bars toward the filament. While holding the two puller bars in place with one hand, use the other hand to slightly loosen the clamp on the glass. Carefully slide the capillary through the filament, until it is roughly centered and then tighten both clamps. Ensure that the capillary stays within the groove on the puller bars during this process to avoid damage to the filament.
3. Press "pull" on your desired program; once the glass has been pulled into two tips, remove them from the tip puller (*see* **Note 10**). After pulling, needles can be stored in either a petri dish, with a double-sided adhesive pad attached to the bottom or a pipette tip box, with the adhesive pad attached to the tip rack. The adhesive pad will hold the needles in place, above the

bottom of the container, which will stop them from getting damaged.

4. Optional: for our setup, the needles pulled in the previous steps are too long, so we cut 2–3 cm of the open end of the capillary using a glass cutter before loading.

3.6 Getting Spray and Tuning the MS

3.6.1 Mass Calibration

The Q Exactive UHMR is calibrated using a 2 mg/mL solution of cesium iodide (CsI) prepared in 50:50 water:isopropanol. To calibrate, the heated electrospray ionization (HESI) source should be installed, along with the sweep cone. Fresh CsI solution should be prepared immediately before calibration, and the signal needs to be stable before attempting calibration. To mass calibrate, follow the protocol outlined in the Q Exactive UHMR quickstart guide. The frequency of calibration will depend on the experimental needs. It should be noted that CsI can contaminate the front end of the instrument; typically we use a dedicated ion transfer tube for calibration and switch to a different one for running samples.

3.6.2 Acquiring MS Spectra for Protein Complexes

1. For an SID in-house modified instrument, set the circuit so that the C-trap offset is being applied by the instrument power supply, turn on the external power supplies, and set the voltage to appropriate transmission mode settings. Table 2 below gives typical voltages applied to the front-end instrument optics along with the SID device for transmission mode.
2. Set your trap gas to the desired value; typically we use between 2 and 9 Arb., with higher pressure, and hence higher trap gas being required for larger protein complexes.
3. Set an appropriate mass range for the sample being analyzed. The expected average charge state (Z_{av}) can be estimated from Eq. 1 below [49]. We recommend initially setting the mass range so you are scanning from 500 *m/z* to ~2 times above the expected *m/z* of your species. For example, if the protein complex is expected to be ~6000 *m/z*, acquire from 500 to 12,000 *m/z* initially. This is to ensure that you may observe any smaller or larger species than may be present in your sample. The mass range can then be narrowed as required.

$$Z_{\mathrm{AV}} = 0.0778\sqrt{(M)} \tag{1}$$

4. Remove the peek screw at the back of the sample holder and thread a platinum wire through the sample holder. Make sure that there is contact between the sample holder and the wire by carefully bending the wire to form a loop and covering the loop in aluminum foil to provide extra stability (Fig. 3). Replace the peek screw on the back of the sample holder. Once this setup is established, the wire can be left in place and does not need to be removed between samples. To avoid cross contamination the

Table 2
Typical voltages applied during transmission mode and SID mode on a Q Exactive UHMR

	Transmission mode (voltage unless otherwise stated)	SID 45 V (voltage unless otherwise stated)
Source temperature	200–250 °C	200–250 °C
Injection flatapole	8	8
Inter flatapole	5	5
Bent flatapole	2	2
C-trap entrance lens inject	1.8	−29
C-trap exit lens inject	16	−20
Entrance 1–3	−5	−5
Entrance 2	5	6
Front top	0	−23.5
Front bottom	0	−9.5
Surface	0	−40
Middle bottom	0	−110
Back top	0	−154
Back bottom	0	−90
Exit 1–3	−20	−30
Exit 2	2	−50
C-trap offset	–	−35

Note that these voltages should be tuned and optimized for each instrument and after venting and surface replacement

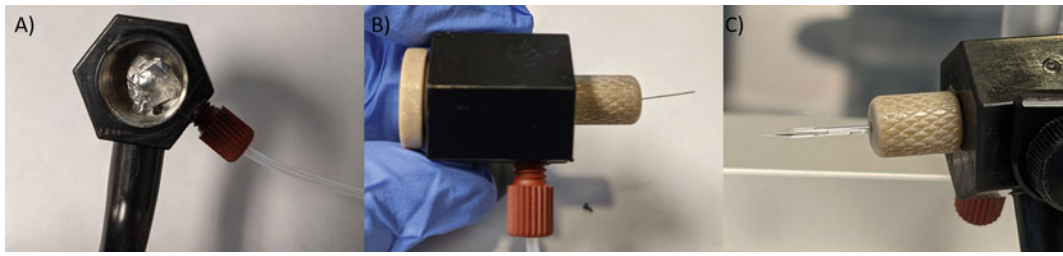


Fig. 3 Photographs of the static nESI sample holder for a nano-flexspray source with platinum wire inserted. (**a**) Back view with peek screw removed. The piece of aluminum foil, inserted in the metal housing, can be observed. (**b**) After assembly of the sample holder, the platinum wire must be in contact with the aluminum inside the metal housing and protrude from the front of the peek nut. The red peek screw is connected to a 3 mL syringe for backing pressure. (**c**) Sample holder with filled nESI needle, ready for analysis. The platinum wire is inserted inside the needle and is in direct contact with the sample

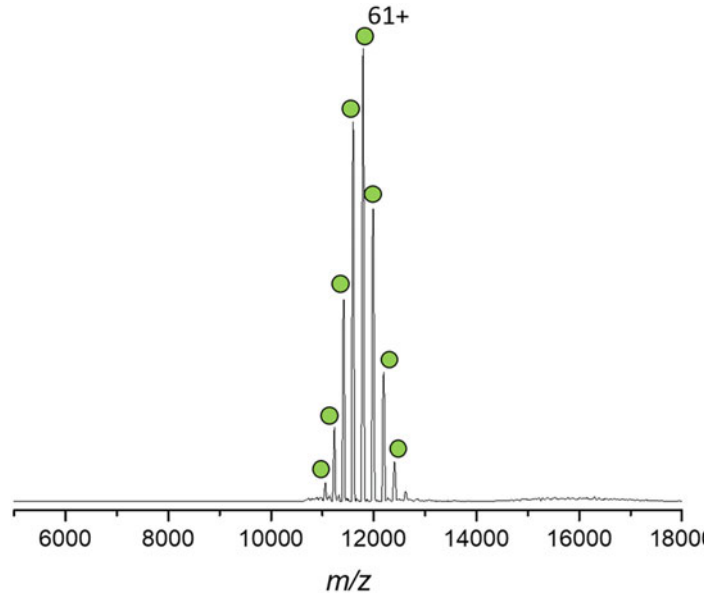


Fig. 4 Mass spectrum of 1 μM rat 20S proteasome prepared in 150 mM AmAc and acquired with −50 V in-source trapping, trap gas 9, and a resolution of 3125 K. Measured mass is 719, 173 ± 160 Da

wire can be carefully rinsed with water or methanol between samples.

5. Load 2–3 μL of your sample (~1 μM concentration or lower) into a pulled needle. For this step you can either use a glass syringe (e.g., a Hamilton 10 μL 1700 series with an N termination) or gel loader tips (e.g., 0.5–10 μL Micro gel tips from Genesee Scientific). To remove air bubbles and ensure that the solution is at the bottom of the tip, glass capillaries can be carefully spun down using a mini centrifuge. If using a glass syringe to fill nESI needles, the syringe should be flushed with a 50:50 water:methanol solution several times between samples.
6. Carefully push the back of the needle onto the platinum wire so that the wire is in contact with solution and push the needle through the peek nut (note: if backing pressure is required, a soft ferrule should be installed between the needle holder and the peek nut to allow for proper sealing; a 3 mL syringe can be connected to the sample holder to provide the necessary backing pressure), *see* Fig. 3. Tighten the peek nut so the sample is held securely.
7. Push source bottom into the instrument so the source is sealed, and the interlock is fulfilled. Using the adjustment knobs on the XYZ manipulator of the source, adjust the tip position so that it is centered with the sample inlet and approximately 5–10 mm away from the inlet.

8. Optional: If using backing pressure, slowly push on the syringe until a droplet appears at the end of the nESI needle.
9. Turn the instrument into operate and apply a capillary voltage to begin the spray. We find that a higher spray voltage (~0.9–1.1 kV) is often needed to start the spray after which the capillary voltage can be dropped to ~0.6–0.9 kV for analysis. Once you have sprayed, the tip position should be optimized (optimizing both for intensity and spectral quality); typically we find moving further back from the inlet is preferred as well as moving the tip either up or down, so it is not directly level with the inlet.
10. Tune the instrument to optimize signal while avoiding activation (*see* **Note 11**) and dissociation. Detailed information on the modifications that have been made to the Exactive platform to optimize it for high-mass transmission can be found elsewhere [50, 51]. We typically begin in the front end of the instrument and work our way toward the analyzer, *see* Fig. 2a. Normally, we work with fixed injection times (of 100–200 ms), which are optimized based on signal to avoid overfilling and hence space-charge effects. When tuning for the first time, we recommend acquiring the data while you tune, so you can easily note any changes (e.g., activation) as you tune.
    (a) Source temperature: increasing the source temperature can aid in desolvation; however, high-source temperatures could result in activation. We typically work at 200–250 °C.
    (b) In-source CID: applying low in-source CID voltages can aid desolvation but can also cause activation.
    (c) In-source trapping: For improved desolvation of protein complexes, we typically use in-source trapping as opposed to in-source CID, as this gives better kinetic control of the ions. Higher in-source trapping voltages will strip off more adducts and, however, can result in dissociation and/or activation and resturcturing (*see* **Note 11**) [17]. Therefore, we recommend keeping this as low as possible (5–50 V works well for most of our soluble protein complexes prepared in AmAc without additives).
    (d) Flatapoles: The voltages on the injection, inter, and bent flatapoles should be optimized for transmission, but care should be taken to minimize the voltages in this region to avoid activation. For SID, the SID voltage is defined as the difference between the bent flatapole and SID surface which is typically ~5 V; hence, for SID, the SID voltage is the surface voltage plus 5 V. Significantly higher voltages here will therefore impact the definition of the SID

voltage (*see* Subheading 3.7 *Tuning SID for a range of energies*).

(e) Ion transfer target *m/z:* The ion transfer target controls the RF amplitudes on the injection flatapole, bent flatapole, transport multipole, and HCD cell and can be set to high or low or manually set. For the majority of protein complexes, we work in high *m/z* mode; however, if smaller peptides or ligands are also present in the sample, the RF amplitudes may need to be manually tuned to allow both to be observed.

(f) Detector *m/z* optimization: This setting changes the voltages and timing by which ions are pulsed into the Orbitrap. In our experience, "low" tends to work well for most protein complexes until you reach very large ions (>15,000 *m/z*).

(g) Extended trapping: Extended trapping allows ions to be transferred into the HCD cell before being transferred into the C-trap and is useful for collisionally cooling large ions [51].

(h) HCD voltage: Applying low HCD voltages can increase spectral quality by removing any residual adducts. Typically, we prefer to use in-source trapping or in-source CID so the sample is cleaned up before mass selection in the quadrupole when selecting a single species.

(i) Trap gas: The trap gas, which controls the pressure of the HCD cell, should be optimized for each system. Generally increasing the trap gas helps with trapping and collision cooling of larger systems; however, when doing HCD to generate monomers, this setting may need to be decreased, as discussed below.

(j) Resolution: The resolution setting is the resolution at 400 *m/z*, and increasing the resolution increases the transient time allowing higher resolution data to be acquired. However, signal decay can be observed using long transients due to collisions of the protein ions with background gas molecules [52] and hence spectral quality may decrease with increased resolution. For intact protein complexes, we typically use resolution settings of ≤25K.

11. Once tuned, long acquisitions are not necessary, with sufficient signal intensity (normalized level of ≥ mid E4). Typical acquisitions for full MS are 50 scans. The length of time will depend on the injection time and resolution setting used. A typical MS spectrum for the 20S proteasome, biochemically purified from rat liver [7], a ~717 kDa hetero 28-mer acquired at 3125 K resolution, with trap gas 9, and −50 V in source trapping is shown in Fig. 4.

3.6.3 CID for Characterization of Protein Complex

Almost all instruments used for native MS are equipped to perform CID. CID involves accelerating the ion of interest into a neutral collision gas, the ions undergo multiple collisions with the collision gas and dissociation typically proceeds via an unfolding/rearrangement mechanism producing unfolded and highly charged monomer and the complementary N-1mer, sometimes with sequential monomer losses. CID is therefore a useful tool for confirming the stoichiometry of the complex, determining the molecular weight of the monomers, and identifying proteoforms present [53].

1. Optional: it is not necessary to mass select a single charge state to perform CID. Depending on the signal intensity, we either select a single charge state of species using a narrow window (width of 2–20 m/z units) or select the entire charge state distribution (CSD) for dissociation. Selecting the entire CSD is beneficial when signal is too low to select a single peak but multiple species are present.
2. Gradually increase the HCD CE voltage in steps of ~20 V until dissociation is observed. For confirming stoichiometry, we aim to achieve balance between monomers and N-1mers. In order to achieve balance between the products, the trap gas should be optimized (Fig. 5a), along with the HCD field gradient purge (the electrical field by which ions are ejected from the HCD cell). Resolution can also bias mass distributions with lower resolutions typically providing better balance. For proteoform characterization, we aim to produce as abundant monomers as possible, using higher HCD voltages and tuning with lower trap gases to better transmit these species (Fig. 5b).

3.7 Tuning SID Over a Range of Energies

In order to perform SID, the ion beam has to be guided toward, and collide against, the surface following which the ions need to be collected and transmitted to the rest of the instrument. In the modified Q Exactive UHMR and EMR (*see* **Note 12**), this is achieved by tuning the front deflector such that the ions are attracted toward the front top deflector and hence toward the surface and the back deflector more attractive to pull ions off the surface and towards the rest of the instrument (*see* Table 2). The SID voltage is defined as the difference between the bent flatapole (typically held ~5 V) and the surface, and hence SID voltage is generally defined as the surface voltage plus 5 V (*see* Fig. 6) [18]. The SID energy however is defined as the SID voltage multiplied by the charge state (e.g., for an 11+ precursor SID 45 V would equal an SID energy of 495 eV). If multiple charge states are selected for dissociation, a single SID energy cannot be defined; however, using the average charge state, the average energy can be determined (average charge state multiplied by the voltage).

When initially tuning in SID, it is recommended to use a protein standard, of known structure, which has previously been

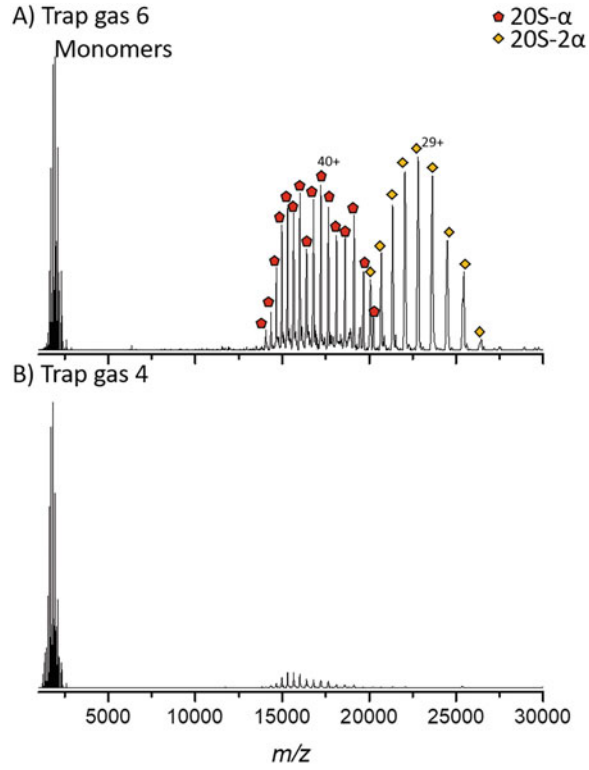


Fig. 5 250 V HCD of 20S proteasome from rat prepared in 150 mM AmAc, acquired with −50 V in-source trapping at 3125 K resolution. (**a**) Acquired with trap gas 6 and (**b**) acquired with trap gas 4. Identity of the monomers observed in B will be discussed below in the Subheading 3.8

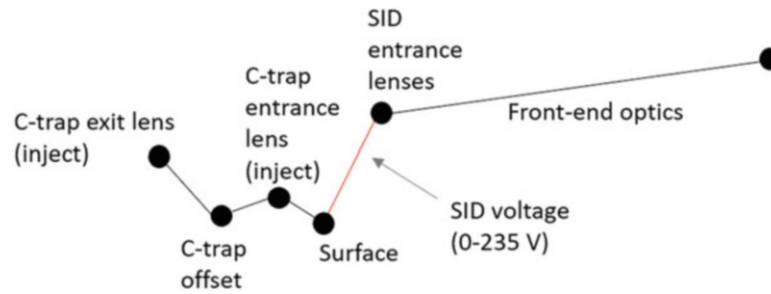


Fig. 6 Relative voltage diagram showing how the SID voltage is defined (the voltage difference between the last optic before the SID device or the SID entrance lens and the surface). Increasing SID voltages are achieved by lowering the C-trap offset (during the injection) and lowering the SID devices from the surface onwards. (Adapted with permission from VanAernum et al. [18]. Copyright (2019) American Chemical Society)

studied with SID and hence has well known dissociation products, e.g., streptavidin (dimer of dimers produces dimers with SID) or C-reactive protein (cyclic pentamer produces monomer to tetramer

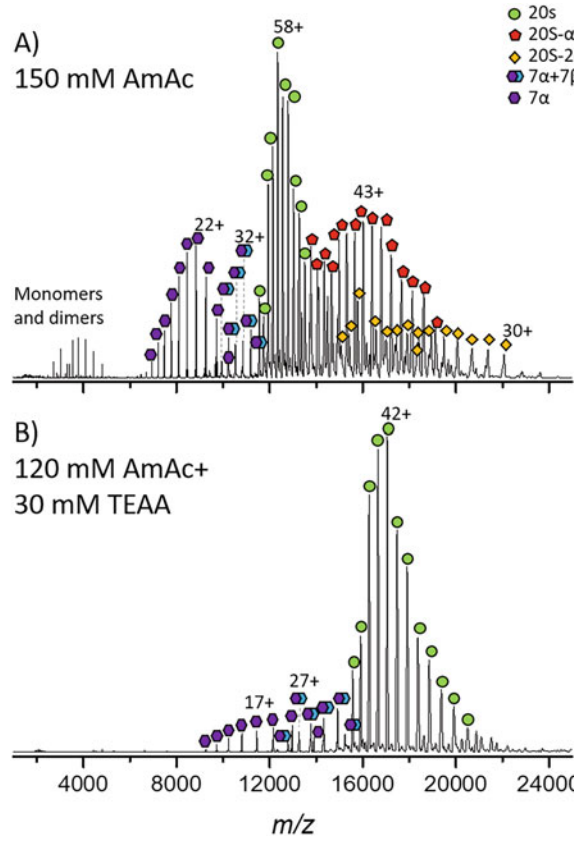


Fig. 7 Lower energy SID of rat 20S proteasome acquired with −50 V in-source trapping at 3125 K resolution and a trap gas of 9. Prepared in either (**a**) 150 mM AmAc (average charge state of precursor +61, determined from UniDec [59]) and acquired with 85 V SID (corresponding to an average energy of 5185 eV) or (**b**) 120 mM AmAc +30 mM TEAA (average charge state of precursor +45, determined from UniDec) and acquired with 105 V SID (corresponding to an average energy of 4725 eV)

with SID) [14, 16, 54]. For structural information, it is recommended to prepare the sample under charge-reducing conditions (typically containing 20% by volume TEAA [11, 55]). Low-voltage (or energy) SID is particularly useful for studying the structure of protein complexes as they typically cleave at the weakest interfaces first, particularly when samples are prepared under charge-reducing conditions [14, 54]. The rat 20S proteasome was subjected to low-voltage SID under both normal charge and charge-reduced conditions, as shown in Fig. 7a, b. When prepared under charge-reducing conditions, the dominant product is the 7α ring (Fig. 7b), an observation that is consistent with previous SID studies and the cryo-EM structure [7]. When the sample is prepared without a charge reducing agent, the 7α ring is observed; however, competing pathways are also present such as the production of α

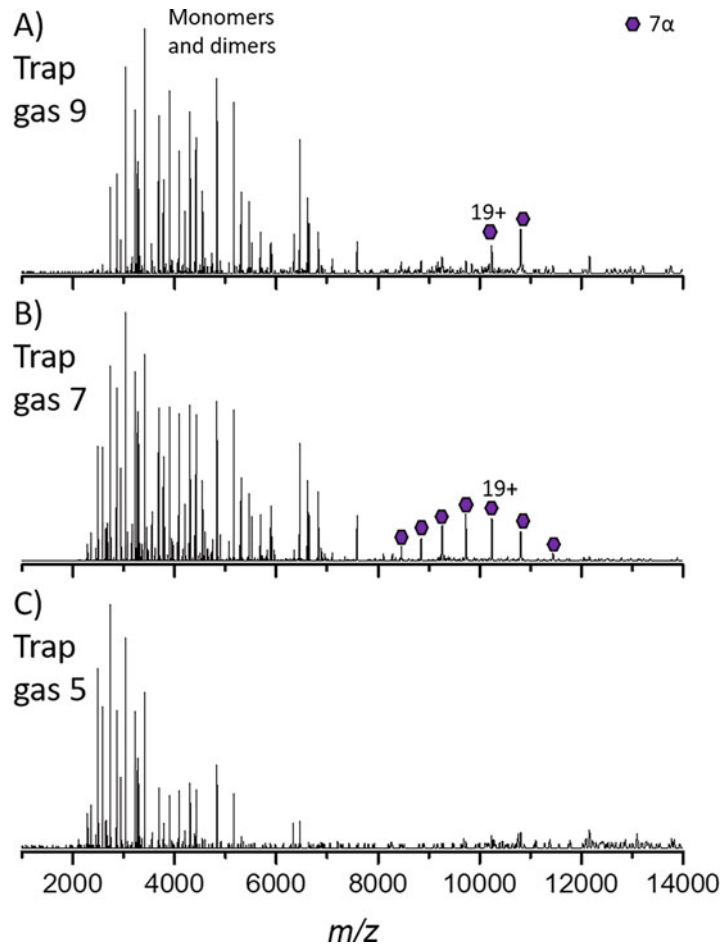


Fig. 8 Higher energy SID of rat 20S proteasome acquired with −50 V in-source trapping at 3125 K resolution prepared in 150 mM AmAc (average charge state of precursor +61, determined from UniDec [59]) and acquired with 185 V SID (corresponding to an average energy of 11,346 eV) with (**a**) trap gas 9, (**b**) trap gas 7, and (**c**) trap gas 5. With trap gas 7, greater coverage of the monomers is obtained, as demonstrated with the increased number of peaks in the low mass region. The identity of the monomers is discussed below in the interpreting data Subheading 3.8

monomers and the complementary 20S-α (Fig. 7a), highlighting the importance of charge reduction in structural studies by SID.

When increasing the SID voltage on the Orbitrap platforms, as the front-end optics cannot be increased to very high voltages, the back end of the SID device and the C-trap entrance are therefore made more negative. Making the voltage on the surface more negative will increase the voltage difference between the bent flatapole and the surface (therefore increasing SID voltage). To collect and transmit the ions, the back deflector, exits, C-trap entrance and exit (inject) voltages, and the C-trap offset are also made more

negative to aid transmission [18], as discussed in more detail in the following protocol. High-energy SID is particularly useful when proteoform information is required, as it can be used to produce higher intensity of monomer (compared to low-energy SID). Given the higher charge states observed in AmAc (and therefore higher SID energies accessible), we have found that the most extensive proteoform characterization by SID is obtained using samples prepared in AmAc without the addition of a charge reducing reagent [56]. During SID, ions are not typically transmitted into the HCD cell, so for proteoform characterization, the trap gas does not need to be dropped as low as with HCD; however, the trap gas can still affect the SID spectra observed and therefore should be optimized for proteoform characterization. Figure 8 below shows a comparison for high-energy SID of rat 20S proteasome acquired with trap gas 9, 7, and 5.

The protocol for a typical SID experiment, after achieving spray and optimizing the instrument for MS of the complex of interest (*see* Subheading 3.6.2), is outlined below.

1. Switch the instrument into standby and adjust the circuit so that the C-trap offset is being supplied by the external power supply as opposed to the onboard power supply [18].
2. Restart spray and check for MS signal (*see* Subheading 3.6.2, **Step 8**).
3. Optional: if desired, select a single *m/z* species with a narrow isolation window (typically 5–20 *m/z*). If signal is low, however, the entire charge state distribution can be selected instead by applying a wide isolation around the region of interest. If signal is low, it is often beneficial to increase the injection time to increase the number of ions.
4. Set the SID device voltages, and C-trap offset and C-trap entrance and exit (inject) voltages as shown in Table 2. SID can then be optimized by fine-tuning the device as needed. To fine–tune, we recommend starting with the front deflectors and then working step-wise toward the back of the device. This process may need to be repeated more than once. Care should be taken at this stage to ensure unintentional CID is not occurring along with SID. CID can be distinguished from SID as it typically produces highly charged monomer and complementary N-1mer, whereas SID produces a greater variety of subcomplexes, with more symmetrical charge partitioning [57]. It is recommended to tune the SID device initially with a complex of known structure that has been previously studied with SID.
5. Optimize injection time and trap gas for signal. Care should be taken when selecting the resolution at which to acquire SID data—higher resolutions will typically bias toward monomers

and lower resolutions will typically provide greater balance between product pairs. We typically acquire multiple SID voltages and multiple resolutions to check product distribution before determining final conditions.

6. To increase SID voltage by 10 V, decrease the C-trap entrance inject, C-trap exit inject, and C-trap offset by 10 V. Increase front top and front bottom deflectors on the C-trap by 0.5 V and decrease the surface, middle bottom, back top deflector, back bottom deflector, and exit lenses by 10 V. Typically, we acquire SID at 10 V intervals from 45 to 225 V SID.

3.8 Interpreting the Data

There are multiple software packages that can be used to deconvolute native MS data, aiding in interpretation and reporting. Within our laboratories, the most commonly used packages are UniDec, Intact Mass, and Biopharma Finder. All three allow for mass matching of deconvoluted species which can assist in interpretation. These are briefly introduced below, and protocols for nMS data analysis using these three software packages can be found elsewhere [41, 58].

UniDec (Michael Marty, University of Arizona) UniDec is a free, open source, software that employs a Bayesian framework [59]. UniDec is well suited to nMS data analysis including ion mobility data and is well suited to analyze data with overlapping species. MetaUniDec also allows for high-throughput batch processing [60], assisting in the analysis of large data sets. UniDec can directly read data from Waters and Thermo Instruments, as well as opening data from other instruments by converting to a .txt or mzML format.

Intact Mass (Protein Metrics) Intact Mass is based on a parsimonious algorithm [61] and is well suited to the analysis of protein and protein complex spectra. Intact Mass can be used to interpret data from multiple different instrument vendors and can batch process.

BioPharma Finder (Thermo Scientific) Biopharma finder is a software package that can be used for deconvolution of data acquired on Thermo Instruments. In addition to being able to handle intact mass data, it can also be used for top-down data analysis for protein identification.

Manually interpreting SID spectra for complex species, including those with multiple subunits or proteoforms, can be a very time-consuming process, and therefore using one of the software packages listed above can greatly assist in this process. An example of this is identifying the monomers present in CID or higher-energy SID spectra of the rat 20S proteasome (Figs. 5b and 8b

respectively). In order to identify the different monomers, dimers, and trimers present in the low *m/z* region of these spectra, we employed the use of UniDec. Data was imported into UniDec and analyzed with the "high-resolution native" preset method, using the automatic *m/z* peak width option, sampling mass every 1 Da, peak detection range of 1 Da, and peak detection threshold of 0.01. Peak peaking was then manually validated. Major proteoforms of the α-subunits were identified with both CID and SID (Fig. 9 and Table 3). However, under these conditions, a greater number of β-subunits were identified by SID. In addition, SID can provide information on the dimers in the 7α ring (Table 4) and therefore increased substructural information than CID here [7].

The protocol presented here demonstrates that when coupled with a high-resolution analyzer, SID is a useful tool to study both substructure (using low-energy SID) and proteoforms (using high-energy SID). Although demonstrated here for a single system, this approach is applicable to systems of different sizes and can be used for soluble and membrane proteins.

4 Notes

1. Due to the hygroscopic nature of ammonium acetate, we typically purchase smaller quantities at a time, and once opened, prepare a 5M stock from the entire bottle. This ensures an accurate concentration can be achieved. The stock can be stored at −20 °C until needed. Prior to sample preparation, the stock solution can be diluted to the working concentration; for most of our systems, this is 100–200 mM AmAc, but concentrations of 10–1000 mM can be used if required.
2. For buffer exchange, we follow the protocol as listed; however, we perform four washes, which will result in >99.9% buffer exchange.
3. If using a fixed rotor centrifuge, care should be taken to rotate the spin column between centrifuge steps to keep the column bed as even as possible.
4. Ensure that the column bed looks dry before loading the sample or you may experience sample dilution.
5. For Amicon-ultra 0.5 mL concentrators, trace amounts of glycerine can remain on the filter; therefore, for nMS applications, it is recommended to follow the suggested wash step prior to sample loading. Do not let the filter dry out between the wash step and loading your sample.
6. For the most common biochemistry buffers, we find 2–3 buffer exchange (dilution) steps are sufficient to remove the salt and non-volatile components; however, up to 10 exchanges may be

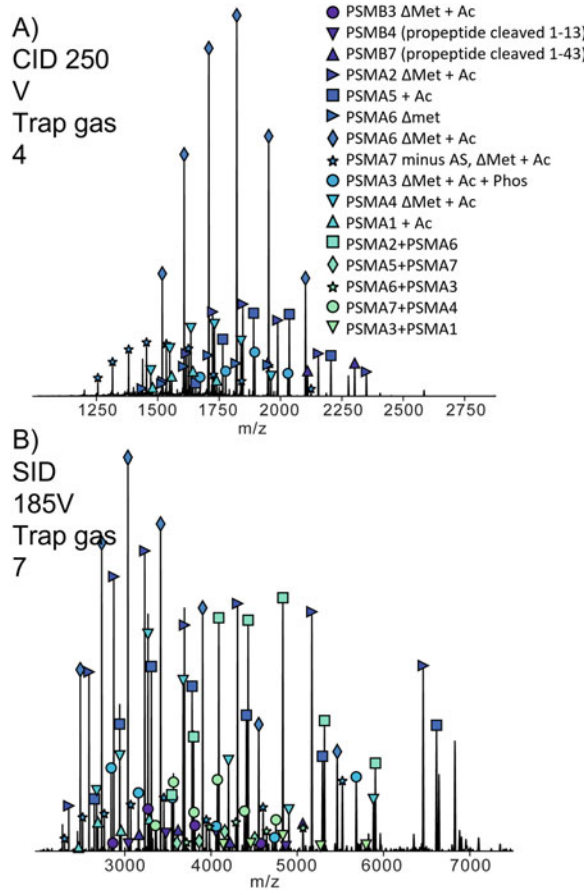


Fig. 9 Monomers and dimers identified with (**a**) 250 V CID trap gas 4 and (**b**) 185 V SID with trap gas 7. Data analyzed using UniDec. In both cases, data was acquired with – 50 V in source trapping and at 3125 K resolution

required for some harder-to-remove components or when samples are prepared with high concentrations of non-volatiles.

7. For nMS applications, we recommend following the optional wash step for Slide-A-Lyzer. However, do not let the membrane dry out between the wash and sample loading steps. We also typically change the dialysis buffer at least once (after 1–2 h) before an overnight dialysis.
8. The microdialysis cassettes can be prewashed by pipetting 100 μL of water into the cassette and then carefully removing it before loading your sample. Do not let the membrane dry out between the wash and sample loading steps. A typical dialysis for nMS using microdialysis will require changing the dialysis buffer 2–3 times after 1–2 h for each step. If necessary, an overnight dialysis can be performed after the initial exchanges; however, an overnight dialysis is typically only

Table 3
Theoretical and observed molecular weights of α-monomers and β-monomers of rat 20S proteasome

	Theoretical mass (Da)	CID (250 V trap gas 4)	SID (185 V trap gas 7)
PSMA1 + Ac	29, 560	29, 560 ± 1	29, 558 ± 1
PSMA2 ΔMet + Ac	25, 838	25, 837 ± 1	25, 836 ± 1
PSMA3 ΔMet + Ac + Phos	28, 410	28, 410 ± 1	28, 409 ± 3
PSMA4 ΔMet + Ac	29, 409	29, 409 ± 1	29, 407 ± 4
PSMA5 + Ac	26, 453	26, 453 ± 1	26, 452 ± 1
PSMA6 ΔMet	27, 268	27, 264 ± 2	–
PSMA6 ΔMet + Ac	27, 310	27, 311 ± 1	27, 309 ± 1
PSMA7 minus AS, ΔMet + Ac	27, 608	27, 609 ± 1	27, 609 ± 2
PSMB1 (propeptide cleaved 1–27)	23, 547	–	23, 541 ± 9
PSMB3 ΔMet + Ac	22, 876	–	22, 875 ± 1
PSMB4 (propeptide cleaved 1–13)	24, 336	–	24, 337 ± 6
PSMB7 (propeptide cleaved 1–43)	25, 314	25, 313 ± 1	25, 316 ± 4

Theoretical molecular weights were determined based on the previously observed most abundant proteoform [7]. Observed molecular weights and errors were determined manually

Table 4
Theoretical and observed molecular weights of α-dimers of rat 20S proteasome

	Theoretical mass (Da)	SID (185 V trap gas 7)
PSMA5 + PSMA7	54, 061	54, 061 ± 5
PSMA7 + PSMA4	57, 017	57, 013 ± 4
PSMA2 + PSMA6	53, 148	53, 142 ± 4
PSMA6 + PSMA3	55, 720	55, 719 ± 1
PSMA3 + PSMA1	57, 970	57, 969 ± 4

Theoretical molecular weights were determined based on the previously observed most abundant proteoform [7]. Observed molecular weights and errors were determined manually

needed for removal of very high concentrations of non-volatiles.

9. The heat value needed to melt the glass will vary for different types of glass and between filaments. To determine the heat value required for the glass/filament combination on a Sutter P97 puller, you run a "ramp test," which systematically raises the temperature until the glass starts to soften and puller bars

move apart. The heat is turned off when a factory-set ramp test velocity is achieved. The heat value required to reach this velocity is the ramp test temperature. The ramp test temperature will vary for different types of glass and between filaments. It is therefore important to rerun the ramp test when changing the filament or changing glass types.

10. When pulling tips that are open and do not require clipping, we have found that more consistent pulling is achieved when the filament is allowed to cool completely between uses. It can take up to 5 min for the filament to completely cool after pulling a capillary.
11. Care should be taken with all instrument parameters to avoid unintentional activation. It is possible to activate and unfold the complex before dissociation, so even if the complex is intact, it may not still be native [62]. Previous reports have shown that it is possible to rearrange the structure with high in-source trapping voltages [56]. It is likely that high voltages, voltages drops, and high capillary temperature could also have an effect and therefore care should always be taken to keep instrument conditions as "soft" and in low voltage as possible.
12. Instrument modifications should be made only if the customer and instrument vendor agree on steps to be followed.

Acknowledgments

We gratefully acknowledge the current and former Wysocki group members who have contributed to the development and understanding of SID. In particular, we would like to acknowledge Zachary VanAernum and Joshua Gilbert for their hard work incorporating SID into the Exactive platforms and optimizing tuning and Benjamin Jones for helpful discussions during the preparation of this protocol. We would like to acknowledge NSF Grants DBI1455654 and DBI0923551 for SID instrument development. We also acknowledge NIH Grant P41GM128577 for the development of an integrated MS-based structural biology workflow and dissemination of SID.

References

1. Ben-Nissan G, Sharon M (2018) The application of ion-mobility mass spectrometry for structure/function investigation of protein complexes. Curr Opin Chem Biol 42:25–33. https://doi.org/10.1016/j.cbpa.2017.10.026
2. Heck AJR, van den Heuvel RHH (2004) Investigation of intact protein complexes by mass spectrometry. Mass Spectrom Rev 23:368–389. https://doi.org/10.1002/mas.10081
3. Liko I, Allison TM, Hopper JT, Robinson CV (2016) Mass spectrometry guided structural biology. Curr Opin Struct Biol 40:136–144. https://doi.org/10.1016/j.sbi.2016.09.008

4. Eschweiler JD, Kerr R, Rabuck-Gibbons J, Ruotolo BT (2017) Sizing up protein–ligand complexes: the rise of structural mass spectrometry approaches in the pharmaceutical sciences. Annu Rev Anal Chem 10:25–44. https://doi.org/10.1146/annurev-anchem-061516-045414
5. Sarni S, Biswas B, Liu S et al (2020) HIV-1 Gag protein with or without p6 specifically dimerizes on the viral RNA packaging signal. J Biol Chem 295:14391–14401. https://doi.org/10.1074/jbc.RA120.014835
6. Laganowsky A, Reading E, Allison TM et al (2014) Membrane proteins bind lipids selectively to modulate their structure and function. Nature 510:172–175. https://doi.org/10.1038/nature13419
7. Vimer S, Ben-Nissan G, Morgenstern D et al (2020) Comparative structural analysis of 20S proteasome ortholog protein complexes by native mass spectrometry. ACS Cent Sci 6:573–588. https://doi.org/10.1021/acscentsci.0c00080
8. Melani RD, Skinner OS, Fornelli L et al (2016) Mapping proteoforms and protein complexes from king cobra venom using both denaturing and native top-down proteomics. Mol Cell Proteomics 15:2423–2434. https://doi.org/10.1074/mcp.M115.056523
9. Franc V, Zhu J, Heck AJR (2018) Comprehensive proteoform characterization of plasma complement component C8αβγ by hybrid mass spectrometry approaches. J Am Soc Mass Spectrom 29:1099–1110. https://doi.org/10.1021/jasms.8b05825
10. Schwartz BL, Bruce JE, Anderson GA et al (1995) Dissociation of tetrameric ions of noncovalent streptavidin complexes formed by electrospray ionization. J Am Soc Mass Spectrom 6:459–465. https://doi.org/10.1016/1044-0305(95)00191-F
11. Hall Z, Politis A, Bush MF et al (2012) Charge-state dependent compaction and dissociation of protein complexes: insights from ion mobility and molecular dynamics. J Am Chem Soc 134:3429–3438. https://doi.org/10.1021/ja2096859
12. Popa V, Trecroce DA, McAllister RG, Konermann L (2016) Collision-induced dissociation of electrosprayed protein complexes: an all-atom molecular dynamics model with mobile protons. J Phys Chem B 120:5114–5124. https://doi.org/10.1021/acs.jpcb.6b03035
13. Stiving AQ, VanAernum ZL, Busch F et al (2019) Surface-induced dissociation: an effective method for characterization of protein quaternary structure. Anal Chem 91:190–209. https://doi.org/10.1021/acs.analchem.8b05071
14. Harvey SR, Seffernick JT, Quintyn RS et al (2019) Relative interfacial cleavage energetics of protein complexes revealed by surface collisions. Proc Natl Acad Sci 116:8143–8148. https://doi.org/10.1073/pnas.1817632116
15. Zhou M, Wysocki VH (2014) Surface induced dissociation: dissecting noncovalent protein complexes in the gas phase. Acc Chem Res 47:1010–1018. https://doi.org/10.1021/ar400223t
16. Quintyn RS, Yan J, Wysocki VH (2015) Surface-induced dissociation of homotetramers with D2 symmetry yields their assembly pathways and characterizes the effect of ligand binding. Chem Biol 22:583–592. https://doi.org/10.1016/j.chembiol.2015.03.019
17. Harvey SR, VanAernum ZL, Wysocki VH (2021) Surface-induced dissociation of anionic vs cationic native-like protein complexes. J Am Chem Soc 143:7698–7706. https://doi.org/10.1021/jacs.1c00855
18. VanAernum ZL, Gilbert JD, Belov ME et al (2019) Surface-induced dissociation of noncovalent protein complexes in an extended mass range orbitrap mass spectrometer. Anal Chem 91:3611–3618. https://doi.org/10.1021/acs.analchem.8b05605
19. VanAernum ZL (2020) Novel native mass spectrometry-based fragmentation and separation approaches for the interrogation of protein complexes. Electronic Thesis or Dissertation, The Ohio State University
20. Cooks RG, Terwilliger DT, Ast T et al (1975) Surface modified mass spectrometry. J Am Chem Soc 97:1583–1585. https://doi.org/10.1021/ja00839a056
21. Beck RD, John PS, Homer ML, Whetten RL (1991) Impact-induced cleaving and melting of alkali-halide nanocrystals. Science 253:879–883. https://doi.org/10.1126/science.253.5022.879
22. Laskin J, Futrell JH (2003) Surface-induced dissociation of peptide ions: kinetics and dynamics. J Am Soc Mass Spectrom 14:1340–1347. https://doi.org/10.1016/j.jasms.2003.08.004
23. Stone E, Gillig KJ, Ruotolo B et al (2001) Surface-induced dissociation on a MALDI-ion mobility-orthogonal time-of-flight mass spectrometer: sequencing peptides from an "in-solution" protein digest. Anal Chem 73:2233–2238. https://doi.org/10.1021/ac001430a
24. Schultz DG, Hanley L (1998) Shattering of SiMe3+ during surface-induced dissociation. J

Chem Phys 109:10976–10983. https://doi.org/10.1063/1.477737

25. Volný M, Elam WT, Ratner BD, Tureček F (2005) Preparative soft and reactive landing of gas-phase ions on plasma-treated metal surfaces. Anal Chem 77:4846–4853. https://doi.org/10.1021/ac0505019
26. Mohammed S, Chalmers MJ, Gielbert J et al (2001) A novel tandem quadrupole mass spectrometer allowing gaseous collisional activation and surface induced dissociation. J Mass Spectrom 36:1260–1268. https://doi.org/10.1002/jms.217
27. Castoro JA, Nuwaysir LM, Ijames CF, Wilkins CL (1992) Comparative study of photodissociation and surface-induced dissociation by laser desorption Fourier transform mass spectrometry. Anal Chem 64:2238–2243. https://doi.org/10.1021/ac00043a010
28. Galhena AS, Dagan S, Jones CM et al (2008) Surface-induced dissociation of peptides and protein complexes in a quadrupole/time-of-flight mass spectrometer. Anal Chem 80:1425–1436. https://doi.org/10.1021/ac701782q
29. Jones CM, Beardsley RL, Galhena AS et al (2006) Symmetrical gas-phase dissociation of noncovalent protein complexes via surface collisions. J Am Chem Soc 128:15044–15045. https://doi.org/10.1021/ja064586m
30. Zhou M, Huang C, Wysocki VH (2012) Surface-induced dissociation of ion mobility-separated noncovalent complexes in a quadrupole/time-of-flight mass spectrometer. Anal Chem 84:6016–6023. https://doi.org/10.1021/ac300810u
31. Yan J, Zhou M, Gilbert JD et al (2017) Surface-induced dissociation of protein complexes in a hybrid Fourier transform ion cyclotron resonance mass spectrometer. Anal Chem 89:895–901. https://doi.org/10.1021/acs.analchem.6b03986
32. Snyder DT, Panczyk E, Stiving AQ et al (2019) Design and performance of a second-generation surface-induced dissociation cell for Fourier transform ion cyclotron resonance mass spectrometry of native protein complexes. Anal Chem 91:14049–14057. https://doi.org/10.1021/acs.analchem.9b03746
33. Snyder DT, Panczyk EM, Somogyi A et al (2020) Simple and minimally invasive SID devices for native mass spectrometry. Anal Chem 92:11195–11203. https://doi.org/10.1021/acs.analchem.0c01657
34. Harvey SR, VanAernum ZL, Kostelic MM et al (2020) Probing the structure of nanodiscs using surface-induced dissociation mass spectrometry. Chem Commun. https://doi.org/10.1039/D0CC05531J
35. Zhou M, Dagan S, Wysocki VH (2012) Protein subunits released by surface collisions of noncovalent complexes: nativelike compact structures revealed by ion mobility mass spectrometry. Angew Chem Int Ed 51:4336–4339. https://doi.org/10.1002/anie.201108700
36. Quintyn RS, Harvey SR, Wysocki VH (2015) Illustration of SID-IM-SID (surface-induced dissociation-ion mobility-SID) mass spectrometry: homo and hetero model protein complexes. Analyst 140:7012–7019. https://doi.org/10.1039/c5an01095k
37. Hernández H, Robinson CV (2007) Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. Nat Protoc 2:715–726. https://doi.org/10.1038/nprot.2007.73
38. Dyachenko A, Gruber R, Shimon L et al (2013) Allosteric mechanisms can be distinguished using structural mass spectrometry. Proc Natl Acad Sci 110:7235–7239. https://doi.org/10.1073/pnas.1302395110
39. Ma X, Lai LB, Lai SM et al (2014) Uncovering the stoichiometry of *Pyrococcus furiosus* RNase P, a multi-subunit catalytic ribonucleoprotein complex, by surface-induced dissociation and ion mobility mass spectrometry. Angew Chem Int Ed 53:11483–11487. https://doi.org/10.1002/anie.201405362
40. Lorenzen K, van Duijn E (2010) Native mass spectrometry as a tool in structural biology. Curr Protoc Protein Sci 62:17.12.1–17.12.17. https://doi.org/10.1002/0471140864.ps1712s62
41. VanAernum ZL (2020) Rapid online buffer exchange for screening of proteins, protein complexes and cell lysates by native mass spectrometry. Nat Protoc 15:28
42. Busch F, VanAernum ZL, Lai SM et al (2021) Analysis of tagged proteins using tandem affinity-buffer exchange chromatography online with native mass spectrometry. Biochemistry 60:1876–1884. https://doi.org/10.1021/acs.biochem.1c00138
43. Gan J, Ben-Nissan G, Arkind G et al (2017) Native mass spectrometry of recombinant proteins from crude cell lysates. Anal Chem 89:4398–4404. https://doi.org/10.1021/acs.analchem.7b00398
44. Vimer S, Ben-Nissan G, Sharon M (2020) Direct characterization of overproduced proteins by native mass spectrometry. Nat Protoc 15:236–265. https://doi.org/10.1038/s41596-019-0233-8

45. Ben-Nissan G, Vimer S, Warszawski S et al (2018) Rapid characterization of secreted recombinant proteins by native mass spectrometry. Commun Biol 1:1–12. https://doi.org/10.1038/s42003-018-0231-3
46. Xia Z, Williams ER (2018) Protein-glass surface interactions and ion desalting in electrospray ionization with submicron emitters. J Am Soc Mass Spectrom 29:194–202. https://doi.org/10.1021/jasms.8b05670
47. Susa AC, Lippens JL, Xia Z et al (2018) Submicrometer emitter ESI tips for native mass spectrometry of membrane proteins in ionic and nonionic detergents. J Am Soc Mass Spectrom 29:203–206. https://doi.org/10.1021/jasms.8b05656
48. Kirshenbaum N, Michaelevski I, Sharon M (2010) Analyzing large protein complexes by structural mass spectrometry. J Vis Exp:e1954. https://doi.org/10.3791/1954
49. Fernandez de la Mora J (2000) Electrospray ionization of large multiply charged species proceeds via Dole's charged residue mechanism. Anal Chim Acta 406:93–104. https://doi.org/10.1016/S0003-2670(99)00601-7
50. Fort KL, van de Waterbeemd M, Boll D et al (2018) Expanding the structural analysis capabilities on an Orbitrap-based mass spectrometer for large macromolecular complexes. Analyst 143:100–105. https://doi.org/10.1039/C7AN01629H
51. Rose RJ, Damoc E, Denisov E et al (2012) High-sensitivity Orbitrap mass analysis of intact macromolecular assemblies. Nat Methods 9:1084–1086. https://doi.org/10.1038/nmeth.2208
52. Makarov A, Denisov E (2009) Dynamics of ions of intact proteins in the Orbitrap mass analyzer. J Am Soc Mass Spectrom 20:1486–1495. https://doi.org/10.1016/j.jasms.2009.03.024
53. Benesch JLP, Aquilina JA, Ruotolo BT et al (2006) Tandem mass spectrometry reveals the quaternary organization of macromolecular assemblies. Chem Biol 13:597–605. https://doi.org/10.1016/j.chembiol.2006.04.006
54. Zhou M, Dagan S, Wysocki VH (2013) Impact of charge state on gas-phase behaviors of noncovalent protein complexes in collision induced dissociation and surface induced dissociation. Analyst 138:1353–1362. https://doi.org/10.1039/C2AN36525A
55. Pagel K, Hyung S-J, Ruotolo BT, Robinson CV (2010) Alternate dissociation pathways identified in charge-reduced protein complex ions. Anal Chem 82:5363–5372. https://doi.org/10.1021/ac101121r
56. Harvey S, VanAernum Z, Wysocki V (2021) Surface-induced dissociation of anionic vs cationic native-like protein complexes. https://doi.org/10.26434/chemrxiv.13547837.v1
57. Beardsley RL, Jones CM, Galhena AS, Wysocki VH (2009) Noncovalent protein tetramers and pentamers with " n " charges yield monomers with $n/4$ and $n/5$ charges. Anal Chem 81:1347–1356. https://doi.org/10.1021/ac801883k
58. Kostelic M, Marty M (2020) Deconvolving native and intact protein mass spectra with UniDec. https://doi.org/10.26434/chemrxiv.13417118.v1
59. Marty MT, Baldwin AJ, Marklund EG et al (2015) Bayesian deconvolution of mass and ion mobility spectra: from binary interactions to polydisperse ensembles. Anal Chem 87:4370–4376. https://doi.org/10.1021/acs.analchem.5b00140
60. Reid DJ, Diesing JM, Miller MA et al (2019) MetaUniDec: high-throughput deconvolution of native mass spectra. J Am Soc Mass Spectrom 30:118–127. https://doi.org/10.1007/s13361-018-1951-9
61. Bern M, Caval T, Kil YJ et al (2018) Parsimonious charge deconvolution for native mass spectrometry. J Proteome Res 17:1216–1226. https://doi.org/10.1021/acs.jproteome.7b00839
62. Quintyn RS, Zhou M, Yan J, Wysocki VH (2015) Surface-induced dissociation mass spectra as a tool for distinguishing different structural forms of gas-phase multimeric protein complexes. Anal Chem 87:11879–11886. https://doi.org/10.1021/acs.analchem.5b03441

INDEX

Liangliang Sun and Xiaowen Liu (eds.), *Proteoform Identification: Methods and Protocols*, Methods in Molecular Biology, vol. 2500, https://doi.org/10.1007/978-1-0716-2325-1,

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Science+Business Media, LLC, part of Springer Nature 2022

Q

R

S

T