Maximizing Sequence Coverage in Top-Down Proteomics By Automated Multi-modal Gas-phase Protein Fragmentation

Shliaha, Pavel Vyacheslavovich; Gibb, Sebastian; Gorshkov, Vladimir; Jespersen, Malena Schack; Andersen, Gregers Rom; Bailey, Derek; Schwartz, Jacob; Eliuk, Shannon; Schwämmle, Veit; Jensen, Ole Nørregaard

Published in:
Analytical Chemistry

DOI:
10.1021/acs.analchem.8b02344

Publication date:
2018

Document version
Accepted manuscript

Citation for published version (APA):

Terms of use
This work is brought to you by the University of Southern Denmark through the SDU Research Portal. Unless otherwise specified it has been shared according to the terms for self-archiving. If no other license is stated, these terms apply:
• You may download this work for personal use only.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying this open access version

If you believe that this document breaches copyright please contact us providing details and we will investigate your claim. Please direct all enquiries to puresupport@bib.sdu.dk

Download date: 23. Apr. 2021
Maximizing Sequence Coverage in Top-Down Proteomics By Automated Multi-modal Gas-phase Protein Fragmentation.

Pavel Vyacheslavovich Shliaha, Sebastian Gibb, Vladimir Gorshkov, Malena Schack Jespersen, Gregers R. Andersen, Derek Bailey, Jacob Schwartz, Shannon Eliuk, Veit Schwämmle, and Ole Nørregaard Jensen

Anal. Chem., Just Accepted Manuscript • DOI: 10.1021/acs.analchem.8b02344 • Publication Date (Web): 25 Sep 2018

Downloaded from http://pubs.acs.org on September 27, 2018
Maximizing Sequence Coverage in Top-Down Proteomics By Automated Multi-modal Gas-phase Protein Fragmentation.

Pavel V. Shliaha1,a, Sebastian Gibb2,a, Vladimir Gorshkov1, Malena Schack Jespersen1, Gregers R. Andersen3, Derek Bailey2, Jacob Schwartz3, Shannon Eliuk3, Veit Schwämmle1 and Ole N. Jensen1*. 

1Department of Biochemistry and Molecular Biology and VILLUM Center for Bioanalytical Sciences University of Southern Denmark, DK-5230 Odense M, Denmark
2Department of Anesthesiology and Intensive Care, University Medicine Greifswald, Greifswald, Germany
3Thermo Fisher Scientific, San Jose, California 95134, United States

Abstract: Intact protein sequencing by tandem mass spectrometry (MS/MS), known as top-down protein sequencing, relies on efficient gas-phase fragmentation at multiple experimental conditions to achieve extensive amino acid sequence coverage. We developed the “topdownr” R-package for automated construction of multi-modal (i.e. involving CID, HCD, ETD, ETciD, ETHcD and UVPD) MS/MS fragmentation methods on an orbitrap instrument platform and systematic analysis of the resultant spectra. We used topdownr to generate and analyze thousands of MS/MS spectra for five intact proteins of 10–30kDa. We achieved 90-100% coverage for the proteins tested and derived guiding principles for efficient sequencing of intact proteins. The data analysis workflow and statistical models of topdownr software and multi-modal MS/MS experiments provide a framework for optimizing MS/MS sequencing for any intact protein. Refined topdownr software will be suited for comprehensive characterization of protein pharmaceuticals and eventually also for de novo sequencing and detailed characterization of intact proteins.

Mass spectrometry is routinely used for amino acid sequencing of peptides of length 5-50 amino acid residues. Intact protein sequencing beyond 10 kDa remains a challenge due to physicochemical features of intact proteins (solubility, heterogeneity) and mass spectrometry limitations due to space-charge effects, signal splitting due to natural isotopes and charge state distributions,1 rapid ion signal decay in orbitrap analyzers.2 Nevertheless, top-down mass spectrometry is gaining momentum for the characterization of intact proteins, including protein pharmaceuticals/biologics.3,4 One of the major challenges in top-down mass spectrometry is achieving comprehensive amino acid sequence coverage of intact proteins molecule ions.5 Briefly, gas-phase fragmentation of intact protein molecule ions leads to cleavage at selected peptide bonds. Preferably only one peptide bond is broken per protein molecule, which generates N- and C-terminal fragment ions. Thus, fragmentation of a population of identical protein molecule ions will lead to a distribution of fragments, which are then mass analyzed to generate a tandem mass spectrum that reflects the amino acid sequence.6 However, any particular fragmentation method and MS/MS parameter setting will lead to cleavage of only a subset of bonds. Amino acid sequence coverage approaching 100% currently relies on the combination of multiple MS/MS fragmentation methods and various associated MS/MS parameter settings.7 Collisionally activated dissociation (fragmentation) of gas-phase ions is based on precursor ion acceleration and collision with neutral gas molecules to produce high intensity fragment ions, however, often with limited amino acid sequence coverage. On hybrid ion trap-orbitrap instruments, CAD can be performed in a low electrostatic field in an ion trap analyzer through slow ion heating (referred to as CID, or resonance-type CAD) or at higher electrostatic field in a collision cell through higher energy collisions (referred to as HCD or beam-type CAD). HCD is identical to CAD performed on triple-quadrupole (QqQ) or hybrid quadrupole-time-of-flight (Q-TOF) instruments.8 Fragmentation by electron capture dissociation (ECD)9,10 or electron transfer dissociation (ETD)11 produces a more uniform distribution of fragments thereby increasing amino acid sequence coverage.12,13 However, ECD and ETD fragmentation efficiency and resultant fragment ion intensity is generally lower than in CID or HCD4,14 and the spectra are dominated by non-fragmented or charge reduced precursor ion signals. This can be explained by intramolecular interactions holding the ETD and ECD products together. These interactions become more prominent as the charge density decreases.16 To address this problem, ETD products and remaining precursor ions including the charge reduced species can be subjected to an additional round of fragmentation by CID12 or HCD.17 More recently, concomitant infrared photoactivation18–20 has been suggested as a method to increase ETD efficiency, presumably through unfolding of gas phase protein structures.21 Ultraviolet photodissociation (UVPD) is an emerging technique for intact protein fragmentation.22–24 It has been reported, however, that UVPD spectra can be complicated to interpret due to multiple types of fragments (abc and xyz) being produced and post-fragmentation molecule ion rearrangements.25,26
Several studies investigated some combinations of different fragmentation techniques and conditions for improving amino acid sequence coverage in intact proteins.\textsuperscript{13,23,26} Attempts to use LC-MS/MS, by repeated injections of intact proteins or their mixtures, were also reported.\textsuperscript{22} However, a truly systematic and comprehensive analysis requires testing hundreds to thousands of fragmentation conditions, which is not compatible with the LC-MS/MS timescale. Hence, other studies focused on the analysis of proteins delivered by direct infusion ESI.\textsuperscript{7,13,20,27,28} These studies involved manual interrogation of fragmentation conditions, and, hence, were limited to the analysis of tens to hundreds of spectra with a human operator going through fragmentation conditions and recording spectral annotations. This can lead to significant bias, as an operator has to choose a small set of conditions to test based on limited a priori evidence, potentially missing optimal and/or previously untested conditions.

The remarkable improvement of data analysis software for top-down proteomics in recent years has addressed many issues of label-free quantification, proteofrm identification and PTM annotation in LC-MS/MS.\textsuperscript{29–33} However, systematic investigation of protein fragmentation by direct infusion ESI requires a novel approach. We here propose a workflow that we call “multimodal MS/MS” for systematic optimization of intact protein sequencing by combination of different modes of fragmentation: CID, HCD, ET D, ETcID, EThcD and UVPD. Each of these fragmentation methods (ETD in particular) rely on multiple parameters that affect the quality of the generated MS/MS spectra, i.e. the number and intensity of fragment ion signals. Systematic testing of these parameters and their combinations requires obtaining thousands of MS/MS spectra. Hence we developed \textit{topdownr} R package which allows: 1) effortless construction of MS/MS methods with multiple fragmentation modes and settings; 2) systematic investigation of thousands of fragmentation spectra for achieving maximum protein sequence coverage; 3) deriving rules for efficient protein fragmentation. \textit{topdownr} was implemented using R that offers a plethora of statistical methods and visualization tools developed by the R-community. We combine the \textit{topdownr} R software with the Orbitrap Fusion Lumos MS, the only currently available platform that combines six fragmentation modes, i.e. CID, HCD, enhanced ET D,\textsuperscript{34} EThcD, EThcD and UVPD. We applied \textit{topdownr} to the systematic investigation of fragmentation of five proteins with different molecular weight (MW). The generated data provided a deeper understanding of the fundamentals of protein fragmentation and suggested ways for achieving maximum possible protein sequence coverage for comprehensive top-down protein characterization.

MATERIALS AND METHODS

Direct infusion of protein analytes. Supplementary Table 1 describes the source of proteins used in this work. Prior to infusion proteins were desalted on PD MiniTrap G-25 columns (GE healthcare), as recommended by the manufacturer, but limiting elution volume to 850 µL to completely avoid salt. Proteins were infused into the ESI source from 10 µM solution (assuming complete recovery from the desalting step) of 50% ACN and 0.1% formic acid solution using 1002 LTN 2.5mL Hamilton syringe at 600 nL/min through 50 µm ID silica line and electrosprayed through FS360-20-10-N-20-6.35CT emitter (New Objective). Silica line and the emitter were joined through 700002843 union (Waters) with 700002715 (Waters) peek nuts. To ensure spray stability solvent was filtered inline through 0.5µm solvent filter (Idex M-520). MS/MS was acquired with the following global settings: ion transfer tube spray voltage at 3kV, temperature - 400 °C, HCD cell collision pressure - 2mTorr, quadrupole precursor isolation width - 1 m/z, maximum injection time 200 msec, Orbitrap resolution – 120k. The rest of the settings were varied as described in the “Construction of multimodal MS/MS methods” section.

Computational procedures. Detailed description of the algorithm and computational procedures is given in Supporting Information

RESULTS AND DISCUSSION

Algorithm and experiment overview.

The general overview of the \textit{topdownr} algorithms and workflow is presented in Figure 1. A user specifies the set of values for fragmentation parameters to be tested. The \textit{topdownr} software creates MS/MS method files with all meaningful combinations of the chosen values of parameters. We refer to such combinations as fragmentation conditions. Two or more replicate ESI MS/MS spectra are obtained for each fragmentation condition.

The acquired MS/MS spectra are deisotoped by the Xtract\textsuperscript{35} node in Proteome Discoverer 2.1 and abc and xyz type fragment ions are matched against the input protein sequence (FASTA format) with user specified tolerance (default value used in this study is 5 ppm). After ions to fragment matching the dataset is filtered to preserve MS/MS spectra with reproducible injection times and ions with reproducible and high intensity signals (see methods section). Then for every fragmentation condition replicate MS/MS spectra are combined and the number of observed fragments and the proportion of assigned ion signal intensity is computed. This allows subsequent investigation of the effects of individual or combinatorial MS/MS parameter settings by statistical and visualization methods.

The data is then investigated in detail by statistical methods and visualization techniques to reveal the effects of individual and combinatorial MS/MS parameter settings (the detailed explanation is given below). The tools are presented in logical order: starting with methods that allow quality control (linear discriminant analysis) and comparisons of whole datasets (random forest regression), towards investigating importance of individual fragmentation parameters (LDA scaling factors and RFR variable contributions) and fragmentation parameters interactions (bubble plots) and finally visualization of fragmented ions matched in individual spectra (fragmentation maps and annotated spectra). The software also allows selecting a subset of fragmentation conditions that in combination provide the highest amino acid sequence coverage for a protein.

The outcome of this systematic analysis is a set of simple rules for efficient protein fragmentation, achieving maximum amino acid sequence coverage in a given timeframe.

Construction of multimodal MS/MS methods

We systematically investigated MS/MS fragmentation of five proteins: C3a recombinant protein (9.8 kDa), histone H4 (11.2 kDa), histone H3.3 (15.2 kDa), horse myoglobin (16.9 kDa), bovine carbonic anhydrase (29kDa) (see Supplementary Table 1)
Ion isolation and injection parameters:

1) Ion injection (ion density in ion trap and orbitrap analyzers) expressed as AGC (automatic gain control) target: 1e5, 5e5, 1e6.

2) Three charge states (low m/z, middle m/z, high m/z). Exact m/z differ between proteins (Supplementary Table 1).

3) Amount of ETD reagent, expressed as ETD Reagent AGC Target – 1e6, 5e6, 1e7.

Fragmentation types and energies:

1) ETD Reaction Time - 2.5, 5, 10, 15, 30, 50 msec.

2) Collisional activation type – Beam type (HCD) or resonance type (CID)

3) Normalised Collision energy (NCE) – 7, 14, 21, 28, 35.

4) UVPD reaction time 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60 msec (carbonic anhydrase only)

This set of fragmentation parameters values generates 3024 combinations (see Supplementary Table 2) of which 1872 were meaningful and 1152 were nonsensical (e.g. having ETD reagent target of higher than 0 for a CID scan). Nonsensical combinations were automatically pruned. We acquired two or more replicate spectra for each fragmentation condition for the combinations (see Figure S3) of which 1872 were aggregated into 1872 fragmentation conditions.

**Investigation of fragment types**

The **topdownr** ion-to-fragment matching is highly computationally effective and allows for fast surveys of different mass tolerance settings and post fragmentation molecule ion rearrangements crucial for top-down proteomics. For example, it has been reported that c fragments tend to abstract hydrogen radicals from z fragments in ETD (resulting in c+1 and z-1 fragments). UVPD has recently been implemented on the Orbitrap Fusion Lumos and reports suggest that a, x and y ions are prone to hydrogen abstractions resulting in a+1, a+2, x+1, x+2, y-1 or y-2 ions. We have initially matched ions to a, b, c and x, y, z fragments with four possible fragment ion hydrogen abstractions (minus two, minus one, plus one and plus two hydrogen radicals) for the six fragmentation conditions (CID, HCD, ETD, ETciD, EThcD and UVPD) for carbonic anhydrase. Figure S3 illustrates that the results of our analysis are in line with the previous reports with a+1, a+2, x+1, x+2, y-1 and y-2 rearrangements present in UVPD and c+1 and z+1 rearrangements in ETD, ETciD and EThcD spectra.

**Systematic analysis of the effects of fragmentation parameter values and their interactions**

The analysis of the resulting data poses a significant challenge. The purpose is to understand the contribution of tuning individual fragmentation parameters, yet the parameters are not studied individually, but convoluted together in the fragmentation conditions. In order to address this challenge we developed several novel visualization tools and applied statistical methods, which are not normally used in proteomics studies. We present the tools in the order, which progresses from very general overview of the whole datasets toward focus on smaller parts of the data in increasingly higher details.
Linear Discriminant Analysis (LDA) serves as an initial quality control to confirm that the values of fragmentation parameters tested have an effect on the number of fragments matched.

LDA is a linear transformation technique similar to principal component analysis (PCA). Unlike PCA, however, LDA aims to maximize separation between data points belonging to different classes. In the present analysis every fragmentation condition is a data point. Its features are the values of fragmentation parameters. It is assigned to one of the ten classes based on the number of matched fragments relative to the best performing condition, rounded to one decimal place (see online methods for more details). Figure 2 displays the LDA analysis of data for the five standards proteins. The best performing fragmentation conditions cluster together on the right side of the panels. This indicates that fragmentation optimization experiment was overall successful, since conditions with the highest number of ions matched to fragments share some underlying optimal values for fragmentation parameters. The ovals are drawn to contain fragmentation conditions belonging to a particular fragmentation method (CID, HCD, ETD, ETciD, EThcD). It is noteworthy that the MS/MS methods involving CID (CID and ETciD) have much lower spread than HCD based methods (HCD and EThcD). This indicates that the outcome of fragmentation is much less dependent on a particular value of NCE for CID than for HCD.

We then applied Random Forest Regression (RFR) to our dataset to examine to what degree fragmentation parameter optimization performed on one protein can be extrapolated to other proteins. RFR is a machine learning algorithm that can generate models based on a set of features and their outcomes. As with LDA the features are fragmentation parameters, while the outcomes are the number of fragments for each fragmentation condition. Using RFR, we created a fragmentation model for each of the five proteins under investigation. Next we applied every model to every protein dataset. Each model performed best against the protein dataset it was derived from (Figure 3, diagonal). This demonstrates that optimization performed on one protein cannot be perfectly extrapolated to other proteins and the highest coverage can only be achieved by optimizing fragmentation for each protein individually. The H3.3 model performed very well for the myoglobin dataset. On the contrary, the CA model demonstrated poor predictive power for the C3A dataset. This is probably due to similarity in MW and charge state distributions between H3.3 and myoglobin and large difference in these properties between CA and C3A.

Next we investigated the importance of tuning individual fragmentation parameters for maximizing sequence coverage. Firstly, we extracted the scaling factors from the LDA model (Figure 4A). The magnitude of a scaling factor for a particular fragmentation parameter reflects it contribution for separating of optimal and sub-optimal fragmentation conditions in LDA space and therefore for achieving maximum sequence coverage. Secondly, we extracted the variable importance from the RFR models, which indicates the contribution of individual parameter values to the predictions of the number of fragments that RFR models make (Figure 4B). Both the approaches identified ETD reaction time, molecular ion charge state and

Figure 2. Linear discriminant analysis (LDA) for the five standard proteins investigated. The ovals are drawn to contain 90% of fragmentation conditions for a particular mode.
HCD NCE as the most important parameters. This means that the user should focus optimization experiments on these parameters, while varying ETD reagent target and CID activation NCE parameters will have less of an effect on MS/MS outcome in the MW range tested.

![Figure 3](image-url) Heatmap for evaluation of random forest regression models performance against the datasets for the five proteins in the study (see online methods for further details). Every model is evaluated against every dataset. Comb dataset and model is based on combination of all proteins. The numbers in the cells are mean square errors. Low mean square error indicates good performance of the RFR model against the corresponding dataset.

The next stage of analysis includes bubble plots to visualize combinatorial effects of fragmentation parameter values. Each fragmentation condition is represented as a point, for which coordinates represent fragmentation parameter values, while its size and color reflect the number of matched fragments and the proportion of ion current accounted by them. Figure 5 displays a bubble plot for 608 fragmentation conditions (out of 1872) of the myoglobin molecule ion with charge state 24+ (see figure legend for a detailed explanation of the bubble plot). Figure S4 displays the bubble plots for additional four proteins and for other myoglobin charge states. Several patterns are evident from the bubble plot for this myoglobin 24+ charge state. Firstly, as discussed above, ETD reagent target values have little effect on the fragmentation outcome. Secondly, CID on its own and as a supplementary activation technique (ETciD) is more robust than HCD, meaning that overfragmentation is far less likely. Thirdly, while HCD peak performance is at 7-14 NCE, CID performs better at 28-35 NCE. Finally, the performance of ETciD seems to peak at 5 msec reaction time and then decreases until 50 msec.

The main hypothesis of this study is that different MS/MS fragmentation conditions provide complementary protein sequence information and to the down software provides fragmentation maps for visualization of this complementarity. Prior to building fragmentation maps, we convert a, b and c fragments to a single type called N-terminal and x, y and z fragments to a single type called C-terminal. If the bond is covered in both C- and N-directions, we refer to this as Bidirectional fragments (see Figure S5 for illustration). Having only three fragment types (N-terminal, C-terminal and Bidirectional) allows building easy-to-interpret fragmentation maps visualizing many conditions side-by-side, as shown in Figure 6 for 152 (out of 1872) conditions of myoglobin charge state 24+, AGC Target of 1e6 and EtdReagent Target of 1e7 (for ETD, ETciD and EThcD modes). The map demonstrates that CID and HCD contribute little amino acid sequence information (panels 1 and 2), mostly at the protein N and C-termini, whereas ETD provides coverage throughout the protein sequence (panel 3-5). Thus, ETD clearly outperforms CID and HCD for myoglobin (24+). While, CID and HCD spectra are very similar, HCD achieves optimal fragmentation at lower NCE. A similar trend is observed when ETciD and EThcD are compared, with EThcD resulting in overfragmentation at NCE higher than 14. CID with or with out supplemental activation forms a C-shaped pattern with longer fragmentation events covering bonds closer to N and C termini. This is most likely due to secondary fragmentation events.

![Figure 4](image-url) Contribution of individual parameters optimization to the fragmentation outcome. A. measured as scaling factors from the LDA model B. measured as the variable importance from the RFR models. The higher the column the higher the impact of that fragmentation parameter.

Selecting a subset of complementary fragmentation conditions to achieve maximum coverage.

The combined coverage from all 1872 CID, HCD, ETD, ETciD and EThcD fragmentation conditions tested approaches 100% for four out five proteins: C3A (95%), H3.1 (99%), H4 (99%) and myoglobin (99%). Carbonic anhydrase combined sequence coverage was 80%. We applied UVPD to carbonic anhydrase, which increased fragment coverage to 89%. Importantly, we use very stringent criteria for observing a fragment: the mass accuracy has to be within 5ppm and the CV for the intensity of the fragment has to be within 30% in replicate spectra. This indicates that multimodal MS/MS scanning will allow achieving close to 100% sequence coverage for any protein in the MW range tested.

ACS Paragon Plus Environment
Figure 5. Bubble plot for myoglobin charge 24+. Each point is a single fragmentation condition. Point size represents the number of assigned fragments (relative to maximum observed for the protein), while its colour represents the proportion of ion current accounted by them. The position on the plot reflects its fragmentation parameters. The panel of the point specifies the ion AGC target (shown on the right), ETD reagent target and type of supplemental activation (shown on the top), while position within the panel reflects ETD reaction time (increasing along the x-axis) and supplemental activation NCE (increasing along y-axis). See example point for explanation. To investigate the effect of ETD reagent targets, compare the columns of panels as shown with green brackets. To find optimal ETD reaction time, move right within a panel until a column with the largest points is found. To find optimal NCE, move up within a panel until the row with largest points is found.

It is, however, impractical to scan so many fragmentation conditions and topdownr allows selecting a subset of conditions that provide maximum cumulative coverage. The algorithm works by first selecting a condition that provides maximum coverage; it then iteratively selects conditions that add most previously uncovered bonds to the combination, until no further bonds could be covered. The result of this analysis is presented in Supplementary Table 3. Maximum coverage can be achieved with 12, 6, 9, 8 and 31 conditions for C3A, H4, H3.1, myoglobin and CA respectively. As an example, in case of myoglobin the first condition provides the best coverage of 96 bonds (63% of the sequence). The second condition covers 69 bonds overall, with 25 bonds not observed in condition one (conditions #1 and 2 together cover 121 bonds, which provide 73% cumulative coverage). The third condition has 77 bonds overall (higher than the condition #2), with only 13 bonds complementary to the coverage of conditions #1 and #2. A total of eight conditions will allow maximum coverage of 99%. The table can be visualized by fragmentation maps as shown in the Figure S6. Figure S7 shows spectra for CA conditions #1 (ETcID), #3 (UVPD) and #5 (CID) annotated by topdown.

Almost all fragmentation conditions selected for the 5 proteins have AGC for precursor at 1e6 (47 instances) or 5e5 (16), and only 4 of 1e5, illustrating that at least in this range there is no space charge effects, confirming previous reports on high capacity ETD.34 There is no such clear preference for ETD reagent AGC target, but there appears to be a slight trend towards lower ETD reagent targets: 1e6 (18 instances), 5e6 (17 instances), 1e7 (13 instances). While the best fragmentation condition (having the highest number of fragments) involves ETD for all five proteins; four of the five proteins: H4, H3.1, myoglobin and carbonic anhydrase have either HCD or CID as one of the conditions in their combination. All three protein specific charge states (low, middle and high) are present in combinations for all five proteins, illustrating that targeting different charge states provides complementary information. Optimal CID and HCD results are achieved at lower charge states, while optimal ETD fragmentation at higher charge states.

CONCLUSIONS

Here we present the topdownr package for automatic optimization of intact protein fragmentation on the Orbitrap Fusion Lumos MS platform. We applied it to five proteins of different MW, producing the largest top-down dataset to date for systematic investigation of protein fragmentation. The analysis of this data allowed us to generalize principles of achieving high coverage in top-down proteomics presented below.
Figure 6. Multi-modal MS/MS fragmentation map for myoglobin. Each column represents a fragmentation condition, while each row represents a bond. The colors in the cells reflect the type of fragments observed for this bond at this fragmentation condition (green – N-terminal, orange – C-terminal, purple – Bidirectional). The conditions are split into five panels representing the five fragmentation methods: CID, HCD, ETD, ETciD and EThcD. The labels for the conditions represent a combination of three numbers in order: ETD reaction time (msec), CID NCE and HCD NCE energies. Thus 2.5_7_0 label represents ETciD with 2.5msec ETD activation and 7 NCE CID activation. The conditions are organized so that moving left to right the ETD reaction time and the energy of supplemental activation increase.

Importantly, results from different tools are largely in agreement, increasing the confidence of our conclusions:

1) For best results fragmentation has to be optimized for each protein individually. See RFR analysis (Figure 3).

2) Maximum AGC target should be used whenever possible, meaning there is no saturation of the ion trap capacity due to space-charge effects. See bubble plots (Figure 5) and “selection of complementary fragmentation conditions” section.

3) Different charge states (low, medium, high) should be analyzed by MS/MS since they provide complementary amino acid sequence information, generally starting with high or middle charge state for ETD and lower charge states for CID and HCD. See “selection of complementary fragmentation conditions” section.

4) ETD reagent target has little effect of fragmentation in the range tested (1e6 to 1e7), See LDA scaling factors and RFR weights (Figure 4).

5) Both short and extended ETD times should be used (extended reaction times cover bonds closer to N and C-termini, shorter times cover protein center). See fragmentation maps (Figure 6 and Figure S6).

6) Prioritize CID over HCD, since CID performance is less dependent on the specified NCE.

7) To achieve maximum amino acid sequence coverage, ETD with supplemental activation (ETciD or EThcD) should be used in addition to CID and if possible UVPD. See “selection of complementary fragmentation conditions” section and fragmentation maps (Figure 6 and Figure S6).

Some of these rules (e.g. two, three, seven) confirm previous knowledge, while others (e.g. four) provide new insight, yet others (e.g. six) challenge the previous reports, potentially suggesting distinct behavior of the Orbitrap Fusion Lumos MS platform.

Using these rules, it is possible to dramatically decrease the number of fragmentation parameters to be tested in an optimization experiment. In particular, if only maximum analyte AGC is used (e.g. fixed at 1e6), a fixed ETD reagent target is specified (1e6) and the analysis uses CID (omits HCD) as a supplemental activation technique the amount of instrument time can be reduced to under 2 hours per protein (compared to 30 hours per protein used in the current work). In fact, limiting analysis to these conditions results in almost no loss of fragment coverage: for C3A (88%), H3.1 (97%), H4 (97%) and myoglobin (99%) and CA (85%).

In this study we averaged 40 microscans for every fragmentation spectrum, which corresponds to roughly 30 seconds acquisition time for every fragmentation condition. This scan rate is readily compatible with direct infusion experiments that are often used for intact protein characterization by top-down mass spectrometry. Given that signals for some fragments is low (in particular for ETD and UVPD spectra) it is possible that optimal fragmentation conditions and rules for maximizing coverage may be different for online LC-MS/MS, which allows for less microscans.

Data generated through topdownr guided MS/MS experiments of intact proteins can be used to build more sophisticated mathematical models of fragmentation using third party software.

In conclusion, our topdownr approach for intact protein sequencing by tandem mass spectrometry achieves near-complete amino acid sequencing of intact proteins in the mass...
range 10-30 kDa. Future efforts in software development will focus on implementation of multistage activation methods and characterization of larger proteins with post-translational modifications.

**ASSOCIATED CONTENT**

**Supporting Information**

Supporting information.docx: detailed description of bioinformatics procedures and supplementary figures; ST1_proteins.xlsx: proteins used in this study; ST2_fragmentation_conditions.xlsx: all potential fragmentation conditions, including nonsensical ones; ST3_best_conditions_table.xlsx: combinations of conditions providing maximum coverage for the five proteins in the study.

**REFERENCES**
