Published in final edited form as:

Anal Chem. 2018 February 06; 90(3): 2369–2375. doi:10.1021/acs.analchem.7b04999.

IMTBX and Grppr: Software for Top-Down Proteomics Utilizing Ion Mobility-Mass Spectrometry

Dmitry M. Avtonomov^{†,§}, Daniel A. Polasky^{‡,§}, Brandon T. Ruotolo[‡], and Alexey I. Nesvizhskii^{†,¶}

[†]Department of Pathology, University of Michigan, Ann Arbor, Michigan, USA

[‡]Department of Chemistry, University of Michigan, Ann Arbor, Michigan, USA

[¶]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA

Abstract

Top-down proteomics has emerged as a transformative method for the analysis of protein sequence and post-translational modifications (PTMs). Top-down experiments have historically been performed primarily on ultra-high resolution mass spectrometers due to the complexity of spectra resulting from fragmentation of intact proteins, but recent advances in coupling ion mobility separations to faster, lower resolution mass analyzers now offer a viable alternative. However, software capable of interpreting the highly complex 2-dimensional spectra that result from coupling ion mobility separation to top-down experiments is currently lacking. In this manuscript we present a software suite consisting of 2 programs, IMTBX (“IM Toolbox”) and Grppr (“Grouper”), which enable fully automated processing of such data. We demonstrate the capabilities of this software suite by examining a series of intact proteins on a Waters Synapt G2 ion-mobility equipped mass spectrometer, and compare the results to the manual and semi-automated data analysis procedures we have used previously.

Introduction

Rapid assessment of protein sequence by mass spectrometry (MS) has enabled a revolution in the analysis of biological and biochemical samples,¹ powering the rapidly-expanding field of proteomics. Typical proteomic workflows involve the enzymatic digestion of proteins into peptides, which can be separated by chromatography to enable rapid analysis by MS.^{2,3} Despite the advances made with these techniques, the requirement for enzymatic digestion in these “bottom-up” approaches often complicates identification and complete analysis of the post-translational modifications (PTMs) and proteoforms that are critical to biological

Correspondence to: Dmitry M. Avtonomov; Alexey I. Nesvizhskii.

[§]Contributed equally to this work

The authors declare no competing financial interests.

Supporting Information Available

User-defineable parameters used for “semi-automated” and IMTBX/Grppr processing. The description of “Averagine” isotopic grouping algorithm used in Grppr.

function of proteins.^{4,5} In response, “top-down” MS techniques have been developed to sequence intact proteins without enzymatic digestion.^{6–8}

Top-down MS holds great promise, but remains less common than bottom-up approaches due to a number of technical challenges associated with achieving efficient separation and fragmentation of intact proteins. One of the major challenges associated with top-down techniques is the low signal-to-noise ratio (SNR) of sequence-informative fragment ions.⁹ Fragmentation of intact proteins typically results in a population of ions distributed amongst charge states ranging from 1⁺ to 5⁺ or higher, with masses extending from 100 to over 10,000 Da. For larger proteins in particular, the need to generate hundreds of fragments from a single precursor in order to achieve sufficient sequence coverage often results in poor SNR for fragment ions, as well as many instances of overlapping isotopic distributions. To resolve these overlapping distributions, most top-down MS workflows utilize Fourier transform (FT) instruments (both ion cyclotron resonance (ICR) and Orbitrap) that possess extremely high MS resolving power. While capable of producing exceptionally narrow line widths for MS measurements, such equipment also requires significant acquisition times per spectrum to achieve high performance results, creating a situation that gives rise to low instrument duty cycles and difficulties in coupling to chromatographic separations. Slow acquisitions compound the challenge of low SNR for fragment ions, as fewer spectra can be acquired to average and improve spectral quality. Significant efforts have been made to mitigate these challenges,^{10–12} but they remain a barrier to widespread adoption of top-down approaches.

Ion mobility-mass spectrometry (IM-MS) offers a promising alternate path to alleviate several of the challenges detailed above by providing additional separation of the product ions created during top-down sequencing experiments. IM separates ions based on their migration time through a chamber pressurized with inert neutrals under the influence of a relatively weak electric field. Ions of different orientationally-averaged sizes (collision cross sections, CCSs) and charges take different amounts of time to travel through the IM device, offering significant improvements to the peak capacity of a given mass analyzer for a complex fragment ion population¹³. The improvements offered by IM have seen adoption for bottom-up proteomics,^{14–16} and initial proof-of-concept demonstrations for top-down protein sequencing have been reported,^{17,18} but the routine use of IM-MS for top-down proteomics remains rare. This trend runs counter to a dramatic increase in the use of IM-MS in many other MS-based fields, such as native and structural MS,¹⁹ metabolomics,²⁰ drug discovery,²¹ which have appeared since the introduction of commercial IM-MS instruments.

Perhaps the most significant reason for the lack of adoption of IM-MS for top-down proteomics is the general lack of software for processing IM-MS data containing the complex fragment ion populations inherent in top-down analyses. It has been well established for MS data that top-down data requires specialized analysis tools that differ from those used for bottom-up data,^{22,23} and the same need for specialized tools is clearly required for top-down IM-MS data. While many software tools have been developed for automated handling of top-down MS data without IM, they are generally incapable of handling IM-MS data, due both to differences in the spectral characteristics of the raw data and a general inability to read raw IM-MS data from any format in which it is natively generated. Indeed, all reports of IM-MS top-down to date, to our knowledge, rely largely or

entirely on manual interpretation of spectra. The complexity of fragment ion populations generated in top-down experiments makes this a massively time-consuming process, as hundreds to thousands of peaks must be identified in each dataset, severely limiting the scope and speed of viable experiments. Here, we describe a software suite capable of processing top-down IM-MS data in a rapid and automated fashion to enable routine top-down IM-MS analyses. The suite consists of two tools, IMTBX (“IM Toolbox”) and Grppr (“Grouper”), which together can completely process complex top-down fragmentation data into an accurate monoisotopic peak list in seconds. The list can be annotated in a sequencing program of choice. A plotting module enables flexible visualization of IM-MS data and detected peaks to evaluate the impact of peak detection parameters. The suite is modular, with a robust scripting interface to enable extension into any IM-MS data processing workflow. Both Grppr and IMTBX can be downloaded from github repositories: <https://github.com/chhh/Grppr>, <https://github.com/chhh/IMTBX> (the latter includes a combined package with both tools).

Methods

Experimental

Ubiquitin (bovine), Myoglobin (bovine), Carbonic Anhydrase (bovine), Transthyretin (human), β -lactoglobulin (bovine), Avidin (chicken), Aldolase (rabbit), and ammonium acetate were purchased from Sigma-Aldrich (St. Louis, MO). Serum amyloid p-component (human) protein was purchased from EMD Millipore (Billerica, MA). Proteins were buffer exchanged into 100mM ammonium acetate using Micro Biospin 6 spin columns (BioRad, Hercules, CA) prior to analysis by IM-MS. All data was collected using a Synapt G2 HDMS IM-Q-ToF mass spectrometer (Waters, Milford, MA). Intact protein ions were generated using a direct infusion nESI source in positive mode. Applied capillary voltage was 1.5 kV, and the sample cone and extraction cones were operated at 40V and 0V, respectively. Instrument settings were optimized to preserve native-like protein ions and noncovalent complexes prior to CID, as described previously.²⁴ A single charge state protein ion was selected in the quadrupole and collisionally activated in the trap traveling-wave ion guide (voltage applied depending on protein molecular weight, from 75V for Ubiquitin to 200V for Aldolase), which was pressurized to $2 - 4 \times 10^{-2}$ mbar with argon gas. Ion mobility separation was performed at a pressure of 3 mbar with 20V wave height and 150m/s wave velocity. The ToF mass analyzer was operated over the range of 100–6000 m/z for Ubiquitin and Myoglobin, and 100–8000 m/z for all other proteins, at a pressure of 1.5×10^{-6} mbar.

Data processing

Data was analyzed with IMBTX and Grppr and a “semi-automated” workflow comprised of several software packages. For IMTBX analysis, IMTBX version 2.9.1 and Grppr version 0.3.6 were used. Parameters used can be found in Supporting Table 1. A wrapper script written in Python 3.5 was used to operate both IMTBX and Grppr in a batch processing mode prior to the development of a user interface. Isotopic cluster output from Grppr was annotated using ProSight Lite²⁵ or a custom peak annotation program in Python 3.5 for batch processing. For the semi-automated processing comparison workflow, regions of two-dimensional IM-MS data space corresponding to fragment charge states were extracted

using TWIMExtract.¹⁵ Extracted data were smoothed (Savitsky-Golay, 0.2 m/z window size, 3 cycles), peak-picked (Intensity threshold 500, picking height 90%), and de-isotoped (max charge 5, isotope mass tolerance 0.05 m/z, isotope intensity tolerance 100%) in mMass v5.5.0.^{26–28} Final output was annotated with the same custom annotation script as IMTBX data. Charge state comparisons were generated by comparing the annotated charge state against the m/z and drift time of each detected cluster. ProSight Lite, TWIMExtract, and mMass are all publicly available free software packages.

Processing raw data is a two-stage process: peak picking and isotopic cluster detection. First single 2D peaks (m/z, ion mobility drift time) in raw data are detected and written to a file, followed by grouping these peaks into isotopic clusters. Two in-house built software tools perform these steps: IMTBX (Ion Mobility Toolbox), which does raw peak picking, and Grppr (Grouper), which performs isotopic grouping.

Raw feature detection (IMTBX)

IMTBX is written in C# (.NET 4.5.1), it can currently read data from Waters IM-MS or SONAR enabled instruments, such as Synapt G2-Si or Xevo systems. It can also handle regular LC-MS data without IM. An API to support custom data format adapters is under development. For assistance with unsupported data formats, please contact the authors. IMTBX can operate in two modes: scan summation/averaging followed by 2D peak detection in a single combined scan, the mode that is used throughout this manuscript, or 3D peak detection for LC-MS data where retention time is added to the model. If the data contains a lock-mass channel (that is Waters specific), it can be processed as well - time dependent mass calibration correction can be calculated and applied to spectra as they are being processed.

Each IM scan in Waters data is represented as a sparse matrix of N rows and M columns, where N is the number of IM drift bins (typically 200), and M is the number of bins for m/z axis (varies, typically 100,000 – 300,000 bins). Data from other instruments that can be represented in such a simple format can be accommodated for IMTBX analysis. IMTBX data processing steps are schematically shown in the left part of Figure 1. For top-down applications, all ion mobility scans are added together or averaged (IMTBX has options for both), then a 2D filter of user-selected size and shape (typically Gaussian) is applied for smoothing. Optionally, a step of removing “lone” data points can be applied, i.e. non-zero intensity data points that do not have many neighboring non-zero points. The exact meaning of “having many neighbors” is, again, user selectable, with the default settings optimized for data observed in this and similar studies. After that, local maxima in the scan are found and used as seed points for fitting the data with 2D Gaussians. Even though the peaks are actually not ideal 2D Gaussians, especially in the IM dimension, variations from this expectation do not negatively influence the data analysis described here to a significant extent. In addition, the actual maxima locations are also reported for each detected 2D peak, not only the means in m/z and IM dimensions. Several stages of data filtration are shown in Figure 1, a more detailed overview can be found in Supporting Information on page S-4.

As top-down proteomics data often contains low SNR features that possess many local maxima even after filtration, our fitting procedure may result in multiple peaks in the

proximity of the true peak apex location. Thus, the resulting list of detected peaks is further filtered to remove such spurious signals. For each peak, multiple values are reported including: m/z , m/z standard deviation, drift time, drift time standard deviation, intensity at apex, total intensity (the sum of intensities over all data points assigned to the identified peak), and area (the number of data points used to define the peak). Several output text formats are available. For each peak, SNR is established using root mean square (RMS) noise estimation. Depending on user-defined parameters, peak lists that result from the above procedure may contain peaks that result from noise rather than analyte signals. In such cases, our isotopic grouping program – Grppr (described below) - uses 2D isotopic clustering as a stringent filter, ignoring peaks for which other isotopologues are not reliably detected. Most of these processing steps can be visualized in the software GUI (graphical user interface) that can display 1D and 2D spectra both before and after signal processing, with options to enable or disable individual steps. The GUI for our software can overlay detected peaks and/or isotopic clusters over raw data as shown in Figure 2.

Isotopic clustering (Grppr)

Grppr is a general purpose 2D de-isotoping algorithm implemented in Java 8, which is not tied to any particular input data format. While these two dimensions refer primarily to the IM and MS, as discussed above, Grppr can also be used for LC-MS datasets, where 2D features appear in m/z and retention time space. As input, Grppr requires a set of 2D peaks containing peak intensities as well as locations and widths in both dimensions (m/z and drift time for IM-MS data). Grppr has a simple application programming interface (API) facilitating adaptation of any input peak list format for Grppr analysis.

Grppr provides two algorithms for isotopic cluster detection: “Convex” and “Averagine”. The “Convex” implementation searches for intensity-wise convex sets of neighboring peaks, i.e. groups of peaks that are spaced in the m/z dimension in a manner consistent with a single charge state assignment for all peaks and with intensities that consecutively decrease when moving starting from the highest peak of the set. While this approach works reasonably well for LC-MS bottom-up proteomic data, analyses of training datasets revealed systematic errors in this approach for top-down protein sequencing datasets. For example, after summation of hundreds of individual IM-MS spectra contained within a typical top-down protein sequencing dataset, chemical noise from low abundance fragments results in small peaks at virtually each unit mass, which leads to incorrect isotope cluster assignments. In contrast to “Convex” isotope cluster detection, “Averagine” clustering implementation is based on the Averagine²⁹ model amino acid. For each input peak Grppr receives, a theoretical isotopic distribution is calculated and compared to the data. The advantage of this algorithm is that it can infer the correct mass of the monoisotope in low SNR datasets containing significant noise backgrounds, even when the monoisotopic peak is not detectable. For top-down proteomics data, we have found that the “Averagine” implementation is superior to the “Convex” approach described above. A high-level overview of the Averagine based deisotoping algorithm used in Grppr is presented in the right half of Scheme 1, while a more detailed description with pseudo-code can be found in Supporting Information.

Results and discussion

Typical top-down proteomics workflows involve the separation and subsequent fragmentation of individual intact proteins in a mass spectrometer. IM offers substantial advantages for these experiments by providing a gas-phase separation technique that can separate the peptide product ions, analogous to the separation of tryptic peptides in bottom-up proteomics. Separation of such fragments using IM allows for accurate peak detection and isotopic clustering in highly complex datasets. Figure 3 demonstrates the utility of IM separation of fragments for a particular region of a plot of drift time versus m/z containing eight isotopic clusters derived from the top-down collision induced dissociation (CID) of Ubiquitin. When all IM drift time spectra are combined together, two clear isotopic clusters are observed (Fig. 3B), along with an overlapped set of peaks that cannot be readily resolved into individual clusters. The two-dimensional IM-MS plot, however, clearly shows the presence of eight distinct clusters having significant relative abundance values (plotted on a square root intensity scale) due to the separation of these ions in the IM dimension (Fig. 3A). Individual mass spectra can be plotted by averaging spectra from IM bins close to drift apexes of the various species (Fig. 3C–E), showing accurate determination of the isotopic clusters by Grppr.

By processing the data natively in two dimensions, IMTBX and Grppr are able to analyze more peaks, corresponding to more fragment ions of interest, in a given m/z range than the corresponding analysis performed with MS alone. While the examples shown here were acquired using an IM device coupled to a time of flight (ToF) mass analyzer, the multidimensional peak detection and isotopic clustering performed by these tools can be applied to IM-MS instruments of any configuration and resolution. IM-MS currently lacks a suitable open data format, so IMTBX is capable of reading Waters file format. Despite this, the algorithms described here can in principle be applied to any IM-MS input files, and Grppr in particular can be applied to any two-dimensional MS-based peak list.

To assess the benefits of IMTBX and Grppr for top-down IM-MS, we benchmarked our new software using top-down data acquired for a range of model proteins. Furthermore, these benchmarking results were compared to a semi-automated workflow that involves manual extraction of IM-MS regions corresponding to fragment charge state trends for analysis by existing MS data processing tools. IM and MS are only partially orthogonal, as IM drift time is a function of collision cross section (size and shape) and charge of an analyte, and mass and size tend to be generally correlated for analytes with similar overall shapes (e.g. protein fragments in top-down analyses). This results in well-defined trend lines within IM-MS data, largely defined by ion charge states^{13, 18} (Fig. 4A, cyan dashed lines). Interestingly, multiple trends were observed for fragment ions of large masses and charges (Fig. 4 inset), likely due to the prevalence of multiple gas-phase conformations for longer amino acid sequences. Extracting each trend line separately allows for MS analysis using existing tools while preserving some of the enhanced peak capacity provided by IM. For our semi-automated analysis, an open-source MS data viewer and peak processor, mMass^{26,27,28}, was used to accomplish signal-to-noise thresholding, peak picking, and isotopic clustering for each charge trend subset of the data. IMTBX and Grppr allow for the analysis of the complete IM-MS dataset natively, by processing the raw data in two dimensions, enabling detection of

off-trend peaks that are missed by the semi-manual workflow, as well as greatly improving the quality of isotopic cluster detection for large fragments.

In the absence of a dataset with known monoisotopic peaks for each feature, evaluating the accuracy of assignments from any computational or manual workflow is challenging. We have thus compared Grpr and the alternative workflow with respect to their relative accuracies in assigning charge states to isotopic clusters using the observed trend lines (Fig. 4A) as an objective comparison metric. For all isotopic clusters detected in eight analyses of Ubiquitin monomer and Avidin tetramer fragmentation data, using both IMTBX/Grpr and the semi-automated workflow, the charge state determined by the analysis was compared to the position (in m/z and drift time) of the fragment relative to the observed trend lines in order to evaluate its correctness (Fig. 4). For higher molecular weight fragments, where multiple trends were observed, positioning of the signal within all observed trends was allowed. Grpr had an average charge state assignment accuracy of $99.0 \pm 0.6\%$ across approximately 50 datasets processed (24,158 assignments), vs. $68 \pm 12\%$ for the semi-manual workflow (8 datasets, 4292 assignments). The natively 2D analysis of IMTBX and Grpr results in high and reproducible accuracy, allowing for unsupervised and fully automated analysis of top-down fragmentation datasets.

To demonstrate the capabilities of IMTBX and Grpr for top-down IM-MS analysis, a series of proteins and protein complex standards, ranging from 8.5 to 158 kDa, were fragmented to generate datasets for top-down analysis, collecting 300–600 IM-MS scans for each. Each dataset was processed with IMTBX (requiring approximately 10–30 seconds for each analysis) and Grpr (requiring less than 5 seconds per analysis) and the resulting peak lists were exported into ProSight Lite for annotation to their respective protein sequences (Fig. 5). IMTBX detected an average of just over 9,000 peaks per dataset (Fig. 5A), which were clustered into several hundred isotopic clusters. It is normal for many peaks not to get matched to isotopic clusters, as the peak list includes precursors and fragments at high m/z for which isotopic peaks cannot be easily resolved. Two sequence fragmentation maps are shown for reference (Fig. 5B): Ubiquitin, demonstrating that the annotated clusters result in excellent sequence coverage for a small protein, and Carbonic Anhydrase, showing coverage for a much larger protein, along with identification of the post-translationally acetylated serine at the n-terminus (red box). Processing even this relatively small dataset manually would have required days to weeks of analysis time (Supporting Table 1), given the need to manually assign thousands of peaks into hundreds of isotopic clusters to generate a peak list for annotation. IMTBX and Grpr allow rapid (seconds to minutes), unsupervised, and accurate handling of this data to enable a wide range of top-down experiments on IM-MS instrumentation.

Conclusion

The potential benefits of IM separation for top-down proteomics have been demonstrated previously.^{14,18} Perhaps the largest obstacle preventing the realization of this potential has been a lack of software capable of automating even the initial step of translating two-dimensional IM-MS data into a list of monoisotopic peaks, which can be fed into an annotation program of choice (e.g. ProSight Lite). Fragmentation of intact proteins typically

generates hundreds to thousands of fragment ions, each consisting of multiple isotopic peaks (as many as 20 for high molecular weight fragments), resulting in spectra that can contain many thousands of features. Manually annotating data of this complexity is extremely time-consuming, and the lack of automated processing tools severely limits the scope of experiments that can be performed on IM-MS platforms. With rapid and accurate determination of isotopic clusters, analysis of complex top-down fragmentation data collected on IM-MS instrumentation can be fully automated using IMTBX and Grppr. The same software suite without modifications can also be applied to bottom-up proteomics LC-IM-MS data (not described in this manuscript). This enables experiments that require processing of hundreds or thousands of fragmentation spectra to be realistically performed on these instruments, including characterization of fragmentation mechanisms and typical top-down proteomics analyses with coupled intact protein separations. Both Grppr and IMTBX can be downloaded from github repositories: <https://github.com/chhh/Grppr>, <https://github.com/chhh/IMTBX> (the latter includes a combined package with both tools). There is also a supporting website with examples and documentation at <https://chhh.github.io/IMTBX> (capitalization of IMTBX is important for the link to work properly).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Scott Geromanos and Steve Ciavarini for providing the software library for reading raw data from Waters' instruments and valuable discussions regarding Waters specific mass calibration using the lockmass channel. Ruotolo group research in the area of top-down protein sequencing is supported by the National Institutes of Health, through the National Institute of General Medical Sciences (R01 GM095832).

References

1. Yates JR. Journal of the American Chemical Society. 2013; 135:1629–1640. [PubMed: 23294060]
2. Mayne J, Ning Z, Zhang X, Starr AE, Chen R, Deeke S, Chiang CK, Xu B, Wen M, Cheng K, Seebun D, Star A, Moore JI, Figeys D. Analytical Chemistry. 2016; 88:95–121. [PubMed: 26558748]
3. Zhang Y, Fonslow BR, Shan B, Baek MC, Yates JR. Chemical Reviews. 2013; 113:2343–2394. [PubMed: 23438204]
4. Beck M, Claassen M, Aebersold R. Current Opinion in Biotechnology. 2011; 22:3–8. [PubMed: 20888217]
5. Nesvizhskii AI, Aebersold R. Molecular & Cellular Proteomics. 2005; 4:1419–1440. [PubMed: 16009968]
6. Catherman AD, Skinner OS, Kelleher NL. Biochemical and Biophysical Research Communications. 2014; 445:683–693. [PubMed: 24556311]
7. Savaryn JP, Catherman AD, Thomas PM, Abecassis MM, Kelleher NL. Genome medicine. 2013; 5:53. [PubMed: 23806018]
8. Siuti N, Kelleher NL. Nature Methods. 2007; 4:817–821. [PubMed: 17901871]
9. Compton PD, Zamdborg L, Thomas PM, Kelleher NL. Analytical Chemistry. 2011; 83:6868–6874. [PubMed: 21744800]
10. Huang TY, McLuckey SA. Proteomics. 2010; 10:3577–3588. [PubMed: 20848674]
11. Cannon JR, Cammarata MB, Robotham SA, Cotham VC, Shaw JB, Fellers RT, Early BP, Thomas PM, Kelleher NL, Brodbelt JS. Analytical Chemistry. 2014; 86:2185–2192. [PubMed: 24447299]

12. Tran JC, Zamdborg L, Ahlf DR, Lee JE, Catherman AD, Durbin KR, Tipton JD, Vellaichamy A, Kellie JF, Li M, Wu C, Sweet SMM, Early BP, Siuti N, LeDuc RD, Compton PD, Thomas PM, Kelleher NL. *Nature*. 2011; 480:254–258. [PubMed: 22037311]
13. McLean JA, Ruotolo BT, Gillig KJ, Russell DH. *International Journal of Mass Spectrometry*. 2005; 240:301–315.
14. Baker ES, Burnum-Johnson KE, Ibrahim YM, Orton DJ, Monroe ME, Kelly RT, Moore RJ, Zhang X, Théberge R, Costello CE, Smith RD. *Proteomics*. 2015; 15:2766–2776. [PubMed: 26046661]
15. Haynes SE, Polasky DA, Dixit SM, Majmudar JD, Neeson K, Ruotolo BT, Martin BR. *Analytical Chemistry*. 2017; 89:5669–5672. [PubMed: 28471653]
16. Meier F, Beck S, Grassl N, Lubeck M, Park MA, Raether O, Mann M. *Journal of Proteome Research*. 2015; 14:5378–5387. [PubMed: 26538118]
17. Halgand F, Habchi J, Cravello L, Martinho M, Guigliarelli B, Longhi S. *Analytical chemistry*. 2011; 83:7306–15. [PubMed: 21800924]
18. Zinnel NF, Pai PJ, Russell DH. *Analytical Chemistry*. 2012; 84:3390–3397. [PubMed: 22455956]
19. Lanucara F, Holman SW, Gray CJ, Eyers CE. *Nature Chemistry*. 2014; 6:281–294.
20. Paglia G, Astarita G. *Nature Protocols*. 2017; 12:797–813. [PubMed: 28301461]
21. Niu S, Rabuck JN, Ruotolo BT. *Current Opinion in Chemical Biology*. 2013; 17:809–817. [PubMed: 23856053]
22. Park J, Piehowski PD, Wilkins C, Zhou M, Mendoza J, Fujimoto GM, Gibbons BC, Shaw JB, Shen Y, Shukla AK, et al. *Nature Methods*. 2017; 14.
23. Sun RX, Luo L, Wu L, Wang RM, Zeng WF, Chi H, Liu C, He SM. *Analytical Chemistry*. 2016; 88:3082–3090. [PubMed: 26844380]
24. Ruotolo BT, Benesch JLP, Sandercock AM, Hyung SJ, Robinson CV. *Nature Protocols*. 2008; 3:1139–1152. [PubMed: 18600219]
25. Fellers RT, Greer JB, Early BP, Yu X, LeDuc RD, Kelleher NL, Thomas PM. *Proteomics*. 2015; 15:1235–1238. [PubMed: 25828799]
26. Strohalm M, Hassman M, Košata B, Kodíček M. *Rapid Communications in Mass Spectrometry*. 2008; 22:905–908. [PubMed: 18293430]
27. Strohalm M, Kavan D, Novák P, Volný M, Havlíček V. *Analytical Chemistry*. 2010; 82:4648–4651. [PubMed: 20465224]
28. Niedermeyer TH, Strohalm M. *PLoS ONE*. 2012; 7, e44913.
29. Senko MW, Beu SC, McLafferty FW. *Journal of the American Society for Mass Spectrometry*. 1995; 6:229–233. [PubMed: 24214167]

Figure 1.

Top left: Raw unprocessed data from a single IM scan for 3+ and 4+ Ubiquitin ions from a Waters Synapt-G2 platform. A slice of data for m/z 1295 – 1298, drift 30 – 55 containing 2 differently charged ions is shown. Horizontal axis is m/z , vertical axis is IM drift time. **Top right:** Same data slice after Gaussian filtering with default parameters. **Bottom left:** 580 filtered scans added together, bilinear interpolation applied. **Bottom right:** Same as *bottom left* with intensities square rooted. Detected single 2D Gaussian peaks are marked with ovals, semi-axes corresponding to 2 standard deviations. Horizontal lines mark the locations and span of 2 detected isotopic clusters: charge 3+ at the top and 4+ at the bottom.

Figure 2.

Graphical interface for viewing raw IM-MS data, processing stages and ion detection results. The software displays 1D spectra, 2D spectra and the mass axis bin scaling. The various controls shown at the bottom allow to turn on/off and customize individual processing steps available in IMTBX. Shown in the figure is a 2D map view of a portion of an ion mobility scan (ion mobility drift time vs. m/z , intensity as color) processed by IMTBX with default parameters with an overlay of isotopic clustering results by Grprr as series of purple arrows.

Figure 3.

Isotopic clusters that cannot be resolved without IM separation. Two-dimensional representation of IM-MS data shows the presence of multiple isotopic clusters (panel A). The composite mass spectrum summed across all drift times, i.e. the spectrum without IM separation, is shown at the top (panel B). Only the two most intense clusters at 1413 and 1415 can be identified, as the shape of the group at m/z 1419 is distorted due to interference from overlapping clusters. All 8 clusters can be identified utilizing IM separation with charge states ranging from 2+ to 5+, as shown in the mass spectra at 3 different drift times displayed on the right (panels C–E). Vertical red and green lines denote the most intense and monoisotopic peaks of a cluster correspondingly as reported by GrpPr. Notice that fragment $y63^{5+}$ (far right ion in spectrum D) has drift time between that of 3+ and 2+ ions, following the observed trend for 5+ ions (see Fig. 4).

Figure 4.

Left: Distribution of fragment ions from Ubiquitin in IM-MS space. Fragments are observed in “trend lines” due to the correlation between size and mass. Purple markers denote detected isotopic clusters. For Ubiquitin (6+ precursor) fragments, trends are observed for singly, doubly, triply, and quadruply charged fragments. However, several distinct trends can be observed for some charge state families, including 2+, 4+ and 5+, indicating the presence of multiple gas-phase conformations for these large fragments. The inset shows a zoom-in of a small IM-MS region, where yellow numbers indicate the charge state of an isotopic cluster detected at a particular location. **Right:** Charge assignment accuracy comparison. Charge state assignments were assessed by comparing the location of the fragment in IM-MS space and its assigned charge against the observed trend lines. For charge states 2+, 4+ and 5+, where multiple trends were observed, positioning of the signal within any of the observed trends were allowed. IMTBX + GrpPr had an accuracy of 99% in assigning correct charge states, vs. 68% for the semi-automated data processing workflow.

Figure 5.

IMTBX and GrpPr processing of 8 protein standards analyzed by top-down IM-MS. **Panel A:** The number of raw single peaks (2D features) detected by IMTBX for each protein (*blue*), the total number of isotopic clusters detected by GrpPr (*green*) and the number of clusters matched to a fragment ion to annotate the protein sequence (*yellow*). Aldolase - tetramer (158 kDa), SAP = Serum Amyloid protein pentamer (125 kDa), Avidin - tetramer (64 kDa), TTR = transthyretin tetramer (56 kDa), Carb. Anh. = Carbonic Anhydrase monomer (29 kDa), B-lac. = β -lactoglobulin monomer (18 kDa), Myo. = myoglobin monomer (17 kDa), Ubq. = ubiquitin monomer (8.5 kDa). An average of 9040 features were detected for each raw dataset. An average of 490 isotopic clusters were found and an average of 112 were annotated to the protein sequence. **Panel B:** Sequence maps showing sites of fragmentation for Ubiquitin (*top*) and Carbonic Anhydrase (*bottom*). The red box at the n-terminal serine of Carbonic Anhydrase indicates an acetylated serine, with the masses of all fragments containing that residue adjusted accordingly.

Scheme 1.

A schematic overview of the data processing steps taken by IMTBX and Grppr for top-down IM-MS data processing.