Journal of proteome research

pubs.acs.org/jpr

Technical Note

IsoForma: An R Package for Quantifying and Visualizing Positional Isomers in Top-Down LC-MS/MS Data

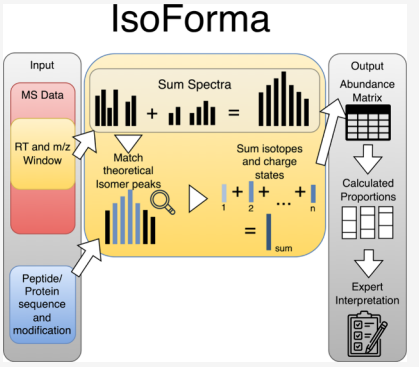
David J. Degnan,[¶] Logan A. Lewis,[¶] Lisa M. Bramer, Lee Ann McCue, James J. Pesavento, Mowei Zhou, and Aivett Bilbao*

Cite This: https://doi.org/10.1021/acs.jproteome.3c00681

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Proteoforms, the different forms of a protein with sequence variations including post-translational modifications (PTMs), execute vital functions in biological systems, such as cell signaling and epigenetic regulation. Advances in top-down mass spectrometry (MS) technology have permitted the direct characterization of intact proteoforms and their exact number of modification sites, allowing for the relative quantification of positional isomers (PI). Protein positional isomers refer to a set of proteoforms with identical total mass and set of modifications, but varying PTM site combinations. The relative abundance of PI can be estimated by matching proteoform-specific fragment ions to top-down tandem MS (MS2) data to localize and quantify modifications. However, the current approaches heavily rely on manual annotation. Here, we present *IsoForma*, an open-source R package for the relative quantification of PI within a single tool. Benchmarking *IsoForma*'s performance against two existing workflows produced comparable results and improvements in speed. Overall, *IsoForma* provides a streamlined process for quantifying PI, reduces the analysis time, and offers an essential framework for developing customized proteoform analysis workflows. The software is open source and available at https://github.com/EMSL-Computing/isoforma-lib.



KEYWORDS: *proteoform quantification, positional isomers, top-down proteomics, R package, LC-MS/MS*

INTRODUCTION

A protein from a single gene can yield many different "proteoforms" via chemical changes known as post-translational modifications (PTMs)[1] and various PTMs serve unique and critically important roles in cellular biology, such as cell signaling and epigenetic regulation. Proteoforms can be detected using mass spectrometry (MS)-based proteomics approaches.[2] Established bottom-up approaches are effective at detecting which PTMs are present but cannot characterize how many PTM combinations are arranged on the intact protein because the link between peptides and their specific modification site is lost, resulting in proteoform ambiguity.

A single PTM type can induce distinct functional changes at different locations in the protein, such as well-known examples of histone methylation/acetylation in gene regulation,[3] and phosphorylation in cell signaling.[4] Therefore, the precise characterization of stoichiometry and site localization are important for understanding proteoform biology. Identifying and quantifying isomeric proteoforms has been a historical challenge due to limitations in MS resolving power and bioinformatic tools.[5] Analyzing intact proteins directly without prior digestion retains the connectivity of PTMs and can distinguish among proteoforms that generate identical peptides in a bottom-up approach. Recent advancements in top-down mass spectrometry have allowed the collection of better-quality MS/MS data of intact proteoforms. Positional isomers (PI) are proteoforms with the same mass, carrying the same type and number of modifications but arranged at different sites. Fragment ion abundances from dissociation of ions created from mixtures of multiply modified histone H4 correspond to their respective intact ion abundances measured.[6] Bioinformatics methods have been demonstrated for the quantification of PI using the abundance of fragments with a modification at each given position relative to each other. Brunner et al.[7] reported the QPI software, however, QPI requires mass deconvolved spectra generated by external software and does not support liquid chromatography (LC) MS data. Rommelfanger et al.[8] reported a workflow supporting LC-MS using several tools in a stepwise fashion: MASH Explorer[9] for database searching, TopFD[10] for mass deconvolution, a custom Python script named SMC for finding and combining corresponding MS/MS, ProsightLite[11] for annotating MS/MS and assigning preliminary PTMs, TDValidator[12] for visualizing

Special Issue: Women in Proteomics and Metabolomics

Received: October 17, 2023
Revised: February 9, 2024
Accepted: February 19, 2024

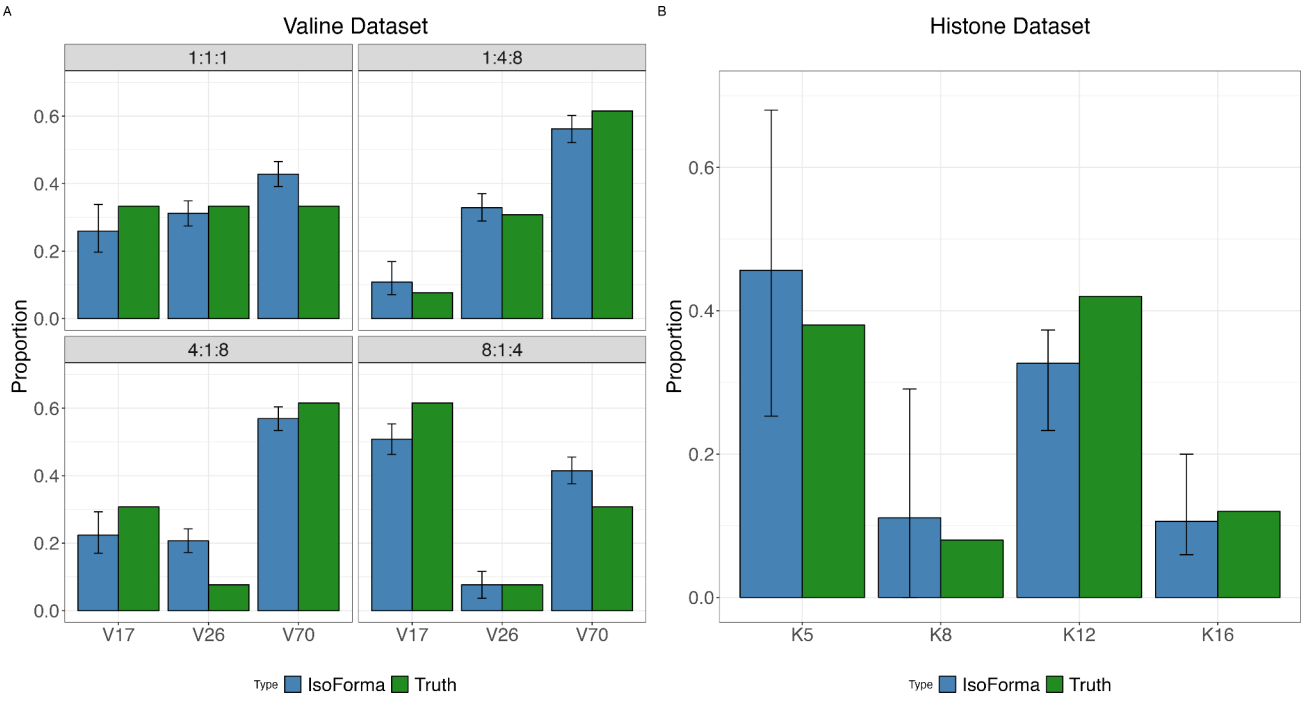


Figure 1. *IsoForma* quantification benchmarking. (A) Expected relative abundance of ubiquitin modified with heavy isotopic labels at positions 17, 26, and 70 compared with the proportions calculated by *IsoForma*. (B) Proportions calculated by *IsoForma* for an LC-MS experiment compared to previously reported results (blue boxes) for 4 isoforms of the histone H4 protein in the green microalga *Chlamydomonas reinhardtii*. Each proteoform includes an alpha-amino acetyl modification and a single internal lysine acetylation at one of the residues: K5, K8, K12, or K16.

MS/MS and filtering PTMs, and finally Microsoft Excel for combining fragment abundances into proteoform quantification and proportions. Besides several tools used in a stepwise fashion, this workflow requires extensive manual analysis.

To facilitate the relative quantification of PI, we built *IsoForma*, an R[13] package that streamlines the workflow for both the relative quantification and visualization of PI. This package provides all the necessary steps for completing a full PI analysis following identification in an automated fashion.

METHODS

Package Design

The *IsoForma* package was built in R 4.2.2 to perform the following steps: detecting and summing MS2 data, generating isoform sequences with user-specified modifications, matching calculated and experimental peaks, summing the abundances from multiple fragment charge states and isotopic peaks, generating a final abundance matrix, and calculating the proportions (i.e., the abundances of each input isoform relative to each other). The wrapper function only requires that users provide MS2 scan data, a relevant retention time (RT) range, the unmodified protein or peptide sequence, and the modifications of interest.

For LC-MS data, the first step of the algorithm automatically detects and sums MS2 spectra from ThermoFisher raw or open format XML-based MS files. *IsoForma* utilizes *pspecterlib* wrapper functions (a package now providing programmatic access to functionality originally implemented as an end-user tool[14]) to allow the construction of peak data from either any file type that can be read into R, or read peak data directly from the MS files using other available packages.[15−18] *IsoForma's* MS2 scan detection algorithm selects MS2 scans with precursor m/z and RT values that fall within the corresponding windows. For each MS2, the precursor theoretical isotopic distribution is matched to the closest MS1 in RT and if the number of matched peaks satisfies the user-specified minimum number of MS1 isotopic peaks (typically 3) within the instrument-specified precursor isolation window, then the MS2 is selected. This step was designed to remove low-quality MS2 spectra. The selected MS2 are then summed, and m/z values are rounded to the user-specified PPM threshold (typically 5 ppm) to improve signal-to-noise ratios.

To specify the modifications of interest in the parameters, a format was created to simplify the evaluation of multiple isoforms at once. For example, "Methyl,(3,5,7)[2]" generates every possible combination of an input amino acid sequence with two methyls on either the third and the fifth, the third and the seventh, or the fifth and the seventh residues. To fix a position and filter the results, a "^" can be inserted next to the number. The maximum number of modifications in a combination could be fixed by changing the value in parentheses, e.g., "[2]". A full list of examples can be found in the vignette at https://emsl-computing.github.io/isoforma-lib.

The main *IsoForma* pipeline function generates the user-specified isoforms with their theoretical fragments and matches them to the summed MS2 using the *pspecterlib* algorithm, which incorporates the *MSnbase*,[18] *Rdisop*,[19] and *isopat*[20] packages. All charge states and isotopic peaks per fragment are then summed together to maximize relative abundance values and, thus, improve the quality of the results. Intensities are divided by their corresponding charge states before summing. *IsoForma* returns an abundance matrix with rows by a selected ion type (a, b, c, x, y, or z) for every residue and each column as a fragment. To calculate relative proportions, every fragment ion in the abundance matrix is divided by the highest

https://doi.org/10.1021/acs.jproteome.3c00681
J. Proteome Res. XXXX, XXX, XXX−XXX

abundance per row, which is always the last isoform in a series. The mean relative proportion and their 95% confidence intervals are then reported as each isoform's proportion by fitting proportions to a beta distribution with the *stats4*[13] R package. If all other isoforms have a proportion calculation that is nonzero, then the last isoform is calculated and otherwise becomes NA (not available value). All quantification results can be saved as csv files. A figure with proportions and an interactive annotated spectrum in an HTML file can also be generated.

IsoForma Installation

The following steps should be followed to install the IsoForma package and necessary dependencies:

1. Download and install R: https://ftp.osuosl.org/pub/cran/
2. Download, install, and open RStudio (Free version): https://www.rstudio.com/products/rstudio/download/
3. If reinstalling, remove any previous versions of pspecter:remove.packages("pspecterlib")
4. Install packages:
   install.packages("devtools")
   devtools::install_github("EMSL-Computing/pspecterlib")
   devtools::install_github("EMSL-Computing/isoforma-lib")
5. Test by completing the steps with example data in the vignette at: https://emsl-computing.github.io/isoforma-lib.

MS Data

Two public data sets were used to benchmark *IsoForma*'s performance. The first, from Brunner et al.[7] contains known proportions of isoforms, where the positional isomers of synthetic ubiquitin with a heavy isotope-labeled Valine positioned at residue 17th, 26th, or 70th, were mixed in known concentrations. The second data set, used in Rommelfanger et al.,[8] can be found in the MassIVE database under accession code MSV000088458 (doi:10.25345/C5J286). As opposed to the Brunner data set which has known proportions, the proportions in this second data set were calculated using the manual pipeline described in the introduction, which uses MASH Explorer,[9] TopFD,[10] ProSightLite,[11] and TDValidator.[12] These values were then treated as the true proportions for the purpose of comparison to the *IsoForma* results. This example, including the input parameter files and shorter versions of the LC-MS raw files in mzML format, can be found in the example code and vignette of the *IsoForma* package.

RESULTS & DISCUSSION

When compared to a data set where the positional isomer concentrations are known, *IsoForma* reasonably recovers the proportions of the ubiquitin proteoforms (Figure 1A). In the 1:1:1 ratio where the expected relative proportions of the proteoforms are all 0.333, *IsoForma* calculated the V17 modification proportion as 0.234 with a 95% confidence interval (CI) of 0.177−0.308, the V26 as 0.330 (95% CI of 0.289−0.370), and the V70 as 0.436 (95% CI of 0.396−0.477). In the 1:4:8 ratio where the proportions are 0.077, 0.308, and 0.615 for V17, V26, and V70 respectively, *IsoForma* calculated 0.072, 0.362, and 0.566. For the 4:1:8 ratio, where the true values are 0.308, 0.077, and 0.616, *IsoForma* produced 0.204 (95% CI of 0.141−0.294), 0.231 (95% CI of 0.191−0.272), and 0.565 (95% CI of 0.524−0.605). In the 8:1:4 ratio the expected proportions were 0.615, 0.077, and 0.308 for V17, V26, and V70 respectively, and *IsoForma* calculated 0.481 (95% CI of 0.422−0.540), 0.0900 (95% CI of 0.045−0.134), and 0.429 (95% CI of 0.385−0.475). To improve the proportion measurements, especially in cases where there is an over- or underestimation, we suggest that users increase their sample size. Here, measurements are from a single run of a sample through a mass spectrometer and, thus, have an n of 1. Despite the small sample size, in the unequal proportion comparisons (e.g., 1:4:8, 4:1:8, and 8:1:4), *IsoForma* correctly ranked the proteoforms in terms of most to least abundant.

Similarly, for the LC-MS data with 4 isoforms of the histone H4 protein, *IsoForma* generated comparable results of isoform proportions calculated using MASHExplorer and TopFD (Figure 1B). The *IsoForma* algorithm for selection of relevant MS2 successfully retrieved 4 of the 6 expected scans, but 2 scans were missed due to the erroneous precursor charge assigned by the acquisition software (charges 36 and 21, rather than 17). Overall, K5 and K12 are more abundant than K8 and K16, as expected. For both analyses, *IsoForma* completed a set of isoforms (*e.g.*, 1:1:1) in under 1 min using a 2.3 GHz 8-Core Intel Core i9MacBook Pro.

IsoForma is a single tool for calculating and visualizing the relative quantification of PI that is faster than hand annotation and eliminates the need for multiple tools. The modular design of the package allows for experienced bioinformaticians to adapt the pipeline as necessary, like skipping the MS2 scan selection step for direct infusion experiments. For all new users, *IsoForma* contains the full example code and documentation in vignettes.

CONCLUSION

IsoForma is an R package for calculating and visualizing the proportion or relative abundance of multiple proteoforms representing PI in a sample. Although domain knowledge and experimental context are needed to specify relevant PTMs and get the most benefit from this tool, *IsoForma* serves to simplify the relative quantification of PI by automating and significantly reducing the time investment needed for data processing. Our automated results show comparable accuracy for infusion data with expected concentrations and LC-MS data analyzed by the traditional hand-annotation method. Proteomics via top-down mass spectrometry is a continuously growing field that needs computational tools to keep pace, and *IsoForma* is a technology built to make such analyses easier. Future work will involve code refinements to quantify more complicated groups of isoforms, improved determination of precursor charge for MS2 selection, and a graphical user interface. Our tool can facilitate the uncovering of PTM relative abundance of protein isoform groups and is expected to assist in the study of biologically relevant modification sites to understand functional roles of proteoforms.

AUTHOR INFORMATION

Corresponding Author

Aivett Bilbao − *Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Washington 99354, United States;* orcid.org/0000-0003-2985-8249; Email: Aivett.Bilbao@pnnl.gov

https://doi.org/10.1021/acs.jproteome.3c00681
J. Proteome Res. XXXX, XXX, XXX−XXX

Authors

David J. Degnan − *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99354, United States;* orcid.org/0000-0001-5737-7173

Logan A. Lewis − *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99354, United States;* orcid.org/0000-0003-2301-300X

Lisa M. Bramer − *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99354, United States;* orcid.org/0000-0002-8384-1926

Lee Ann McCue − *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99354, United States;* orcid.org/0000-0003-4456-517X

James J. Pesavento − *Saint Mary's College of California, Moraga, California 94575, United States;* orcid.org/0000-0001-6107-3666

Mowei Zhou − *Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Washington 99354, United States;* Present Address: Department of Chemistry, Zhejiang University, Hangzhou, Zhejiang 310058, China; orcid.org/0000-0003-3575-3224

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.3c00681

Author Contributions

[¶](D.J.D., L.A.L.) These authors contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to thank the cohort of individuals who tested the package: Kelly Stratton, Dr. Javier Flores, Bea Meluch, Sam Obermiller, and Natalie Winans. This work was performed on a project award DOI: 10.46936/intm.proj.2021.60149/60000429 from the Environmental Molecular Sciences Laboratory, a DOE Office of Science User Facility sponsored by the Biological and Environmental Research program under Contract No. DE-AC05-76RL01830. J.J.P. was supported by the National Science Foundation under Grant No. MCB-1943493.

■ REFERENCES

(1) Smith, L. M.; Agar, J. N.; Chamot-Rooke, J.; Danis, P. O.; Ge, Y.; Loo, J. A.; Pasa-Tolic, L.; Tsybin, Y. O.; Kelleher, N. L. The Human Proteoform Project: Defining the human proteome. *Sci. Adv.* **2021**, *7* (46), 0734.

(2) Halder, A.; Verma, A.; Biswas, D.; Srivastava, S. Recent advances in mass-spectrometry based proteomics software, tools and databases. *Drug Discov Today Technol.* **2021**, *39*, 69−79.

(3) Lu, C.; Coradin, M.; Porter, E. G.; Garcia, B. A. Accelerating the Field of Epigenetic Histone Modification Through Mass Spectrometry-Based Approaches. *Mol. Cell Proteomics* **2021**, *20*, 2257.

(4) Latorraca, N. R.; Masureel, M.; Hollingsworth, S. A.; Heydenreich, F. M.; Suomivuori, C. M.; Brinton, C.; Townshend, R. J. L.; Bouvier, M.; Kobilka, B. K.; Dror, R. O. How GPCR Phosphorylation Patterns Orchestrate Arrestin-Mediated Signaling. *Cell* **2020**, *183* (7), 1813.

(5) Melby, J. A.; Roberts, D. S.; Larson, E. J.; Brown, K. A.; Bayne, E. F.; Jin, S.; Ge, Y. Novel Strategies to Address the Challenges in Top-Down Proteomics. *J. Am. Soc. Mass Spectrom.* **2021**, *32* (6), 1278−1294.

(6) Pesavento, J. J.; Mizzen, C. A.; Kelleher, N. L. Quantitative analysis of modified proteins and their positional isomers by tandem mass spectrometry: human histone H4. *Anal. Chem.* **2006**, *78* (13), 4271−4280.

(7) Brunner, A. M.; Lossl, P.; Geurink, P. P.; Ovaa, H.; Albanese, P.; Altelaar, A. F. M.; Heck, A. J. R.; Scheltema, R. A. Quantifying Positional Isomers (QPI) by Top-Down Mass Spectrometry. *Mol. Cell Proteomics* **2021**, *20*, No. 100070.

(8) Rommelfanger, S. R.; Zhou, M.; Shaghasi, H.; Tzeng, S. C.; Evans, B. S.; Pasa-Tolic, L.; Umen, J. G.; Pesavento, J. J. An Improved Top-Down Mass Spectrometry Characterization of Chlamydomonas reinhardtii Histones and Their Post-translational Modifications. *J. Am. Soc. Mass Spectrom.* **2021**, *32* (7), 1671−1688.

(9) Wu, Z.; Roberts, D. S.; Melby, J. A.; Wenger, K.; Wetzel, M.; Gu, Y.; Ramanathan, S. G.; Bayne, E. F.; Liu, X.; Sun, R.; et al. MASH Explorer: A Universal Software Environment for Top-Down Proteomics. *J. Proteome Res.* **2020**, *19* (9), 3867−3876.

(10) Kou, Q.; Xun, L.; Liu, X. TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. *Bioinformatics* **2016**, *32* (22), 3495−3497.

(11) Fellers, R. T.; Greer, J. B.; Early, B. P.; Yu, X.; LeDuc, R. D.; Kelleher, N. L.; Thomas, P. M. ProSight Lite: graphical software to analyze top-down mass spectrometry data. *Proteomics* **2015**, *15* (7), 1235−1238.

(12) Fornelli, L.; Srzentic, K.; Huguet, R.; Mullen, C.; Sharma, S.; Zabrouskov, V.; Fellers, R. T.; Durbin, K. R.; Compton, P. D.; Kelleher, N. L. Accurate Sequence Analysis of a Monoclonal Antibody by Top-Down and Middle-Down Orbitrap Mass Spectrometry Applying Multiple Ion Activation Techniques. *Anal. Chem.* **2018**, *90* (14), 8421−8429.

(13) *R: A language and environment for statistical computing*; https://www.R-project.org/, 2023.

(14) Degnan, D. J.; Bramer, L. M.; White, A. M.; Zhou, M.; Bilbao, A.; McCue, L. A. PSpecteR: A User-Friendly and Interactive Application for Visualizing Top-Down and Bottom-Up Proteomics Data in R. *J. Proteome Res.* **2021**, *20* (4), 2014−2020.

(15) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **2012**, *30* (10), 918−920.

(16) Kockmann, T.; Panse, C. The rawrr R Package: Direct Access to Orbitrap Data and Beyond. *J. Proteome Res.* **2021**, *20* (4), 2028−2034.

(17) Trachsel, C.; Panse, C.; Kockmann, T.; Wolski, W. E.; Grossmann, J.; Schlapbach, R. rawDiag: An R Package Supporting Rational LC-MS Method Optimization for Bottom-up Proteomics. *J. Proteome Res.* **2018**, *17* (8), 2908−2914.

(18) Gatto, L.; Gibb, S.; Rainer, J. MSnbase, Efficient and Elegant R-Based Processing and Visualization of Raw Mass Spectrometry Data. *J. Proteome Res.* **2021**, *20* (1), 1063−1069.

(19) Bocker, S.; Letzel, M. C.; Liptak, Z.; Pervukhin, A. SIRIUS: decomposing isotope patterns for metabolite identification. *Bioinformatics* **2009**, *25* (2), 218−224.

(20) *Isopat: Calculation of Isotopic Pattern for a given Molecular Formula*; https://cran.r-project.org/web/packages/isopat/index.html, 2012

https://doi.org/10.1021/acs.jproteome.3c00681
J. Proteome Res. XXXX, XXX, XXX−XXX