ProteoCombiner: integrating bottom-up with top-down proteomics data for improved proteoform assessment

Diogo Lima, Mathieu Dupré, Magalie Duchateau, Quentin Giai Gianetto, Martial Rey, Mariette Matondo, Julia Chamot-Rooke

Application Note

ProteoCombiner: integrating bottom-up with top-down proteomics data for improved proteoform assessment

Diogo B Lima[1+], Mathieu Dupré[1], Magalie Duchateau[1], Quentin Giai Gianetto[1,2], Martial Rey[1], Mariette Matondo[1], Julia Chamot-Rooke[1+]

[1]Mass Spectrometry for Biology Unit, Institut Pasteur, CNRS USR 2000, Paris, France

[2]Bioinformatics and Biostatistics HUB, Computational Biology Department, Institut Pasteur, CNRS USR 3756, Paris, France

+ To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

We present a high-performance software integrating shotgun with top-down proteomic data. The tool can deal with multiple experiments and search engines. **Motivation:** Enable rapid and easy visualization, manual validation, and comparison of the identified proteoform sequences including the PTM characterization. **Results:** We demonstrate the effectiveness of our approach on a large-scale *E. coli* dataset; ProteoCombiner unambiguously shortlisted proteoforms among those identified by the multiple search engines.

Availability: ProteoCombiner, a demonstration video and user tutorial are freely available at https://proteocombiner.pasteur.fr, for academic use; all data are thus available from the ProteomeXchange consortium (identifier PXD017618).

Contact: diogobor@gmail.com or julia.chamot-rooke@pasteur.fr **Supplementary information:** Supplementary material is available at *Bioinformatics* online.

1 Introduction

Proteoforms correspond to the different forms of a protein arising from all combinatorial sources of variation from a single gene (including combinations of genetic variation, alternative splicing, and post-translational modifications)(Smith and Kelleher, 2018). Protein characterization at the proteoform level has a crucial importance to fully understand biological processes since specific proteoforms can carry particular biological functions. This includes both the precise determination of protein sequence and identification and localization of Post Translational Modifications (PTMs). Several proteomics strategies can be used to identify proteins, either from their peptides (bottom-up proteomics – BUP) or directly at the intact protein level (top-down proteomics – TDP). Ultimately, combining these approaches results in improved confidence in the identification (Ntai *et al.*, 2016). Although there are tools capable of visualizing TDP data, such as VisioProt-MS, which is able to visualize plots of molecular weights of eluting proteins as a function of their retention time (Locard-Paulet *et al.*, 2019; Lesne *et al.*, 2019), there are very few computational tools that combine data from these different strategies. For instance, Proteome Discoverer enables to perform BUP and TDP database searches independently, but does not provide modules to check the number of identified peptides corresponding to an identified proteoform (Scheffler). Another example is TBNovo (Liu *et al.*, 2014), which allows the integration of data obtained from *de novo* sequencing of proteins both at the peptide and intact protein levels without database search. The limitation in this case, is the assessment of the results, since there is no easy way to improve the confidence of the proteoform identification, based on the identified peptides. Finally, a recent development in the Proteoform Suite enables the large-scale integration of BUP and TDP data to help PTM localization (Schaffer *et al.*, 2020). Here we introduce ProteoCombiner, a software that combines BUP and TDP data from various search engines, *i.e.*, Andromeda (Cox *et al.*, 2011), PatternLab for Proteomics (Carvalho *et al.*, 2015), Comet (Eng *et al.*, 2013), SEQUEST (Eng *et al.*, 1994), ProSightPD, pTop (Sun *et al.*, 2016), TopPIC (Kou *et al.*, 2016), and any other one that export its results to mzIdentML 1.2 format (Vizcaíno *et al.*, 2017). ProteoCombiner allows assessing a wealth of proteomics data to increase confidence in proteoform characterization, including PTMs. Our tool allows a rapid and easy visualization, manual validation and comparison of identified proteoform sequences. The software was programmed in C# with .NET Framework 4.8 and requires a computer with Windows 10 or later, and at least 8 GB of RAM. In what follows, we use ProteoCombiner to

integrate results and simplify the analysis of an *E. coli* protein lysate.

2 Material and methods

To evaluate our software, three biological replicates of an *E. coli* lysate were analyzed using BUP and TDP approaches.

Bottom-up experiments were performed using an Orbitrap Q Exactive™ Plus mass spectrometer with HCD fragmentation, whereas top-down experiments were performed using an Orbitrap Fusion™ Lumos™ mass spectrometer with EThcD fragmentation, as described in Supplementary Material. Raw files were processed using MaxQuant (v. 1.5.3.8) and Patternlab for Proteomics (v. 4.1.1.13) for BUP analyses; and using ProsightPD (v. 2.1) for TDP analyses, with the same database and default parameters for both experiments. Processed results were exported and used as input data for ProteoCombiner.

In ProteoCombiner, peptides that match an identified proteoform can be selected and displayed in a visualization module to easily see sequences identified by each proteomics strategy (Figure 1). A new score (noted *CombScore*) is calculated and allows to rank proteoforms from the best to the worst characterized by combining TDP and BUP. For this, TDP and BUP related scores are first calculated: i) S_{tdp} – a score related to the TDP identification software; and ii) S_{bup} – a score related to the BUP identification software, taking into consideration common and unique peptides that match proteoforms. The *CombScore* is then calculated (between 0 and 1) by applying the Fisher's method to merge S_{tdp} and S_{bup} (Suppl. Material), assuming they are independent of each other (see Figure S2 – Suppl. Material). Finally, each identified proteoform can be assessed by the user using the *CombScore* to evaluate its confidence. The Figure S1 – Suppl. Material proposes an overview of ProteoCombiner principle.

3 Results & Discussion

For bottom-up experiments, MaxQuant and PatternLab for Proteomics led respectively to the identification of 2,443 proteins (39,149 peptides) and 2,922 proteins (25,261 peptides), while for top-down, ProsightPD led to 257 identified proteins (789 proteoforms).

After ProteoCombiner processing, we found 14.4 peptides per proteoform on average. All proteoforms were associated with a *CombScore* expressing the confidence in their identification based on both bottom-up and top-down data. The higher the *CombScore*, the better the confidence in the proteoform identification. For instance, in our results, a proteoform of the HupB protein (P0ACF4), had a normalized ProsightPD score of 0.125 (S_{tdp}), as shown in Figure S3b – Suppl. Material. After the association of 16 unique peptides, the *CombScore* of this proteoform is 0.9545. So, this proteoform, initially identified with low confidence using only Prosight PD (ranked 625th among the 789 identified proteoforms), is now characterized with a much higher confidence after incorporation of BUP data (37th / 789) (Figure S3c).

Moreover, ProteoCombiner is the first software that allows a rapid, automated and easy visualization of a family of identified proteoforms with their associated identified peptides. All proteoforms are annotated with PTMs, sequence coverage, and also monoisotopic and average masses. The type of proteoform that has been identified (expected – described in database, non-expected – not described in database or tagged proteoform – not described in database but contains a gap mass that might represent a PTM) is also clearly indicated (Figure 1) and can be exported to a PDF® file. Finally, an Excel file can be generated with all identifications in order to facilitate data handling, and to eventually create new proteoform databases. We strongly encourage viewing further details and functionalities in our online supplementary video available at https://proteocombiner.pasteur.fr/video. All data are stored on ProteomeXchange consortium via PRIDE (Vizcaíno *et al.*, 2013) repository (data set identifier PXD017618) and are readily available to the scientific community. The remaining files, which include the search results and the protein databases, are available at the project's website (https://proteocombiner.pasteur.fr).

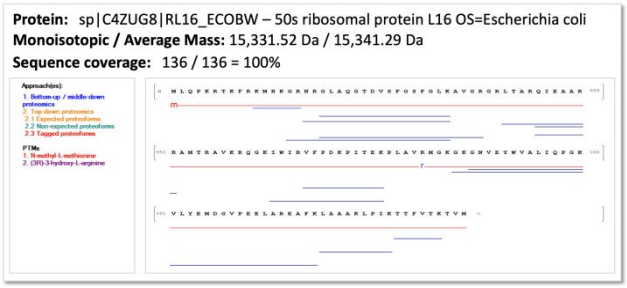

Fig. 1 *A screenshot of the Protein Coverage graphical user interface, showing in the top panel some information about the protein, such as, description, monoisotopic mass, and sequence coverage, as well as, the identified proteoform and the peptides with the respective identified PTM.*

4 Final Remarks

ProteoCombiner can successfully integrate bottom-up datasets to top-down proteomics ones in order to improve proteoform identification. We anticipate that it will play a key role in the analyses of proteomes at the proteoform level, notably for biomarker identification. To facilitate its use, a protocol (Borges Lima *et al.*, 2020) is included in the tool's website.

5 Funding

This work has been supported by the ANR (project ANR-15-CE18-0021). It is also part of the European Joint Programme One Health EJP from the European Union's Horizon 2020 research and innovation programme (Grant Agreement 773830).

Conflict of Interest: none declared.

6 References

Borges Lima,D. *et al.* (2020) Using ProteoCombiner to integrate bottom-up and top-down proteomics data to improve proteoform identification. *Protoc. Exch.*

Carvalho,P.C. *et al.* (2015) Integrated analysis of shotgun proteomic data with PatternLab for proteomics 4.0. *Nat. Protoc.*, **11**, 102–117.

Cox,J. *et al.* (2011) Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J. Proteome Res.*, **10**, 1794–1805.

Eng,J.K. *et al.* (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.*, **5**, 976–989.

Eng,J.K. *et al.* (2013) Comet: an open-source MS/MS sequence database search tool. *Proteomics*, **13**, 22–24.

Improving Proteoform Identifications in Complex Systems Through Integration of Bottom-Up and Top-Down Data | Journal of Proteome Research.

Kou,Q. *et al.* (2016) TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. *Bioinforma. Oxf. Engl.*, **32**, 3495–3497.
Lesne,J. *et al.* (2019) Top-Down and Intact Protein Mass Spectrometry Data Visualization for Proteoform Analysis Using VisioProt-MS. *Bioinforma. Biol. Insights*, **13**, 1177932219868223.
Liu,X. *et al.* (2014) De Novo Protein Sequencing by Combining Top-Down and Bottom-Up Tandem Mass Spectra. *J. Proteome Res.*, **13**, 3241–3248.
Locard-Paulet,M. *et al.* (2019) VisioProt-MS: interactive 2D maps from intact protein mass spectrometry. *Bioinformatics*, **35**, 679–681.
Ntai,I. *et al.* (2016) Integrated Bottom-Up and Top-Down Proteomics of Patient-Derived Breast Tumor Xenografts. *Mol. Cell. Proteomics*, **15**, 45–56.
Schaffer,L.V. *et al.* (2020) Improving Proteoform Identifications in Complex Systems Through Integration of Bottom-Up and Top-Down Data. *J. Proteome Res.*, **19**, 3510–3517.
Scheffler,K. Combination of Bottom-Up and Top-Down Characterization of Biologics Using a High Throughput Capable Workflow in Proteome Discoverer Software. 3.
Smith,L.M. and Kelleher,N.L. (2018) Proteoforms as the next proteomics currency. *Science*, **359**, 1106–1107.
Sun,R.-X. *et al.* (2016) pTop 1.0: A High-Accuracy and High-Efficiency Search Engine for Intact Protein Identification. *Anal. Chem.*, **88**, 3082–3090.
Vizcaino,J.A. *et al.* (2017) The mzIdentML Data Standard Version 1.2, Supporting Advances in Proteome Informatics. *Mol. Cell. Proteomics*, **16**, 1275–1285.
Vizcaino,J.A. *et al.* (2013) The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.*, **41**, D1063–1069.