Research Square Preprint: Please note that this article has not completed peer review.

Using ProteoCombiner to integrate bottom-up and top-down proteomics data to improve proteoform identification

CURRENT STATUS: POSTED

Diogo Borges Lima
Mass Spectrometry for Biology Unit, CNRS USR 2000, Institut Pasteur, Paris, France
diogobor@gmail.com*Corresponding Author*
ORCiD: https://orcid.org/0000-0001-6056-0825

Mathieu Dupré
Mass Spectrometry for Biology Unit, CNRS USR 2000, Institut Pasteur, Paris, France
ORCiD: https://orcid.org/0000-0002-1845-0048

Magalie Duchateau
Mass Spectrometry for Biology Unit, CNRS USR 2000, Institut Pasteur, Paris, France

Quentin Giai Gianetto
Mass Spectrometry for Biology Unit, CNRS USR 2000, Institut Pasteur, Paris, France

Martial Rey
Mass Spectrometry for Biology Unit, CNRS USR 2000, Institut Pasteur, Paris, France

Mariette Matondo
Mass Spectrometry for Biology Unit, CNRS USR 2000, Institut Pasteur, Paris, France

Julia Chamot-Rooke
Mass Spectrometry for Biology Unit, CNRS USR 2000, Institut Pasteur, Paris, France
julia.chamot-rooke@pasteur.fr*Corresponding Author*

DOI:
10.21203/rs.2.10327/v1

SUBJECT AREAS
Computational biology and bioinformatics *Biological techniques*

KEYWORDS
proteomics, bioinformatics, top-down, bottom-up, mass spectrometry

Abstract

Here we present a high-performance software for proteome analysis that combines different mass spectrometric approaches, such as, top-down for intact protein analyses and bottom-up, for proteolytic fragment characterization. **ProteoCombiner** capitalizes on the data arising from different experiments and proteomics search engines and presents the results in a user-friendly manner. Our tool also provides a rapid and easy visualization, manual validation and comparison of the identified proteoform sequences, including post-translational modifications (PTM) characterization. Thus, ProteoCombiner is recommended for studies dealing with different proteomics strategies, in order to increase confidence in proteoform identification including PTMs.

Introduction

Proteoforms describe all combinatorial sources of variation from a single gene, including genetic variations, alternative splicing, and post-translational modifications [1]. Proteoform analysis is nowadays of crucial importance since they have been proven to have a key role in biological systems [2] [3]. This includes both precise determination of the expressed sequence of the protein and the characterization of all Post Translational Modifications (PTM), in particular their precise localization and determination of their combination (combinatorial PTM). Several proteomics strategies, including bottom-up, middle-down and top-down proteomics, have been developed to analyze the PTM at the peptide level or to directly target proteoforms. Nevertheless, there is a lack of bioinformatics tools able to combine such proteomics results and it is still difficult to map the PTMs identified in bottom-up data onto proteoforms obtained by top-down proteomics.

In this protocol, we describe the key steps for using **ProteoCombiner**, a powerful tool capable of integrating proteomics datasets obtained from different experiments (bottom-up and top-down proteomics) and different search engines, such as, PatternLab for Proteomics[4], MaxQuant[5], Comet[6], ProsightPD[7], pTop[8] and TopPIC[9], in order to increase the confidence of proteoform identification and facilitate the characterization of PTMs. A video demonstrating ProteoCombiner v1.0 in action is available at https://proteocombiner.pasteur.fr/video.

Equipment

Hardware

- A computer with at least 8 GB RAM and 2 computing cores is recommended.

Software

- Windows 10 (32 or 64 bits) or later.
- The .NET framework 4.7, which will be automatically updated by ProteoCombiner if necessary.
- **ProteoCombiner** software is available at https://proteocombiner.pasteur.fr.

Data files

- **ProteoCombiner** v1.0 is compatible with output data files from PatternLab for Proteomics[4], MaxQuant[5], Comet[6], ProsightPD[7], pTop[8] and TopPIC[9] and can work directly with Thermo®, ABSciex®, Agilent® and Waters® RAW files. The software is also compatible with mzML 1.1.0, MS2 and Mascot Generic Format (MGF).
- **ProteoCombiner** saves results in its own format (*i.e.*, **.pcmb*) and exports them as Excel® *(*.xlsx*) and PDF® format files.

Procedure

1. Software installation:

Download **ProteoCombiner** by clicking on the *Download* button at https://proteocombiner.pasteur.fr.

2. Workflow

The following workflow demonstrates how to combine proteomics data using **ProteoCombiner**.

2.1. Execute the *ProteoCombiner* tool (**Figure 1**)

[Figure 1]

2.2. Specify the directory containing all result files from bottom-up proteomics experiments. In this directory can also have the database used in the search and the original RAW files in one of these formats: mzML 1.1.0, MS2, Mascot Generic Format (MGF), ABSciex®, Agilent®, Waters® and

Thermo® RAW.

2.2.1. *Bottom-up proteomics*

2.2.1.1. For PatternLab for Proteomics and Comet output files, we recommend to use SEPro[1] tool for filtering results, and use the **.sepr* file(s) as input of ProteoCombiner.

2.2.1.2. For MaxQuant output files, we recommend using the *txt folder*, that contains the following required files: proteinGroups.txt, peptides.txt and msms.txt.

2.2.2. *Database*

2.2.2.1. We recommend using the FASTA and/or XML database formats obtained from Uniprot.

2.3. Specify the directory containing all result files from top-down proteomics experiments. In this directory can also have the database used in the search and the original RAW files in one of these formats: mzML 1.1.0, MS2, Mascot Generic Format (MGF), ABSciex®, Agilent®, Waters® and Thermo® RAW.

2.3.1. *Top-Down proteomics*

2.3.1.1. For ProsightPD results, we recommend exporting only PSM identifications, which contain all information of each proteoform (**_PSM.txt*). This corresponding file must have the following columns: Checked, Confidence, Identifying Node Type, Identifying Node, Search ID, Identifying Node No, PSM Ambiguity, Sequence, Annotated Sequence, Modifications, # Protein Groups, # Proteins, Master Protein Accessions, Protein Accessions, Protein Descriptions, # Missed Cleavages, Charge, Original Precursor Charge, DeltaScore, DeltaCn, Rank, Search Engine Rank, m/z [Da], MH+ [Da], Theo. MH+ [Da], DeltaM [ppm], Deltam/z [Da], Matched Ions, Total Ions, Intensity, Activation Type, MS Order, Isolation Interference [%], Ion Inject Time [ms], RT [min], First Scan, Last Scan, Master Scan(s), Spectrum File, Ions Matched, Annotation, -Log P-Score, -Log E-Score, C Score, Corrected Delta Mass (Da), Corrected Delta Mass (ppm).

2.3.1.2. For pTop output files, we recommend using only the file(s) ending with *_filter.csv*.

2.3.1.3. For TopPIC output files, we recommend using only the file(s) ending with *_prsm.csv*.

2.3.2. *Database*

2.3.2.1. We recommend using the FASTA and/or XML database formats obtained from Uniprot.

2.4. The *Parameters* tab allows to access various parameters that are not usually required to be changed for combining all data.

2.4.1. *Remove Contaminants*: This option allows to remove all protein sequences that represent a contaminant.

2.4.2. *Remove Reverse Sequence*: This option allows to remove all decoy protein sequences.

2.5. To start combining data, click on the *OK* button in the *Combine* tab.

PS: Although ProteoCombiner capitalizes on the data arising from different proteomics search engines, we recommend using a single software-tool to analyze all experiments for proteolytic fragment characterization, and another to analyze intact proteins experiments.

3. Exploring the results

Note: At this point we recommend saving results by selecting *Save* from File menu or by pressing CTRL + S.

3.1. *Filter*

3.1.1. All results are pre-filtered according to the following parameters designated on the top of the *Results Browser* window, as shown in **Figure 2**:

3.1.1.1. *CombScore*: Results containing identification scores greater than or equal to this value will be displayed.

3.1.1.2. *Peptide Count*: Only identified proteins containing at least this value as identified peptide amount (by bottom-up) will be displayed.

3.1.1.3. *Spectral Count*: Only identified proteins containing at least this value as identified spectra amount will be displayed.

3.1.1.4. *Unique peptide*: Only identified proteins containing at least this value as identified unique peptide amount will be displayed.

3.1.1.5. *Search*: Only results from peptides or proteins containing the sequence input to this field will be displayed. The user can further search by ProteinID (protein accession number), protein description or file name.

3.1.1.6. By clicking on *Filter* button, this operation is accomplished using all of the parameters described above.

3.1.1.7. By clicking on *Reset* button, all initial values are restored.

[Figure 2]

3.2. *Combined results*

All identified proteins that contain a valid sequence* will be displayed on this tab sorted by *Score* followed by *Sequence Coverage*. The protein score is represented by the best proteoform score.

**Protein sequence present in the database.*

3.2.1. By clicking on an identified protein, all respective identified proteoforms will be displayed* below the protein table sorted by *CombScore***.

**If there is no identified proteoform for a respective protein, all possible identified peptides will be displayed instead.*

***CombScore is calculated by summing two different scores: i) a score related to the TDP identification software (which is normalized between 0 and 1; and ii) the percentage of the proteoform sequence coverage based on the peptides, obtained in BUP approach, that match to this proteoform. This score also ranges between 0 and 1.*

3.2.1.1. The user can assess each identified proteoform by clicking on the *Is valid* checkbox.

3.2.1.1.1. At this point, we recommend saving results once again so that the personal assessments

can be included. This is done by selecting *Save* from *File* menu or by pressing CTRL + S.

3.2.1.2. By clicking on an identified proteoform, all respective identified peptides will be displayed below the proteoform table sorted by *Peptide Score*.

3.2.1.2.1 By double-clicking on an identified peptide, the tandem mass spectrum, which contains the best identification score, will be displayed on *Spectrum Viewer*. (**Figure 3**)

[Figure 3]

3.2.2. By double-clicking on an identified protein, a new window will be opened that shows the *Protein Coverage* (**section 3.5**).

3.3. *Bottom-up proteomics results*

All identified proteins and the corresponding peptides will be displayed on this tab.

3.3.1. By clicking on an identified protein, all respective identified peptides will be displayed below the protein table.

3.3.1.1. By clicking on an identified peptide, all respective identified tandem mass spectra will be displayed below the peptide table.

3.3.1.1.1. By double-clicking on an identified tandem mass spectrum, the *Spectrum Viewer* will be opened. (**Figure 3**)

3.3.2. By double-clicking on an identified protein, a new window will be opened that shows the *Protein Coverage* (**section 3.5**).

3.4. *Top-down proteomics results*

All identified proteoforms will be displayed on this tab grouped by *Theoretical Mass* in Da.

3.4.1. By double-clicking on the *Scan Number* column of an identified proteoform, the *Spectrum Viewer* will be opened. (**Figure 3**). By clicking on any other column, a new window will be opened that shows the *Protein Coverage* (**section 3.5**).

3.5. *Protein Coverage*

3.5.1. Once the window is opened, the information will be displayed on the top tab: Protein description, Monoisotopic and Average protein mass* and Sequence coverage. (**Figure 4**)

**If Protein coverage window is opened from the click of a specific proteoform, its monoisotopic and average mass will be displayed instead of the protein mass.*

[Figure 4]

3.5.2. The left box will display all approaches used to identify the proteoforms and/or peptides. By clicking on each item, all respective lines will be high-lighted in the right box.

3.5.2.1. Bottom-up and middle-down approaches, represented by the blue color,

3.5.2.2. Top-down approach is represented in three different colors: Expected proteoforms in orange, which means all identified proteoforms by the full theoretical mass; non-expected proteoforms in cyan, which represents all identified truncated proteoforms; and tagged proteoforms in red, which means all proteoforms that were identified by a part of the protein sequence.

3.5.2.3. All identified PTMs will also be displayed in this box.

3.5.3. The right box will display all identified proteoforms and/or peptides that will be represented by different lines. All of them will be displayed sorted by *CombScore*. On the top will be displayed the full protein sequence and all theoretical modifications (present in the database) as can be seen in **Figure 4**. All theoretical chains will be shown below protein sequence in gray dash lines. The user is able to check all information about the modification or theoretical chain by passing the mouse over the line or the modified amino acid. (**Figure 4**)

3.5.3.1. Each proteoform will be displayed according to the classification: expected, non-expected or tagged proteoform. By right-clicking on each line, the user is able to assess the proteoform identification (valid or invalid); in addition, it’s possible to highlight only the peptides that fit into the proteoform. (**Figure 5**). By hovering over each line, some information will be shown: Proteoform sequence, Score, Search engine that identified this specific sequence, Start and End positions.

3.5.3.2. All identified PTMs will also be displayed and by hovering over each one, it is possible to check their position and description.

[Figure 5]

3.5.3.3. The identified proteoforms and/or peptides can be displayed in a single line or in multiple lines. The user can change the way to visualize in *Utils* menu, by selecting *Custom Protein Visualization* option in *Results Browser* window (or by pressing ALT + C).

3.6. *Loading results*

3.6.1. ***ProteoCombiner*** loads results in its own format (*.*pcmb*). This can be accomplished in three ways, the easiest one is by double-clicking on a ProteoCombiner results file. If the *Results Browser* window is opened, another way to launch the file is by clicking on *Load* option from the *File* menu or by pressing CTRL + O, as seen in (**Figure 6**). Otherwise, if the main window is opened, select *Load Results* from *File* menu (or press CTRL + O), as seen in **Figure 7**.

[Figure 6]

[Figure 7]

3.7. *Exporting results*

3.7.1. ***ProteoCombiner*** also allows to export all combined results to Excel® (*.*xlsx*) or PDF® file. This is done by selecting *Excel file* from the File menu Þ Export results (or by pressing ALT + E). Or by selecting *PDF file* from the File menu Þ Export results (or by pressing ALT + P). (**Figure 6**)

References

[1] L. M. Smith and N. L. Kelleher, “Proteoforms as the next proteomics currency,” *Science*, vol. 359, no. 6380, pp. 1106–1107, Mar. 2018.

[2] Z. R. Gregorich and Y. Ge, “Top-down proteomics in health and disease: Challenges and opportunities,” *PROTEOMICS*, vol. 14, no. 10, pp. 1195–1210, 2014.

[3] J. Chamot-Rooke *et al.*, “Posttranslational Modification of Pili upon Cell Contact Triggers N.

meningitidis Dissemination," *Science*, vol. 331, no. 6018, pp. 778–782, Feb. 2011.

[4] P. C. Carvalho *et al.*, "Integrated analysis of shotgun proteomic data with PatternLab for proteomics 4.0," *Nat. Protoc.*, vol. 11, no. 1, pp. 102–117, Dec. 2015.

[5] J. Cox, N. Neuhauser, A. Michalski, R. A. Scheltema, J. V. Olsen, and M. Mann, "Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment," *J. Proteome Res.*, vol. 10, no. 4, pp. 1794–1805, Apr. 2011.

[6] J. K. Eng, T. A. Jahan, and M. R. Hoopmann, "Comet: an open-source MS/MS sequence database search tool," *Proteomics*, vol. 13, no. 1, pp. 22–24, Jan. 2013.

[7] "ProSightPD 1.1: Integrating Top Down Searching in Proteome Discoverer 2.1." .

[8] R.-X. Sun *et al.*, "pTop 1.0: A High-Accuracy and High-Efficiency Search Engine for Intact Protein Identification," *Anal. Chem.*, vol. 88, no. 6, pp. 3082–3090, Mar. 2016.

[9] Q. Kou, L. Xun, and X. Liu, "TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization," *Bioinforma. Oxf. Engl.*, vol. 32, no. 22, pp. 3495–3497, Nov. 2016.

Acknowledgements

The authors thank the Agence Nationale de la Recherche (project ANR-15-CE18-0021) for financial support. This work 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).

Figures

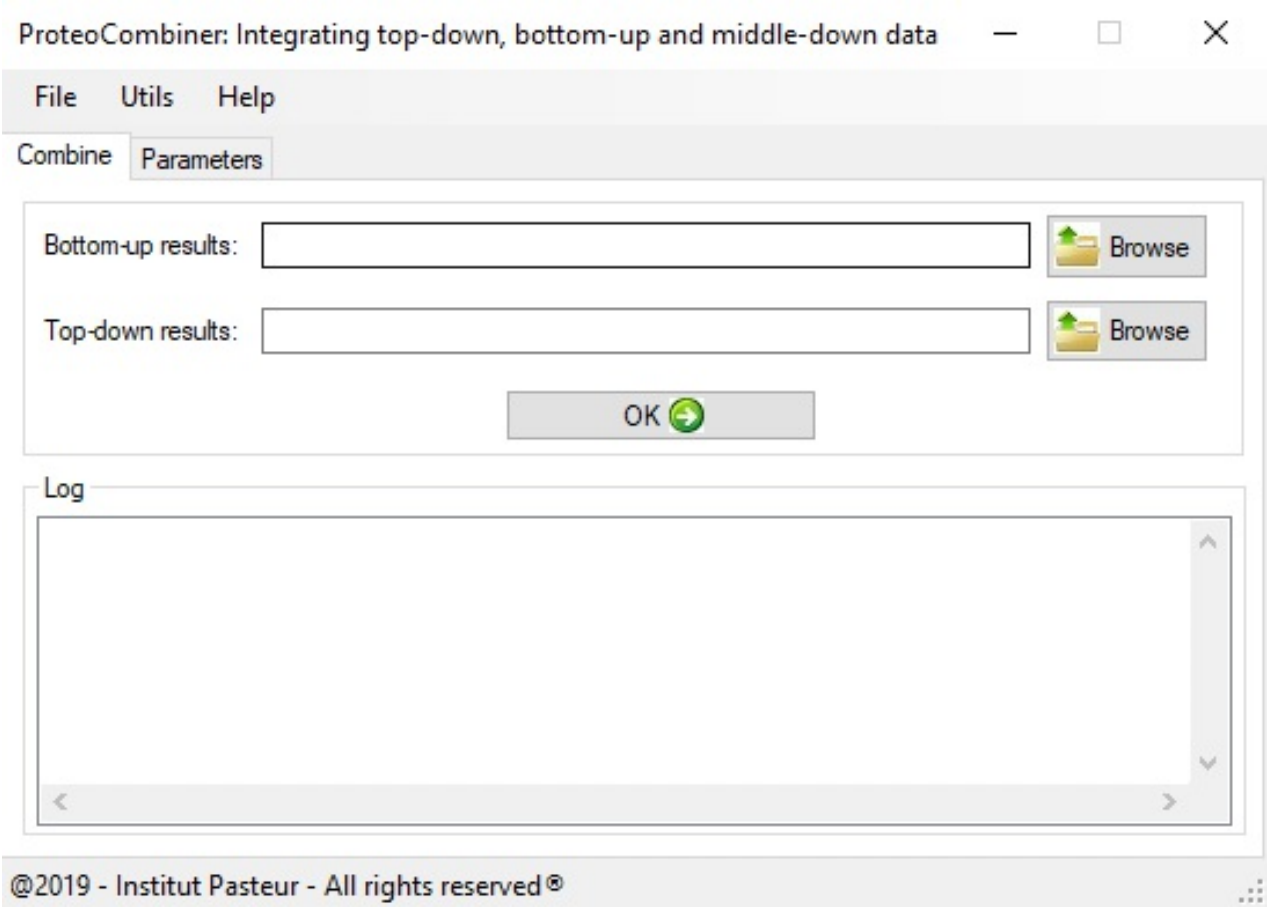

Figure 1

Graphical User Interface of the main window of ProteoCombiner.

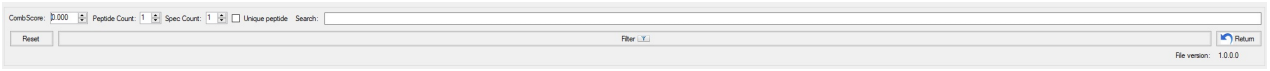

Figure 2

Filter parameters responsible for selecting results.

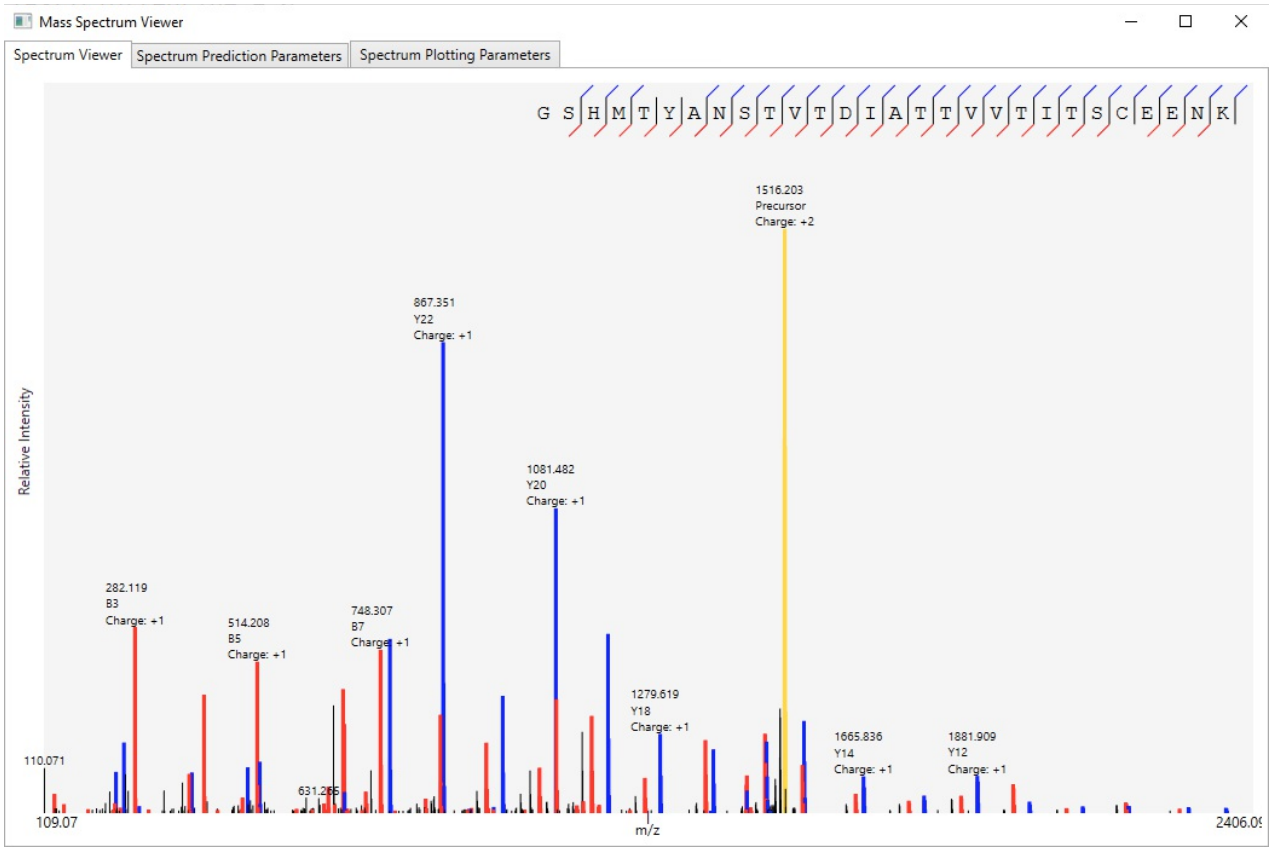

Figure 3

Spectrum viewer – Blue lines represent all y or x or z peaks matched, whereas red lines represent all a or b or c peaks matched. Yellow lines represent the precursor ion.

Figure 4

Protein coverage window: All identified proteoforms and/or peptides are displayed here in different lines. The theoretical sequences (proteoforms) are displayed in gray dash lines and the theoretical PTMs above protein sequence. The information about the PTM and theoretical sequences are shown when the mouse is hovered over them.

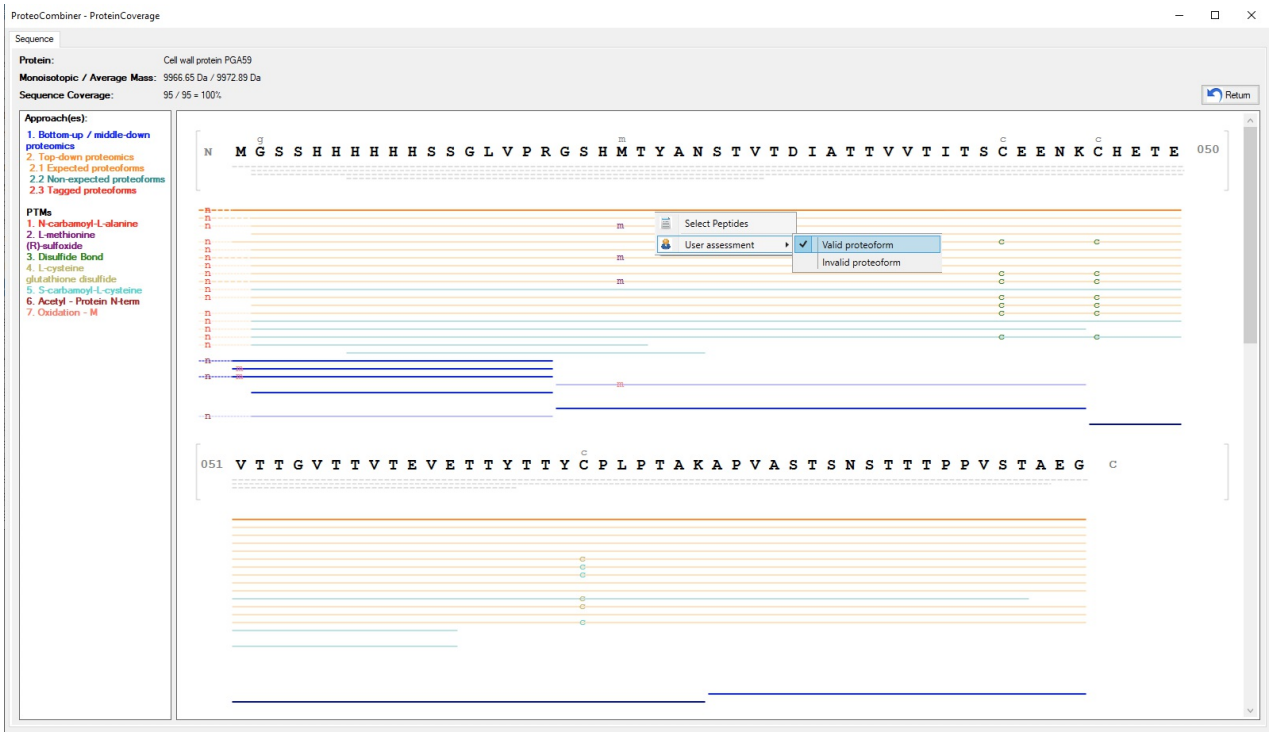

Figure 5

An assessment can be applied for each identified proteoform in order to increase the confidence of the evaluation. The user can also highlight only peptides that fit with a specific proteoform.

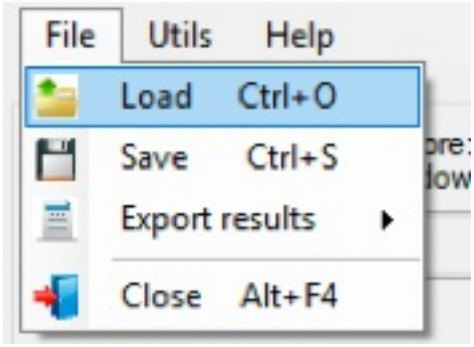

Figure 6

Save or Load ProteoCombiner results from Results Browser window. The user can also export the results as an Excel® or PDF format file.

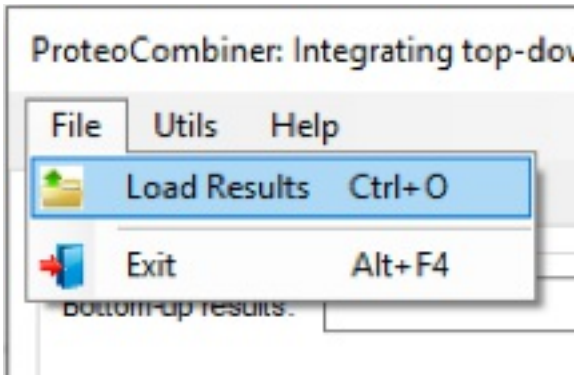

Figure 7

Load ProteoCombiner results from main interface.