



**HAL**  
open science

## Validation of MS/MS identifications and label-free quantification using Proline

Véronique Dupierris, Anne-Marie Hesse, Jean-Philippe Menetrey, David Bouyssié, Thomas Burger, Yohann Couté, Christophe Bruley

► **To cite this version:**

Véronique Dupierris, Anne-Marie Hesse, Jean-Philippe Menetrey, David Bouyssié, Thomas Burger, et al.. Validation of MS/MS identifications and label-free quantification using Proline. *Statistical Analysis of Proteomic Data*, 2426, Springer US, pp.67-89, 2023, *Methods in Molecular Biology*, 10.1007/978-1-0716-1967-4\_4 . hal-03244240

**HAL Id: hal-03244240**

**<https://hal.science/hal-03244240>**

Submitted on 1 Jun 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Validation of MS/MS identifications and label-free quantification using Proline

Véronique Dupierris<sup>1</sup>, Anne-Marie Hesse<sup>1</sup>, Jean-Philippe Menetrey<sup>2</sup>, David Bouyssie<sup>3</sup>,  
Thomas Burger<sup>2</sup>, Yohann Couté<sup>1</sup> and Christophe Bruley<sup>1,†</sup>

<sup>1</sup> Univ. Grenoble Alpes, INSERM, CEA, UMR BioSanté U1292, CNRS FR2048, 38000, Grenoble, France

<sup>2</sup> Univ. Grenoble Alpes, CNRS, INSERM, CEA, FR2048 38000, Grenoble, France

<sup>3</sup> Institut de Pharmacologie et de Biologie Structurale (IPBS), Université de Toulouse, CNRS, UPS, Toulouse, France

† [christophe.bruley@cea.fr](mailto:christophe.bruley@cea.fr)

## Abstract

In the proteomics field, the production and publication of reliable mass spectrometry (MS)-based label-free quantitative results is a major concern. Due to the intrinsic complexity of bottom-up proteomics experiments (requiring aggregation of data relating to both precursor and fragment peptide ions into protein information, and matching this data across samples), inaccuracies and errors can occur throughout the data-processing pipeline. In a classical label-free quantification workflow, the validation of identification results is critical since errors made at this first stage of the workflow may have an impact on the following steps and therefore on the final result. Although false discovery rate (FDR) of the identification is usually controlled by using the popular target-decoy method, it has been demonstrated that this method can sometimes lead to inaccurate FDR estimates. This protocol shows how Proline can be used to validate identification results by using the method based on the Benjamini-Hochberg procedure and then quantify the identified ions and proteins in a single software environment providing data curation capabilities and computational efficiency.

**Key words** Mass spectrometry software, Discovery proteomics, Benjamini-Hochberg False Discovery Rate, Statistics, Label-free quantification, Data visualization, Software engineering

## 1 Introduction

High-throughput mass spectrometry-based proteomics is continuously evolving toward increased complexity through the analysis of larger sample cohorts, with more sophisticated experimental designs, and deeper proteome coverage. This complexity has dramatically increased the data volumes to be processed, so that the processing efficiency has become a major concern for many labs or core facilities. Another effect of this complexity is that the examination of the data and the ability to review and correct the data processing steps requires a solution allowing experimentalists to visualize and navigate a considerable amount of data in a very effective way. To address this, we developed a software suite called Proline [1] which provides an efficient and integrated way to process, visualize and publish proteomic datasets.

The validation process of tandem mass spectrometry (MS/MS) identification in Proline was originally based on predefined filters used to accept or reject a peptide-spectrum match (PSM) and on the widely used target-decoy competition (TDC) method [2] to control the false discovery rate (FDR). We recently proposed an alternative to TDC validation by a totally different method to control the FDR at the PSM, peptide, and protein levels, while benefiting from the theoretical guarantees of the Benjamini-Hochberg (BH) framework [3]. Since version 2.1.2, the BH FDR method is available in the official release cycle of Proline, in addition to the target-decoy competition based method. The BH method provides a simple and pragmatical way to validate identification results without requiring the use of decoy protein databases. Following validation, Proline can then be used to quantify peptides and proteins as shown in this protocol.

## 2 Material

### 2.1 Requirements

Proline is a client-server application that can either be installed on a machine running Linux, Windows or MacOS. An all-in-one package called Proline-Zero, including both the client and the server parts as well as the required dependencies is available. This package does not require any complicated installation procedure (*see*

Subheading 2.3). In that particular case, the amount of RAM available is the only requirement: a minimum of 8GB of RAM is recommended for the server especially to be able to run quantification processes; while 1GB of RAM is enough for the client (see **Note 1**). This protocol is based on Proline version 2.1.2 installed from the Proline-Zero distribution.

For more intensive usage, the server part of Proline can be installed on a centralized server allowing multiple clients to connect to the same Proline server. In this deployment scheme, Java SE 8 Runtime Environment is required for client and server parts and a PostgreSQL database server (versions above 9.4) must be installed and configured on server side.

## 2.2 Data format

Identification results from different search engines can be imported into Proline in their native data format. This includes .dat files from Mascot, .omx files from OMSSA, .xml files from X!Tandem and folders of .txt files from Maxquant. In addition, Proline supports mzIdentML files, enabling to import results from any search engines that support this standard format.

## 2.3 Software install

To install Proline from the Proline-Zero distribution, select the correct archive file for your operating system from the Proline website (<http://www.profiroteomics.fr/proline/#downloads>) and unarchive it. On first launch, the different components are automatically initialized and configured, including the database schema that will be used by Proline.

## 2.4 Sample data

This protocol focuses on validation of identification results based on the BH method to control the FDR. To do so, input data must be MS/MS identification searches performed by Mascot search engine on databases containing only target proteins: no decoy is needed using this strategy. Identification data from [3] are used throughout this protocol to illustrate the workflow. This dataset (.dat and .raw files) is available on the proteomeXchange repository [4] with the identifier PXD016669. For the sake of simplicity, the subset of data needed for this protocol as well as the Proline-Zero distribution version 2.1.2 are available at <ftp://ftp.cea.fr/pub/edyp/Proline/MiMB/>. This FTP folder contains:


1. Replicates 1 to 4 of the 10 replicates of samples analyzed with a Q-Exactive Plus instrument with HH settings and searched against a database containing only target proteins will be used (files QEx\_HH\_no\_decoy\_R1 .dat to QEx\_HH\_no\_decoy\_R4 .dat).
2. The Thermo .raw files corresponding to the MS analysis of these four replicates (QEx2\_020296.raw, QEx2\_020300.raw, QEx2\_020322.raw and QEx2\_020419.raw) that will be used in the second part of this protocol.
3. The protein sequence database (.fasta file) that has been used to perform the MS/MS search.

# 3 Methods

## 3.1 Starting and configuration

1. To start Proline, double-click Proline-Zero.exe on Windows or type Proline-Zero.sh on Linux-like OS. The different components are started by the launcher. Afterwards, the graphical user interface process (named ProlineStudio) starts.
2. At the first start, the default configuration is initialized and a default user as well as a project are created.
3. The connection window appears. Fill the requested fields with the default authentication information (host: localhost and user/password: proline/proline).

The initial configuration creates a default user and a first project but it also generates some predefined descriptions such as instrument configurations or software used to generate the peaklist submitted to the search engine. Once connected, multiple projects can be created and some of these definitions can be modified as described below:



4. To create a new project, click on the  icon in the left panel. In the creation dialog, indicate the name and the description of the project. Click .
5. Click on the   menu. A dialog appears showing four tabs allowing to: create new users; create a new peaklist software definition; view all existing projects with their associated data; and add new fragmentation rule sets (see **Note 2**).

By default all data used and generated by Proline-Zero are located in a sub-folder of the installation folder named `data`. For security reasons, the server side is only allowed to browse the content of a restricted list of folders that can however be manually modified by editing the server configuration. By default the folders added to this list are sub-folders of `<proline-install>/data`. Files must be manually copied into the following folders to be accessible to the server:





6. Copy the Mascot .dat files into the `<proline-install>/data/mascot` sub-folder.
7. Copy the fasta file into `<proline-install>/data/fast` sub-folder.
8. Copy the Thermo .raw files into `<proline-install>/data/mzdb` sub-folder.


## 3.2 Import search results

Importing a search engine results consists in parsing the search results file to extract information and meta-information and store them into the Proline databases. Neither filtering nor thresholding is applied at this stage: all submitted spectra, peptide-spectrum matches and protein hits are stored in the database, enabling the subsequent validation of putative identifications. As Mascot is a prerequisite to validate identification results using the BH method, this protocol focuses on Mascot identifications results. However, Proline also supports other search engines (*see Note 3*) that can be validated through the classical target-decoy approach.

1. Right-click on the  **Identifications** node in the upper left panel and select `Import Search Result`.
2. The import dialog appears (*see Figure 1*).
3. Select the files to import by clicking on the  file button.
4. Set the different options:
  - (a) *Software engine* `Mascot`: the search engine used to generate the file to import.
  - (b) *Instrument* `Q EXACTIVE (A1=FTMS F=HCD A2=FTMS)`: the mass spectrometer used to perform the MS/MS analysis.
  - (c) *Fragmentation Rule Set* `ESI-TRAP`: The fragmentation rules specified in the search engine. The button on the right can be used to visualize all rules of a specific rule set and new rules can be created through the admin dialog (*see Subheading 3.1, Step 5*). This information is required to generate theoretical ion fragments and then annotated MS/MS spectra.
  - (d) *Peaklist software* `Mascot Distiller`: the software used to generate the peaklist submitted to the search engine. This information is used during the optional quantification step to extract scan number or retention time from the MS/MS spectrum titles. Additional software can be configured in the admin dialog (*see Subheading 3.1, Step 5*).
  - (e) *Decoy* `No Decoy`: the target-decoy strategy used during the search: `No Decoy` if the search was performed using a database containing only target protein sequences, `Concatenated Decoy` if the search was performed against a database containing both target and decoy proteins or `Software Engine Decoy` if the decoy search is performed on the fly by the search engine.

If a wrong fragmentation rule set or a wrong peaklist software is selected at this step, they can be modified afterward.

5. Click `OK`.
6. Processes are all executed on the server side of Proline. The graphical user interface communicates with the server by submitting the tasks to be performed and modifies the user interface status while waiting for the tasks completion. The submitted and running tasks can be displayed in the `logs` tab: submitted tasks are represented by an hourglass icon () , running tasks by a blue arrow () , successfully completed tasks by a green tick () or a red cross () when a task failed.
7. Wait for the task completion.

Once imported, the search result appears in the upper left panel under the `Identification` node. User's actions on these results are reachable by right-clicking on the icon  representing them in the left panel. By default a search result is designated by the name of the imported file but can be renamed.

## 3.3 Combine identification results

Identification results of different MS/MS spectrum searches, for example from replicates or sample fractionation, can be combined or merged to generate a non-redundant list of identified peptides and proteins. This merge can be performed before (on search results) or after validation (on identification summaries) and can be applied recursively, leading in a hierarchical organization of the datasets.

If applied before validation, all PSMs identified by the search engine are taken into account and are combined to create new PSMs into the aggregated dataset. This is useful to combine analytical replicates or fractions since peptides belonging to a single protein could be spread across different fractions. Then the created dataset can be validated as if it were a single search engine identification result.

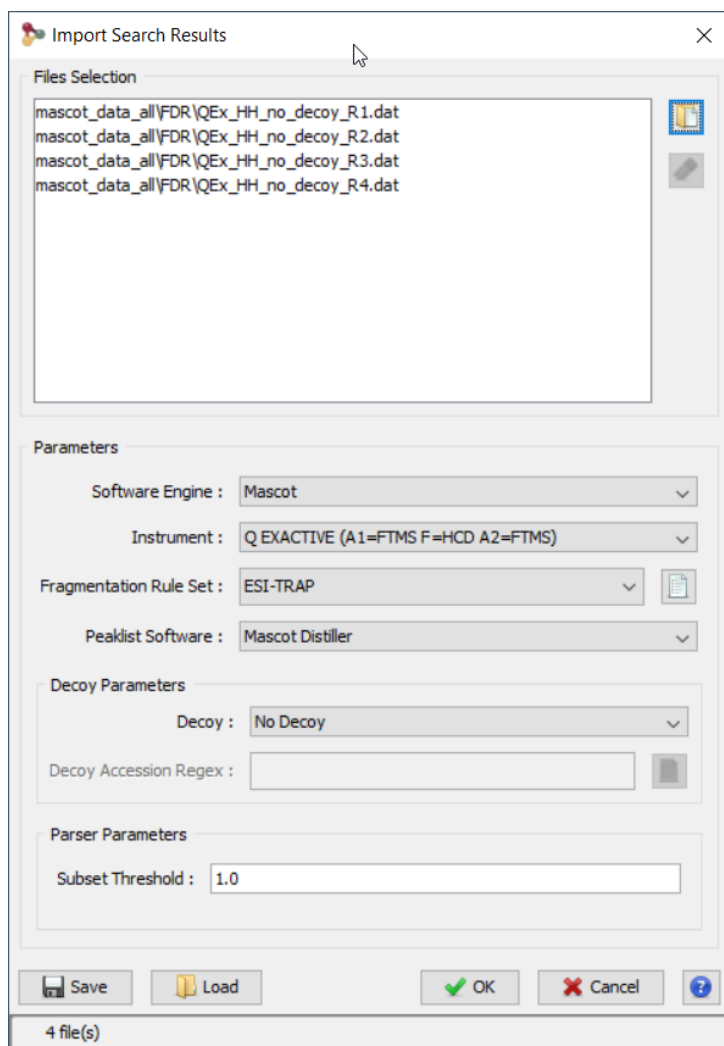





Figure 1: Import dialog.

1. To create a new dataset, **Right-click** on the **Identification** node and select **Add Dataset**.
2. Name the new dataset.
3. Each imported file can be reused with no need to import it again. The **All Imported** node allow to retrieve the already imported files. **Right-click** on **All Imported** and then **Display List** (or double click on **All Imported**). Select the search results to be merged, drag and drop them in the newly created dataset. The search results are now hierarchically organized: the newly created dataset is the parent of the hierarchy and the dragged and dropped search results are the children.
4. To merge the search results, **right-click** on the parent dataset and select **Merge Datasets** > **Union** and wait for the completion of the merge task (*see Note 4*).
5. Once the merge is completed, the icon representing the dataset changes to give a feedback of the type of combination the dataset is made of. While  represents an empty dataset,  is used for an imported result set or a dataset just merged from non validated result sets. The right part of the icon indicates properties before validation whereas the left part indicates properties after validation. A letter is added in the blue half circle to indicate the type of combination that has been made: U for **Union** and A for **Aggregate** (*see Note 4*). In our case, since the merge is done by union and before validation, the icon changes to .

### 3.4 Validate identification results

Validation of a search result can be performed at PSM, peptide and protein levels independently. At each level, a set of predefined filters can be combined and applied to accept or reject PSMs or proteins based on a user specified threshold value applied to various properties such as rank, peptide length, minimum score for PSM and minimum score or number of peptides for protein sets. In addition to these filters, a procedure can be applied to limit

Table 1: PSM filters.

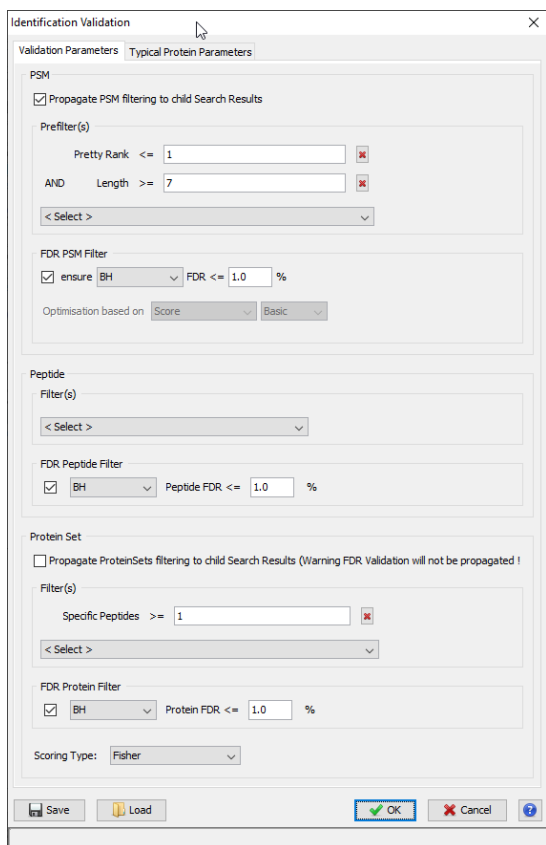
Filter	Description
Minimum peptide length	PSMs corresponding to peptide sequences shorter than the cut-off stipulated will be discarded when this parameter is applied.
Pretty rank	This filter is applied after having temporarily joined target and decoy PSMs corresponding to the same query. For each query, target/decoy PSMs are then sorted by score. As in Mascot, a pretty rank is computed for each PSM depending on their ranking: PSM with almost equal score (difference < 0.1) are assigned the same rank. All PSMs with a pretty rank greater than the cut-off specified are discarded.
Single PSM per rank	This filter selects only one PSM per pretty rank, which is already the case when a given pretty rank is associated with a single PSM. When multiple PSMs have the same pretty rank, Proline retains the peptide associated with the protein that has the highest number of MS/MS events. Thus, if this filter is combined with the “Pretty rank” filter, the result obtained should be identical to the result of the “Single PSM per MS query” filter.

the false discovery proportion. This validation step can be done by using the popular target-decoy competition method [2] (see **Note 5**) or by using a method based on the Benjamini-Hochberg procedure, as proposed in [3].

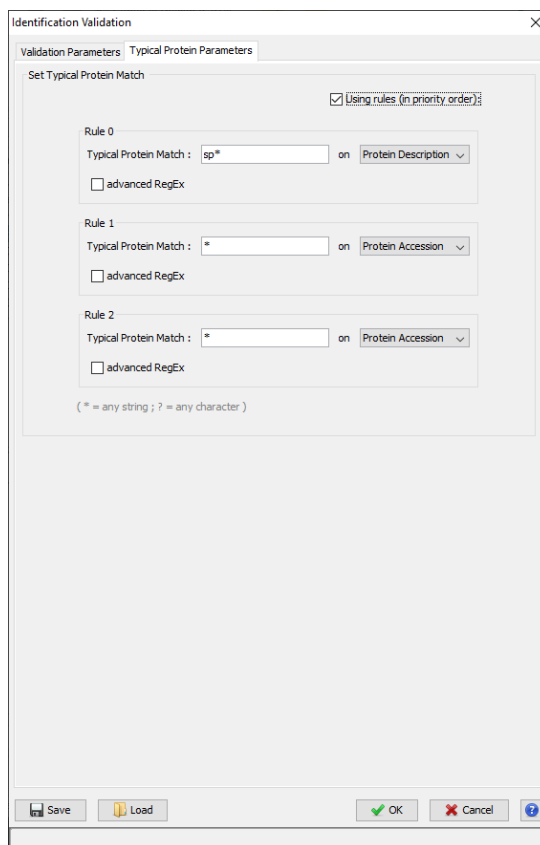
1.  on the node representing the merged search results and select .
2. In the validation dialog (see Figure 2), add filters to accept only PSM of rank 1 and length  $\geq 7$ . Ensure that only one PSM is accepted per MS/MS spectrum (also named a query) by adding a **Single PSM per Query** filter (details of these filters are described in Table 1).
3. Select the  procedure to control the PSM FDR and set the expected FDR value to 1%.
4. Select the  procedure to control the Peptide FDR and set the expected FDR value to 1%.
5. Add a filter to invalidate protein set identified without at least one specific peptide (Specific peptides correspond to peptides identifying a unique protein set at the dataset level).
6. Select the  procedure to control the Protein FDR and set the expected FDR value to 1%.
7. Choose  as Scoring Type that relies on Fisher’s test to calculate protein scores and p-values that will be used by the BH procedure (see **Note 6**).
8. Check the **Propagate PSM filtering** option. When validating a merged dataset, the applied filters as well as the threshold determined by the FDR control method can be propagated to the child datasets of the hierarchy that have been merged. Doing this, the global FDR at the top level of the hierarchy is controlled, but the user can still explore the individual validated search results, relying on the same validation criteria.
9. Move on to the **Typical Protein Parameters** tab. In the **Rule 0** panel, fill the **Typical protein match** text field with `sp*` on **Protein Description**. A set of identified peptides can match to multiple proteins that cannot be distinguished [5]. This depends on the sample but also on the redundancy of the protein database used for the identification. The typical protein is the protein selected to be the representative of the identified group. The selection can be customized by specifying selection rules. In that particular example, the selection tries to select proteins from the SwissProt database instead of Trembl one if any (see **Note 7**).
10. Click  and wait for the completion of the validation task. Once validated, the left half-circle is highlighted in orange (🟡) (see **Note 8**).
11. Protein sequences are usually absent from the search engine identification result files. However, Proline can retrieve amino acid sequences from the fasta file (see **Note 9**). Once validation is completed  on the parent dataset and select  (see **Note 10**).

### 3.5 Navigate through identification summaries

MS/MS identification results before and after validation can be selected to display their content. Before validation, the navigation path through the data can start from submitted MS/MS spectrum, PSMs or proteins. After validation, the validated content can alternatively be displayed and the navigation then starts from PSMs, peptides or protein sets. In any case, the selected data are displayed in views composed of one or more tables. The



(a) Validation parameters tab.



(b) Change typical protein tab.

Figure 2: Validation dialog.

navigation principle is consistent across the graphical interface: the first table contains the data to start from, while the content of the following tables of the view is induced by the selected row in the preceding table. Each table can be searched, filtered and sorted and the column visibility can be modified to focus on the information of interest.

1. Right-click on the parent dataset node, select **Display >> Identification Summary >> Protein sets**. The same view as represented in Figure 3 appears in the right panel of the main window. The first table shows the identified protein sets of the identification summary. Each row is a protein set represented by one of its protein. The menu bar on the left allows to search for a specific value in the table (🔍), to filter rows (🔼), to modify the display (column visibility, order and width) (🔧) or to export the content of the table to a CSV or XLS file (📄) (see Note 11).
2. Select or search for the FDOG\_ECOLI protein set in the list of identified protein sets in the first table: Click on the search button, indicate FDOG\* in the search popup panel.
3. The second and third tables of the view show the content of the selected protein set: the different proteins identified by the same set of peptides (or by a subset of them) is represented in the second table, while the set of peptides identifying the proteins is displayed in the third table (see Figure 3). If a protein is selected in the second table, for example a protein set marked as a subset by the icon (📌), the set of peptide displayed in the third table is restricted to peptides identifying the selected protein.
4. The third table shows all peptides identifying the selected protein. In addition to the sequence and the post-translational modifications associated to each peptide, information related to the highest scoring PSM identifying the peptide are shown. Click on a peptide in the third table to change the selection.
5. The protein sequence retrieved by Proline from the fasta file (see Subheading 3.4, Step 11) is displayed in the last panel at the bottom of the view. Amino acids matched by an identified peptide are represented on a gray background and the selected peptide is highlighted in blue. Underneath the amino-acid sequence, a graphical display of the sequence layed out on a single line shows the position of identified peptides on the sequence.
6. Move on to the Spectrum tab representing the MS/MS spectrum submitted to the search engine. Click on

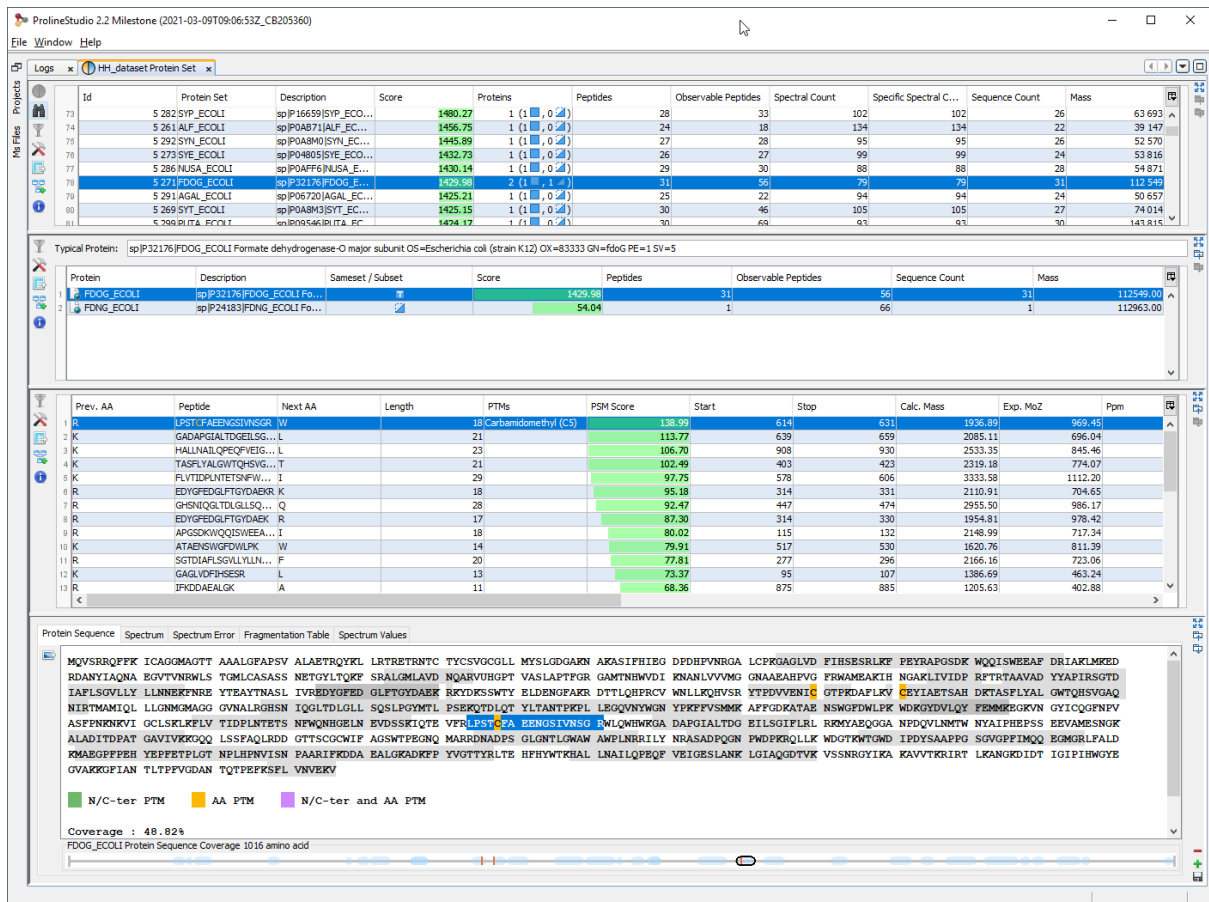



Figure 3: Navigation through identification summary data.

the Generate and Store Spectrum Match icon () on the left. In the dialog, select ESI-TRAP as the fragmentation rule set and click **Ok**. The fragments matching the theoretical spectrum generated using the specified fragmentation rules are displayed over the MS/MS spectrum (see Figure 4). Move on to the Fragmentation table tab to display the fragment matches in a tabular view (see Note 12).

7. Proline systematically stores and keeps track of metadata from processing steps, used parameters and generated data. These metadata are available in different views and export outputs. Metadata from multiple dataset can be easily displayed, allowing to compare them (see Figure 5). Select the four replicates in the left panel (left-click) on the first dataset, hold the **Shift** key and then left-click on the last replicate) then right-click and select **Properties**. In the **Properties** view, each row is a property or a metadata (grouped by theme) and each column represents a dataset. In the second column (referred to as Type) the property name is displayed on with an orange-light color if the values in the row are different. As an example, in the Search properties group, the Result File Name is different in each dataset but the other search parameters are all identical.

### 3.6 MS1 label-free quantification

Once validated, identification summaries can be used for label-free quantification, either using spectral counting, or after detection of chromatographic peaks from MS1 signals. The different steps of the quantification process, the algorithms implemented in Proline and their parameters have been extensively described in [1]. In this protocol, the focus is made on how to start a MS1 quantification from an identification summary.

1. The signal extraction is based on the mzDB format [6]. Raw files must be first converted by using the conversion tool which is embedded in Proline-Zero (see Note 13). In the left panel move on the MS Files tab. Browse the local folders to the Proline installation path, then into the <proline-install>/data/mzdb sub-folder. Select the .raw files than Right-click and select **Convert to mzDB**.
2. In the conversion dialog, indicate the path to the converter: <proline-install>/raw2mzDB\_0.9.10\_build20170802/raw2mzDB.exe. Then choose as output path the <proline-install>/data/mzdb folder.



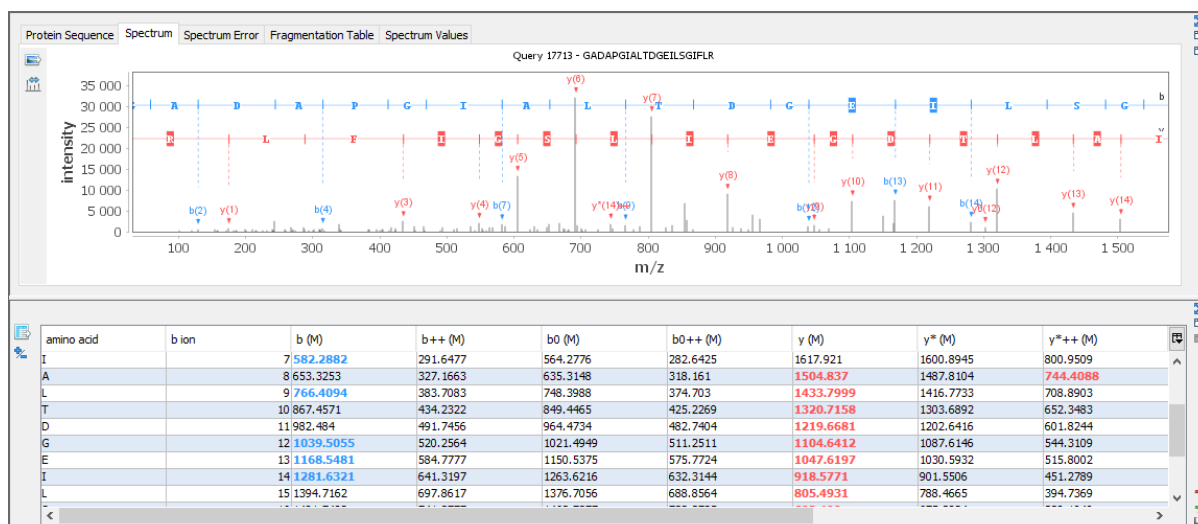



Figure 4: Annotated MS/MS spectrum and fragmentation table.

- Click **OK** and wait for the end of the conversion task (see Subheading 3.2, Step 6).
- Move back to the **Projects** tab and select the parent dataset. **Right-click** on the dataset and select **Quantify** **Label-Free** (see Note 14).
- In the dialog (see Figure 6a), the first step is the experimental design definition. Groups and samples corresponding to conditions to compare and to biological samples can be created manually by right-clicking on the **Quant** node. However, a more automated solution is available: drag and drop the parent dataset node from the right panel into the left panel, on the **Quant** node. An experimental design is automatically created and appears under the sample node (in this case, it contains a single group, a single sample and the four replicates). **Right-Click** on the **Quant** node then select **Rename**. Modify the default dataset name and click **OK**. Click **Next**.
- In step 2 (Associate MS files to sample analyses), mzdb files must be associated to their respective datasets (see Figure 6b). In the right panel, browse the **mzdb\_files** folder to display the four mzdb files. Select these files (**Shift**+**left-click**) and drag and drop them into the panel named **Drop Zone**. Proteome takes advantage of the available metadata, especially the name of the peaklist associated to each dataset to match the peaklist name and the name of the mzdb files dropped in the **Drop Zone**. In this case, the names perfectly match so that there is nothing more to do. Click **Next**.
- The last step consists in setting the parameters used during the quantification process (see Figure 6c). By default the dialog shows a simplified set of parameters with default values suitable for high resolution mass spectrometry analysis: the mass over charge (moz) tolerance is set to **5.0 ppm**, the alignment process is performed by using a maximum shift of retention time of **600.0s** and the cross assignment is enabled between all runs with **60.0s** tolerance to match predicted retention time (see Note 15).
- Click **OK** to start the quantification and wait for the completion of the quantification task.
- Once completed, the quantification dataset node icon changes to , the dataset is ready for browsing and navigation by right-clicking on this icon.
- By default, peptide ion measurements are summarized as protein abundances using a simple sum operation. However, additional operations such as excluding peptides or ions based on their characteristics (missed cleavages, variable modifications, sequence specificity, etc.) or normalizing peptide and protein abundances between runs can be performed. These post-processing steps can be executed on-demand using different parameters or methods; there is no need to repeat the whole quantification process when changes are made. **Right-click** on the quantification dataset node then select **Compute Post Processing on Abundances**. Choose how ions and peptides abundances are combined into peptide and protein abundances in the parameter dialog and click **OK**. Wait for the task completion.

### 3.7 Navigate through quantification datasets

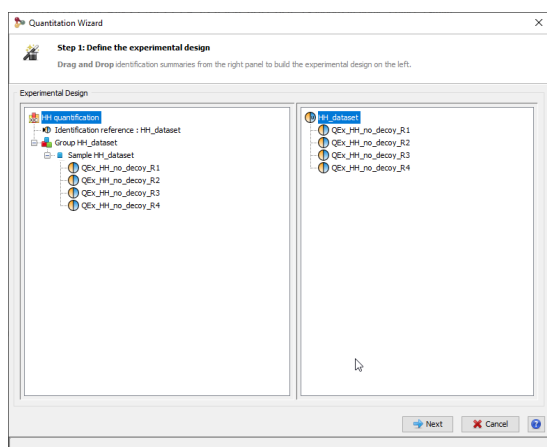
- Select the created quantification dataset and **right-click** then select **Display Abundances** **Protein Sets**. The predefined view opens (see Figure 7). This view starts from the quantified list of protein sets shown in the first table. Below this table, the quantified peptides of the selected protein set are represented in a tabular display on the left hand side; and in a graphical view on the right one. The graphical view represents the

Group	Type	QEx_HH_no_decoy_R1	QEx_HH_no_decoy_R2	QEx_HH_no_decoy_R3	QEx_HH_no_decoy_R4
<b>General Information</b>					
Raw File Name		QEx2_020296.mgf	QEx2_020300.mgf	QEx2_020322.mgf	QEx2_020419.mgf
Fasta Files		local/mascot/mascot_2.6/sequence/LP_K1...	local/mascot/mascot_2.6/sequence/LP_K1...	local/mascot/mascot_2.6/sequence/LP_K1...	local/mascot/mascot_2.6/sequence/LP_K1...
Search Result Name		YOC_SP-Tr_Cerevisiae	YOC_SP-Tr_Cerevisiae	YOC_SP-Tr_Cerevisiae	YOC_SP-Tr_Cerevisiae
Instrument Name		Q EXACTIVE (A1+FTMS F+HCD A2+FTMS)	Q EXACTIVE (A1+FTMS F+HCD A2+FTMS)	Q EXACTIVE (A1+FTMS F+HCD A2+FTMS)	Q EXACTIVE (A1+FTMS F+HCD A2+FTMS)
Fragmentation Rule Set		ESI-TRAP	ESI-TRAP	ESI-TRAP	ESI-TRAP
Target Decoy Mode					
Peaklist Software		Mascot Distiller	Mascot Distiller	Mascot Distiller	Mascot Distiller
<b>Search Properties</b>					
Result File Name		QEx_HH_no_decoy_R1.dat	QEx_HH_no_decoy_R2.dat	QEx_HH_no_decoy_R3.dat	QEx_HH_no_decoy_R4.dat
Search Date		8 février 2019	8 février 2019	8 février 2019	8 février 2019
Software Name		Mascot	Mascot	Mascot	Mascot
Software Version		2.6.0	2.6.0	2.6.0	2.6.0
Taxonomy		All entries	All entries	All entries	All entries
Enzyme		Trypsin/P	Trypsin/P	Trypsin/P	Trypsin/P
Max Missed Clivage		2	2	2	2
Fixed Modifications		Carbamidomethyl(C)	Carbamidomethyl(C)	Carbamidomethyl(C)	Carbamidomethyl(C)
Variable Modifications		Acetyl(Protein N-term), Oxidation(M)	Acetyl(Protein N-term), Oxidation(M)	Acetyl(Protein N-term), Oxidation(M)	Acetyl(Protein N-term), Oxidation(M)
Fragment Mass Tolerance		25.0 mmu	25.0 mmu	25.0 mmu	25.0 mmu
Peptide Charge States		2+ and 3+	2+ and 3+	2+ and 3+	2+ and 3+
Peptide Mass Error Tolerance		10.0 ppm	10.0 ppm	10.0 ppm	10.0 ppm
Fragment Charge States		2+ and 3+	2+ and 3+	2+ and 3+	2+ and 3+
Fragment Mass Error Tolerance		25.0 mmu	25.0 mmu	25.0 mmu	25.0 mmu
<b>Search Result Information</b>					
Queries Number		26604	26520	26103	24542
PSM Number		13499	13469	13312	14602
Protein Number		1887	1864	1857	2046
PSM Decoy Number					
Protein Decoy Number					
<b>Identification Summary Information</b>					
Protein Sets Number		1307	1295	1291	1290
PSM Number		10543	10445	10451	10931
Peptide Number		8184	8125	8164	8637
Protein Sets Decoy Number					
PSM Decoy Number					
Peptide Decoy Number					
<b>Validation Parameters</b>					
protein_filters / description		protein set filter on specific peptide (protein ...)	protein set filter on specific peptide (protein ...)	protein set filter on specific peptide (protein ...)	protein set filter on specific peptide (protein ...)
protein_filters / parameter		SPECIFIC_PEP	SPECIFIC_PEP	SPECIFIC_PEP	SPECIFIC_PEP
protein_filters / properties / thresh...		1	1	1	1
protein_filters / description		praline:fisher score	praline:fisher score	praline:fisher score	praline:fisher score
protein_filters / parameter		peptide match rank filter	peptide match rank filter	peptide match rank filter	peptide match rank filter
protein_filters / properties / thresh...		PRETTY_RANK	PRETTY_RANK	PRETTY_RANK	PRETTY_RANK
psm_filters#1 / description		peptide sequence length filter	peptide sequence length filter	peptide sequence length filter	peptide sequence length filter
psm_filters#1 / parameter		PEP_SEQ_LENGTH	PEP_SEQ_LENGTH	PEP_SEQ_LENGTH	PEP_SEQ_LENGTH
psm_filters#1 / properties / thresh...		7	7	7	7
psm_filters#2 / description		peptide match score filter	peptide match score filter	peptide match score filter	peptide match score filter
psm_filters#2 / parameter		SCORE	SCORE	SCORE	SCORE
psm_filters#2 / properties / thresh...		20.77	20.77	20.77	20.77
<b>Validation Results</b>					
protein_results / target_matches_...		1307	1295	1291	1290
psm_results / target_matches_count		10543	10445	10451	10931
<b>Sig IDs</b>					
Project id		2	2	2	2
Dataset id		12	13	14	15

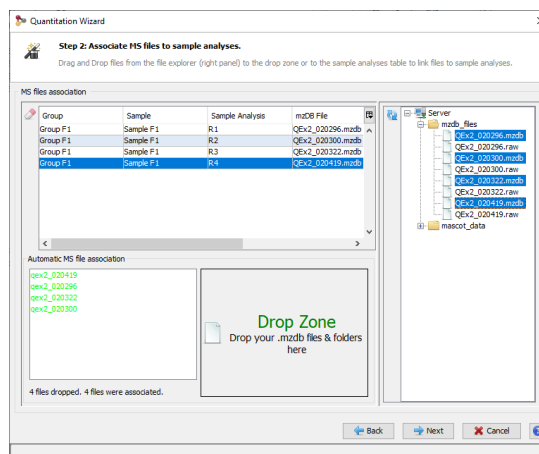
Figure 5: Datasets properties.

quantification profile of each peptide of the protein set (the abundance value in each MS analysis) as well as the protein set abundances (in yellow) calculated from these peptides on a second vertical axis. When a peptide is selected in the table, the graphical representation of the peptide is highlighted accordingly. The bottom panel shows the ions that have been quantified for the selected peptide (Quantification - Peptide Ions tab). The XIC Features tab shows the signal extracted from each MS analysis (called a feature) while the right hand side panel gives a graphical display of these features.

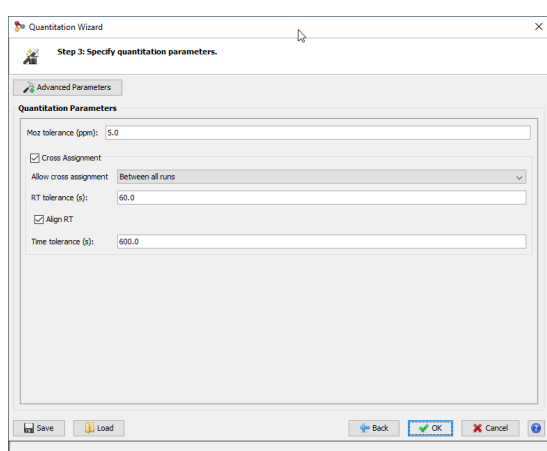
- Click on the Display all isotopes icon to display the signal extracted for each isotope of a currently selected feature. Click again to go back to the previous display showing all features.
- The graphical display uses the identified features stored in the database. However a chromatogram can be extracted from the mzdb files, even if Proline was unable to assign a signal abundance to the selected ion in some MS analyses. Select a peptide (For example, peptide IFNADWVIDGEQQPK of protein PUR4\_ECOLI), then an ion for which an abundance value is missing, then move on to the XIC Features. The table shows the abundance values extracted for each MS analysis. If Proline fails to find an abundance value for an MS analysis, the corresponding row is filled with zero values. Right-click on any row then select Extract All XIC. A task is created, requesting the extraction of the ion  $m/z$  chromatogram in all MS analysis. Once the response is received, the chromatograms are visible in the graphical display as dotted lines overlaid on the existing features.
- In addition to abundance values, the alignment between MS analyses, which is a critical step of the label free quantification process, can be visualized. Right-click on the quantification dataset then select Display Exp. Design >> Map alignment. The upper panel of the view shows the alignment curve between two selected MS analyses represented as the time difference (in seconds) between these two MS analysis over the overall analysis time (in minutes). The lower panel shows the alignment curves of all MS analysis with an analysis choose as the alignment reference.
- In both panels the alignment curves give a general trend of the retention time shift. However in the upper panel the retention time difference of each individual ion can be represented as a scatter plot. Click on the icon. After a few seconds, the Loading Data tool-tip disappears and the points representing the ions are shown overlaid on the curve. Filled circles represent ions that have been identified in



(a) Experimental design.




(b) MzDB files association.



(c) Label-free parameters.

Figure 6: Label-free quantification dialog.

both MS analyses while black circles represent ions that have been cross-assigned in one of the two MS analyses (cross-assigned ions can be removed from the display by clicking on the  icon). The upper and lower curves added to the alignment curve indicate the retention time tolerance used to perform the cross-assignment. Visualizing this information enables the user to adjust the retention time tolerance used during the cross-assignment step to reduce the likelihood of errors or, conversely, to take into account the retention time discrepancy between the MS analyses. To modify only this single parameter, relaunch the quantification by right-clicking on the quantitation dataset then select **Clone & Extract Abundances**: the quantification parameters dialog appears (see Subheading 3.6, Step 5) pre-filled with the experimental design and parameters of the selected quantification dataset. The parameter of interest can then be modified leaving all other parameters unchanged.

### 3.8 Customized Graphical display

Any dataset in Proline can be browsed using different predefined views, starting from an initial list of objects: PSMs, peptides, protein sets, MS/MS queries, etc. Additionally, new views can be created dynamically and saved for future browsing.

1. **Right-click** on the quantification dataset and select **Display Abundances** **New User Window**. The list of objects that can be used as starting point are shown. Select **Quant Protein Sets** and click **OK**. A new view appears, composed of a single table showing the list of quantified protein sets (see **Note 16**).
2. **Click** on the **+** icon in the right border of the panel to add a new panel into the view. A list of panels that are linked to the protein sets table is shown as well as an option to organize the panel into the view. Select the **Customisable Graphical Display** panel and add it **Below** the protein sets table.

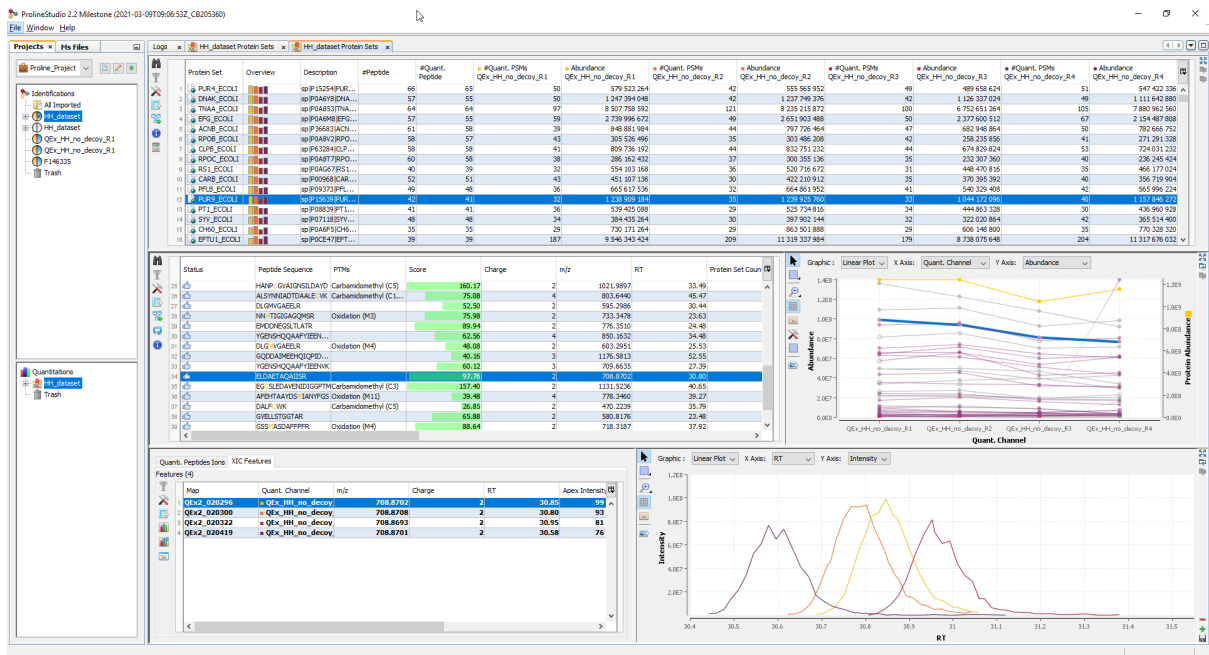







Figure 7: Navigation through quantification dataset.

- In the graphical view, set the Graphic type to **Scatter Plot**, X Axis to **Abundance QEx\_HH\_no\_decoy\_R1** and Y Axis to **Abundance QEx\_HH\_no\_decoy\_R2**. As expected, the two replicates are similar so that abundances are linearly correlated. **Right-click** on the X Axis and select **Log10 Axis**. Do the same with the Y Axis. To select points in the scatter plot where the abundances are different between replicates, **click** on the selection tool icon () to select a rectangular region (see **Note 17**) containing these points (see **Figure 9**). Selected points are then framed in black.
- The selection can be transferred between tables and graphical display. To select the protein sets corresponding to the selected points in the scatter plot, **click** on the **Export selection** icon (). The corresponding rows are selected in the protein sets table, which can then be filtered to display only the content of the selection: **Right-click** on the first selected row then select **View Selected Data**. To return to the original display **right-click** on a row and select **View All Data**.
- Add a new table to the view by clicking on the **+** icon. Select the **Quanti Peptides** panel and add it **Below**. The content of this panel is now synchronized with the selected protein set allowing to display the peptide quantification underlying the protein calculated abundance.
- Select a protein from the first table: the panel containing the peptides quantification information is updated accordingly, showing the quantified peptides belonging to the selected protein. Transfer this selection to the graphical display by clicking on the (): the corresponding point in the scatter plot is then selected.
- This view can be saved to be reused with any other quantification dataset: **Click** on the  icon in the left border of the panel to save the view. Choose a name for this view and click **OK**.
- Right-click** on the quantification dataset and select **Display Abundances**: the new view now appears in the sub-menu next to the existing predefined views.

### 3.9 Export data

Identification and quantification results as well as metadata related to the different processing steps performed by Proline are stored in the Proline database and can be exported into various formats. MSEXcel (.xlsx) and text (.csv) output formats contain data at PSMs, peptide ions, peptides and protein sets levels that can be customized before export. In addition, results can be exported in standard-compliant (mzIdentML) formats for publication of data in public repositories such as PRIDE and ProteomeXchange.

- Right-Click** on a dataset and select **Export**  **Excel**.
- Indicate the output file name.
- Select **Excel (.xlsx)** as Export type.

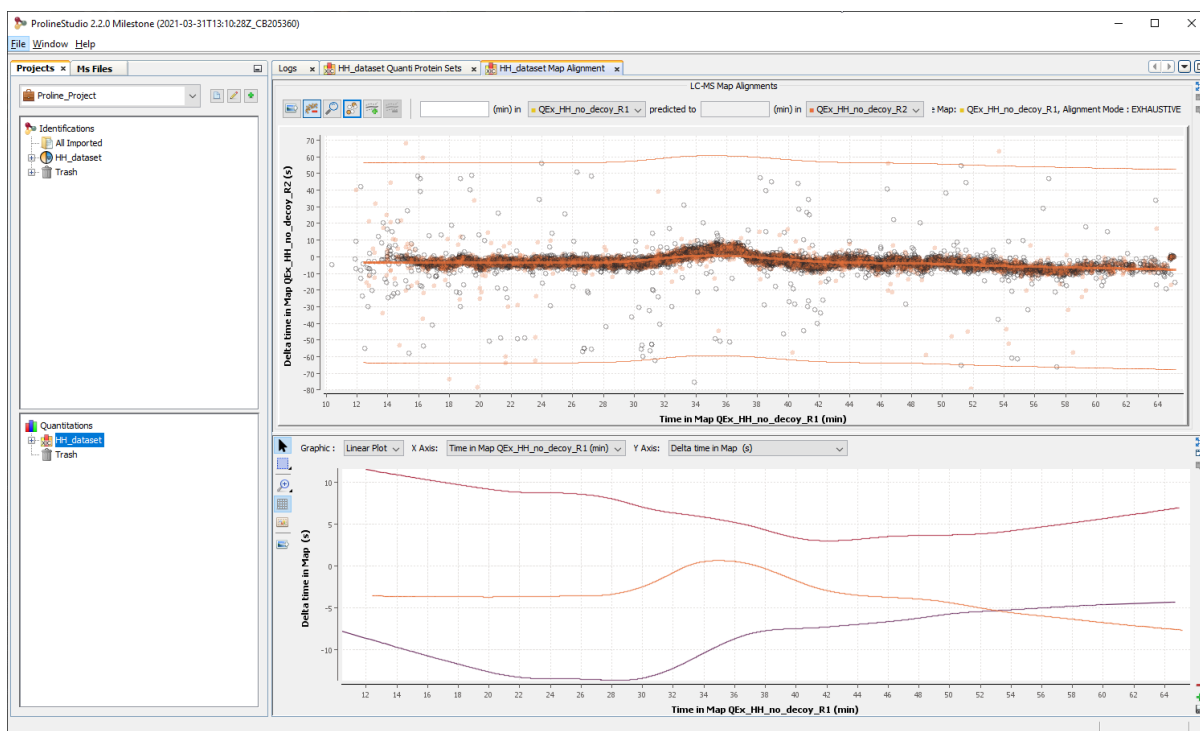


Figure 8: View MS analyses alignments.

4. Click on Custom options to choose the sheets, columns and column names that must be exported (Figure 10). Each exported sheet is represented by a tab in the customization dialog that can be unchecked to be removed from the exported file.
5. Click Export.
6. In addition to the identified PSM, peptides and proteins, Proline offers more specialized exports. Export Sequence Fasta exports a fasta file corresponding to the identified sequences.
7. Export Spectra List exports a spectral library containing the list of identified peptides with their precursor mass, their observed fragment masses and their retention time. This export can be achieved from identification or quantification datasets, however, MS/MS spectrum annotation must be generated beforehand (by Right-click on the dataset in the left panel and select Generate Spectrum Matches). Mainly used for Data Independent Acquisition (DIA), this output can be formatted to be compliant with PeakView or Spectronaut software (see Note 18).
8. Finally, the whole dataset can be exported in mzIdentML format for publication in ProteomeXchange. Right-click on the parent dataset icon and select Export MzIdentML. Fill the different fields with the needed administrative data and click Next. Set the output file path and click OK.

## 4 Notes

1. These values obviously depend on the data size.
2. To modify these settings, use an administrator user profile (user: admin and password: proline).
3. Results files produced by Mascot, X! Tandem, OMSSA and Andromeda search engines can be imported into Proline in their native format.
4. Combining datasets can be done in two different ways: Union duplicates all PSMs from the child datasets and creates a copy in the parent dataset; this is similar to what can be obtained by merging peak lists before submission to the search engine. Aggregate selects a representative PSM among PSMs matching the same peptide (same sequence and same post-translational modifications).
5. In this case, the search must have been performed against a target-decoy database (see Subheading 3.2) while in this protocol a classical search using a target database is sufficient. In the case of using a target-decoy strategy, any identification result performed against a target-decoy database from any search engine supported by Proline can be used. In addition, search results using the decoy option available in Mascot can also be imported and validated as a target-decoy search.

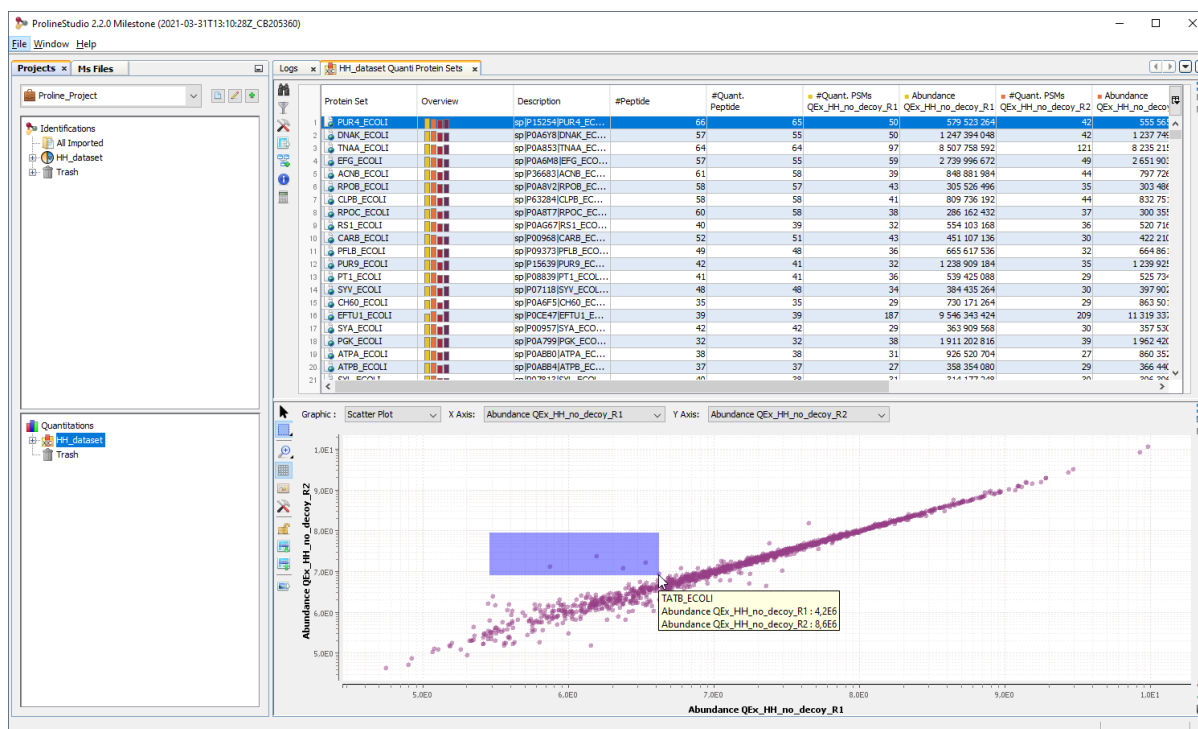


Figure 9: Custom quantification view.


6. Proline applies the BH filters in the bottom-up order (PSM > peptides > proteins), however, it is not necessary to use them all: it is possible to validate the identification results using only one or any combination of two of these filters, depending on the user's need and objectives.

7. For Advanced users, a fully regular expression can be specified. In this case, check the corresponding option. A maximum of three rules can be specified. They are applied in priority order, *i.e.*, if no protein of a protein set satisfies the first rule, the second one is tested and so on. If needed, the selection of the typical protein can be changed after the validation by right-clicking on the dataset and select **Change Typical Protein**. This operation can also be performed on multiple datasets at once by selecting multiple datasets (**Control** + **left click**).

8. In the same way as for non validated results, a dataset combined or merged after validation is annotated with letters U or A, which are added in the orange half-circle.

9. The pre-requisites are: (i) the fasta file name must be the same as the one used during the MS/MS identification search; (ii) a regular expression must be supplied to extract the protein accession from the fasta file to match the protein identification of the search result. Predefined regular expressions are configured into Proline so that uniprot accessions for example can be easily retrieved without modifying the configuration.

10. The default configuration can be modified by adding additional matching rules between fasta entries and protein accessions in the `parsing-rules.conf` file, located in the `SequenceRepository` sub-folder of ProlineZero. The syntax of this file is explained in Proline Installation Guide, section "Sequence Repository Configuration", subsection "Protein description parsing rule".

11. When displaying a target/decoy identification summary, only target proteins are shown in this table. However, the decoy results are also available and can be displayed in a similar view by clicking on the  icon in the menu bar. In this protocol, the dataset does not contain any decoy protein so the decoy dataset icon is disabled.

12. Instead of requesting the spectrum annotation PSM by PSM, this can be done systematically as follows: **Right-click** on a dataset in the left panel and select **Generate Spectrum Matches**. The same dialog appears and spectrum matches are generated and stored in the database for every validated PSM.

13. The converter is a standalone tool based on ProteoWizard (ensuring compatibility with a wide range of instrument vendors) available at <https://github.com/mzdb/pwiz-mzdb>. Note that the converter is only available on Windows platforms.

14. There are two ways to start a quantification process: from the identification dataset or from the Quantitations node in the lower left panel. Starting from the identification dataset ensures that the

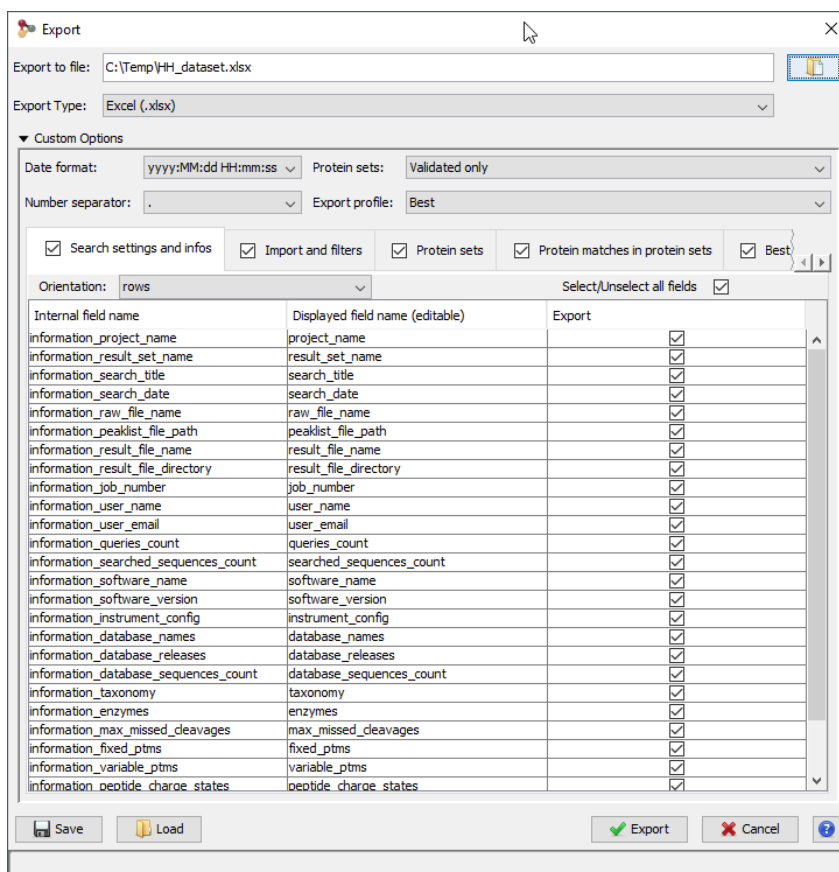


Figure 10: Customize Export dialog.

quantification process will start from a set of PSM/peptides and proteins that are controlled by the validation process and by the selected validation criteria.

15. The complete set of parameters, as described in [1], is accessible by clicking on the **Advanced Parameters** button.

16. Any predefined view can also be customized: as an example from the Protein sets view, in the right border of the bottom panel **click** on the minus icon (—) to remove the XIC feature graphical display. Possibly, **Click** repeatedly until the protein sets table remains the only element in the view (the minus icon will then disappear).

17. Clicking on the arrow at the bottom right corner of this icon allows to modify the selection tool. The default selection tool selects the points within a rectangular region drawn by the user. The alternative selection tool allows the user to draw a freehand selection region. Holding the **Ctrl** key allows points from disconnected regions to be added to the current selection.

18. Even if the spectra list can be exported from identification or quantification dataset, we recommend to start from a quantification node. Indeed, in such a case, the retention time of identified peptides corresponds to the apex of the extracted signal, which is more accurate than the retention time of the MS/MS spectrum (used for identification datasets).

### Acknowledgement

This work was supported by grants from the "Investissement d'Avenir Infrastructures Nationales en Biologie et Santé" program (ProFI project, ANR-10-INBS-08), by the French National Research Agency (GRAL project, ANR-10-LABX-49-01) and by MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

## References

- [1] Bouyssié D, Hesse AM, Mouton-Barbosa E, Rompais M, Macron C, Carapito C, Gonzalez de Peredo A, Couté Y, Dupierriis V, Burel A, Menetrey JP, Kalaitzakis A, Poisat J, Romdhani A, Burlet-Schiltz O, Cianférani S,

- Garin J, Bruley C (2020) Proline: an efficient and user-friendly software suite for large-scale proteomics. *Bioinformatics* 36(10):3148–3155, <https://doi.org/10.1093/bioinformatics/btaa118>
- [2] Elias JE, Gygi SP (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* 4(3):207–214, <https://doi.org/10.1038/nmeth1019>
- [3] Couté Y, Bruley C, Burger T (2020) Beyond Target–Decoy Competition: Stable Validation of Peptide and Protein Identifications in Mass Spectrometry-Based Discovery Proteomics. *Analytical Chemistry* 92(22):14898–14906, <https://doi.org/10.1021/acs.analchem.0c00328>
- [4] Vizcaíno JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Ríos D, Dianas JA, Sun Z, Farrah T, Bandeira N, Binz PA, Xenarios I, Eisenacher M, Mayer G, Gatto L, Campos A, Chalkley RJ, Kraus HJ, Albar JP, Martinez-Bartolomé S, Apweiler R, Omenn GS, Martens L, Jones AR, Hermjakob H (2014) ProteomeXchange provides globally co-ordinated proteomics data submission and dissemination. *Nature biotechnology* 32(3):223–226, <https://doi.org/10.1038/nbt.2839>
- [5] Nesvizhskii AI, Aebersold R (2005) Interpretation of Shotgun Proteomic Data The Protein Inference Problem. *Molecular & Cellular Proteomics* 4(10):1419–1440, <https://doi.org/10.1074/mcp.R500012-MCP200>
- [6] Bouyssie D, Dubois M, Nasso S, Peredo AGd, Bulet-Schiltz O, Aebersold R, Monsarrat B (2015) mzDB: A File Format Using Multiple Indexing Strategies for the Efficient Analysis of Large LC-MS/MS and SWATH-MS Data Sets. *Molecular & Cellular Proteomics* 14(3):771–781, <https://doi.org/10.1074/mcp.O114.039115>