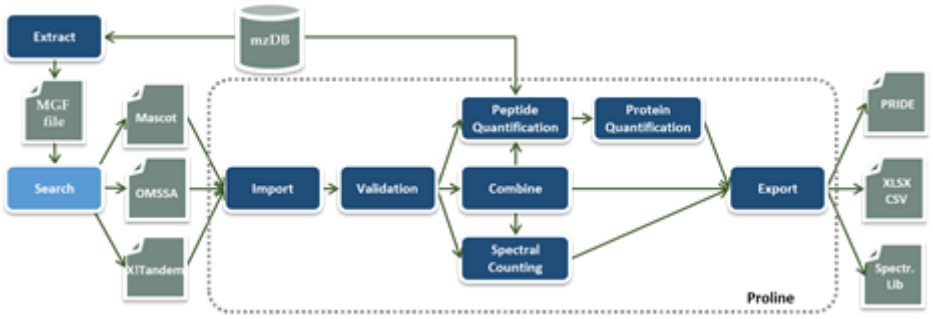
Proline User Guide

Release 2.3.0

Proline Software Suite

Proline is a production grade software suite, which provides an environment for large-scale MS data management, visualization, analysis and curation with the main objective of promoting the production and sharing of high quality proteomic datasets. Proline can be used (i) to produce reliable identification and quantification results through robust automated processes, (ii) for data curation, (iii) to systematically save and keep track of metadata from processing steps, parameters and generated data, and (iv) to submit highly qualified datasets to public repositories.



A workflow in Proline is implemented as a collection of tasks (see Figure) that can be performed by the user through the graphical user interface. Users can import multiple identification results corresponding, for example, to fractions and replicates of a biological sample and combine them before or after validation. The resulting datasets can then be compared or quantified using spectral counting or DDA label-free quantification, before exporting the results in different file formats.

The software suite is based on two main components: a **server** handling processing tasks and based on **relational database management system** storing the data generated and two different graphical user interfaces, both allowing users to start tasks and visualize the data: **Proline Studio** which is a rich client interface and **Proline Web** the web client interface. An additional component called **ProlineAdmin** is used by system administrators to set up and manage Proline. **Proline Zero** is an all-in-one, "zero installation" solution containing the **server** and **Proline Studio**.

Document organization

This document is organized in two main sections:

- The Concepts & Principles section presents the main concepts and algorithms implemented in Proline as well as the different parameters of these algorithms.
- The How to section gives more details on how to proceed with the Proline Studio graphical interface.

You could also find a description of available converters to create files in **mzDB format** (Bouyssié et al., 2015) from spectrometers raw files.

Proline Concepts & Principles

- Dataset types:
  - Result File
  - Search Results
  - Decoy Searches
  - Identification Summary
  - Modification Dataset
- Data Processing:
  - Protein Inference
  - Protein and Proteins Sets scoring
  - Validation Algorithm
  - FDR Estimation
  - Protein Sets Filtering
  - Combining datasets
  - Identify Modification Sites or Clusters
  - Compare with Spectral Count
  - Label Free LC-MS quantitation workflow
  - Isobaric labelling quantitation
  - Post-processing of quantitative results
  - Aggregation of quantitative results
- Data Import/Export:
  - Identification Summary Export
- Advanced Features:
  - Allow multiple imports in parallel

Identification results

Proline considers different types of identification data: Result Files, Search Results and Identification Summaries which will be defined in the following sections.

Result File

A **result file** produced by a search engine can be imported into Proline in their native format. OMSSA (`.omx` files), Mascot (`.dat` files) and X!Tandem (`.xml` files) search engines are currently supported. In addition, the mzIdentML format is supported to allow the output from any other search engine compatible with this standard to be imported (e.g. MS-GF+). A first version for MaxQuant support has been implemented. It is possible to import only search results or to import search results as well as quantitation (beta version) values from MaxQuant files.

Search engines may provide different types of searches for MS and MS/MS data. It is important to highlight that Proline only supports MS/MS ions searches at this point.

Search Result

A **search result** is the raw interpretation of a given set of MS/MS spectra given by a search engine. It contains one or many peptides matching the submitted MS/MS spectra (**PSM**, i.e. Peptide Spectrum Match), and the protein sequences these peptides belong to. The **search result** also contains additional information such as search parameters, used protein sequence databank, etc.

A **search result** is created when a result file is imported in Proline. During this step, no filtering or thresholding is applied: along with the search parameters, all submitted spectra, peptide spectrum matches (PSMs) and protein hits suggested by the search engine are retained in the Proline database to allow subsequent validation of putative identifications. In the case of a target-decoy search, two **search results** are created: one for the target PSMs, one for decoy PSMs.

Content of a search result

Importing a result file creates a new **search result** in the database which contains the following information:

- Search Settings: software name and version, parameters values
- Peaklist and Spectrum information: file name, MS level, precursor m/z, …
- Search result data:
    - Protein sequences
    - Peptide sequences
    - Spectra
    - Two kinds of Matches:
        - Peptide Spectrum Matches (PSM), i.e. the matching between a peptide and a spectrum, with some related data such as the score, fragment matches…
        - Protein Matches, i.e. the proteins in the databank corresponding to the PSMs identified by the search engine

Search engine specificities

- Mascot
    - The PSM score corresponds to the Mascot ion score.
- OMSSA
    - The PSM score corresponds to the negative common logarithm of the E-value:
      Score = -log10(E-value). Note that Proline only supports OMSSA Result Files generated with the 2.1.9 release.
- X!Tandem
    - The X!Tandem standard hyperscore is used as a PSM score. Note that Proline supports X!Tandem Result Files generated with the Sledgehammer release (or later).

Decoy Searches

Proline handles decoy searches performed from two different strategies:

- Concatenated searches:
    - A protein databank is created by concatenating target protein sequence to decoy protein sequence. Decoy could be created using reverse or random strategy. A unique search is done using that databank.

- When importing a Search Result from a decoy concatenated databank, decoy data are extracted from the Result File and stored in Proline databases as a decoy Search Result independent of the target Search result. Nevertheless both searches are linked to each other.

- Separated searches:
  - Two searches are done using the same peaklist, one on a target protein databank and one on a decoy protein databank. These searches are then combined to retrieve useful information such as FDR. Mascot allows the user to check a decoy option and automatically creates a decoy databank.

  - The two performed searches are stored in Proline databases and are linked together.

Identification Summary

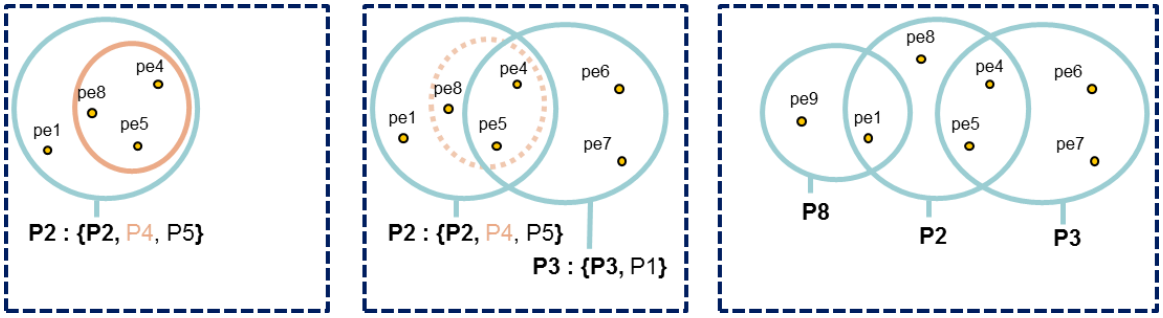
An **Identification Summary** is a set of identified proteins inferred from a subset of the PSM contained in the search result that have been declared valid. The subset of PSM taken into account are the PSM that have been validated by a filtering process (example: PSM fulfilling some specified criteria such as score greater than a threshold value).

Content of an identification summary

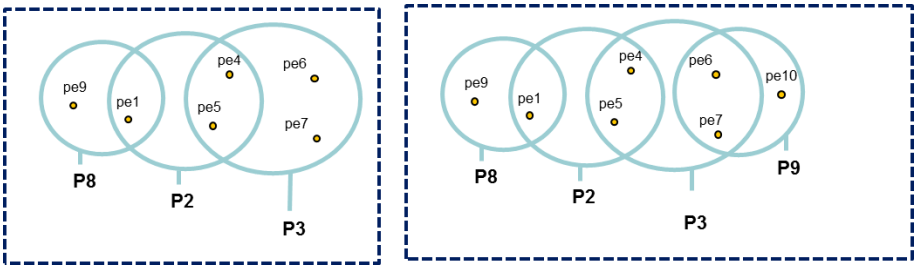
- Peptide Set
  - A set of peptides identifying one or more proteins.
- Protein Set
  - Typical Protein or representative protein. Protein "chosen" to represent the ProteinSet.
  - Sameset : all proteins identified by the same set of peptides than the typical one.
  - Subset: Proteins identified by a subset of peptides. All these peptides should also identify typical protein.

Protein Inference

All peptides identifying a protein are grouped in a `Peptides Set`. A same Peptides Set can identify many proteins, represented by one `Proteins Set`. In this case, one protein of this Protein Set is chosen to represent the set, it is the `Typical Protein`. If only a subset of peptides identify a (or some) protein(s), a new Peptide Set is created. This Peptide Set is a subset of the first one, and identified Proteins are `Subset Proteins`.



- In the first example, P2 and P5 are identified by the same peptide set {pe1, pe4, pe5, pe8}. P2 was chosen as typical protein. One subset composed of {pe4, pe5, pe8} identifies subset protein P4.
- In the second example, another protein set represented by P3 shares some peptides with the protein set represented by P2. Both protein sets have specific peptides.
- Sharing could involve many protein sets as shown in example 3.



All peptides sets and associated protein sets are represented, even if there are no specific peptides. In both cases above, no choice is done on which protein set / peptide set to keep. These protein sets could be filtered after inference (see Protein sets filtering).

Modification Dataset

From an Identification Summary, users can ask to Identify Modification Sites and Clusters, thus creating a Modification Dataset containing the result. This dataset contains

- Modification Sites: a modification type at a given location on a given protein. They are only extracted for the proteins that are representative of a validated protein set. Only modifications of interest are used in this process.
- Modification Clusters: co-localized modification sites are grouped into clusters. Identify Modification Sites and Clusters describe this grouping process.

Proteins and Protein sets scoring

There are multiple algorithms that could be used to calculate the Proteins and Protein Sets scores. Proteins scores are computed during the importation phase while Protein Sets scores are computed during the validation phase.

Protein

Each individual protein match is scored according to all peptide matches associated with this protein, independently of any validation of these peptide matches. The sum of the peptide matches scores is used as protein score (called standard scoring for Mascot result files).

Protein Set

Each individual protein set is scored according to the validated peptide matches belonging to this protein set (see inference).

Scoring schemes

Mascot Standard Scoring

The score associated with each identified protein (or protein set) is the sum **of the score of all peptide matches** identifying this protein (or protein set). In case of duplicate peptide matches (peptide matched by multiple queries) only the match with the best score is considered.

Mascot MudPIT Scoring

This scoring scheme is also based on the sum of all non-duplicate peptide matches score. However the score for each peptide match is not its absolute value, but the amount that it is above the threshold: the score offset. Therefore, peptide matches with a score below the threshold do not contribute to the protein score. Finally, the average of the thresholds used is added to the score. For each peptide match, the “threshold” is the homology

threshold if it exists, otherwise it is the identity threshold. The algorithm below illustrates the MudPIT score computation procedure:

```
Protein score = 0
For each peptide match {
 If there is a homology threshold and ions score > homology threshold {
        Protein score += peptide score - homology threshold
 } else if ions score > identity threshold {
        Protein score += peptide score - identity threshold
 }
}
Protein score += 1 * average of all the subtracted thresholds
```

- if there are no significant peptide matches, the protein score will be 0.
- homology and identity threshold values depend on a given p-value. By default Mascot and Proline compute these thresholds with a p-value of 5%.
- In the case of separated target-decoy searches we obtain two values for each threshold: one for the target search and another one for the decoy search. In order to obtain a single value we apply the following procedure:
    - the homology threshold is the decoy value if it exists else the target value
    - the identity threshold is the mean of target and decoy values.

The benefit of the MudPIT score over the standard score is that it removes many of the junk protein sets, which have a high standard score but no high scoring peptide matches. Indeed, protein sets with a large number of weak peptide matches do not have a good MudPIT score.

Mascot Modified MudPIT Scoring

This scoring scheme, introduced by Proline, is a modified version of the Mascot MudPIT one. The difference with the latter is that it does not take into account the average of the subtracted thresholds:

```
Protein score = 0
For each peptide match {
 If there is a homology threshold and ions score > homology threshold {
        Protein score += peptide score - homology threshold
 } else if ions score > identity threshold {
        Protein score += peptide score - identity threshold
 }
}
```

This score has the same benefits as the MudPIT one. The main difference is that the minimum value of this modified version will be always close to zero while the genuine MudPIT score defines a minimum value which is not constant between the datasets and the proteins (i.e. the average of all the subtracted thresholds).

Fisher

This scoring scheme relies on the Fisher's test to define protein scores and p-value from the scores of the best subset of peptides. The scoring is extensively described in (https://doi.org/10.1021/acs.analchem.0c00328).

Search result validation

Once a result file has been imported and a search result created, the validation is performed in four main steps:

1. Peptide Matches filtering and Validation
2. Protein Inference (peptides and proteins grouping)
3. Protein and Proteins Sets scoring
4. Protein Sets Filtering and Validation

Finally, the identification summary issued from these steps is stored in the identification database. Different validation of a Search Result can be performed and a new Identification Summary of this Search Result is created for each validation.

When validating a merged Search Result, it is possible to propagate the same validation parameters to all childs Search Results. In this case Peptide Matches filtering and validation will be applied on childs as well as Protein Sets filtering. Note: actually, Protein Sets validation is not propagated to childs Search Results.

Peptide matches filtering

Peptide Matches identified in search results can be filtered using one or multiple predefined filters (described hereafter). Only validated peptide matches will be considered for further steps.

Score filter

All PSMs with a score lower than a given threshold are discarded. For some search engines, Proline computes the PSM score value itself by applying a mathematical transformation to another PSM property. For instance, the score values for PSMs from X!Tandem search results correspond to the log10 transformation of the PSMs expectation values.

Pretty Rank filter

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.

Minimum sequence Length filter (Length)

PSMs corresponding to peptide sequences shorter than the cut-off stipulated will be discarded when this parameter is applied.

Mascot e-Value filter (e-Value)

This filter is used to select PSMs based on the Mascot expectation value (e-value) which reflects the difference between the PSM's score and the Mascot identity threshold (p=0.05). PSMs with an e-value greater than the threshold specified are discarded.

Mascot adjusted e-Value filter (Adjusted e-Value)

Proline can compute an adjusted e-value. It first selects the lowest threshold between the identity and homology e-values (p=0.05). Then, it computes the e-value using this selected threshold. PSMs for which the adjusted e-value is greater than the specified cut-off are discarded.

Mascot p-Value based on identity filter (Identity p-Value)

Given a specific p-value, the Mascot identity threshold is calculated for each query and all peptide matches associated with the query for which the score is lower than the identity threshold calculated are discarded.

Mascot p-Value based on homology filter (Homology p-Value)

Given a specific p-value, the Mascot homology threshold is inferred for each query and all peptide matches associated with the query which have a score lower than the calculated homology threshold are discarded.

Single PSM per MS Query filter

This filter validates only **one PSM per Query**. To select a PSM, following rules are applied:

For each query:

- Select PSM with higher score.
- If several PSM with same score:
    - Choose PSM which identify the protein which have the max number of valid PSM
    - If still equality choose the PSM with the lower delta moz
    - If still equality, select the first PSM using alphabetic order
    - If still equality, select the first one. Warning : may be different in different versions of Prolline...

Single PSM per Rank filter

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, the same selection than “Single PSM per Query” is used

- Select PSM that identifies the protein which has the max number of valid PSM.
- If several PSM are equals
    - Choose the PSM with the lower delta moz.
    - If still equality, select the first PSM using alphabetic order.
    - If still equality, select the first one. Warning : may be different in different versions of Proline...

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.

Single Sequence per Rank filter

This filter selects the best **PSM per pretty rank** and all PSMs that have the **same sequence**, without considering Post translational modification. The best PSM is selected using the same rules as the Single PSM per Rank filter described above.

Isotope Offset filter

This filter validates PSMs which have been validated using an isotope offset less or equal than specified one. This corresponds to identification search engine parameters such as Mascot isotope error and could be used to invalidate PSM identified using one or more 13C .

Peptide matches validation

In addition to these filters a PSM level validation step may be specified using:

- A target-decoy approach (Elias and Gygi, 2007), which can be performed by adjusting a user-specified validation criterion until it reaches a user-specified false discovery rate (FDR). The search engine score can be used as a generic validation criterion for any of the search engines supported by Proline. For results obtained with the Mascot search engine, other criteria can be used to control the FDR: Mascot e-Value, Mascot adjusted e-Value, Mascot p-Value based on identity threshold, or Mascot p-Value based on homology threshold.
- A method based on the Benjamini-Hochberg procedure to control the FDR proposed in (https://doi.org/10.1021/acs.analchem.0c00328)

FDR estimation for Target Decoy validation

There are several ways to calculate FDR depending on the database search type. In Proline the FDR is calculated at PSM and protein levels using the following rules:

- if the Search has been done on a concatenated Target/Decoy bank or if rank filter has been used during validation :

$$FDR = 2 \frac{nbr\ DecoyPSM}{nbr\ TargetPSM + nbr\ DecoyPSM}$$

Note: when computing PSM FDR, peptide sequences matching a Target Protein and a Decoy Protein are taken into account in both cases.

- if the Search has been done on a separated Target/Decoy bank :

$$FDR = \frac{nbr\ DecoyPSM}{nbr\ TargetPSM}$$

Protein sets filtering

Any Identification Summary, generated by a validation process or by merging datasets could be filtered.

Filtering consists in invalidating Protein Sets which don't follow specified criteria. Invalidated Protein Sets are not taken into account for further algorithms or display.

Available filtering criteria are defined below.

Specific peptides filter

This filter invalidates protein sets that don't have at least x peptides identifying only that protein set. The specificity is considered at the DataSet level.

This filtering goes through all Protein Sets from worse score to best score. For each, if the protein set is invalidated, associated peptides properties are updated before going to the next protein set. Peptide property is the number of identified protein sets.

Peptides count filter

This filter invalidates protein sets that don't have at least x peptides identifying that protein set, independently of the number of protein sets identified by the same peptide.

This filtering goes through all Protein Sets. For each, if the protein set is invalidated, associated peptides properties are updated before going to the next protein set. Peptide property is the number of identified protein sets.

Peptide sequence count filter

This filter invalidates protein sets that don't have at least x different peptide sequences (independently of PTMs) identifying that protein set.

This filtering goes through all Protein Sets from worse score to best score. For each, if the protein set is invalidated, associated peptides properties are updated before going to the next protein set. Peptide property is the number of identified protein sets.

Protein set score filter

This filter invalidates protein sets which score is below a given value.

Protein sets validation

Once prefilters (see above) have been applied, a validation algorithm can be run to control the FDR. In the same way as for PSM, FDR may be controlled using :

- the Benjamini-Hochberg procedure, see (https://doi.org/10.1021/acs.analchem.0c00328)
- a target decoy strategy.

Target-Decoy Strategy:

See how FDR is calculated.

At the moment, it is only possible to control the FDR by changing the Protein Set Score threshold. Three different protein set scoring functions are available.

Given an expected FDR, the system tries to estimate the best score threshold to reach this FDR. Two validation rules (R1 and R2) corresponding to two different groups of protein sets (see below the detailed procedure) are optimized by the algorithm. Each rule defines the optimum score threshold allowing to obtain the closest FDR to the expected one for the corresponding group of protein sets.

Here is the procedure used for FDR optimization:

- protein sets are segregated in two groups, the ones identified by a single validated peptide (G1) and the ones identified by multiple validated peptides (G2), with potentially multiple identified PSMs per peptide.
- for each of the validation rules, the FDR computation is performed by merging target and decoy protein sets and by sorting them by descending score. The score threshold is then modulated by using successively the score of each protein set of this sorted list. For each new threshold, a new FDR is computed by counting the number of target/decoy protein sets having a score above or equivalent to this value. The procedure stops when there are no more protein sets in the list or when a maximum FDR of 50% is reached. It has to be noted that the two validation rules are optimized separately:
    - G2 FDR is first optimized leading to the R2 score threshold. The validation status of G2 protein sets is then fixed.
    - final FDR (G1+G2) is then optimized leading to the R1 score threshold. Only the G1 protein sets are here used for the score threshold modulation procedure. However the FDR is computed by taking into account the G2 validated target/decoy protein sets.

The separation of protein sets in two groups allows to increase the power of discrimination between target and decoy hits. Indeed, the score threshold of the G1 group is often much higher than the G2 one. If we were using a single average threshold, this would reduce the number of G2 validated proteins, leading to a decrease in sensitivity for a same value of FDR.

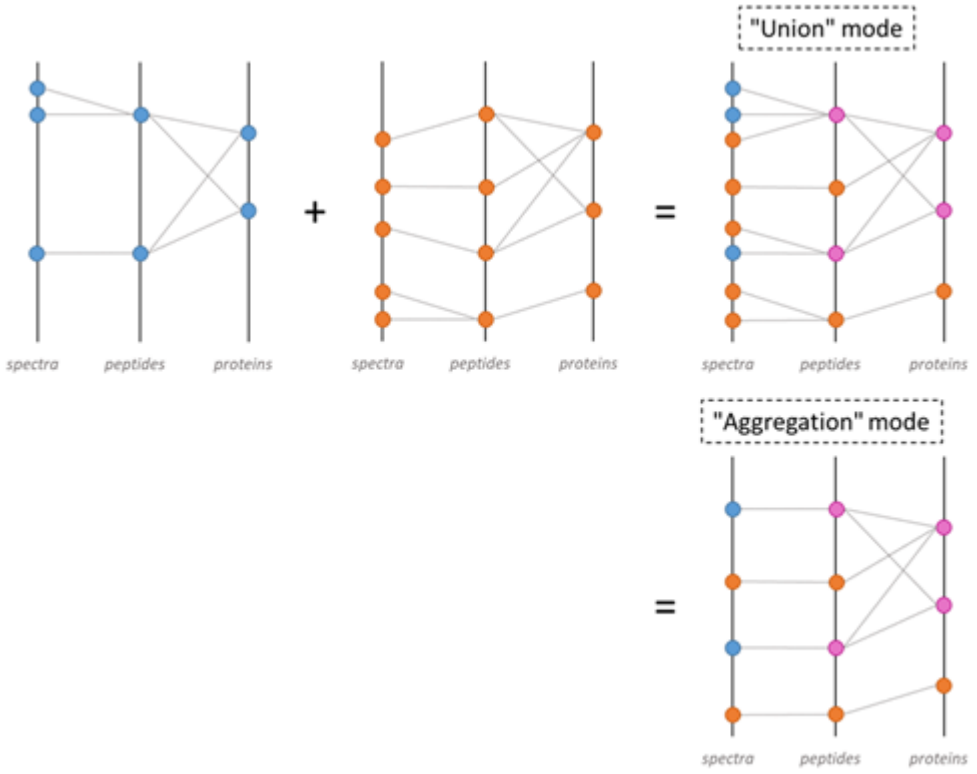
Combining datasets

Identification results can be combined to construct a parent dataset, and create a non-redundant list of identified peptides and proteins. This combination can be performed either before validation (on search results) or after validation (on identification summaries). Since this operation could be recursively performed, it leads to hierarchical structuring of search results and/or identification summaries. On the one hand, combination before validation (taking into account all PSMs identified by the search engine) may, for example, be relevant when analyzing results obtained after peptide fractionation: in that case, several peptides belonging to the same protein may be spread across different result sets; these sets should be merged before protein validation. On the other hand, merging

identification summaries is appropriate when seeking to group the validated results from a series of individual samples to be compared or when combining data from different search engines.

When datasets are combined, their PSMs are collected to generate a non-redundant set of peptides before recomputing protein inference. Additionally, the mappings between peptides and FASTA entries observed across the different datasets are also collected and merged into a single final mapping list. This list reflects thus the whole set of peptide and protein matches that were observed in the individual datasets.

Users can combine search results or identification summaries. The main difference is the set of spectra and peptides (and thus PSMs) considered. When combining search results, all spectra, peptides and PSMs in the dataset are considered, whereas when combining identification summaries, only validated PSMs are taken into account. In addition, Proline can be used to control how PSMs are collected in the parent dataset: in union mode, PSMs originating from combined datasets are added, while in aggregation mode, all PSMs identifying the same peptide are aggregated into a single representative PSM.



Combining datasets in Proline. Datasets are represented as a tripartite graph composed of spectra, peptides and proteins; edges between spectrum and peptides represent PSM. When the blue and orange datasets are combined, PSMs from both datasets are collected together, generating a non-redundant set of peptides. The combination can be performed in 2 modes: union or aggregation mode.

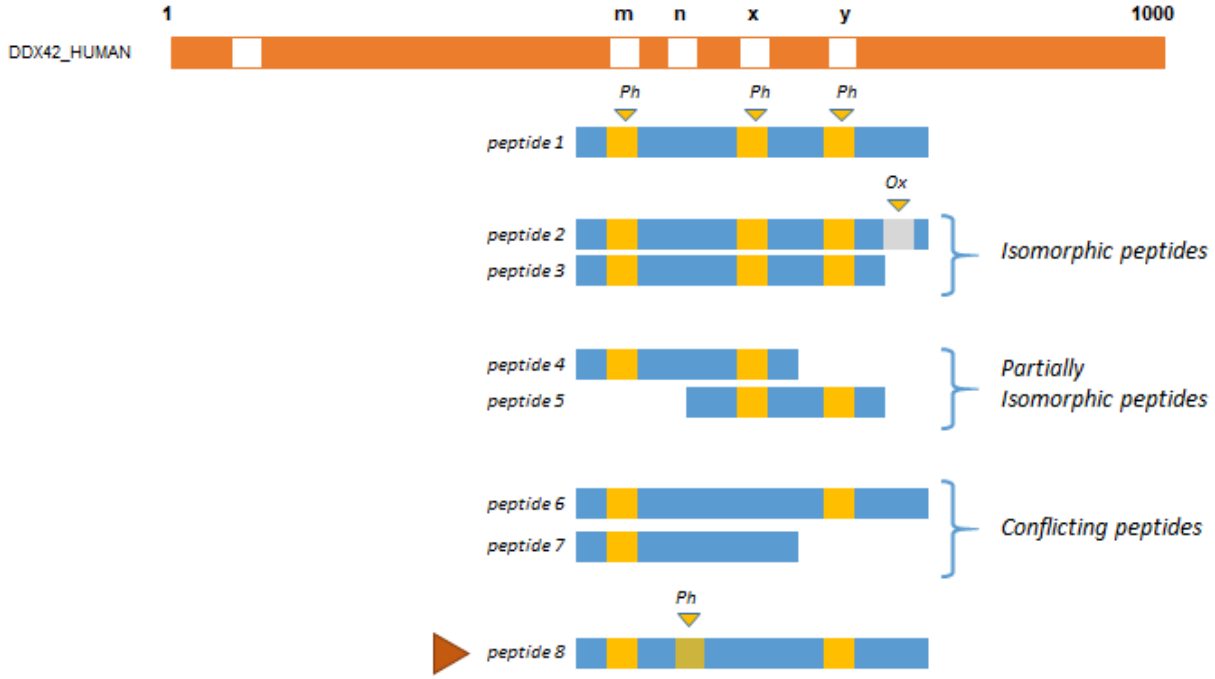
Identifying Modification sites and clusters

A list of Post translational modification (PTM) sites identified among the peptides of an identification summary can be extracted by Proline. A modification site is characterized by a modification type, at a given location on a given protein. The list of modification sites extracted by the software is restricted to the proteins that are representative of a validated protein set and to the modifications of interest specified by the user. This means that a peptide

identified with two different modifications of equal interest to the user will appear twice in the list, one for each modification location.

In a second phase, co-localized modification sites are grouped into clusters as soon as evidence of their co-existence exists. The required evidence is a peptide sequence identified in the dataset with all the clusterized modification sites.

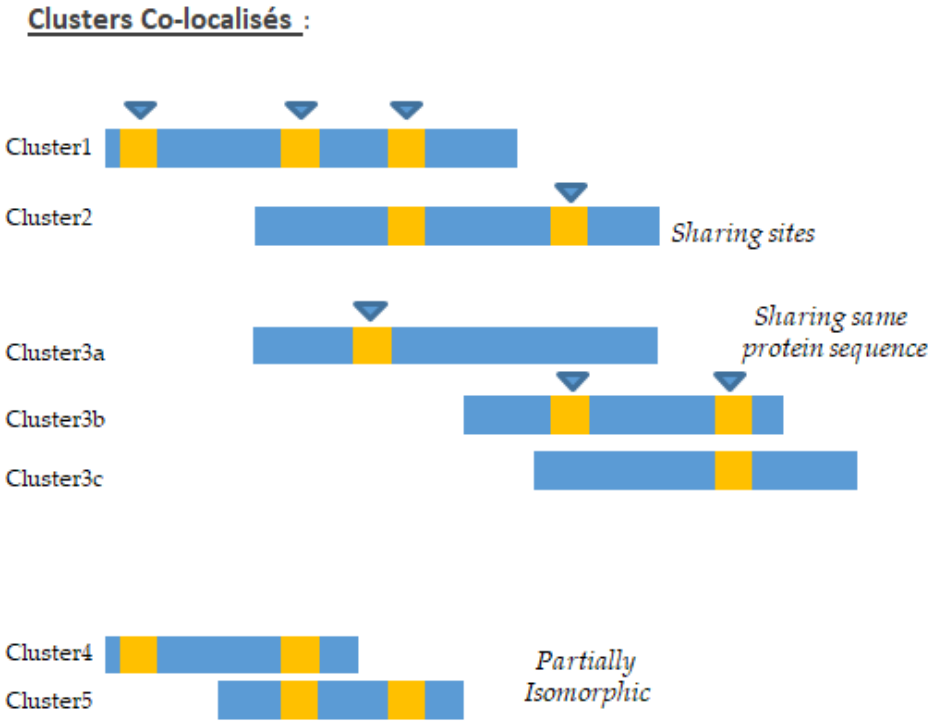
The figure below represents different peptides (blue rectangles), co-localized on a protein sequence (in red). In this example, modifications of interest (Phosphorylation (Ph)) are shown in orange. Peptide 1 proves that Phosphorylation at positions m, x and y occurs simultaneously. Peptides 2 and 3 are considered as isomorphic since the oxidation of peptide 2 is ignored (only Phosphorylations have been declared of interest by the user in this example). Peptides 4 and 5 are partially isomorphic: they confirm Phosphorylation respectively at position (m, x) and (x, y). Conversely peptides 6 and 7 are not in accordance, suggesting that there are two other proteoforms, one with Phosphorylations at position (m,y) but no modification at x and one with a Phosphorylation at position m but not at x. These two peptides could be grouped into another cluster.



The user can choose between two different modification clustering methods:

- Partially Isomorphic Matching (aka Exact Position matching): Two sequence matches are clusterized if they are Isomorphic or Partially Isomorphic. Partially Isomorphic matches could belong to multiple clusters.
- Only Isomorphic Matching: two sequence matches are clusterized if they are Isomorphic.

In Proline, clusters co-localisation concept is used. Co-localized clusters are clusters that have overlapping on protein sequence. These clusters may share one or more sites, but they could also cover the same protein sequence without any common site. See co-localized clusters example below.



Spectral Counting

Principles

Proline can be used to compare protein sets based on spectral counts through a previously presented algorithm (Hesse et al., 2016). This algorithm notably computes a weighted spectral count metric (called adjusted spectral count in the original publication). Basically, the algorithm takes both unique and shared peptides into account, and for each shared peptide, the proportion of MS/MS spectra that should be attributed to the different protein sets is determined. This proportion (also called weight) is based on the spectral counting of proteotypic (or specific) peptides identifying the different protein sets sharing the peptide to be attributed.

Spectral counting is calculated from a hierarchy of `identification summaries`. The parent `identification summary` (at the top of the hierarchy) is where the list of protein sets to compare and the list of specific peptides are created. The list of specific peptides is then used to compute the protein sets' respective weights but users can choose any"child" dataset where the weights must be calculated.

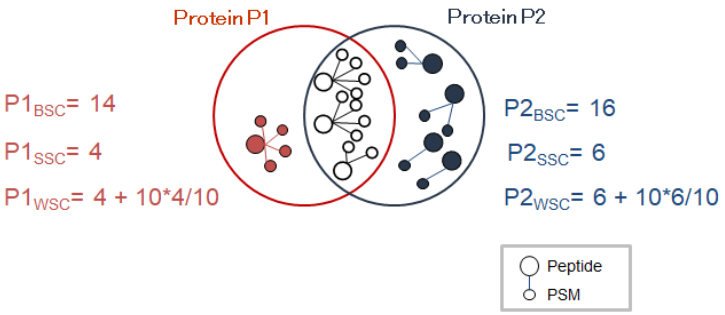
First, Proline compute the `peptide spectral count` at each level of the dataset hierarchy using the following rules:

- If the dataset is a "leaf" `identification summary` of the hierarchy (not issued from a merge, no child dataset), the `peptide spectral count` is the number of MS/MS spectra matching a peptide (equal to the number of peptide spectrum matches).
- If the dataset is a merged `identification summary`, the `peptide spectral count` is the sum of its child `peptide spectral count`.
- If the dataset is a merged of `search results`, the `peptide spectral count` is the sum of **validated children** `peptide spectral count`. Validated children are PSMs that meet validation criteria applied to parent `Identification Summaries`.

Once, `peptide spectral count` is calculated for each peptide, the `protein spectral count` is computed using the following rules:

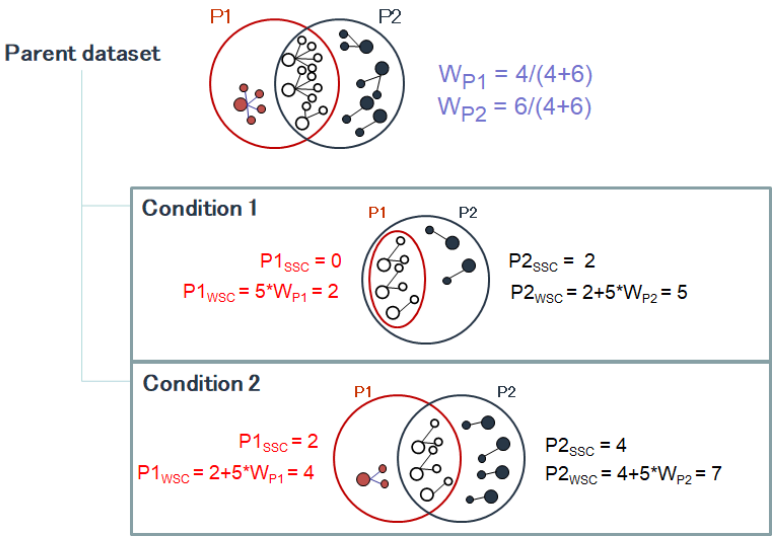
- `Protein basic spectral count` (BSC) is the sum of the `peptide spectral count` of all peptides matching a protein.
- `Protein specific spectral count` (SSC) is the sum of the `peptide spectral count` of specific peptides. A specific peptide is a peptide which does not identify any other protein sets in the context of the identification summaries.
- `Protein weighted spectral count` (WSC) is the `Protein specific spectral count` (SSC) + weighted `peptide spectral count` of shared peptides.

Simple example



Protein sets weights

The protein set respective weights computation is based on the proteotypic peptides. The level in the dataset hierarchy where these weights are calculated can be chosen by the user, it could be the top level of the dataset hierarchy or at a lower level. In the following example, the weights (W_{P1} and W_{P2}) are calculated at the "parent dataset" level. At this level, P1 and P2 are respectively identified by the red and dark blue peptides/psm, each protein set weight is calculated using their specific spectral counting. These weights are thus used to calculate the weighted spectral count of P1 and P2 in the two "child" dataset "Condition 1" and "Condition 2".



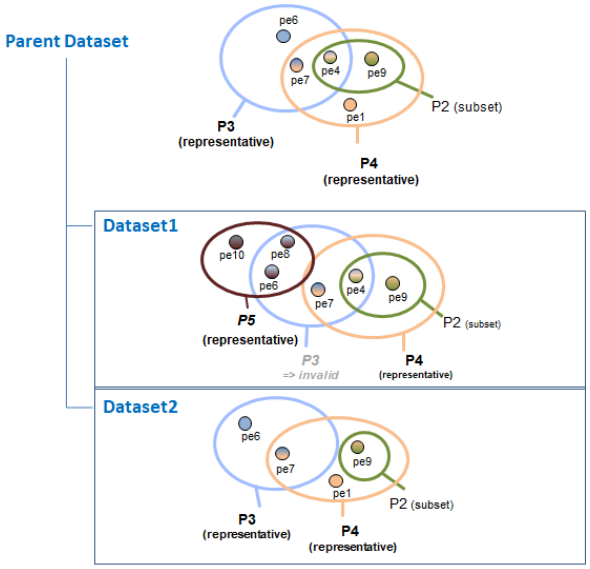
FAQ

Why is the BSC less than Peptide Count ?

When running SC even on a simple hierarchy (1 parent, 2 childs) in some cases we obtain a BSC value smaller than the peptide count of the protein set. This occurs only for *invalid protein sets*. *Invalid protein sets* are the ones that are present at the parent level but are filtered at child level (if a *specific peptide* filter has been applied for example).

Indeed, the peptide count value is read in the child protein sets. On the other hand, the BSC is calculated by getting the spectral count information at child level for each peptide identified at parent level. If a protein set is invalid, its peptides are not taken into account during the merging so some of them could be missing at the parent level if they were not identified in the other child.

This case is illustrated in the following figure:



Parent Protein Sets	Dataset 1			Dataset 2		
	Status	Pep Count	BSC	Status	Pep Count	BSC
P3	Invalid	4	3	representative	2	2
P4	representative	3	3	representative	3	3

Label-free LC-MS quantitation workflow

Proline detects chromatographic peaks from raw data **converted to the mzDB format** (Bouyssié et al., 2015). Depending on the Proline version and the instrument used to acquire data, two converters are available to generate mzdbFile.

After a first **signal extraction** step, the algorithm associates the chromatographic peaks detected with validated PSMs, first by retrieving the corresponding MS/MS spectra acquired during the peptide elution, and then by **matching the precursor** m/z value of these spectra to the chromatographic peak m/z value. After the **deisotoping** step, the abundance of each ion is estimated from the apex of the chromatographic peak, which corresponds to the theoretically most abundant isotopologue (inferred from the peptide's atomic composition). The software then **aligns** the retention time of these annotated ions for all the LC-MS runs to be compared, and uses this information

to **cross-assign** MS/MS data to ions (i.e. chromatographic peaks) that were detected but not identified in other runs . The resulting ion abundances are finally stored in the Proline database, making them available for rapid data visualization and further post-processing.
Finally, peptide ion measurements can be **summarized as protein abundances** using different computational methods. The user can opt to perform 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. 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.

Signal extraction

During an LC-MS experiment, the m/z and intensity values for each peptide ion detected are recorded in MS1 scans acquired during the elution of this peptide from the chromatographic column. Most existing peak picking algorithms analyze these MS scans individually or sequentially. The Proline algorithm performs the signal detection in a different way. It first takes advantage of the mzDB format to detect chromatographic peaks in spectrum slices (5 m/z wide by default) across the whole chromatographic time. Then, in a given slice, m/z peaks are sorted in decreasing order of intensity. Thus, starting from the most intense m/z peak (apex), the algorithm searches for a peak with the same m/z value in the previous and subsequent MS1 scans, while applying a user-defined m/z tolerance. This lookup procedure stops when the ion signal is absent from more than a predefined number of consecutive scans. The [RT, m/z, intensity] peak list obtained, which is comparable to an extracted ion chromatogram (XIC), is then smoothed using a Savitzky-Golay filter (Savitzky and Golay, 1964). The resulting smoothed chromatogram is then split into the time dimension to form chromatographic peaks, by applying a peak picking procedure that will search for significant minima and maxima of signal intensity. When the signals of two ions overlap in the time dimension, a minimum is generally surrounded by two maxima. If the corresponding valley is deep enough, i.e. at least 66% of the lower surrounding maximum, this minimum will be considered significant (and thus will trigger the generation of two peaks). Once the smoothed chromatogram has been fully analyzed, the algorithm removes the corresponding detected peaks from the current spectrum slice, and performs another lookup using the next available apex. The result of this whole procedure is a list of chromatographic peaks defined by an m/z value, an apex elution time and an elution time range.

Parameters

These parameters are used by signal extraction algorithms.

- Extraction m/z tolerance: this corresponds to the error tolerance used when searching for a m/z peak in the previous and subsequent MS1 scans (see above).
- Use previous/last peakel detection: If a quantitation has already been done on a run, use the previously detected peakels file for this run. **Warning**: In this case, the extraction m/z tolerance parameter is ignored since the signal extraction is already done !

PSM assignment and deisotoping

In a single run, validated PSMs are MS/MS spectra assigned to a peptide sequence, and each spectrum is characterized by a precursor mass, a charge state and a retention time (RT). The algorithm assigns PSMs to detected chromatographic peaks by matching the spectrum precursor m/z ratio to the chromatographic peak m/z and verifies that the spectrum retention time falls within the peak time range. The PSM charge state assigned is then used to search for chromatographic peaks corresponding to the ion's isotopologues, considering the precursor mass-to-charge ratio of the spectrum as the monoisotope. The peptide ion intensity is summarized by retaining only

the apex of the peak corresponding to the theoretically most abundant isotopologue (inferred from the peptide's atomic composition). All peptide ion signals (a.k.a. LC-MS features) extracted from an mzDB file and assigned to a PSM are used to construct an LC-MS map. For the sake of simplicity no distinction is made between LC-MS runs and LC-MS maps in this manuscript.

Parameters

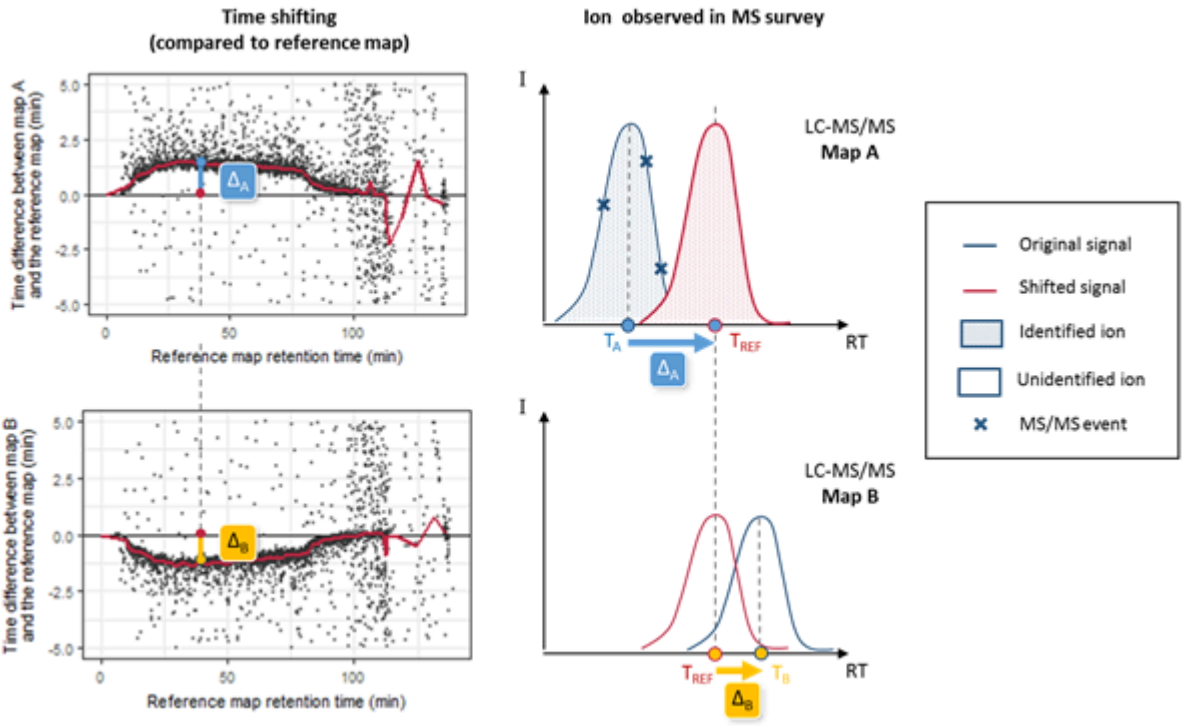
- PSM/Peakel matching m/z tolerance is the m/z error allowed to match a peakel to a PSM

Clustering is applied to group peakels that are matching the same identified ion.

- Cluster time computation: you have the choice between 2 computation methods: most intense or median. For "most intense" method, the cluster time corresponds to the time of the most intense feature composing the cluster. For the "median" method, cluster time is the median of the feature times forming the cluster.
- Cluster intensity computation: you have the choice between 2 computation methods: most intense or sum. For "most intense" method, the cluster intensity corresponds to the intensity of the most intense feature of features forming the cluster. For the "sum" method, cluster intensity is the sum of the intensities of features composing the cluster.

Retention time alignment

As soon as the PSM are matched to peakels, the software then aligns the retention time of the annotated ions for all the LC-MS runs to be compared, and uses this information to cross-assign MS/MS data to ions (i.e. chromatographic peaks) that were detected but not identified in other runs. Because chromatographic separation is not completely reproducible, LC-MS runs must be aligned. The retention time alignment procedure is a critical step in MS1 label-free quantification. Proline's alignment algorithm selects a **reference run** and generates a set of functions that will be used to predict the RT (retention time) for missing features from another run. These functions are obtained by performing pairwise alignments between the different runs to be compared (Bylund et al., 2002; Jaitly et al., 2006; Sadygov et al., 2006). The first step consists in computing a scatter plot (see Figure below) of the observed time difference between the two runs as a function of the reference run's time-scale. This mapping can be based on the **peptide identity** (same sequence and same post-translational modifications) of identified features, or by mapping the detected features of the pair of runs, taking user-defined time and mass error ranges into account (the default feature mapping time and m/z tolerance values are set to 600 seconds and 5 ppm, respectively).



***RT prediction using computed alignments**. The two scatter plots on the left correspond to computed run alignments between the reference run and two other runs (A and B). The red curves on these plots correspond to the median RT prediction for each alignment, obtained by applying a moving median calculation. The graphs on the right illustrate the case of a peptide ion that is present in runs A and B, but has only been fragmented by MS/MS in run A. Knowing the retention time in run A (T_A), we can predict T_B by two consecutive time conversions. T_A is first converted to the reference run scale ($T_{REF} = T_A + D_A$) using the first run alignment, then to $T_B = T_{REF} + D_B$ using the alignment for the second run.*

RT prediction functions are then obtained by **smoothing** these scatter plots using a moving median calculation or a local regression. To decrease the number of alignment combinations (i.e., pairs of maps), the reference run is determined by an iterative method. The algorithms begin by selecting a random run as a reference and compute all alignments against this map. The algorithm then determines a new reference run by selecting the run with the smallest sum of RT differences in the resulting run alignments. The iteration stops after a user-specified maximum number of iterations or when the reference run remains unchanged between two iterations. The software can also be configured to compute all possible RT alignment combinations (all possible pairs of maps, “exhaustive” option), but this can be very computationally expensive when there is a high number of maps to be compared. The alignments computed can then be used to predict the retention times for peptide ions in a specific sample where they were not identified.

Parameters

Alignment

- Map Alignment: alignment can be disabled. In this case the cross assignment step could be executed but without any retention time prediction

- Method : is the name of the method used to determine the reference run
    - Exhaustive: the exhaustive algorithm computes the distance between maps for each possible couple of maps and selects the map with the lowest sum of distances to be the reference map.

- Iterative: the algorithm iteratively selects the best reference run as described above.
    - Maximum number of iterations: this option is available only for the iterative method. This is a stop condition of the iterative algorithm, when the algorithm has reached its maximum number of iterations, it stops.

- Ignore alignment error : allow or not to continue quantitation even if there are some errors in alignment.

Feature mapping

- Feature mapping is needed to build the alignment scatter plot of the observed time difference between two runs. The method used could be:
    - Feature coordinates : the mapping is based on features RT and m/z coordinates taking into account a tolerance for each of these values.
        - m/z tolerance : m/z window used to match features between two compared maps.
    - Peptide Identity : the mapping is based on peptides identification (same sequence and same post-translational modifications)
- Time tolerance in seconds: time window used to match features/peptides between two compared maps.

Alignment smoothing

When features of two runs are matched, a trend can be extracted from the scatter plot by using a smoothing method.

- Smoothing method: you have the choice between three smoothing methods: LOESS, time window or landmark range:
    - LOESS : no specific parameter
    - Landmark range
        - Number of landmarks/window size: time of aligned map is corrected using median computed on windows containing a specified number of landmarks. The run is divided into windows of size the specified number of landmarks. You have to provide the number of landmarks by window. The smoothing method is applied considering the number of landmarks present in the window, and computes the median point for this window.
        - Sliding window overlap: overlap is used to compute the step to move the smoothing window forward to calculate a smoothing point for this new smoothing window. Overlap gives the percentage of overlapping between two consecutive windows. For example, if window size is 200 (seconds or landmarks depending on which smoothing method is selected) and overlap is 20%, the step forward = 200*((100-20)/100) = 160 seconds or landmarks, i.e. the smoothing window is moved forward by a step of 160, so two successive windows overlaps each other by a step of 40 seconds or landmarks corresponding to 20% of 200
    - Time window : time of aligned map is corrected using median in a time window.
        - Window size/time interval: You have to provide the time interval. This time interval corresponds to the window size in which time median will be computed.
        - Minimum number of landmarks: This allows you to specify the minimum number of landmarks a window must contain to compute a median on it, it is not significant to compute a median on less landmarks.

- Sliding window overlap: overlap is used to compute the step to move the smoothing window forward to calculate a smoothing point for this new smoothing window. Overlap gives the percentage of overlapping between two consecutive windows. For example, if window size is 200 (seconds or landmarks depending on which smoothing method is selected) and overlap is 20%, the step forward = 200*((100-20)/100) = 160 seconds or landmarks, i.e. the smoothing window is moved forward by a step of 160, so two successive windows overlap each other by a step of 40 seconds or landmarks corresponding to 20% of 200.

MoZ calibration

Delta moz (between experimental and theoretical moz) may change during a run. By specifying a moz calibration method, you can specify to use the recalibrated theoretical moz to cross assign instead of using the best feature moz. In both cases, a moz tolerance should be specified. see below

Cross Assignment

Proline uses a hybrid approach to retrieve intensity values for ions that were not identified. As indicated above, identified and quantified features are obtained by detecting chromatographic peaks in raw files without a-priori, using an identification-based deisotoping method. During the PSM assignment step, the identification data provides the monoisotopic mass and the charge state for the ion, guiding the deisotoping procedure to group together the detected chromatographic peaks. These grouped peaks are then removed from the list of peaks to be assigned, thereby reducing the data density when annotating subsequent chromatographic peaks during the cross-assignment step. Ions that were not identified in a run can then be sought out in this restricted list of detected peaks using their m/z and RT coordinates. The m/z value is the theoretical m/z value obtained following identification of this ion in another run. The RT value is predicted from the apex RT of the feature detected in the run providing the highest identification score. This RT prediction computation may involve the use of one or two alignment functions. Using these two coordinates, associated with user-defined m/z and RT tolerances, the algorithm seeks a corresponding signal among the chromatographic peaks that have not already been assigned to an identified ion. To avoid the propagation of erroneous cross-assignments between runs, an additional but optional control (named “use only confident features”) is applied to ensure that this peak is the monoisotope of a peptide ion with a charge state identical to the master feature one. This is done by fitting its observed isotope pattern to a theoretical one.

Parameters

- allow cross assignment between all runs or only between runs of the same groups: the user can completely disable the cross assignment or can control runs between which cross assignment is applied:
    - Between all runs
    - Within groups only: cross assignment is applied only between runs belonging to the same group.

Feature mapping

- m/z tolerance (ppm or dalton): when mapping features from two different maps of the map set, the delta m/z between features must be lower than the m/z tolerance to be considered as the same feature seen on two different maps.
- Time/RT tolerance (seconds): when mapping features from two different maps of the map set, delta time between features must be lower than the time tolerance to be considered as the same feature seen on two different maps.

Filtering/Correction

- Use only confident features: is applied to ensure that the cross-assigned peak is the monoisotope of a peptide ion with a charge state identical to the master feature one. This is done by fitting its observed isotope pattern to a theoretical one.
- Master feature intensity filter (optional): a filter can be applied to the map features to keep the best features (above threshold) to build the master map.

    - Two methods are available to filter features: the filter can be applied directly on intensity values (*Intensity method*) or it can be a proportion of the map median intensity (*Relative intensity method*).

    - intensity threshold/value: this provides the threshold for the filtering method. Only features above or below (depending on the *operator*) this threshold are considered for the master map building process.

If you choose Relative intensity for master feature filter type, the only possibility you have is percent, so features which intensities are beyond the relative intensity threshold in percentage of the median intensity are removed. If you choose Intensity for master feature filter type, you also have only one possibility at the moment of the intensity method: basic. Features which intensities are beyond the intensity threshold are removed and not considered for the master map building process.

Normalizing LC-MS maps

The comparison of LC-MS maps is confronted to another problem which is the variability of the MS signals measured by the instrument. This variability can be technical or biological. Technical variations between MS signals in two analyses can depend on the injected quantity of material, the reproducibility of the instrument configuration and also the software used for the signal processing. The observed systematic biases on the intensity measurements between two successive and similar analysis are mainly due to errors in the total amount of injected material in each case, or the LC-MS system instabilities that can cause variable performances during a series of analysis and thus a different response in MS signal for peptides having the same abundance. Data may not be used if the difference is too important. It is always recommended to do a quality control of the acquisition before considering any computational analysis. However, there are always biases in any analytic measurement but they can usually be fixed by normalizing the signals. Numerous normalization methods have been developed, each of them using a different mathematical approach (Christin, Bischoff et al. 2011). Methods are usually split in two categories, linear and non-linear calculation methods, and it has been demonstrated that linear methods can fix most of the biases (Callister, Barry et al. 2006). Three different linear methods have been implemented in Proline by calculating normalization factors as the ratio of the sum of the intensities, as the ratio of the median of the intensities, or as the ratio of the median of the intensities.

Sum of the intensities

How to calculate this factor:

1. For each map, sum the intensities of the features
2. The reference map is the median map
3. The normalization factor of a map = sum of the intensities of the reference map / sum of the intensities of the map

Median of the intensities

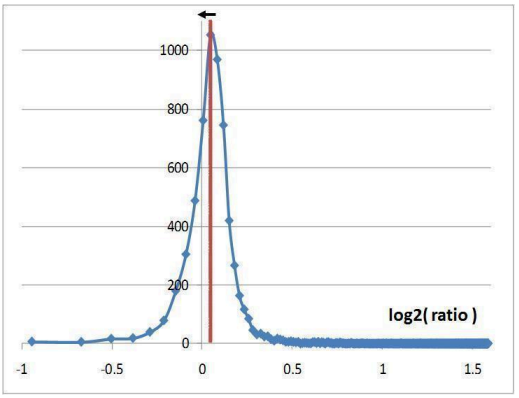
How to calculate this factor:

1. For each map, calculate the median of the intensities in the map
2. The reference map is the median map
3. The normalization factor of a map = median of the intensities of the reference map / median of the intensities of the map

Median of ratios

This last strategy was published in 2006 (Dieterle, Ross et al. 2006) and gives the best results. It consists in calculating the intensity ratios between two maps to be compared then set the normalization factor as the inverse value of the median of these ratios (cf. figure below). The procedure is the following:

1. For each map in a "map set", sum the intensities of the features
2. The reference map is the median map
3. For each feature of the master map, ratio = intensity of the feature in the reference map / intensity of the feature for this map
4. Normalization factor = median of these ratios



Distribution of the ratios transformed in log2 and calculated with the intensities of features observed in two LC-MS maps. The red line representing the median is slightly off-centered. The normalization factor is equal to the inverse of this median value. The normalization process will refocus the ratio distribution on 0 which is represented by the black arrow

Proline makes this normalization process for each match with the reference map and has a normalization factor for each map, independently of the choice of the algorithm. The normalization factor for the reference map is equal to 1.

Isobaric labeling quantitation workflow

Common isobaric labeling quantitation methods, such as TMT 6 to 18plex or iTRAQ 4 or 8plex, are supported in Proline. Adding a missing xPlex method can be easily done. This technique may be used to compare multiple samples in a single run. This is done by reading all reporters ion peak intensity each corresponding to a different sample.

Data is red from the MS2 spectrum stored in mzDB file (see available converters paragraph). These quantitation results are stored in the Proline database and aggregated to define peptide ion abundances which are also stored in the Proline database, making them available for rapid data visualization and further post-processing.

Finally, in the same way as for label free quantitation, peptide ion measurements can be **summarized as peptides and then protein abundances** using different computational methods.

The user can opt to perform additional operations such as excluding peptides or ions based on their characteristics (precursor intensity fraction, missed cleavages, variable modifications, sequence specificity, etc.). A purity correction matrix can also be provided to correct the reporter ion abundances. Indeed some manufacturers provide such matrices indicating the percentages of each reporter ion that have masses differing by +/- n Da from the nominal reporter ion mass due to isotopic variants.

Post-processing of quantitative results

This procedure is used to compute peptide and protein abundances. It is available for LC-MS and isobaric labeling quantitative results. Several filters can also be set to increase the quality of quantitative results.

Here is the description of the parameters that can be modified by the user.

PSM filters

For isobaric labeling only

- Discard PSM on Precursor Ion Fraction value. “The PIF is defined as the fraction of ion current in the isolation window that is due to the targeted precursor ion and therefore ranges between 0 and 1. The PIF is determined for each targeted precursor ion and each tandem mass spectrum based on the closest full scan, which could be the previous or the consecutive one” (J. Proteome Res. 2011, 10, 4, 1785–1793).

Purity correction matrix

For isobaric labelling only

In isobaric labelling quantitation, isobaric tags are used in samples to compare their abundances. Due to isotopic contamination in tags, reporter ion peaks will contribute to those of neighboring reporter ions. A purity correction matrix is provided with the labelling kit by the manufacturer. This matrix can be used to correct the abundance of the reporter ions .

Peptide filters

- **Peptides selection method:** define which peptides are to be considered for quantitation.
    - **All peptides:** all peptides are taken into account at this step. Further filters may invalidate some peptides.
    - **Specific peptides**: peptides shared between different protein sets are discarded for protein set abundance calculation and statistical analysis.
    - **Razor and specific:** In addition to specific peptides, quantitation of shared peptides are considered for the “best” ProteinsSet. The best is actually based on the score.
- **Discard missed cleaved peptides**: if checked, peptides containing missed cleavages are discarded from the statistical analysis. It has to be noted that perfect tryptic peptides whose sequence is included in an observed missed cleaved peptide are also discarded if this option is enabled.
- **Discard modified peptides**: if checked, peptides containing specific modifications are discarded for protein set abundance calculation and the statistical analysis.
    - only those modified peptides or also non-modified peptides whose sequence is the same as these peptides may be discarded depending on the chosen filtering method.

Peptide and protein common parameters

- Normalization: the normalization factors are computed as the median of the ratio distributions between each run and a run of reference. A similar procedure is used for the normalization of LC-MS features.

Summarize PSM (reporter ions) into peptide ions abundance

For isobaric labelling only.

To calculate peptide ions abundance, associated reporter ions abundances can be **summed** or the **median** can be computed.

Summarize peptides ions into peptide abundance

To calculate peptide abundance, associated peptide ion abundances can be summed or the best ion is used. To choose the best peptide ion following rules are defined :

- Select the peptide ions identified and quantified in the maximum of quant channels(runs)
- If more than one, select peptide ions with the maximum of PSMs count
- If more than one, choose the peptide ion with the higher abundance.

Summarize peptides abundances into protein abundance

Peptide abundances can be summarized into protein abundances using several mathematical methods:

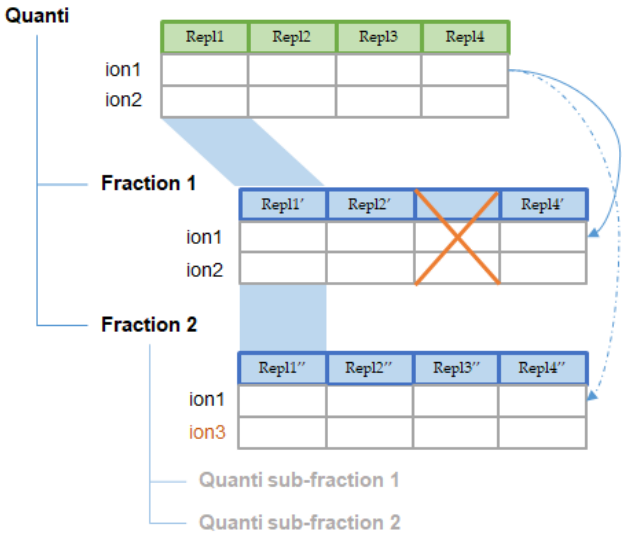
- **sum**: for each quantitative channel (raw file) the sum of observed peptides abundances is computed
- **mean**: for each quantitative channel (raw file) the mean of observed peptides abundances is computed
- **mean of TOP3** peptides: same procedure but applied on the 3 most abundant peptides. Peptides are sorted by descending median abundances (computed across all compared samples for peptide). Then the 3 first peptides are kept.
- **median**: for each quantitative channel, the median of observed peptides abundances is computed
- **median profile**: a matrix of peptide abundance ratios is first computed (rows correspond to peptides and columns to quantitative channels). The median of these ratios is then computed for each column. The relative values are then converted back into absolute values using a scaling factor. This factor is computed as the maximum value from the **means of TOP3 peptides** abundances.
- **normalized median profile**: matrix of peptide abundance ratios is first computed (rows correspond to peptides and columns to quantitative channels). This matrix is then normalized and then summarized using the median method described above. The obtained median abundances are then adjusted by using a scaling factor. This factor is computed as the maximum value from the **means of TOP3 peptides** abundances.
- **Median Ratio Fitting**: MRF computes a matrix of abundance ratios calculated between any two runs from ion abundances for each protein. For each pair-wise ratio, the median of the ion ratios is then calculated and used to represent the protein ratio between these two runs. A least-squares regression is then performed to approximate the relative abundance of the protein in each run in the dataset. This abundance is finally rescaled to the sum of the ion abundances across runs.

Aggregation of quantitative results

Two or more label free quantitations can be combined such that an ion quantified in multiple aggregated quantitations is represented only once in the aggregation result. The abundance of this ion is a combination of its

abundance measured in the different aggregated quantitations. This could be useful to combine for example quantitation of fractions into a single quantitation result.
For Isobaric labelling quantitations, the same process is used but combining reporter ions abundances. These combined PSMs will then be summarized to get peptide ions abundances.

The experimental design of the aggregation is based on the experimental design of the aggregated quantitation: the number of group/condition and the number of replicates per condition remains the same. However, the user can modify the correspondence between the groups and replicates if needed. In the following example the abundance of ion1 in the aggregated quantitation (in green) is based on the quantitation of this same ion in "Fraction 1" and "Fraction 2". Since ion2 is quantified only in "Fraction 1", its abundance values in the aggregated quantitation are the same as the abundances measured in "Fraction 1".



In this simple example, the correspondence between experimental designs is such that the abundance of ion1 in the replicate Repl1 is based on the measured abundance of ion1 in Repl1' in "Fraction 1" and Repl1'' in "Fraction 2". This could be modified by the user to take into account differences in the replicates order in aggregated quantitations or to account for the absence of a replicate (see for example replicate 3 in "Fraction 1").

In the current version, the abundance at the aggregation level is the sum of the abundances in aggregated quantitations.

Identification Summary Export

When exporting a whole Identification Summary in an excel file, the following sheets may be generated:

- *Search settings and info* : Contains information on project and search settings parameters
- *Import and filters* : Summary of used parameters during import, filtering and validation process
- *Protein sets* : List of all Protein Sets, valid or invalidated (configurable in custom option) during Protein Sets Filtering. Some columns description :
    - *#sequences* (*#specific sequences*) : number of different peptide sequences identifying the Protein Set (specific : which does not identify any other valid Protein Set)

- *#peptides* (*#peptides*) : number of different peptide (sequence + PTM) identifying the Protein Set (specific : which does not identify any other valid Protein Set)
- *#spectral_count* (*specific_spectral_count*) : number of different peptide spectrum matches identifying the Protein Set (specific : which does not identify any other valid Protein Set)

- *Protein matches in protein set* : list of Protein Matches in each Protein Set. A same Protein Match could thus appear a few times if it belongs to different Protein Sets. (same column as protein set)
- *Best PSM from protein sets* : List of best peptide spectrum matches (a single PSM per peptide is listed) for each Protein Set. Some columns description :
  - *#psm_prot_sets* : number of Valid Protein Sets identified by this PSM.
  - *#psm_prot_matches* : number of Protein Match, which belong to at least 1 valid Protein Set, identified by this PSM.
  - #psm_db_prot_matches : number of Protein Match, validated or not, identified by this PSM. This is equivalent to the number of proteins in fasta files containing the PSM.
- *All PSMs from protein sets* : List of all peptide spectrum matches for each Protein Set. (same column as best PSM from protein sets)
- PTM Cluster (filled only if "identify Modifications sites" has been run): All clusters (protein-set of localised ptms) identified in this identification summary.
- *Dataset statistics* : Some statistical values for the exported Identification Summary : number of Protein Set, modified peptides …

Advanced features

Allowing multiple imports

By default, Proline does not allow importing multiple results files at the same time. This was due to the obsolete Peptide database which has been removed in Proline version 2.0. Nevertheless, if multiple results files could now be imported at the same time there is still a restriction at project level. Indeed, result files should be inserted sequentially in the same project. In addition, depending on Proline Server configuration, there could not be as many import threads as wanted or as the number of projects.

This is why a configuration allows the user (administrator) to specify groups of projects which will belong to the same "import thread". For instance, if Proline Server enables 3 threads for import services, you can specify 2 groups of projects:

- Import Thread 1 : project 1, project 3, project 5
- Import Thread 2 : project 6, project 7
- Import Thread 3 : all other project (and new projects)

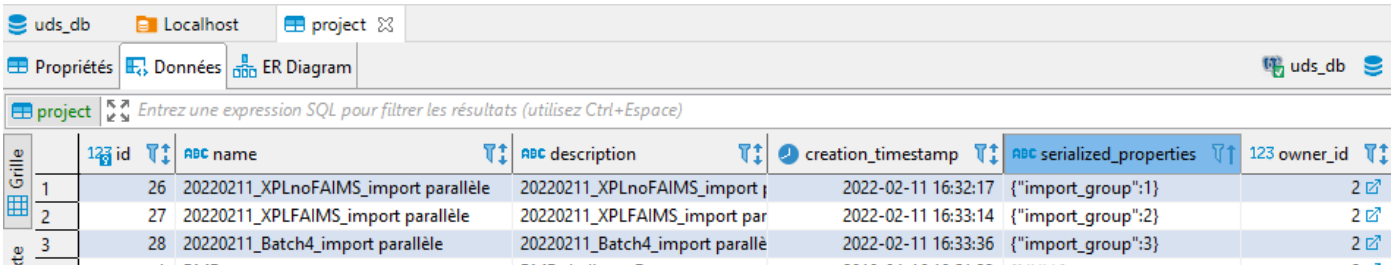
This means that for projects 1, 3 and 5 imports will be sequential but in parallel of import to project 6 or 7…

If you specify more groups than allowed thread, some project groups will be grouped into a single one.

To define a project group, you should add the following string to *project.serialized_properties* in uds_db database. Currently this could be done only with PGAdmin (or any database server administration tool) :

{"import_group":"1"}

In the following example, the user has defined 3 specific groups, and the default group is always defined for all other projects.



	id	name	description	creation_timestamp	serialized_properties	owner_id
1	26	20220211_XPLnoFAIMS_import parallèle	20220211_XPLnoFAIMS_import	2022-02-11 16:32:17	{"import_group":1}	2
2	27	20220211_XPLFAIMS_import parallèle	20220211_XPLFAIMS_import par	2022-02-11 16:33:14	{"import_group":2}	2
3	28	20220211_Batch4_import parallèle	20220211_Batch4_import parallè	2022-02-11 16:33:36	{"import_group":3}	2

Available converters to mzDB

Raw2mzdb

Raw file conversion to mzDB

This converter, named raw2mzdb, is based on ProteoWizard, ensuring compatibility with a wide range of instrument vendors.

Installation

1. get the zip archive on Proline download section
2. install of MSFileReader from Thermo (here, will install all necessary c++ redistribuables)
3. ensure your regional settings parameters are '.' for the decimal symbol and ',' for the list separator

Use case procedure

Open a command line window in the directory containing raw2mzdb.exe

Enter:

raw2mzdb.exe -i <rawfilename> -o <outputfilename>

By default, the raw file will be converted in the "fitted" mode for the MS1 (MS2 is often in centroid mode and can not be converted in fitted mode). If the MS2 (or superior) are acquired in high resolution (i.e in profile mode), you could specify that you want to convert several MSs in the required mode: **raw2mzdb.exe -i <rawfilename> -o <outputfilename> -f 1-2** will try to convert MS1 to MS2 in fitted mode.

There are two other available conversion modes:

1. "*profile*", the command line is then: **raw2mzdb.exe -i <rawfilename> -o <outputfilename> -p 1** (means you want profile mode for MS1, others MS will be stored as they were stored in the raw file)
2. "*centroid*" : **raw2mzdb.exe -i <rawfilename> -o <outputfilename> -c 1** (means you want centroid mode for MS1, others MS will be stored as they were stored in the raw file)

Warning: for Thermo raw file, -n option is mandatory. Anyway, it is recommended to use **mzDBConverter**. See below

mzDBConverter

New converter based on manufacturer libraries. It currently supports Thermo raw files and Bruker TimsTof files.

This converter is distributed with the automatic installer or can be downloaded from proline download page.
To display a complete list of option type :

- mzdbConverter thermo --help
- mzdbConverter brucker --help

mzdbConverter.bat thermo -i <InputFilename> -o <outputFilename> -c 2,3 -f 1

This will convert Thermo <inputFilename> raw file info <outputFilename> mzdb file, using centroid mode for MS2 and MS3 and fitted mode for MS1 spectra.

Proline Studio

How to

Note: Read the Concepts & Principles documentation to understand main concepts and algorithms used in Proline.

- List of Abbreviations

Creation/Deletion

- Open a session and access to my projects
- Create a new project
- Create a Dataset
- Import a Search Result
- Delete Data
- Connection Management

Display

- Display MSQueries, Peptides/PSM or Proteins of a Search Result
- Display MSQueries, PSM, Peptides, Protein Sets, PTM Protein Sites or Adjacency Matrices of an Identification Summary
- Display Modification Sites or Clusters of a Modification Dataset
- Display Search Result & Identification Summary Properties
- Display Spectral Counts
- Display Quantitation results
- Create and Save a User Window
- Frame Toolbars Functionalities
- Filter tables
- Search tables
- Graphics : Scatter Plot / Histogram
- Statistical Reports (MSDiag)
- MS Files (MzScope)
- General Settings

Save, import and export

- Import Search Result file (Mascot/OMSSA/X!Tandem/MzIdent/MaxQuant)
- Export data
- Generate Spectrum Matches

Algorithm and other operation

- Validate a Search Result
- Filter Protein Sets
- Change Typical Protein of a Protein Set
- Combine datasets

- Data Analyzer
- Calculator
- Update Spectrum using Peaklist software
- Identify Modification Sites/Cluster
- Annotate and Edit Modification Clusters
- Administration

Quantitation

- Spectral Count
- Label Free Quantitation
- Isobaric Labeling Quantitation
- Quantitations Aggregation
- Compute Post Processing on abundances

List of Abbreviations

Calc. Mass: Calculated Mass

Delta MoZ: Delta Mass to Charge Ratio

Exp. MoZ: Experimental Mass to Charge Ratio

Ion Parent Int.: Ion Parent Intensity

Missed Cl.: Missed Cleavage

Modification D. Mass: Modification Delta Mass

Modification Loc.: Modification Location

Next AA: Next Amino-Acid

Prev. AA: Previous Amino-Acid

Protein Loc.: Protein Location of the Modification

Protein S. Matches: Protein Set Matches

PSM: Peptide Spectrum Match

PTM: Post Translational Modification

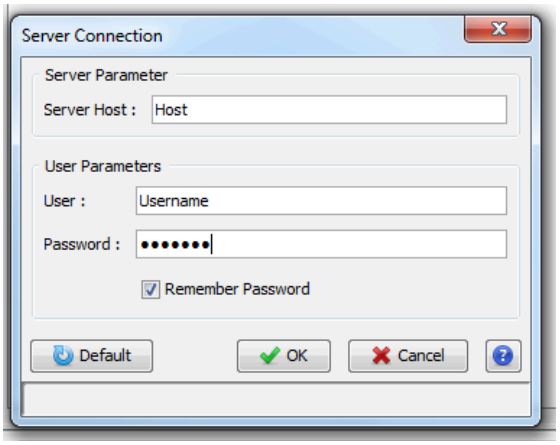
PTM D. Mass: PTM Delta Mass

RT: Retention Time

SC: Spectral Counting

Server Connection

When you start Proline Studio for the first time, the Server Connection Dialog is automatically displayed.

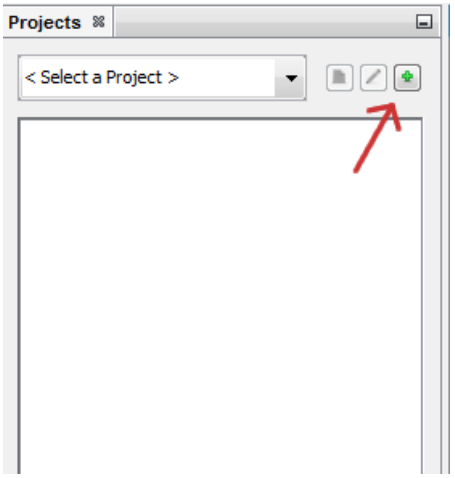


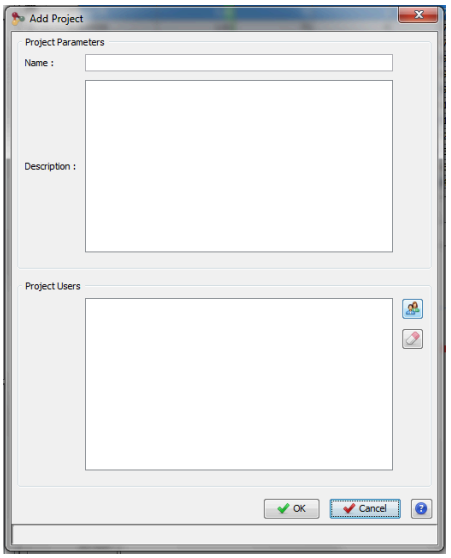
You must fill the following fields:

- **Server Host**: this information must be asked to your IT Administrator. It corresponds to the Proline server *name*
- **User**: your username (an account must have been previously created by the IT Administrator).
- **Password**: password corresponding to your account (username).

If the field “Remember Password” is checked, the password is saved for future use. Server connection dialog continues to open with Proline Studio, the user though does not need to fill in his password, unless the last one is changed after his last login.

Create a New Project





To create a Project, click on **“+“** button at the right of the Project Combobox. The Add Project Dialog opens.

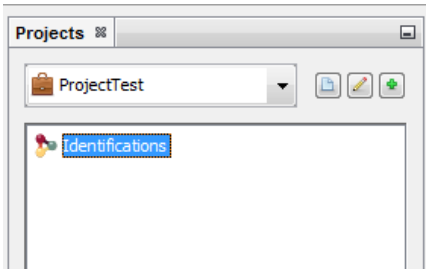
Fill the following fields:

- **Name**: name of your project

- **Description**: description of your project

You can specify other people to share this new project with them. Then click on OK Button

Creation of a Project can take a few seconds. During its creation, the Project is displayed grayed with a small hourglass over it.



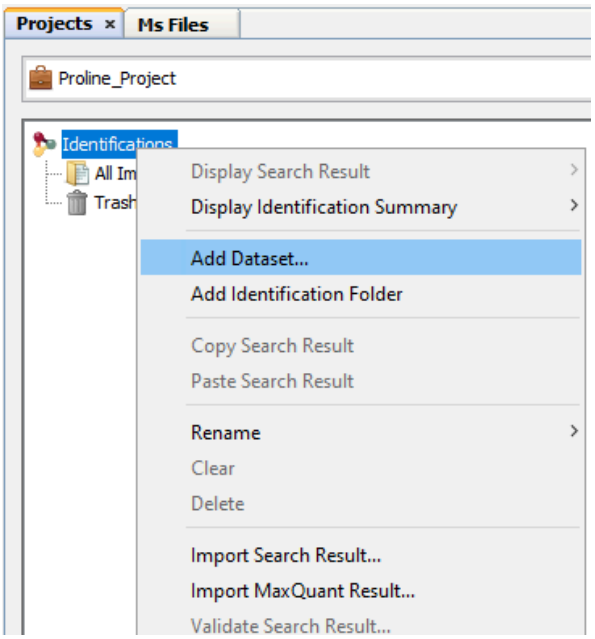
Create a Dataset

In the Identification tree, you can create a Dataset to group your data

To create a Dataset:

- right click on Identifications or on a Dataset to display the popup.

- click on the menu “Add Dataset…”

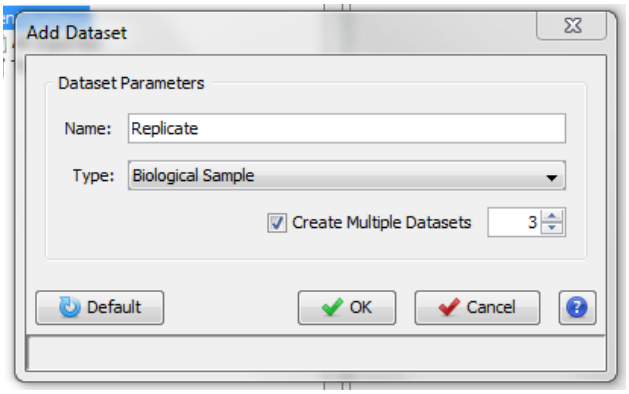


On the dialog opened:

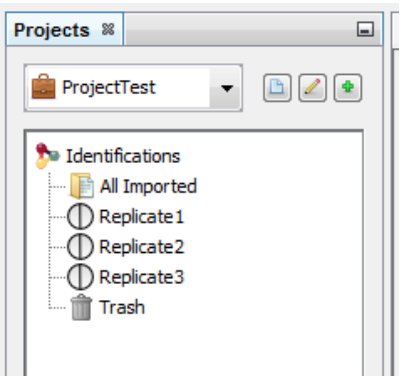
- fill the name of the Dataset

- choose the type of the Dataset

- optional: click on “Create Multiple Datasets” and select the number of datasets you want to create



Let's see the result of the creation of 3 datasets named “Replicate”:



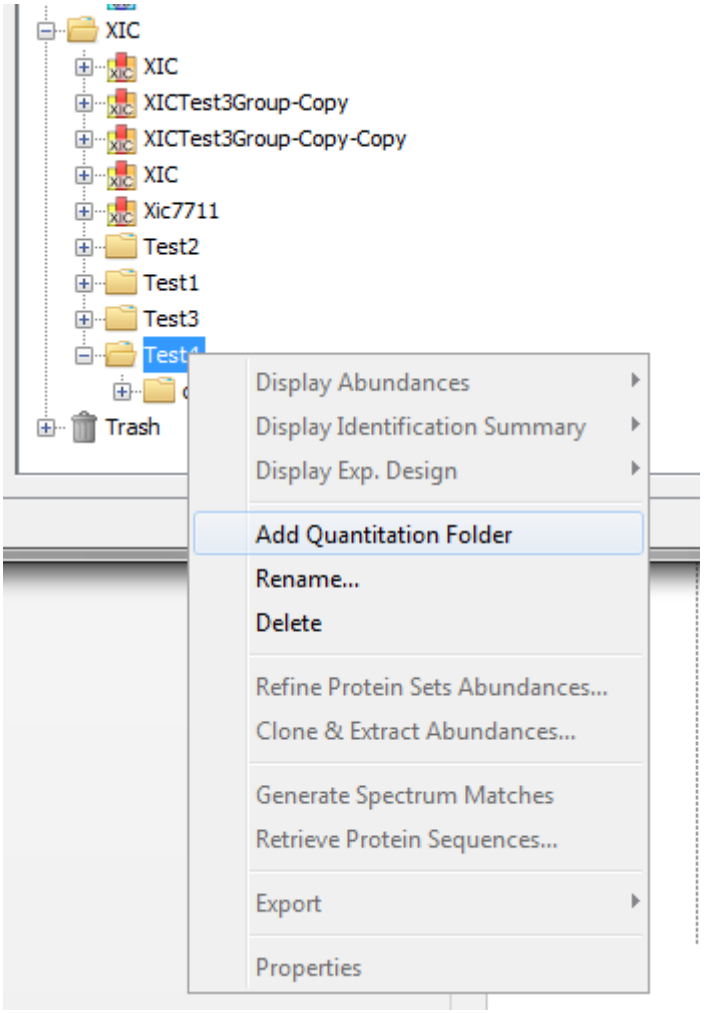
Create a Folder

In both Identification and Quantitation tree, you can create Folders to organize your data

To create a Folder :

- right click on Identifications, Quantitations or on a Folder to display the popup.

- click on the menu “Add Identification Folder…” or “Add Quantitation Folder…”



Import a Search Result

See Concept & Principle section

Import Mascot/X!Tandem/OMSSA/MzIdentML

There are two possibilities to import Search Results:

- import multiple Search Results in “All Imported” and put them later in different datasets.

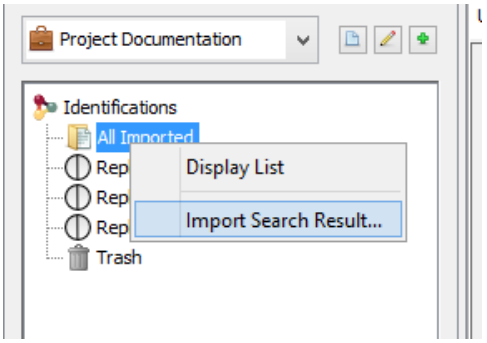
- import directly a Search Result in a dataset.

Import in "All Imported"

To import in “All Imported”:

- right click on “All Imported” to show the popup

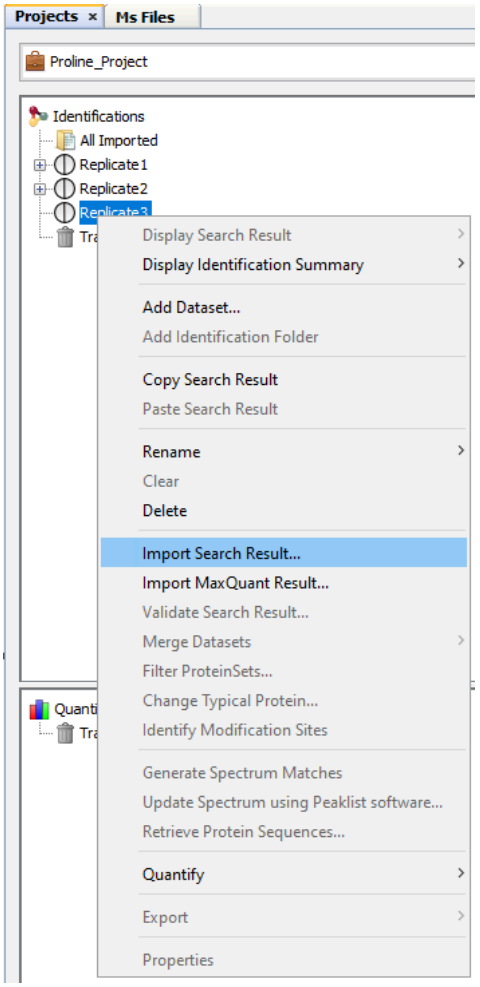
- click on the menu “Import Search Result…”



Import directly in a Dataset

It is possible to import Search Results directly in a Dataset. Even in this case, Search Results are available in “All Imported”.

To import a Search Result in a Dataset, right click on a dataset and then click on “Import Search Result…” menu. Same dialog and parameters as in “Import in “All Imported”” above will be displayed.



Import Search Result Parameters

In the Import Search Results Dialog:

- select the file(s) you want to import thanks to the file button (the Parser will be automatically selected according to the type of file selected)

- select the different parameters (see description below)

- click on OK button

Note 1: You can only browse the files accessible from the server according to the configuration done by your IT Administrator. Ask him if your files are not reachable. (Look for "*Setting up Mount-points"* paragraph in Installation & Setup page).

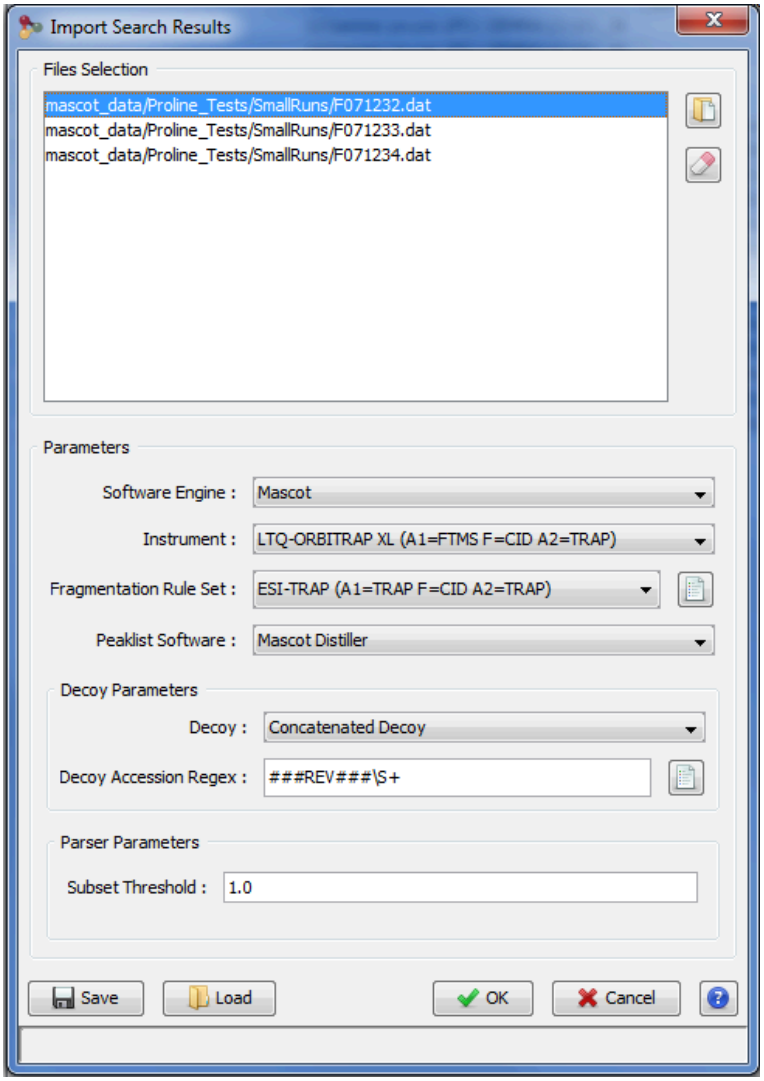
Note 2: Proline is able to import OMSSA files compressed with BZip2.

Parameters description:

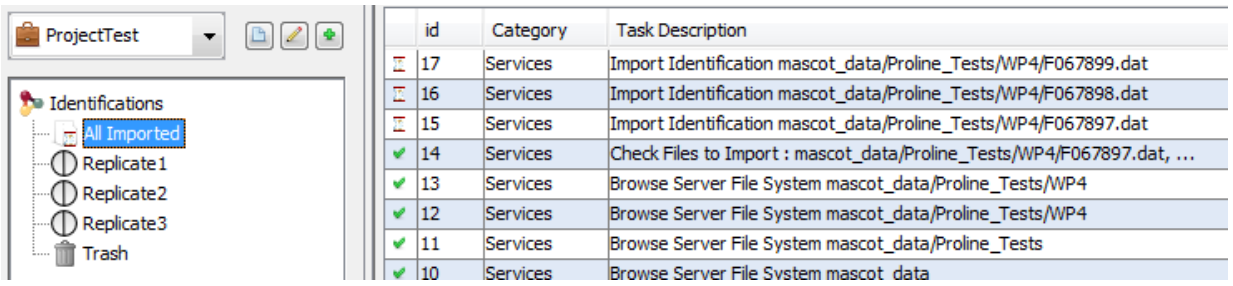
- Software Engine: the software which generated your Result File (this parameter will be automatically set when files are selected or you can select it)
- Instrument: mass-spectrometer (with specific configuration) used for sample analysis
- Fragmentation Rule Set: The fragmentation rules specified in software. by clicking on  you could visualize all rules for a specific rule set. This is necessary to generate spectrum matches.
- Peaklist Software: the software used for the peaklist creation (mandatory for LCMS XIC quantitation)
- Decoy Strategy: The type of decoy search which was performed.
    - "No Decoy": if the search was performed against a target database only.
    - "Concatenated Decoy": if target and decoy sequences were merged into a single database.
    - "Software Engine Decoy": if the decoy sequences were generated on-the-fly by your search engine.
    - Decoy Accession Regex: for concatenated searches only. Select the rule to apply for the discrimination of target and decoy protein matches.
- Parser Parameters: according to your Software Engine, some extra-parameters are displayed:
    - Mascot:
        - Subset Threshold: the percentage of score between a given protein match and the master protein match (superset). Protein matches with a relative score lower than
          `Master_protein_score * (1-subset threshold)`
          won't be imported.
    - Omssa:
        - Usermods file path: an XML file containing the definitions for each user defined PTM used in the OMSSA search.
        - PTM Composition file path: a text file containing the chemical composition for each user defined PTM. This is required for PTMs not already imported in another Search Result. The format is the following:
          `PTM name=<PTM composition> (one per line).`

          Example:
          `Acetyl peptide N-term=H(-6) C(-7) O(-1)`
    - X!Tandem:

- Protein parsing rule: Specify the rule (regular expression) to be used to retrieve protein accession from protein info. As an example, to use the Uniprot *Entry Name* as protein accession, use the following rule: `..\|[^|]*\|([^ ]*)`

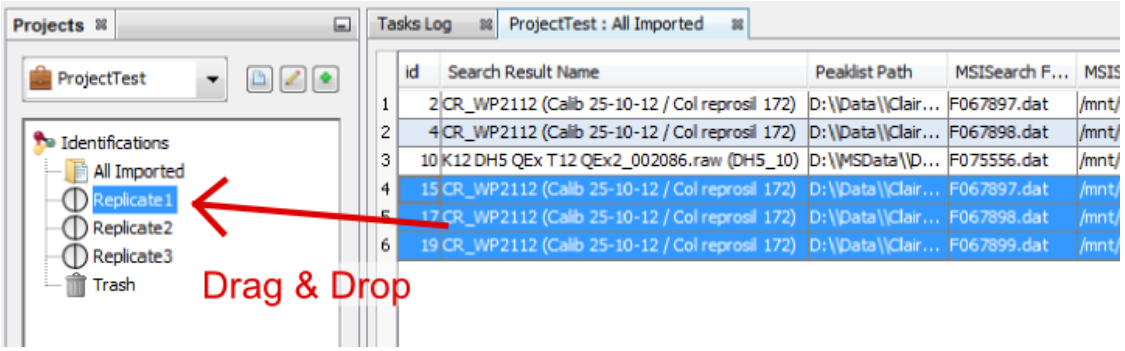


Importing a Search Result can take some time. While the import is not finished, the "All Imported" or "selected dataset" is shown grayed with an hourglass and you can follow the imports in the Tasks Log Window (Menu Window > Tasks Log to show it).



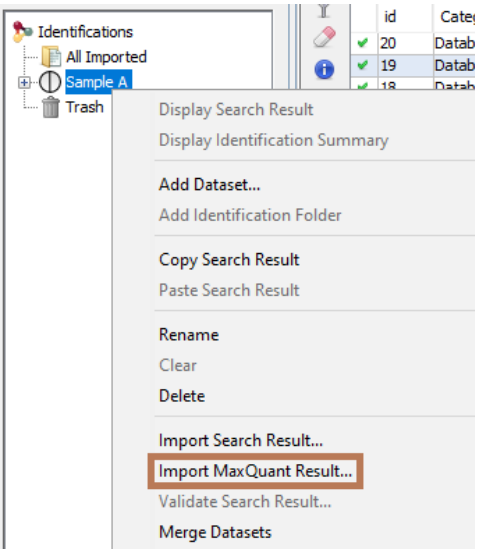
To show all the Search Results imported, double click on "All Imported", or right click to popup the contextual menu and select "Display List"

If needed, from the All Imported window, you can drag and drop one or multiple Search Result to an existing dataset.



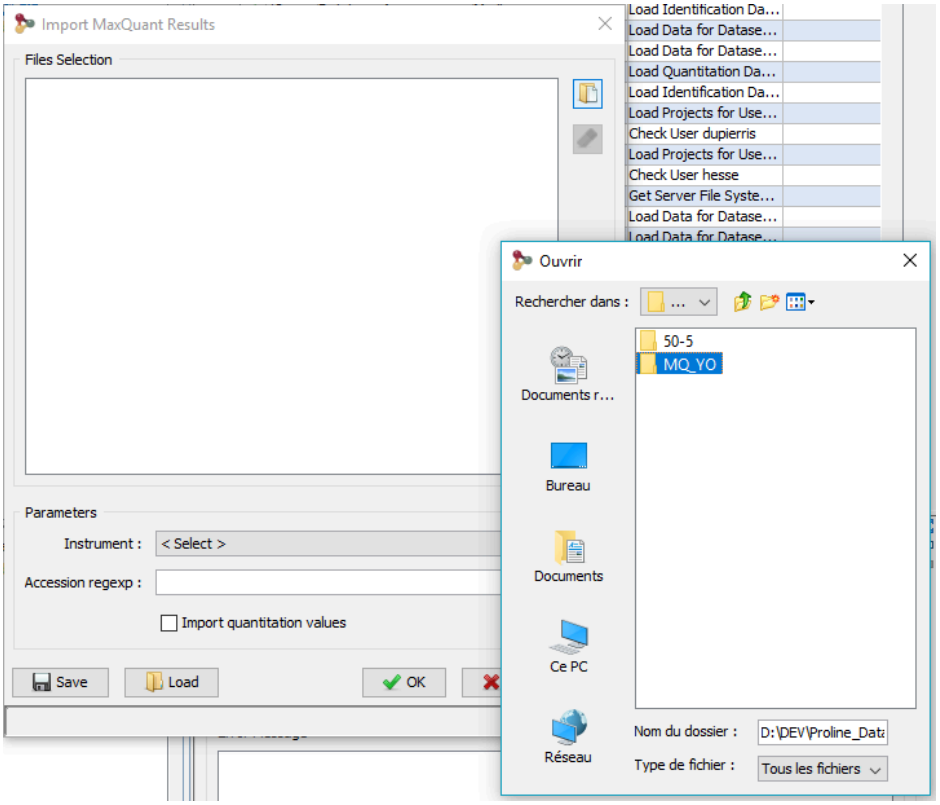
Import MaxQuant result

To import a MaxQuant Search Result, right click on a dataset and then select “Import MaxQuant”



Note 1: MaxQuant import will generate a dataset hierarchy with the result from the different acquisition.

The following dialog will be displayed



- select the directory containing the files generated by MaxQuant. This folder should look like:

<root_folder>\mqpar.xml
<root_folder>\combined\txt\summary.txt
<root_folder>\combined\txt\proteinGroups.txt
<root_folder>\combined\txt\parameters.txt
<root_folder>\combined\txt\msmsScans.txt
<root_folder>\combined\txt\msms.txt

- select the Instrument: mass-spectrometer used for sample analysis different parameters

- specify, if needed, the regular expression to extract protein accessions from MaxQuant protein ids.

- you can choose to import also quantitative data

- click on OK button

Delete Data

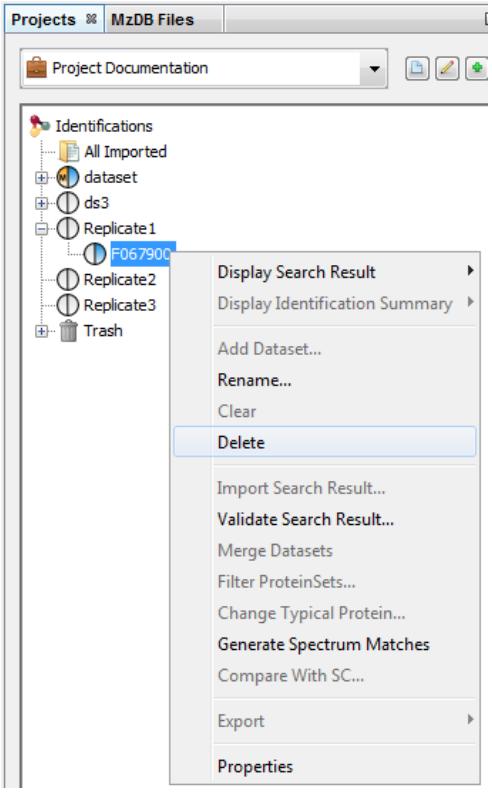
You can delete Search Results, Identification Summaries and Datasets in the data tree. You can also delete XIC or Spectral Counts in the quantitation tree.

Delete the Datasets (identification or quantitation…) **from the tree view** (Search Result always accessible from "All Imported" view…).

There are two ways to delete data: use the contextual popup or drag and drop data to the Trash.

Delete Data from the contextual popup

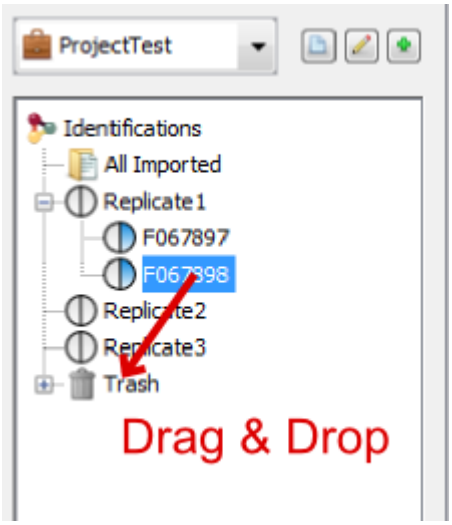
Select the data you want to delete, right-click to open the contextual menu and click on delete menu.



The selected data is put in the Trash. So it is possible to restore it while the Trash has not been emptied.

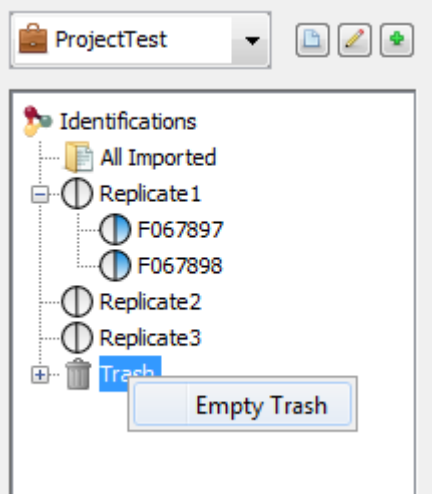
Delete Data by Drag and Drop

Select the data you want to delete and drag it to the Trash. It is possible to restore data while the Trash has not been emptied



Empty the Trash

To empty the Trash, you have to Right click on it and select the “Empty Trash” menu.



A confirmation dialog is displayed and if accepted Dataset will be removed from the Trash.

Search Results are not completely removed, you can retrieve them from the “All Imported” window.

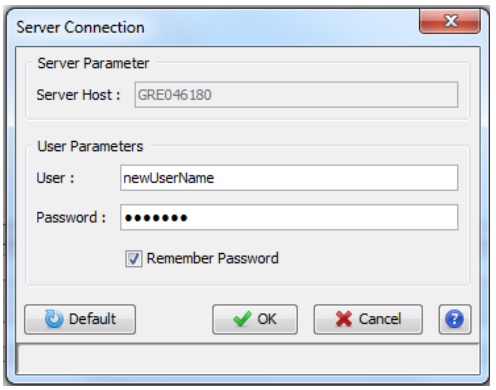
Delete a Project

It is not possible to delete a Project by yourself. If you need to do it, ask your IT Administrator.

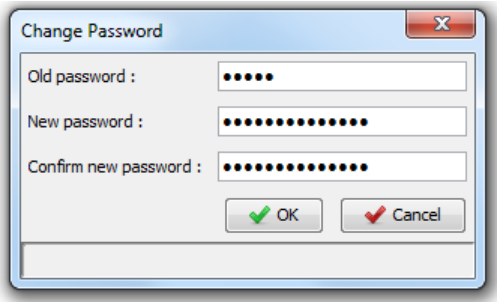
Connection Management

Once user is connected (see Server Connection), it is possible to:

- Reconnect with a different login



- Change password



Display MS Queries, Peptides/PSM or Proteins of a Search Result

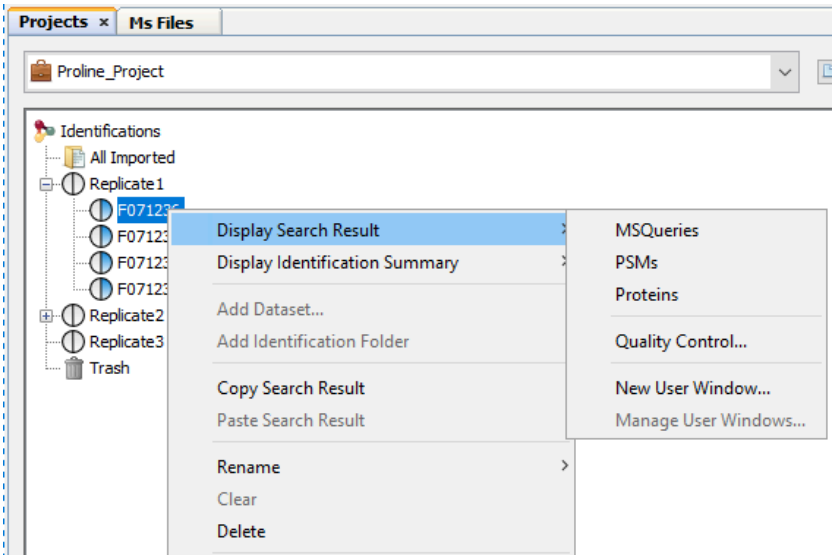
All information, validated or not, can be accessible from this menu. Indeed, Search Result contains all data imported from a result file without any validation consideration.

Functionality Access

To display data of a Search Result:

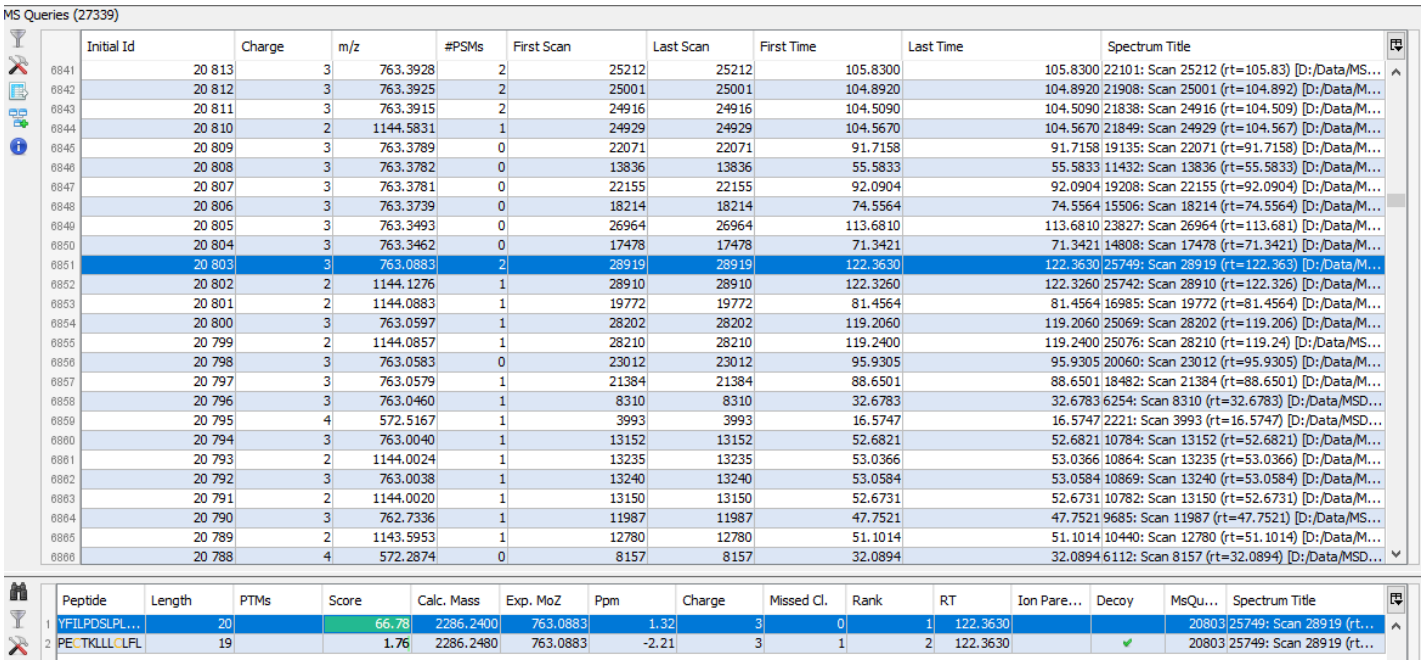
- right click on a Search Result

- click on the menu “Display Search Result >” and on the sub-menu “MSQueries” or “PSM” or “Proteins”



MSQueries Window

If you click on MSQueries sub-menu, you obtain this window:



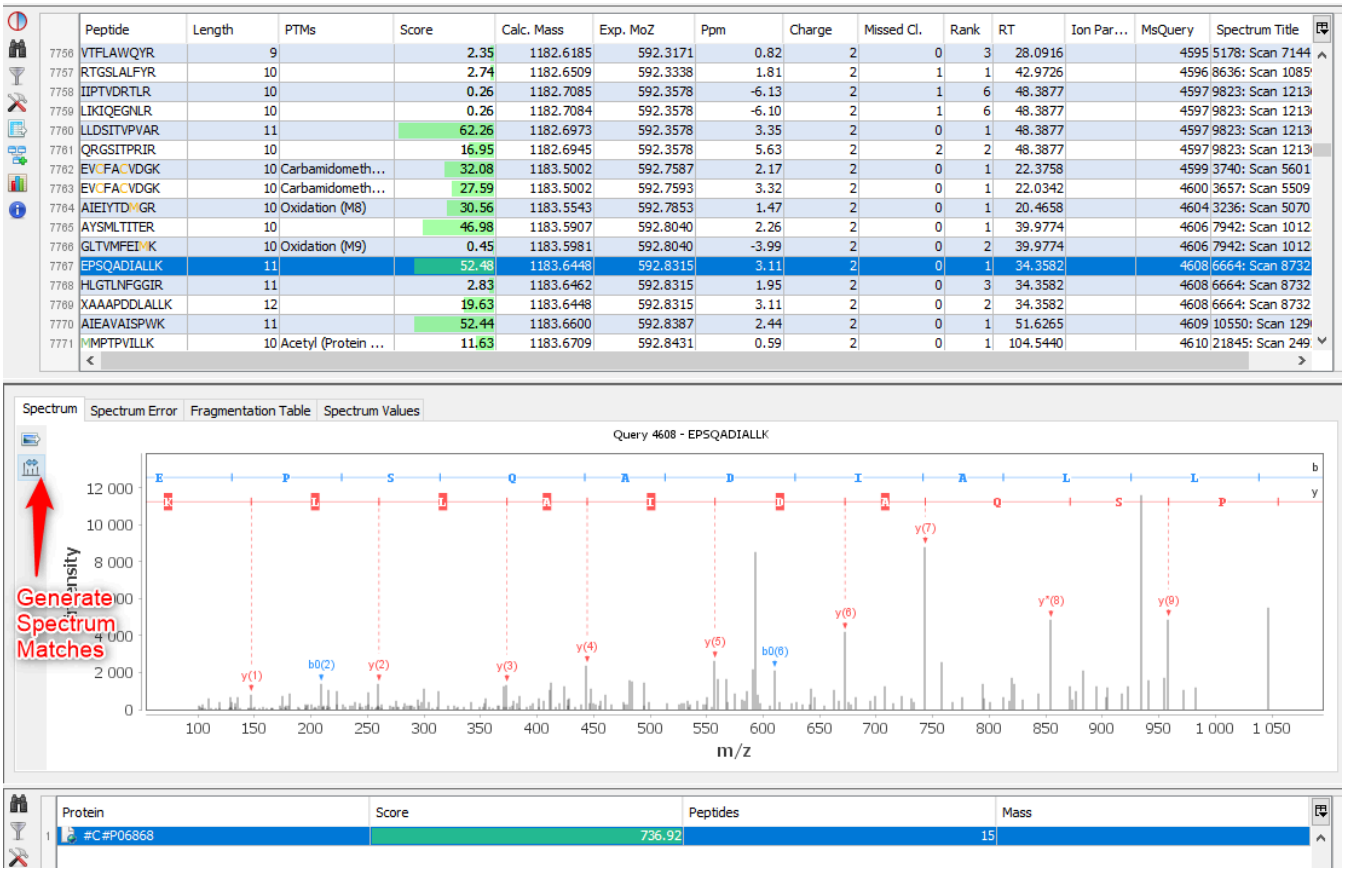
Upper View: list of MSQueries. Some columns may not be (correctly) filled if the Peaklist software were not correctly specified during import. It is possible to change this information using '[Update Spectrum ..]'

Bottom Window: list of all Peptides linked to the current selected MSQuery.

Note: Abbreviations used are listed [here]

PSMs (Peptides) Window

If you click on PSMs sub-menu, you obtain this window:



Upper View: list of all Peptide Spectrum Matches

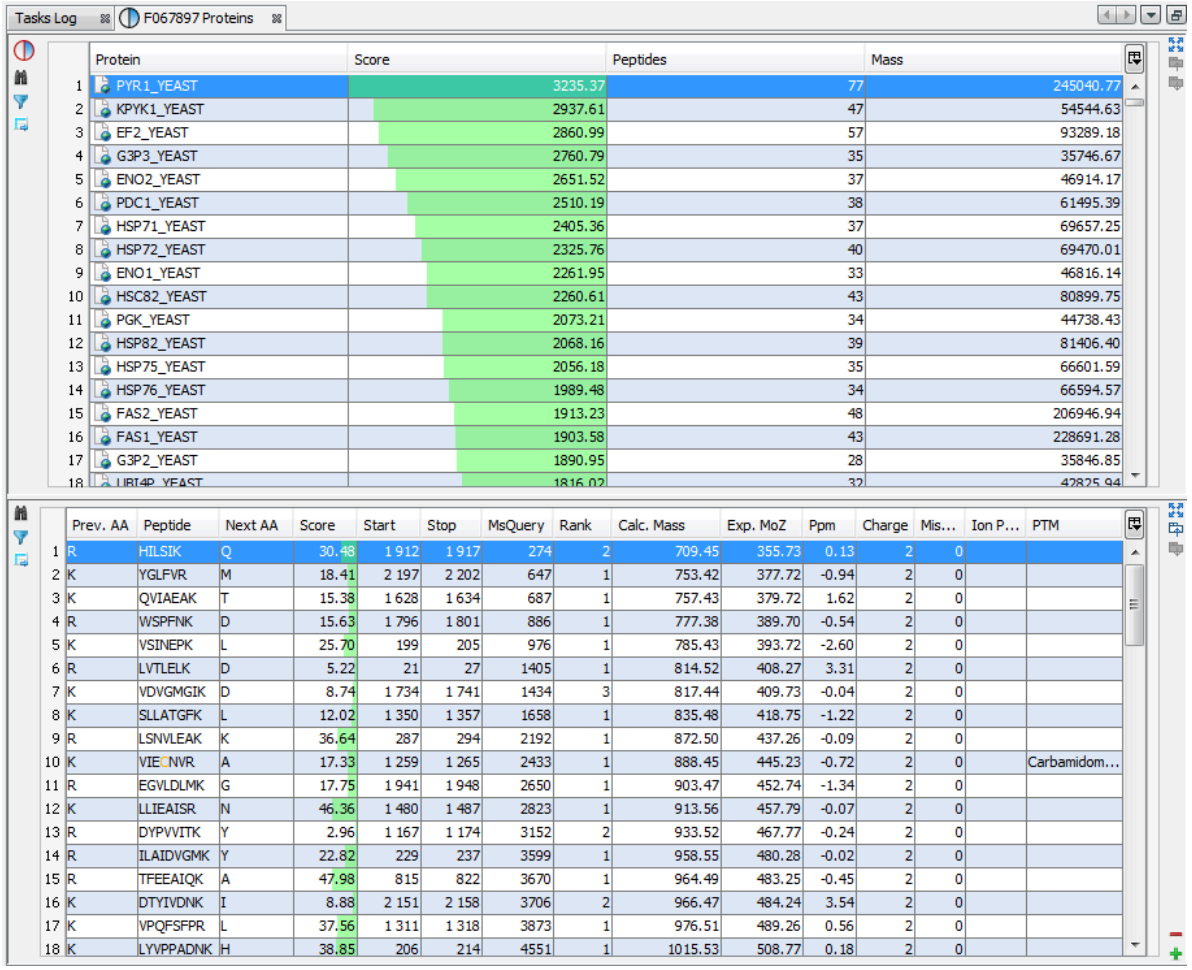
Middle View: Spectrum, Spectrum Error, Spectrum Values and Fragmentation Table of the selected PSM. If no annotation is displayed, you can generate Spectrum Matches by clicking on the according button

Bottom Window: list of all Proteins identified by the currently selected Peptide.

Note: Abbreviations used are listed [here]

Proteins Window

If you click on Proteins sub-menu, you obtain this window:



Upper View: list of all Proteins

Bottom View: list of all Peptides identifying the selected Protein.

Note: Abbreviations used are listed here

Display MS Queries, PSM, Peptides, Protein Sets or Adjacency Matrices of an Identification Summary

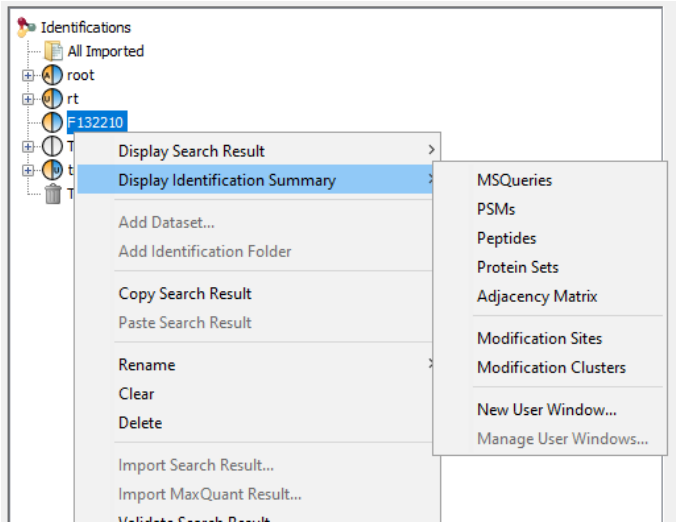
From this menu, all displayed information is Identification Summary data, which has been validated according to user specified rules. To view the raw information as defined at import, use the Search Result sub menu.

Functionality Access

To display data of an Identification Summary:

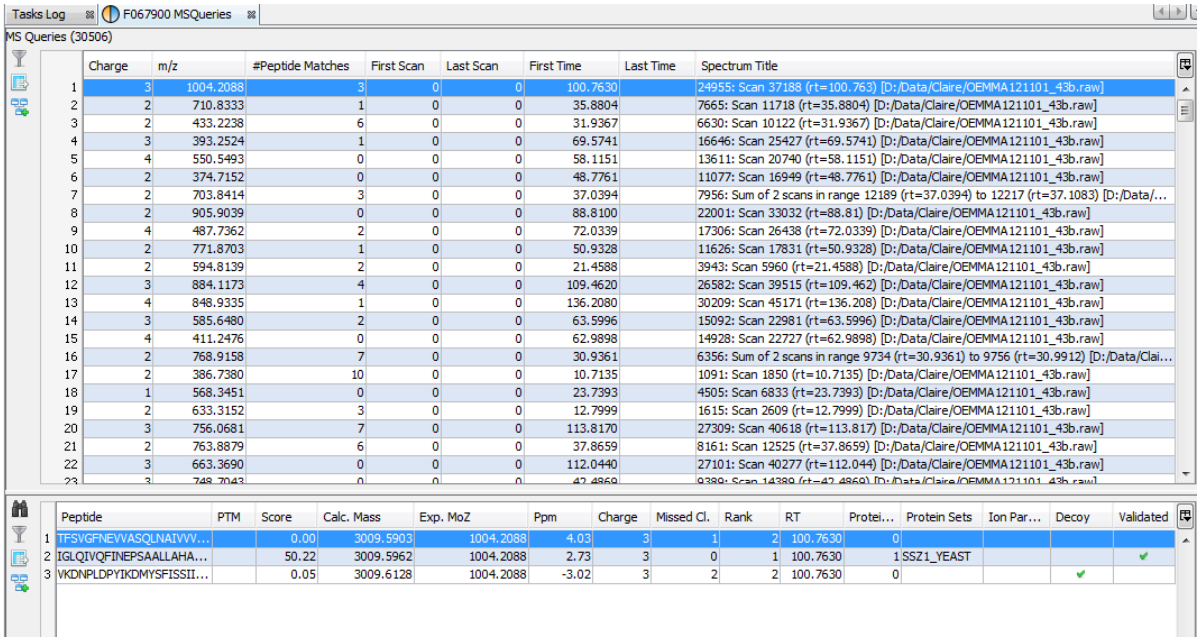
- right click on an Identification Summary

- click on the menu “Display Identification Summary >” and on the sub-menu “MSQueries”, “PSM”, “Peptides”, “Protein Sets”, “PTM Protein Sites” or “Adjacency Matrix”



MSQueries Window

If you click on MSQueries sub-menu, you obtain this window:



Upper View: list of MSQueries.

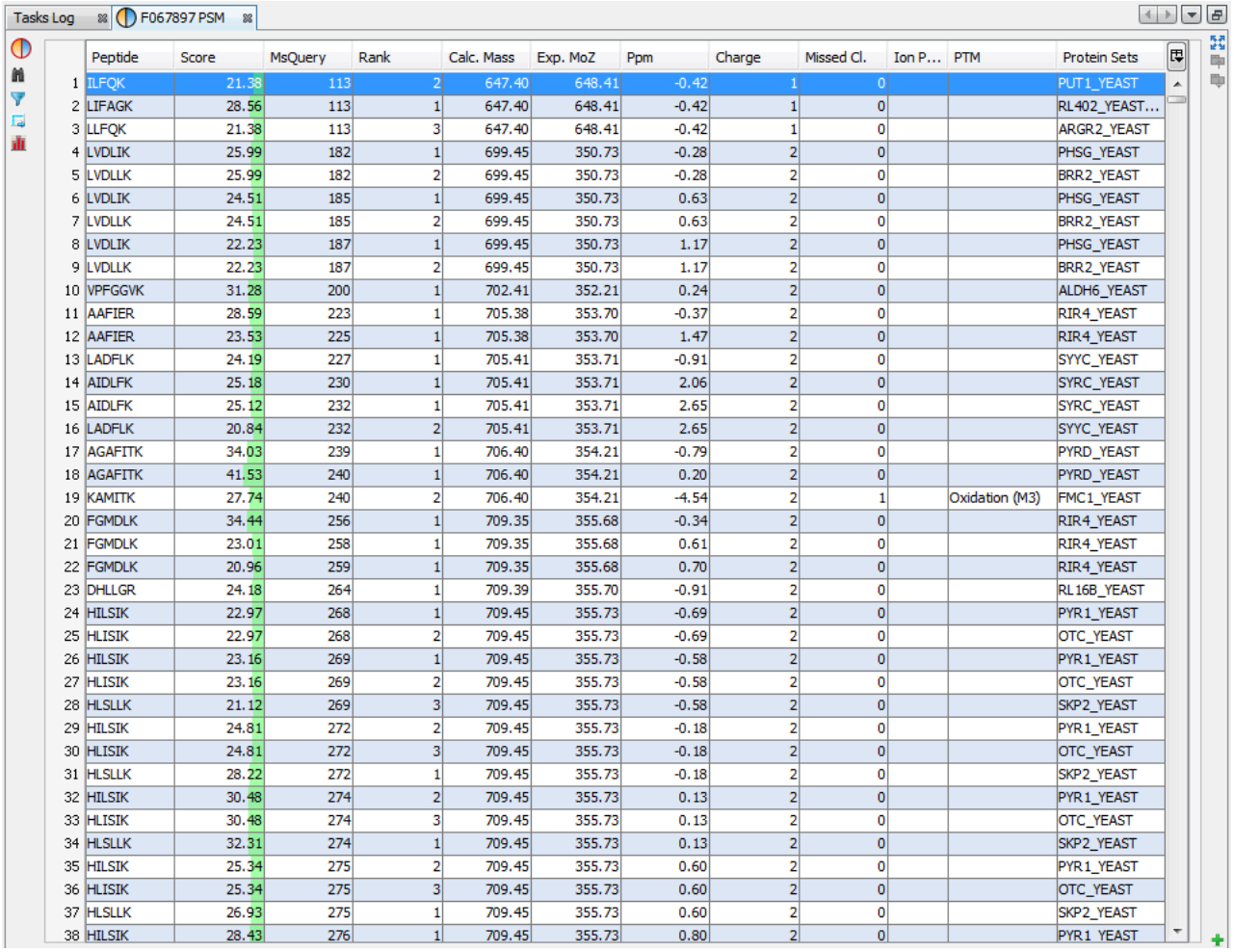
Bottom Window: list of all Peptides linked to the current selected MSQuery.

Note: Abbreviations used are listed here

This view contains all MSQueries even if it doesn't bring an identification.

PSMs Window

If you click on PSM sub-menu, you obtain this window:

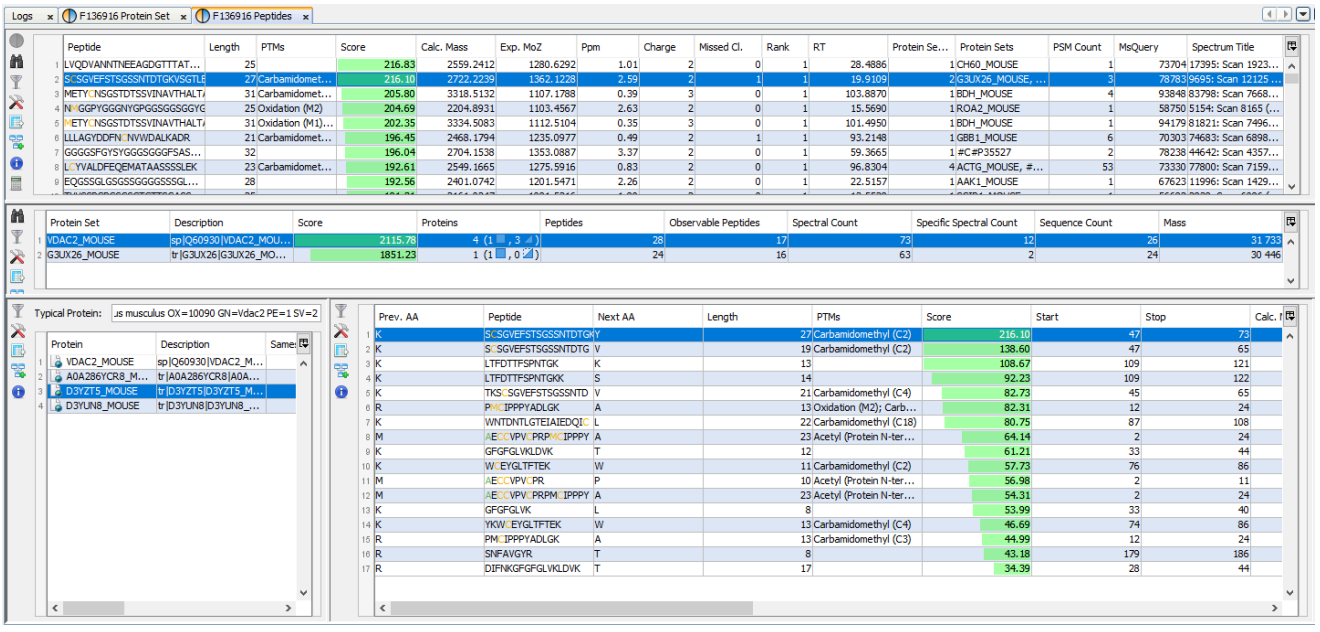


	Peptide	Score	MsQuery	Rank	Calc. Mass	Exp. MoZ	Ppm	Charge	Missed Cl.	Ion P…	PTM	Protein Sets
1	ILFQK	21.38	113	2	647.40	648.41	-0.42	1	0			PUT1_YEAST
2	LIFAGK	28.56	113	1	647.40	648.41	-0.42	1	0			RL402_YEAST…
3	LLFQK	21.38	113	3	647.40	648.41	-0.42	1	0			ARGR2_YEAST
4	LVDLIK	25.99	182	1	699.45	350.73	-0.28	2	0			PHSG_YEAST
5	LVDLLK	25.99	182	2	699.45	350.73	-0.28	2	0			BRR2_YEAST
6	LVDLIK	24.51	185	1	699.45	350.73	0.63	2	0			PHSG_YEAST
7	LVDLLK	24.51	185	2	699.45	350.73	0.63	2	0			BRR2_YEAST
8	LVDLIK	22.23	187	1	699.45	350.73	1.17	2	0			PHSG_YEAST
9	LVDLLK	22.23	187	2	699.45	350.73	1.17	2	0			BRR2_YEAST
10	VPFGGVK	31.28	200	1	702.41	352.21	0.24	2	0			ALDH6_YEAST
11	AAFIER	28.59	223	1	705.38	353.70	-0.37	2	0			RIR4_YEAST
12	AAFIER	23.53	225	1	705.38	353.70	1.47	2	0			RIR4_YEAST
13	LADFLK	24.19	227	1	705.41	353.71	-0.91	2	0			SYYC_YEAST
14	AIDLFK	25.18	230	1	705.41	353.71	2.06	2	0			SYRC_YEAST
15	AIDLFK	25.12	232	1	705.41	353.71	2.65	2	0			SYRC_YEAST
16	LADFLK	20.84	232	2	705.41	353.71	2.65	2	0			SYYC_YEAST
17	AGAFITK	34.03	239	1	706.40	354.21	-0.79	2	0			PYRD_YEAST
18	AGAFITK	41.53	240	1	706.40	354.21	0.20	2	0			PYRD_YEAST
19	KAMITK	27.74	240	2	706.40	354.21	-4.54	2	1		Oxidation (M3)	FMC1_YEAST
20	FGMDLK	34.44	256	1	709.35	355.68	-0.34	2	0			RIR4_YEAST
21	FGMDLK	23.01	258	1	709.35	355.68	0.61	2	0			RIR4_YEAST
22	FGMDLK	20.96	259	1	709.35	355.68	0.70	2	0			RIR4_YEAST
23	DHLLGR	24.18	264	1	709.39	355.70	-0.91	2	0			RL16B_YEAST
24	HILSIK	22.97	268	1	709.45	355.73	-0.69	2	0			PYR1_YEAST
25	HLISIK	22.97	268	2	709.45	355.73	-0.69	2	0			OTC_YEAST
26	HILSIK	23.16	269	1	709.45	355.73	-0.58	2	0			PYR1_YEAST
27	HLISIK	23.16	269	2	709.45	355.73	-0.58	2	0			OTC_YEAST
28	HLSLLK	21.12	269	3	709.45	355.73	-0.58	2	0			SKP2_YEAST
29	HILSIK	24.81	272	2	709.45	355.73	-0.18	2	0			PYR1_YEAST
30	HLISIK	24.81	272	3	709.45	355.73	-0.18	2	0			OTC_YEAST
31	HLSLLK	28.22	272	1	709.45	355.73	-0.18	2	0			SKP2_YEAST
32	HILSIK	30.48	274	2	709.45	355.73	0.13	2	0			PYR1_YEAST
33	HLISIK	30.48	274	3	709.45	355.73	0.13	2	0			OTC_YEAST
34	HLSLLK	32.31	274	1	709.45	355.73	0.13	2	0			SKP2_YEAST
35	HILSIK	25.34	275	2	709.45	355.73	0.60	2	0			PYR1_YEAST
36	HLISIK	25.34	275	3	709.45	355.73	0.60	2	0			OTC_YEAST
37	HLSLLK	26.93	275	1	709.45	355.73	0.60	2	0			SKP2_YEAST
38	HILSIK	28.43	276	1	709.45	355.73	0.80	2	0			PYR1_YEAST

Note: Abbreviations used are listed here

Peptides Window

If you click on Peptides sub-menu, you obtain this window:



Upper View: list of all Peptides with best PSM information (charge, score ...)

Middle View: list of all Protein Sets identified by the selected peptide.

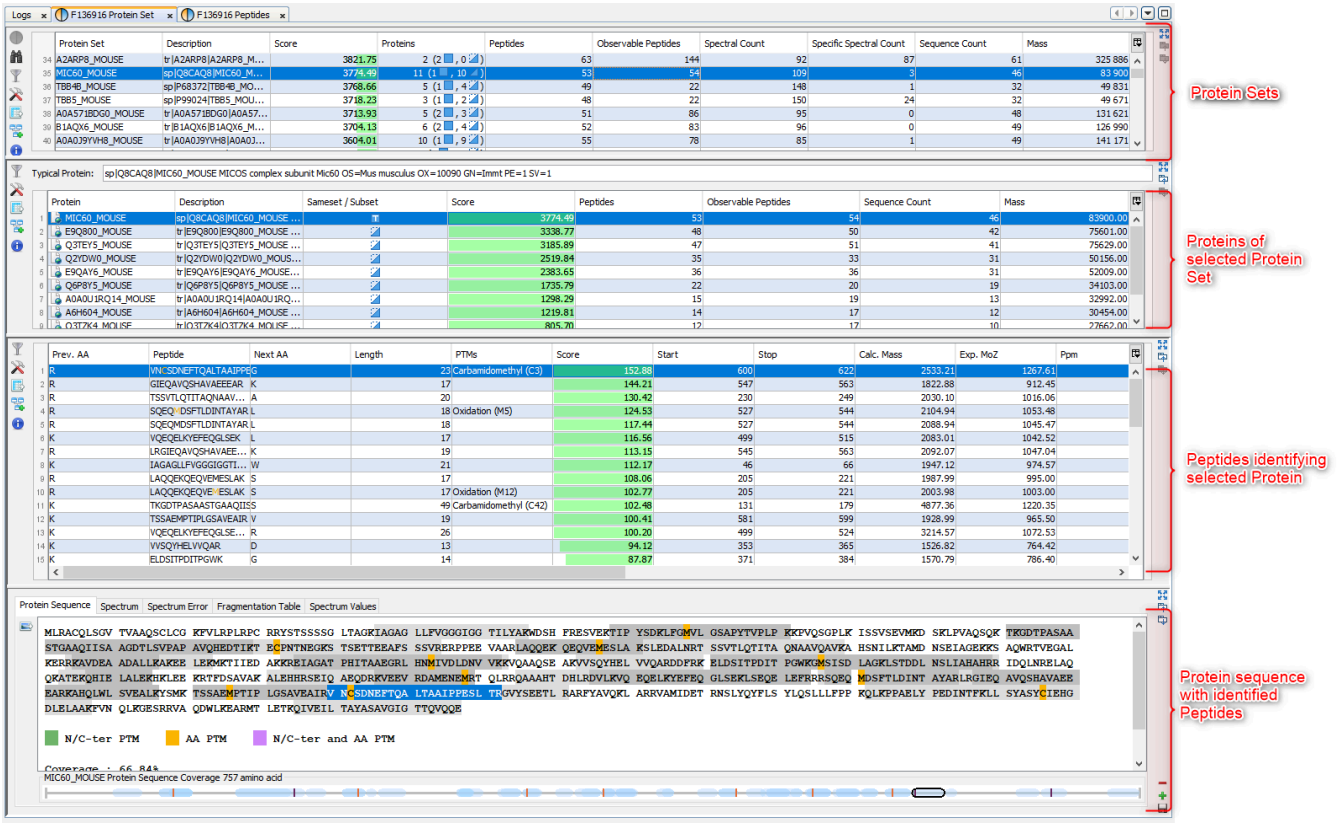
Bottom Left View: list of all Proteins of the selected Protein Set

Bottom Right View: list of all Peptides of the selected Protein

Note: Abbreviations used are listed here

Protein Sets Window

If you click on Protein Sets sub-menu, you obtain this window:



View 1 (upper): list of all Protein Sets of the identification Summary

Note: In the column *Proteins*, 8 (2, 6) means that there are 8 proteins in the protein set : 2 in the sameset, 6 in the subset.

View 2: list of all Proteins of the selected Protein Set, sameset or subset.

View 3: list of all Peptides of the selected Protein. If a subset is selected only peptides matching that protein will be listed.

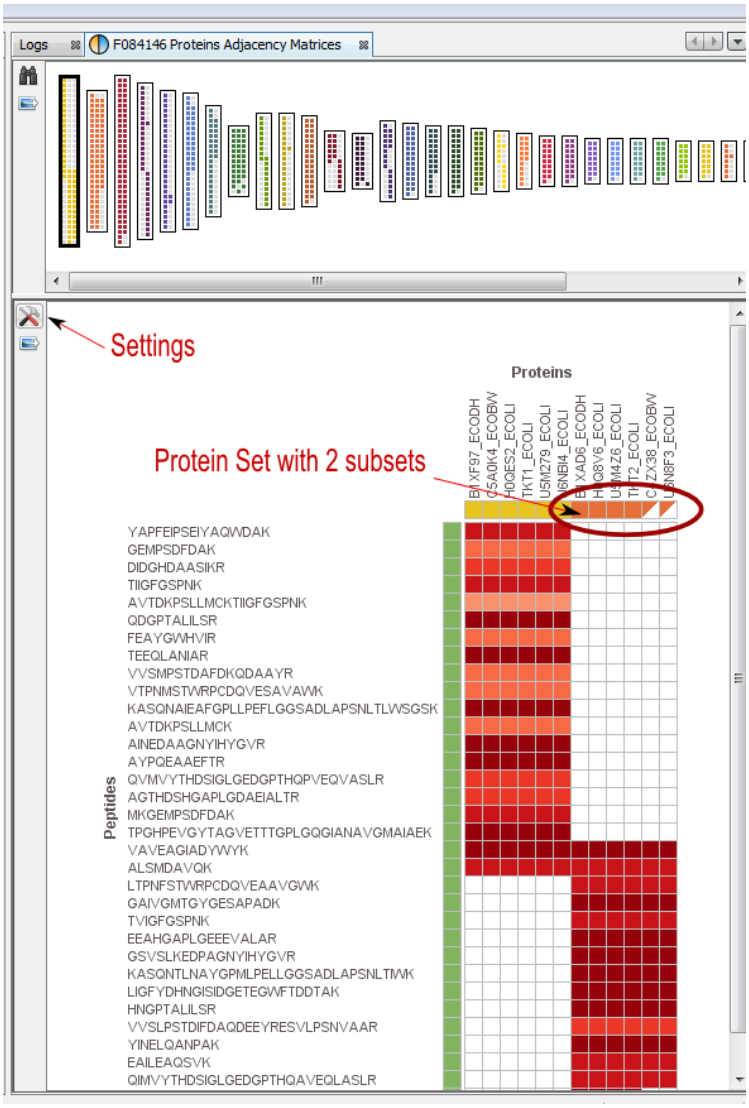
View 4a: Protein Sequence of the previously selected Protein and Spectrum of the selected Peptide. Other tabs display Spectrum, Spectrum Error and Fragmentation Table.

View 4b: Graphic representation of the Protein with matching peptide and associated modifications.

Note: Abbreviations used are listed here

Adjacency Matrix Window

If you click on Adjacency Matrix sub-menu, you obtain this window:



View 1: All the matrices. Each matrix corresponds to a cluster composed of linked Proteins/Peptides.

Note: use the Search tool to display an Adjacency Matrix for a particular Protein or Peptide

View 2: The currently selected matrix.

In the example, you can see two different protein sets which share only two peptides.

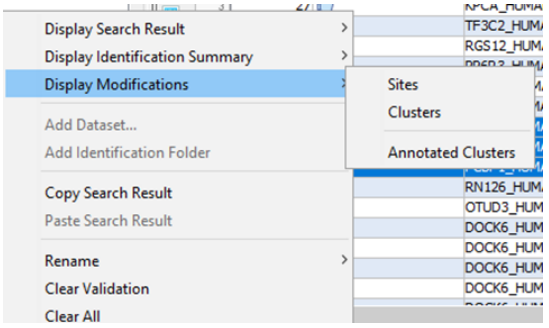
Thanks to the settings you can hide proteins with exactly the same peptides.

Display Modification Site or Cluster of a Modification Dataset

Functionality Access

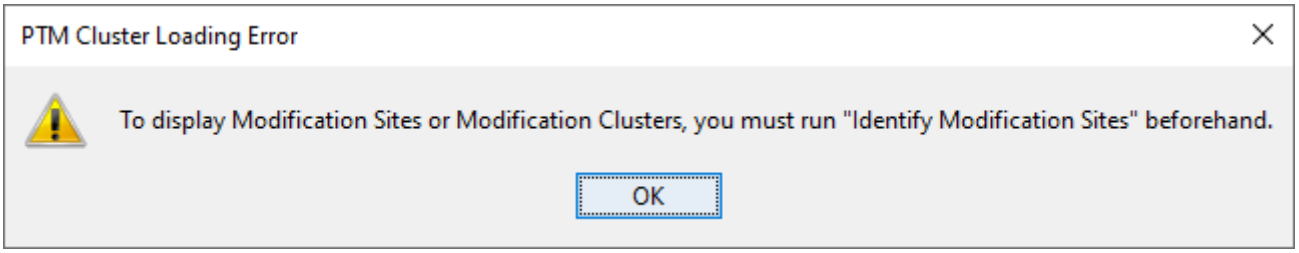
To display modification data associated to an Identification Summary:

- right click on the Dataset
- click on the Display Modification > “Sites”; “Clusters”; “Annotated Clusters”



If you click on one of “*Display Modifications”* sub-menu, you can obtain the following warning dialog.

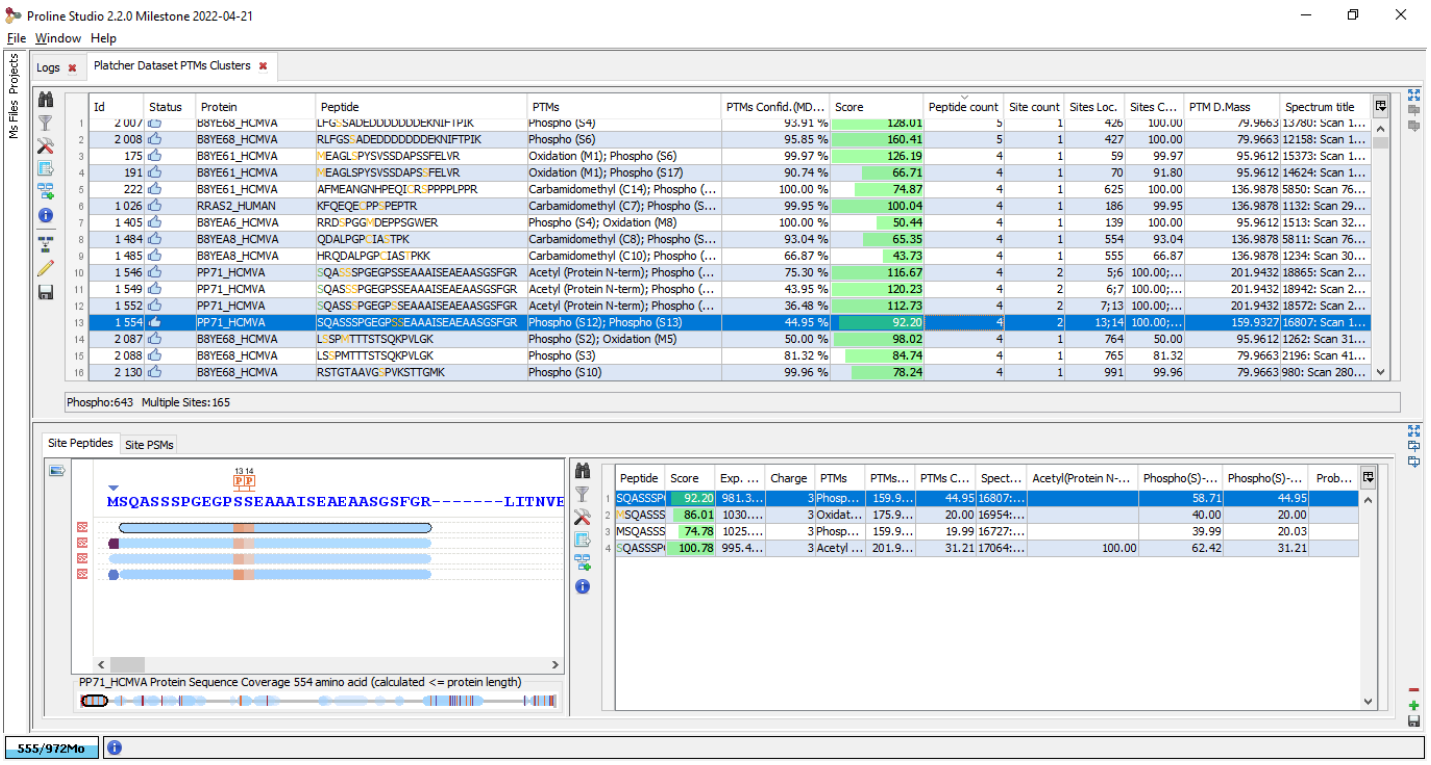
This is due to the fact that you must run beforehand the “Identify Modification Sites” process. To do that, mouse right click on your Identification Summary and select the “Identify Modification Sites” menu.



The Annotated Clusters menu may also cause the display of this warning. Indeed, Annotated Clusters display is only available after Clusters have been modified and saved.

All displays, Sites and Clusters, are structured in the same way. In Sites windows, the upper view will list all individual Sites while in Clusters windows, Sites will be clustered using rules specified by user (see Identify Modification Sites)

Clusters/Sites Windows



Upper View: This view lists all Modification Sites or Clusters. In this example, a Cluster view is shown. For each of them, many information is given : the number of Peptides belonging to the Cluster, how many sites have been clustered, The Confidence for the site combination, ...

Bottom Left view: Graphic illustration of the cluster/site peptides mapped to the protein sequence. All modification locations are represented above the protein sequence. An overview graphic is also given below.

Bottom Right view: Cluster/Site peptides list. The PSMs list is accessible in a hidden tab.

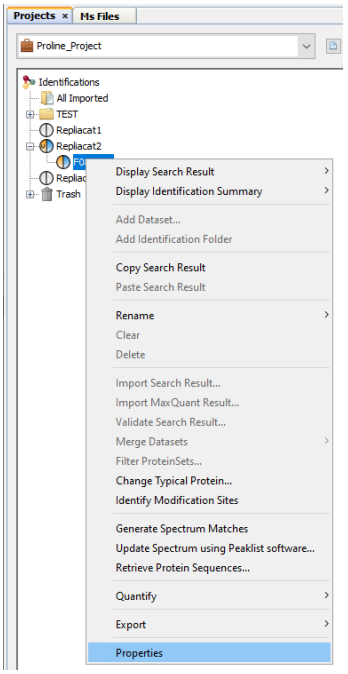
Display Additional Information on Search Result/Identification Summary

Functionality Access

To display properties of a Search Result or Identification Summary:

- right click on a Search Result/Identification Summary
- click on the menu “Properties”

Note: it is possible to select multiple Search Results/Identification Summaries to compare the values.



Properties Window

Property window opened:

General Information: Various information on the analysis (instrument name, peaklist software…)

Search Properties: Information extracted from the Result File (date, software version, search settings...)

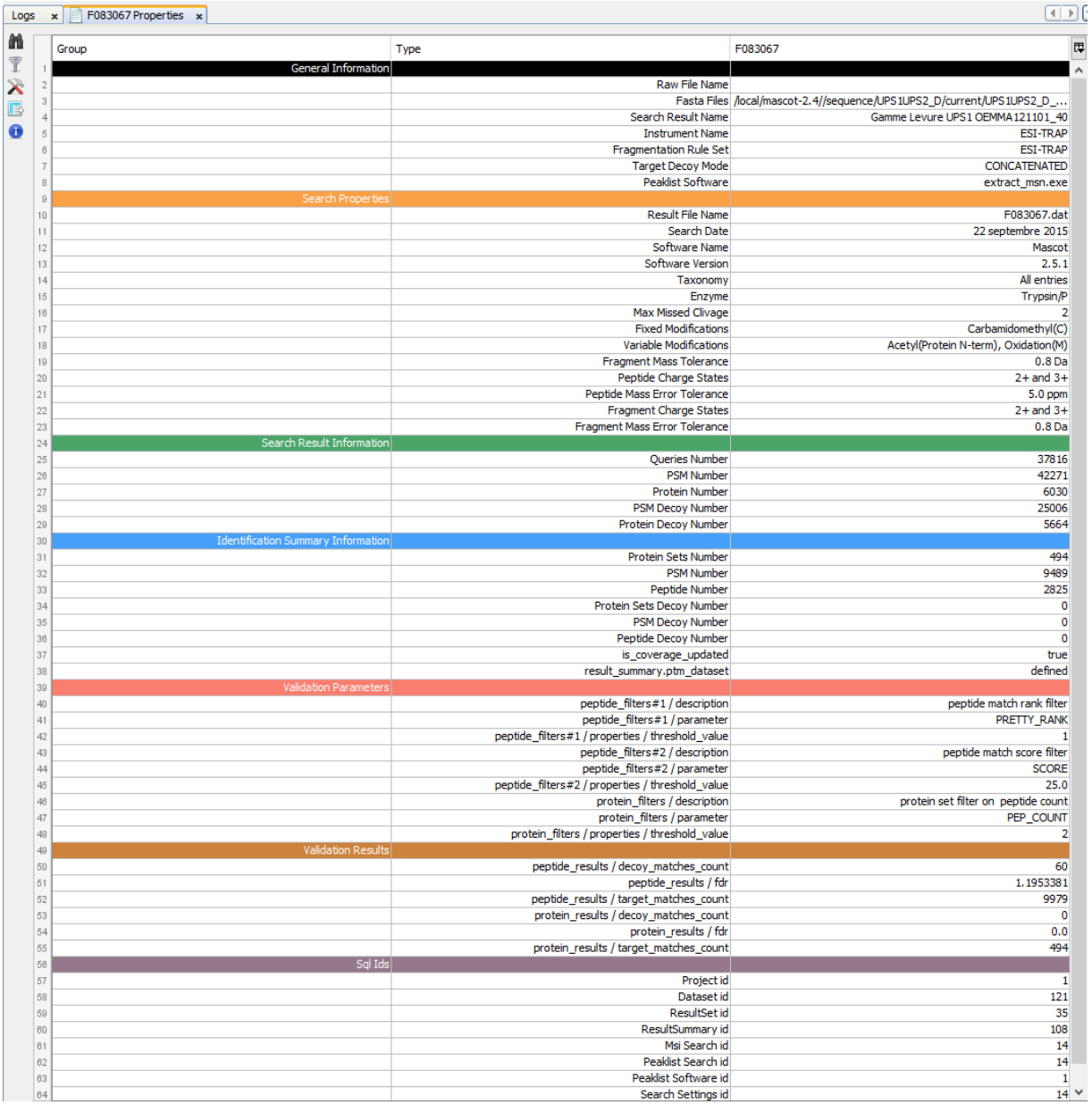
Search Result Information: Amount of Queries, PSM and Proteins in the Search Result.

Identification Summary Information: Information obtained after validation process

Validation xxx: Information on validation process : parameters used to validate and result

Sql Ids: Database ids related to this item

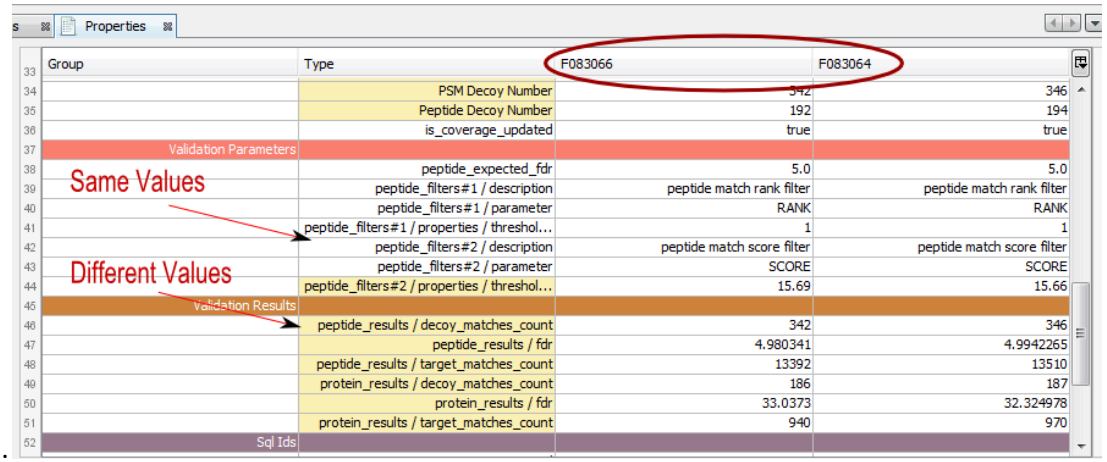
Note: Identification Summary Number may differ from Validation Results. Indeed, on one hand, peptide matches count in Validation Results takes into account all PSMs that have been validated . On the other hand, the PSM Number in the “Identification Summary Information” section considers only PSMs that identify a valid Protein Sets.



Group	Type	F083067
General Information		
	Raw File Name	
	Fasta Files	/local/mascot-2.4//sequence/UPS1UPS2_D/current/UPS1UPS2_D_...
	Search Result Name	Gamme Levure UPS1 OEMMA121101_40
	Instrument Name	ESI-TRAP
	Fragmentation Rule Set	ESI-TRAP
	Target Decoy Mode	CONCATENATED
	Peaklist Software	extract_msn.exe
Search Properties		
	Result File Name	F083067.dat
	Search Date	22 septembre 2015
	Software Name	Mascot
	Software Version	2.5.1
	Taxonomy	All entries
	Enzyme	Trypsin/P
	Max Missed Clivage	2
	Fixed Modifications	Carbamidomethyl(C)
	Variable Modifications	Acetyl(Protein N-term), Oxidation(M)
	Fragment Mass Tolerance	0.8 Da
	Peptide Charge States	2+ and 3+
	Peptide Mass Error Tolerance	5.0 ppm
	Fragment Charge States	2+ and 3+
	Fragment Mass Error Tolerance	0.8 Da
Search Result Information		
	Queries Number	37816
	PSM Number	42271
	Protein Number	6030
	PSM Decoy Number	25006
	Protein Decoy Number	5664
Identification Summary Information		
	Protein Sets Number	494
	PSM Number	9489
	Peptide Number	2825
	Protein Sets Decoy Number	0
	PSM Decoy Number	0
	Peptide Decoy Number	0
	is_coverage_updated	true
	result_summary.ptm_dataset	defined
Validation Parameters		
	peptide_filters#1 / description	peptide match rank filter
	peptide_filters#1 / parameter	PRETTY_RANK
	peptide_filters#1 / properties / threshold_value	1
	peptide_filters#2 / description	peptide match score filter
	peptide_filters#2 / parameter	SCORE
	peptide_filters#2 / properties / threshold_value	25.0
	protein_filters / description	protein set filter on peptide count
	protein_filters / parameter	PEP_COUNT
	protein_filters / properties / threshold_value	2
Validation Results		
	peptide_results / decoy_matches_count	60
	peptide_results / fdr	1.1953381
	peptide_results / target_matches_count	9979
	protein_results / decoy_matches_count	0
	protein_results / fdr	0.0
	protein_results / target_matches_count	494
Sql Ids		
	Project id	1
	Dataset id	121
	ResultSet id	35
	ResultSummary id	108
	Msi Search id	14
	Peaklist Search id	14
	Peaklist Software id	1
	Search Settings id	14

Property window opened with multiple Identification summaries selected:

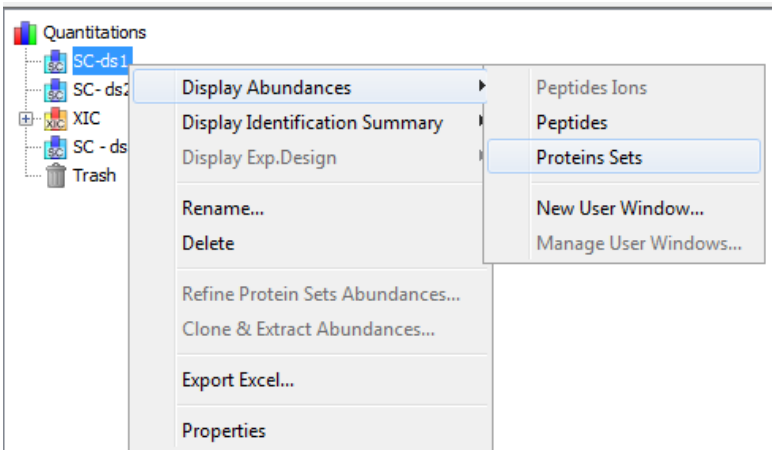
The color of the type column indicates if the values are the same (white) or different (yellow)



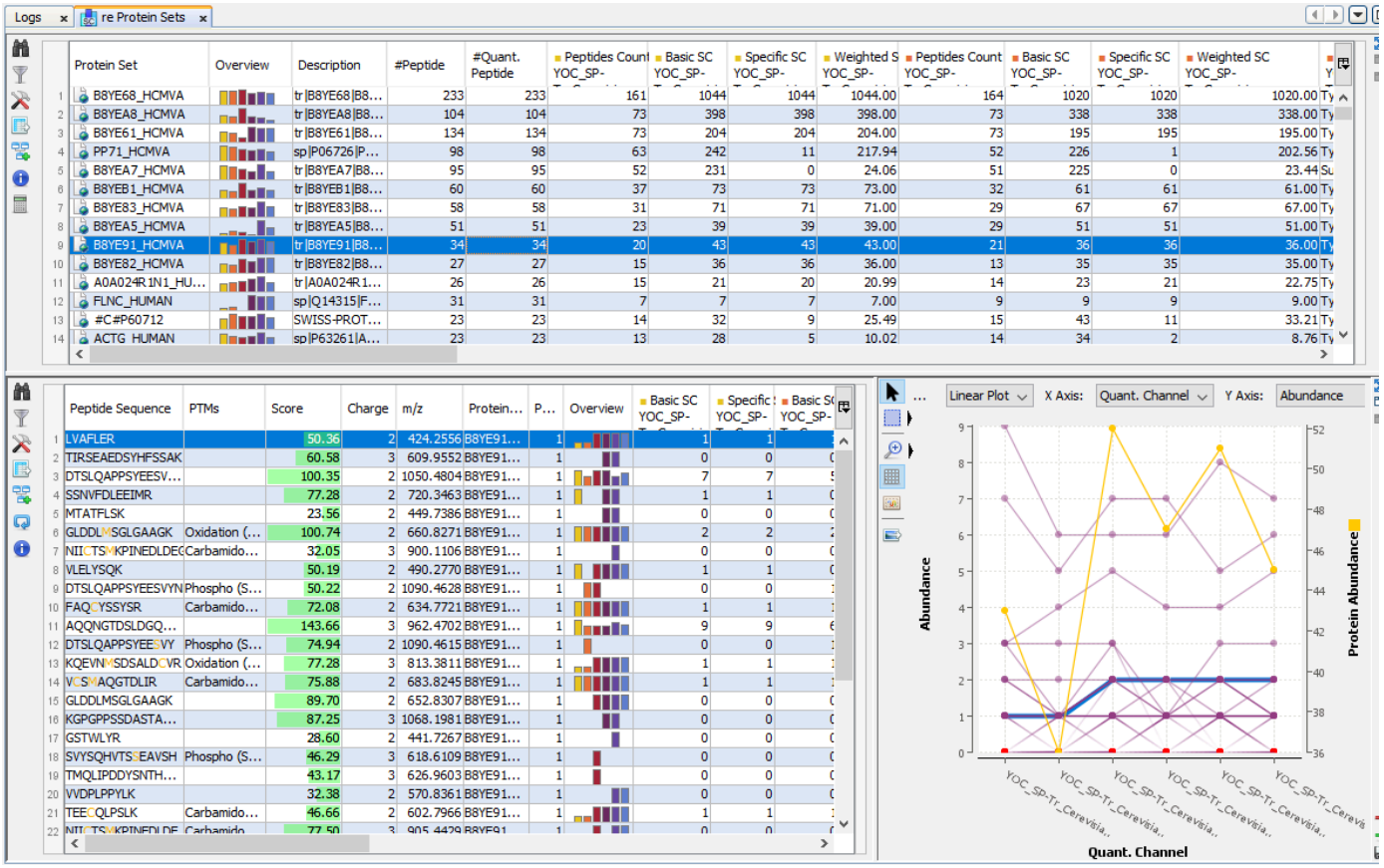
Group	Type	F083066	F083064
	PSM Decoy Number	342	346
	Peptide Decoy Number	192	194
	is_coverage_updated	true	true
Validation Parameters			
	peptide_expected_fdr	5.0	5.0
	peptide_filters#1 / description	peptide match rank filter	peptide match rank filter
	peptide_filters#1 / parameter	RANK	RANK
	peptide_filters#1 / properties / threshol...	1	1
	peptide_filters#2 / description	peptide match score filter	peptide match score filter
	peptide_filters#2 / parameter	SCORE	SCORE
	peptide_filters#2 / properties / threshol...	15.69	15.66
Validation Results			
	peptide_results / decoy_matches_count	342	346
	peptide_results / fdr	4.980341	4.9942265
	peptide_results / target_matches_count	13392	13510
	protein_results / decoy_matches_count	186	187
	protein_results / fdr	33.0373	32.324978
	protein_results / target_matches_count	940	970
Sql Ids			

Display a Spectral Count

You can display a generated Spectral Count by using the right mouse popup.



To have more details about the results, see [spectral_count_result]



The overview is based by default on the weighted spectral count values. (Note: if you sort on the overview column, the sort is based on max (value-mean (values))/mean (values). So, you will obtain the most homogenous and confident rows first)

For each compared dataset, are displayed:

- status (typical, sameset, /)

- peptide numbers

- the basic spectral count

- the specific spectral count

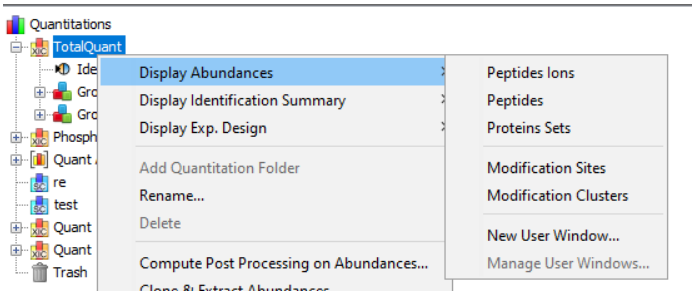
- the weighted spectral count

- the selection level

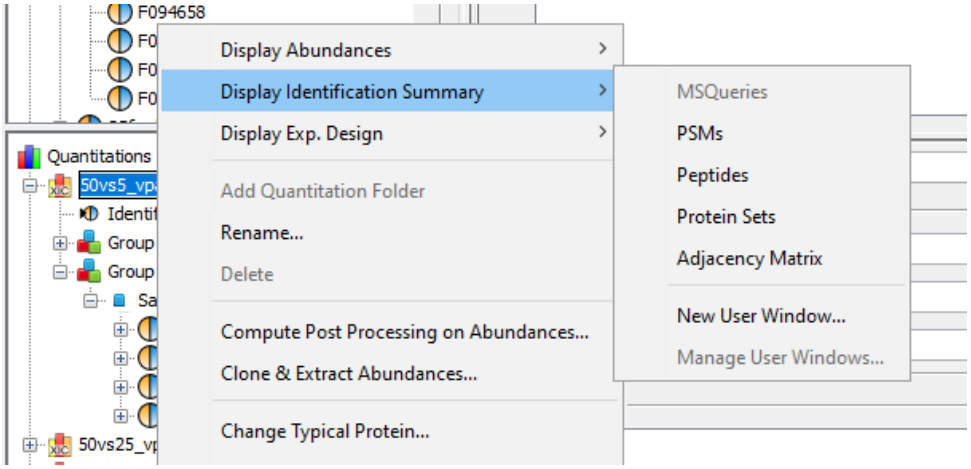
User can change the information displayed by the overview using the table settings icon ( ) .

Display a quantitation result

To display a quantitation, other than spectral count, right click on the selected node in the Quantitation tree, and select “Display Abundances”, and then the level you want to display:



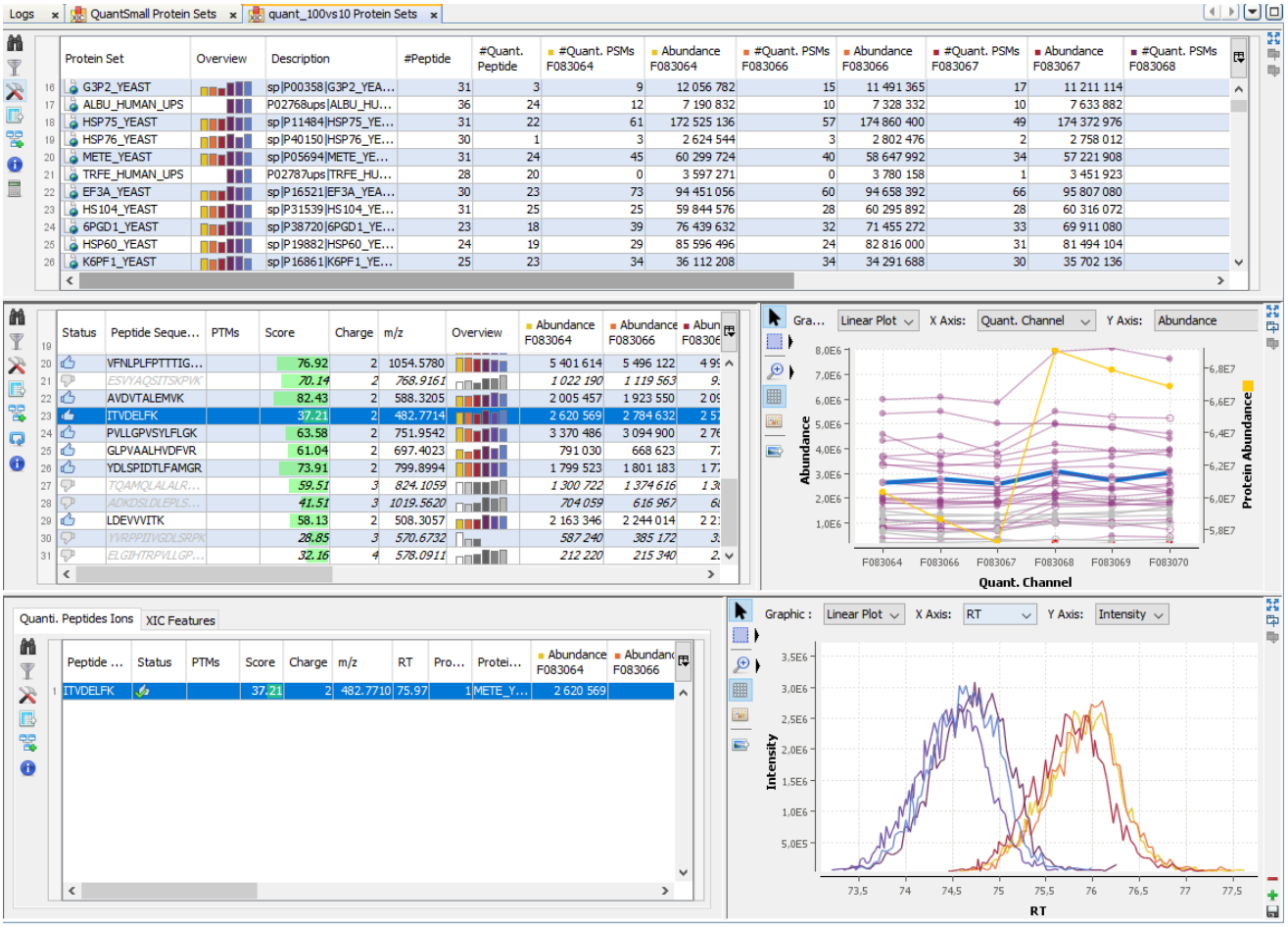
Note: You can also display the identification summary used as reference for the quantitation from the popup menu in the quantitation tree:



Display Protein Sets

Protein Sets

By clicking on “Display Abundances” / “Protein Sets”, you can see all quantified protein sets. For each quantified protein set, you can see below all peptides linked to the selected protein set and peptides Ions linked to the selected peptide. For each peptide Ion, you can see the different features and the graph of the peakels in each quantitation channel.



The overview is based by default on the abundance values.

Note: if you sort on the overview column, the sort is based on max (value-mean (values))/mean (values). So, you obtain the most homogenous and confident rows first.

For each quantitation channel, are displayed:

- the raw abundance

- the peptide match count

- the abundance

- the selection level

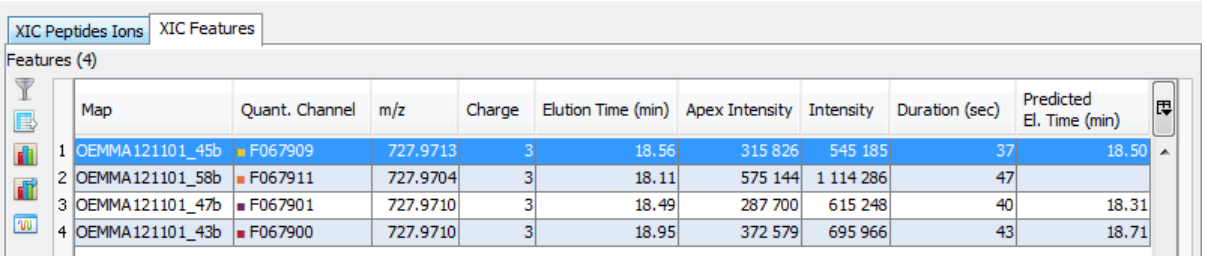
By clicking on the using the “table setting” icon , you can choose the information you want to display or change the overview.

Peptides

The middle part of the window lists all peptides of the selected Protein set with the same kind of quantitative data. The status column indicates whether the peptide was used or not for protein set abundances. On the right, a graph allows you to see the variations of the abundance (or raw abundance) of a peptide in the different quantitation channels.

Features and Peakels

You can see the different features in the different quantitation channels and the graph of the peakels:

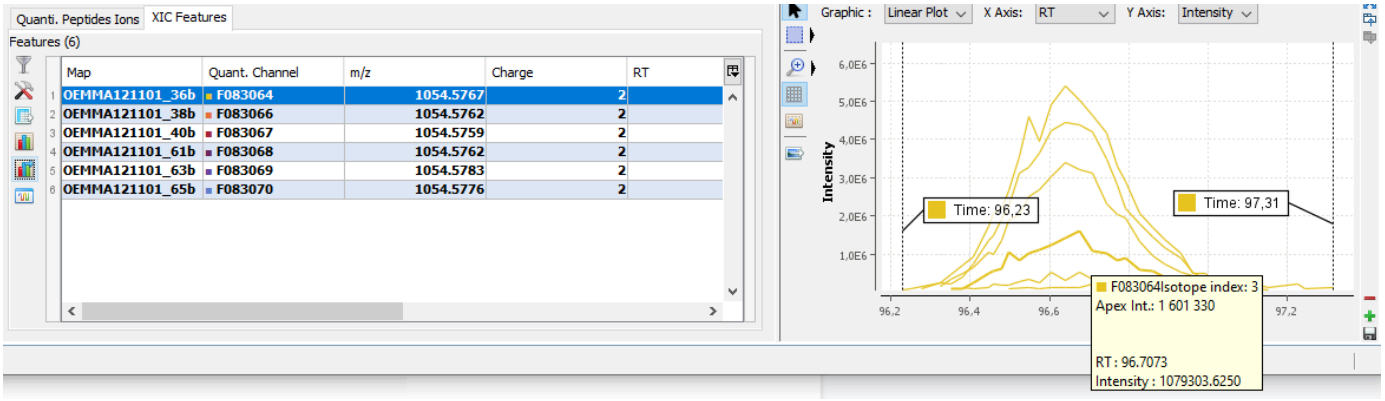


	Map	Quant. Channel	m/z	Charge	Elution Time (min)	Apex Intensity	Intensity	Duration (sec)	Predicted El. Time (min)
1	OEMMA121101_45b	F067909	727.9713	3	18.56	315 826	545 185	37	18.50
2	OEMMA121101_58b	F067911	727.9704	3	18.11	575 144	1 114 286	47	
3	OEMMA121101_47b	F067901	727.9710	3	18.49	287 700	615 248	40	18.31
4	OEMMA121101_43b	F067900	727.9710	3	18.95	372 579	695 966	43	18.71

By clicking on  you can display either:

- the peaks of isotope 0 in all quantitation channels

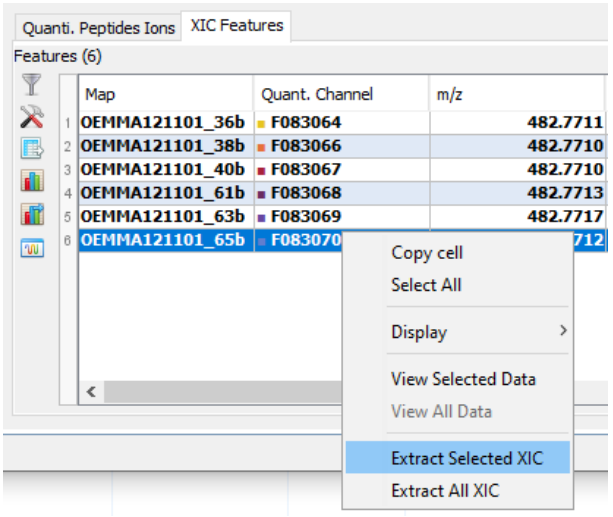
- all isotopes for the selected quantitation channel:



	Map	Quant. Channel	m/z	Charge	RT
1	OEMMA121101_36b	F083064	1054.5767	2	
2	OEMMA121101_38b	F083066	1054.5762	2	
3	OEMMA121101_40b	F083067	1054.5759	2	
4	OEMMA121101_61b	F083068	1054.5762	2	
5	OEMMA121101_63b	F083069	1054.5783	2	
6	OEMMA121101_65b	F083070	1054.5776	2	

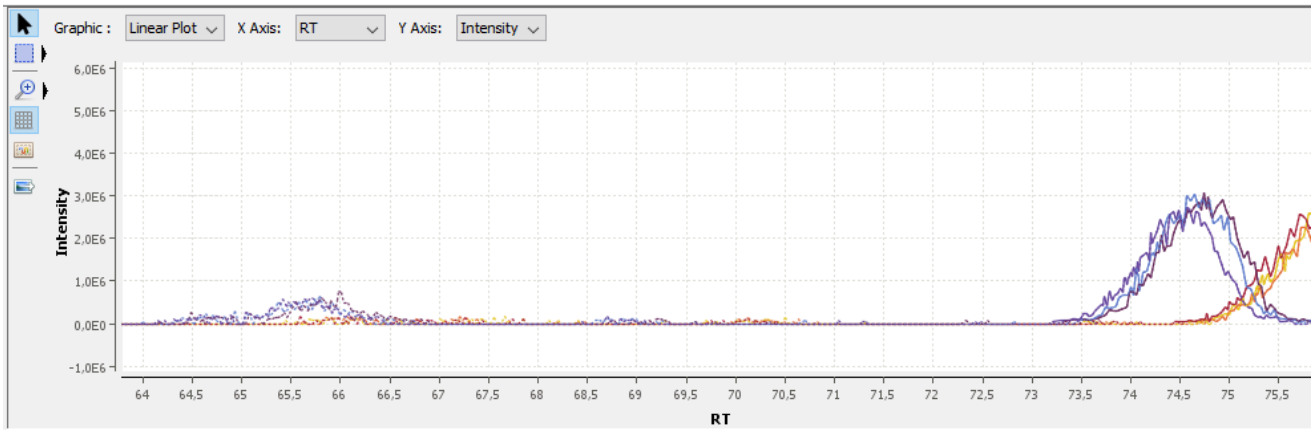
By clicking on  you can see the chromatograms of the features and their first time scan and last time scan in mzScope. For more details see the mzScope section.

It is also possible to extract the corresponding chromatogram for one or all of the features.



	Map	Quant. Channel	m/z
1	OEMMA121101_36b	F083064	482.7711
2	OEMMA121101_38b	F083066	482.7710
3	OEMMA121101_40b	F083067	482.7710
4	OEMMA121101_61b	F083068	482.7713
5	OEMMA121101_63b	F083069	482.7717
6	OEMMA121101_65b	F083070	712

The resulting chromatograms will be displayed in the same windows as peakel.



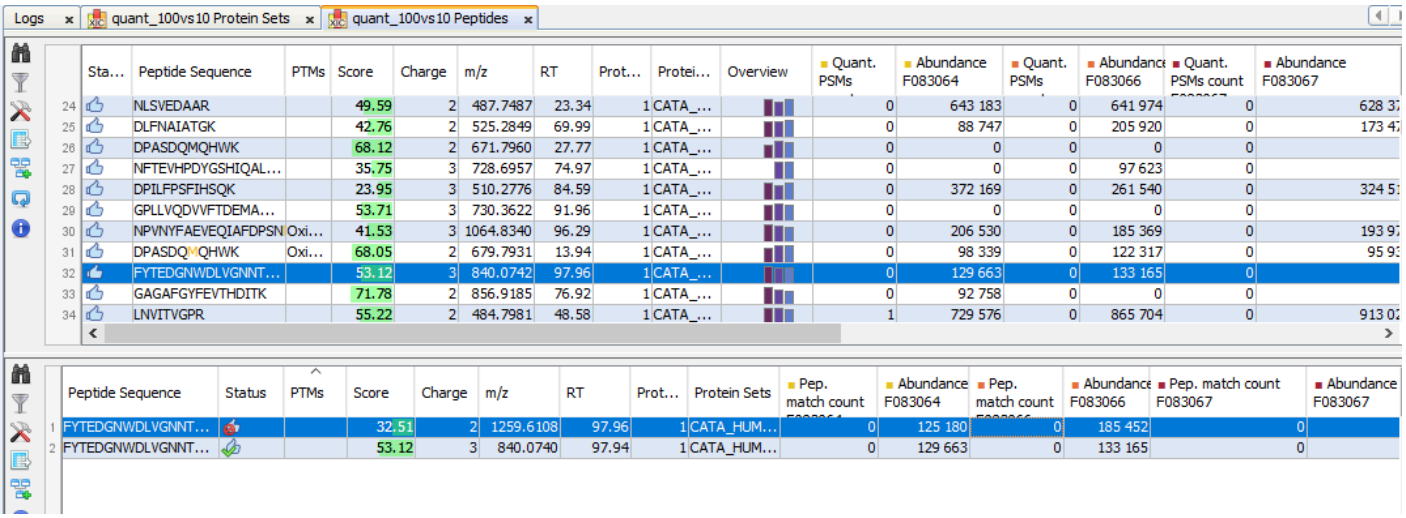
Display Peptides

By clicking on “Display Abundances“ / “Peptides”, you can see:

- identified and quantified Peptides

- non identified but quantified peptides

- identified but not quantified peptides (linked to a quantified protein)



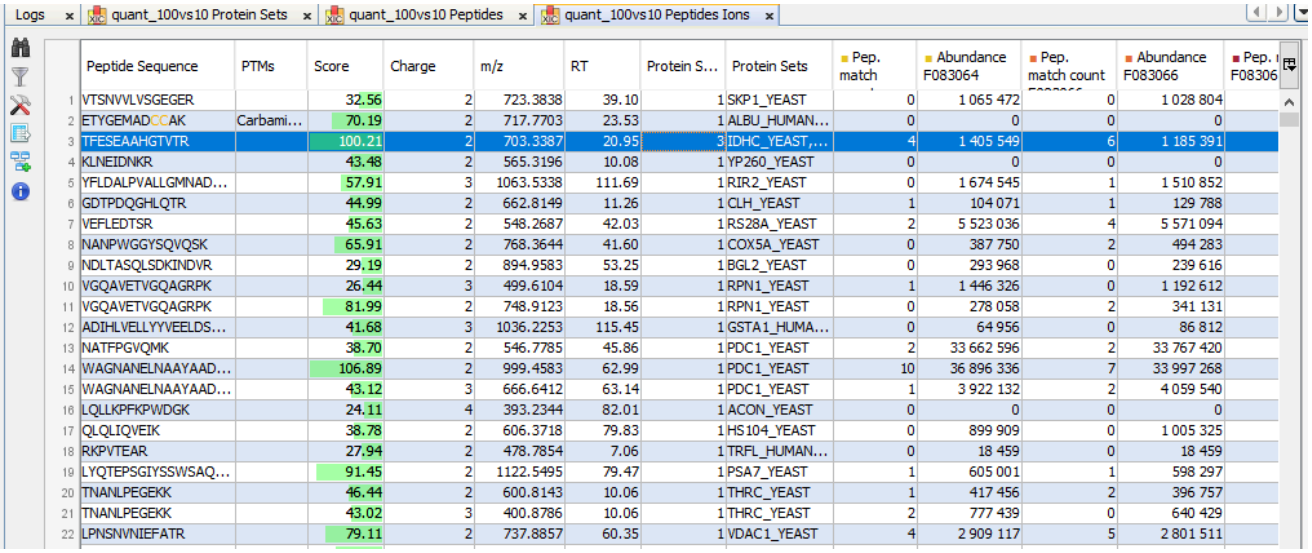
The lower view lists all peptide ions (specific charge) of selected peptide. The status column indicates if the ion is valid or not and if it was used for peptide quantitation.

Display Peptides Ions

By clicking on “Display Abundances” / “Peptides Ions”, you can see:

- all identified and quantified Peptides Ions

- non identified but quantified peptides Ions

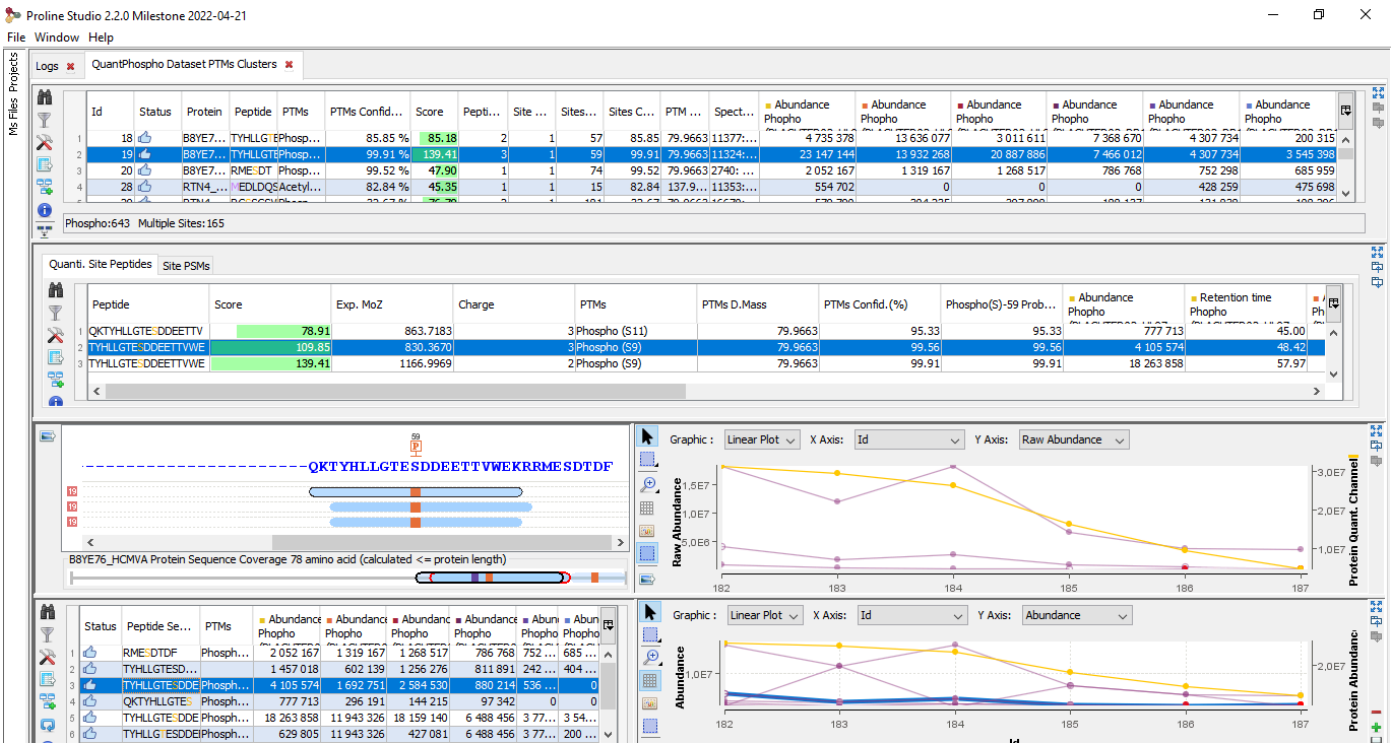


	Peptide Sequence	PTMs	Score	Charge	m/z	RT	Protein S...	Protein Sets	Pep. match	Abundance F083064	Pep. match count	Abundance F083066	Pep. F08306
1	VTSNVVLVSGEGER		32.56	2	723.3838	39.10	1	SKP1_YEAST	0	1 065 472	0	1 028 804	
2	ETYGEMADCCAK	Carbami...	70.19	2	717.7703	23.53	1	ALBU_HUMAN...	0	0	0	0	
3	TFESEAAHGTVTR		100.21	2	703.3387	20.95	3	IDHC_YEAST,...	4	1 405 549	6	1 185 391	
4	KLNEIDNKR		43.48	2	565.3196	10.08	1	YP260_YEAST	0	0	0	0	
5	YFLDALPVALLGMNAD...		57.91	3	1063.5338	111.69	1	RIR2_YEAST	0	1 674 545	1	1 510 852	
6	GDTPDQGHLQTR		44.99	2	662.8149	11.26	1	CLH_YEAST	1	104 071	1	129 788	
7	VEFLEDTSR		45.63	2	548.2687	42.03	1	RS28A_YEAST	2	5 523 036	4	5 571 094	
8	NANPWGGYSQVQSK		65.91	2	768.3644	41.60	1	COX5A_YEAST	0	387 750	2	494 283	
9	NDLTASQLSDKINDVR		29.19	2	894.9583	53.25	1	BGL2_YEAST	0	293 968	0	239 616	
10	VGQAVETVGQAGRPK		26.44	3	499.6104	18.59	1	RPN1_YEAST	1	1 446 326	0	1 192 612	
11	VGQAVETVGQAGRPK		81.99	2	748.9123	18.56	1	RPN1_YEAST	0	278 058	2	341 131	
12	ADIHLVELLYYVEELDS...		41.68	3	1036.2253	115.45	1	GSTA1_HUMA...	0	64 956	0	86 812	
13	NATFPGVQMK		38.70	2	546.7785	45.86	1	PDC1_YEAST	2	33 662 596	2	33 767 420	
14	WAGNANELNAAYAAD...		106.89	2	999.4583	62.99	1	PDC1_YEAST	10	36 896 336	7	33 997 268	
15	WAGNANELNAAYAAD...		43.12	3	666.6412	63.14	1	PDC1_YEAST	1	3 922 132	2	4 059 540	
16	LQLLKPFKPWDGK		24.11	4	393.2344	82.01	1	ACON_YEAST	0	0	0	0	
17	QLQLIQVEIK		38.78	2	606.3718	79.83	1	HS104_YEAST	0	899 909	0	1 005 325	
18	RKPVTEAR		27.94	2	478.7854	7.06	1	TRFL_HUMAN...	0	18 459	0	18 459	
19	LYQTEPSGIYSSWSAQ...		91.45	2	1122.5495	79.47	1	PSA7_YEAST	1	605 001	1	598 297	
20	TNANLPEGEKK		46.44	2	600.8143	10.06	1	THRC_YEAST	1	417 456	2	396 757	
21	TNANLPEGEKK		43.02	3	400.8786	10.06	1	THRC_YEAST	2	777 439	0	640 429	
22	LPNSNVNIEFATR		79.11	2	737.8857	60.35	1	VDAC1_YEAST	4	2 909 117	5	2 801 511	

Display Modification Sites/Clusters

By clicking on “Display Abundances” / “Modification XXX” (Site, Clusters or Annotated Clusters), you can see the same Clusters/Sites information as in “Identification Summary/Display / Modification XX” , See Display Modification description

In addition, quantitative information is computed and displayed.



Upper View: This view lists all Modification Sites or Clusters. For each of them, in addition to cluster information, quantitative data is given. Abundance is calculated using peptide’s abundance sum.

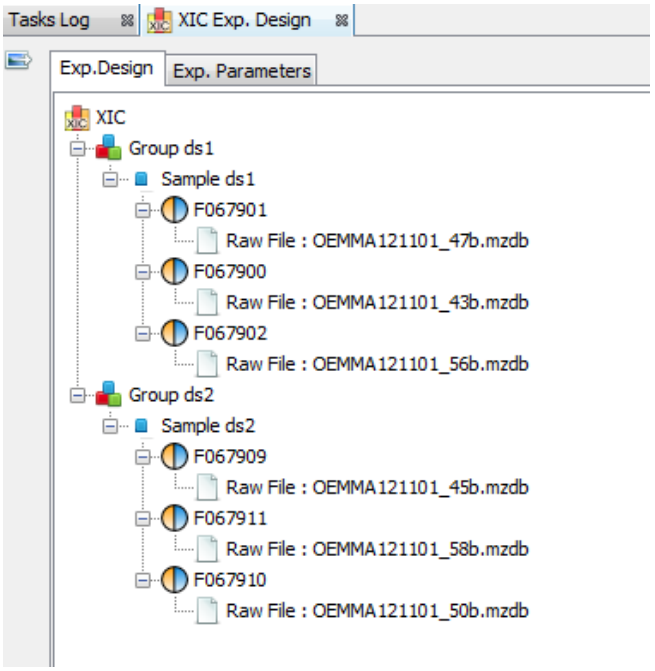
2nd view: Cluster/Site peptides list. The PSMs list is accessible in a hidden tab. Peptide quantitative information are displayed the same way it is done in other quantitative views.

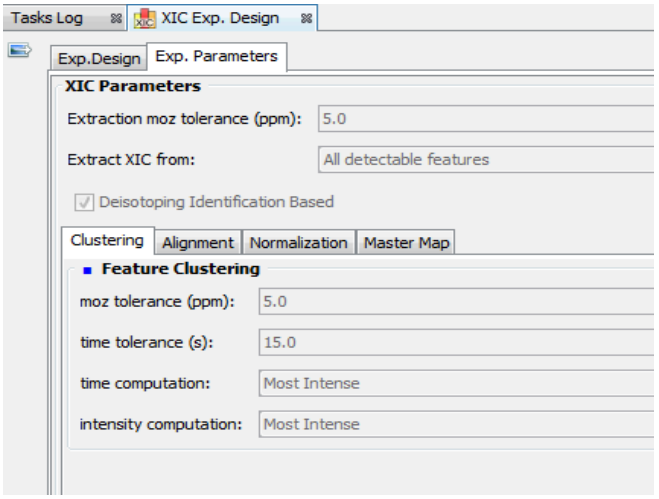
3rd view : Graphic illustration of the cluster/site peptides mapped to the protein sequence is shown on the left. On the right, a graph allows you to see the variations of the abundance (or raw abundance) of these peptides in the different quantitation channels.

Bottom Right view: This view is the same as the "Peptide" view in "Display Xic Protein Sets" above.

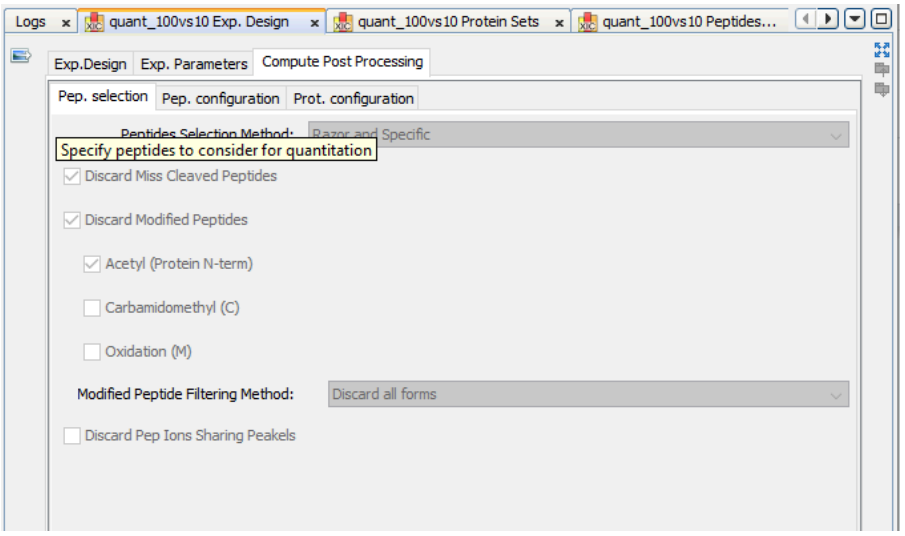
Display Experimental Design and Parameters

By clicking on "Exp. Design > Parameters", you can see the experimental design and the parameters of the selected XIC.



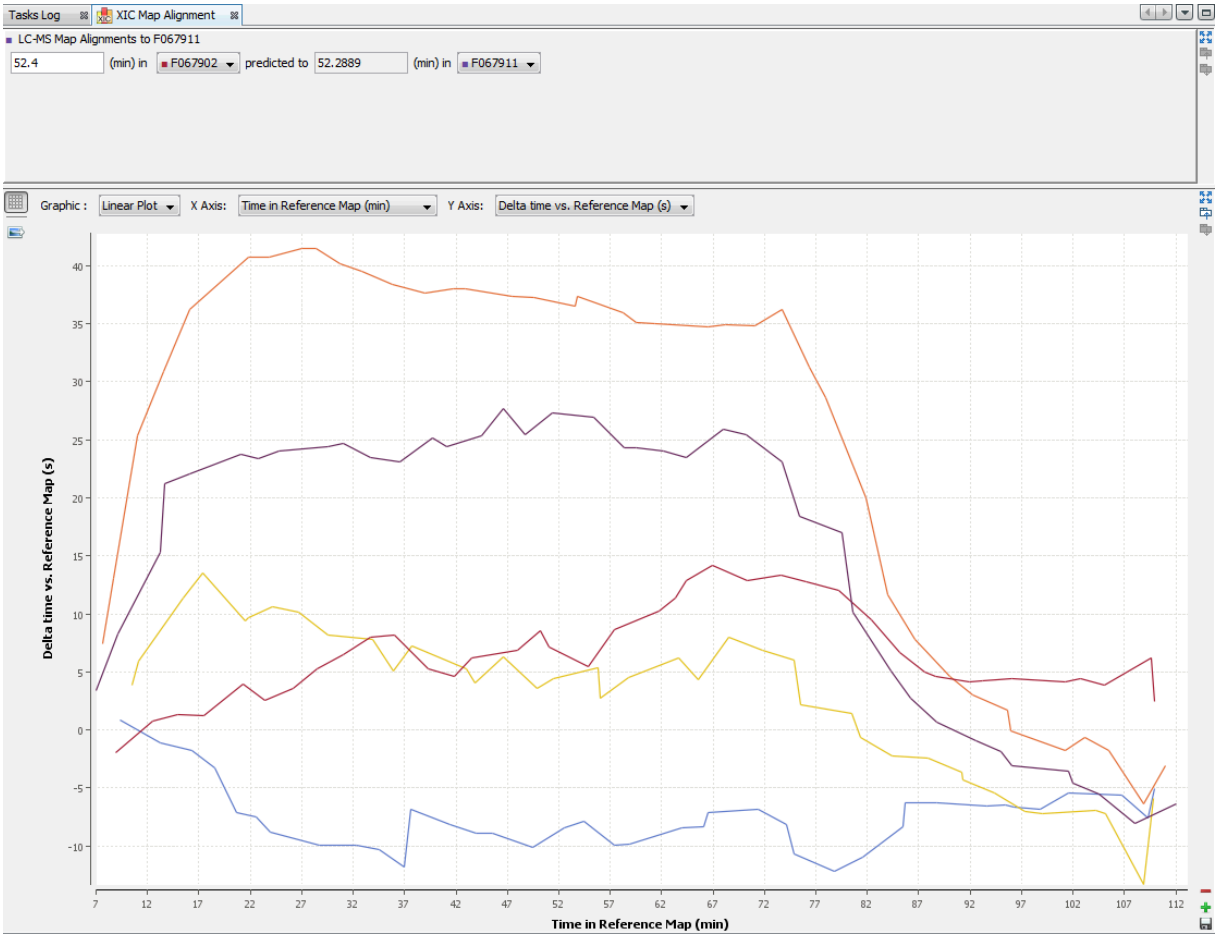


If you have launched the "compute post processing ..." on the XIC, you can also display the corresponding parameters.

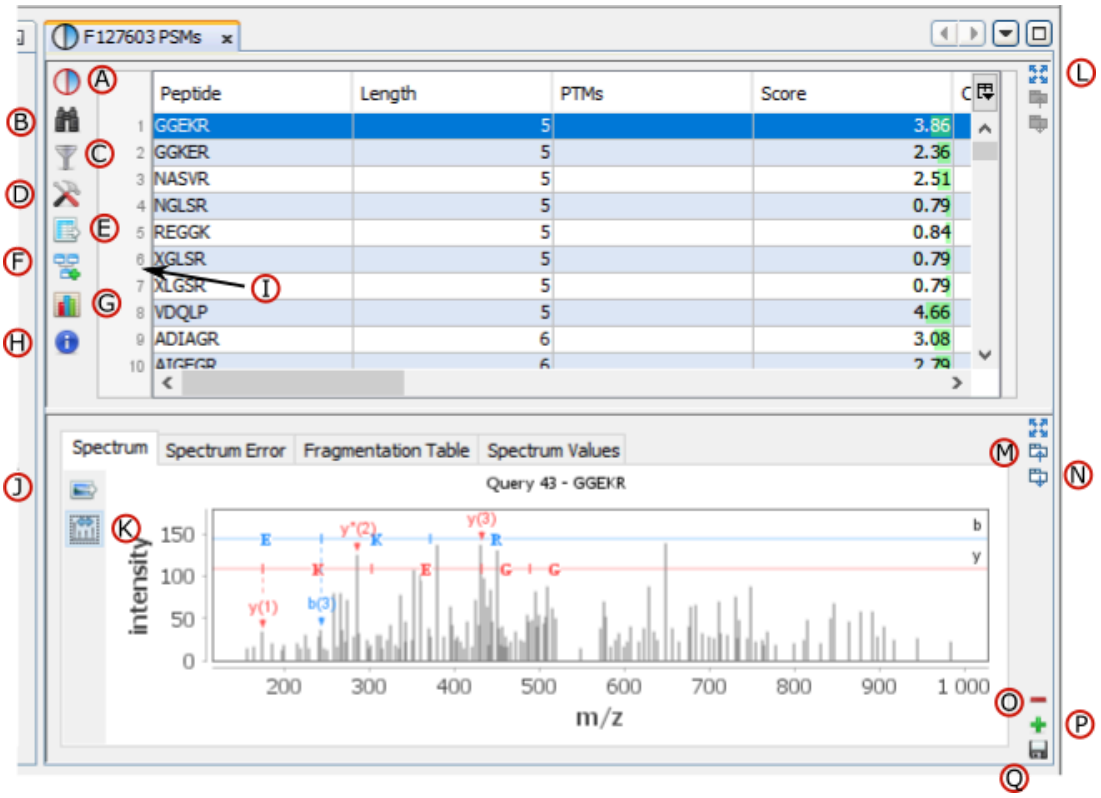


Display Map Alignment

By clicking on “Exp. Design > Map Alignment”, you can see the map of the variation of the alignment of the maps compared to the map alignment of the selected XIC. You can also calculate the predicted time in a map from an elution time in another map.



Frame Toolbars Functionalities



A: Display Decoy Data.

B: Search in the Table. (Using * and ? wild cards)

C: Filter data displayed in the Table

D: Display settings dialog (you can modify displayed columns and perform double sorting)

E: Export data displayed in the Table

F: Send to Data Analyzer to compare data from different views

G: Create a Graphic : histogram or scatter plot . Only on PSMs table

H: Display number of entities in the table (number of PSMs / Peptides / Proteins…)

I: Right click on the marker bar to display Line Numbers or add Annotations/Bookmarks

J: Export view as an image

K: Generate Spectrum Matches (specific to spectrum grahic)

L: Expands the frame to its maximum (other frames are hidden). Click again to undo.

M: Gather the frame with the previous one as a tab.

N: Split the last tab as a frame underneath

O: Remove the last Tab or Frame

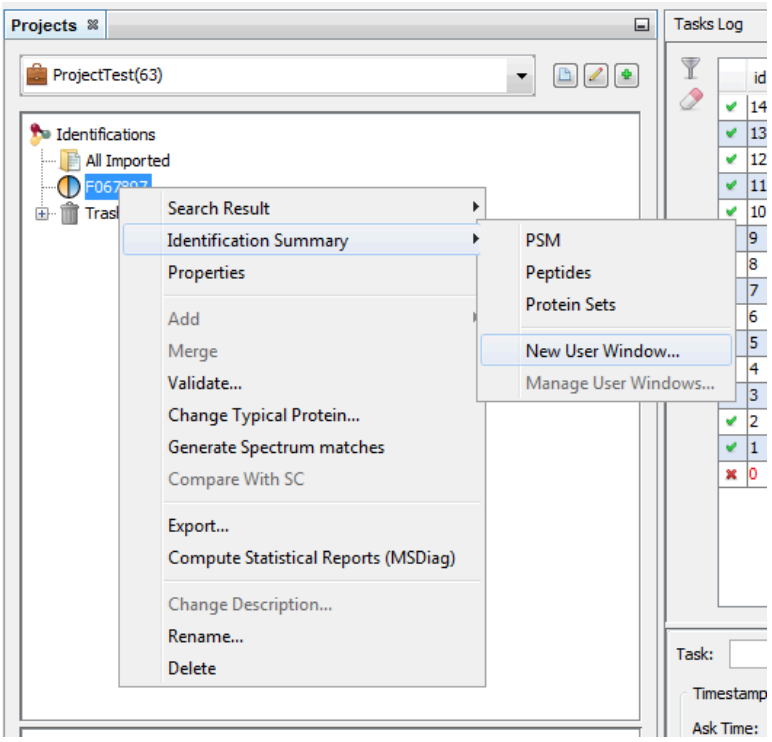
P: Open a dialog to let the user add a View (as a Frame, a Tab or a splitted Frame)

Q: Save the window as a user window, to display the same window with different data later

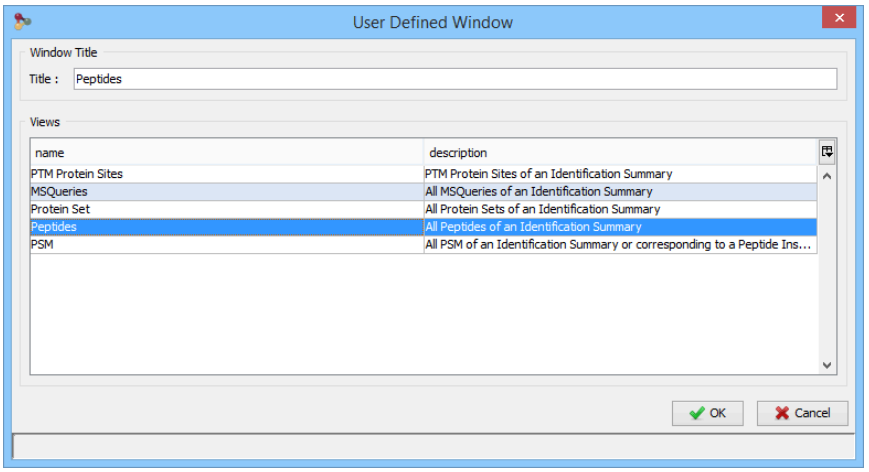
Create a custom User Window

You can lay out your own user window with the desired views.

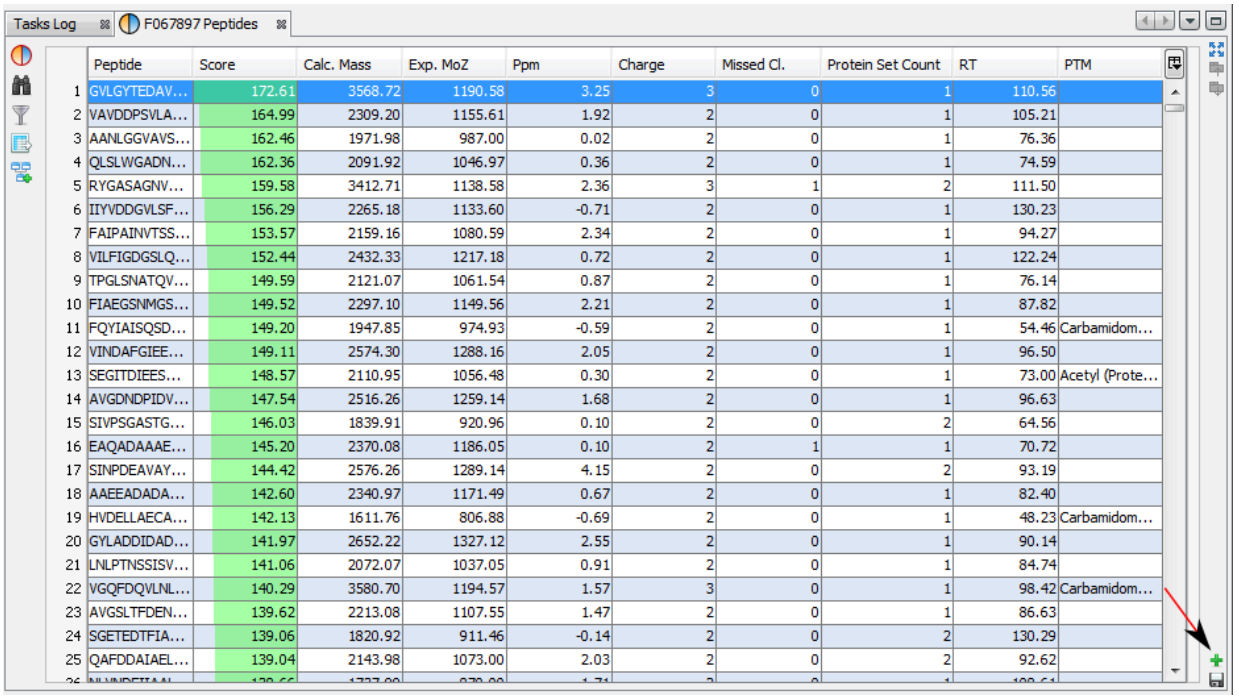
You can do it from an already displayed window, or by using the right click mouse popup on a dataset like in the following example (Use menu “Search Result>New User Window…” or “Identification Summary>New User Window…”)



In the example, the user has clicked on “Identification Summary>New User Window…” and selects the Peptides View as the first view of his window.

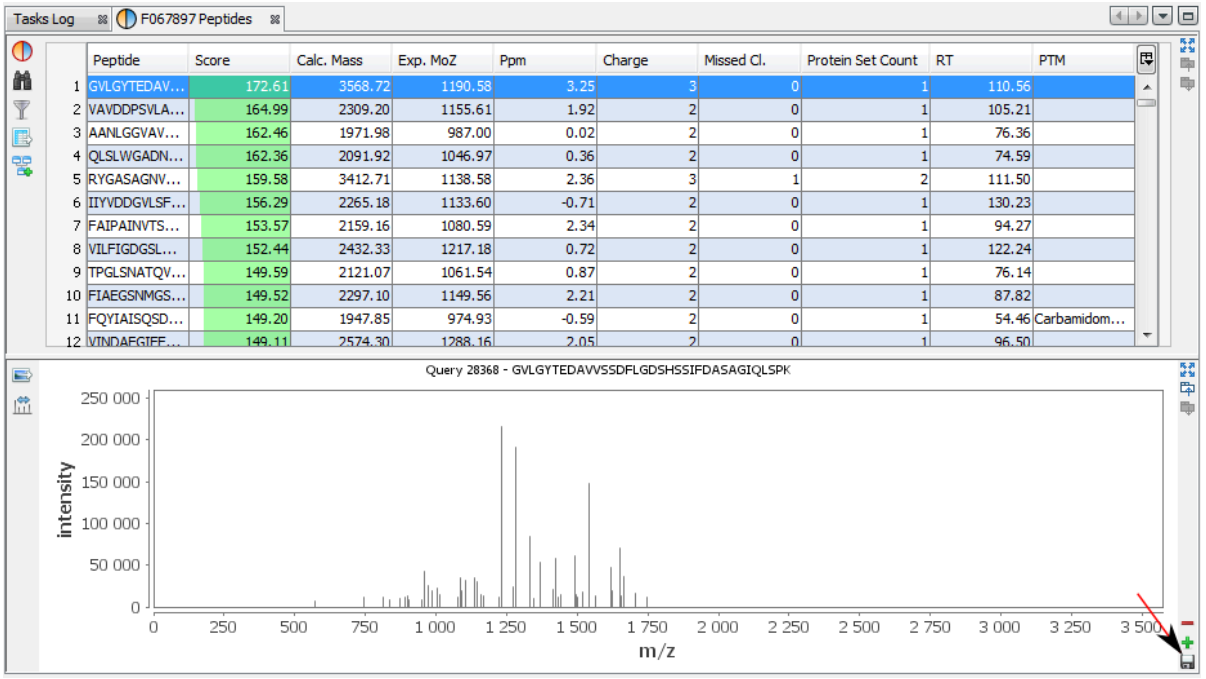


You can add other views by using the '+' button.



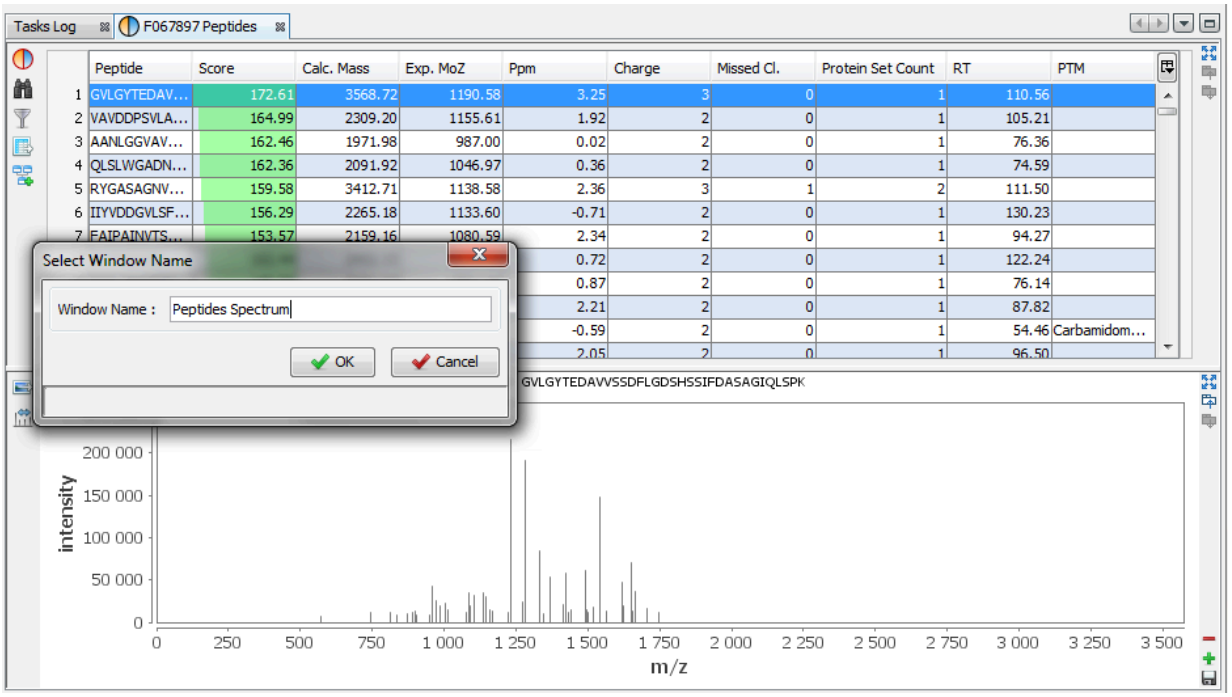
	Peptide	Score	Calc. Mass	Exp. MoZ	Ppm	Charge	Missed Cl.	Protein Set Count	RT	PTM
1	GVLGYTEDAV...	172.61	3568.72	1190.58	3.25	3	0	1	110.56	
2	VAVDDPSVLA...	164.99	2309.20	1155.61	1.92	2	0	1	105.21	
3	AANLGGVAVS...	162.46	1971.98	987.00	0.02	2	0	1	76.36	
4	QLSLWGADN...	162.36	2091.92	1046.97	0.36	2	0	1	74.59	
5	RYGASAGNV...	159.58	3412.71	1138.58	2.36	3	1	2	111.50	
6	IIYVDDGVLSF...	156.29	2265.18	1133.60	-0.71	2	0	1	130.23	
7	FAIPAINVTSS...	153.57	2159.16	1080.59	2.34	2	0	1	94.27	
8	VILFIGDGSLQ...	152.44	2432.33	1217.18	0.72	2	0	1	122.24	
9	TPGLSNATQV...	149.59	2121.07	1061.54	0.87	2	0	1	76.14	
10	FIAEGSNMGS...	149.52	2297.10	1149.56	2.21	2	0	1	87.82	
11	FQYIAISQSD...	149.20	1947.85	974.93	-0.59	2	0	1	54.46	Carbamidom...
12	VINDAFGIEE...	149.11	2574.30	1288.16	2.05	2	0	1	96.50	
13	SEGITDIEES...	148.57	2110.95	1056.48	0.30	2	0	1	73.00	Acetyl (Prote...
14	AVGDNDPIDV...	147.54	2516.26	1259.14	1.68	2	0	1	96.63	
15	SIVPSGASTG...	146.03	1839.91	920.96	0.10	2	0	2	64.56	
16	EAQADAAAE...	145.20	2370.08	1186.05	0.10	2	1	1	70.72	
17	SINPDEAVAY...	144.42	2576.26	1289.14	4.15	2	0	2	93.19	
18	AAEEADADA...	142.60	2340.97	1171.49	0.67	2	0	1	82.40	
19	HVDELLAECA...	142.13	1611.76	806.88	-0.69	2	0	1	48.23	Carbamidom...
20	GYLADDIDAD...	141.97	2652.22	1327.12	2.55	2	0	1	90.14	
21	LNLPTNSSISV...	141.06	2072.07	1037.05	0.91	2	0	1	84.74	
22	VGQFDQVLNL...	140.29	3580.70	1194.57	1.57	3	0	1	98.42	Carbamidom...
23	AVGSLTFDEN...	139.62	2213.08	1107.55	1.47	2	0	1	86.63	
24	SGETEDTFIA...	139.06	1820.92	911.46	-0.14	2	0	2	130.29	
25	QAFDDAIAEL...	139.04	2143.98	1073.00	2.03	2	0	2	92.62	

In this example, the user has added a Spectrum View and he saves his window by clicking on the “Disk” Button.

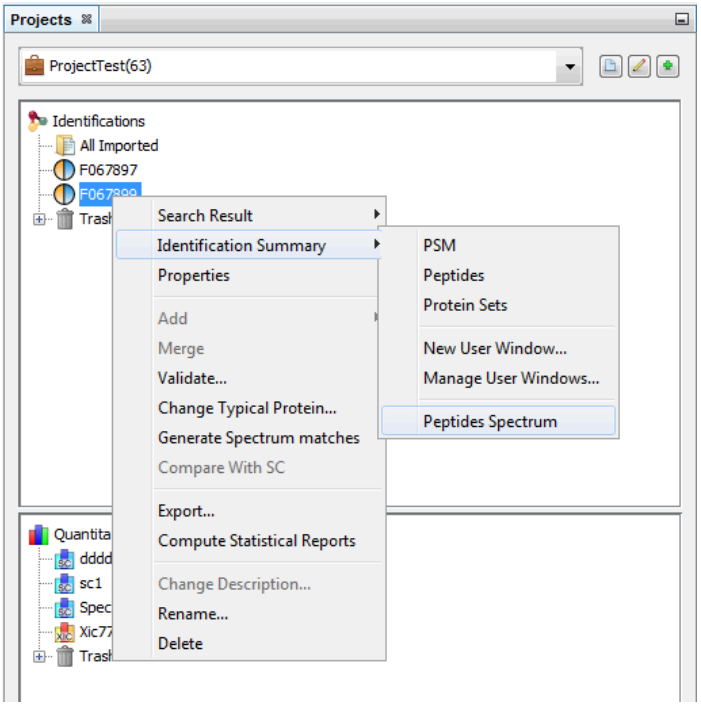


	Peptide	Score	Calc. Mass	Exp. MoZ	Ppm	Charge	Missed Cl.	Protein Set Count	RT	PTM
1	GVLGYTEDAV...	172.61	3568.72	1190.58	3.25	3	0	1	110.56	
2	VAVDDPSVLA...	164.99	2309.20	1155.61	1.92	2	0	1	105.21	
3	AANLGGVAV...	162.46	1971.98	987.00	0.02	2	0	1	76.36	
4	QLSLWGADN...	162.36	2091.92	1046.97	0.36	2	0	1	74.59	
5	RYGASAGNV...	159.58	3412.71	1138.58	2.36	3	1	2	111.50	
6	IIYVDDGVLSF...	156.29	2265.18	1133.60	-0.71	2	0	1	130.23	
7	FAIPAINVTS...	153.57	2159.16	1080.59	2.34	2	0	1	94.27	
8	VILFIGDGSL...	152.44	2432.33	1217.18	0.72	2	0	1	122.24	
9	TPGLSNATQV...	149.59	2121.07	1061.54	0.87	2	0	1	76.14	
10	FIAEGSNMGS...	149.52	2297.10	1149.56	2.21	2	0	1	87.82	
11	FQYIAISQSD...	149.20	1947.85	974.93	-0.59	2	0	1	54.46	Carbamidom...
12	VINDAFGIEE...	149.11	2574.30	1288.16	2.05	2	0	1	96.50	

The user selects 'Peptides Spectrum' as his user window name

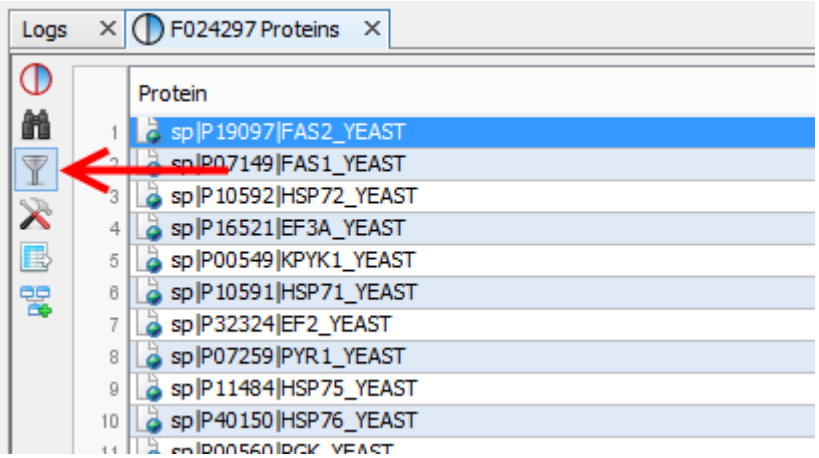


Now, the user can use his new 'Peptides Spectrum' on a different Identification Summary.



Filter Tables

You can filter data displayed in the different tables thanks to the filter button at the top right corner of a table.

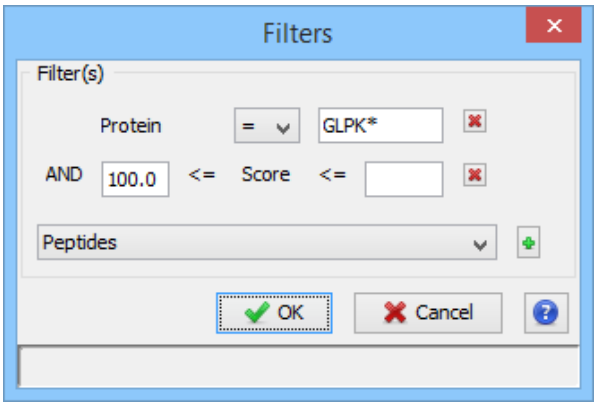


When you have clicked on the filter button, a dialog is opened. In this dialog you can select the columns of the table you want to filter thanks to the "+" button.

In the following example, we have added two filters:

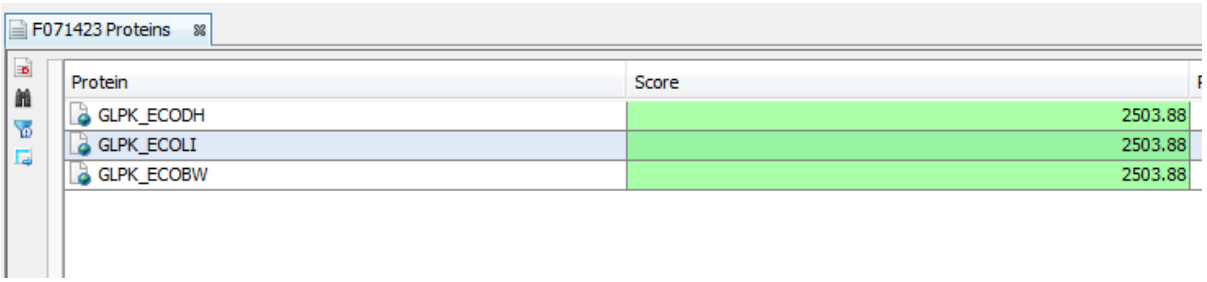
- **one on the Protein Name column** (available wildcards are * to replace multiple characters and ? to replace one character)

- **one on the Score Column** (Score must be at least 100 and there is no maximum specified).



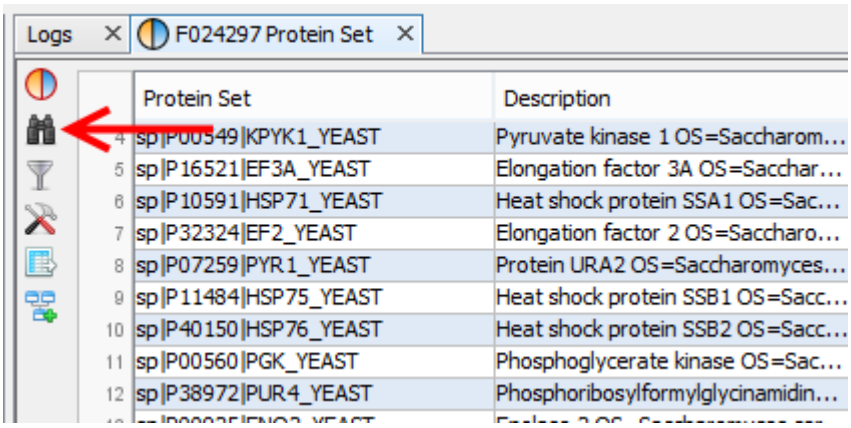
The result is all the proteins starting with GLPK (correspond to GLPK*) and with a score greater or equal than 100.

Note: for String filters, you can use the following wildcards: * matches zero or more characters, ? matches one character.



Search Tables

In some tables, a Search Functionality is available thanks to the search button at the top right corner.

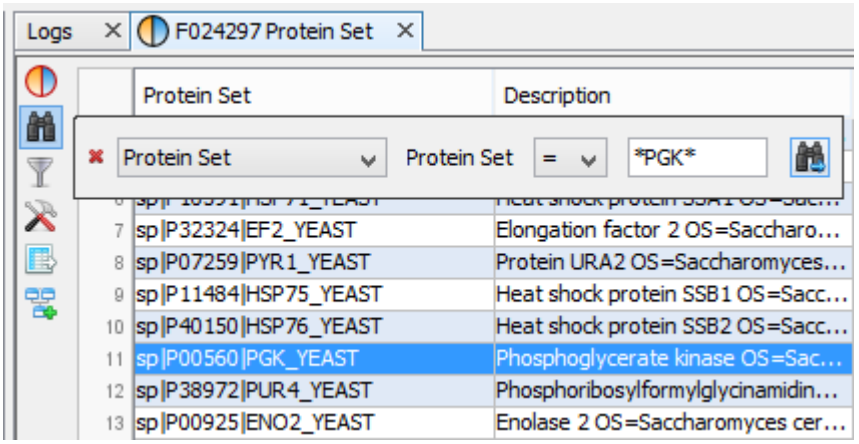


When you have clicked on the search button, a floating panel is opened. In this panel you can select the column searched and fill in the searched expression, or the value range.

For searched expressions, two wild cards are available:

- '*' : can replace all characters
- '?' : can replace one character

In the following example, the user searches for a protein set whose name contains “PGK”.



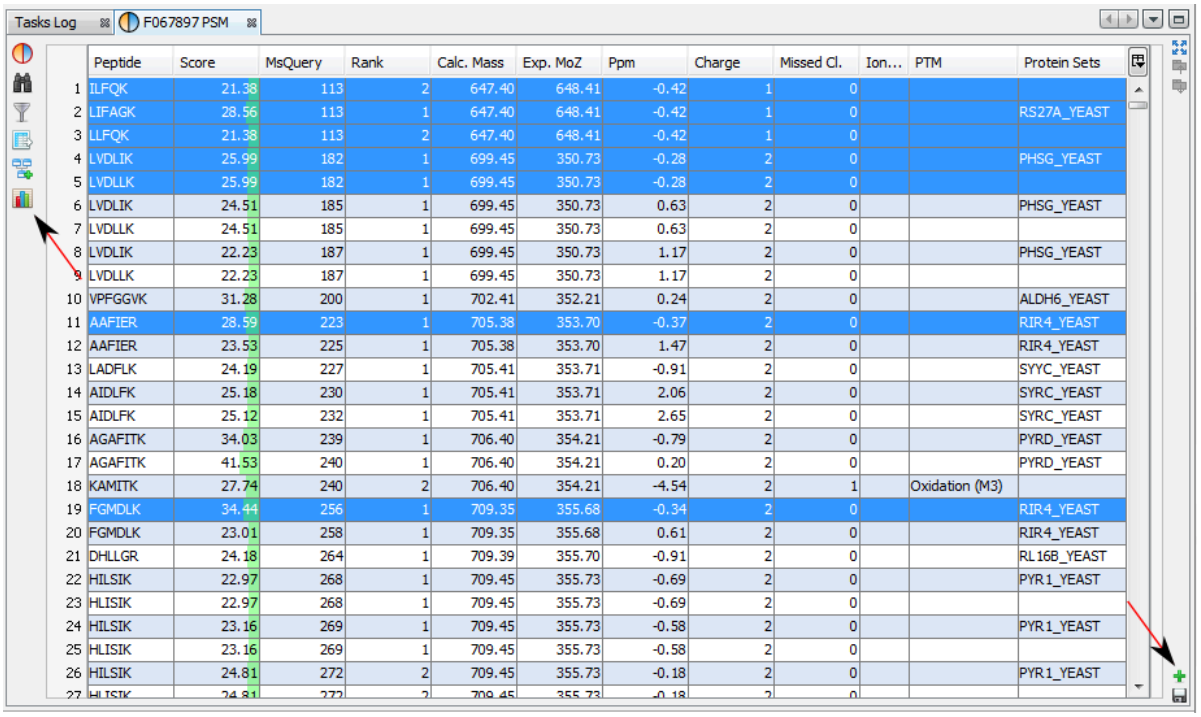
You can do an incremental search by clicking again on the search button of the floating panel, or by pressing the Enter key.

Graphics

Create a Graphic

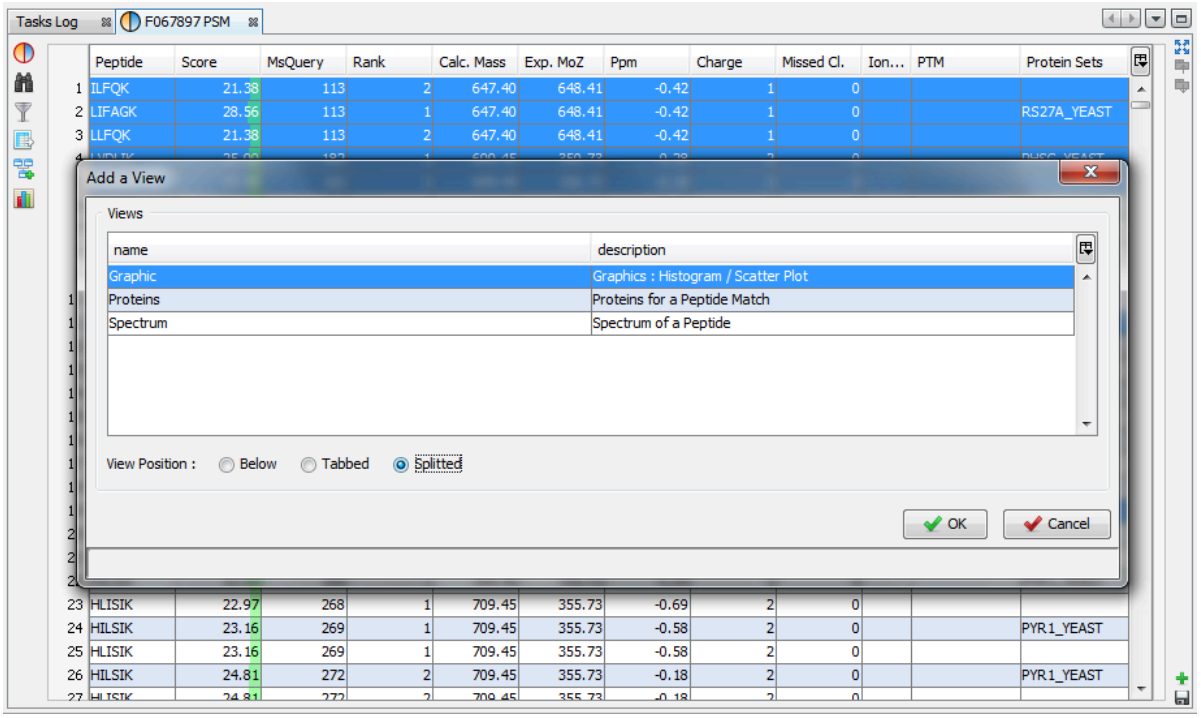
There are two ways to obtain a graphic from data:

1. In the windows with PSM of a Search Result or of an Identification Summary, you can ask for the display of a histogram in a new window to check the quality of your identification.
2. In any window, you can click on the '+' button to add a graphic (Scatter Plot or Histogram) as a view in the same window



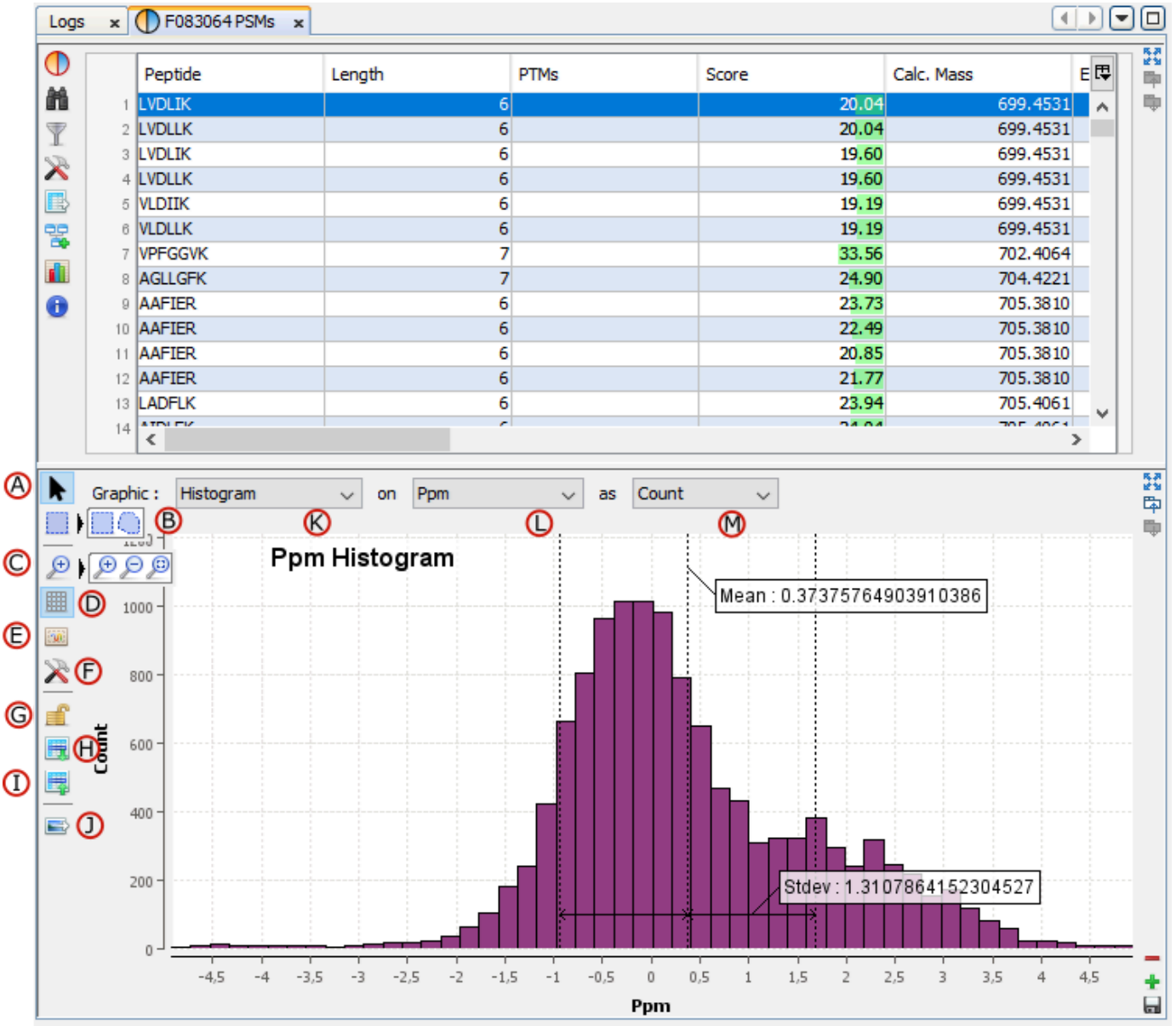
	Peptide	Score	MsQuery	Rank	Calc. Mass	Exp. MoZ	Ppm	Charge	Missed Cl.	Ion...	PTM	Protein Sets
1	ILFQK	21.38	113	2	647.40	648.41	-0.42	1	0			
2	LIFAGK	28.56	113	1	647.40	648.41	-0.42	1	0			RS27A_YEAST
3	LLFQK	21.38	113	2	647.40	648.41	-0.42	1	0			
4	LVDLIK	25.99	182	1	699.45	350.73	-0.28	2	0			PHSG_YEAST
5	LVDLLK	25.99	182	1	699.45	350.73	-0.28	2	0			
6	LVDLIK	24.51	185	1	699.45	350.73	0.63	2	0			PHSG_YEAST
7	LVDLLK	24.51	185	1	699.45	350.73	0.63	2	0			
8	LVDLIK	22.23	187	1	699.45	350.73	1.17	2	0			PHSG_YEAST
9	LVDLLK	22.23	187	1	699.45	350.73	1.17	2	0			
10	VPFGGVK	31.28	200	1	702.41	352.21	0.24	2	0			ALDH6_YEAST
11	AAFIER	28.59	223	1	705.38	353.70	-0.37	2	0			RIR4_YEAST
12	AAFIER	23.53	225	1	705.38	353.70	1.47	2	0			RIR4_YEAST
13	LADFLK	24.19	227	1	705.41	353.71	-0.91	2	0			SYYC_YEAST
14	AIDLFK	25.18	230	1	705.41	353.71	2.06	2	0			SYRC_YEAST
15	AIDLFK	25.12	232	1	705.41	353.71	2.65	2	0			SYRC_YEAST
16	AGAFITK	34.03	239	1	706.40	354.21	-0.79	2	0			PYRD_YEAST
17	AGAFITK	41.53	240	1	706.40	354.21	0.20	2	0			PYRD_YEAST
18	KAMITK	27.74	240	2	706.40	354.21	-4.54	2	1		Oxidation (M3)	
19	FGMDLK	34.44	256	1	709.35	355.68	-0.34	2	0			RIR4_YEAST
20	FGMDLK	23.01	258	1	709.35	355.68	0.61	2	0			RIR4_YEAST
21	DHLLGR	24.18	264	1	709.39	355.70	-0.91	2	0			RL16B_YEAST
22	HILSIK	22.97	268	1	709.45	355.73	-0.69	2	0			PYR1_YEAST
23	HLISIK	22.97	268	1	709.45	355.73	-0.69	2	0			
24	HILSIK	23.16	269	1	709.45	355.73	-0.58	2	0			PYR1_YEAST
25	HLISIK	23.16	269	1	709.45	355.73	-0.58	2	0			
26	HILSIK	24.81	272	2	709.45	355.73	-0.18	2	0			PYR1_YEAST

If you have clicked on the '+' button, the Add View Dialog is opened and you must select the Graphic View



Graphic options

Graphic Toolbar



A: When this button is selected, you have the "**Pointer Mode**" activated.

In this mode :

- If you move with the left mouse button pressed on the middle of the graphic, you can scroll along the X and Y Axis.

- If you move with the right mouse button pressed from the top/left corner to the bottom/right corner, a zooming rectangle is displayed. When you release the mouse button, a zoom in according to the zooming rectangle is performed.

- If you move with the right mouse button pressed from the bottom/right corner to the left/top corner, a view all is done.

B: When this button is selected, you have the "**Selection Mode**" activated.

By clicking on the black right arrow, you can switch between the square selection mode and the lasso selection mode.

In this mode:

- Use the left mouse button to do a square or lasso selection.
- The right mouse button works as with the “Pointer Mode”

C: Zoom out / Zoom in / View all. Click on the black right arrow to select the zooming mode.

D: Display/Remove Grid toggle button

E: Display/Hide View All Map. The goal of this map is to display the whole graphic in a small zone even when you have zoomed

F: Open a settings dialog for the graphic. You can modify for example colors or bins of an histogram.

G: Lock/Unlock incoming data. If it is unlocked, the graphic is updated when the user applies a new filter to the previous view (for instance Peptide Score >= 50) If it is locked, changing filtering on the previous view does not modify the graphic.

H: Select Data in the graphic according to data selected in the table in the previous view.

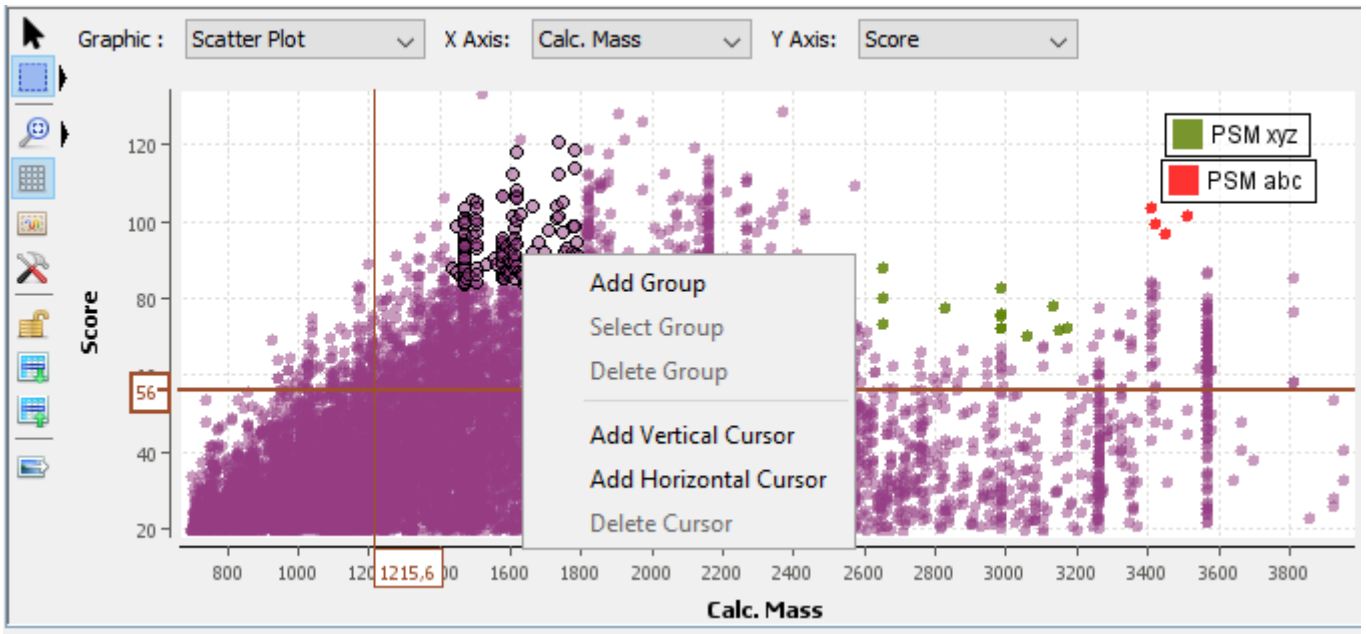
I: Select data in the table of the previous view according to data selected in the graphic.

J: Export graphic to image

K: Select the graphic type: Scatter Plot / Histogram

L/M: Select data used for X / Y axis.

Popup Menu on Graphic Area

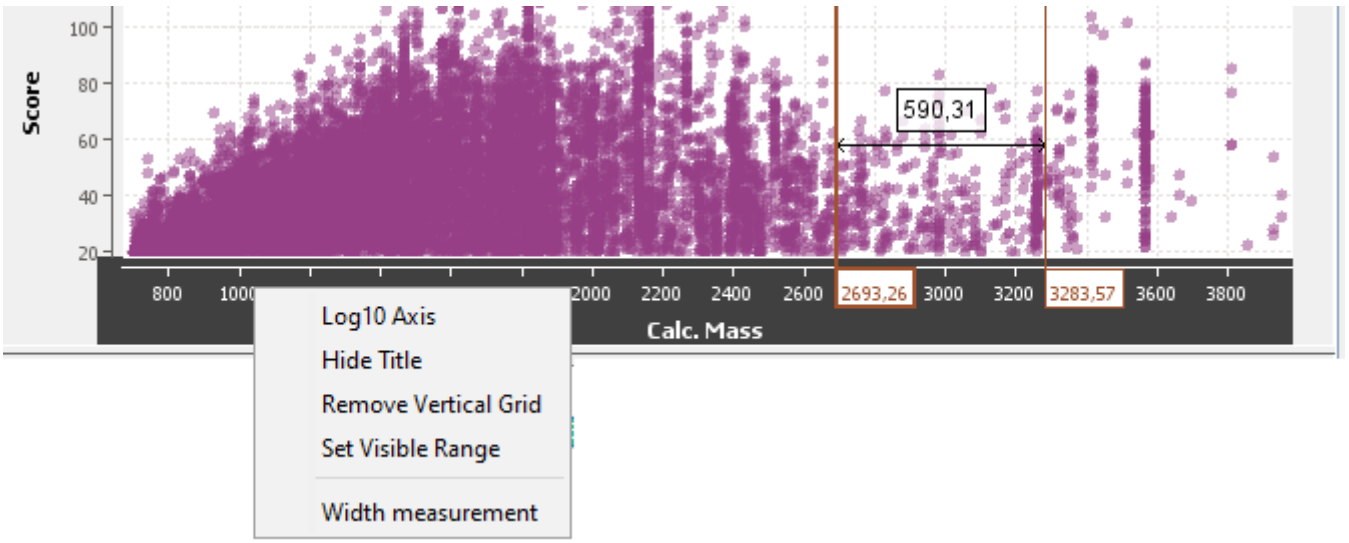


By right mouse click on the graphic area, you get a popup with several menus:

- Add Group: after selecting some data, you can create a group defined by a name and a color (there are two groups in the example : PSM xyz and PSM abc)
- Select Group: Right click on a data group to select them.
- Delete Group: Right click on a data group and delete it
- Add Vertical Cursor: Add a vertical cursor, its X position is displayed on the X Axis.
- Add Horizontal Cursor: Add an horizontal cursor, its Yposition is displayed on the Y Axis.

- Delete Cursor: Right click on a cursor and delete it.

Popup Menu on Axis



By right mouse click on an axis, you get a popup with several menus:

- Log 10 / Linear Axis : toggle the display of the axis between log and linear.
- Hide / Display Title : hide the axis title if you want more space for the graphic
- Remove / Add Vertical Grid
- Set Visible Range : to set the minimum and maximum values of the axis
- Width measurement : display 2 cursors and the measurement of the space between the cursors.

Zooming

There are several ways to perform zoom actions.

Zoom in:

- Click on the Zoom in button in the toolbar.
- Press the right mouse button and drag to the right bottom direction. A red box is displayed. Release the mouse button when you have selected the area to zoom in.
- Put the mouse over an axis and roll upward the mouse wheel. The zoom in is centered around the middle of the axis. If you press on Ctrl Key at the same time, the zoom in is performed around the mouse position.

Zoom out:

- Click on the Zoom out button in the toolbar.
- Put the mouse over an axis and roll downward the mouse wheel. The zoom out is centered around the middle of the axis. If you press on Ctrl Key at the same time, the zoom out is performed around the mouse position.

View All:

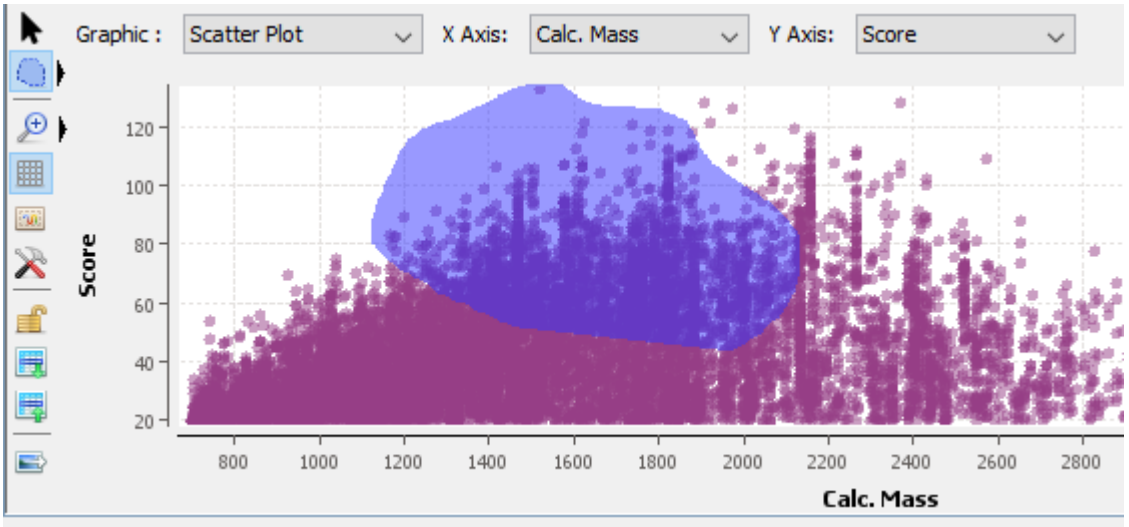
- Click on the View All button in the toolbar.
- Press the right mouse button and drag to the left top direction. When you release the mouse button, the zooming is reset to view all.

Selecting

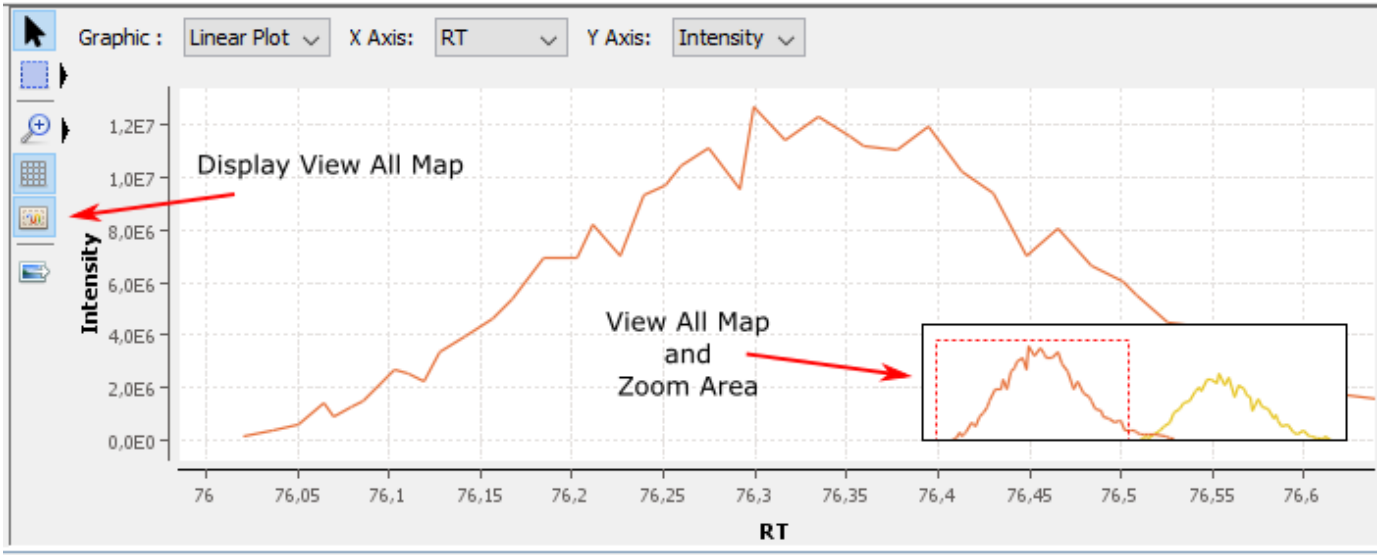
To be able to select data, you must be in “Selection Mode”

Select: You can select with a rectangle area or with a lasso according to the selected button. Press the left mouse button and drag the mouse to surround the data you want to select. When you release the button, the selection is done. Or left click on the data you want to select. It is possible to use the Ctrl key to add to the previous selection.

Unselect: Left click on an empty area to clear the selection.



View All Map



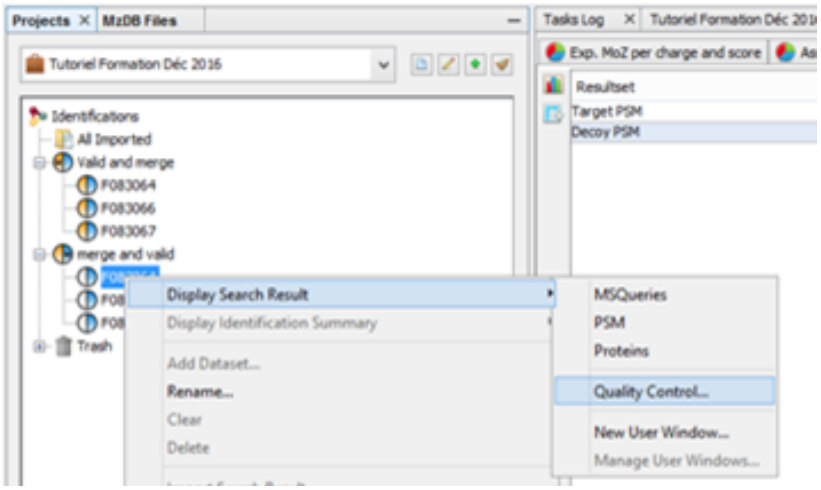
Click on View All Map button to display the map. This map always displays the whole graphic and the zoomed area. You can directly zoom on the view all map. You can resize it and move it.

Quality Control

Search Result QC

You can run a Quality Control on any leaf Search Result, that is to say an imported Result File not a merged search result. It consists in a transversal view of the imported data: rather than visualising the results per PSM or Proteins, results are sorted according to the score, charge state...

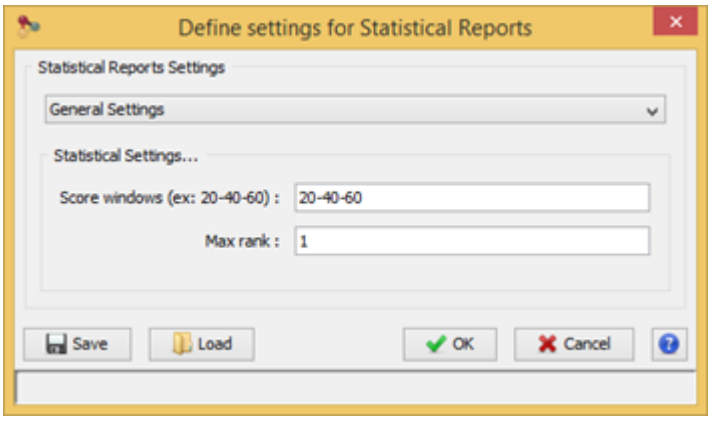
Choose the menu option:



Settings

Configure some settings before launching the process

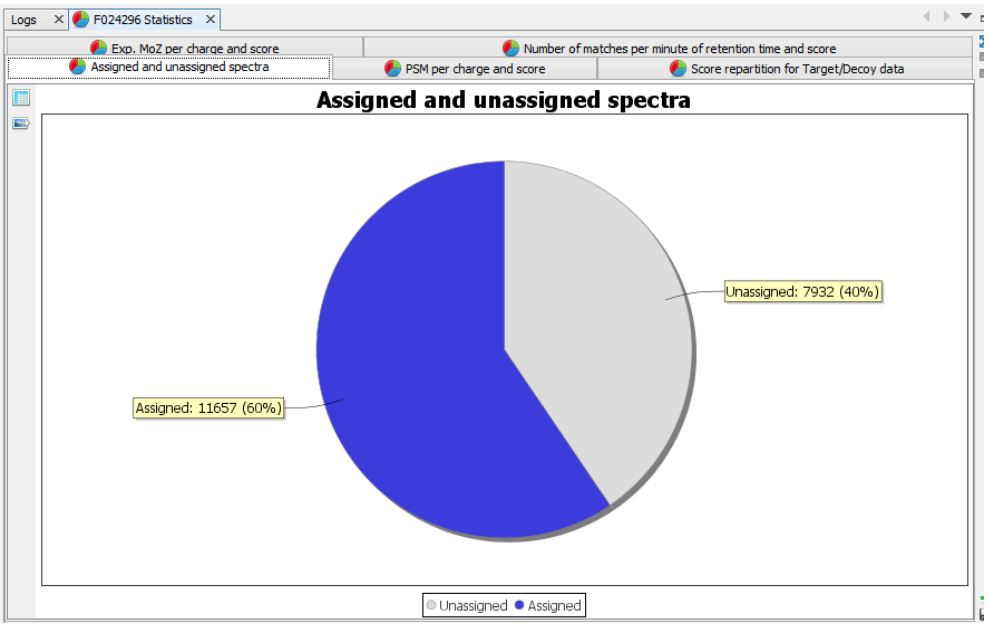
- Score windows: you can split your data in different groups based on the score. The default groups are : less than 20, between 20 and 40, between 40 and 60, over 60
- Max rank: data can be filtered to get a view focused on the best ranks. Default is to consider only the first rank.



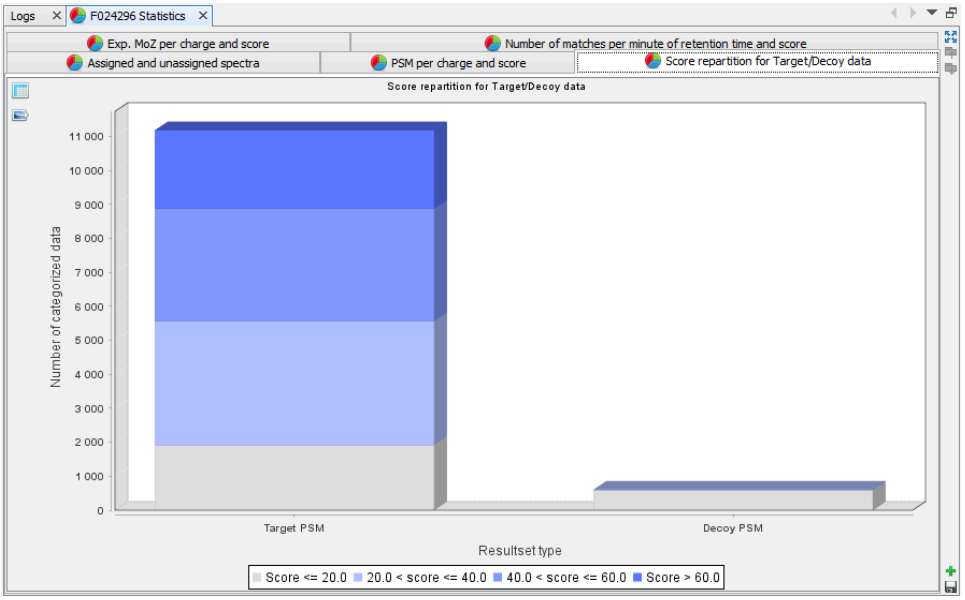
QC results

The report will appear in a matter of seconds (depending on the amount of data to be processed). You will get the following tabs:

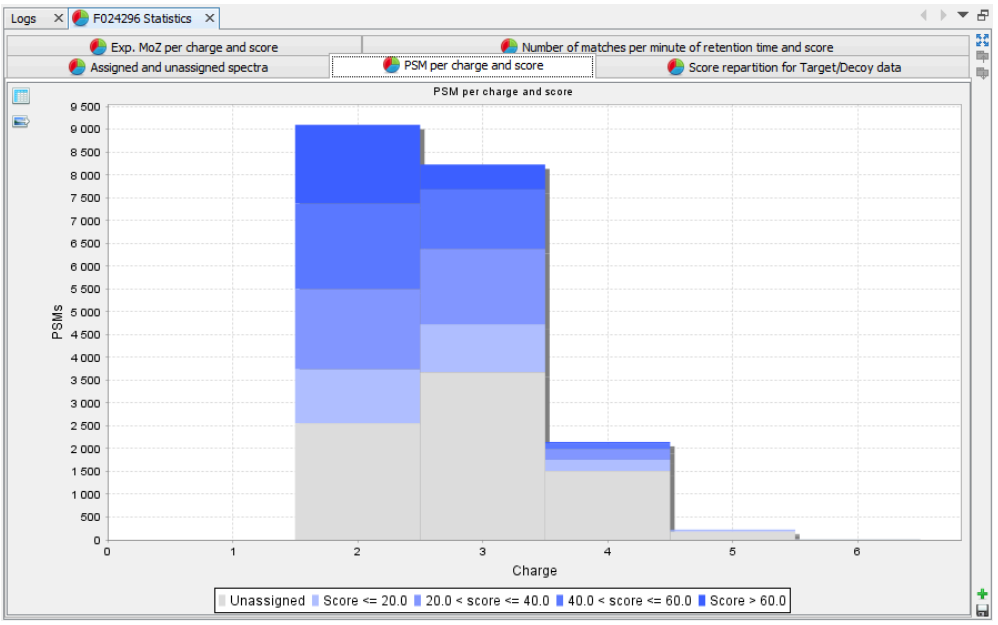
Assigned and unassigned spectra: Pie chart presenting the ratio of assigned spectra



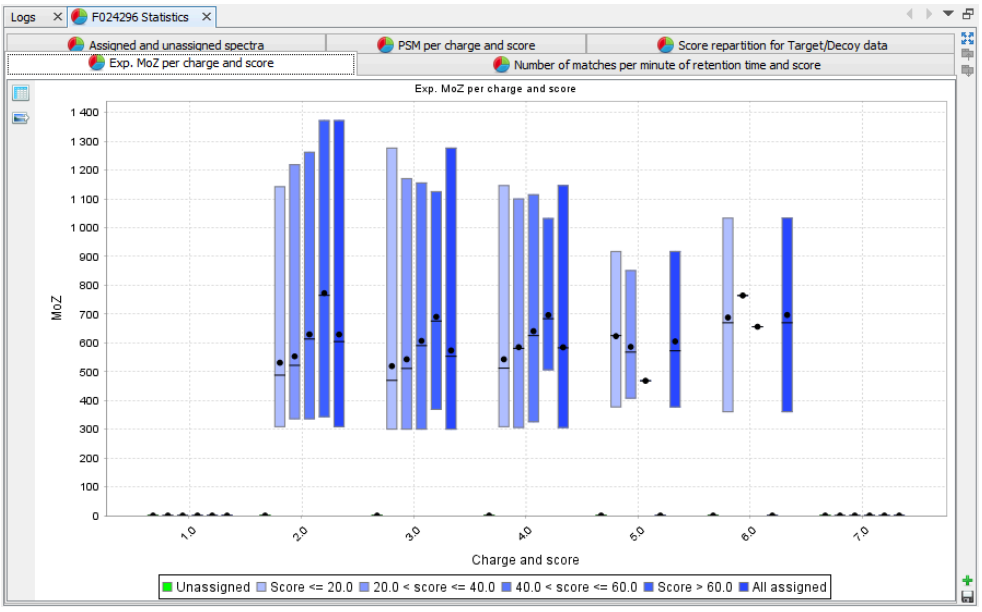
Score repartition for Target/Decoy data: Histogram presenting the amount of PSM per group of score, separating target and decoy data



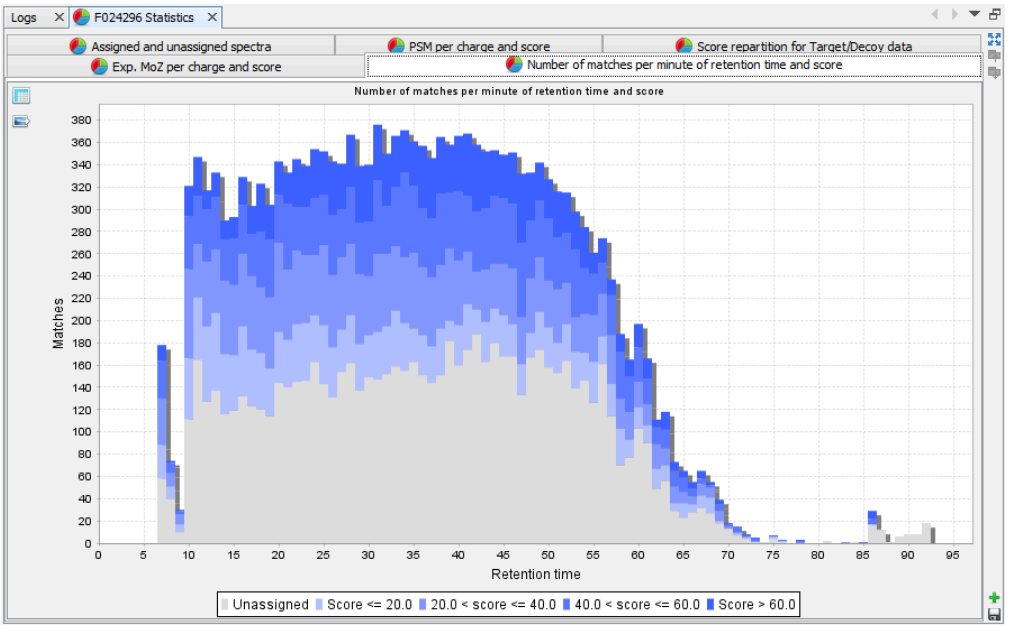
PSM per charge and score: Histogram presenting the amount of PSM per group of score and charge state



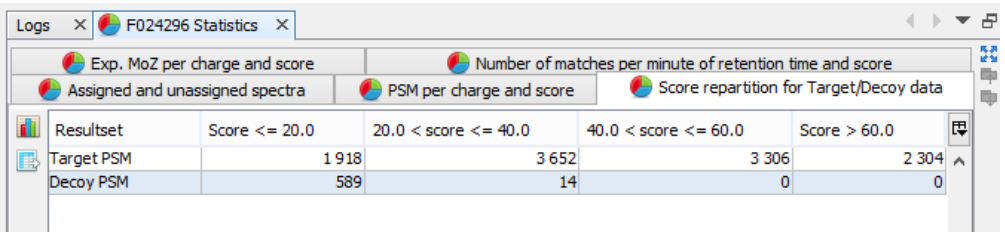
Experimental M/z per charge and score: Box plot presenting M/z information for each category of score and charge state



Number of matches per minute of RT and score: histogram presenting the amount of PSM per score and retention time. This view is only calculated when retention time is available.



Each graph is also available in a table view



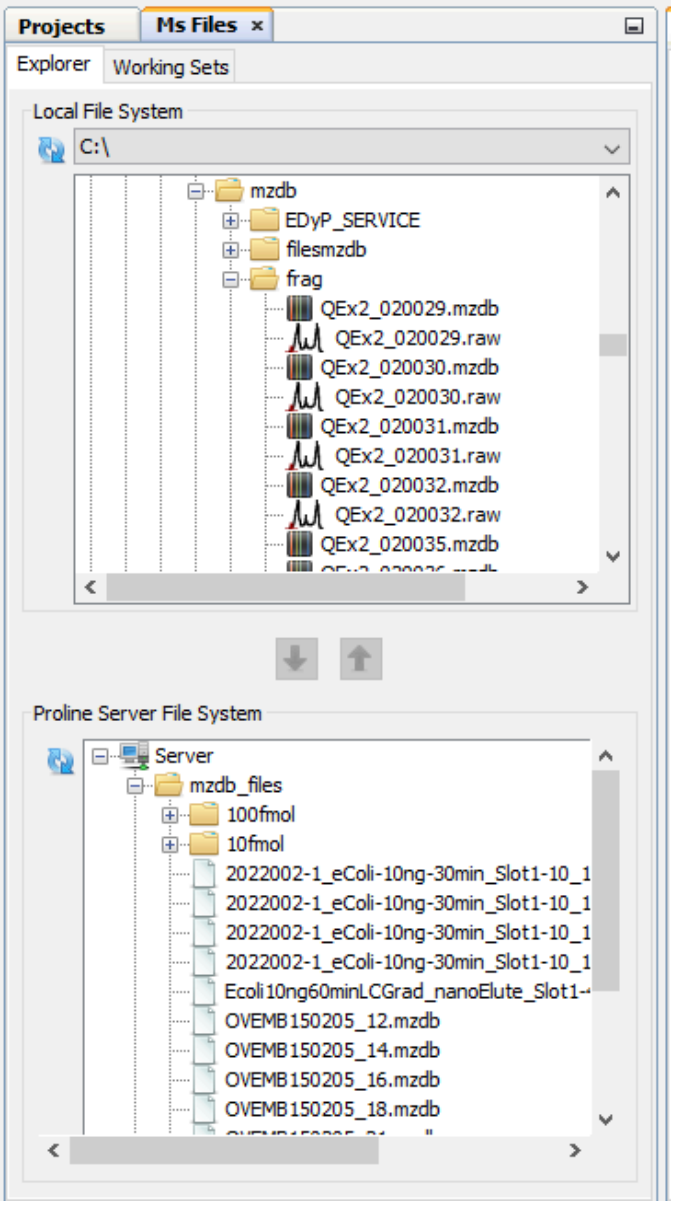
Resultset	Score <= 20.0	20.0 < score <= 40.0	40.0 < score <= 60.0	Score > 60.0
Target PSM	1 918	3 652	3 306	2 304
Decoy PSM	589	14	0	0

Ms Files Tab

In order to facilitate different actions on Ms Files, Proline Studio contains an homonym tab providing the end user with a view over his local and server remote file system, called Local File System and Proline Server File System respectively.

Furthermore, on local file system a series of actions can take place, through an appropriate popup menu, on the encountered .mzdb and .raw files, including among others the:

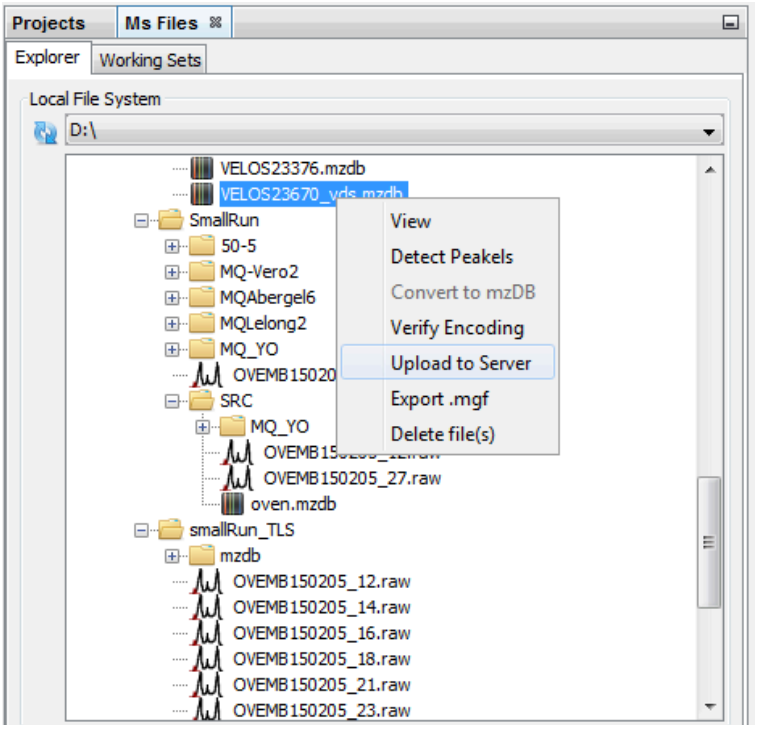
- Conversion of a .raw file to an .mzdb file
- The upload of an .mzdb file
- View .mzdb files
- Detect its peakels



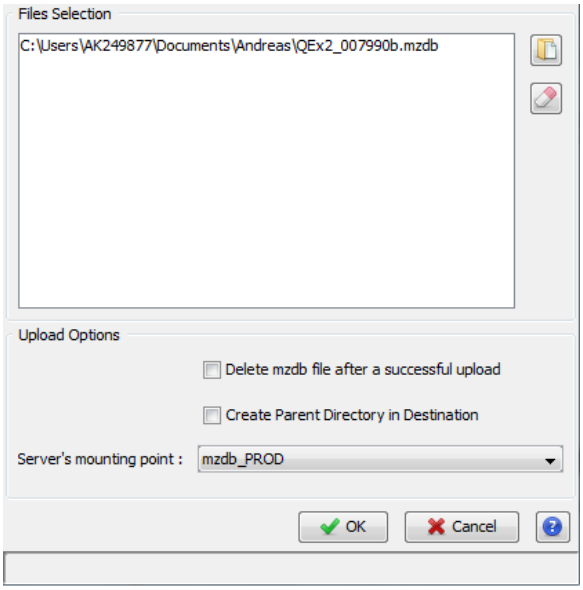
Apart from the popup menu supported functionality, since Proline Studio 1.5, uploads can be triggered via drag and drop mechanism.

mzDB File Upload

As mentioned earlier, after selecting a number of files, the user can either drag and drop them inside the remote site, or use the popup menu as shown in the following screenshot. It is important to precise that both approaches are not compatible with a selected group consisting of different file types.



As we can see, clicking on upload opens a dedicated dialog packing a series of uploading options:

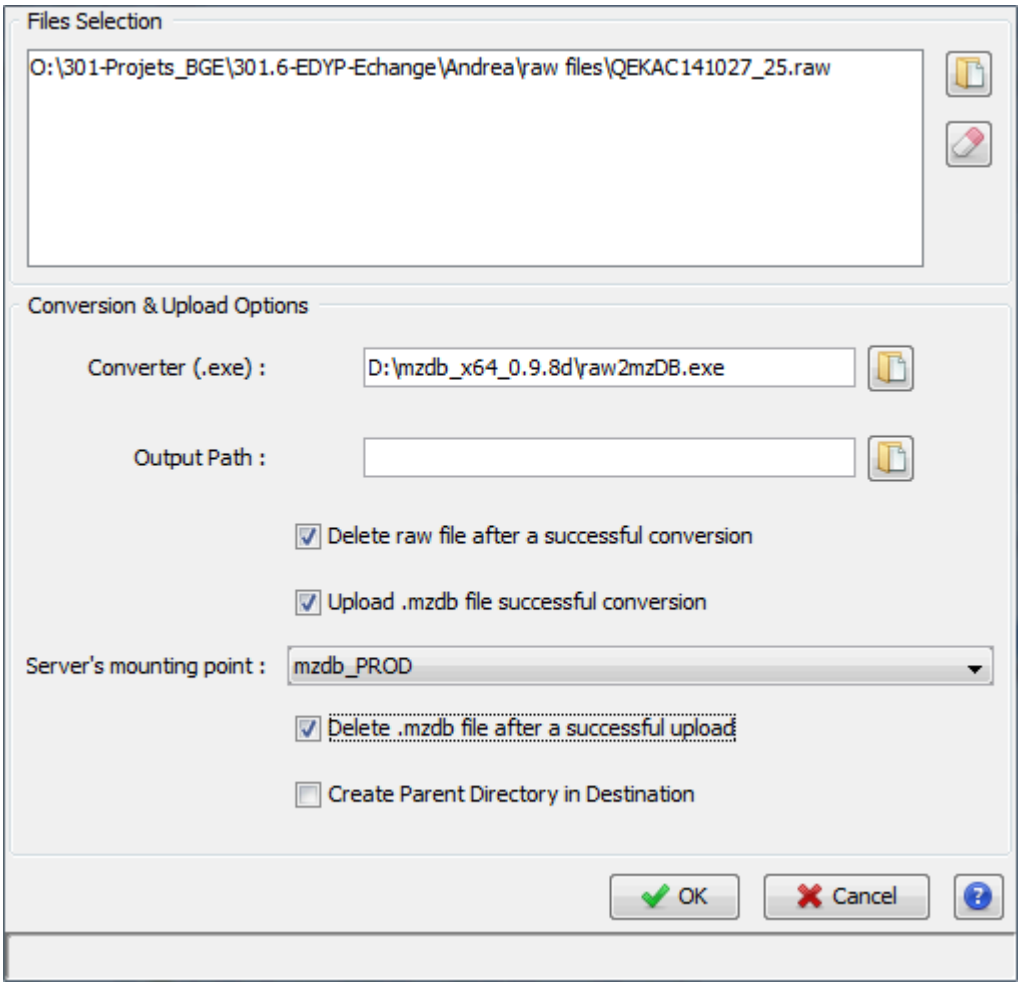


- The deletion of the file(s) after the successful upload
- The creation of the file(s) parent directory in destination: only the direct parent folder will be created. Otherwise, the file will be uploaded at the mount point root.
- The mounting point at the server: depending on server mounting points configuration.

Furthermore, the dialog permits us to add or remove .mzdb files to upload. The uploading tasks status could be viewed in the Logs tab.

Raw File Conversion

In the same way, when the user desires to convert and upload a raw file, he or she can use the respective dialog through the popup menu.



For the upload step, the same options as described above are shown.

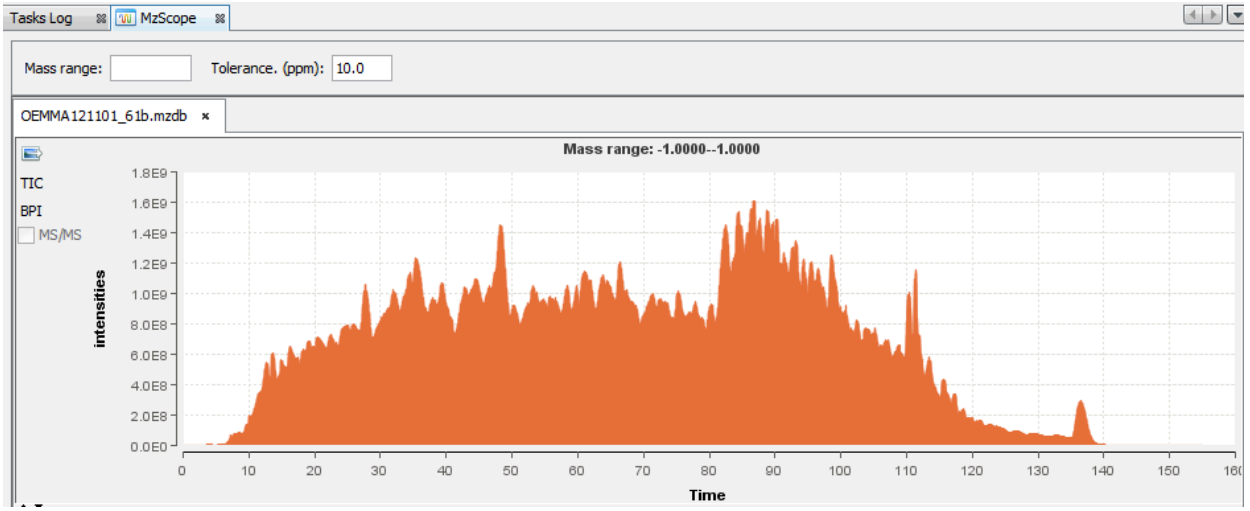
For the conversion, the path to the converter exe file should be specified. This value will be saved upon different executions. A default path may be specified in the general settings dialog. The same way mzdb file may be deleted after a successful upload, raw files could be deleted after a conversion .

TIC or BPI chromatogram

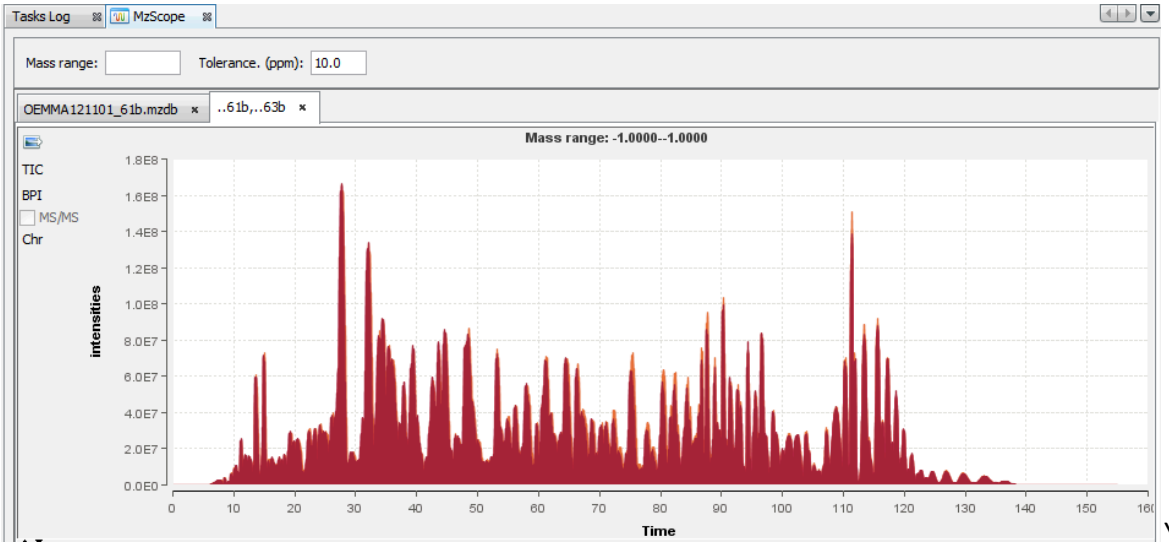
When the user chooses to “View” an mzDB file, the MzScope window is opened.

By default, the TIC chromatogram is displayed. You can click on “BPI” to see the best peak intensity graph.

By clicking in the graph, you can see below the scan at the selected time.

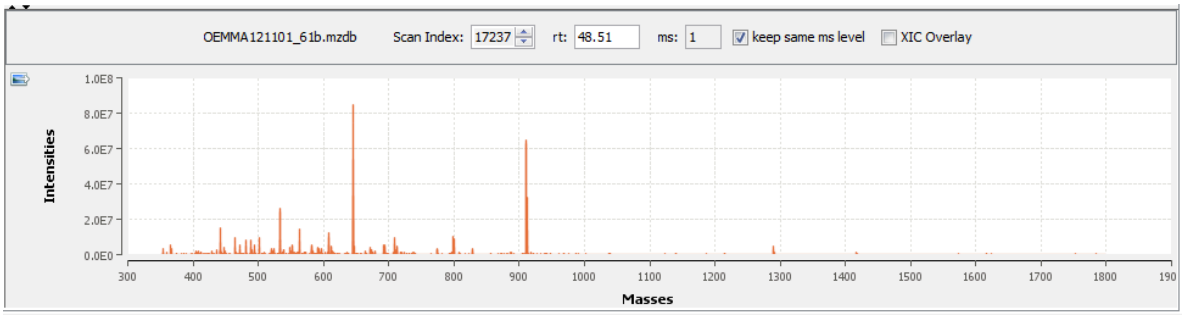


You can choose to display 2 or more chromatograms on the same graph, by selecting 2 files and clicking on “View”



You can extract a chromatogram at a given mass by entering the specified value in the panel above.

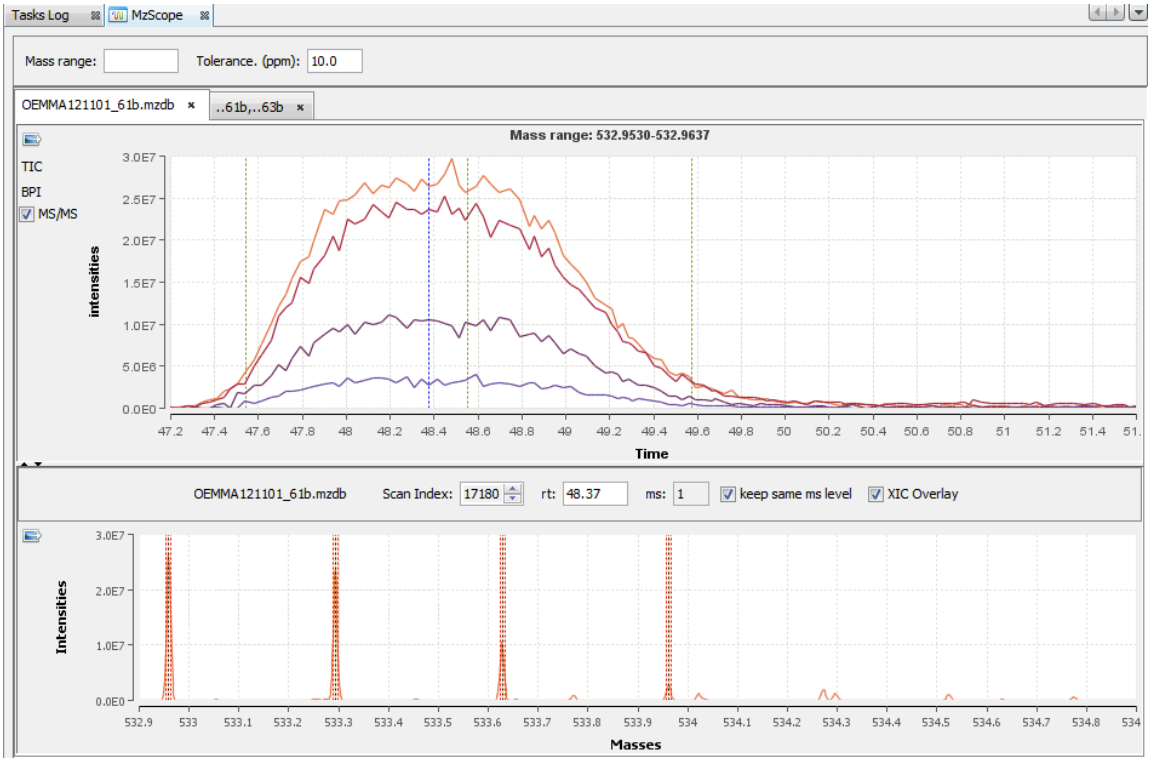
Scan



You can navigate through the scans

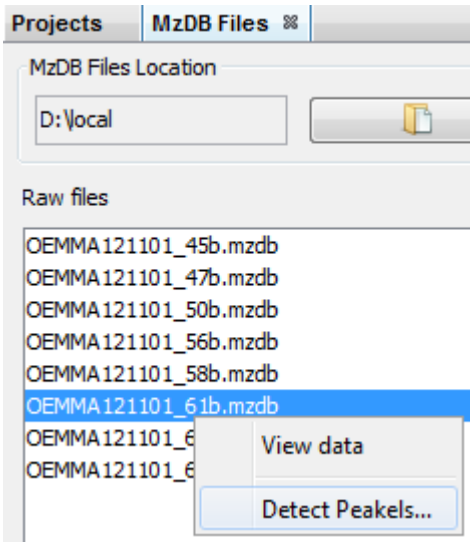
- by increasing or decreasing the scan Ids
- by entering a retention time
- by clicking the keys arrows on the keyboard (Ctrl+Arrows to keep the same ms level)

By double clicking on the scan, the corresponding chromatogram is displayed above (The Alt key or the check box “XIC overlay” allows you to overlay the chromatograms in the same graph).

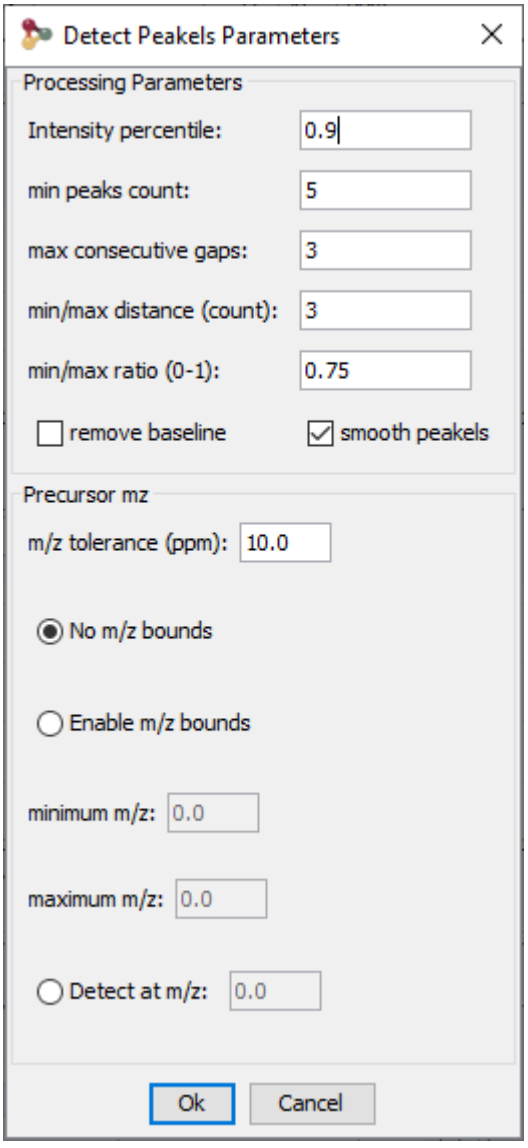


Peakels

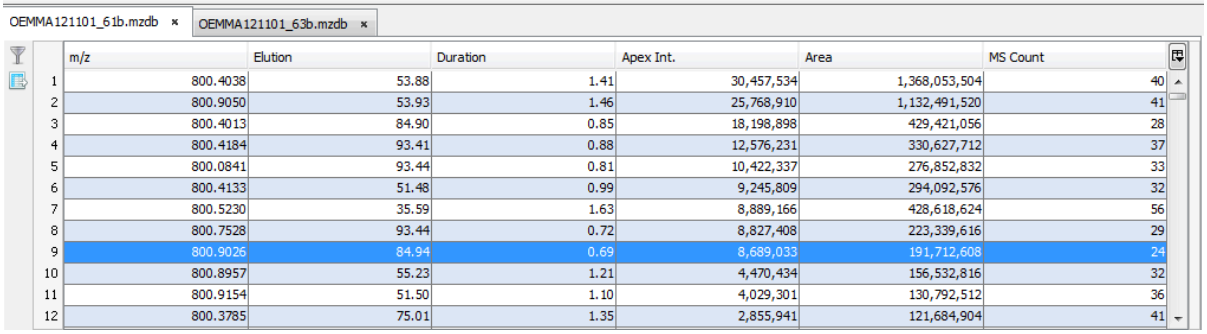
By selecting a file, you can click on “Detect Peakels” in the popup menu.



A dialog allows you to choose the parameters of the peakels detection: the tolerance and eventually a range of m/z, or a m/z value:

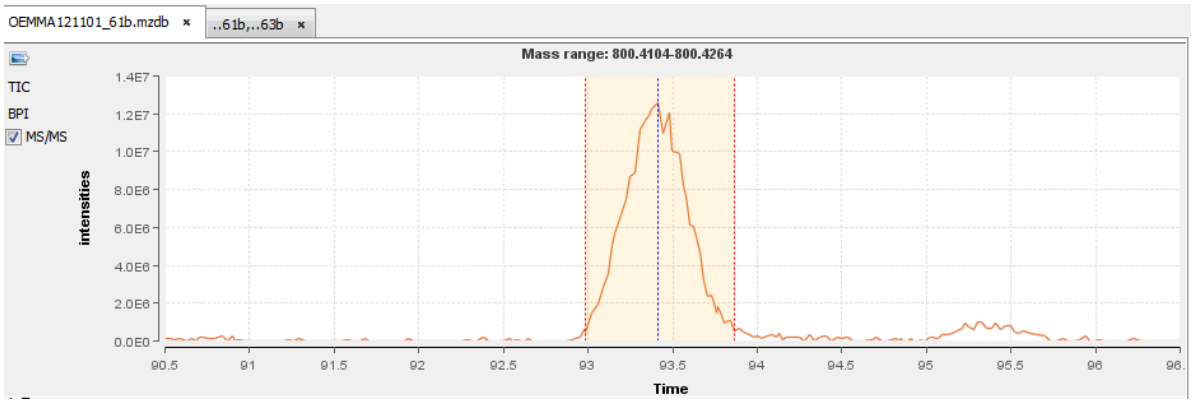


The results are displayed in a table:



	m/z	Elution	Duration	Apex Int.	Area	MS Count
1	800.4038	53.88	1.41	30,457,534	1,368,053,504	40
2	800.9050	53.93	1.46	25,768,910	1,132,491,520	41
3	800.4013	84.90	0.85	18,198,898	429,421,056	28
4	800.4184	93.41	0.88	12,576,231	330,627,712	37
5	800.0841	93.44	0.81	10,422,337	276,852,832	33
6	800.4133	51.48	0.99	9,245,809	294,092,576	32
7	800.5230	35.59	1.63	8,889,166	428,618,624	56
8	800.7528	93.44	0.72	8,827,408	223,339,616	29
9	800.9026	84.94	0.69	8,689,033	191,712,608	24
10	800.8957	55.23	1.21	4,470,434	156,532,816	32
11	800.9154	51.50	1.10	4,029,301	130,792,512	36
12	800.3785	75.01	1.35	2,855,941	121,684,904	41

You can double-click (or through the popup menu) on a row to display the peakel in the corresponding raw file:

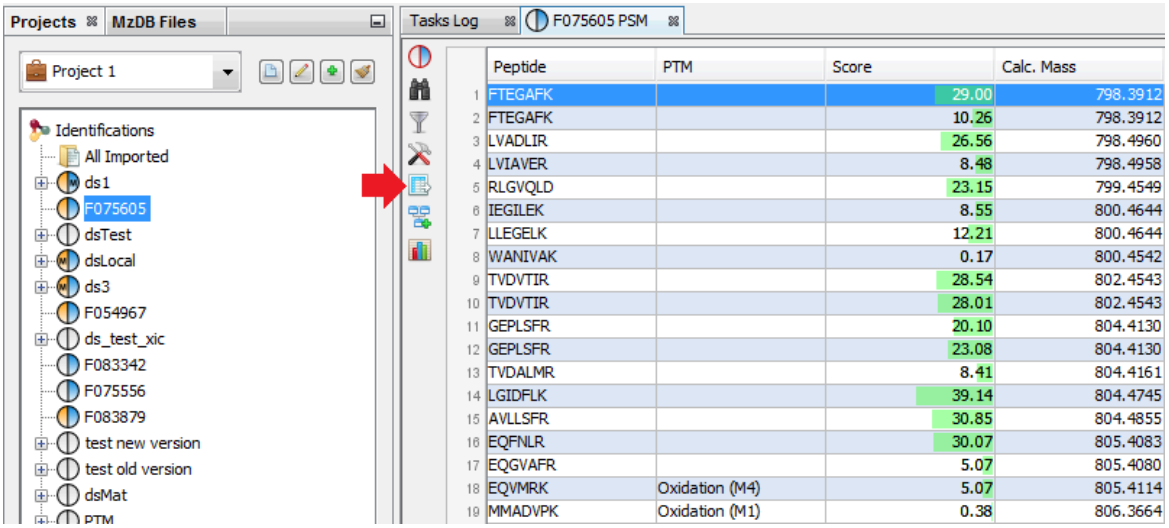


Export Data / Image

There are many ways to do an export:

- Export a Table using the export button (supported formats: {xlsx, xls, csv})
- Export data using Copy/Paste from the selected rows of a Table to an application like Excel.
- Export all data corresponding to an Identification Summary, XIC or Spectral Count
- Export an image of a view
- Export Identification Summary data into MzIdentML format (for ProteomeXchange) .
- Export Identification Summary spectra list.

1. Export a Table

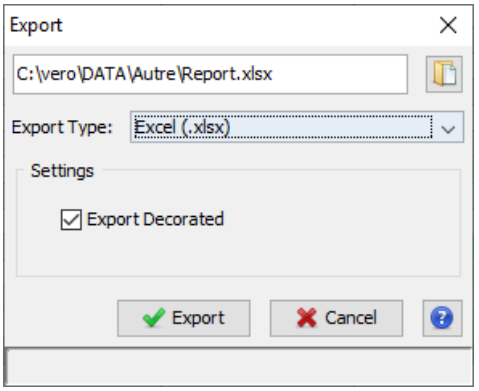


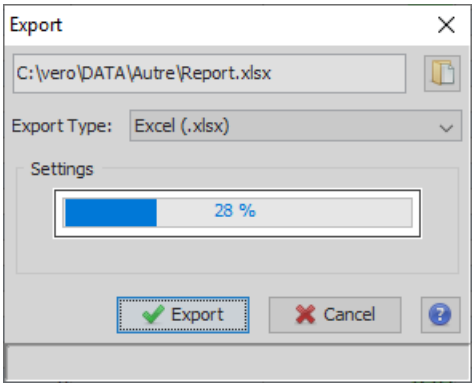
	Peptide	PTM	Score	Calc. Mass
1	FTEGAFK		29.00	798.3912
2	FTEGAFK		10.26	798.3912
3	LVADLIR		26.56	798.4960
4	LVIAVER		8.48	798.4958
5	RLGVQLD		23.15	799.4549
6	IEGILEK		8.55	800.4644
7	LLEGELK		12.21	800.4644
8	WANIVAK		0.17	800.4542
9	TVDVTIR		28.54	802.4543
10	TVDVTIR		28.01	802.4543
11	GEPLSFR		20.10	804.4130
12	GEPLSFR		23.08	804.4130
13	TVDALMR		8.41	804.4161
14	LGIDFLK		39.14	804.4745
15	AVLLSFR		30.85	804.4855
16	EQFNLR		30.07	805.4083
17	EQGVAFR		5.07	805.4080
18	EQVMRK	Oxidation (M4)	5.07	805.4114
19	MMADVPK	Oxidation (M1)	0.38	806.3664

To export a table, click on the Export Button at the left top of a table.

An export dialog is opened, you can select the file path for the export and the format of the export (supported formats: {xlsx, xls, csv}).

In case that the selected format is either .xls or .xlsx, the user has now the ability to maintain in his exported excel document any rich text format elements (color, font weight etc.) apparent on the original table in Proline Studio. Choice is done using the checkbox shown on the following screenshot.





To perform the export, click on the Export Button. The task can take a few seconds if the table has a lot of rows and so a progress bar is displayed.

2. Copy/Paste a Table

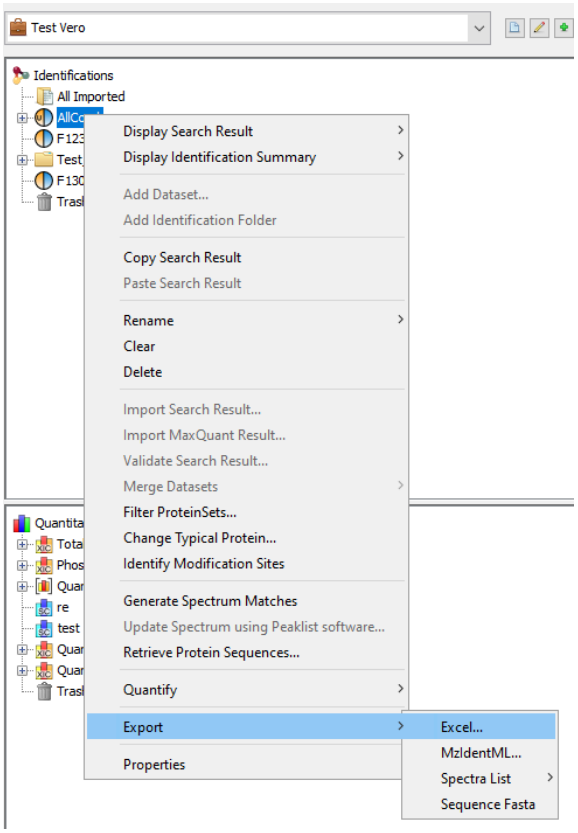
To copy/Paste a Table:

- Select rows you want to copy

- Press Ctrl and C keys at the same time. The column titles are also copied

- Open your spreadsheet editor and press Ctrl and V keys at the same time to paste the copied rows. If paste is done in a text editor, the column separator used is the tabulations.

3. Export an Identification Summary, a XIC or a Spectral Count

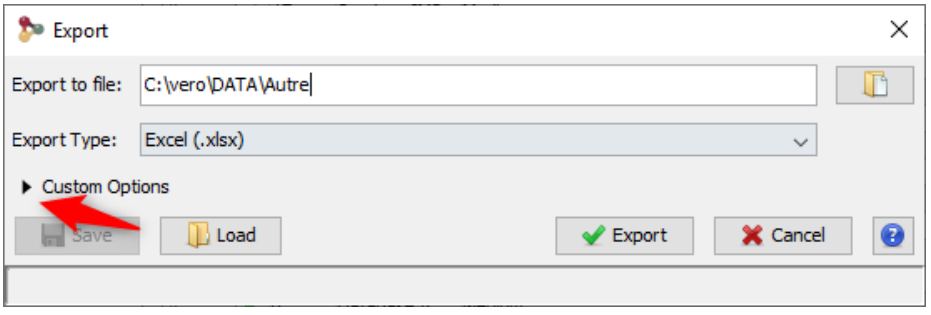


To Export all data of a dataset (Identification Summary, XIC or Spectral Count), right-click on the dataset to open the contextual menu and select the “Export” menu and then “Excel...” sub-menu.

You can also export multiple dataset simultaneously, if they have the same type (Identification Summary or XIC or Spectral Count).

An Export Dialog is opened, you can select the file path and the type of the export : Excel (.xlsx) or Tabulation separated values (.tsv).

You can export with the default parameters or perform a custom export. To enable custom export, click on the tick box located on the right of the dialog:



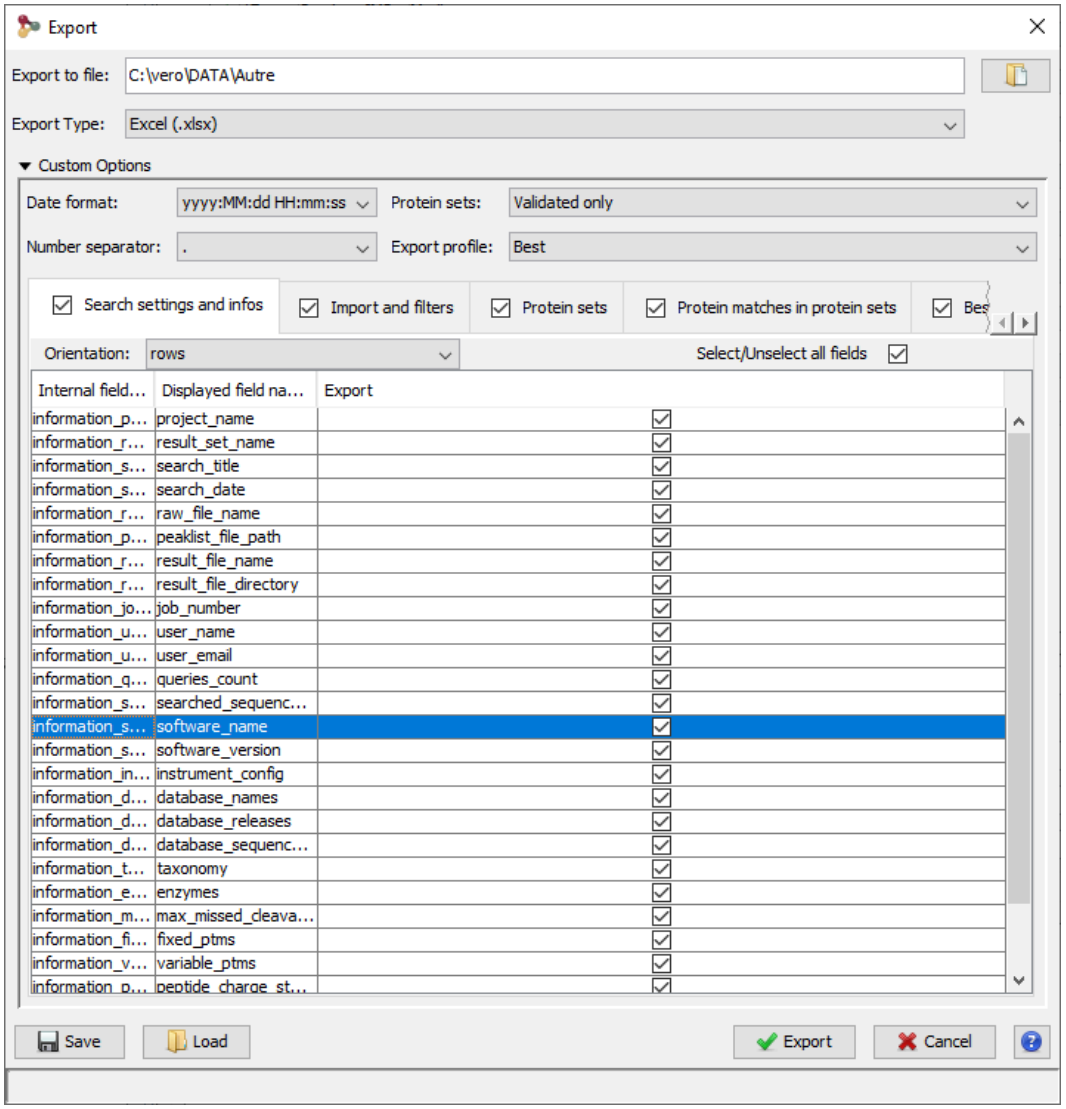
Custom export allows a number of parameters in addition to the file format to be chosen.

- Sheets to be created. You can **enable/disable** sheets to be exported, **rename** them (by double clicking on the title) or **change their order** by drag and drop. Available sheets are dependent on the identification summary type, but you can export which parameters were used (search settings or import and filters), the list of protein sets, all proteins, PSMs… , modification Clusters or statistics data.
- For each sheet, you can define the fields to be exported: change the title used in the exported file if needed, change the order of the fields using drag and drop of the corresponding row at the desired location.
- Symbol to use for decimal output (comma ',' or point '.')
- Timestamp format to use

- Protein sets to export ('all' or 'validated only')
- Export profile ('best' or 'all')

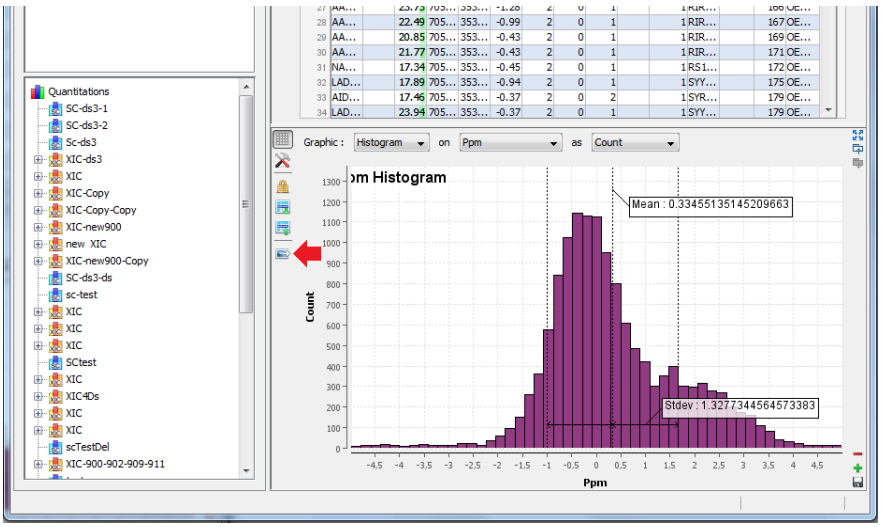
Description of the exported file is available here.

Note: The parameters could be saved and loaded in further export using the Save / Load buttons of the export dialog.

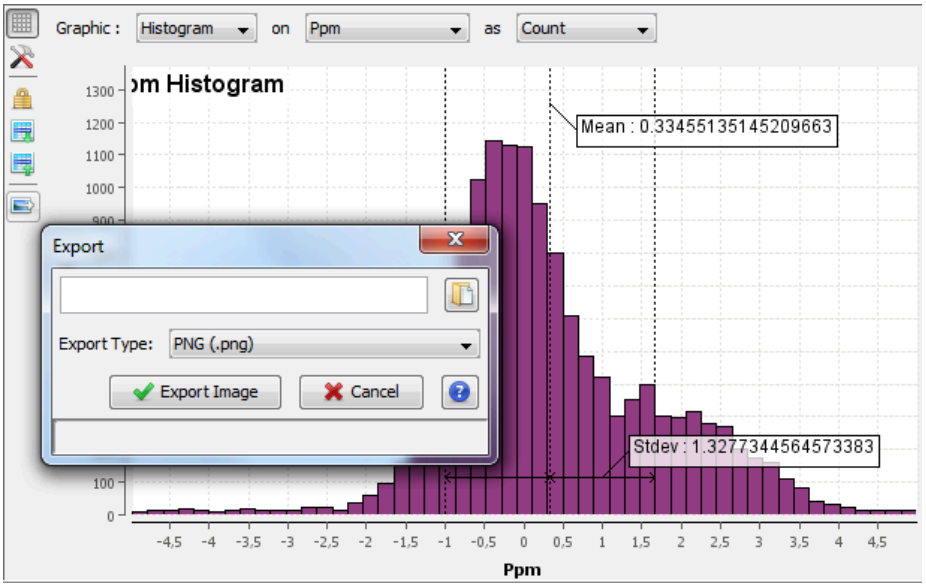


4. Export an Image

Any graphics in proline can be exported. Click on the Export Image Button at the left top of the image.



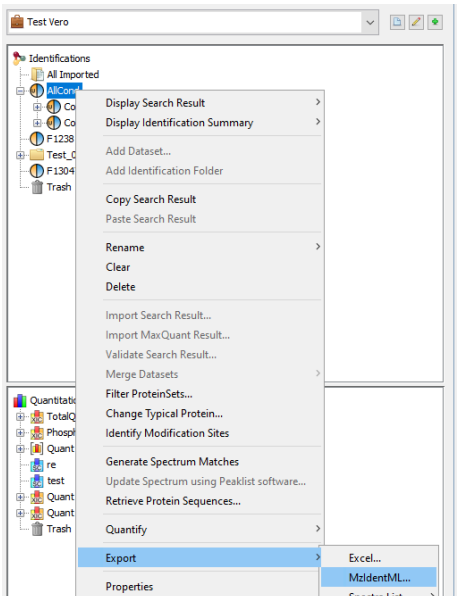
An Export Dialog is opened where you can select the file path and the export type. Available formats are PNG or SVG formats. SVG format produces a vector image that can be edited and resized afterwards.



5. Export Identification Summary for ProteomeXchange

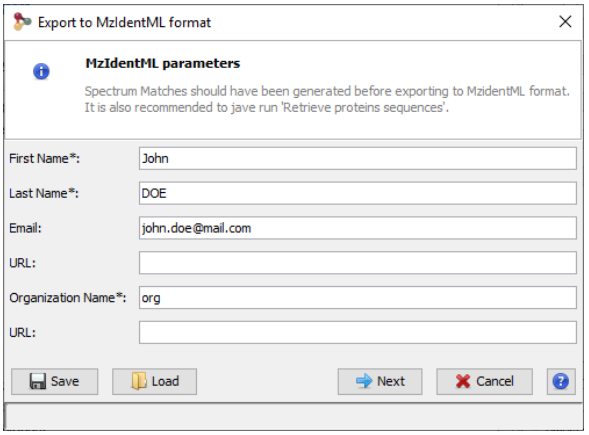
Actually it is possible to export Identification Summary into MzIdentML, Pride isn't supported any more.

Note: Before exporting data all spectrum matches should have been generated. To do so, right click on the dataset and select "Generate Spectrum Matches".



Right click on the dataset you want to export and select the "Export" menu and then "MzIdentML..." sub-menu

A dialog is opened where user information may be specified (name, organization …)



The file name and path should be specified in the next step. A progress bar is shown until the file is generated. The generated file contains identification and validation data issues from the dataset. All meta information including instrument configuration as well as search engine parameters are also extracted from dataset associated data.

6. Export Identification Summary spectra list

To export valid PSM Spectra from an Identification Summary or from a XIC Dataset. The exported tsv file is compatible with Peakview.

Note: all Spectrum Matches must be generated first.

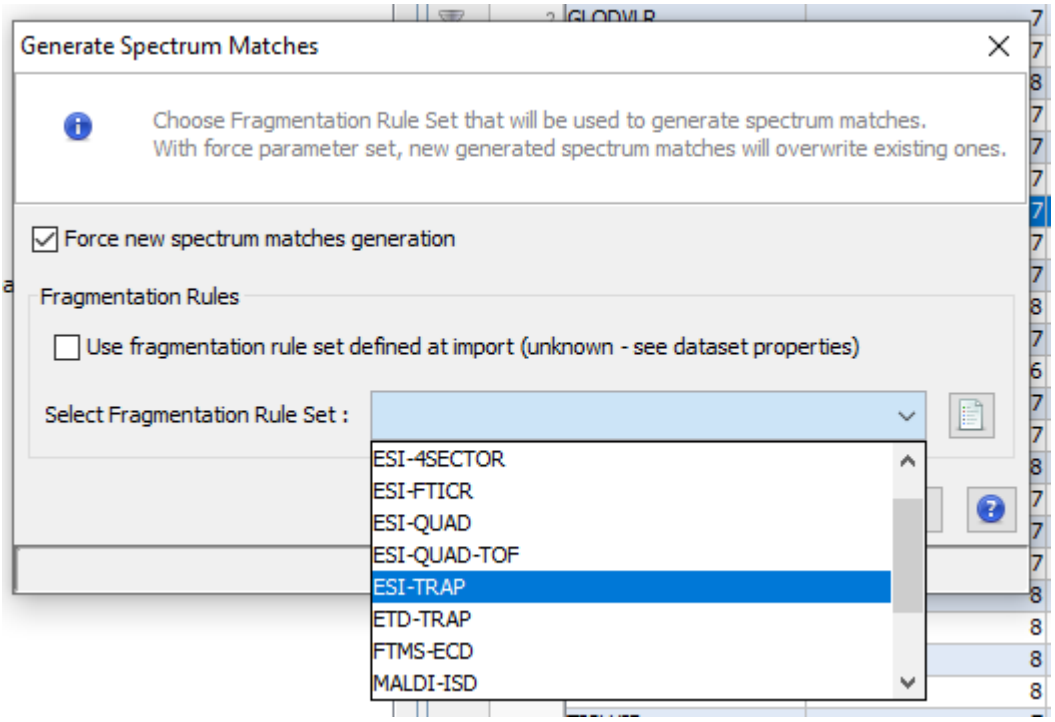
Generate Spectrum Matches

When importing a Search Result in Proline, users can view PSM with their associated Spectrum but by default no annotation is defined. Users need to generate (and save) this information explicitly.

- For a single PSM, select the icon near the Spectrum (see Display Peptide and PSM)
- For a whole Search Result, Identification Summary or Quantitation Result :
    - right click on a Dataset
    - select “Generate Spectrum Matches”

In both cases, the following dialog will be opened. User can

- Choose to force spectrum matches to be calculated even if it is defined. This could be useful if previous spectrum matches generation was done with the wrong rules set.
- Use the fragmentation rules set defined at import. In some cases, this information is not accessible and the user should then specify it.
- Specify the fragmentation rules set to use without taking into account the rules specified at import.



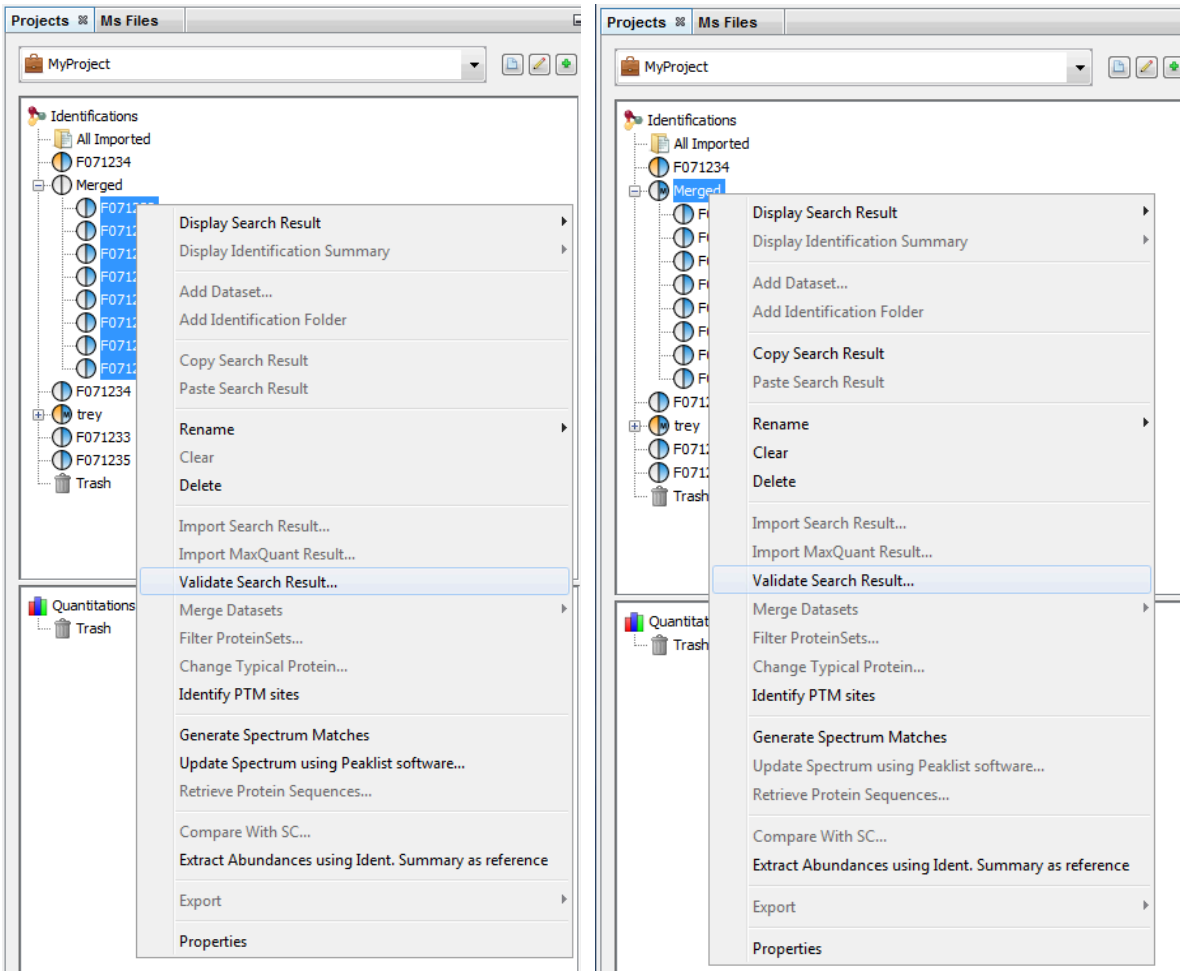
Once executed, the dataset views need to be loaded again to effectively view the spectrum matches.

How to validate a Search Result

See description of Validation Algorithm.

It is possible to validate identification Search Result or merged ones. In the latest case, the filters and validation threshold can be propagated to child Search Results.

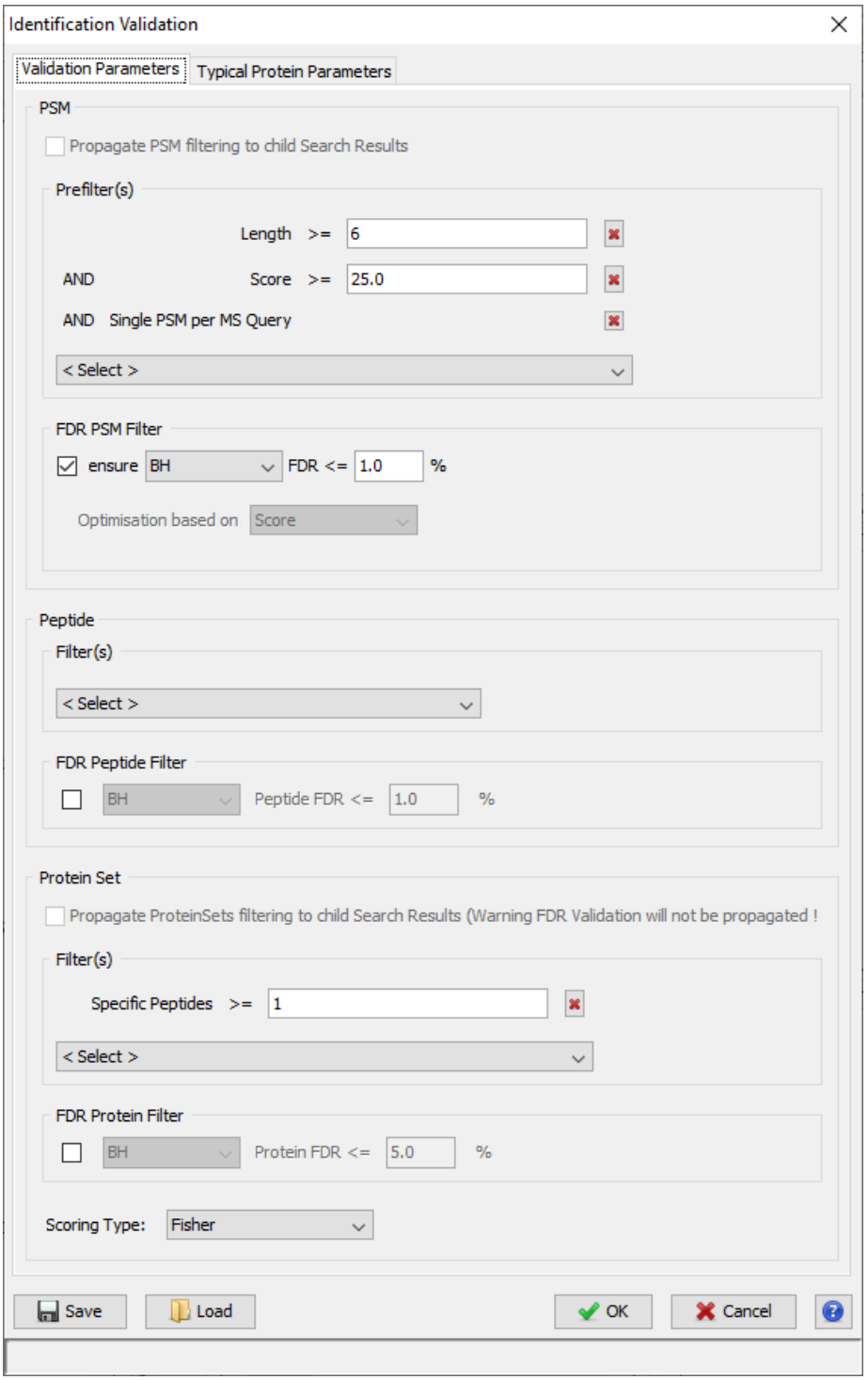
Starting Validation



To validate a Search Result:

- Select one or multiple Search Results to validate
- Right Click to display the popup
- Click on “Validate…” menu

Validation Dialog



In the Validation Dialog, fill the different Parameters (see Validation description):

- you can add multiple PSM Prefilter Parameters (Rank, Length, Score, e-Value, Identity p-Value, Homology p-Value) by selecting them in the combobox.

- you can ensure a FDR on PSMs using BH or Taget/Decoy strategy. If using the latest, FDR will be reached according to the variable selected (Score, e-Value, Identity p-Value, Homology p-Value,…)

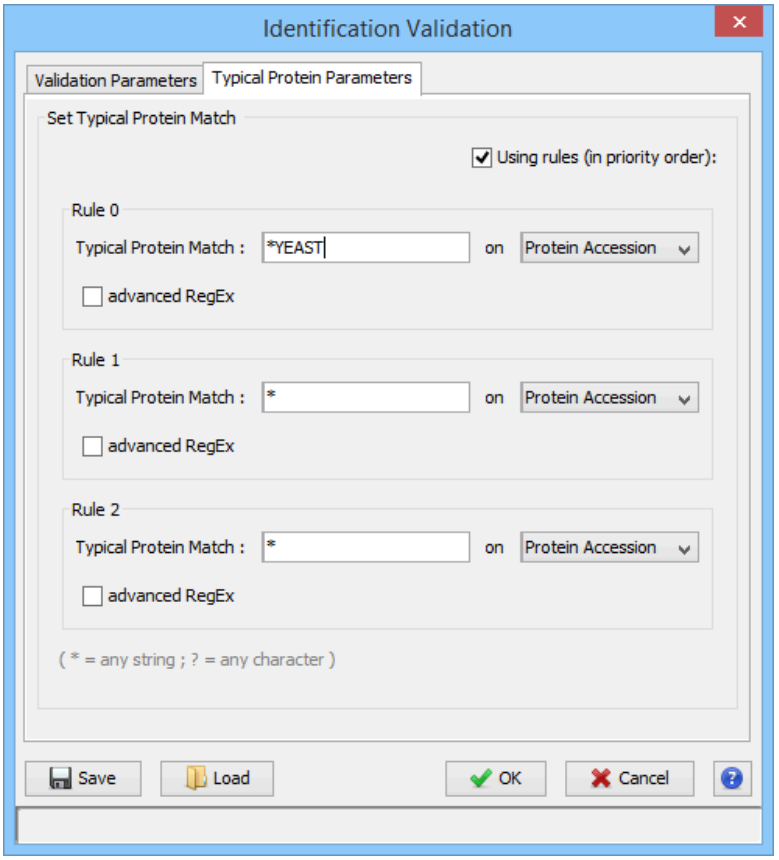
- you can add a Protein Set Prefilter on Specific Peptides count, peptides or peptides sequence count or on Protein Sets score.

- you can also ensure a FDR on protein Sets or peptides using BH or Target/Decoy strategy. The strategy must be consistent between all levels where an FDR control is done (PSM, Peptide and ProteinSet)

Note: FDR can be used only for Search Results with Decoy Data.

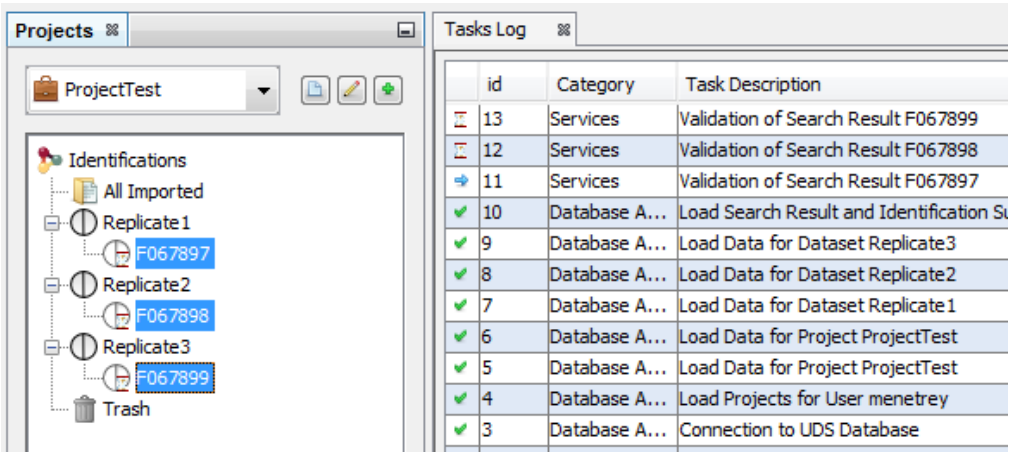
If you run validation on a merged Search Result, you can choose to propagate it to child Search Result. Specified prefilters will be used as defined. For the FDR Filter, it is the threshold found by the validation algorithm which will be used for childs, as a prefilter.

In the second tab, you can define rules for choosing the Typical Protein of a Protein Set by using a match string with wildcards (* or ?) on Protein Accession or Protein Description. (see Change Typical Protein of Protein Sets).



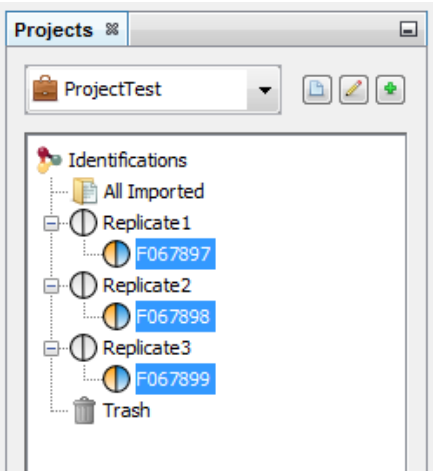
Note: All validation parameters can be saved and loaded using appropriated buttons.

Validation Processing



Validating a Search Result can take some time. While it is not finished, the Search Results are shown greyed with an hourglass over them. The tasks are displayed as running in the “Tasks Log Dialog”.

Validation Done



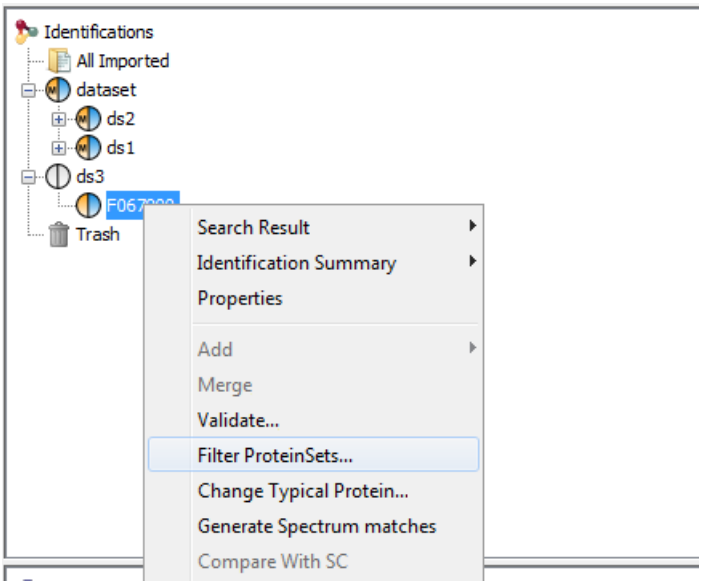
When the validation is finished, the icon becomes orange and blue. Orange part corresponds to the Identification Summary. Blue is for the Search Result part.

How to filter Protein Sets

See description of Protein Sets Filtering.

The protein sets windows are not updated after filtering Protein Set. You should close and reopen the window

Starting filtering



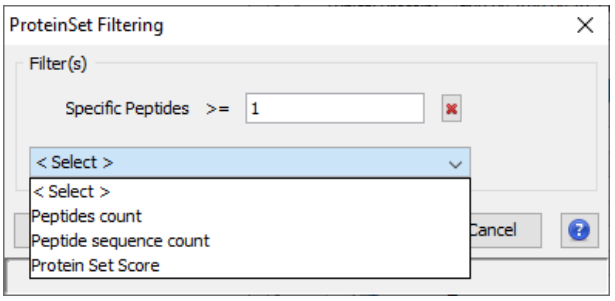
To filter Protein sets of Identification Summaries:

- Select one or multiple Identification Summaries to filter

- Right Click to display the popup

- Click on “Filter ProteinSets…” menu

Filtering Dialog



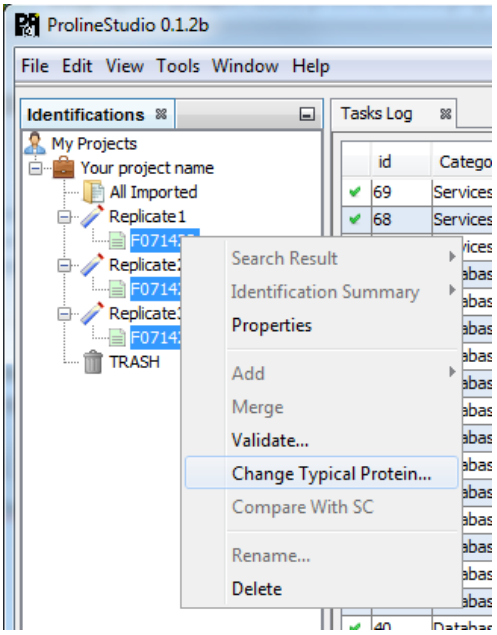
you can add multiple filters (Specific Peptides, Peptide count, Peptide sequence count, Protein Set Score) by selecting them in the combobox.

Once the filtering is done, you will have to open a new protein sets window in order to see modification.

Change Typical Protein of Protein Sets

The protein sets windows are not updated after changing Typical Protein. You should close and reopen the window

Open the Dialog



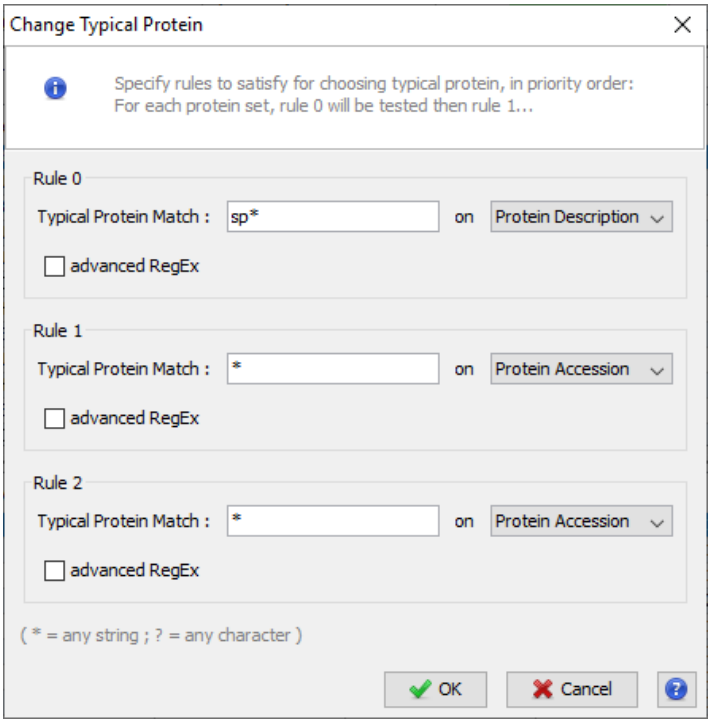
To change the Typical Protein of the Protein Sets of an Identification Summary:

- Select one or multiple Identification Summaries

- Right Click to display the popup

- Click on “Change Typical Protein…” menu

Dialog Parameters



You can set the choice for the Typical Protein of Protein Sets by using a match string with wildcards (* or ?) on Protein Accession or Protein Description.

For Advanced users, a fully regular expression could be specified. In this case, check the corresponding option.

Three rules could 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.

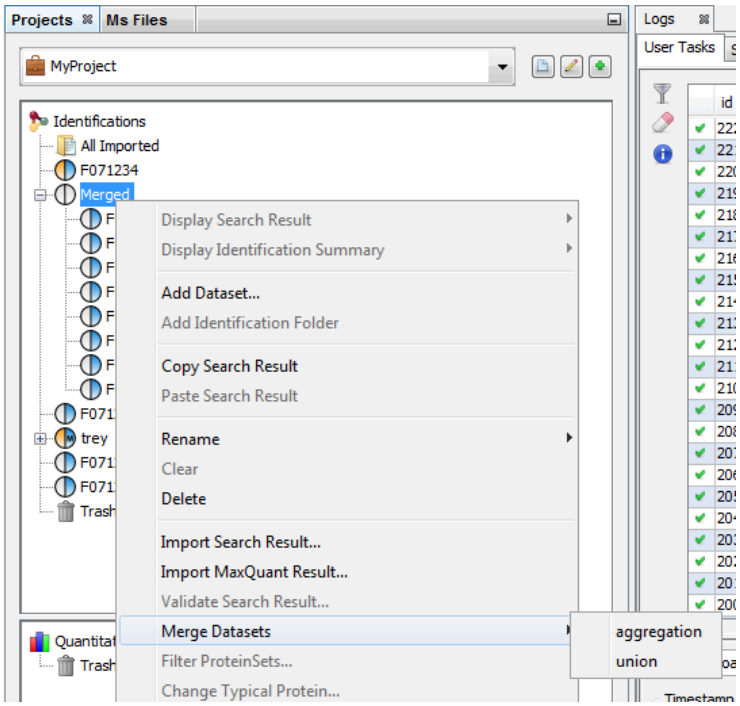
Processing

The modification of Typical Proteins can take some time. During the processing, Identification Summaries are displayed grayed with an hourglass and the tasks are displayed in the Tasks Log Window

Combine datasets

Merge can be done on Search Results or on Identification Summaries. You have also to specify which merge mode is to be used (aggregation or union). See description for combining Search Results or Identification Summaries.

Merge on Search Results

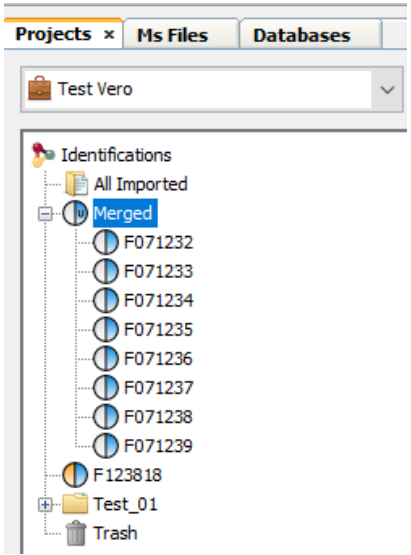


To merge a dataset with multiple Search Results:

- Select the parent dataset

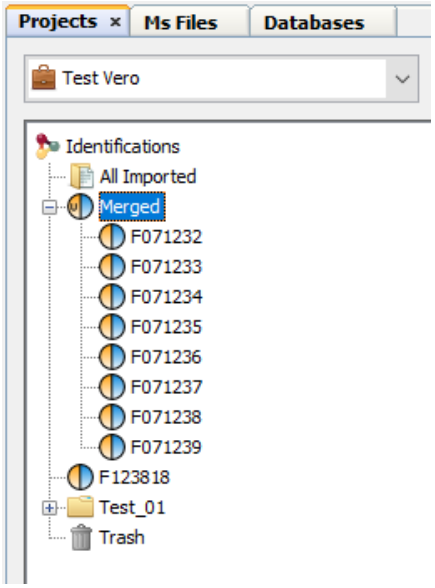
- Right Click to display the popup

- Click on “Merge” menu



When the merge is finished, the dataset is displayed with an U or A in the blue part of the icon, indicating that the merge has been done using Union or Aggregation at a Search Result level.

Merge on Identification Summaries



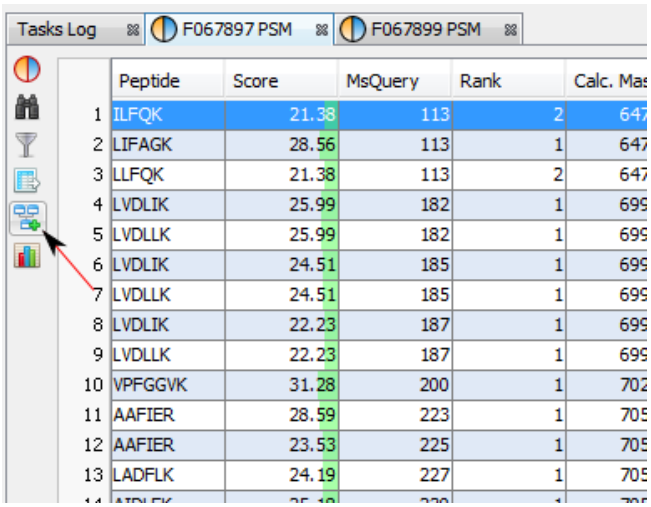
If you merge a dataset containing Identification Summaries. The merge is done on an Identification Summary level. Therefore the dataset is displayed with an U or A in the orange part of the icon.

Data Analyzer

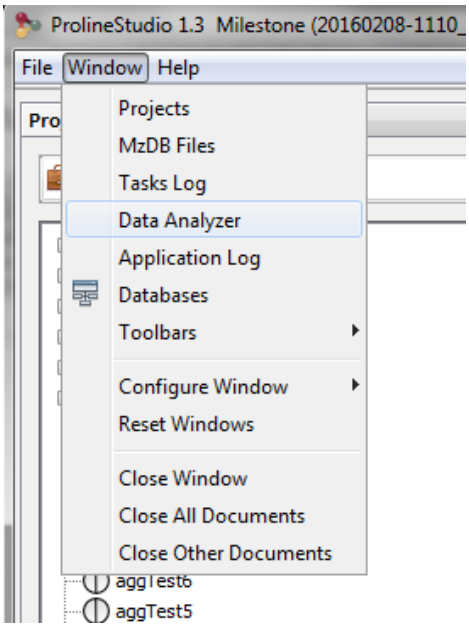
The purpose of the Data Analyzer is to easily do calculations/comparisons on data.

To open the data analyzer, you have two possibilities:

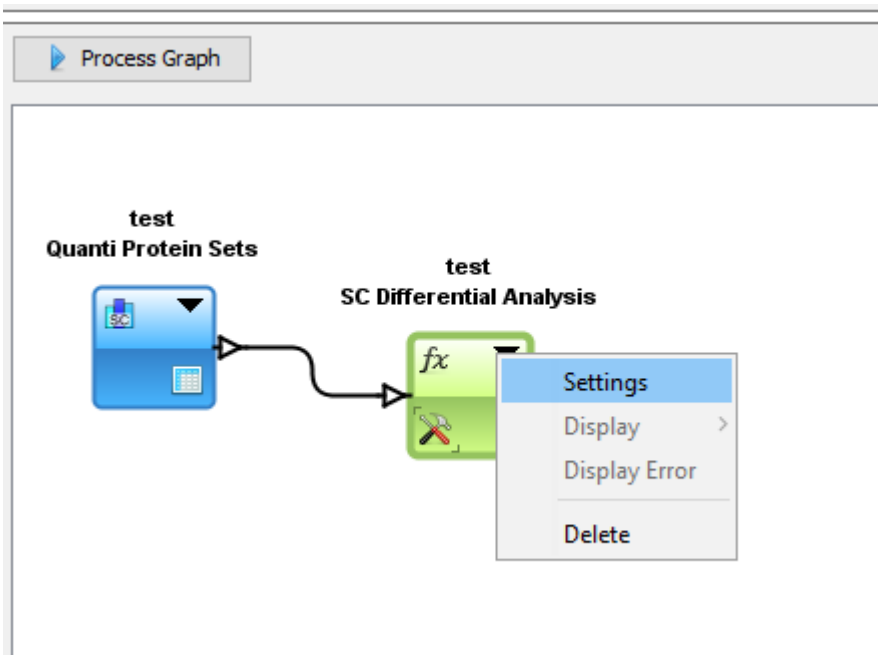
- you can use the dedicated button that you can find in the toolbar of all views. If you use this button, the corresponding data is directly sent to the data analyzer.



- you can use the menu “Window > Data Analyzer”

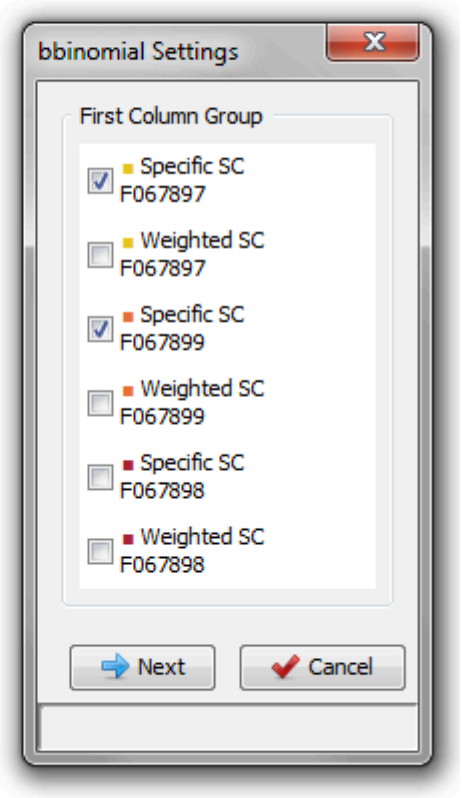


In the Data Analyzer view, you can access all data views, to some functions and graphics. In the following example, we create a graph by adding by Drag & Drop the Spectral Count Data and the corresponding differential analysis function (beta-binomial BBinomial). Then we link them together.

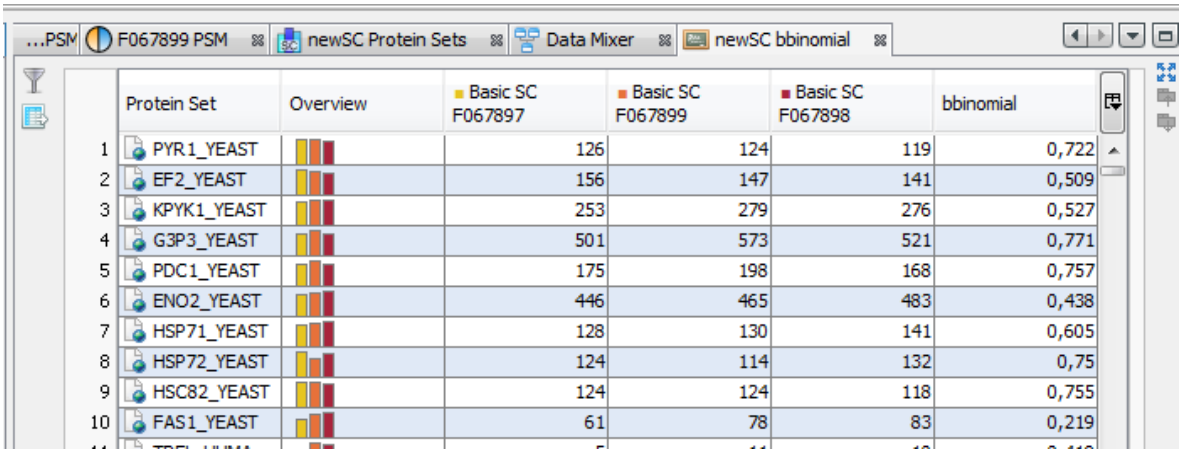


You have to specify the parameters of the Function: right click on the function and select the “settings” menu

In the settings menu, select the two groups of columns on which you want to perform the BBinomial function. When the parameters are set, the calculation is started immediately and an hourglass icon is shown.



When the calculation is finished: the hourglass icon becomes a green tick, and the user can right click and select the “Display” menu to see the result (or click on the “table" icon).



	Protein Set	Overview	Basic SC F067897	Basic SC F067899	Basic SC F067898	bbinomial
1	PYR1_YEAST		126	124	119	0,722
2	EF2_YEAST		156	147	141	0,509
3	KPYK1_YEAST		253	279	276	0,527
4	G3P3_YEAST		501	573	521	0,771
5	PDC1_YEAST		175	198	168	0,757
6	ENO2_YEAST		446	465	483	0,438
7	HSP71_YEAST		128	130	141	0,605
8	HSP72_YEAST		124	114	132	0,75
9	HSC82_YEAST		124	124	118	0,755
10	FAS1_YEAST		61	78	83	0,219

Available Functions

STATISTICS FUNCTIONS

FDR Computation Function

This function is used by ProStar Macro to compute the FDR.

More information:
http://bioconductor.org/packages/release/bioc/vignettes/Prostar/inst/doc/Prostar_UserManual.pdf

PValue Adjustment Function / calibration Plot

Calibration Plot for Proteomics is described here: https://cran.r-project.org/web/packages/cp4p/index.html

SC Differential Analysis Function

beta binomial function, useful for Spectral Count Quantitations

Xic Differential Analysis Function

This function is used by ProStar Macro. Two tests are available: Welch t-test and Limma t-test.

More information:
http://bioconductor.org/packages/release/bioc/vignettes/Prostar/inst/doc/Prostar_UserManual.pdf

Missing values filter Function

This function is used by ProStar Macro to remove rows with too many missing quantitative values.

The available missing values algorithm are:

- Whole Groups: The lines (across all groups) in the quantitative dataset which contain less non-missing value than a user-defined threshold are deleted.
- For every group: The lines for which each condition contains less non-missing value than a user-defined threshold are deleted.
- At least one group: The lines for which at least one condition contains less non-missing value than a user-defined threshold are deleted.

Missing values imputation Function

This function is used by ProStar Macro to impute missing values.

More information:
http://bioconductor.org/packages/release/bioc/vignettes/Prostar/inst/doc/Prostar_UserManual.pdf

Normalization Function

This function is used by ProStar Macro to normalize quantitative values.

More information on algorithms:
http://bioconductor.org/packages/release/bioc/vignettes/Prostar/inst/doc/Prostar_UserManual.pdf

TABLE FUNCTIONS

Join Function

Join data from two tables according to the selected key.

Difference Function

Perform a difference between two joined table data according to a selected key. When a key value is not found in one of the data source tables, the line is displayed as empty. For numerical values a difference is done and for string values, the '<>' symbol is displayed when values are different.

Columns Filter Function

Columns filter, let the user remove unnecessary columns in a matrix. A combobox, with prefix and suffix of the columns allows to select multiple similar columns to filter them rapidly.

Rows Filter Function

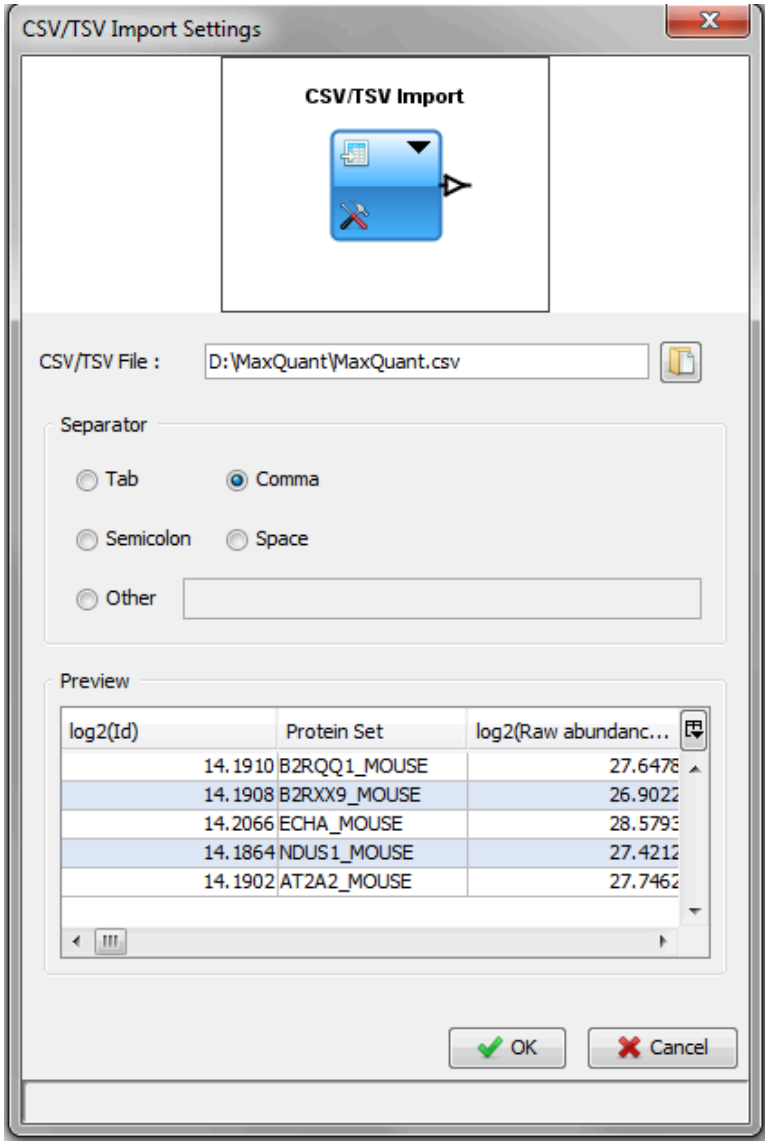
Rows filter function lets the user filter some rows of a matrix according to settings on columns.

Log Function

Create a column by calculating the Log (2 or 10) of an existing column.

CSV/TSV Import

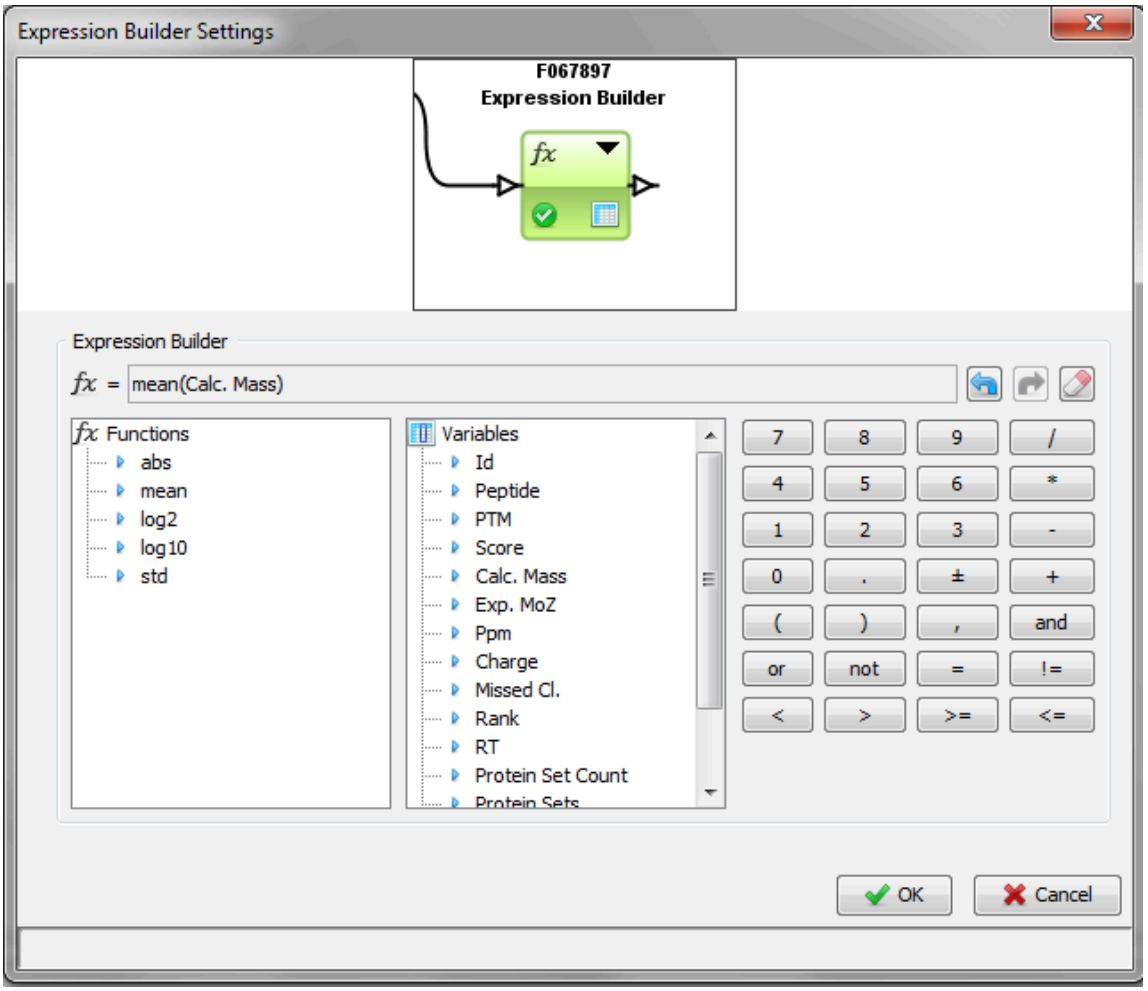
This module lets you import data from a CSV or TSV file. Then you can do calculations and display these data directly in Proline Studio.



The separator is automatically selected according to the csv file. But you can modify it.

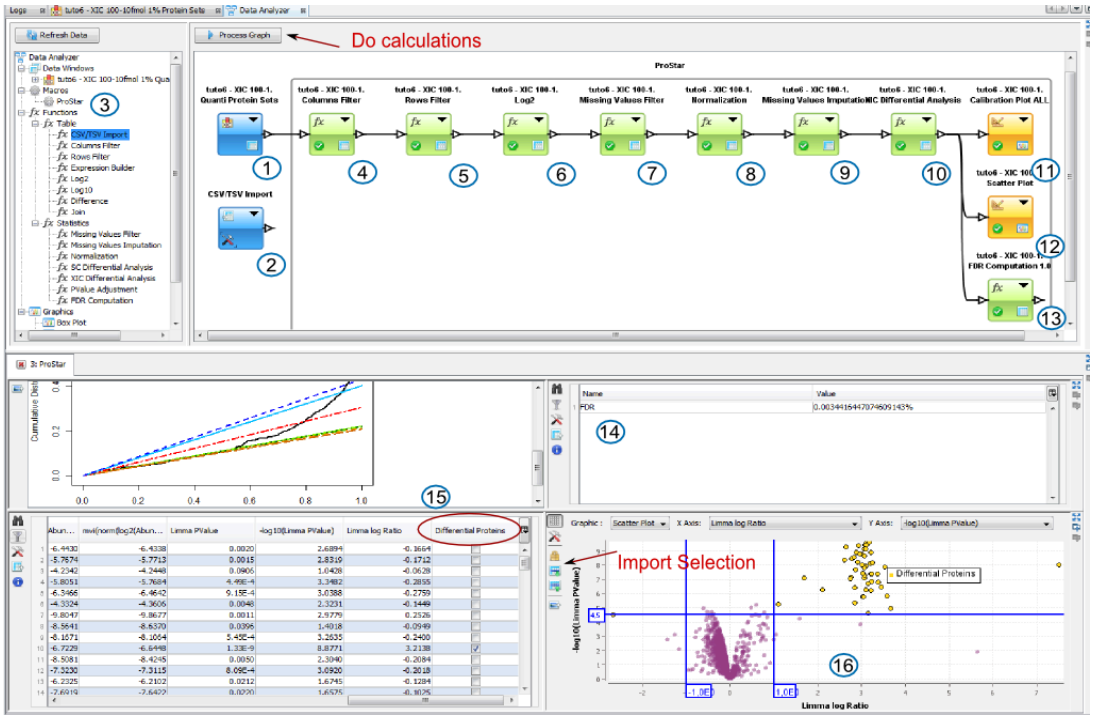
The preview zone displays the first lines of the file as it will be loaded.

Expression Builder



The expression builder lets you create an expression with built-in functions or comparators and variables (columns from the linked matrix). In the example, we calculate the mean of a column in the matrix.

Prostar Macro



1 or 2: Add XIC Data to Data Analyzer from the Protein Set View or by importing data from a csv file.

3: Add Prostar Macro by a drag and drop and link XIC Data to the Macro. And do the calculation by clicking on the button Process Graph.

During the process, the Data Analyzer will ask you settings for each function.

4: Filter unnecessary columns from your data if. Settings can be validated with no parameters if you don't need it.

5: Filter is needed only if you want to remove contaminants. Settings can be validated with no parameters if you don't need it.

6: Log is needed to log abundances (Data from Proline). For Data coming from MaxQuant, data is already logged.

7 to 13: follow the settings asked (you can find some help in Prostar documentation, or information in corresponding functions.)

During the process, results will be automatically displayed:

14: FDR Result

15: Calibration Plots

16: Result Table with differential Proteins Table and the corresponding scatter plot. You can select differential proteins in the table, to import them in the scatter plot and create a colored group with them.

If you want to look at other results, right click on a function and select “Display in New Window”

Prostar User Manual:

http://bioconductor.org/packages/release/bioc/vignettes/Prostar/inst/doc/Prostar_UserManual.pdf

Prostar Tutorial :

http://bioconductor.org/packages/release/bioc/vignettes/Prostar/inst/doc/Prostar_Tutorial.pdf

Calculator

Calculator lets you write python scripts to manipulate freely viewed data.

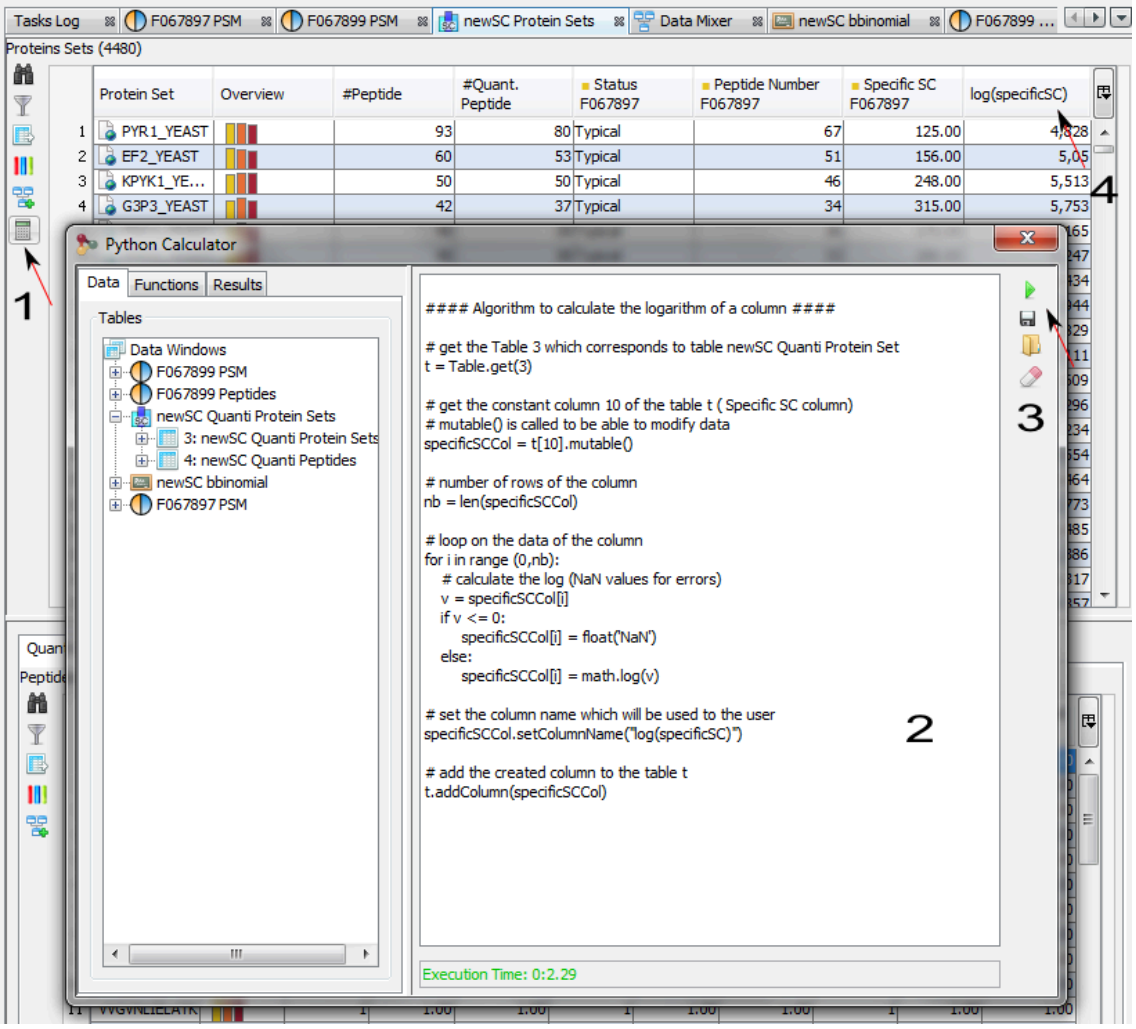
1) To open the calculator, click on the calculator icon (not available on all views for the moment)

On the left part of the calculator, you can access all viewed data, double click to add a table or a column to the script.

2) Write your python script on the text area

3) Execute it by clicking on the green Arrow.

4) When the script has been executed, the results of the calculations (variables, new columns) are available in the "Results" tab. Double click on a new column to add it to the table. Or like in the example, directly add the column to the table programmatically.



Examples

Script to calculate a log column

```
#### Algorithm to calculate the logarithm of a column ####

# get the Table 3 which corresponds to table newSC Quanti Protein Set
t = Table.get(3)

# get the constant column 10 of the table t ( Specific SC column)
# mutable() is called to be able to modify data
```

```
specificSCCol = t[10].mutable()

# number of rows of the column
 nb = len(specificSCCol)

# loop on the data of the column
for i in range (0,nb):
 # calculate the log (NaN values for errors)
 v = specificSCCol[i]
 if v <= 0:
  specificSCCol[i] = float('NaN')
 else:
  specificSCCol[i] = math.log(v)

# set the column name which will be used to the user
specificSCCol.setColumnName("log(specificSC)")

# add the created column to the table t
t.addColumn(specificSCCol)
```

Script to perform a difference and a mean between two columns

```
#### Algorithm to perform a difference and a mean between two columns ####


t = Table.get(9)
colAbundance1 = t[3]
colAbundance2 = t[5]

# difference between two columns
colDiff = colAbundance1-colAbundance2

# set the name of the column
colDiff.setColumnName("diff")

# mean between two columns
colMean = (colAbundance1+colAbundance2)/2

# set the name of the column
colMean.setColumnName("mean")

# add columns to the table
t.addColumn(colDiff)
t.addColumn(colMean)
```

Script to perform a perform a pvalue and a ttd on a XIC quantitation table

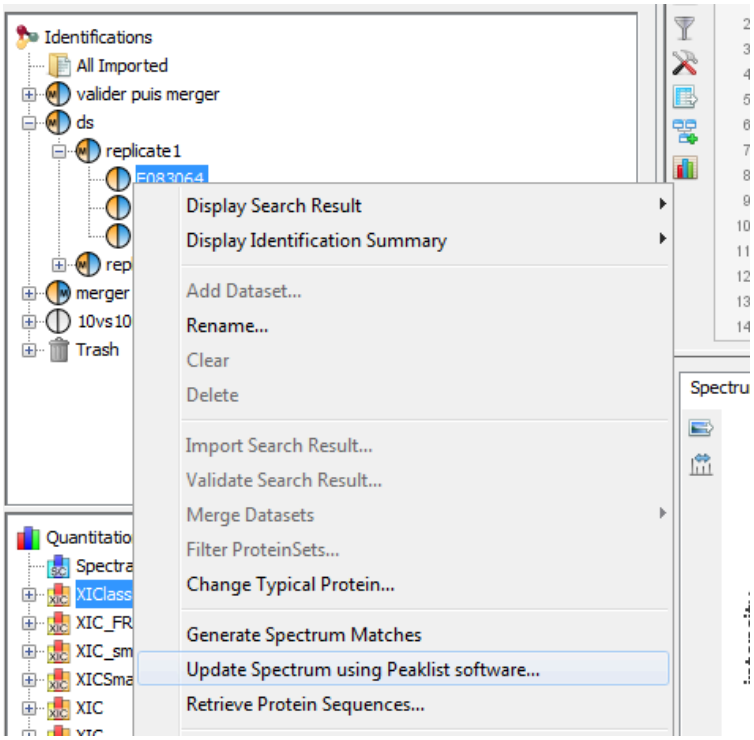
```
#### Algorithm to perform a pvalue and a ttd on abundances column of a XIC quantitation
####
t = Table.get(1)

pvalueCol = Stats.pvalue( (t[2], t[3]), (t[4],t[5]) )
ttdCol = Stats.ttd( (t[2], t[3]), (t[4],t[5]) )

pvalueCol.setColumnName("pvalue")
ttdCol.setColumnName("ttd")

t.addColumn(pvalueCol)
t.addColumn(ttdCol)
```

Update Spectrum using Peaklist software

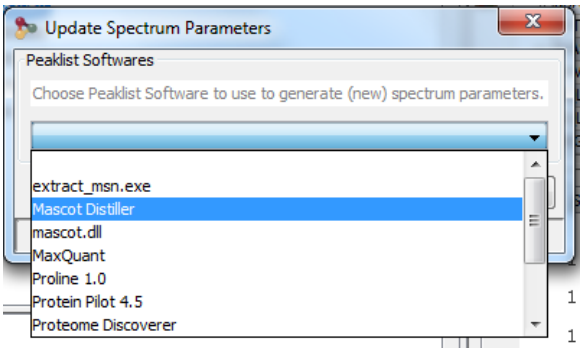
When importing a search result, the software used for the peaklist creation has to be specified. This parameter is mandatory for the XIC quantitation as it is used to find scan number or RT in the spectrum title. Indeed, this information is then used to extract abundances in the *raw files*.

If an invalid software has been specified when importing, it is possible to change the peaklist software afterwards. This option is only valid for Identification DataSets.

Right click on the identification DataSet, and select “Update Spectrum using Peaklist software”



The following dialog will be displayed allowing the user to select the peaklist software to use.

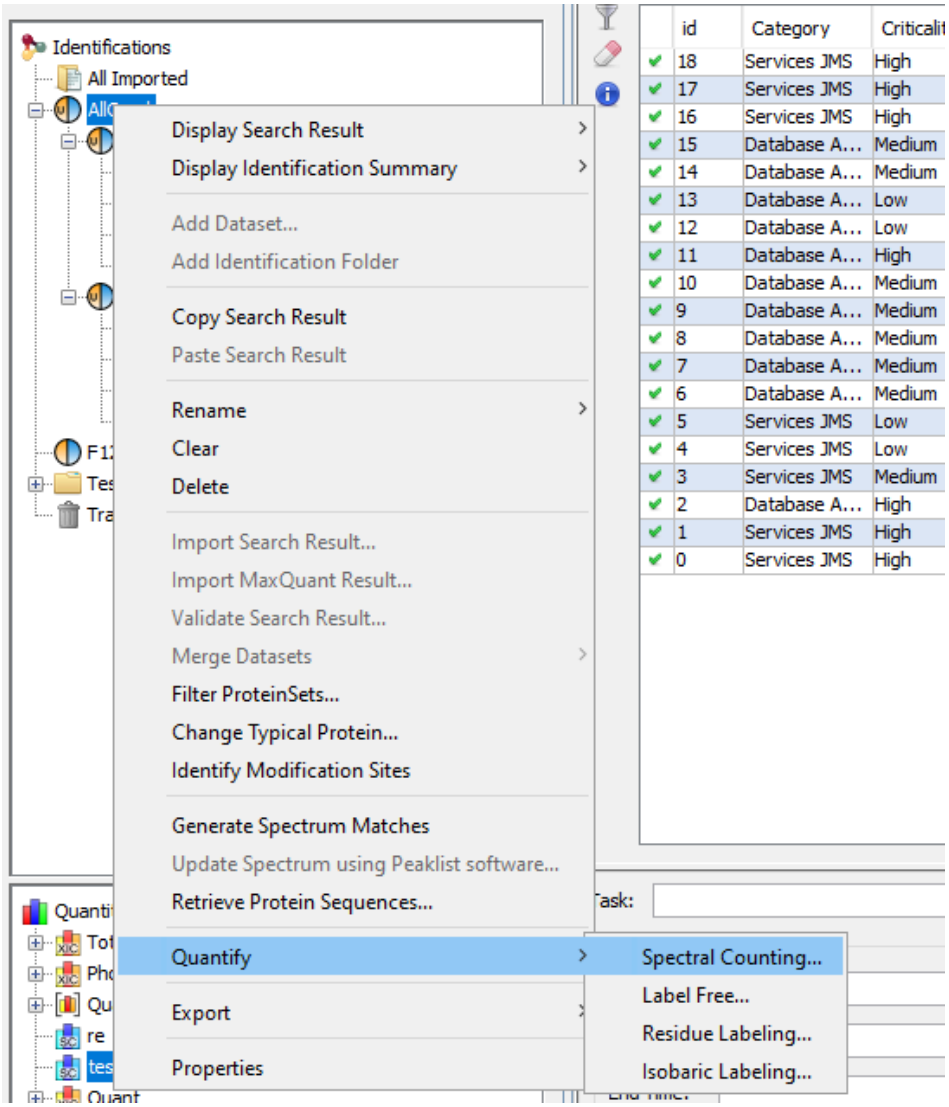


Spectral Count

See description of Compare Identification Summaries with Spectral Count.

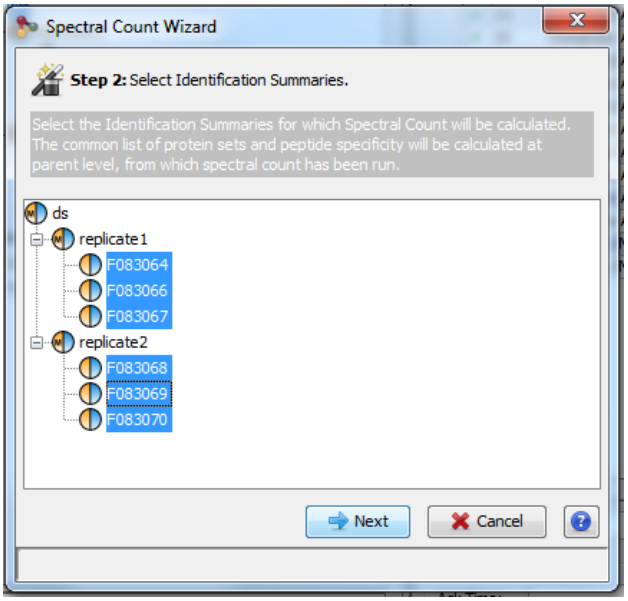
Generate a Spectral Count

To obtain a spectral count, right click on a Dataset with merged Identification Summaries and select the “Quantify → Spectral Counting” menu in the popup. This Dataset is used as the reference Dataset and Protein Set list as well as specifics peptides are defined there.

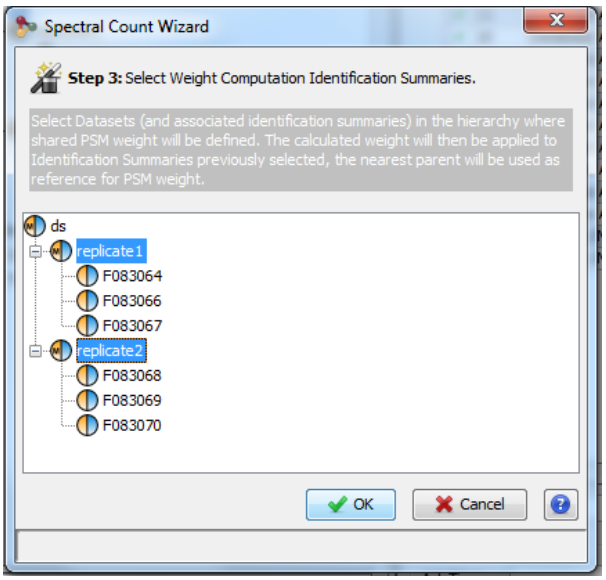


In the Spectral Count window, fill the name and description of your Spectral Count and press Next.

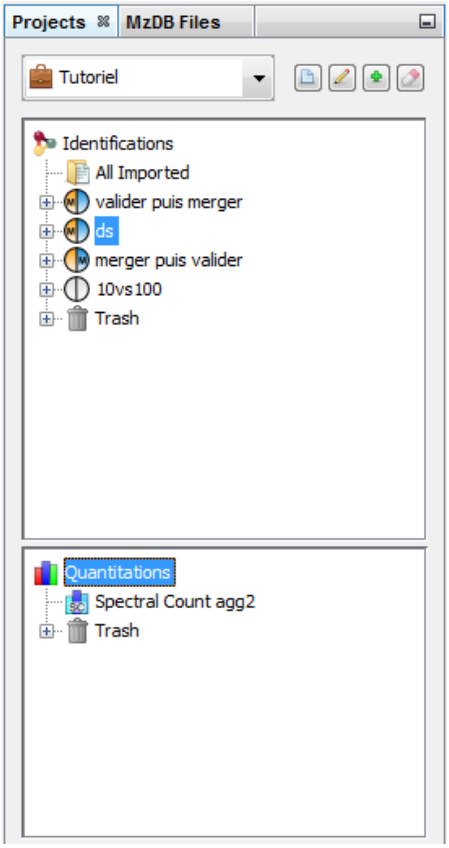
Then select the Identification Summaries on which you want to perform the Spectral Count and press Next.



Finally select the DataSet where shared peptides spectral count weight should be calculated and press OK.



A Spectral Count is created and added to the Quantitations Panel.



Display a Spectral Count

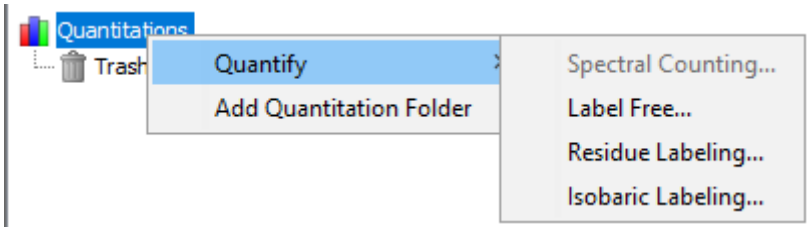
You can then display a Spectral Count, see Display a Spectral Count

Label Free Quantitation

For description on LC-MS Quantitation you can first read the principles in this page: Label-Free quantitation principles

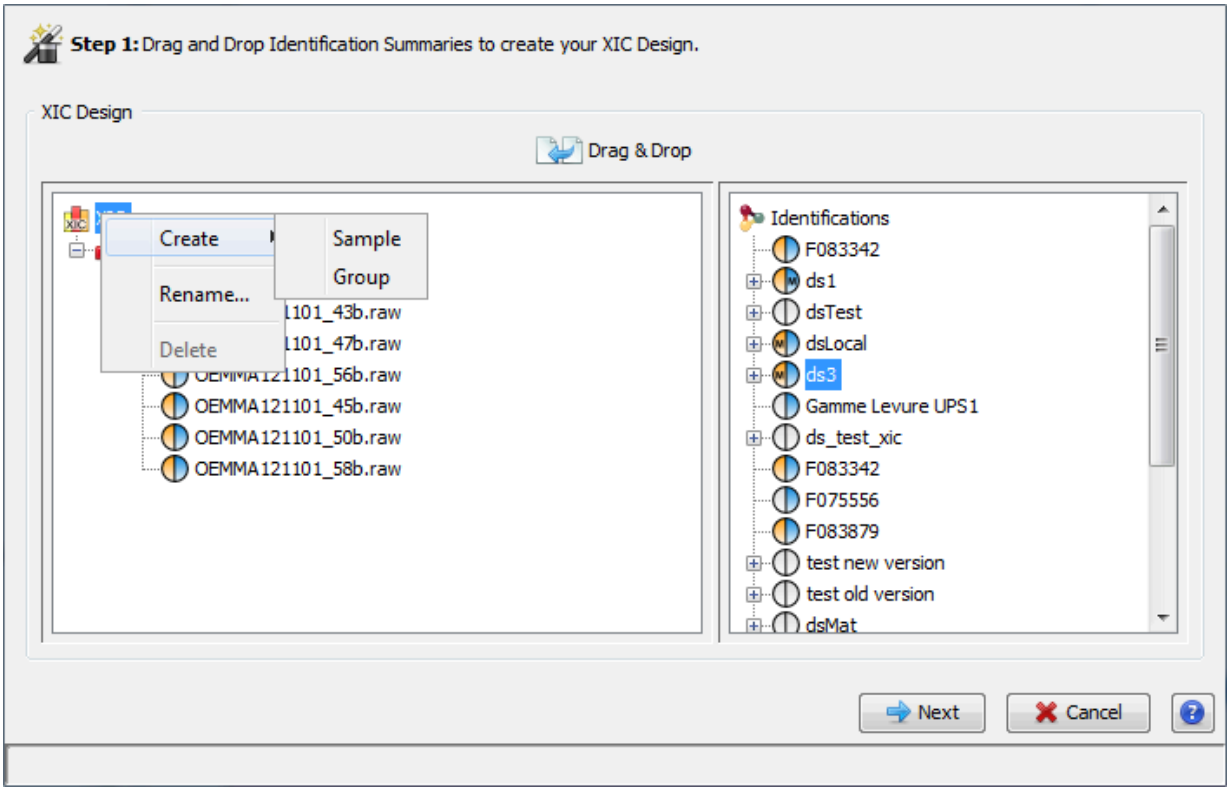
Create a New Label Free Design

You can create a new label-free design by clicking on the "Quantify → Label Free" menu from Quantitation or from a dataset in the Identification tree. In this case, the selected dataset will be used as reference for the quantitation (from where the list of peptides and protein is extracted).



Settle Design

When you start the creation of a new XIC Design, the following dialog is opened:



In the left Area, you can create your design :

- by adding Group and Samples (thanks to mouse right click popup)
- by renaming every items (thanks to F2 or mouse right click popup)
- by rearranging order of groups and samples (thanks to drag & drop)

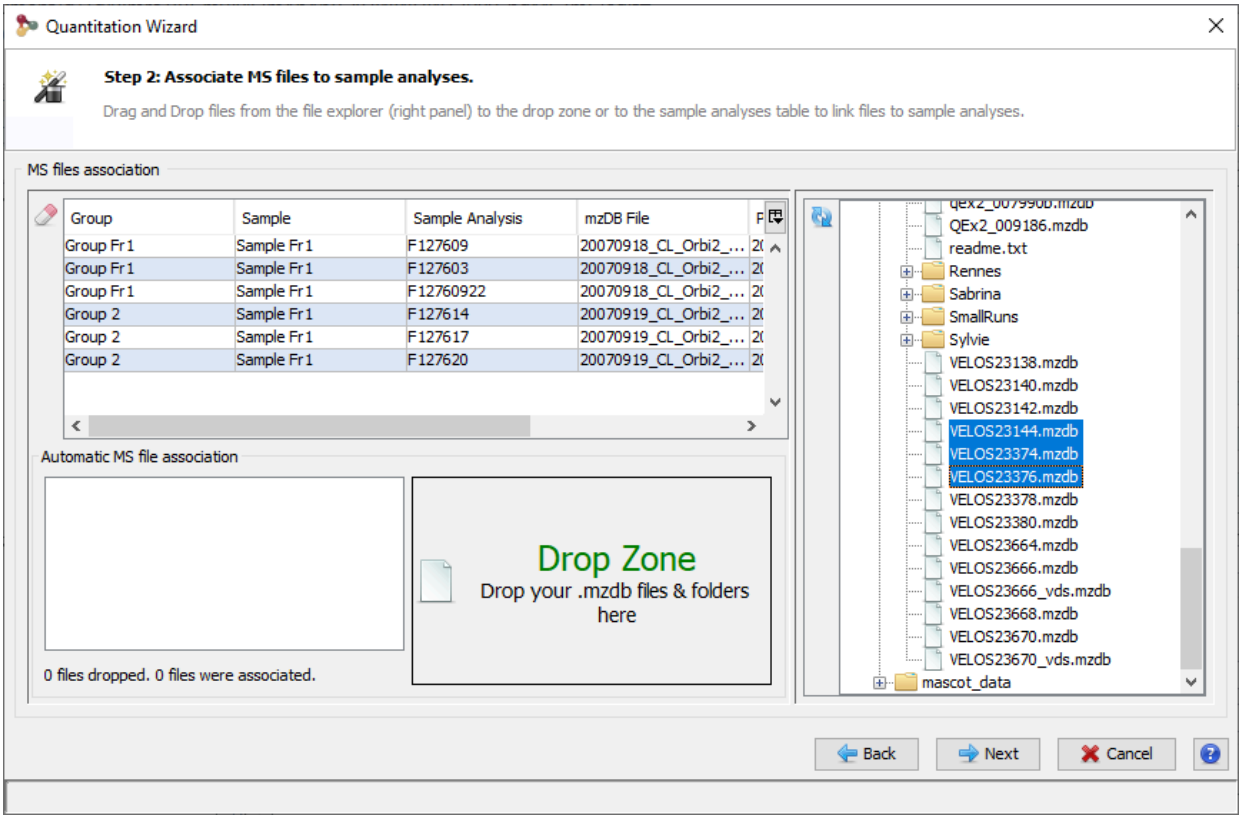
From the right Area, you can drag and drop desired identifications to the left Area. If you drop in the XIC or in a Group node, Group and Sample nodes will be automatically added. But you can drop directly in a Sample node.

Note : it is recommended that the XIC node is renamed.

When your design is ready, click on Next Button.

Link to Raw Files

The following panel is displayed:



In order to be able to perform any XiC design, all participating Sample Analysis must be associated with a corresponding raw file. The association is done at the second step of the XIC design by :

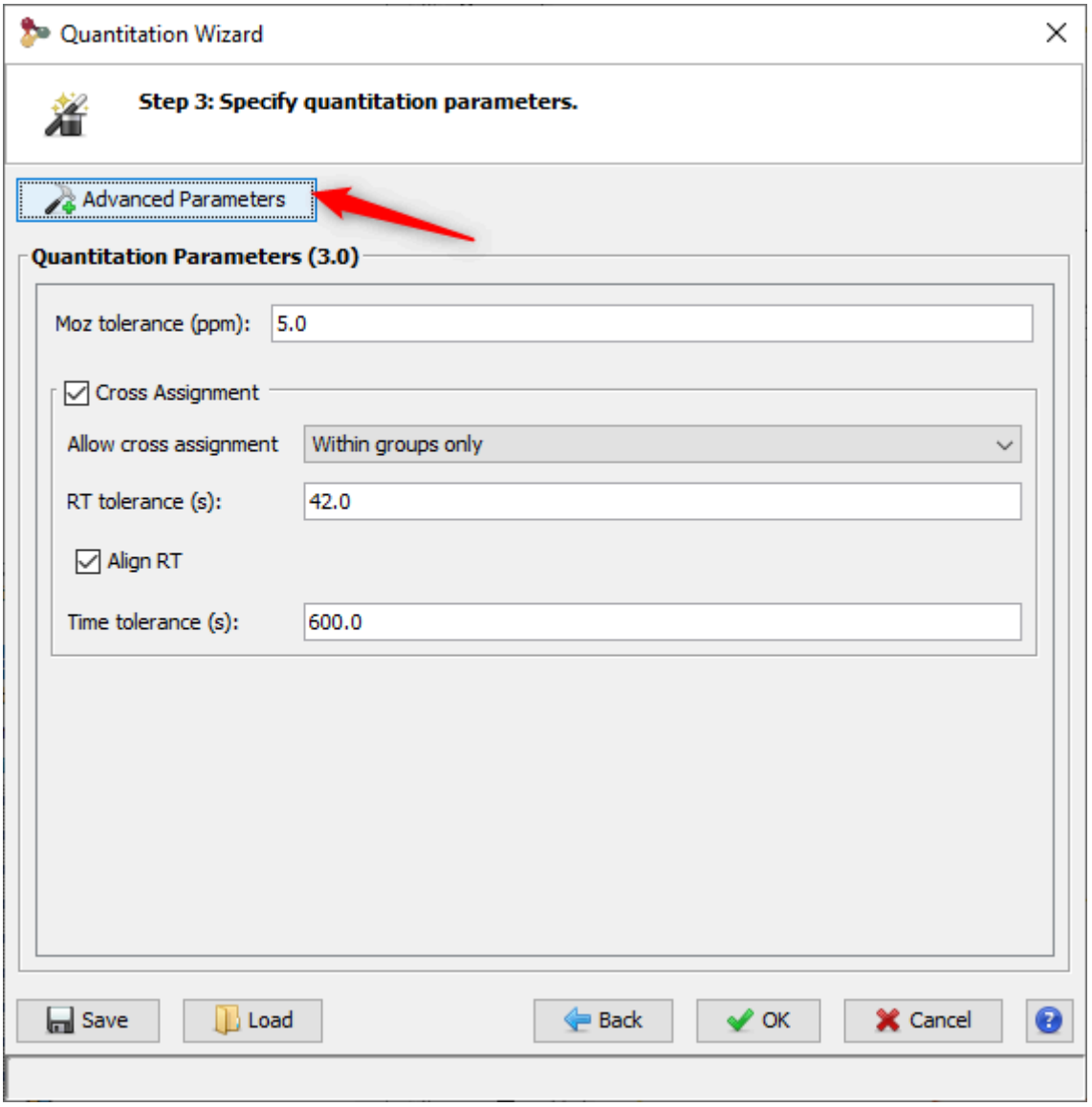
- drag & drop the “mzDB File” into the raw corresponding to the correct sample analysis. It must be noted that although it is possible to overwrite an existing association (done in a previous quantitation using the same sample analysis dataset), there is no verification done to ensure compatibility between the mzDb file and the sample analysis dataset.
- using the Drop Zone. This feature can be quite helpful in cases where multiple associations are missing or when the plethora of uploaded .mzdb files intimidates the user from manually searching the files one by one. The feature itself is extremely easy to use as it just requires dragging a set of files or folders containing .mzdb files from the user part. As soon as a drop takes place, all missing connections will be automatically created as long as a matching .mzdb file has been dropped in the Drop zone. It must also be noted that since version 1.5, users have now at their disposal indices about the association source. Furthermore, in order to protect from a possible corruption of an existing association in the database, the latter ones cannot be overwritten.

When all participating Sample Analysis are associated with a raw file, click on Next Button.

Quantitation Parameters

Now you must select the parameters. See Label-free LC-MS quantitation configuration to have more details about the different parameters.

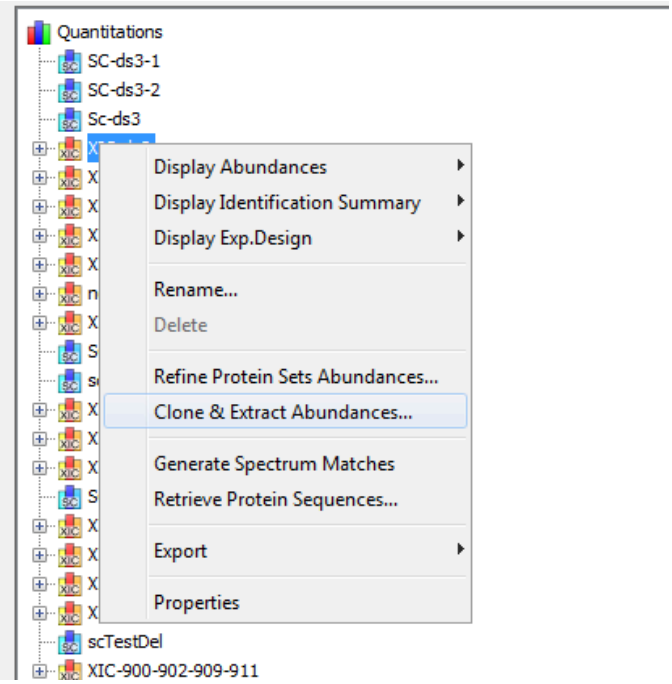
These parameters are not all displayed. You can display a complete set of parameters by clicking on the “Advances Parameters”.



Note: all the parameters are already set with default values, or with previous values if a quantitation has already been run.

Clone a Label Free Design

You can copy a XIC design by selecting the “Clone & Extract Abundances” option from an existing quantitation result. The new XIC is generated using an existing Experimental Design and parameters setting.



Displays

Newly generated quantitation designs are immediately added to the Quantitation Tree. Through the latter one, and via a popup menu, the end user has the capacity either to view a design's properties as seen to the following screenshot, or to apply a series of actions on it, including among others:

- Display experimental Design parameters, results and abundances : see how to Display a quantitation result
- Delete a quantitation Design, see how to Delete Data
- Rename a quantitation Design, by clicking on "Rename…" in the popup menu.
- Export the quantitation results, see how to Export Data

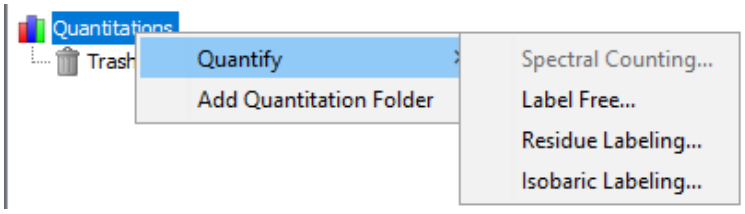
	Group	Type	XIC
1	General Information		
2		XIC id	3615
3	Identification Summary		
4		Identification Summary id	141
5		Description	
6		Date	20 mai 2015
7	Quantitation Processing Config		
8		aln_method_name	EXHAUSTIVE
9		aln_params / ft_mapping_params / moz_tol	5.0
10		aln_params / ft_mapping_params / moz_tol_...	PPM
11		aln_params / ft_mapping_params / time_tol	600.0
12		aln_params / mass_interval	20000
13		aln_params / max_iterations	4
14		aln_params / smoothing_method_name	TIME_WINDOW
15		aln_params / smoothing_params / min_windo...	50
16		aln_params / smoothing_params / window_o...	20
17		aln_params / smoothing_params / window_size	200
18		clustering_params / intensity_computation	MOST_INTENSE
19		clustering_params / moz_tol	10.0
20		clustering_params / moz_tol_unit	PPM
21		clustering_params / time_computation	MOST_INTENSE
22		clustering_params / time_tol	15.0
23		detect_features	false
24		detect_peakels	true
25		extraction_params / moz_tol	5.0
26		extraction_params / moz_tol_unit	PPM
27		ft_filter / name	INTENSITY
28		ft_filter / operator	GT
29		ft_filter / value	0.0
30		ft_mapping_params / moz_tol	10.0
31		ft_mapping_params / moz_tol_unit	PPM
32		ft_mapping_params / time_tol	120.0
33		normalization_method	MEDIAN_INTENSITY
34		start_from_validated_peptides	true
35		use_last_peakel_detection	false
36	Quantitation Method		
37		Name	label free based on the extraction of feature...
38		Type	label_free
39		Abundance Unit	feature_intensity
40	Master Quantitation Channel 1		
41		Master Quantitation id	734
42		Name	XIC
43		Serialized Properties	
44	Quantitation Channel 1		
45		Quantitation Channel id	2882
46		Result File Name	OEMMA121101_43b.raw
47		Raw File Path	D:\Data\Claire\OEMMA121101_43b.raw
48		Mzdb Raw File Name	OEMMA121101_43b.mzdb
49		Identification Summary Id	117
50	Biological Sample 1		
51		Biological Sample id	2027
52		Biological Sample Name	Group ds3-1Sample ds3-1
53	Quantitation Channel 2		
54		Quantitation Channel id	2883
55		Result File Name	OEMMA121101_45b.raw
56		Raw File Path	\\DSV_D01\Shares\301-Projets_BGE\301.1-...
57		Mzdb Raw File Name	OEMMA121101_45b.mzdb
58		Identification Summary Id	123
59	Biological Sample 2		
60		Biological Sample id	2028
61		Biological Sample Name	Group ds3-2Sample ds3-2

Isobaric labeling Quantitation

For description on isobaric labeling quantitation you can first read the principles in this page: Isobaric labeling quantitation workflow

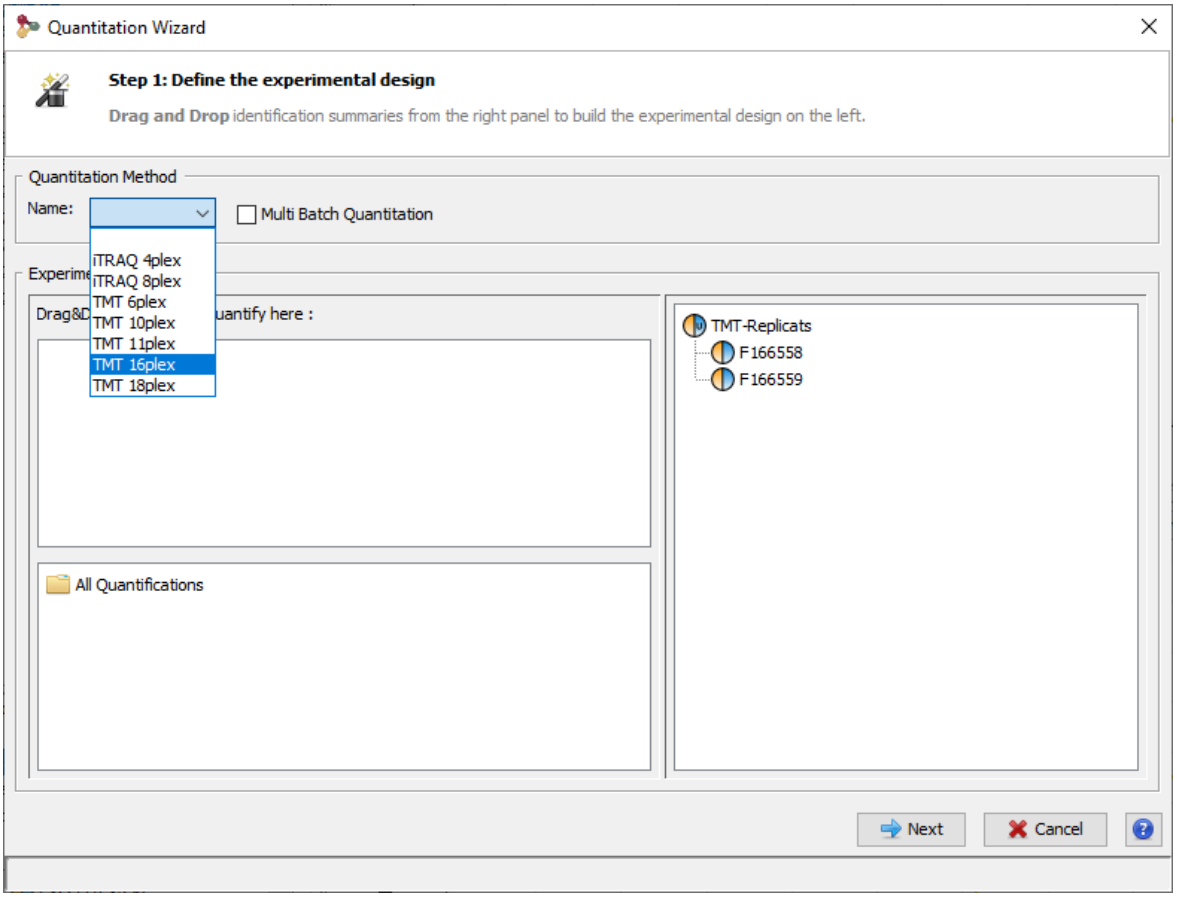
Create a New Design

You can create a new label-free design by clicking on the "Quantify → Label Free" menu from Quantitation or from a dataset in the Identification tree. In this case which is recommended, the selected dataset will be used as **reference** for the quantitation (from where the list of peptides and protein is extracted).



Settle Design

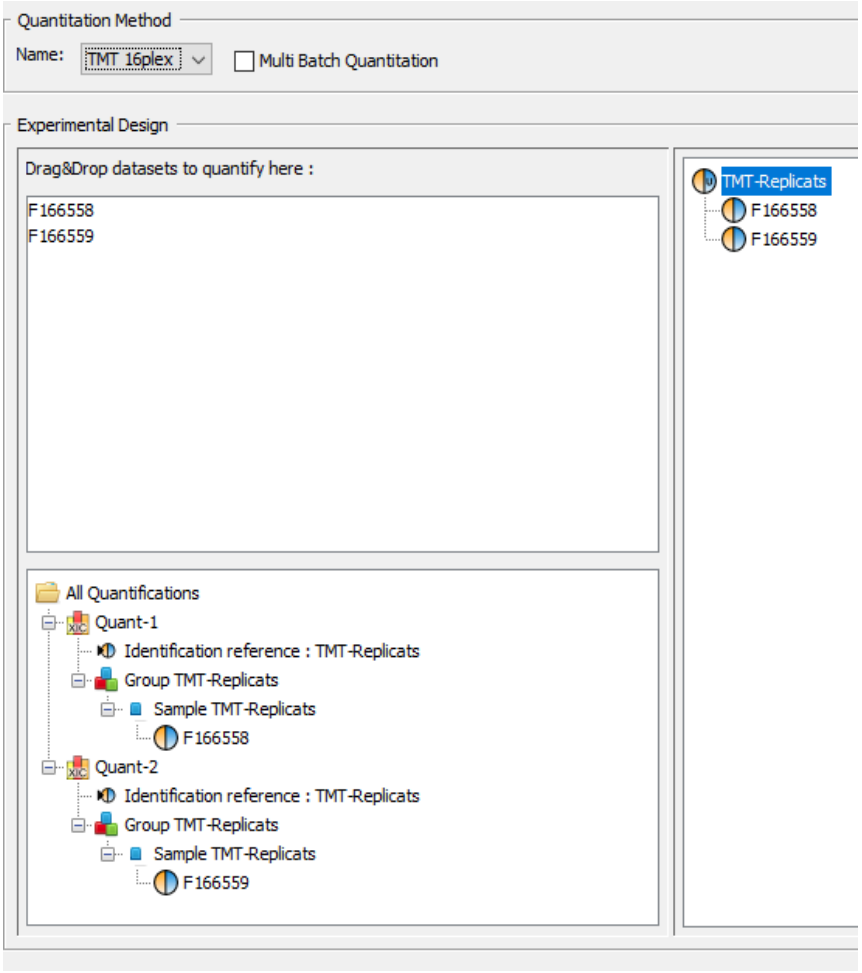
When you start the creation of a new Design, the following dialog is opened:



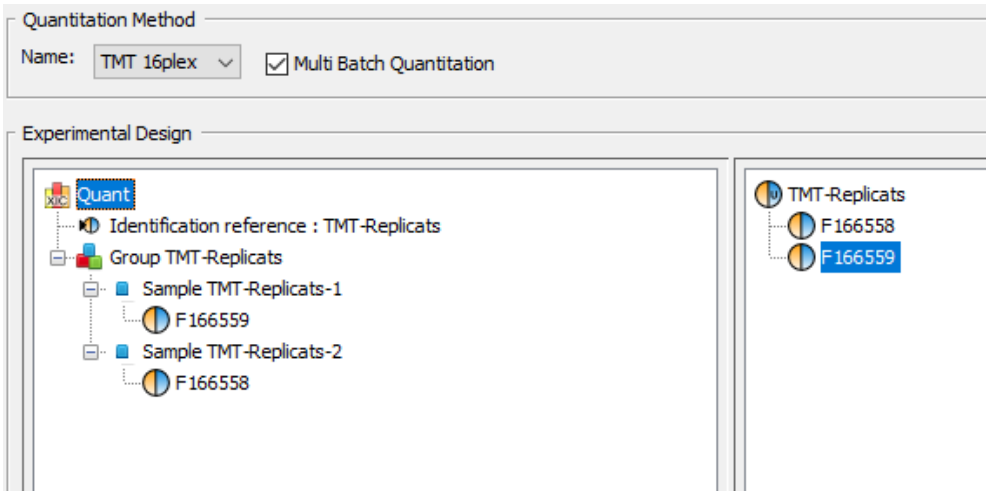
First select the method name in the drop-down list.

In case of "simple" isobaric quantitation, each run is quantified separately, the different conditions are indeed pulled in the same sample with a specific tag. However, Proline allows you to configure the quantisation of several runs at

once, in the case of fractions for example. Each run will produce a separate quantitation result as if you had run them individually.

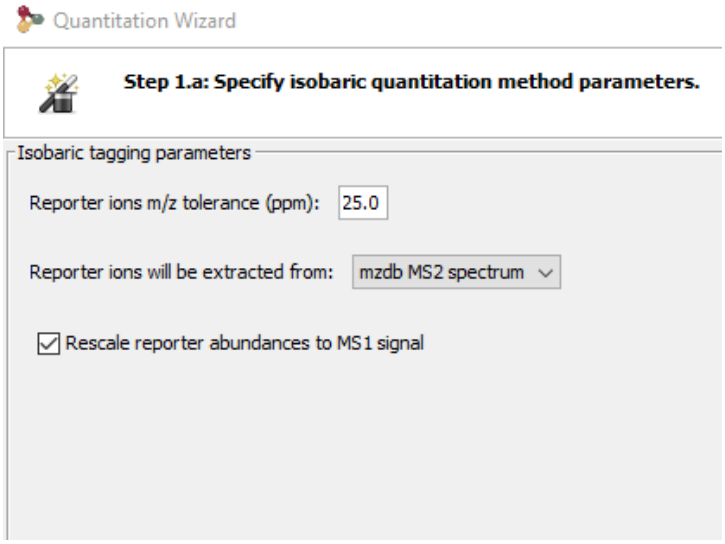


But it is also possible to execute a multi-batch quantitation. In this case, a single quantitation will be performed and the result of all the runs will be concatenated as different channels.
Warning: If multi-batch quantitation is selected, at the next step it is mandatory to use “Rescale reporter abundances…”



Isobaric labeling quantitation parameters

When you click next, a view with isobaric labeling specific parameters is displayed.



- **Reporter ions m/z tolerance** (ppm): m/z window used to match a peak in MSMS spectra with an isobaric tag mz
- **Reporter ions will be extracted from**: specify where intensity values are read
    - **Proline spectrum** : from spectrum stored in Proline database and read in result file when importing search result.
    - **mzDB MS2 spectrum** : from MS2 spectrum read in mzDB associated with the run.
    - **mzDB MS3 spectrum** (beta version): from MS3 spectrum read in mzDB associated with the run.
- **Rescale reporter abundance to MS1 signal** (mandatory for multi-batch) : all reporter ions abundance (read in MS2 or MS3 spectra) are rescaled on the MS1 APEX. If this option is selected, you will be prompted to give “label-free parameters” in order to compute MS1 quantitation, and thus MS1 APEX.

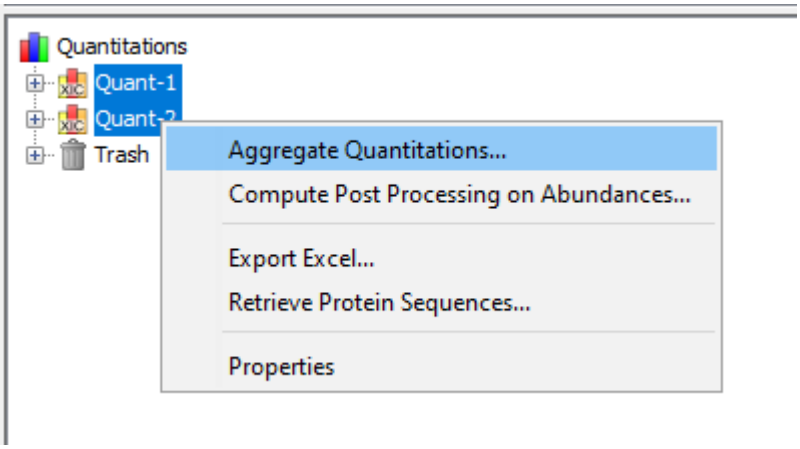
The next steps will be:

- Associate mzDB files to sample analysis: see Link raw files
- Label-free parameter, if rescale to MS1 was selected : see Quantitation parameters.

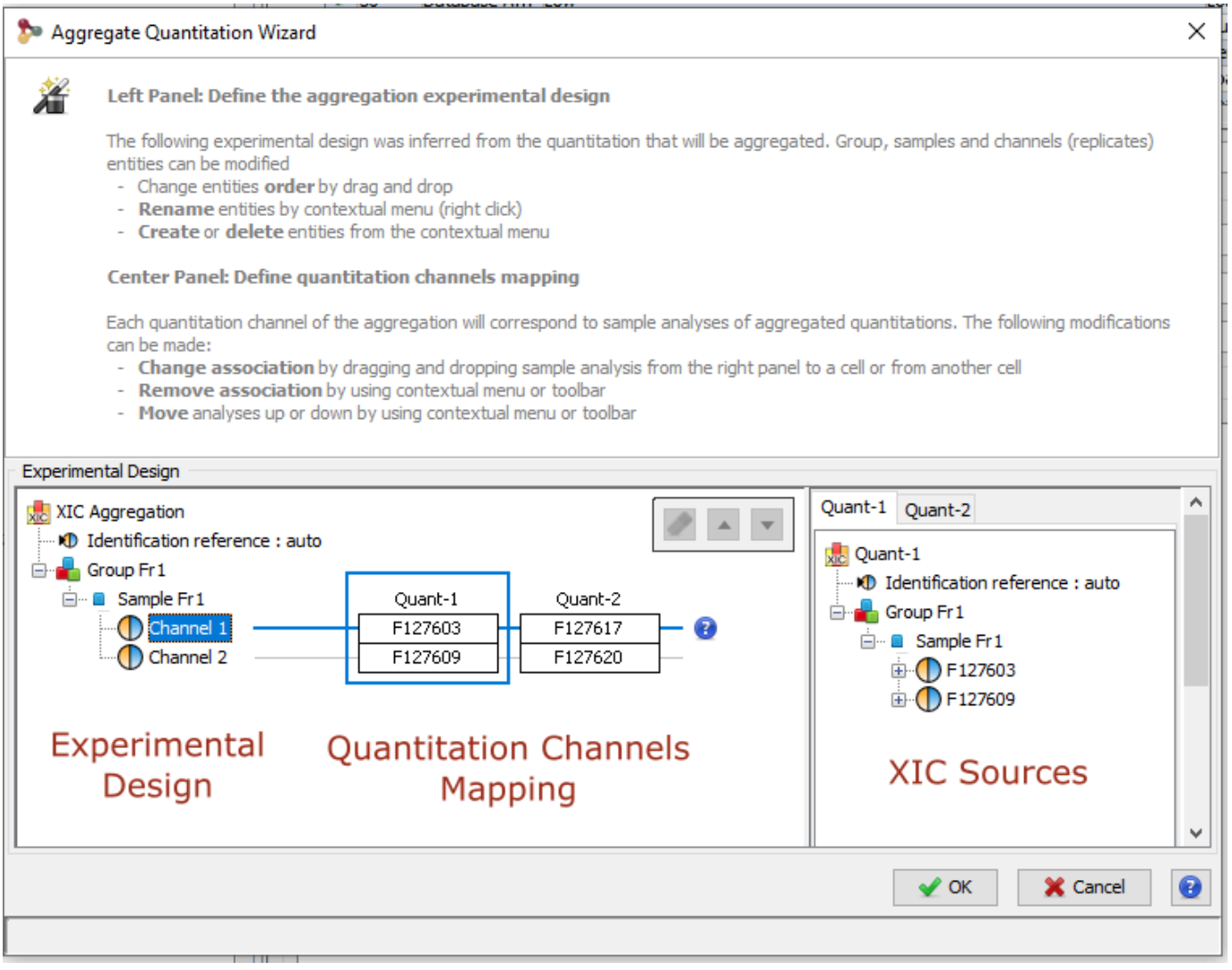
Quantitations Aggregation

It is possible to aggregate data from different label-free or isobaric labeling designs.

First you must select two or more XIC Designs in the Quantitation tree and do a mouse right click to obtain a popup and then select the menu “Aggregate Quantitations”.



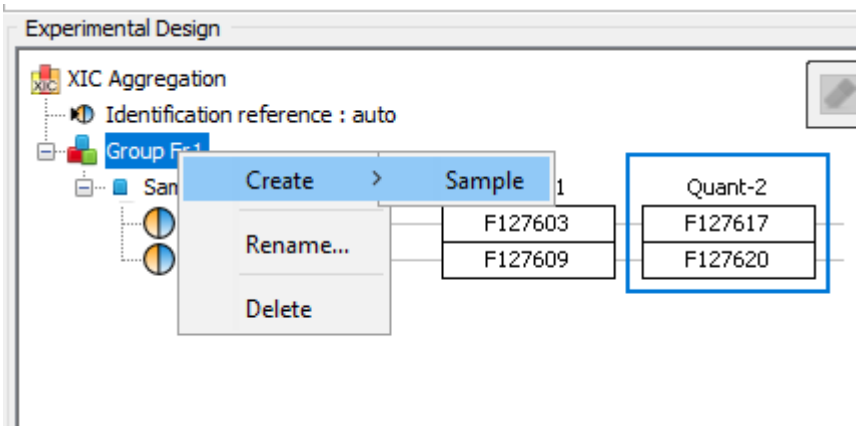
The following dialog is opened:



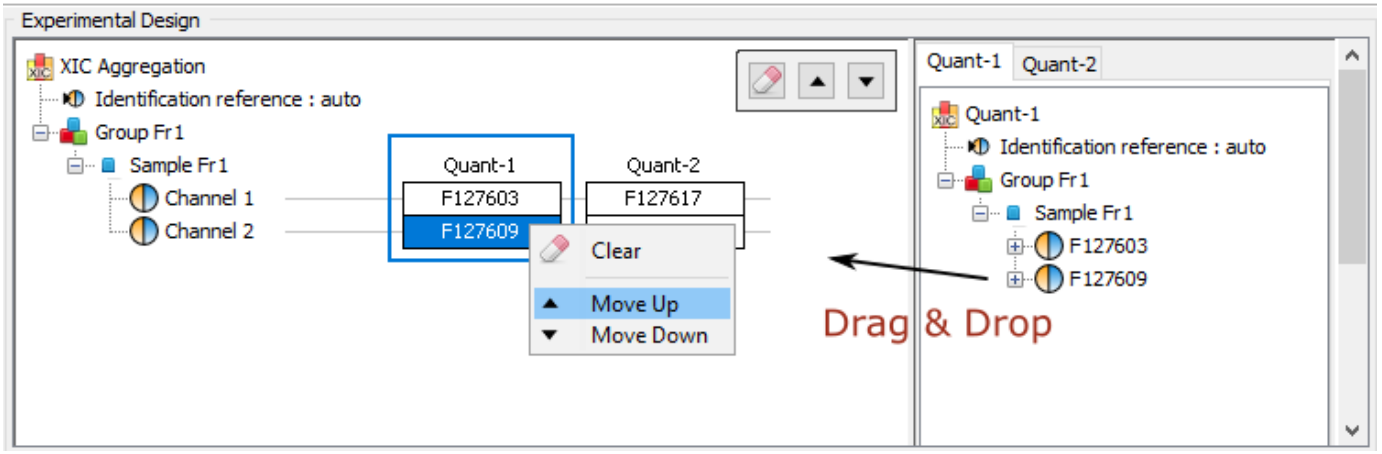
This dialog is divided in three areas:

- **left area**: aggregated experimental design. This design is inferred from the quantitations that will be aggregated.
- **center area** : quantitation channels mapping. Each quantitation channel of the aggregation will correspond to sample analysis of aggregated quantitations.
- **right area** : XIC Sources

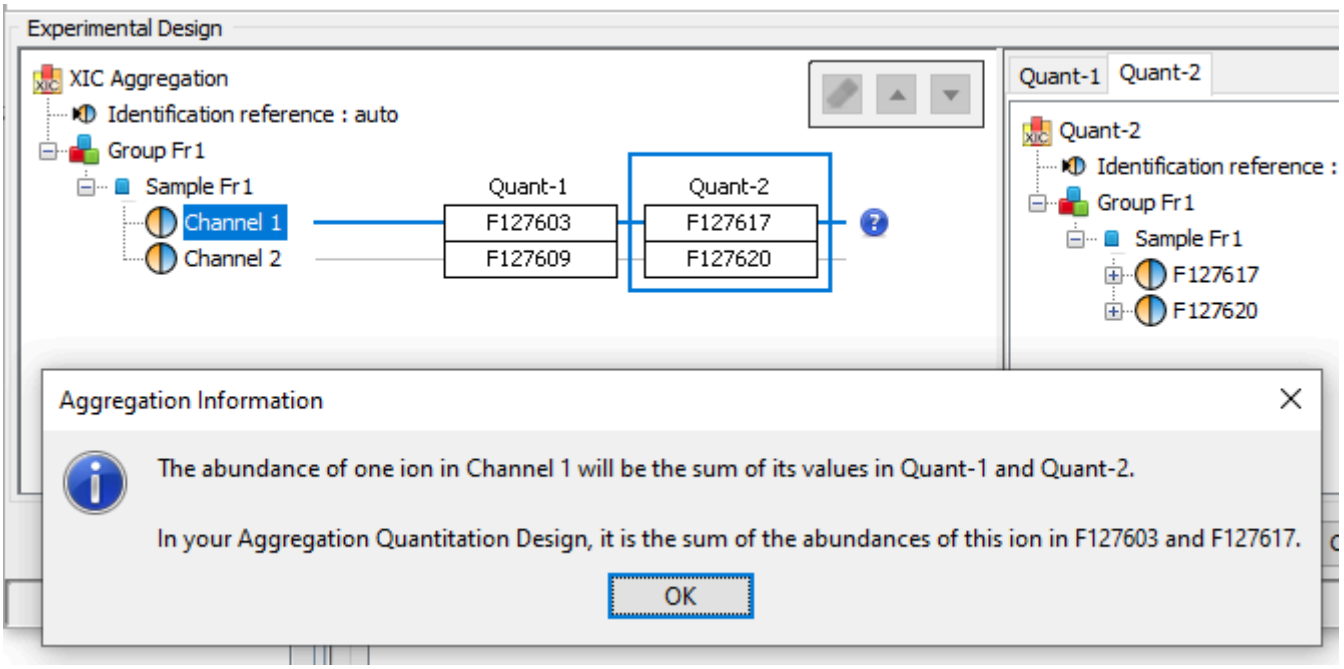
You can completely modify the inferred aggregated experimental design. You can use drag & drop to modify the order of the items and the right mouse popup to create, delete or rename items.



You can modify the quantitation channels mapping by drag and drop from the right area to the center area. You can also remove or reorder the mapping thanks to the right mouse popup or thanks to the floating panel.



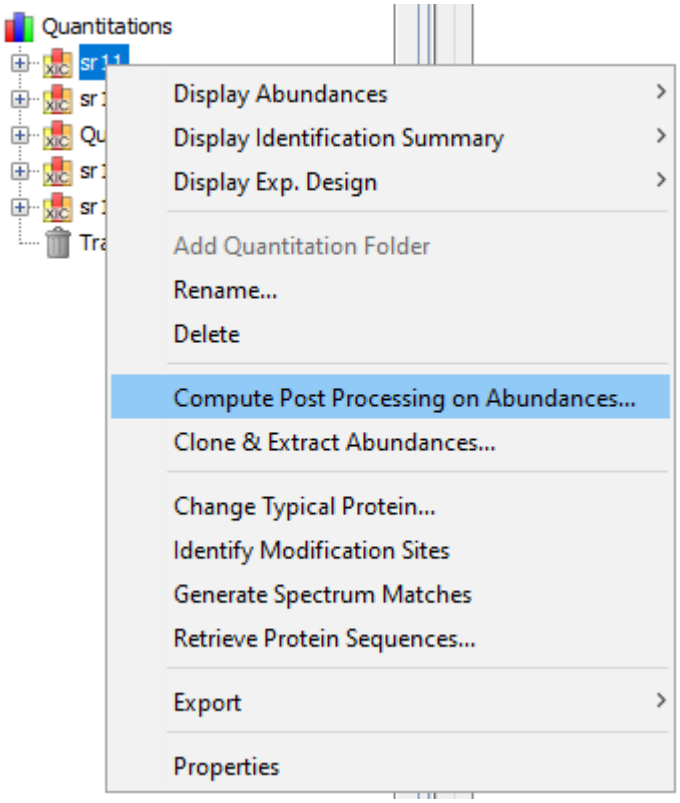
When you select a Channel, you can click on the help button to check what the aggregation will exactly do.



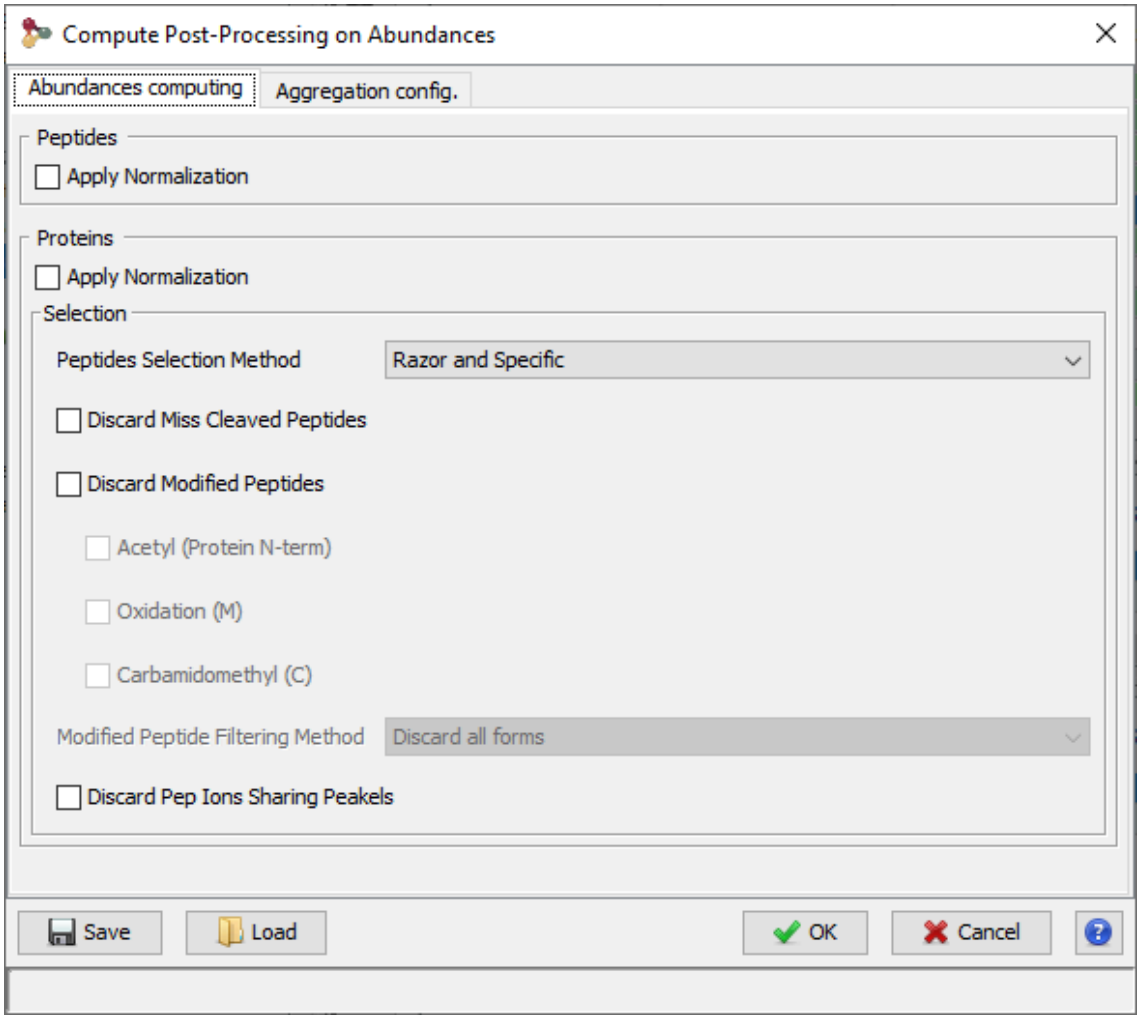
Compute Post Processing on abundances

Advanced Protein Sets abundances

Right click on the selected XIC node in the Quantitation tree, and select “Compute Post Processing on Abundances…”


Quantitations
Display Abundances
Display Identification Summary
Display Exp. Design
Add Quantitation Folder
Rename...
Delete
Compute Post Processing on Abundances...
Clone & Extract Abundances...
Change Typical Protein...
Identify Modification Sites
Generate Spectrum Matches
Retrieve Protein Sequences...
Export
Properties

Configuration



In the dialog, you can:

- specify peptides to consider for quantitation
- configure abundances computing parameters
- select aggregation mode

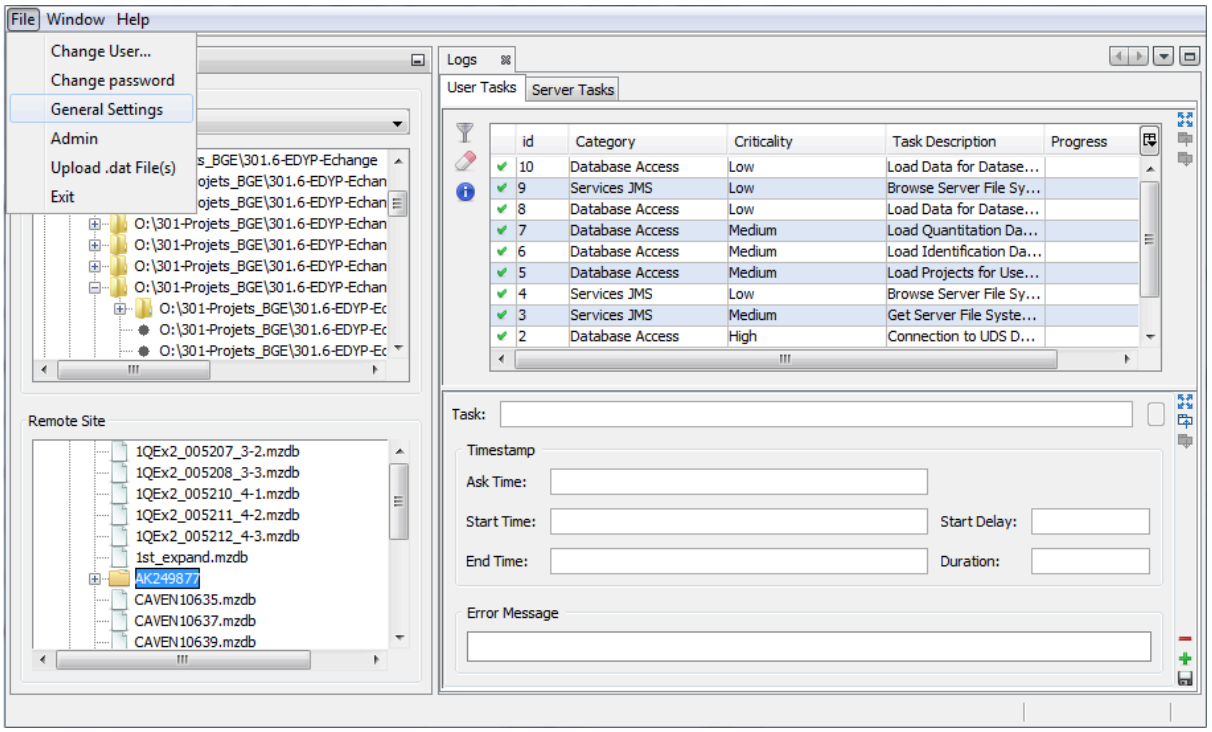
For more details, see Post-processing of LC-MS quantitative results

Advanced XIC results

You can see the results by displaying the XIC (Display a XIC) or export them Export Data

General Settings

Since version 1.4 Proline Studio includes a general settings dialog which can be accessed from the top menu bar clicking on “General Settings” as shown in the screenshot below.

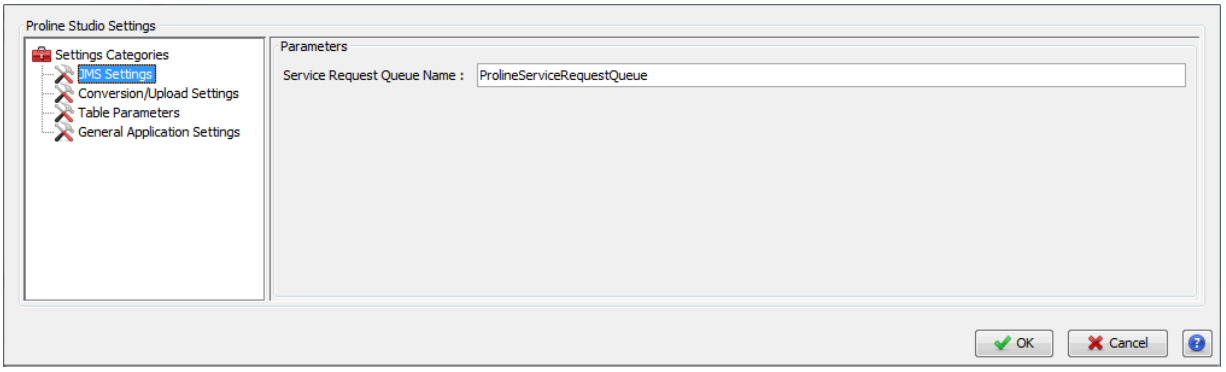


The latter one consists of a constantly but slowly growing number of user preferences regarding various aspects of the utilization of Proline Studio. Based on their context, for the time being, preferences are organized into the following four tabs:

- JMS Settings
- Conversion/Upload Settings
- Table Parameters
- General Application Settings

JMS Settings

JMS Settings tab contains parameters that concern the exchange of messages between your local machine and the JMS Server. It should be made clear that compared to other preferences, preferences that are included in this tab should be treated with caution. Mistreating a communication preference can lead either to communication/connection problems or to users’ confusion to whether they are connected to the correct server version.

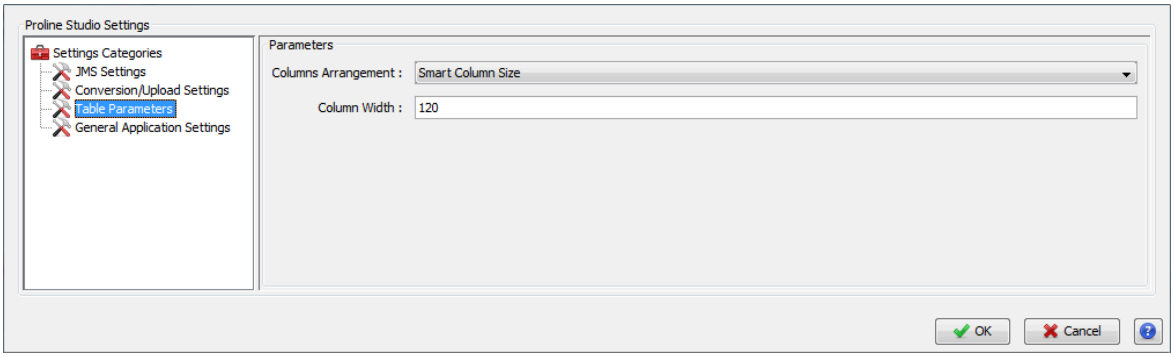


Service Request Queue Name

Parameter can be seen as a name which represents a server address. The parameter's existence is justified by the fact that multiple server versions might run on the same server machine imposing the need to be able to canalize Studio's messages appropriately.

Table Parameters

Table parameters' tab encapsulates a short list of preferences regarding all tables generated throughout Proline Studio. More specifically those preferences control the arrangement of the participating columns as well as their respective width.

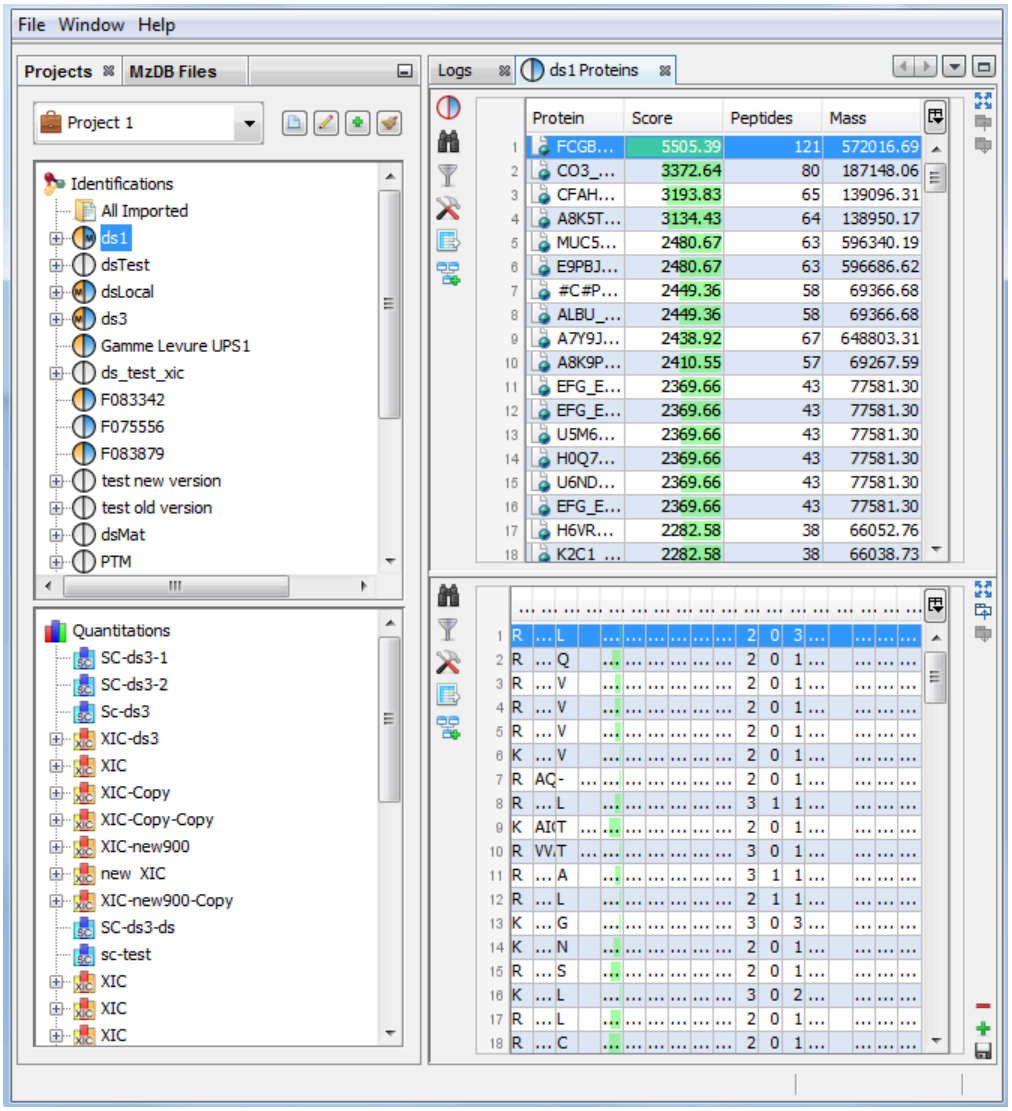


Columns Arrangement

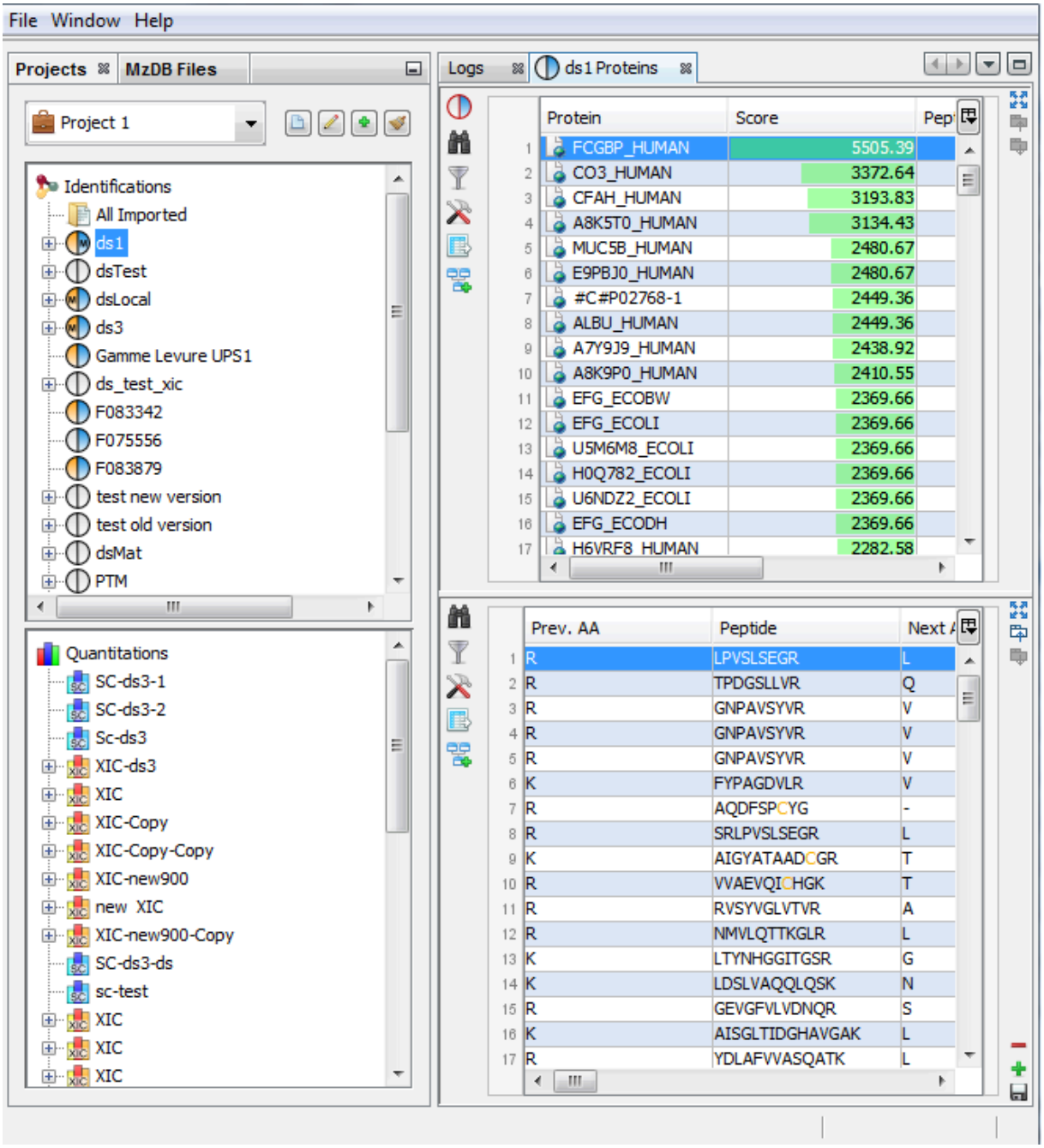
This field dictates the spatial arrangement of table columns. Three arrangements are possible:

- Fixed Column Size
- Automatic Columns Size
- Smart Columns Size

When "Automatic Columns Size" is used, all columns are width-wise readjusted in a way that they all fit to their container. Given that it is a "fit-to-screen" approach, it lacks scrollbars and does not guarantee the readability of the presented date, especially when the number of columns is high.



On the other hand a simpler approach that guarantees readability is to select “Fixed Column Size”. In this case all table columns have a fixed width, explicitly dictated by the user using the parameter “Column Width”.

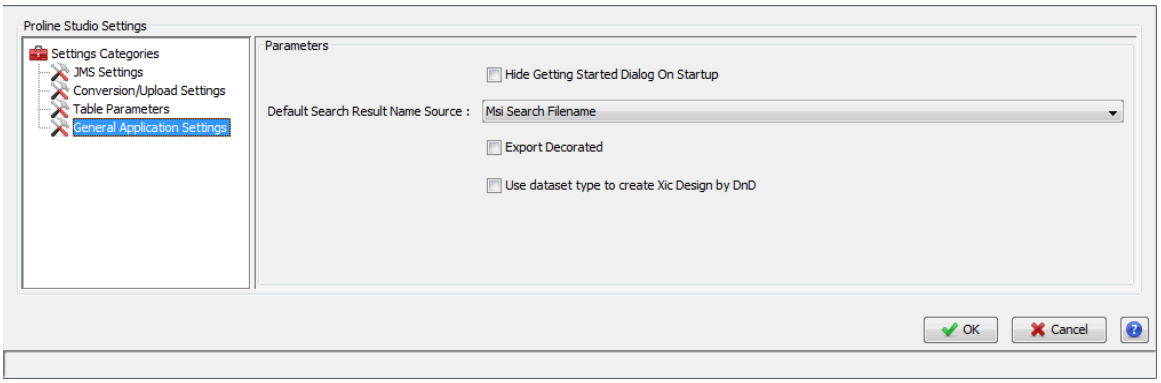


The less clear option is “Smart Column Size” which serves as a trade off between the aforementioned ones. It tackles with the cases that either we have too many columns to visualize using the “Automatic” approach or too few and the selected default and globally applied width imposes unneeded scrollbars ending in hiding some columns at the same time. In this context, “Smart Column Size” can be seen as a simple rule based on the ratio between the mean column width needed in case of “Automatic Column Size” for a specific table and the globally selected width. For the sake of simplicity we have set a threshold of 0.7 or 70% which on its turn determines which one of the two first modes will be used given a table. If the ratio is smaller than 0.7 then the table in question will be presented in “Fixed Column Size” mode. On contraire if ratio is equal or greater than 0.7, then we consider that using “Automatic Column Size” mode is more appropriate as it balances between a possible slightly smaller than desired width and the possibility of hiding a column using scrollbars.

Column Width

The second preference is more or less self explanatory and corresponds to the globally desired columns' width when fixed rune is applied, either directly or as a result of the smart mode.

General Application Settings



In this tab we can find a diverse set of preferences regarding various tasks encountered in Proline Studio. For the time being those preferences are:

- ➔ Hide gettings Started Dialog
- ➔ Default Search Name Source
- ➔ Export Decorated
- ➔ Use dataset type to create a XiC design by DnD

Default Search Name Source

Unlike the last three, Hide Getting Started Dialog is pretty much self explanatory. The second preference on the other hand, "Default Search Name Source" affects the way identification datasets are named on importation. For this preference we have three possible options:

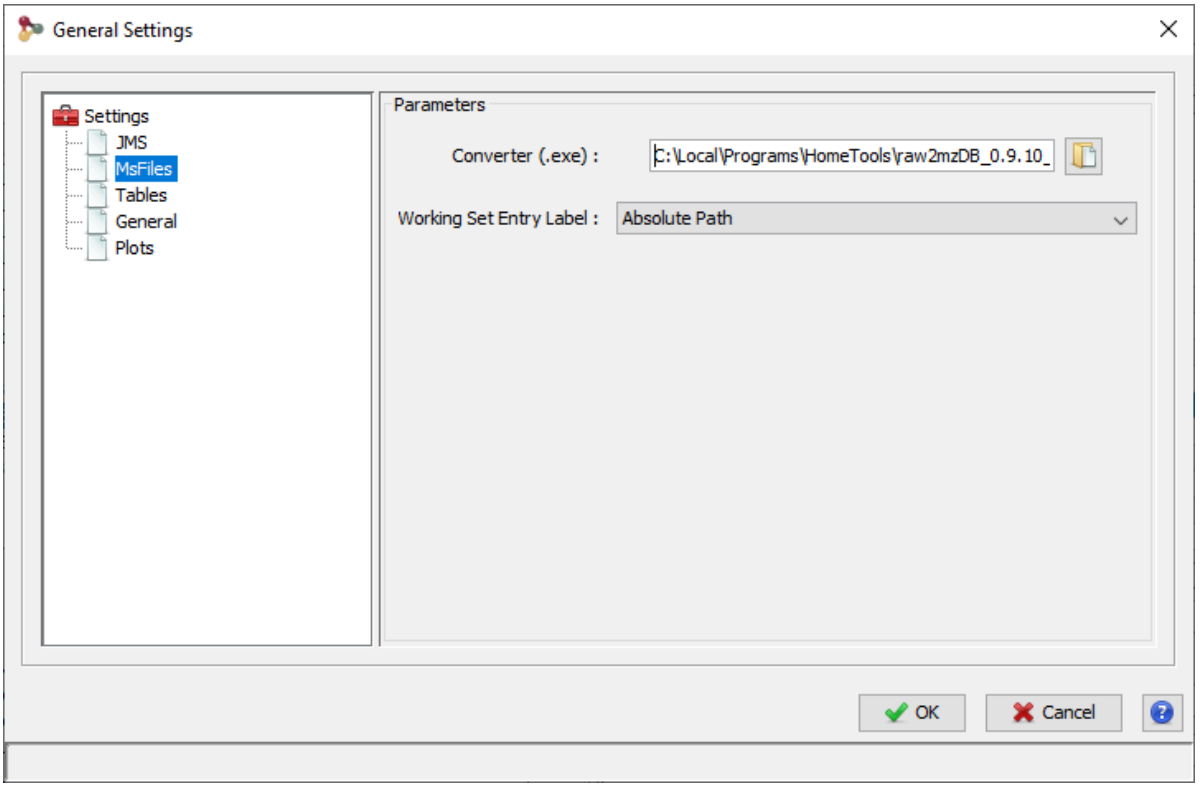
- ➔ Search Name (E.g Gamme Levure UPS1)
- ➔ Peaklist (E.g OEMMA121101_36.raw)
- ➔ MSI Search Filename (E.g F054967)
- ➔ Mascot Rule (...)

Export Decorated

This parameter affects the .xls and .xlsx files that are produced in the process of export (client side). It could be easily described as a preservation of any existing Rich Text Feature in a table. (Colors, Font Weight etc.)

Use dataset type to create a XiC design by DnD

MsFiles



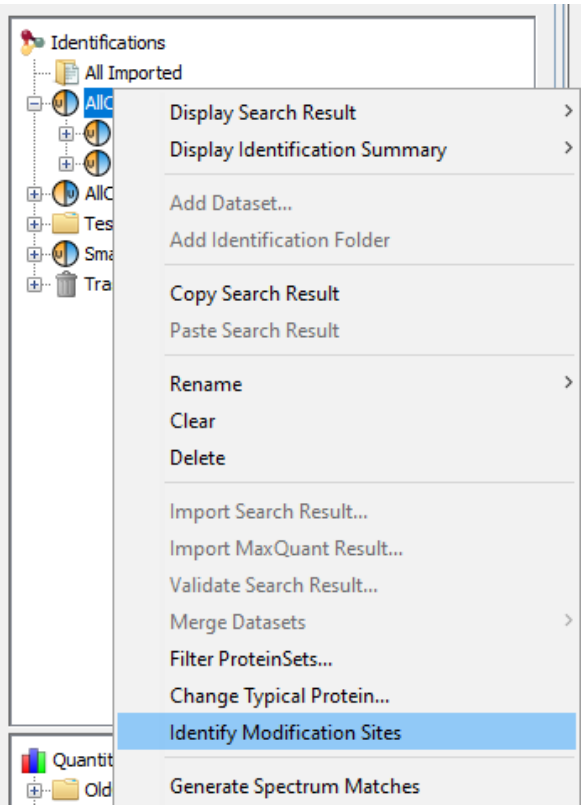
Converter (.exe)

Corresponds to the default raw2mzDB converter. While it is left at the discretion of the user, which version to choose, it must be noted that different versions do tend to work better with specific type raw files. Said that, it is also important to understand that in order for a conversion to be successful within Proline Studio, all system requirements set by the specific raw2mzDB version must be met.

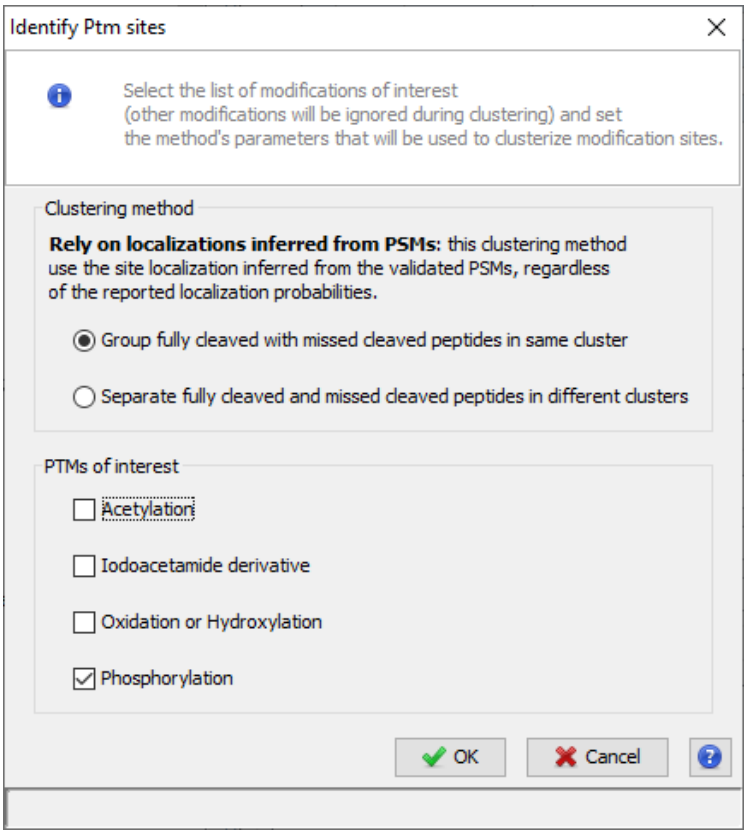
Identify Modification Sites or Clusters

For description on sites and clusters you can first read the principles in this page: Identifying Modification sites and clusters

To identify Modification sites and create Modification clusters, right click on the Identification Summary and select *Identify Modification sites*.



A dialog is displayed where you can choose Modification of interest and configure the clustering method (see [Identifying Modification sites and clusters] for more details).

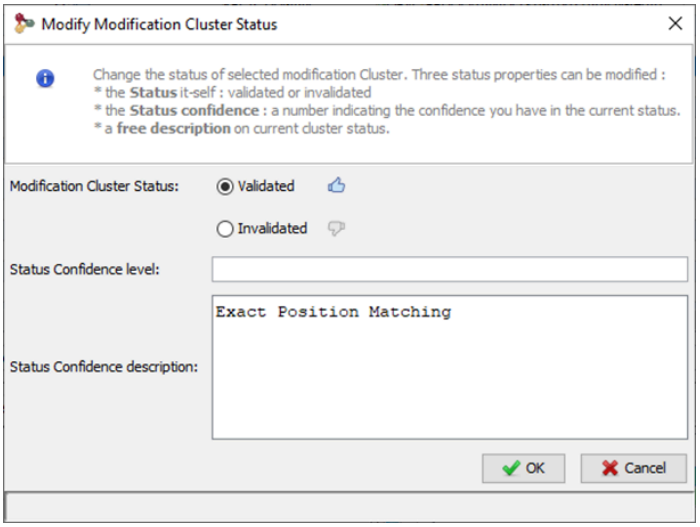


Annotate and Edit Modification Clusters

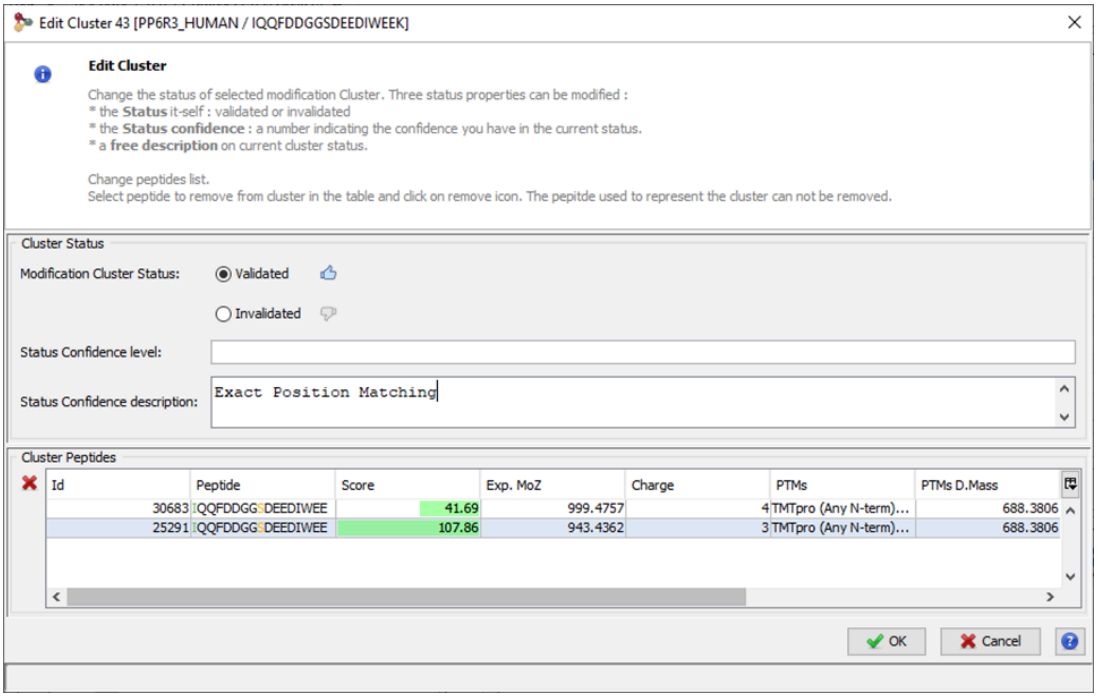
After running Identify Modification for a specific Identification Summary or Quantitation dataset (see above and Identifying Modification sites and clusters for more details), users can annotate and edit these clusters.

In the associated display, few operations are allowed.

- Change the status of the cluster by clicking on the icon. The following dialog appears, allowing user to
  - Validate/Invalidate the cluster
  - Add a confidence level : any number which signification is user specific
  - Add a free description



- Edit the selected cluster by clicking on the icon. In this case, the status dialog is displayed in addition with the list of peptides of the cluster. Same status annotation as described above can be done. Users can also remove some peptides from the cluster. The removed peptides will be removed from the whole “PTM Dataset”. They will not be added to an existing or new cluster. Note: Peptide used as reference could not be deleted, a single peptide could be deleted at a time.



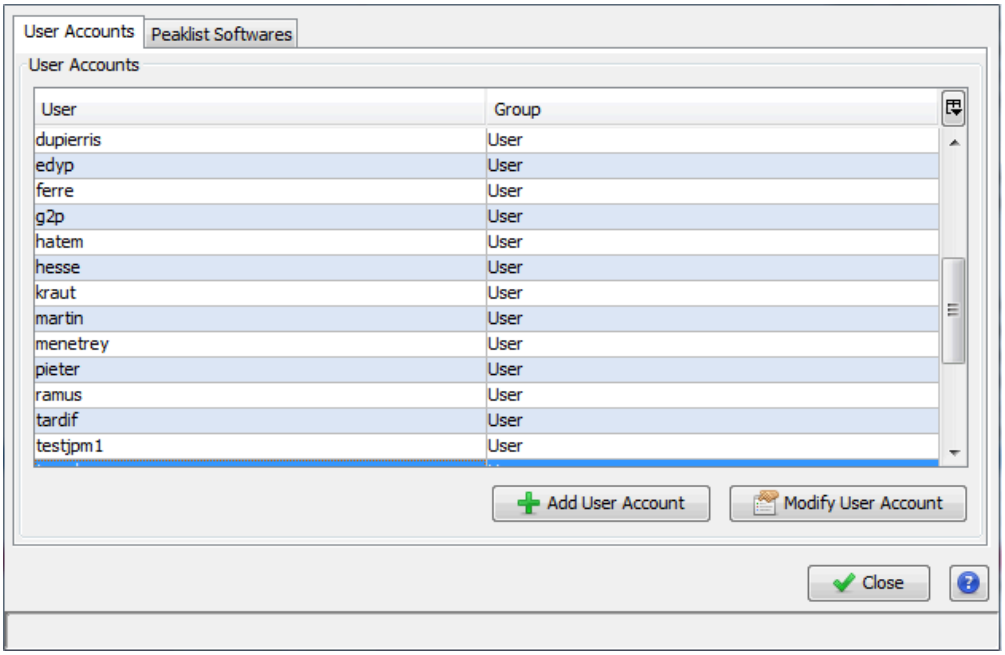
- Merge co-localized clusters by clicking on icon. Users must select clusters that should be merged. These clusters must be co-localized, see Identify Modification Sites or Clusters .
- Save edition by clicking in icon. All modifications done on clusters will not be saved in datastore until user explicitly ask for it. When saving annotation, an Annotated Modification Dataset is created and will be accessible through the *Annotated Clusters* menu. This permits users to still access the initial Modification Dataset. Warning : there is only one Annotated Modification Dataset per Modification Dataset. If new annotation is done and saved from initial Modification Dataset, previous Annotated Modification Dataset will be erased

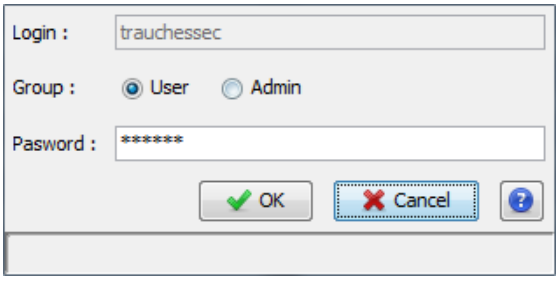
Administration

Some administration views and operations are accessible through the Proline Studio *File > Admin* menu. Edit functions are only permitted to advanced users, who at the same time hold the status of Admin..

User Accounts

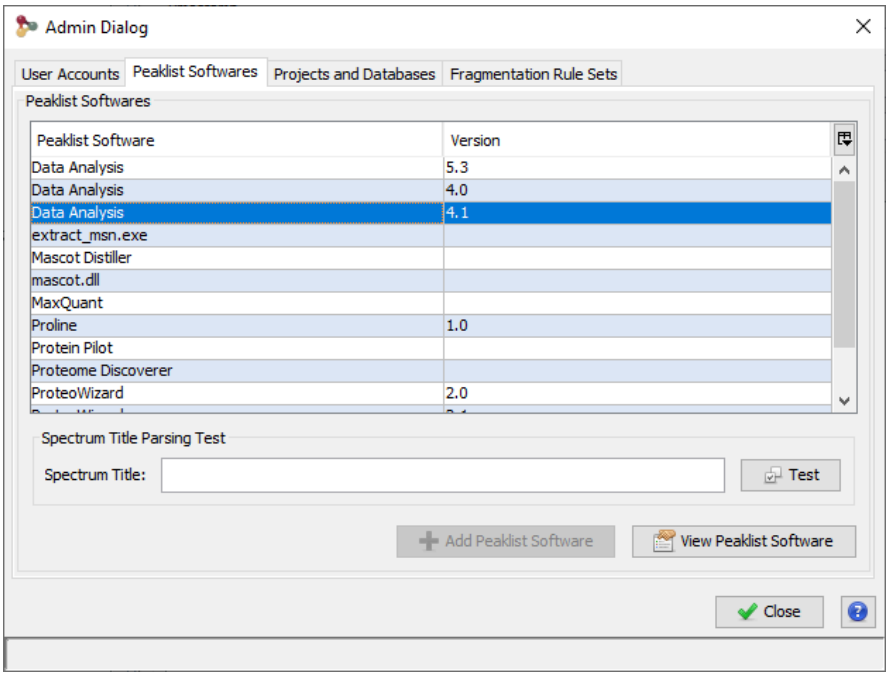
The *"User Accounts"* tab, list all registered users with the group they belong to. *Admin* users can add new users or modify existing one. Modifications consist in changing user groups or passwords.



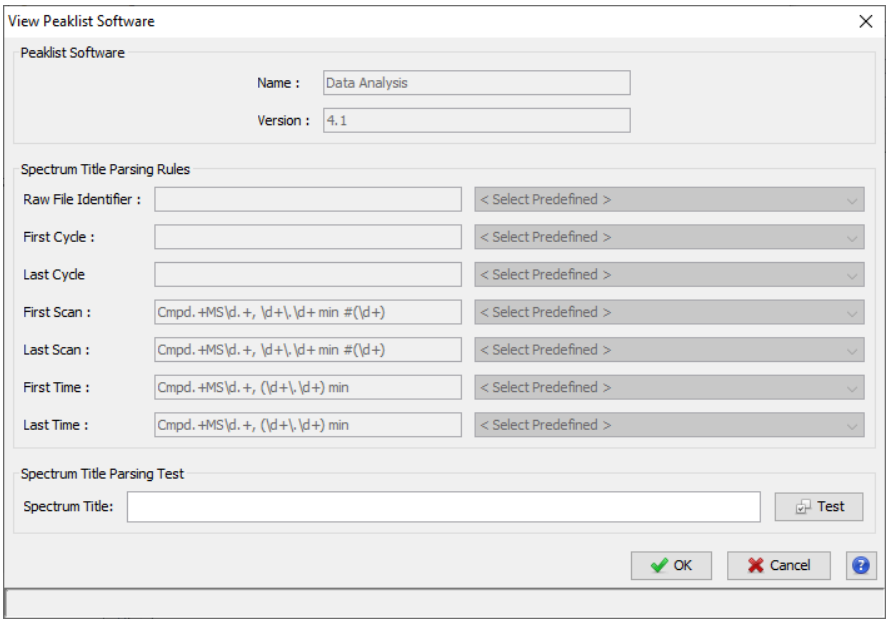


User account modification dialog

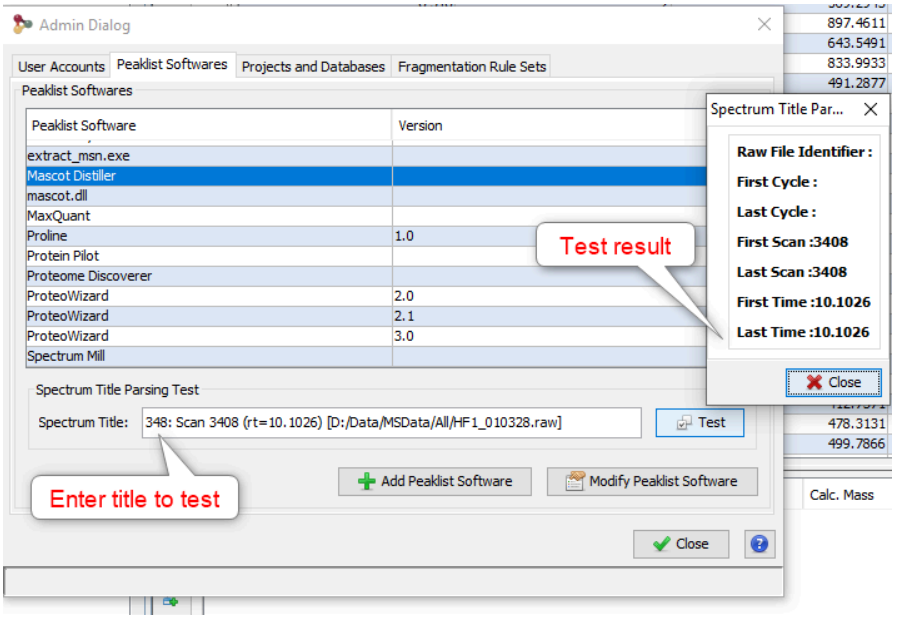
Peaklist Softwares



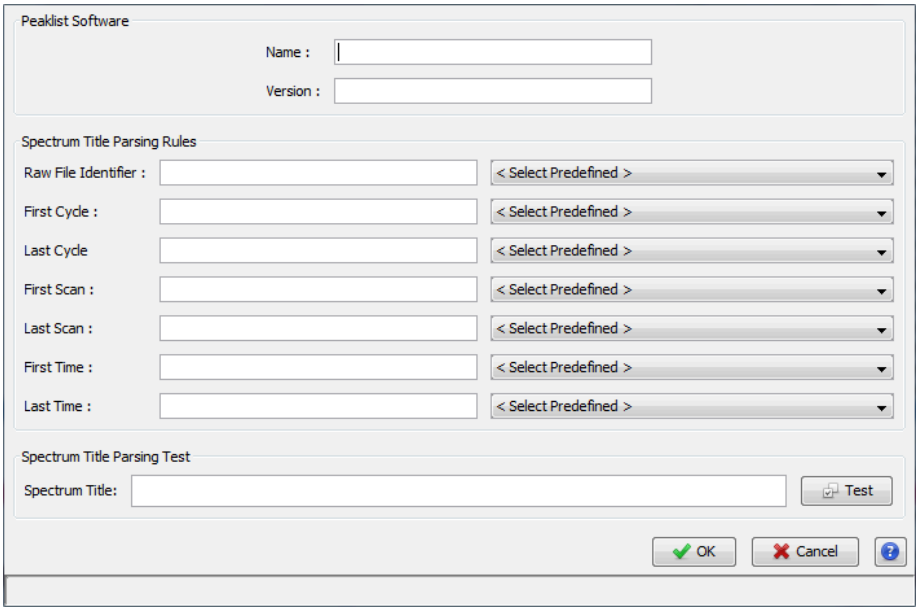
Peaklist Softwares tab permits users to view the list of all available Peaklist Software. A detailed view could be displayed when selecting the “View Peaklist Software” button.



In both windows, It is possible to test a spectrum title to verify which data are extracted for a given Peaklist Software.

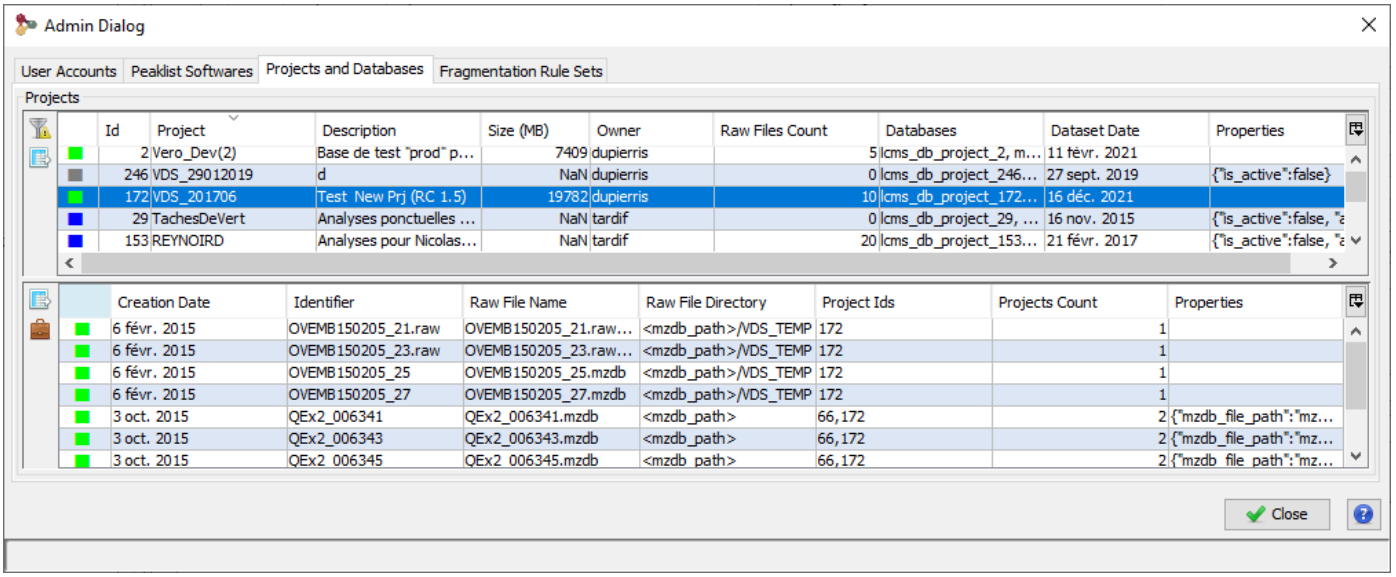


Admin users also have the possibility to add new Peaklist Software definition (name, version and extraction rules using regular expression) or to modify existing definitions by changing the name or version. Rules are not modifiable since some already parsed data may not be consistent with the new definition. New rules definition may be done by creation of a new peaklist software.



Projects and Databases

This tab lists all projects defined on the Proline Suite.

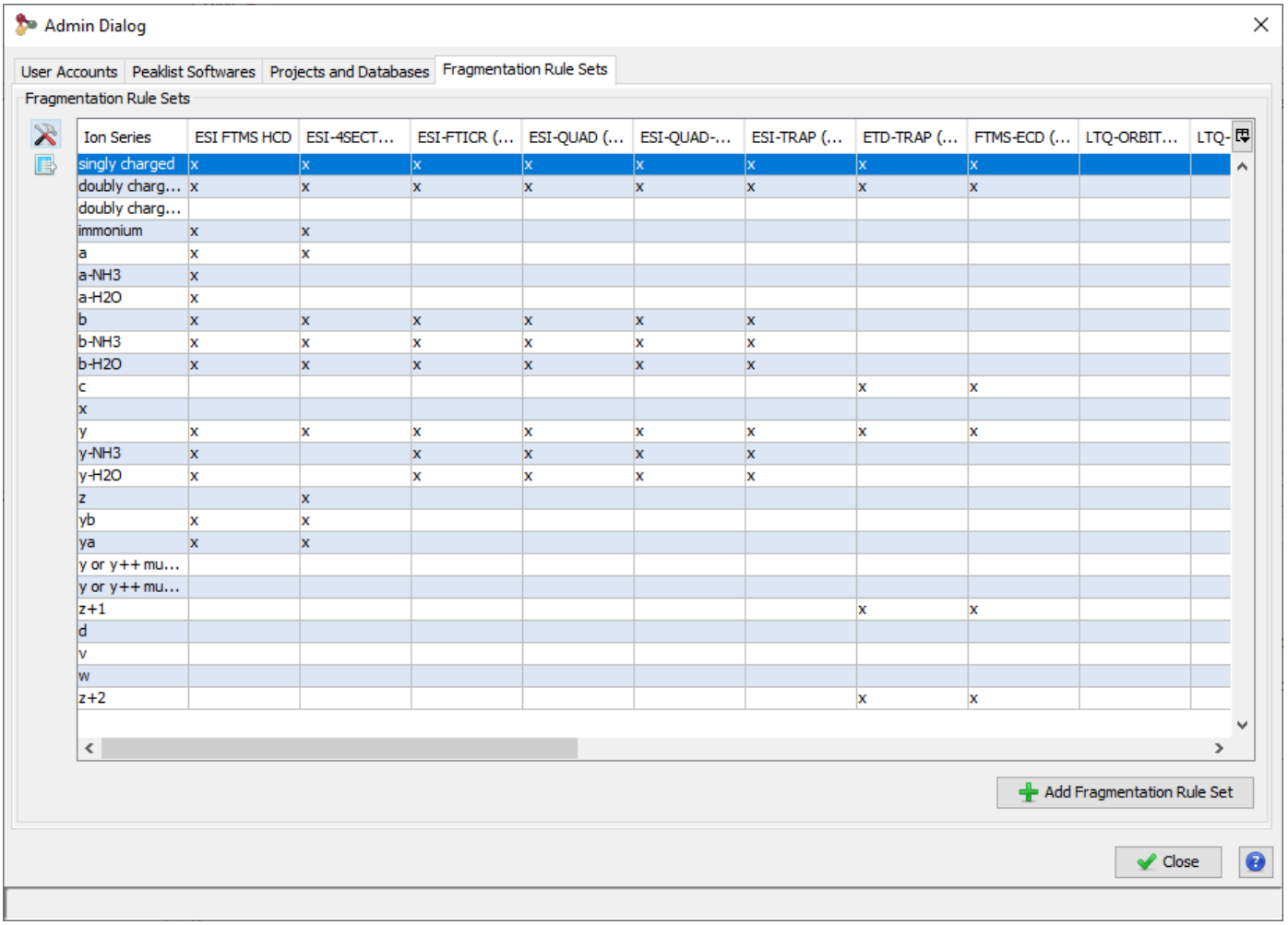


Upper view: List of projects with associated generic information (owner, description) and more administration information such as the name of the databases containing data, size of these databases, last dataset creation… The satus column, first one, indicates if the project is active (green), inactive (blue) or archived (grey)

Lower view: for selected project, list of the mzdb files used for quantitation. This is useful to do some cleanup on disk when archiving projects. The status indicated if this mzdb file is used in active projects only (green), in active and archived projects (yellow) or in archived projects only (grey).

Fragmentation Rule Sets

This tab show all fragmentation rule sets defined in Proline Suite and the rules they use. The display is similar to the one defined by Mascot ®



Ion Series	ESI FTMS HCD	ESI-4SECT...	ESI-FTICR (...	ESI-QUAD (...	ESI-QUAD-...	ESI-TRAP (...	ETD-TRAP (...	FTMS-ECD (...	LTQ-ORBIT...	LTQ-
singly charged	x	x	x	x	x	x	x	x		
doubly charg...	x	x	x	x	x	x	x	x		
doubly charg...										
immonium	x	x								
a	x	x								
a-NH3	x									
a-H2O	x									
b	x	x	x	x	x	x				
b-NH3	x	x	x	x	x	x				
b-H2O	x	x	x	x	x	x				
c							x	x		
x										
y	x	x	x	x	x	x	x	x		
y-NH3	x		x	x	x	x				
y-H2O	x		x	x	x	x				
z		x								
yb	x	x								
ya	x	x								
y or y++ mu...										
y or y++ mu...										
z+1							x	x		
d										
v										
w										
z+2							x	x		

This information is used for [generating spectrum matches.](#) Custom fragmentation rule set can be added if needed by “admin” user.