*my*ProMS

User's Guide

(02/03/18)

Table Index

Introduction

Proteomic Mass Spectrometry (MS) generates complex data that require multiple steps of computational and manual processing to be translated into meaningful biological information. To be successful, this process requires the skills of MS specialists, bioinformaticians and biologists. To facilitate such collaboration, we have developed myProMS (Poullet et al., 2007), a web-based tool that rationalizes this data processing workflow while allowing multiple users to interact with the data according to their expertise level.

Typically, output files from MS tools such as Mascot or Proteome Discoverer are imported into myProMS database within a defined experimental context. Spectrum interpretations, protein attributions and variable modification (e.g. phosphorylation) positions can be validated by MS specialists either automatically through dedicated algorithms or manually by visual inspection of each spectrum. Only validated data become accessible to biologists for further investigation. Different quantification and differential analysis methods are provided through intuitive interfaces. Results are displayed as interactive graphics to help users visualize and mine the data. Further biological interpretation is possible using integrated Gene Ontology analyses and extensive linking to external resources.

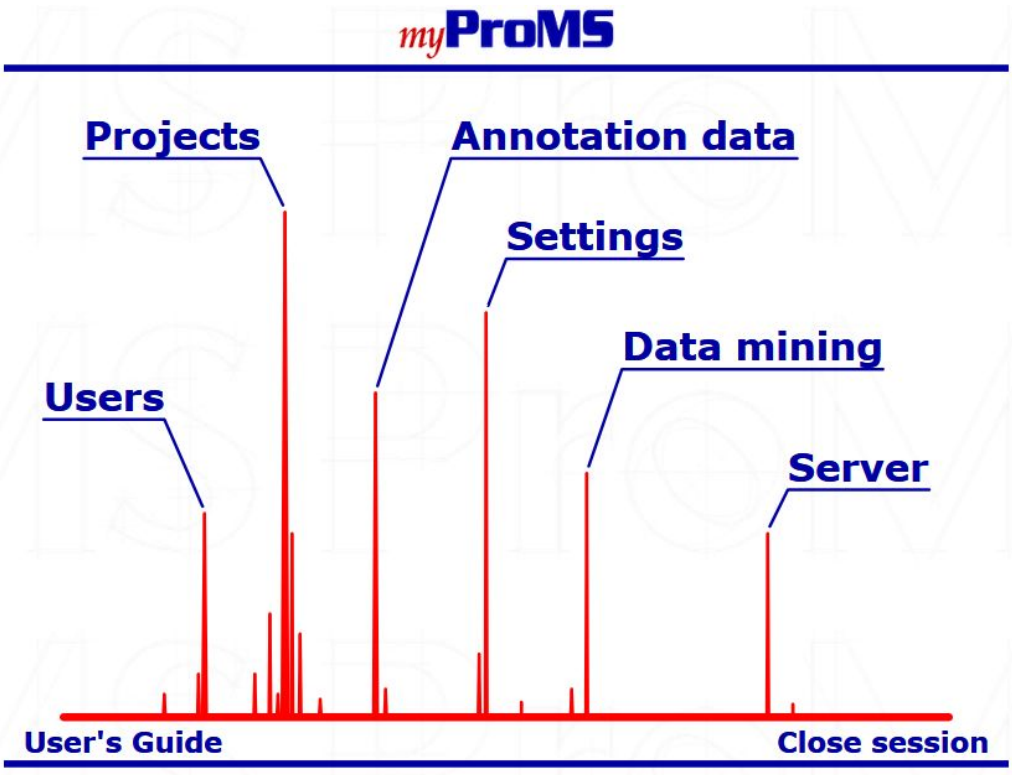
Connection to myProMS server

Login

Access to myProMS data requires a login and password. Contact your local myProMS administrator(s) to request an account. You will then be able to login to the server by clicking on the “Start Session” button from the home page. During login you can choose between tab- or full screen- display modes.

Server main sections

Once logged to myProMS, your login, user class (see the **Users management** chapter below for more information) and connection date are displayed at the top of the browser window. Depending of your user class, either **myProMS main window** (Massist and bioinformatician) or the **Project selection window** (biologist) will be displayed as shown below. Users (including biologists) can access the main window at any time by selecting “*Main Window*” from the Project selection window.



The main window displays links to the 6 areas of myProMS:

- Users management,
- Projects access and management,
- Protein annotation data management,
- Settings management,
- *Data mining section,
- Server management,

**Not yet available.*

Each of these sections is described in a dedicated chapter (see below).

Some sections might not be accessible users depending on the access privileges. Typically, most end users will have access only to their account and projects. See **Users management** chapter below for more information.

Users management

User classes and access privileges

Data access in myProMS is tightly controlled at the user-level to insure data privacy and integrity. Multiple classes of users are defined based on expertise required to perform the different data/users management and processing tasks available in myProMS. In addition, within certain classes, users can be granted additional privileges if their expertise MS data processing justifies it.
There are 4 classes of users defined in myProMS ordered by decreasing access privileges: bioinformaticians, massists, data managers and biologists.

Bioinformaticians

This class of users is intended for server administration and annotation data management. Although bioinformaticians have full access to all functionalities of myProMS, they should not be used to perform routine data processing such as MS data validation has they might not have the necessary expertise. We recommend to keep the number of bioinformatician accounts as low as possible due to their extended ability to modify the data,

Massists

The massist class represents MS experts who are in charge of MS data import, validation and reporting. Massists also manage user accounts and projects creation. By default massists have access to all myProMS functionalities except those normally dedicated to bioinformaticians.

Data managers and workgroups

Data managers have the same privileges as massists but restricted to projects and users of their workgroup. Workgroup usage is optional but is particularly useful for multiple MS-based research labs sharing a common MS facility. In that case, a single instance of myProMS with a workgroup attributed to each lab will insure data privacy while maintaining management centralisation by the MS facility.

Biologists

Biologists are end users of myProMS. They have access to the projects they participate to with various levels of privileges depending of their expertise and involvement in each project. Project access privileges for biologists come in multiple flavors:

+ Project involvement-based privileges:

- **Guest**: A guest user can only access the project data but cannot modify them.
- **User**: Can access and modify project data.
- **Administrator**: Same as a **user**. In addition, a project Administrator can grant other users access to project.

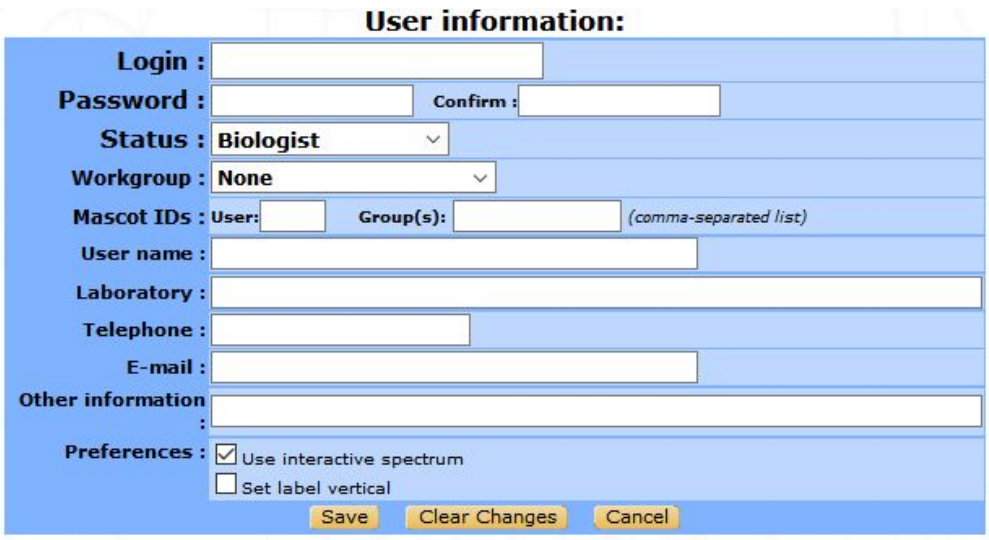
+ Expertise-based privileges:

Biologists (users and administrators) can be granted partial access to MS data validation if their knowledge of the procedures involved is judged sufficient:

- **Power** (User/Administrator): Can enter “Validation” mode to validate protein identification data only.
- **Super** (User/Administrator): Can enter “Validation” mode to validate both protein and peptide identification data.

Account management

From myProMS main window, follow the **Users** link (this link is named **Account** for biologists).



Once in the User section, users can either edit their account and change their password. Massists and managers can view and modify other users accounts and create new ones. However a user cannot create an account classe higher than its own (eg. a manager cannot create a massist account). A bioinformatician can be grant extended privileges to a massist or manager to allow this user to perform specified annotation management tasks. in addition, if Mascot is used with group and user ids enabled, this information can be specified for each massist and manager to insure equivalent data access restriction when importing Mascot data directly from myProMS.

Projects

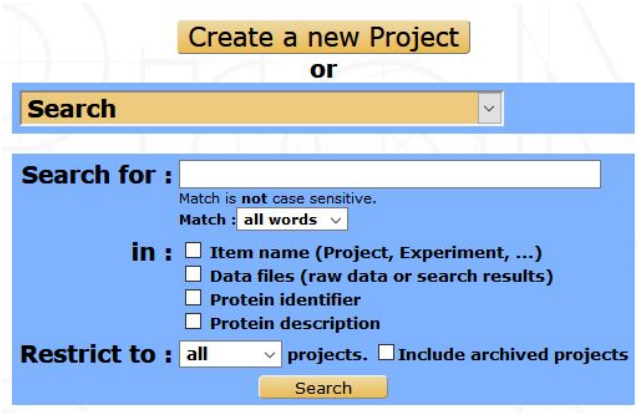
All MS search results and data subsequently generated are organized in projects. A project regroups sets of data that belong to the same user or group of users and generated in the context of a defined scientific project. End-user (biologists and managers outside their workgroup) accessibility to the data is defined at the project level.

Project selection

From myProMS main window, follow the **Projects** link to display the **Project selection** interface.

<Figure Project selection>

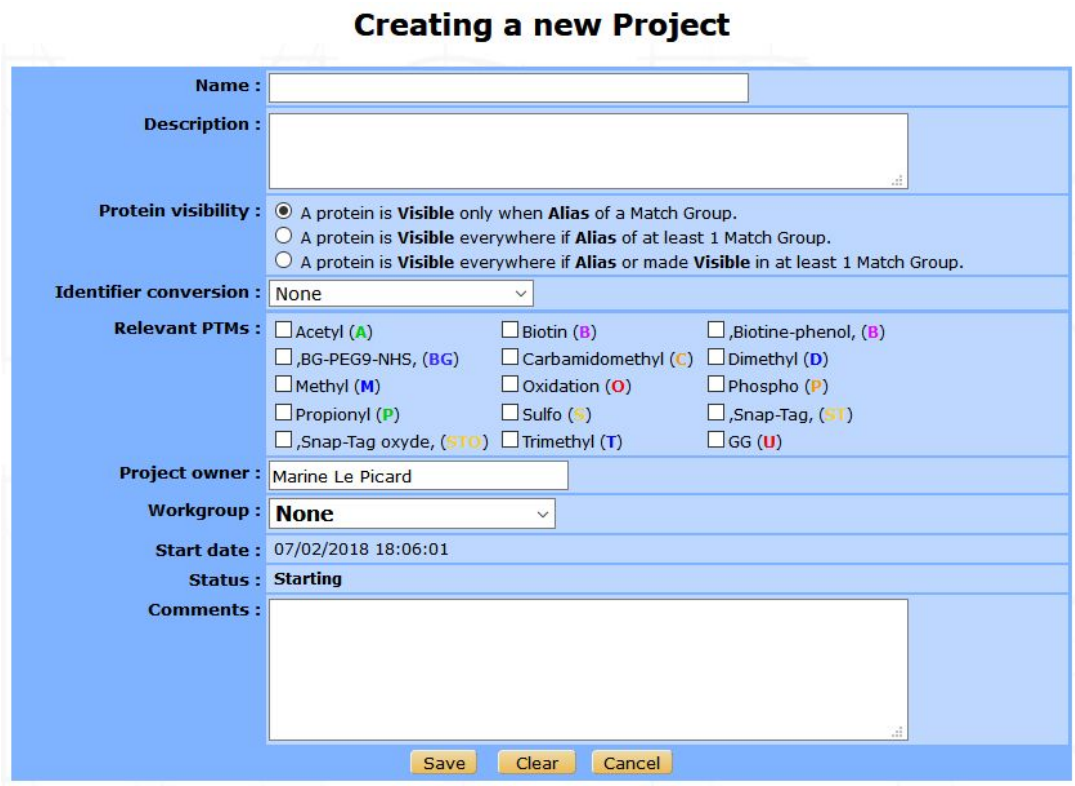
For biologists, a straightforward list of the projects they have access to will be displayed.
For other classes of users, projects can be organized based on the following topics: *Workgroups* (default), *Project owners*, *Active projects* and *Archived projects*. Corresponding projects are listed by ascending name. Their description, owner and/or workgroup together with the access credentials of current user.
If *On-going analyses* is selected, the list generated is composed of Analyses still undergoing validation. Their name, description, data file name, creation date and corresponding project are displayed. This list can be sorted by Import date, Name, MS type, Validation status and Data file name.
Alternatively, a *Search* can be performed using various criterias:



Projects are then listed together with the items that were matched during the search.
Once listed, click on the *Open* button corresponding to the project wanted to access it.

Project creation and settings

Only bioinformaticians, massists and data managers can create projects.
From the **Project selection** interface click on the "*Create a new Project*" button. The following form is then displayed:



- **Name**: Provide a mandatory name for the project.
- **Description**: an optional description for the project.
- **Protein visibility**: Specify the project-wide protein visibility rule to be used. See **Match groups and protein visibility** below for detailed information on this concept.
- I**dentifier conversion**: myProMS tries to map protein identifiers to their synonyms in multiple biological resources. If a conversion is selected, the default identifiers used for protein sequence identification during the MS search process can be replaced by synonyms more meaningful to end users. Unmapped identifiers will not be changed.
- **Relevant PTMs**: Post-translational modifications relevant to project can be selected here. The list of selectable modifications can be modified in **Sequence Modification** section. Information regarding relevant modifications will then be available when performing multiple tasks such as listing, comparing, quantifying and displaying proteins or modification sites.
- **Project owner**: Specify the owner of the project here. This information can be used to sort projects in the projects selection window.
- **Workgroup**: Specify which workgroup this project should belong to if any.

- **Comments**: an optional comments for the project.
Click on the *Save* button to create project.
Projects can be edited at any time to modify any of these settings.

Project life span

- **On-going**: Once created, a project is set as **active** and **on-going**. This means that it can be populated with new items and data. On-going projects are flagged with a yellow icon in the project selection window.
- **Ended**: If the project is judged completed, it can be edited and **ended** by clicking on the *End* button at the bottom of the edition form. Ending a project will automatically end all partially-validated analyses without new reporting (see **Validations** and **Reporting** sections in the **Analysis management** chapter below for more information). Once ended, a project is still active and accessible but can no longer be edited or populated. Ended projects are flagged with a green icon in the project selection window.
- **Archived**: As time passes, some project might no longer be accessed by any users. These projects can be archived to save space on the server. All data files stored outside the database will be compressed. Archived projects are flagged with a red icon and are no longer accessible for data exploration. They can however be listed in the Project selection window by selecting "*List of: Archived projects*".
- **Restoration**: Archived and Ended projects can be fully restored to any activity state if necessary by clicking on the appropriate button in the project home page.

Accessibility

Bioinformaticians and massists have full access to all projects recorded in myProMS. Data managers have full access to all projects within their workgroup. Biologists and managers outside their workgroup must be explicitly granted access to projects when needed. The project access management interface is accessible from the project's home page by clicking on the "*Project Accessibility*" button in the **option frame**.

The interface summarizes the list of users able to access the project together with their credentials. New users can be added one at a time. Once added to the access list, users are guests by default. Select the credentials you wish to provided each user with. The access rights available are listed below the user access form. See also **User classes and access privileges** above for more information.
Click on the *Save* button to validate any changes.

Project navigation

Navigation frame

Sub-navigation frame

Option frame

Results frame

Project organization

Data in a project are hierarchically organized as shown in the figure below:

<Figure Project hierarchy>

Experiments

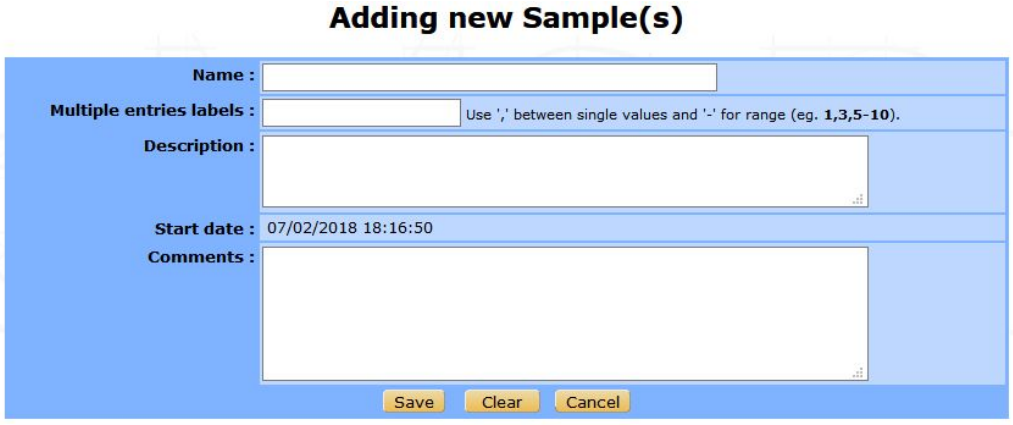
An **Experiment** item represents an actual biological experiment for which MS data will be collected.

To create a new experiment, select the project element in the top left navigation frame and click on *Add Experiment(s)* in the option frame.

Provide a name and optional description and/or comments. Multiple experiments can be created at once if the field Multiple entries labels is filled in. Labels defined in this field will be sequentially appended to the name each experiment created. Labels can be defined individually using a comma-separated string (eg. “A,D,G”) or a range string using a ‘-’ (eg. “1-5”);

Samples

A **Sample** item is a “loose” entity that can represent a single or multiple mixed (e.g. for labelled quantification) biological samples. It can be viewed as a sub-experiment or Analysis-containing item. It is up the user to define its function depending on the experimental context of the analyses it contains.
To create a new sample, select its parent experiment in the navigation frame and click on *Add Sample(s)* in the option frame.



Provide a name and optional description and/or comments. Multiple samples can be created as described for experiments (see the **Experiments** paragraph above).

Analyses

An **Analysis** corresponds to a dataset imported from a single search engine result file: mostly the MS/MS spectra (except for PMF runs), the peptide/protein identifications and associated quantifications when present in the file. Analysis data must be imported, validated and reported before end users can access them and further process their results. These procedures are described in the chapter **Analysis data import and validation** below.

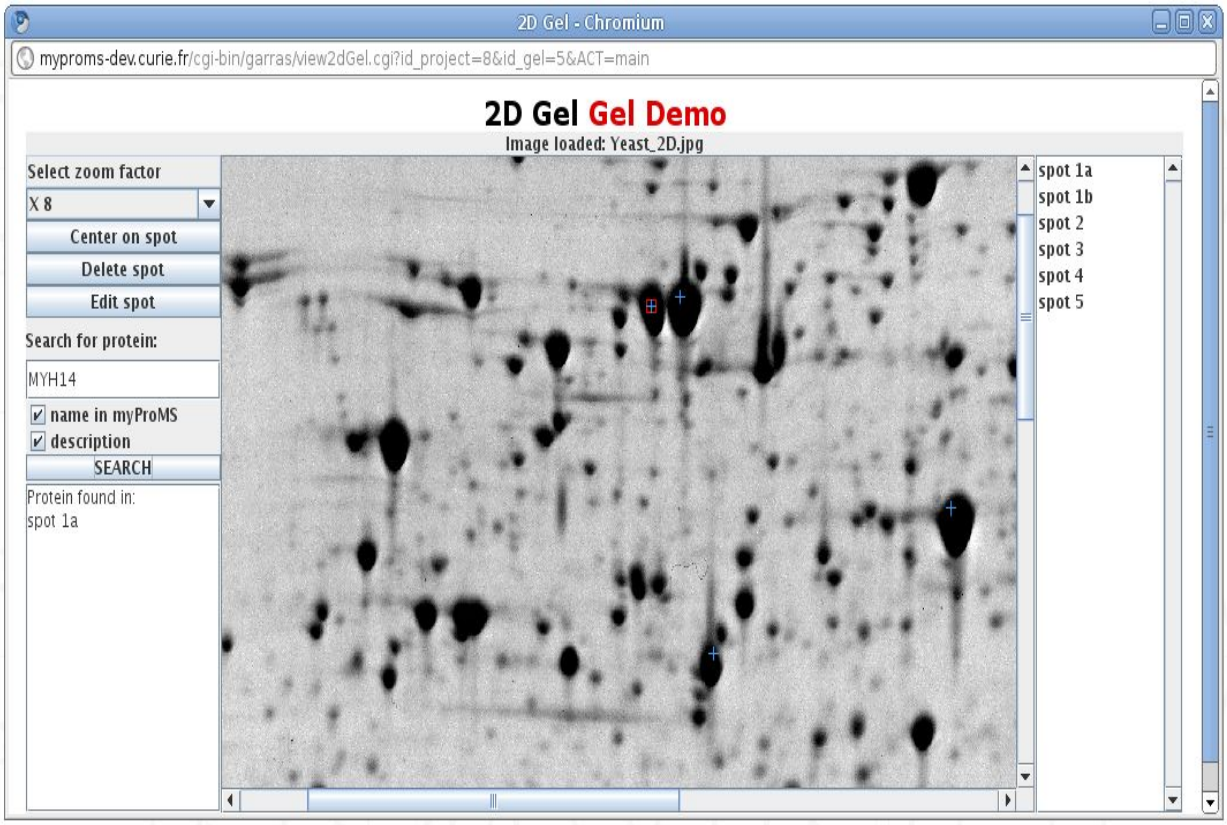
2D gels and spots

2D-Gel (Two-dimensional gel electrophoresis) can be recorded and visualized in myProMS. A picture of the gel (JPEG format only) must be uploaded into myProMS so as to keep a visual record.

To create a 2D-gel, select its parent experiment in the upper navigation frame and click on *Add 2D-Gel* in the option frame. The form shown below will be displayed.

Provide a name, optional description/comments and select the JPEG image file of the gel to be uploaded. Click on *Save* to store the new 2D-Gel element in myProMS.

To visualize a 2D-gel, select the gel element in the upper left navigation frame and click on *Display 2D Gel* in the option frame. A pop-up window will display the gel image as shown below (JAVA must be enabled in your browser and you must accept the security issues).



To record a new spot, double click on the gel image where the spot is located. use the pop-up box to provide a mandatory name and other optional information (isoelectric point, molecular weight, external identifier if you use another image processing software like ImageMaster,...). A blue-cross is then displayed to represent the sport recorded. You can select/unselect a spot by clicking on the corresponding cross. When selected, a spot can be edited or deleted. Protein identification data can be appended to a spot by linking this spot to an existing **Sample**. This sample will no longer be listed in the navigation window and identification data will become accessible through the spot only. When mousing-over a spot with linked identification data, the top-protein (best identification in the associated analysis) is displayed is the spot pop-up box.

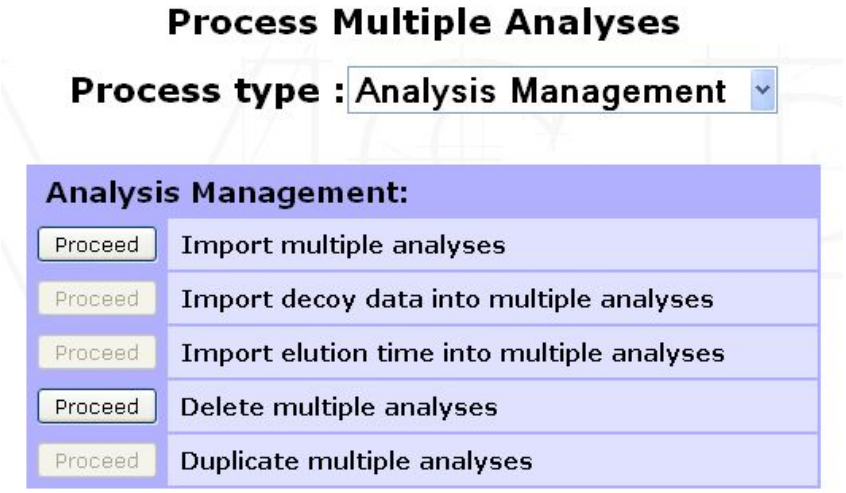
You also can search a protein of interest (in the above example, MYH14) and spot(s) containing this protein will be highlighted with a red square.

In addition to the graphical visualization of the gel, all spots are listed in the lower navigation frame once a 2D-gel is selected. Each spot can be edited/deleted from the usual interface.

Search results data import: MS Analysis

The collection of spectra/peptides/proteins(/quantification) data contained in a search result file are imported into myProMS as an **Analysis**. Only bioinformaticians, massists and managers can import Analyses.

Select the **Experiment** or **Sample** or **2D Gel** into which the Analyses must be imported and click *Process Analyses* in the option frame. From the selection menu displayed, select *Analysis Management* to display the list of available options.

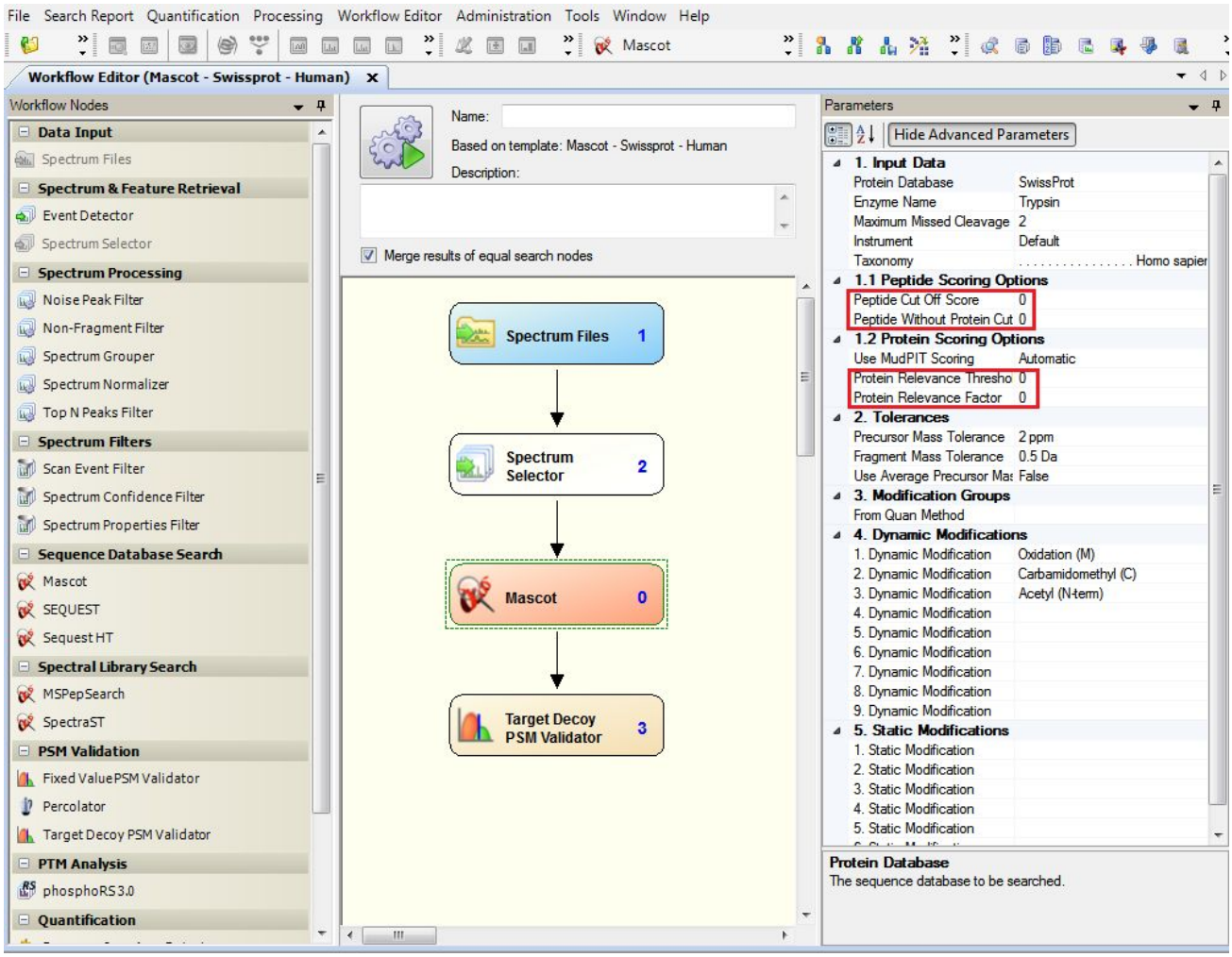


Supported search engines

myProMS allows to import from various search engines :
- Mascot (DAT file or MSF file from Proteome Discoverer Software by Thermo Scientific).
- Paragon (XML generated from ProteinPilot™ Software by AB SCIEX, group2xml.exe).
- Sequest (MSF file from Proteome Discoverer Software by Thermo Scientific).
- Phenyx (XML file generated through Phenyx platform by GeneBio). DEPRECATED!
- Andromeda/MaxQuant (mqpar.xml and 3-4 txt files are required).
- X!Tandem (XML file from X! Tandem pipeline (PAPPSO) or from Trans-Proteomic Pipeline (TPP)).
- PeakView (Exported Excel XLXS file for SWATH data).
- OpenSwath (TSV file from OpenSWATH workflow).
- Spectronaut (TSV file generated from Spectraunot™ by Biognosys).

Important note: if you perform Mascot searches with Proteome Discoverer (PD) Software, make sure you dit not use Protein and Peptide filters. In the **Workflow Editor**, click on Mascot node and then, set the four filters **Peptide Cutt Off Score**, **Peptide Without Protein Cut**, **Protein Relevance Threshold** and **Protein Relevance Factor** to 0. If these filters are not turned-off myProMS import options such as predefined False Discovery Rate (FDR) will not be

accurate.



Collecting search files

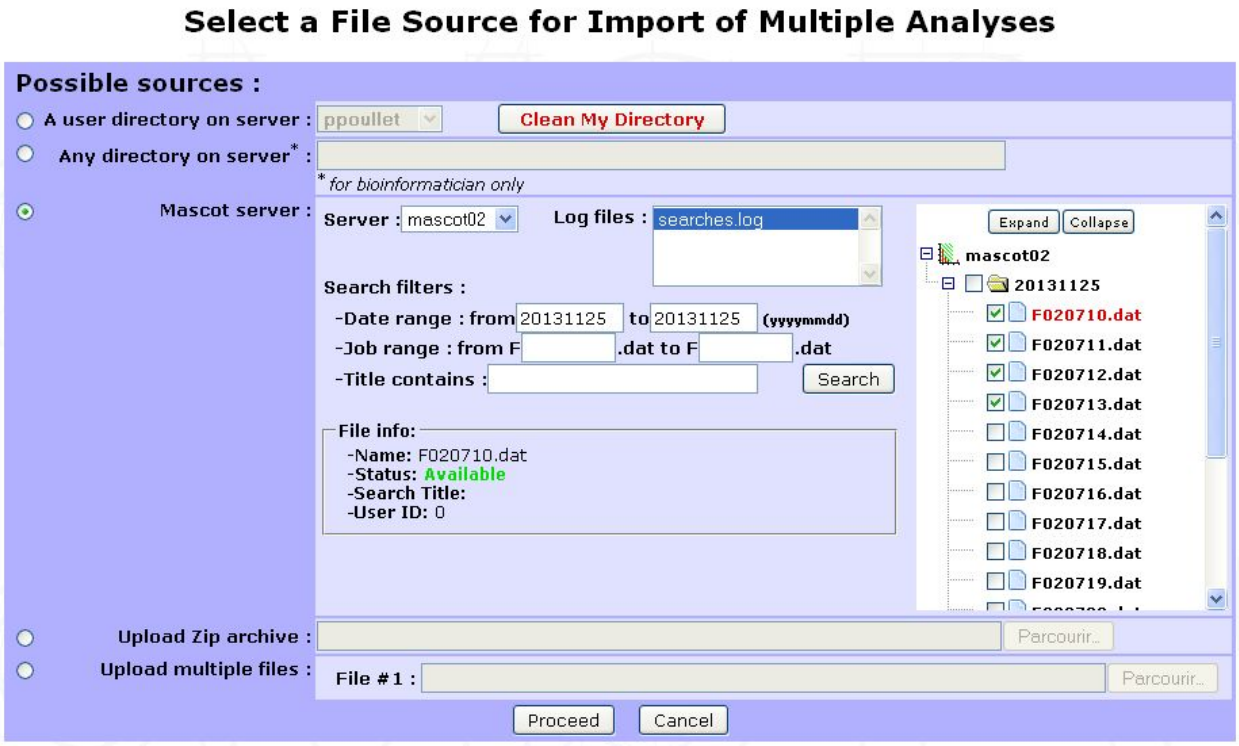
Multiple **Analyses** can be imported at once as long as the corresponding searches were performed with the same search engine and protein databank(s).

Selecting data files (Mascot, Proteome Discoverer and X!Tandem)

Click on *Proceed* next to the **Import multiple analyses** process as shown below. The following form will be displayed to select the source of the search files to be imported.

Multiple import sources are available:

- **A user directory on server**: Following upload, files are stored in a user-dedicated directory on server. These files will stay available to the user until he decides to delete them; either just after import or later. In the later case the user can still access this directory for file management purpose by clicking on the ***Clean My Directory*** button.
- **Any directory on server**: This option is available to bioinformaticians only. The user can provide any path on the server where myProMS should look for search results files.
- **Mascot server**: If a Mascot server is declared in myProMS configuration file, it can be accessed, searched for specific search results and files directly uploaded into myProMS user directory.



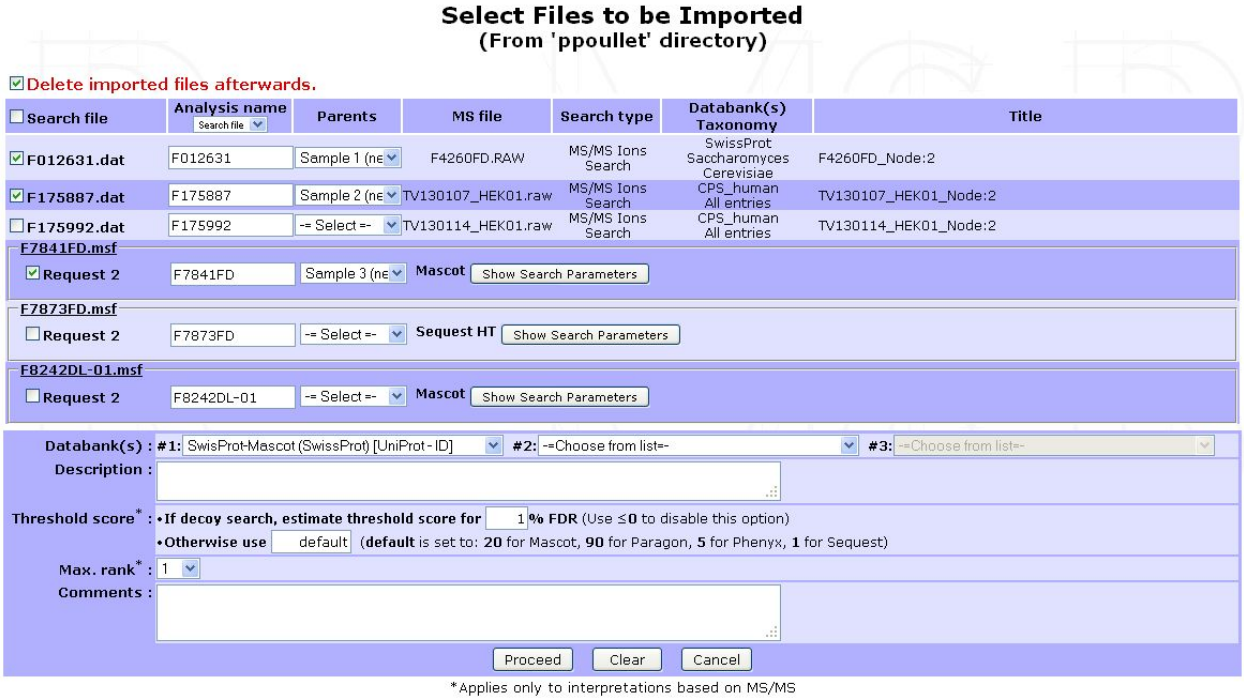
User can search results files by **date** or **job number** range or keywords in the files **search title**. The list matching files is then displayed and grouped by day of creation. Specific information on a file (name, availability, search title and user ID) can be displayed by clicking on the file name. If access to Mascot is restricted (user accounts set up), the Mascot userID must be defined in

myProMS as well (see Account management section above). In this case, only Mascot files accessible to the user will be displayed.
- **Upload Zip archive**: If a large number of files must be imported, they can be uploaded at once as the zip archive. The archive will be unzipped on the server.
- **Upload multiple files**: Alternatively, up to 10 files can be uploaded as séparate files.
Once your files have been selected, click on *Proceed* to initiate file retrieval from the selected source. This procedure may take a few minutes depending of the number and size of the files. Once the transfer is complete, a file import interface will be displayed.
Important Note: Most browsers do not support upload of files with (total) size > **2 Gb**. If files larger than 2 Gb must be uploaded, we recommend to use **Google Chrome**. This limitation does not apply when retrieving files directly from a Mascot server.

Importing analyses

Import parameters (Mascot, Proteome Discoverer and X!Tandem)

Files retrieved are listed in alphabetical order together with pertinent information about the search performed: the MS file, search type (MS2, MS1 or mix of both), databank(s) and taxonomy and search title used. See figure below.

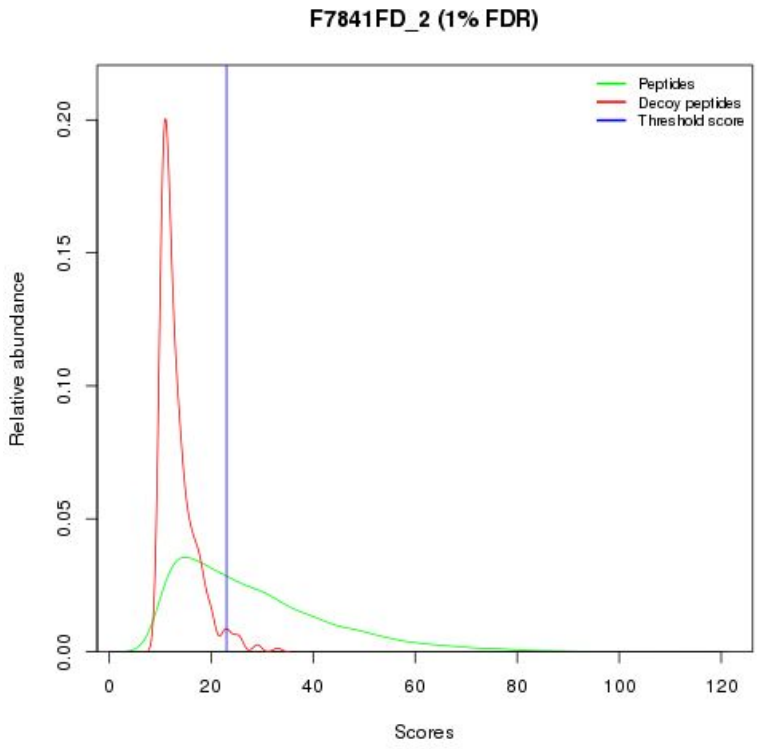


Note: The same Proteome Discoverer msf file can contain multiple searches results (e.g. a search performed with Mascot and another with Sequest). In this case, separate entries will be listed for each search performed together with more details on the parameters used. Each search result can be imported separately and distinct Analysis items will be created.

Proceed as follow to continue data import:

1. Select the files to be imported from the list by checking the boxes on the left-hand side of the files name.
2. Provide a name for the Analysis to be created (**Analysis name** column). This name can be typed by the user or set to match either the name of the search file or that of the original MS file used for the search.
3. If the parent item was an Experiment (or a 2D Gel), each new Analysis must be associated with a specific Sample (or Spot). Pre-existing Samples (or Spots) can be selected from a dropdown menu or the can be created on the fly (Samples only): To create a new Sample, select "New" from the dropdown menu of **Parents** column. A popup window will ask you to provide a name for the new Sample.
4. Select the databanks to be used to extract protein annotations (**Databank(s)** field). If multiple databanks were used during the search (possible with Mascot for instance), the corresponding number should also be selected here. All search files to be imported in the same batch should have been performed using the same or equivalent (set of) databank(s).
5. Define a filtering rule for the data to be imported (**Threshold score**):

- If a decoy search was performed, data can be filtered to based on a user-defined **False Discovery Rate (FDR)** on peptide identification (default is 1%). A threshold score for peptide identification will be determined so that the data imported will (tentatively) match the defined FDR value as illustrated in the figure below. Threshold score calculation can either uses the **qvality** algorithm (Käll et *al.* Bioinformatics 2008,25(7)), the **Mayu** algorithm or the **DT count** algorithm. In DT count mode, decoy (D) and target (T) peptides are simultaneously counted in descending score order until the proportion of the 2 populations matches the selected FDR value.



- If no decoy search was performed or the FDR value was set equal or less than 0, the filtering will be performed according to a minimum (**threshold**) score for peptide identification. A default (search engine-specific) threshold score will be applied unless a different one is provided by the user.

6. Select the maximum number of interpretations allowed of the same fragmentation spectrum (**Max. rank**). The default is 1, but up to 10 can be chosen.

Note: When performing FDR filtering, it is recommended to set this value to 1 since the FDR calculation is based on 1 interpretation per spectrum.

7. Provide optional **description** and **comments**.

8. Decide whether the files should be deleted after import or not (**Delete imported files afterwards**). Unless selected for deletion, file will remain on the server for new import until the user decides to delete them.

9. Click on *Proceed* to initiate the data import into myProMS database.

Importing MaxQuant data

3 to 5 files are required to import a MaxQuant search/quantification into myProMS:

1. mqpar.xml (usually located in the root directory of the MaxQuant search),
2. evidence.txt (file from the Combined/txt directory),
3. Peptides.txt (idem),
4. proteinGroups.txt (idem. Optional, only to import protein quantification data),
5. msms.txt (idem. Optional, only to display peptide fragmentation spectra).

Files 2 to 5 must be compressed in a common archive before import.
Select an Experiment in which to import the data. From the *Process Analyses* window, select either the **Import multiple Analyses** or I**mport quantification** processes and click on *Proceed* next to **Import MaxQuant quantification** to display the form below.

<Image>

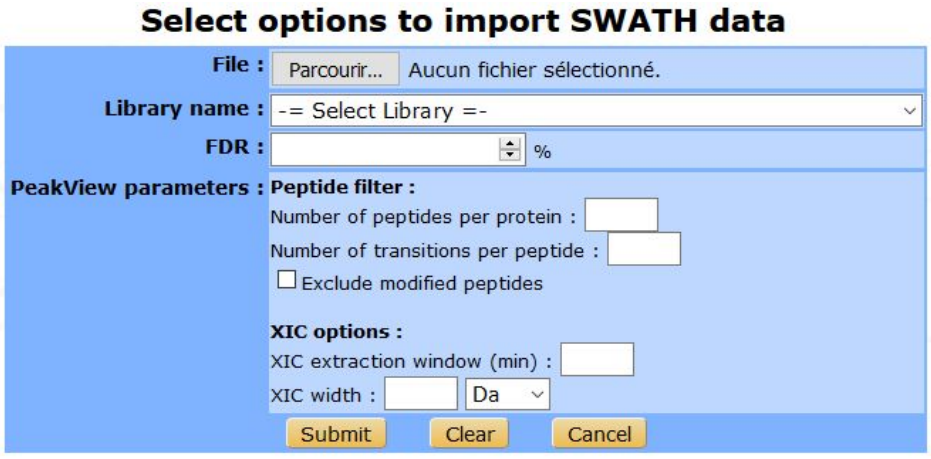
Provide the files mentioned above and the protein sequence databank(s) used for the search. If contaminants were searched, provide a matching contaminant databank. Finally, select the rule you wish to use for protein aggregation into match groups: myProMS or MaxQuant. Specifiy also if you wish to import protein quantification data (the proteinGroups.txt file must be provided in the archive in that case). Submit the form to start data import. Data import will take a few minutes. Samples, Analyses and an experimental Design, a peptide quantification and 1 to multiple protein quantifications will be added to the selected Experiment according to the information extracted from the files uploaded. Peptides and proteins will be automatically validated since they were used in protein quantification.

Importing DIA quantification data

User can import DIA quantification data from three different software : PeakView, OpenSWATH and Spectronaut. Into myProMS select an Experiment and from the Process Analyses window, select the **Analysis quantification** process and click on Proceed next to **Import PeakView/OpenSWATH/Spectronaut data** to display the associated form.

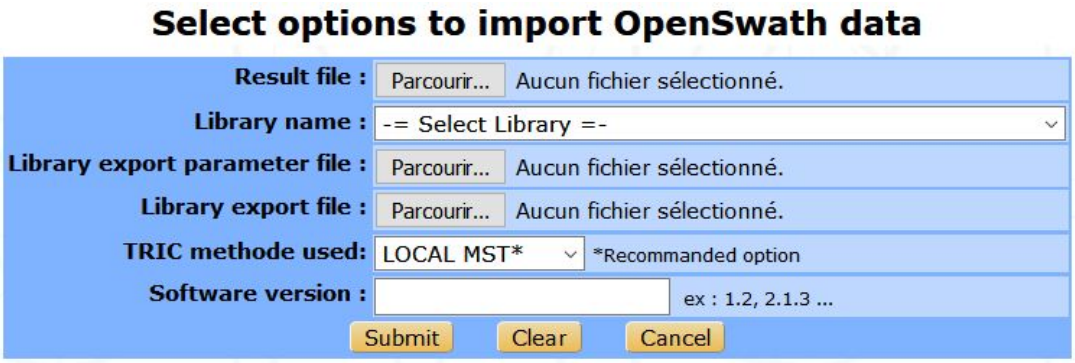
- *From PeakView*

Two files are required : the Excel worksheet file generated by PeakView and the spectral library. Into PeakView, once the experimental SWATH data analysis is over, you can export the result into an Excel file by clicking on the **Quantitation** tab on the toolbar and selecting **SWATH Processing/Export/All**.
The PeakView search parameters can be filled in the following form to be saved in myProMS database to ensure traceability. Then submit the form to launch data import.



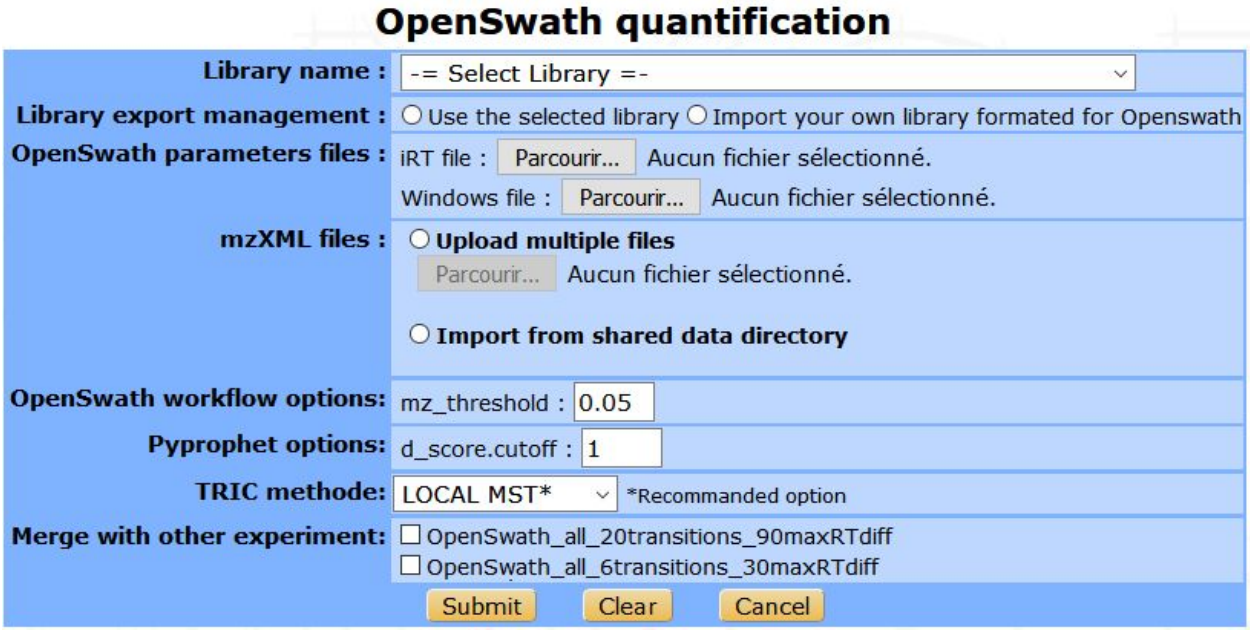
From OpenSWATH

The file from OpenSWATH is required, this is a TSV file generated by the last step of the workflow (TRIC). The library that will be used to analyse the experimental data is also needed as the associated export parameter file. Then, the user must provide the TRIC method used and the number version of OpenSWATH.



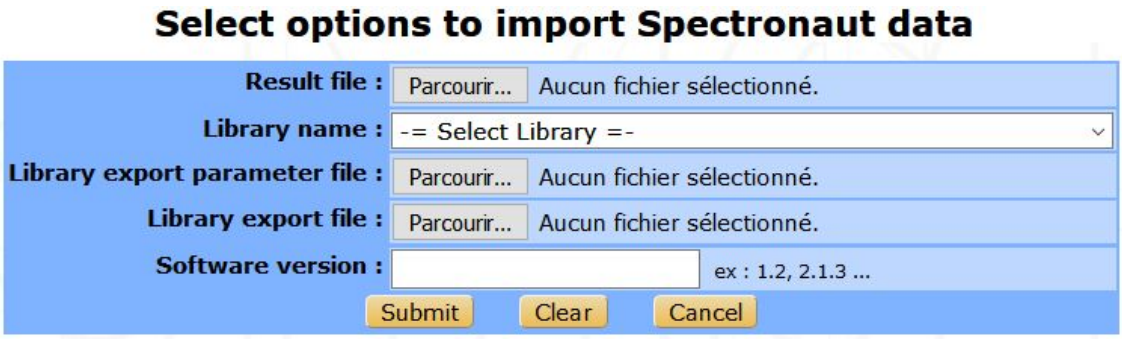
Running OpenSWATH quantification

The OpenSWATH workflow can be launched directly from myProMS. The process will analyse the experimental files and import the result in a same step. Some parameters are required as the library name, the mzXML results files, the iRT file (in TraML format) and the DIA windows file. The user can import his own library converted for OpenSWATH, and choose to merge this result with other existing analysis.



From Spectronaut

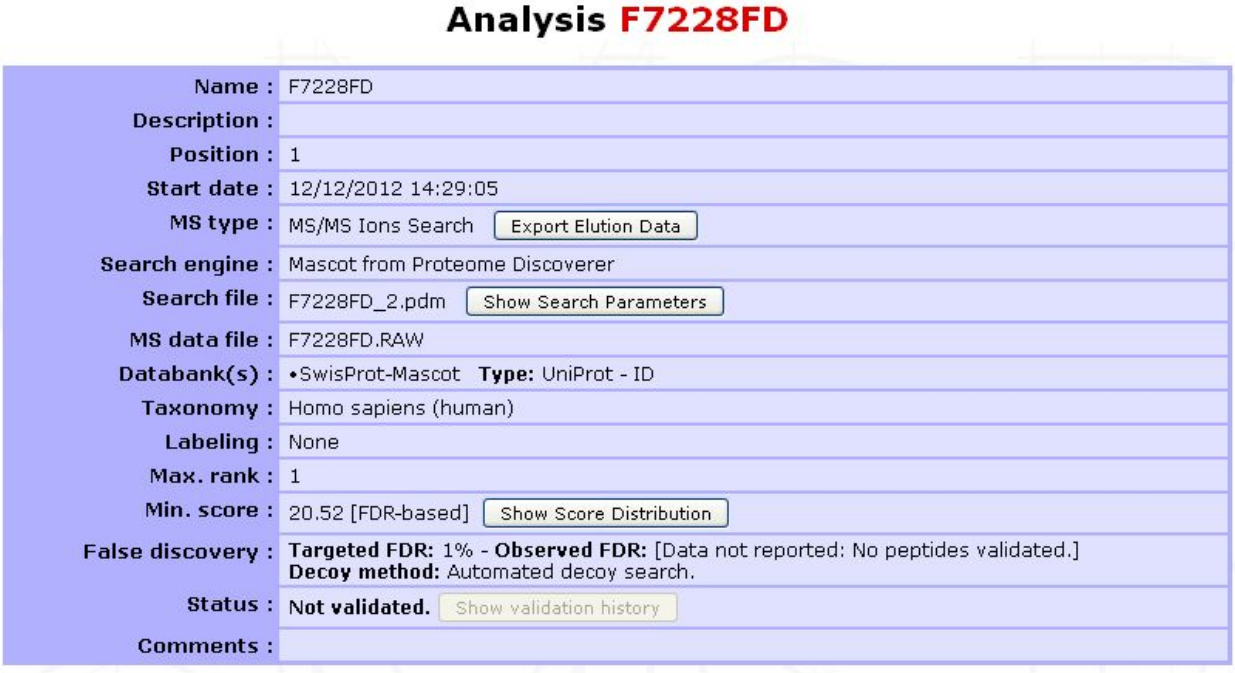
The import Spectronaut data form is similar to OpenSwath's form.



Once the process is over, samples and analyses will be added to the selected Experiment. Peptides and proteins will be automatically validated. Transition quantification data will be imported but not the peptide or the protein quantification results as a dedicated pipeline is available in myProMS to perform this task (see the **Protein quantification** chapter below for more information).

Analysis summary

Files will be imported sequentially and the Analyses (and possibly Samples) newly generated will appear in the top left navigation window. The Analysis summary is shown below:



Diverse information is available to the user such as the MS type, search engine, databank(s) (selected in myProMS), protein identifier type, taxonomy, labeling method if any, threshold score value and strategy (FDR-based or user-defined), validation status,... More search parameters and score distribution for FDR computation can be displayed on demand.

Analysis validation

Search result data associated with an Analysis must be validated before being accessible by end-user for further interpretation. Data processing of Analysis data is a multi-step process. Analyses icons are color-coded based on their data processing status to help users to easily determine the validation level of each of them:

-  : Protein annotation import not yet completed.
-  : Data not yet validated.
-  : Data partially validated.
-  : Data validated and reported.

Note: Protein annotation import from the databank file(s) is not part of the validation process. It is triggered as a background task immediately after search result data import. This process can take several minutes depending on the number of proteins identified. However, validated data cannot be reported (see the Reporting section below for more details) before protein annotation data import has been completed.

Automated peptide/protein validation

FDR (False discovery rate) - based validation

FDR-based validation is initiated at the data import step to filter out data below the corresponding threshold score (see **Importing Analyses** above).

Qualitative validation

Comparative validation

Validation templates

Manual peptide/protein validation

Peptides selection/exclusion

Protein exclusion and filtering

Lower-scoring peptides activation

Clear peptide/protein selections

Validation traceability

Sequence modification validation

Phosphorylation sites validation with PhosphoRS

PhosphoRS (T. Taus et al., J. Proteome Res., 2011) is an algorithm used to determine phosphorylation positions on a peptide, based on its sequence and MS2 fragmentation spectrum. This tool is included into myProMS and can be used to correct imported phosphorylation data.
From "**Process analyses**" menu, select "**Peptide/Protein Selection**" process type, and click on "**Start PhosphoRS analysis**". The following form is then displayed:

<Figure phosphoRS form>

PhosphoRS parameters can be set in "**PhosphoRS Analysis Rules**" section:
- :
- **Activation type**: select the fragmentation mechanism of the analysis instrument. PhosphoRS is optimized for each activation type.
- **Mass Deviation**: mass error tolerance when PhosphoRS matches experimental spectra with theoretical spectra.

Manual validation of modifications

Validation traceability

Reporting

Validated proteins

Match groups and protein visibility

Match groups are a **key feature** in myProMS. Understanding and setting properly the rules that control match group organization and protein visibility is essential for accurate data analysis and interpretation.

Because the shotgun mass spectrometry technique (used to generate all Analysis data) allows to identify peptides and not protein, multiple proteins can be matched with the same (set of) peptides. This creates an inherent ambiguity on the identity of the proteins contained in the extract analysed. myProMS deals with this problem by organizing the proteins identified in **match groups** representing groups of proteins sharing the same (sub-)set of peptides. To avoid protein inflation, by default only **1 (top) protein** per match group is made **visible** and all other will be **hidden** (not considered as identified in the sample studied). The top protein is also used as **alias** for the match group. Only visible proteins are considered as present in the sample analyzed. Hidden proteins will not be listed (unless the user specifies otherwise) and not used in all subsequent data processing such as protein quantification or Gene Ontology analyses.

Top protein selection rules

For each match group, myProMS attempts to set the protein most likely identified by the corresponding set of peptides as top protein using the following rules sequentially until only 1 protein remains:

1. the protein is matched by all peptides in the set.
2. *the protein is the most often found as top proteins in previously validated analyses.
3. the best scoring protein among those meeting the previous criteria.
4. the protein with best sequence coverage among those meeting the previous criteria.
5. the best annotated protein among those meeting the previous criteria. Annotation quality is estimated as follows:
    i. SwissProt identifier.
    ii. trEMBL identifier.
    iii. none of the following keywords found in the protein description.
    iv. the *hypothetical* keyword is found in the protein description.
    v. the *unknown* or *unnamed* keywords are found in the protein description.
    vi. the protein description is missing.
6. the shortest protein among those meeting the previous criteria.
7. the protein with identifier first in alphabetic order.

* myProMS will preferentially select a protein that has been identified often in previous samples.

Project-wide protein visibility rules

By default, only the top protein of each match group is visible. However, the user can use 1 of 3 predefined project-wide protein visibility rules to alter this default behavior. **Edit** the

corresponding project. The following section is part of the edition form:

<Figure protein visibility rules in project edition form>

These rules are self explanatory and ordered by decreasing stringency. Selecting rule #2 or #3 will alter myProMS default behavior and can potentially lead to multiple visible proteins per match group. Click on **Save** to validate your changes. Visibility of all identified proteins will re-evaluated based on the rule selected except if involved in quantifications or GO analyses. Keep also in mind that **protein lists** and saved **comparisons** (see corresponding chapters below) can be modified by the resulting changes in protein visibility.

Checking for conflicting match groups

It is possible to check for inconsistency in match groups across multiple analyses; meaning to detect proteins with inconsistent visibility (**visible** vs **hidden**) across multiple analyses.
From the *Summary* view of any project item containing at least 2 validated analyses, click on the "*Scan for conflicts*" button right of the number of visible/total proteins validated. A list of such proteins (if any) will be displayed with the number of analyses where each protein is found visible or hidden as shown in figure below.

<Figure list conflicts>

Click on the *[+]* icon to display the list of the analyses involved. From this list, you can either:
- Edit a specific match group by selecting an analysis (radio list) and clicking on the "*Edit match group*" button at the top of the table (see the **Manual edition** paragraph below for help).
- Display detailed information on a protein in the context of the analysis of your choice by directly clicking on that analysis' name.

Displaying match group composition

List Proteins at Analysis-level (see the **Project lists** chapter below for more information) and check *Show Match groups* at the top of the list. Click on the *[+]* icon next to the alias protein identifier to display the content of the match group. No *[+]* icon indicates that the protein is alone in its group. As shown in the screen capture below, visible proteins are listed in bold while hidden ones appear in light font.

<Figure match group in protein list>

Manual edition

Aside from project-wide protein visibility rules, any match group can be manually edited to modify the top protein selection as well as the visibility of any protein in the group. Modification is allowed only if corresponding analysis is not associated with protein quantifications nor GO analyses. Click on the "*Edit Match Group*" button at the bottom of the match group list to enter editing mode.

<Figure edit Match group part 1)>

In part 1) of the form, the top (alias) protein can be changed and any other protein can be made visible meaning that you believe they are indeed present in the biological sample analysed.

<Figure edit Match group part 2) & 3)>

In part 2) of the form, you can propagate the changes made in part 1) upward in the project tree. You can chose to propagate independently the alias, visible and hidden protein selection.
Finally - part 3) - you must decide whether your changes can contradict or not the current **project-wide protein visibility rule**. If this option is unchecked, any changes made that do not agree with the project-wide rule will be **ignored**.
Click on *Proceed* to validate your changes.

Peptide distribution in match group

It is possible to visualize graphically the set of peptides of the match group and where they match each protein sequence. To do so, click on the "*Peptide Distribution*" button at the bottom of the match group list.

<Figure Peptide distribution in MG>

See **Peptide Distribution** in the **Single protein view** section below for more information on how to interpret the data displayed.

Identifier mapping

Project-based protein lists

User-defined protein lists

Exporting a protein list

Protein list comparison

Full protein-level comparison

Pair-wise protein-level comparison

Pair-wise peptide-level comparison

Saving a comparison

Search for proteins

Single protein view

Peptide quantification

Peptide quantification is a necessary step for peptide-based protein quantification; whether the quantification is based on MS-spectra (SILAC, TnPQ, XIC-based label-free quantification,...) or on MS/MS fragments (iTRAQ, TMT, SWATH...).

Data import from search results file

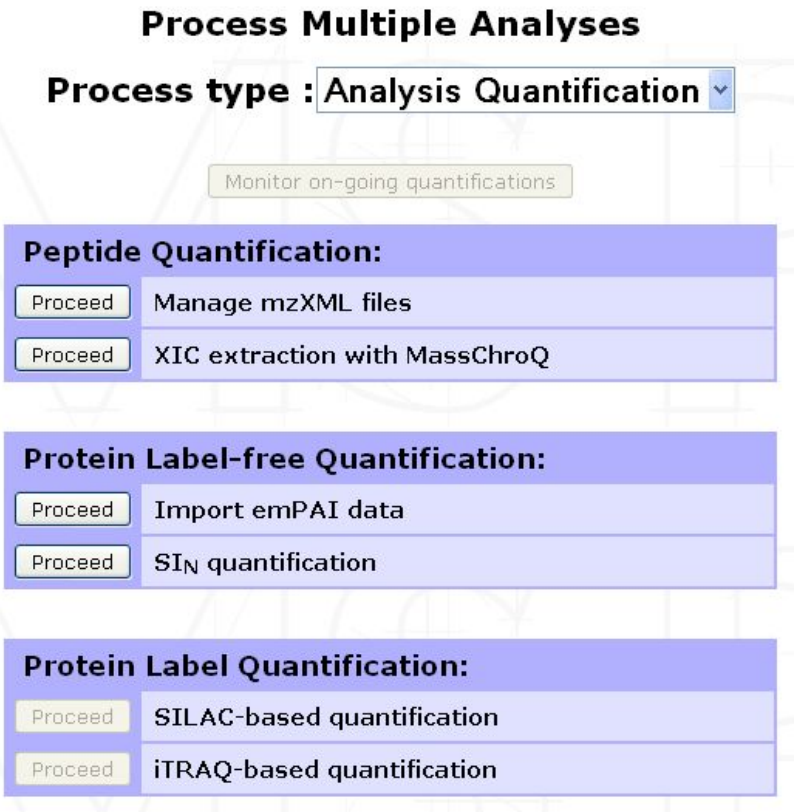
Some search result files already contain peptide quantification data. It is always the case for MS/MS fragment-based quantification such as iTRAQ for which the peptide ion intensities are part of the MS/MS spectrum data. Some search results files (Proteome Discoverer MSF or MaxQuant) may also contain peptide XIC data if a quantification was performed after the search process.
When peptide quantification data are contained in the imported search results file(s) myProMS will automatically import these data either during search data import or at the "**Validated data Report**" step if data validation must be performed. Only quantification data related to validated peptide will be kept (see the Virtual peptides section below for **important** additional information).

Data extraction from LC/MS file with MassChroQ

XIC-based peptide quantification can be performed within myProMS whether or not peptide quantification data were already available in the search results file. myProMS uses the tool MassChroQ (Valot B et *al*, Proteomics, 2011) to perform this task. However, the corresponding LC/MS file(s) must be provided in **mzXML** format.

Managing mzXML files

To manage the list of mzXML files available within a given project, select any **Experiment** or **Sample** from the project navigation window and click on the *Process Analyses* button in the option frame. From the selection menu displayed, select *Analysis Quantification* to display the list of available options. Click on *Proceed* next to the **Manage mzXML files** process as shown below.

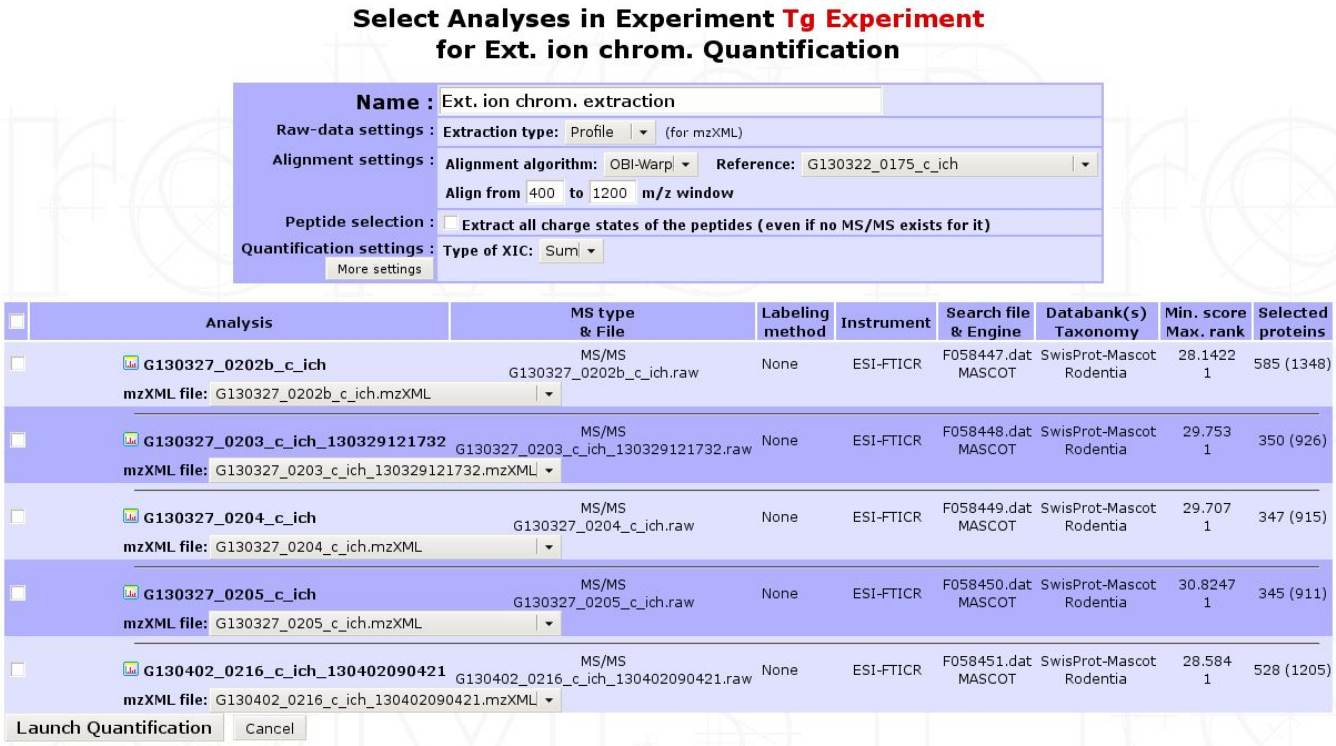


The following form will be displayed to allow you to either import a new mzXML file or delete already imported ones.
fdf

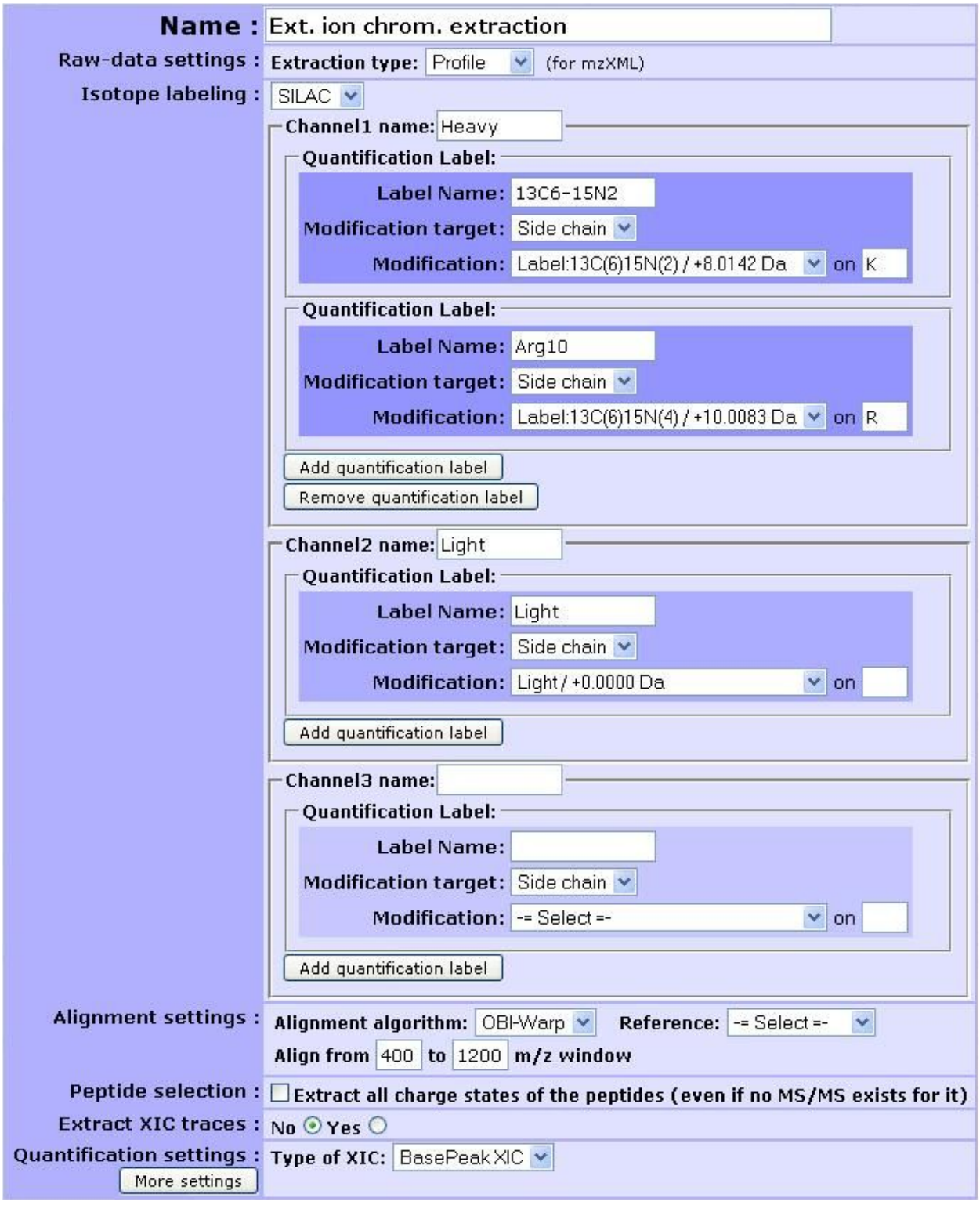
Note: We perform LC/MS files (RAW & WIFF formats) conversion to mzXML with ProteoWizard tool using default settings. Other format conversion tools were not tested. We also recommend not to change the files name (except for the mzXML extension) to ease Analysis/mzXLM file matching in the quantification launch step.

Running XIC extraction

Go to the *Analysis Quantification* options (as shown above) and click on *Proceed* next to the **XIC extraction with MassChroQ** process to display the form shown below.



In the first part of the form, multiple parameters can be set for the extraction:
-Name of the quantification: All extraction data collected will be regrouped in a single quantification carrying this name.
-Extraction type: Profile or centroid
-Isotope labeling: If isotope labeling was performed on your sample, it is possible to use XIC extraction to retrieve it. To do so, you need to choose 'SILAC'. Up to 3 different channels can be retrieved at a time (e.g. heavy, light and medium) that have to be named. For each channel, one or more **quantification label** can be added given the experimental design. Each **quantification label** is linked to a post-translational modification that explains it. Specify the **modification target** on which it occurs (*side chain*, *n-ter* or *c-ter*). If *side chain* is chosen, don't forget to give the residue where the post-translational modification occurs.
Here is provided an example of the filled form in a SILAC experiment where lysine and arginine residues were designed as heavy isotope. **13C6-15N4** was renamed to **Arg10** for clarity.
NB : it is really important at this stage to define a *light* channel if a biological experiment/condition/analysis match the light version. Otherwise, the *light* version of the peptide will not be retrieved in the end.



Note : your modification is not selectable in the **modification target** option ? Check the status of the modification (see **Sequence modification** section below). Maybe the modification you are using is not tagged as label. If not, change this and save it by editing the modification.

-Alignment settings: Multiple LC/MS runs can be quantified at once. MassChroQ can align all

runs to match features across different runs. User must provide an **alignment algorithm** (OBI-Warp or ms2), a **reference** run by selecting the corresponding analysis and an **m/z window** (for OBI-Warp algorithm).
-Peptide selection: Whether to extract or not all charge states of a given peptide.
-Type of XIC extraction to be performed: basePeak area (most intense peak in the range of masses) or TIC area (summed intensity across the range of masses).
-More settings are also available by clicking on the corresponding button.
Finally, click on *Launch Quantification* to start the extraction. A pop-up window will appear to allow you to monitor the quantification progress. XIC extraction is a long process that can last up to an hour or more depending on the number of Analyses to be aligned, the complexity of the LC/MS run and the computer power available. You can continue using myProMS in the mean time and even launch other quantifications. All on-going quantification jobs are displayed in the **monitor quantifications** window (see figure below). As new jobs are launched or old ones completed, they will appear or disappear from the list. Additionally, if an error occurs during quantification, a message will appear for the corresponding job. The user will be able to display the content of the error message and delete the failed job and all associated temporary data.

<Figure Monitor Quantification window>

If this window is closed inadvertently (closing it has no effects on the on-going jobs) or did not appear (pop-up windows for myProMS URL must be enabled in your browser), it can be displayed again by clicking on the *Monitor on-going quantifications* button in the *Analysis Quantification* options (Process Analyses > Analysis Quantification).

Note 1: Check the MassChroQ manual for help on setting these parameters properly.
In the second part of the form, you must select the Analyses for an extraction will be performed and associated the it to the proper mzXML file. If the mzXML file name matches the **MS data file** recorded for the **Analysis**, myProMS will do the job for you.

Note 2: Only runs with reproducible retention-time values (e.g biological or technical replicates) should be selected for alignment. Runs potential very different set of feature (e.g.sample fractions separated on a gel) should be extracted separately.

Virtual peptides and proteins

During the quantification process, intensities of parent ions can be calculated even though the corresponding peptide did not end up in the list of peptides validated. For instance, in the case of a SILAC-labeled analysis, the label-free form of a peptide can be validated but not its labeled counterpart; either because the later falls under the threshold score used or was not identified at all. However, these data are valuable for the quantification since both peptide forms are required for ratio calculation. myProMS solves this issue by adding these missing peptides to the list of validated peptides but with a special status: **virtual peptides**. This strategy of peptide addition also applies to alternative charge states of a given peptides (e.g. if the 2+ charge state

of a peptide is validated, all other quantified charge states will be added as virtual peptides). Virtual peptides remain hidden unless their presence is required for proper data interpretation.
When applied to a label-free quantification where 2 or more analyses are aligned, a peptide validated in the reference Analysis but missing (or not validated) in an aligned one can be "rescued" as a virtual peptide. If this peptide does not belong to any validated proteins of the aligned Analysis, the protein(s) matching this peptide in the reference will be added to the Analysis aligned as **virtual protein(s)**. Virtual proteins appear in italics in most protein lists.

Displaying peptide quantification data

Once the peptide quantification data are available (after a **Report** for Search file extraction or an **XIC extraction** within myProMS), they can be displayed for individual **Analysis** by selecting the corresponding Analysis in the Project navigation frame and clicking on the *Internal Quantifications* button in the option frame. From new window displayed in the result frame, select the name of the quantification (in the Peptide quantification bloc). A window similar to the one below will be displayed showing a summary of the quantification parameters used (if any: no parameters are displayed in case of a direct extraction from a search results file) and a list of proteins with identified peptides and corresponding XICs or fragments area for DIA extraction. In case of labeled quantification, peptide sets (label isoforms) are grouped into a single peptide row. The peptide set sequence, variable modification, position, charge, score(s) and XIC(s) are displayed.

Single-Analysis Peptide Quantification

Select: SILAC 2plex (Arg10, Lys6) (Custom) [SILAC] Edit Delete

Label :	SILAC	
Channel :	1	2
Signal name :	- MMS	+MMS
Isotope(s) :	None	Lys6: Label:13C(6) (K)
Total signal :	4.13e+09	2.93e+09

TIM44_YEAST: RecName: Full=Mitochondrial import inner membrane translocase subunit TIM44; AltName: Full=Inner membrane import site protein 45; Short=ISP45; AltName: Full=Membrane import machinery protein MIM44; AltName: Full=Mitochondrial protein import protein 1 *Saccharomyces cerevisiae* (431 aa)

#	Peptide sets	Start	Charge	Scores	- MMS	+MMS
1	AQRGSTIVGK	98	2^+	34.95/-	3318730	2113160
2	EVSEVIDDGESSRYGGFITK	158	2^+	50.71/44.31	10081900	6240680
3	EVSEVIDDGESSRYGGFITK	158	3^+	-/-	3749330	2450730
4	KLDESFEPVRQTK	142	2^+	51.11/56.01	32805400	19407800
5	KLDESFEPVRQTK	142	3^+	-/-	80268600	48058300
6	KTGETMEHIATK	111	2^+	78.48/63.09	3106750	1802680
7	KTGETMEHIATK + Oxidation (M:6)	111	2^+	61.28/56.6	1453800	820991
8	KVEDFKEK	221	2^+	39.32/-	6068770	2933450
9	LGESEAYKK	82	2^+	45.56/51.64	13590500	8835730
10	LWDESENPLIVVMRK + Oxidation (M:13)	242	2^+	32.59/-	1222880	851545
11	LWDESENPLIVVMRK + Oxidation (M:13)	242	3^+	-/-	6090870	3179090
12	SNEDAGTAVVATNIESK	199	2^+	98.55/88.3	89985600	59983200
13	SNEDAGTAVVATNIESKESFGK	199	2^+	80.62/64.9	15554300	9468700
14	SQELQENIK	65	2^+	44.17/-	64900700	46250700
15	TGETMEHIATK	112	2^+	56.32/-	2087380	1623040
16	TGETMEHIATK + Oxidation (M:5)	112	2^+	36.41/42.8	2555150	1624440
17	TLQDASGKLGESEAYK	74	2^+	98.2/75.46	7019810	4470020
18	TLQDASGKLGESEAYK	74	3^+	-/-	10041000	6254880
19	TLQDASGKLGESEAYKK	74	2^+	67.4/-	7262500	3228410
20	TLQDASGKLGESEAYKK	74	3^+	-/-	20886900	9840560
21	TVVGRSIQSLK	229	2^+	32.13/34.17	65403200	44655600
22	VGGFFAETESSRVYSQFK	261	2^+	44.78/30.41	1827420	1197130
23	VGGFFAETESSRVYSQFK	261	3^+	-/-	4285890	2825770

Legend: Case of a direct extraction of SILAC-labeled peptide XIC from a search result file

Virtual peptides can be easily identified as they do not have score.

Multi-Analysis Quantification

Select: XIC Tg ms2 [Ext. ion chrom.] Edit Delete

XIC quantification Name :	XIC Tg ms2 Export Results
Raw-data settings :	**Extraction type:** profile (for mzXML)
Alignment settings :	**Alignment algorithm:** ms2 **Reference:** G130322_0175_c_ich **Tendency:** 10 - **Smoothing:** 5 **(MS/MS) and** 3 **(MS)**
Charge states :	Validated charge states extracted
Quantification settings :	**Type of XIC:** sum
More settings	

Legend: Case of a MassChroQ extraction with alignment of multiple Analyses

Protein quantification

Absolute abundance quantification

emPAI (label-free)

The **E**xponentially **M**odified **P**rotein **A**bundance **I**ndex (emPAI) is a spectral-count method that estimates the relative quantitation of proteins in a complex mixture (Hishima et *al*, Mol Cell Proteomics, 2005) based on protein coverage by peptide matches. myProMS uses the built-in implementation of the Mascot server 2.3 software which is a slightly modified version of the original emPAI value (for more details, have a look to mascot help). As this value is retrieved from Mascot web-server, this label-free method can only be applied to Analyses generated from Mascot DAT files directly imported from a connected Mascot server.

...

SIn (label-free)

The **S**pectral **I**ndex **N**ormalized (SIn) is a normalized label-free quantitative method which combines three abundance features : peptide and spectral count with fragment-ion (MS/MS) intensity (for more details, see Griffin NM et *al*, Nat Biotechnol., 2009). This label-free method is currently available only for Analyses generated from Mascot DAT files. Support for other search results formats is plan in future versions of myProMS.

…

MaxQuant: Intensity, LFQ, iBAQ

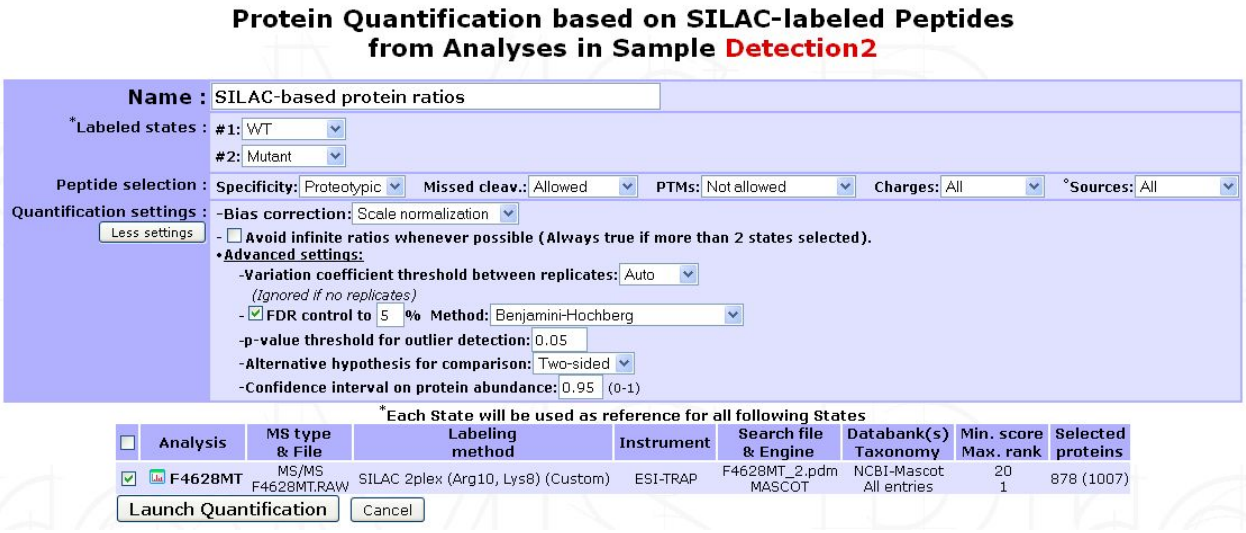
Displaying single abundance quantification data

...

Relative abundance quantification

Single-Analysis quantification (labeled)

If a labeled Analysis has to be quantified, labeling parameters and all peptide XIC data should be readily available in the corresponding search results file. Therefore a straightforward protein quantification can be performed as follow: Go to the *Analysis Quantification* options (Process Analyses > Analysis Quantification) and click on *Proceed* next to the **(SILAC/iTRAQ)-based quantification** process to display the quantification form shown below.



Name: A name for the quantification.

Labeled states: Select the different conditions to be compared. Available labeled states are identified based on labeling design extracted from the search result file. Each condition defined will be used as a reference for the following one(s). 1 state is usually associated with 1 condition. However, if more than 2 states are identified (e.g. iTRAQ 4/8-plex) an additional option will be displayed for grouping different states as replicates of the same condition. In addition, if more than 2 conditions are defined, all corresponding ratios will be calculated except reverse ratios (cond B/cond A but not cond A/cond B).

Note: It is possible to quantify multiple Analyses at once. Make sure they share identical labeling design. If not, they should be quantified separately.

Multiple filter can be apply on **Peptide selection**:

- **Specificity:** Whether to restrict quantification to proteotypic peptides or not.
- **Missed cleav.:** Include or not miss-cleaved peptides.
- **PTMs:** Peptides with sequence modification can be allowed, not allowed or extend exclusion to corresponding non-modified peptide.
- **Charges:** Include all charge states of a peptide set or restrict to set that gives the best signal (set containing peptide with highest XIC value).
- **Sources:** If the search results files is a merge of multiple LC/MS runs (e.g. Proteome Discoverer), use peptide sets from all runs or use only the one with best signal.

Quantification settings: Additional options are available to control experimental bias, outliers detection and differential analysis.

- **Bias correction:** Select whether to correct or not for signal bias between label states and which method to apply: If **Scale normalization** is selected, the assumption is made that the total XIC signal between all states should be equal. Alternatively if **Reference protein(s)** is selected, a pre-recorded **List** of proteins must be provided. When using this option, it is assumed that a subset of proteins (e.g. House keeping proteins) is unchanged amongst all states and therefore only the sums of the XICs matching these proteins are set equal. In both cases, a state-specific correction factor is computed and applied to each individual peptide XIC.

- **Avoid infinite ratios:** Infinite ratios (log values) can occur when XIC values are missing in 1 of the 2 conditions being compared. When a mixture of normal and infinite peptide ratios exists for the same protein, myProMS must either use the most abundant type of ratios to quantify the protein (e.g. set protein ratio to +/-infinite (log values) if more than 50% of matching peptides have infinite log ratios) or only use the "normal" ratios even if they are less frequent that the infinite ones (to **Avoid infinite ratios whenever possible**). This later option is automatically selected if more than 2 conditions are compared to prevent excessive data exclusion.
- More **advanced settings** can be used for **outlier** detection, comparison hypothesis test (Two-sided/Lesser/Greater), **FDR** control, ...

Finally, **select the Analysi(e)s** to be quantified. If multiple peptide quantification datasets are available for an Analysis, one must be selected. Click on the *Launch Quantification* button. Multiple quantifications will be queued and processed as up to 3 parallel jobs. As described above for Peptide quantification, a popup window will appear with the list of all jobs launched with their progress status.

Design-based quantifications

The use of a design for a quantification is highly recommended, even if it requires only single labeled analysis. It is mandatory to create a design for a quantification that requires more than 1 analysis. Designs are automatically generated when importing protein quantification data from MaxQuant analyses.

Conditions
Observations

Displaying relative abundance quantification data

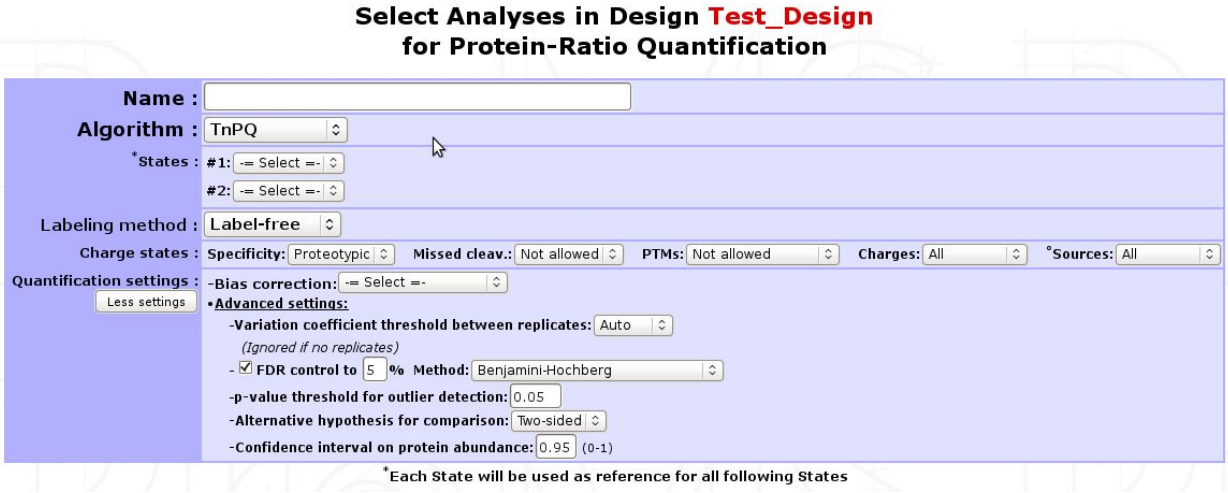
Label-free quantifications

Label-free quantifications are methods that allow to determine the relative amount of proteins in two or more biological samples without any use of stable isotope or chemical tag. It is based on precursor signal intensity or the number of spectra made for each peptide of a protein.
Here is a brief description of several methods available in myProMS that you can use from top panel button **Process Analyses** and then, **Analysis Quantification**.

TnPQ

Silva et *al.* showed in their work on a Q-ToF type instrument that it is possible to quantify unknown protein samples with a known unified signal response factor in absolute manner (Silva et al, Mol Cell Proteomics, 2006). Then, the **T**op **3** **P**rotein **Q**uantification (T3PQ, Grossmann et

al, J Proteomics, 2010) extended this method to ion trap instruments. The method premises that for each protein identified by a set of peptides, the average of the three most efficiently ionized and therefore highest MS signals directly correlated with the input amount of the corresponding protein. In myProMS, we extended this definition to “all available peptides” for a given protein and called it TnPQ.



Steps involved in TnPQ computation:

Step 1 : retrieval of all available XICs (area) of each peptide of the protein for all conditions

Step 2 : removal of incomplete peptide information i.e. peptide with no XIC information in at least one of the replicates of a condition will be removed

Be careful : when creating a quantification, avoid to add too many conditions because you will lose a lot of peptide information given the fact that all conditions must provide a XIC for a peptide to be considered more further

Step 3 : if a bias correction setting was selected (scale or reference protein normalization), a normalization step is introduced by computing bias estimates on unique peptides (for more information in this step, please, go to Yang *et al.* 2002, scale normalization part). All XIC are divided by those bias factors

If **None** was chosen, nothing is done to the data

Step 4 : removal of extreme XIC values (outliers) based on the coefficient of variation (standard deviation divided by the mean) of all identified peptides along the replicates in the conditions.

Step 5 : compute for each protein the geometrical mean of peptide XICs

Step 6 : quality control of the data (normality test on the data and variance sameness)

Step 7 : compute the ratio between paired conditions and make a test to assess equality of mean depending on the design made before

→ for 2 conditions : use Student *t-test* comparison (or Welch *t-test* if variance are not the same)

→ for more than 2 conditions : use Tuckey HSD (honestly significant difference) test

Step 7’ : if chosen, adjust p-values to control FDR level

Gene Ontology analyses

Different types of analyses using **Gene Ontology (GO)** can be performed on validated protein lists. The GO project provides a controlled vocabulary of terms for describing gene products such as proteins. For more details, see the GO website. A GO analysis can regroup proteins into standardized categories of terms belonging to 3 domains: **Biological Process**, **Cellular Component** and **Molecular Function**. In myProMS, all GO analyses need 2 types of GO files that are managed from **GO files management** section (See corresponding chapter below for more information).

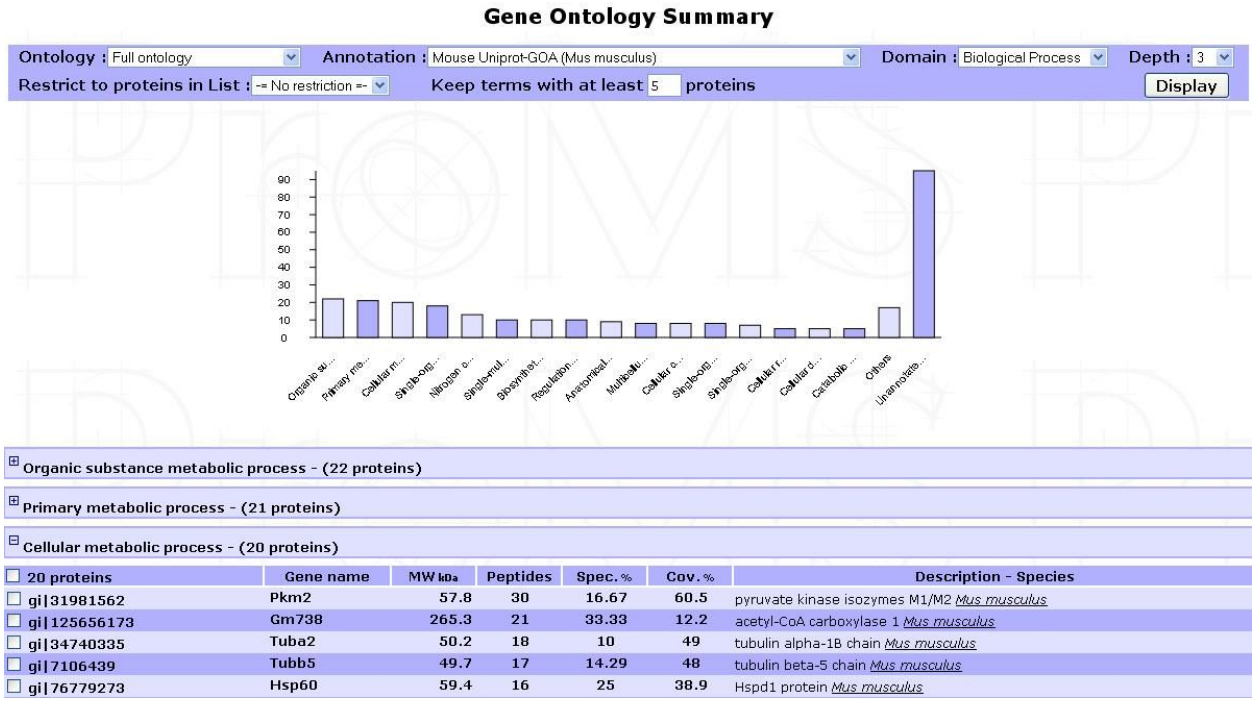
- **Ontology file**: the file that contains all term descriptions and their relationships between each other
- **Annotation file**: the file that maps each protein identifier to the most specific terms that characterize the protein.

GO summary

The GO summary tool can be used to simply regroup proteins sharing common GO terms.
This tool can be run from the **option frame** on any **project item**, by clicking on the "*Gene Ontology summary*" button. The following form is then displayed:

- **Ontology**: the file containing terms that will be used to regroup proteins.
- **Annotation**: the file containing protein annotations to GO terms.
- **Domain**: select one of the 3 GO domains the analysis will be focused on.
- **Depth**: only terms at the specified depth in the GO graph structure will be used. Depth is calculated by counting the distance between a term and the root term of the corresponding ontology domain. If a high depth is selected, a very large number of terms will be displayed and the results may be difficult to read.
- **Minimal protein per term**: if a selected term contains less proteins that this specified value, this term will be ignored and the matching proteins will be added to the "**Other**" category. This parameter is optional.

Click on *Display* to launch the process. After a short calculation time, results are displayed as shown in the example below:



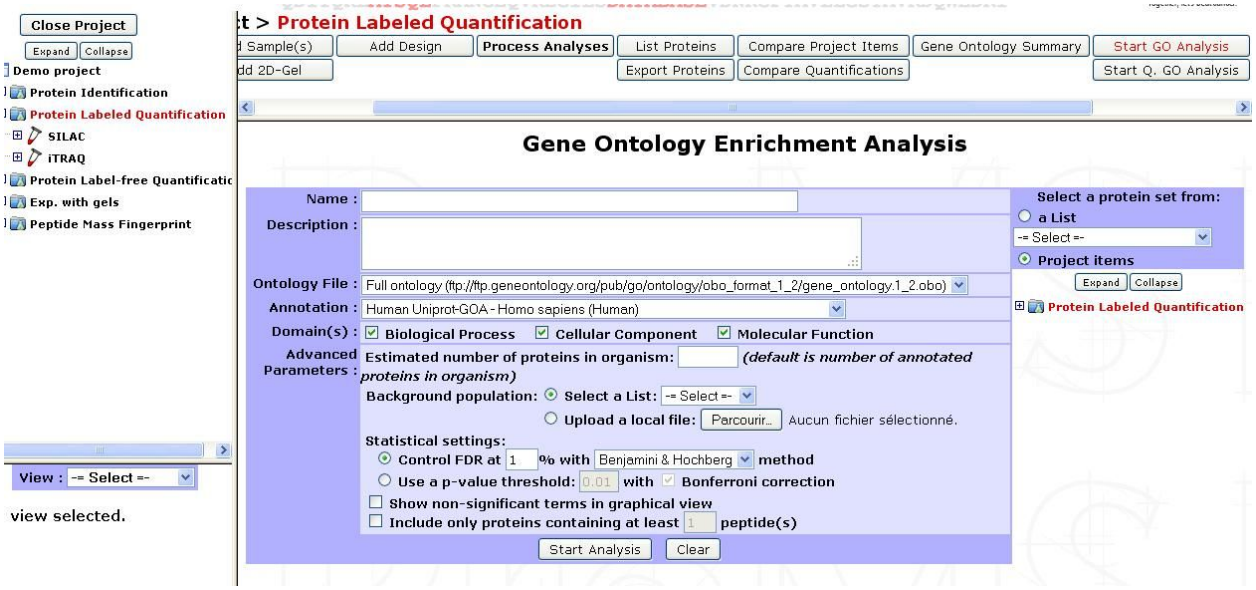
An interactive bar plot shows each term frequency. Click on a bar to display the proteins mapped to the corresponding term. Each protein group can also be viewed by browsing the list of terms displayed below the plot.

GO enrichment analysis

Enrichment analysis is performed to determine which GO terms are significantly enriched in a **tested set** of proteins when compared with a given **background set** (eg. the whole proteome of the species studied). All terms will be tested regardless of their depth.

In myProMS, GO enrichment analysis is calculated with the GO::TermFinder package developed for perl (Boyle *et al.*, Bioinformatics, 2004). Briefly, a *P*-Value using a hypergeometric distribution is computed to determine whether any GO terms annotate a specified list of proteins at a frequency greater than that would be expected by chance. Multiple hypothesis correction is available with FDR computing.

This tool is accessible by a clicking on an experiment and selecting the "*Start GO Analysis*" button in the option frame. The following form in then displayed:



- **Name**: Provide a name for the enrichment analysis. The analysis is saved and can be retrieved by this name in the **GO analyses** tree displayed in the **sub-navigation** frame.
- **Description**: optional description of the current analysis.
- **Ontology File**: the file containing term relationships.
- **Annotation**: the file containing protein annotations to GO terms.
- **Domain(s)**: Select one or more domains to test.
- **Advanced parameters**:

    - **Estimated number of proteins in organism**: If the background population consists of the whole proteome (more exactly the whole protein set contained in the annotation file), this value can be set to calculate properly the enrichment ratio of GO terms in the tested protein set(s), supposing that the annotation file is incomplete. This option artificially adds unannotated proteins to the background.

    - **Background population**: Select the population to which the tested protein set will be compared. A previously built **custom list** can be selected, or a local file can be used instead. This file must contains all protein identifiers that compose the background (1 identifier per row). These identifiers must match the ones contained in the annotation file. If selected background is set to "**Unspecified**", the whole protein set contained in the annotation file will be used as background. In this case, be sure that the annotation file contains only proteins from the current species. This can be considered as a whole proteome background if the annotation has a very good coverage of current species proteome. The background population selection strongly affects the significance of terms and must be chosen carefully and coherently with your biological question.

    - **Statistical settings**: these settings can be set to control the significance cut-off of GO terms. False Discovery Rate (FDR) or p-value criteria can be selected.

    - **Show non-significant terms in graph**: If this option is disabled, non-significant terms will be represented with small dots in graphical view. This can increase significantly the visibility

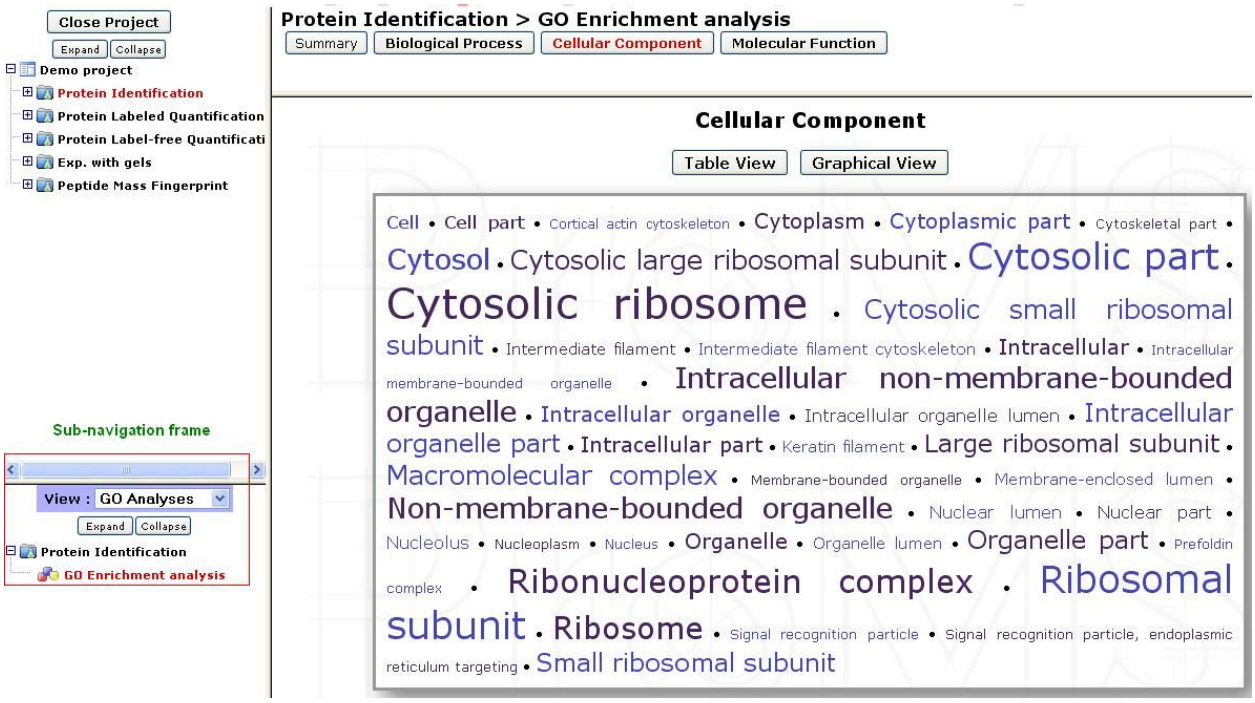
of the graph if the dataset contains a large number of significant terms.

- **Include only proteins with at least n peptide(s)**: Proteins which contains less peptides than the value specified will be excluded from the tested set.

- **Select a protein set**: Select the protein set to be tested. It can be selected from any project item or custom list.

Once all parameters have been set, click on *Start Analysis*. The computation may last several minutes depending on the sizes of the protein sets being compared.

The results are directly displayed after the process but can also be accessed later on by selecting the analysis name in the **GO analyses** tree displayed in the **sub-navigation** frame..



For each domain, results can be displayed in 3 different views accessible at the top of the page:

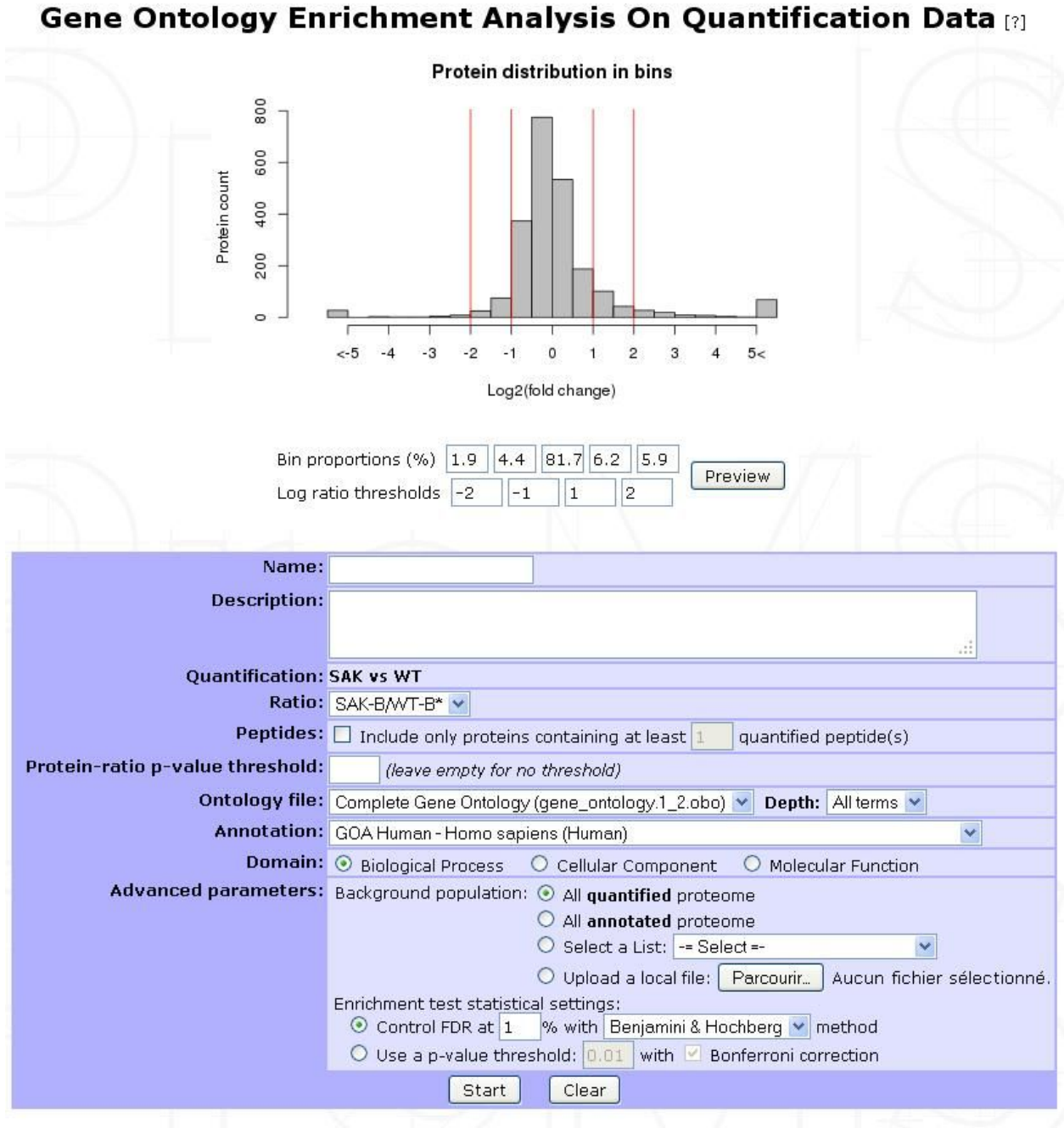
- **Cloud view**: Highly significant terms (low p-value) are represented with a large font, and less significant terms with a small font. The proteins mapped to a term can be listed by clicking and each term.

- **Table view**: More details can be viewed in table format which contains the p-value and enrichment ratio of each term.

- **Graph view**: Displays a graph of the significant terms as **nodes** with their relationships as **edges**. Each node colour is based on the corresponding term's p-value significance. Proteins that are mapped to a term can be viewed by clicking on the corresponding node.

Quantitative gene enrichment analysis

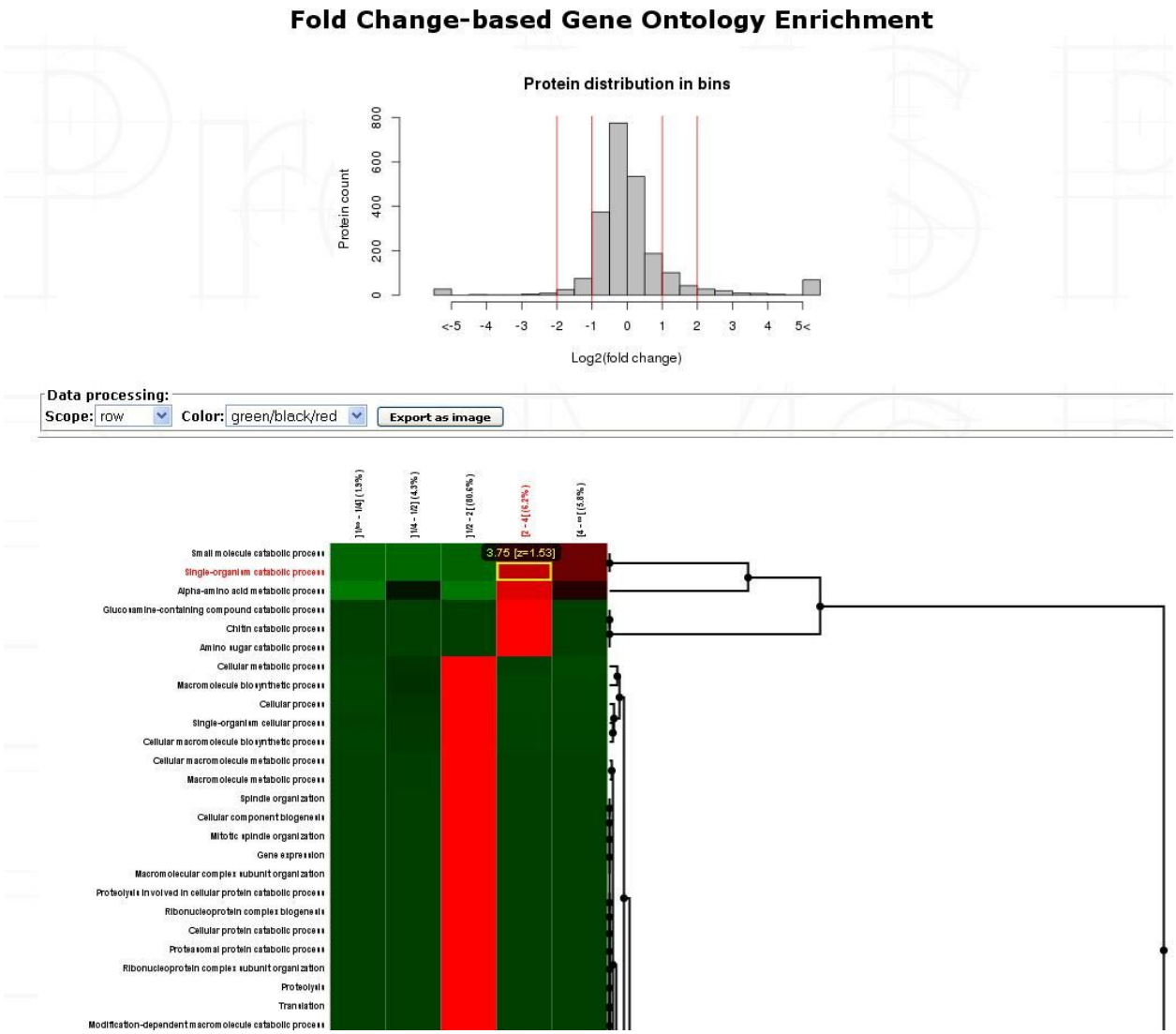
When a quantification is available, a quantitative gene enrichment analysis can be performed as it was originally done for SILAC experiments in an article written by Pan C *et al*, MCP, 2009.
The quantified proteome is divided into five bins corresponding to log2 ratios or bin proportion. Enrichment of GO terms in each bin is then calculated compared to a provided background and a cluster analysis allow to visualize a heatmap of enriched GO-terms in all bins.
Here is how you should proceed to do it.
This option is accessible by clicking on an experiment and selecting the "*Start Q. GO Analysis*" button in the option frame. After loading a protein set of an **Analysis** or a **Design** related quantification, you need to select the parameters in the following form:



- **Name**: Provide a name for the enrichment analysis. The analysis is saved and can be retrieved by this name in the **GO analyses** tree displayed in the **sub-navigation** frame.
- **Description**: Optional description of the current analysis.
- **Ratio**: Choose the ratio considered for the enrichment in the quantitation (like heavy/medium or heavy/light for SILAC experiments).
- **Peptides**: Make a threshold upon the number of peptides used to compute the ratio.
- **Protein-ratio p-value threshold**: Make a selection on the associated p-value of the ratio.
- **Ontology file**: The file containing term relationships.
- **Annotation**: The file containing protein annotations to GO terms.
- **Domain**: Select one domain to test.
- **Advanced parameters**:
    - **Background population**: select the population to which the tested protein set will be compared. See **GO enrichment analysis** section for custom list recommandations.
    - **Enrichment test statistical settings**: these settings can be set to control the significance cut-off of GO terms. False Discovery Rate (FDR) or p-value criteria can be selected.

When the enrichment is done, you can get information of the GO-Analysis by clicking on the sub-navigation frame the item generated and *Summary*.

Click on the *Heatmap* button to see the output you can get:



Each row represent a GO-Term and each cell is the -log10 of the *P*-value of the enrichment test for the GO-Term in the specific bin (put to 1 and then *log-transformed* to 0 if that ontology is not enriched/significant in the bin). Each line is *z-scored*. Then, these *z-scores* are clustered by one-way hierarchical clustering using the function *hclust* in R (the distance function used is 'euclidean' and the agglomeration method used is 'average').

The heatmap is interactive and can be exported as a jpeg image. Clicking on a cell updates the frame and provides the list of proteins containing the annotated GO-Term in the bin.

Exploratory analyses

Launching exploratory analyses

Principal Component Analysis (PCA)

2D-Clustering

Annotation data management

Sequence databanks

The sequence databanks used by the search engines must be referenced in myProMS so that protein annotations (identifier, description, species and sequence) sometimes not present in search result files (eg. Mascot) can be retrieved from the corresponding fasta file during analysis import. A referenced sequence databank is also associated with a specific parse rule that allows myProMS to properly match and extract the annotation from the fasta file.

Databank types

Multiple databank types are available in myProMS depending on the proteomic resource used to download the fasta file:

+ UniProt:

Typical entry in fasta file:

```
>sp|P15311|EZRI_HUMAN Ezrin OS=Homo sapiens (Human) GN=EZR Ezrin
```

- UniProt - ALL

This databank type will extract the entire identifier block `sp|P15311|EZRI_HUMAN` as protein identifier, `Ezrin` as description and `Homo sapiens` as species.

- UniProt - ACC

Same as above except that the protein identifier used is the **Uniprot accession number** `P15311`. This type is also compatible with the Uniprot isoform naming `ACC#-n`.

- UniProt - ID

Same as above except that the protein identifier used is the **Uniprot identifier** `EZRI_HUMAN`.

- SWISSPROT/trEMBL #1, #2 and #3

These 3 types are equivalent of the 3 UniProt types described above except that they recognize the obsolete fasta entry format:

```
>sp|P15311|EZRI_HUMAN Ezrin (p81) (Cytovillin) (Villin-2) - Homo sapiens (Human)
```

+ NCBI:

Typical entry in fasta file:

```
>gi|125987826|sp|P15311|EZRI_HUMAN Ezrin (p81) (Cytovillin) (Villin-2) [Homo sapiens (Human)]
```

- NCBInr - ALL

Extracts `gi|125987826|sp|P15311|EZRI_HUMAN` as protein identifier, `Ezrin (p81) (Cytovillin) (Villin-2)` as descritption and `Homo sapiens` as species.

- NCBI - GI

Same as above except that only the gi number (`gi|125987826)` is kept as protein identifier.

+ *IPI:

Typical entry in fasta file:
`>IPI:IPI00843975.1|SWISS-PROT:P15311|  Tax_Id=9606  Gene_Symbol=EZR Ezrin`

- IPI databank

Extracts `IPI00843975` as protein identifier, `Tax_Id=9606 Gene_Symbol=EZR Ezrin` as description and `9606` as species.
* The IPI resource is no longer maintained. We do not recommend using fasta files from this resource with your MS search engines.

+ Undefined source:

- Protein (User-defined)

This type can be used as a temporary solution for any unknown or custom fasta-compatible entry:
`>pipe_separated_identifier_block any text`
Extracts the identifier block as protein identifier and any following text as description. No species is recorded.

+ Other types:

New databank types can be easily added on demand. Please contact your local myProMS administrator or email to myproms@curie.fr for more information.

Listing databanks

Only bioinformatician and massists/managers with granted appropriate privileges can access and manage the protein databanks.
From **myProMS main window**, select *Annotation data* and follow the *Sequence databanks* link. All active databanks are listed in alphabetic order with a short summary of information as shown in the screen capture below. From this window, you can either **add** a new databank, **edit** or **delete** an existing one.

<Figure of databanks list>

Adding a new databank

From the databank list window, click on the "*Add new Databank*" button at the top or bottom of the list. The following form will be displayed:

<Figure Add new Databank>

Fill out the form to provide information on the databank you want to add. In particular, you must select the databank type so that the server will know how to extract the protein annotation from

the file. Information on the corresponding parse rule is then displayed to help insure the right databank type was selected. You must also provide a **fasta** file containing the protein data. There are multiple ways to do so:

1. Use a databank already referenced by Mascot: myProMS allows you to directly use fasta files stored on the Mascot server to avoid data duplication. In this case, the databank will be automatically synchronized when updated by Mascot.
2. Use a file from a dedicated directory on server (e.g. file was previously uploaded by FTP or the directory is shared between local computer and server).
3. Upload a fasta file from your computer.
4. Download the file from the internet: You must provide an HTTP or FTP link to the file.

For the last 3 options, normal and gzip-/zip-compressed files are handled.
If the databank contains both target and decoy sequences, this must be specified as well as the decoy tag used (eg. REV_).
For the first 2 options (except if a compressed file is used in the 2nd option), it is possible to test the type of annotation rules selected before actually creating the new databank: Select a databank type, the file to be used and click on the *Test rules* button. Annotations from up to 10 entries from the file will be extracted using the selected rules and displayed. Select another set of rules and try again if the extraction did not match your expectations.
If the databank is species-specific, it is recommended to provide the species scientific name even if already specified in the protein entry lines of the fasta file.
Click on the *Save* button to submit the databank creation form. Once the process is completed, you will be redirected to the databank list window.

Editing a databank

You can edit all information concerning an existing databank except its annotation type, the sequence file used and whether it contains decoy sequences.
From the databank list window, click on the *Edit* button on the right side of the databank row. A form similar of that used to add a databank will be displayed. Make the desired changes and click on the *Save* button to validate your changes.
You can test your annotation rules as described above for databank addition but regardless of the databank file origine.
If your databank references a Mascot file, it is possible to check if the file has been updated on the mascot server by clicking on the *Check for update* button. This can take up to a few minutes for large databank files such as NCBI databanks. Checking for file update is not mandatory since it will be performed automatically once the databank is used during an Analysis import.

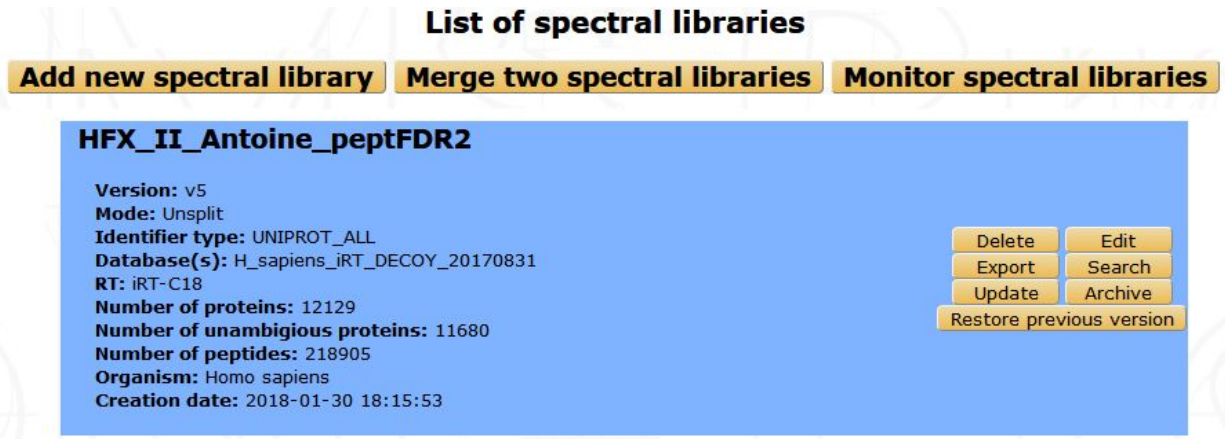
Deleting a databank

We recommend to delete any databank that will no longer be used to keep the list displayed as short as possible. Deletion of a databank as no effect on the traceability information of analyses using this databank. A databank can be deleted at any time except during import of analyses using this databank. From the databank list window, click on the *Delete* button on the right side of the databank row. A prompt will asked you to confirm your decision.

Spectral (SWATH) libraries management

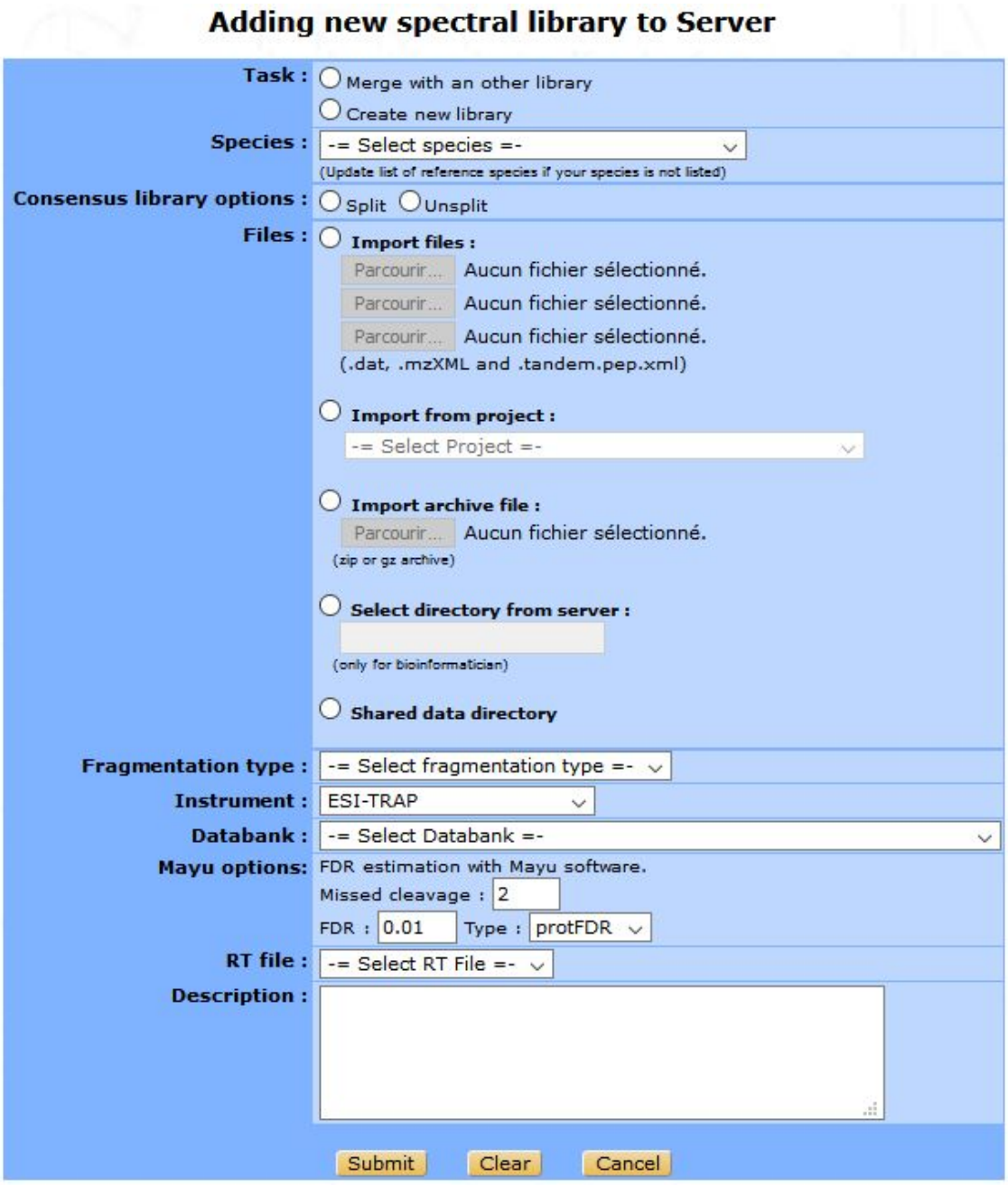
Listing spectral libraries

From **myProMS main window**, select *Annotation data* and follow the *SWATH libraries* link. All available libraries are listed in alphabetic order with few informations as shown in the screenshot below. On the left side are listed all existing libraries with the possibility to delete, export, edit or update them. On the upper part of the window, you can either add a new library, merge two existing ones, or visualize running processes. You can also search if some desired proteins are existing into one library thank to the search link and restore the previous version of an updated library.



Adding a new library

From the libraries list window, click on the *Add new spectral library* button on the top of the list to display the form below.



You need to select the following parameters in the library creation form :

- **Task** : You can create a new library or merge new data files with an existing library (create a new library from an existing one).
- **Library name** : Provide a name for the library.
- **Species** : Select the species scientific name to filter the databank list.
- **Consensus library options** : A consensus library is a spectral library in which MS2 spectrum entries with a redundant peptide sequence assignment have been collapsed into a single entry. Two options are provided for consensus library generation: a simple option that assumes that all fragment ion spectra are correctly assigned (UNSPLIT) and a more sophisticated option that additionally considers retention time when merging

spectra (SPLIT).

- **Files** : Select the DDA data files used to generate the spectral library. Data from 3 search engines can be selected : Mascot files (.dat), X! Tandem files (.xml or .tandem.pep.xml) and Sequest (.xml). For each Mascot, X! Tandem or Sequest file you need to upload the associated mzXML file (with the same name as the Mascot, X! Tandem or Sequest file). You can upload your files from your computer, you can import them from an existing project (only for the .dat files), upload an archive, or select the files in the shared directory.
- **Instrument** : The mass spectrometer used to acquire the data.
- **Databank** : The fasta file used by the search engines (Mascot, X! Tandem and Sequest).
- **Mayu options** : FDR estimation with MAYU. False Discovery Rate (FDR) and number of missed cleavage can be selected.
- **RT file** : The file containing the list of iRT retention time reference peptides.
- **Description** : Optional description of the current library.

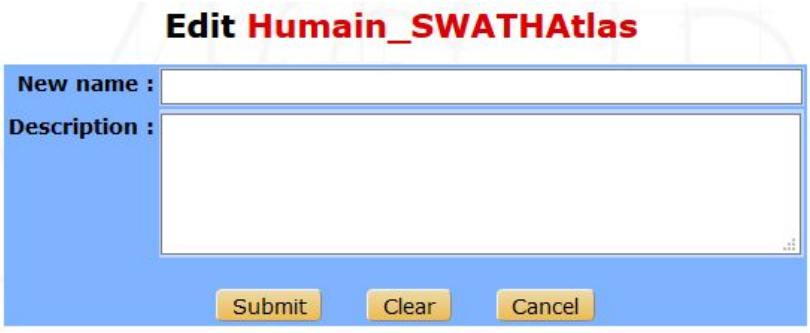
Once the form is filled, click on the *Submit* button to launch the spectral library creation process.

Merging two library

Two libraries can be merged by clicking on the *Merge two libraries* button on the libraries list window. The displayed form requires the names of each of the 2 libraries, the name of the new library and an optional description. Clicking on *Submit* will fuse the selected libraries to create the new library.
Only two libraries with the same iRT file, databank type and consensus library option (SPLIT or UNSPLIT) can be merged.

Editing a library

Note that only the name and the description can be modified. From the libraries list window, click on the *Edit* button on the right side of the library row. The following form will be displayed :



Make the desired changes and click on the *Submit* button to save your changes.

Updating a library

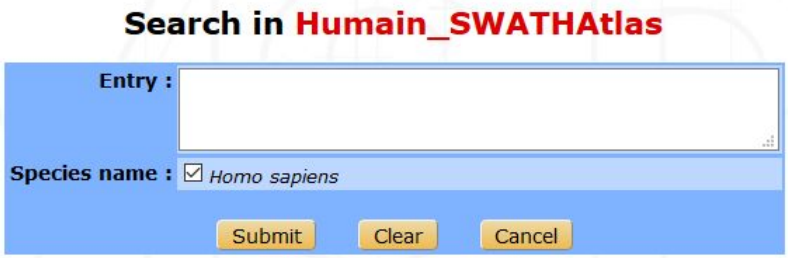
It is also possible to extend a library using another databank-search data from the same organism. From the libraries list window, click on the *Update* button on the right of the library row. A form similar to the library creation one will be displayed. Fill in the parameters and click on the *Submit* button to launch the update process.

Restoring the previous version of a library

An updated library can be downgraded by clicking on the *Restore previous version* button on the right of the library row on the libraries list interface. Every version of a library can be restored by consecutive downgrades.

Searching for proteins in a library

Another available option is to check whether a protein of interest is present in a library and visualize the associated peptides by clicking in the *Search* button of the desired library, on the libraries list window.
Several proteins can be searched at the same time by inserting the accession names, the protein id or the names of the proteins (one per line or separated by either comma or a space character) in the following form.



All the selected terms are searched beforehand in Uniprot, and a list of proteins is displayed. Some information such as the protein name, id, accession number, length and corresponding gene names are shown. The number of associated peptides identified is also indicated.

Results for "histone"

Protein ID (AC)	Gene Names	Protein Names	AA	# Peptides
HDAC1_HUMAN (Q13547)	**HDAC1**, RPD3L1	Histone deacetylase 1 (HD1) (EC 3.5.1.98)	482	**42**
P53_HUMAN (P04637)	**TP53**, P53	Cellular tumor antigen p53 (Antigen NY-CO-13) (Phosphoprotein p53) (Tumor suppressor p53)	393	**13**
KAT5_HUMAN (Q92993)		Tat-interactive protein) (Tip60) (Histone acetyltransferase HTATIP) (HIV-1 Tat interactive protein) (Lysine acetyltransferase 5) (cPLA(2)-interacting protein)	513	**3**

The peptide list and the protein's sequence can be displayed by clicking on the number in the #

Peptides column.

Hide **Peptide list for P04637**

#	Sequence	Modifications	Position	M/Z	Charge	iRT time	Specificity (%)	Found with
1	TYQGSYGFR	-	102-110	539.7513	2+	7.7	100	P04637
2	LGFLHSGTAK	-	111-120	515.7876	2+	-7.2	100	P04637
3	SVTCTYSPALNK	Carbamidomethyl (C:4)	121-132	670.8294	2+	6.4	100	P04637
4	TCPVQLWVDSTPPPGTR	Carbamidomethyl (C:2)	140-156	955.9751	2+	65.8	100	P04637
5	QSQHMTEVVR	-	165-174	607.8010	2+	-20	100	P04637
6	CSDSDGLAPPQHLIR	Carbamidomethyl (C:1)	182-196	833.4043	2+	24.1	100	P04637
7	CSDSDGLAPPQHLIR	Carbamidomethyl (C:1)	182-196	555.9386	3+	25.6	100	P04637
8	RPILTIITLEDSSGNLLGR	-	249-267	690.0635	3+	95.2	100	P04637
9	RTEEENLR	-	283-290	523.7649	2+	-31.3	100	P04637
10	KGEPHHELPPGSTK	-	292-305	505.2634	3+	-35.2	100	P04637
11	ALPNNTSSSPQPK	-	307-319	670.8439	2+	-18.7	100	P04637
12	KKPLDGEYFTLQIR	-	320-333	569.9858	3+	47.4	100	P04637
13	ELNEALELK	-	343-351	529.7900	2+	28.6	100	P04637

Hide **Detailed sequence coverage for P04637**

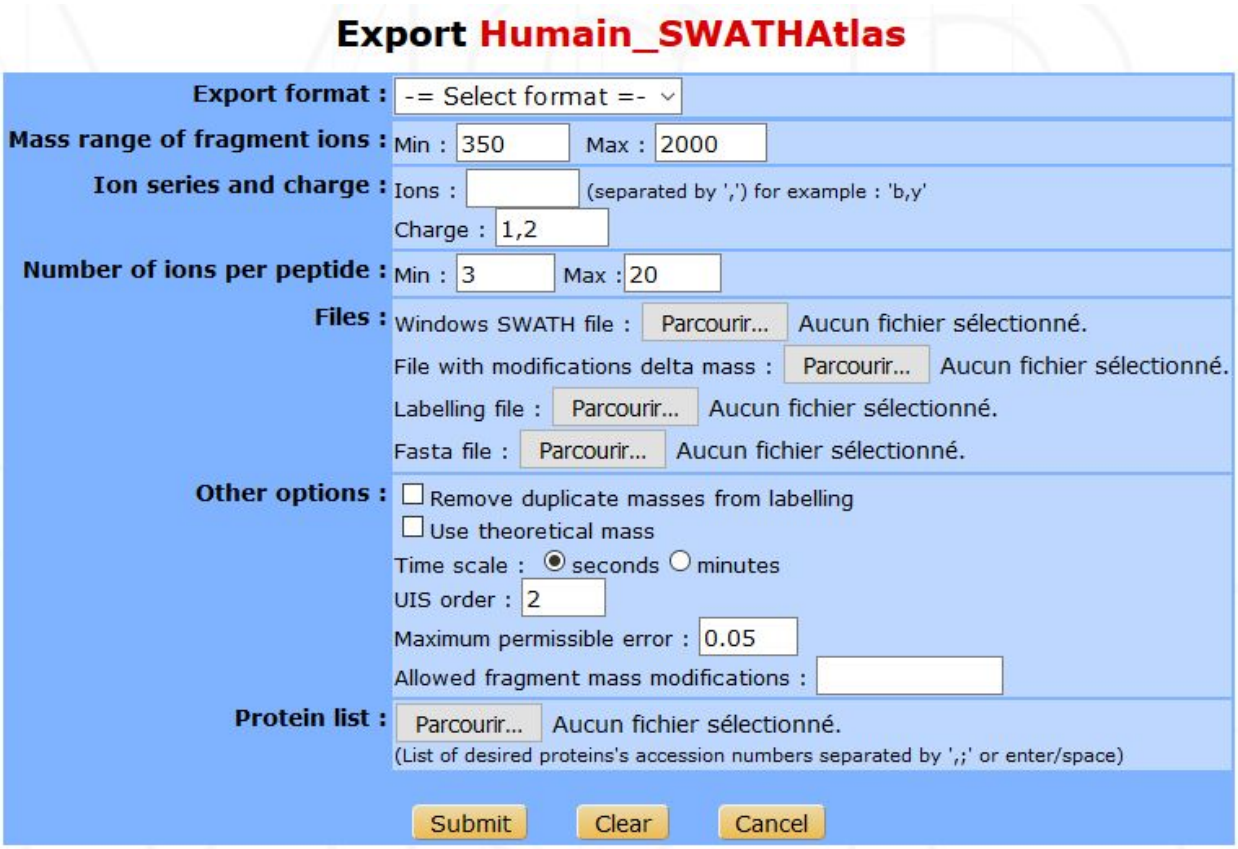
Peptide coverage : 38.2%

```
  1 MEEPQSDPSV EPPLSQETFS DLWKLLPENN VLSPLPSQAM DDLMLSPDDI EQWFTEDPGP

 61 DEAPRMPEAA PPVAPAPAAP TPAAPAPAPS WPLSSSVPSQ KTYQGSYGFR LGFLHSGTAK

121 SVTCTYSPAL NKMFCQLAKT CPVQLWVDST PPPGTRVRAM AIYKQSQHMT EVVRRCPHHE

181 RCSDSDGLAP PQHLIRVEGN LRVEYLDDRN TFRHSVVVPY EPPEVGSDCT TIHYNYMCNS

241 SCMGGMNRRP ILTIITLEDS SGNLLGRNSF EVRVCACPGR DRRTEEENLR KKGEPHHELP

301 PGSTKRALPN NTSSSPQPKK KPLDGEYFTL QIRGRERFEM FRELNEALEL KDAQAGKEPG

361 GSRAHSSHLK SKKGQSTSRH KKLMFKTEGP DSD
```

Some information about each peptide such as sequence, modifications, position on the protein, M/Z, charge, IRT time and specificity are shown.

Exporting a library

You can export a library to use it in a quantification software. From the libraries list interface, click on the *Export* button to display the export form.



You have to fill in the following parameters :

- **Export format** : The library can be exported for PeakView or for OpenSWATH, or you can download the final format of the library (sptxt).
- **Mass range of fragment ions** : Lower and upper mass limits of fragment ions. (min=350 and max=2000 by default).
- **Ion series and charge** : The ion desired type (a, b, c, x, y, or z) and charge separated by a comma. (charge=1+ and 2+ by default).
- **Number of ions per peptide** : Minimum and maximum number of ions per peptide. (min=3 and max=20 by default).
- **Files** :
    - **Windows SWATH file** : Upload the file that contain the SWATH window scheme that has been used for SWATH data acquisition.
    - **File with modifications delta mass** : Optional file containing the modifications not specified by default.
    - **Labelling file** : Optional file containing the amino acid isotopic labelling mass shifts. If this option is used, heavy transitions will be generated.
    - **Fasta file** : Optional databank fasta file used to relate peptides to their proteins.
- **Other options** : You can select another optional options such as the maximum permissible error, the time scale, the UIS order (calculated when using switching modification; if -1 is set, all transitions for each isoform will be reported; default : 2), or

the list of allowed fragment mass modifications.

- **Protein list** : You can select a file containing a protein list to export just these proteins from the library.

Then you can click on the *Submit* button to launch the export process. Once the process is complete, you can download the final file with a download link that will appear.

Deleting a spectral library

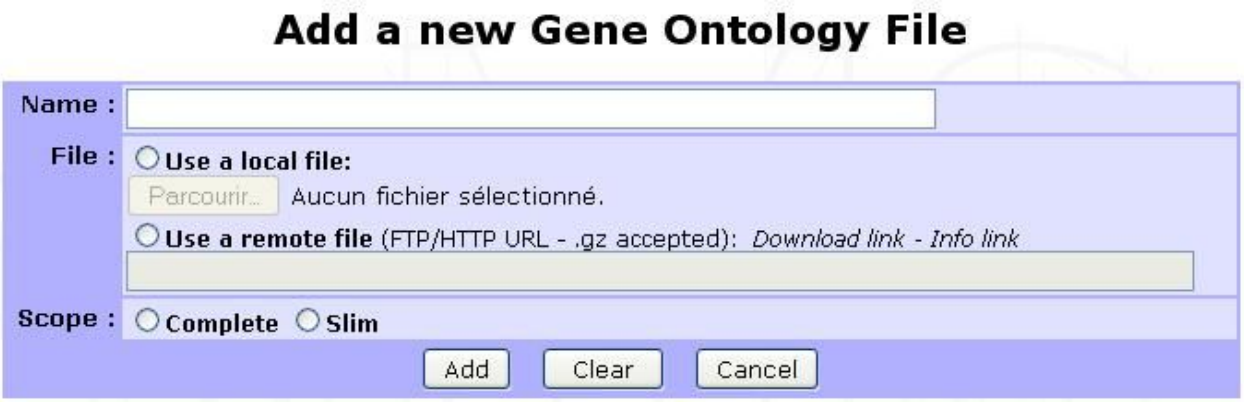
A library can be deleted from the list window (by clicking on the *Delete* button of the corresponding library) only if this library was not used to create another library (merge option, in that case, a prompt will inform you).

GO files management

GO analyses require two types of GO files: an ontology file and an annotation file. These files are not project-specific and are thus managed globally in myProMS. Only bioinformaticians and authorized massists/managers can manage GO files. From **myProMS main window**, select *Annotation data* and follow the "*GO annotations*" link to display the list of GO files recorded.

Ontology files

Ontology files contain the GO terms identifiers, description and relationships between. To **add** a new ontology file, click on *Add new Gene Ontology file*:



The displayed form requires the following information:

- **Name**: A relevant name for the ontology. This name will be displayed in all GO analysis starting forms in ontology selecting section.

- **File**: The file containing the ontology must be in **OBO** format (not XML nor database dump). Daily updated ontology files can be fetched from GO website. The file can be uploaded

directly from user computer or directly retrieved from remote FTP by writing its full URL (e.g. ftp://ftp.geneontology.org/pub/go/ontology/obo_format_1_2/gene_ontology.1_2.obo).

- **Scope**: Specify if the ontology file contains the **full** gene ontology or a **slim** version. A slim version gives a broad overview of the ontology content without the detail of the specific fine grained terms. If a slim file is used, make sure to select the *slim* option. In addition, to be able to use a slim ontology for GO analyses, at least one full ontology file must have been also recorded to allows myProMS to reconstructs missing associations between proteins and the GO terms recorded in the slim file. Running a slim GO analysis without a corresponding full ontology will cause an error!

Saved ontologies can be **edited**. If the file was retrieved by FTP and a most recent version available on the distant server, it can be downloaded again directly by clicking on *Update file*.

Annotation files

Annotation files contains mapping of protein identifiers to GO terms. They are **species-specific** and must be in **Gene Association File (GAF)** format. A large number of updated annotation files for many species can be fetched from the Uniprot-GOA database. To **add** a new annotation, click on *Add new annotation file*:

The displayed form requires the following information:

- **Name**: A relevant name for the annotation, that will be displayed on each GO analysis starting form in annotation selection section.
- **Description**: An optional description for the annotation.
- **Species**: Select the targeted species from the list of available ones (See **Species** below for more information).

- **File**: file can be uploaded from your computer or retrieved remotely from a FTP server (e.g. ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN/gene_association.goa_human.gz for the human annotation file).

- **Identifier used**: Select the protein identifier that must be used in myProMS to match the annotation's one (eg: select Uniprot ID or Uniprot AC for Uniprot-GOA files). If "Default" is selected, the default protein identifier displayed in myProMS will be used. This parameter must be set carefully to insure proper GO annotation mapping.

Species

myProMS automatically records the species associated with any protein validated. Because different strains or variants of the same species are also recorded, it is necessary to manually link these entries to the same **reference** species. Furthermore, reference species must be recorded for Gene Ontology analyses. A species management section is provided so that bioinformaticians and authorized massists/managers can manually record or correct species information. By default, a list of 5 model organisms species data is provided with myProMS as refrence.

Listing species

From **myProMS main window**, select *Annotation data* and follow the "*Species*" link to access the species management interface.

<Figure Species list>

As shown is the above screen capture, a subset of species can be listed either by **scientific** or **common name** by selecting the appropriate initial letter in one of the 2 alphabets displayed.

Adding or editing a species

A species can be added or edited by clicking on "*Add species*" or "*Edit*" buttons respectively. The following form is then displayed:

<Figure Add/edit species>

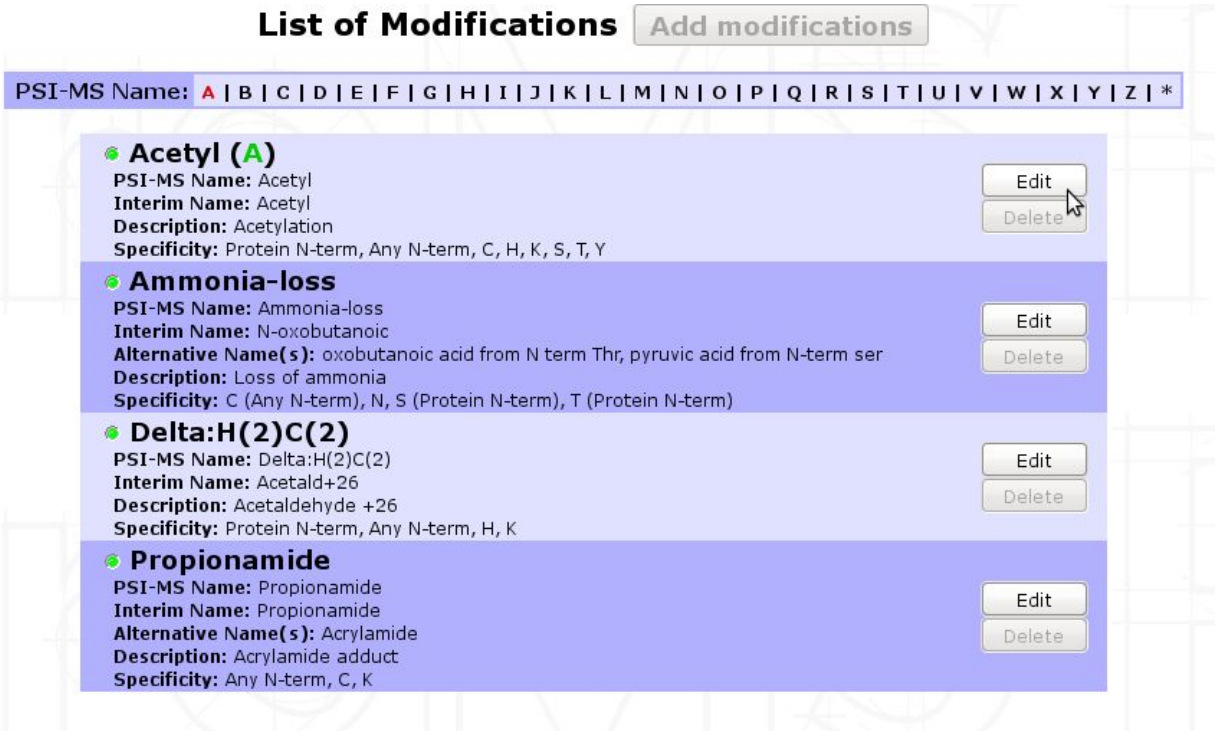
The common name, scientific name and taxonID fields are mandatory. A link to the **NCBI Taxonomy** resource is provided to help you find this information if not known. You can either set this species as reference by checking the "Is reference" or link it to a reference one In addition an optional field allows you to link any species with a reference one by selecting a target species in the drop-down menu.

Deleting a species

A species can be deleted from the list interface (by clicking on the Delete button of the corresponding species) only if this species is no longer associated with validated protein, not set as reference species nor used in a GO analysis.

Sequence modifications

myProMS automatically records the post-translational modifications (PTMs) found in imported analyses on protein sequences. Once an analysis has been imported, PTMs found on this analysis are added to the list and you can edit the properties of those PTMs. myProMS keeps track of every imported modifications and displays by ascending name as defined on UNIMOD website.
In this list, all PTMs are depicted by five informations ; names (PSI-MS and interim name), description, specificity and status (red or green). The specificity describes on which residue the PTM tends to occur. It could be on a specific residue including or not a context (like “Any N-term”, “Protein N-term”,etc.).

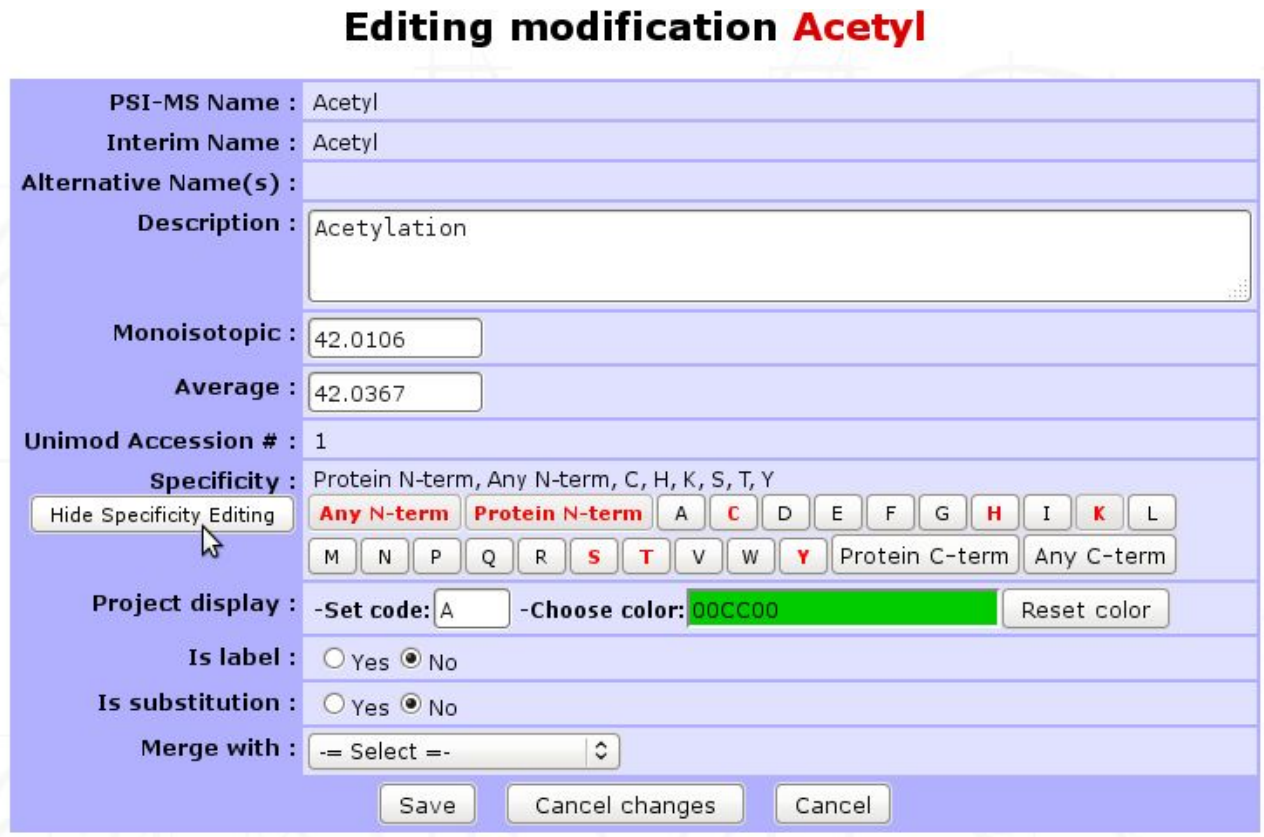


The status displayed on the upper left corner by a circle tells if a PTM is valid or not. To be valid, a PTM should be characterized by a monoisotopic and an average mass like the ones defined on UNIMOD website. If a PTM is not valid ( ), it means that myProMS could not retrieved this PTM through the UNIMOD current list of PTMs. The origin of that issue comes from one reason : you entered a “home-named” modification that is not referenced in UNIMOD. Two solutions exist to solve this issue :
- if this modification was already imported on another referenced name, you should merge this

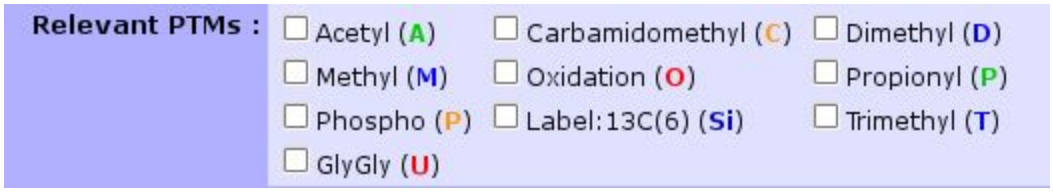
“home-named” modification to this one by editing the non-valid PTM. In the future, myProMS will automatically applies to this “home-named” modification the properties of the referenced one.
- if this modification was not imported through another name, you should edit the PTM and provide mass and specificity.
Make sure that all PTMs retrieved are valid in order to avoid the other features available in myProMS to give wrong output (like fragmentation table of peptides for example).

Editing or merging PTMs

A PTM can be edited by clicking on “*Edit*” button.



In this mode, you can update the description or the delta-mass of this PTM. A link to UNIMOD is provided by giving the Unimod Accession number. Specificity can be updated given your expertise on the PTM and reviews articles you may have read.
The option “*Merge with*” gives the opportunity to merge two PTMs into one single entry. This could be useful if you wish to give an alternative name to a modification. Select the modification you want to merge with the current PTM and click on save. This action will add the name of the current modification to the list of alternative names of the one selected.
For PTMs that you want to make appear in your projects and give special attention to, you need to enter a code (usually, a single letter) and a color. Those PTMs will become relevant and will be choosable in every project you manage.
Here is a list of relevant PTMs and their according code-color designation.



For more information on that topic, please, see the Project creation and settings section.

Settings

Instrument settings

Validation templates management

Server management

Log files

Server statistics

Testing server