
myproms Documentation

Patrick Poulet, Valentin Sabatet, Victor Laigle

Aug 02, 2019

1	Git Repository	3
2	Docker deployment	5
3	Login	7
4	Server Architecture	9
5	Users management	11
5.1	User classes and access privileges	11
5.2	Account management	12
6	Annotation data management	13
6.1	Sequence databanks	13
6.1.1	Databank types	13
6.1.2	Listing databanks	14
6.1.3	Adding a new databank	14
6.1.4	Editing a databank	15
6.1.5	Deleting a databank	15
6.2	Spectral (SWATH) libraries management	15
6.2.1	Listing spectral libraries	15
6.2.2	Adding a new library	16
6.2.3	Merging two library	18
6.2.4	Editing a library	18
6.2.5	Updating a library	19
6.2.6	Restoring the previous version of a library	19
6.2.7	Searching for proteins in a library	19
6.2.8	Exporting a library	21
6.2.9	Deleting a spectral library	23
6.3	GO files management	23
6.3.1	Ontology files	23
6.3.2	Annotation files	24
6.4	Species	25
6.4.1	Listing species	25
6.4.2	Adding or editing a species	25
6.4.3	Deleting a species	25
6.4.4	Sequence modifications	25

6.5	Editing or merging PTMs	26
7	Server administration	29
8	Projects settings	31
8.1	Selection	31
8.2	Creation	32
8.3	Accessibility	34
8.4	Organization	34
8.4.1	Experiments	35
8.4.2	Samples	36
8.4.3	Analyses	37
8.5	Navigation	37
8.6	Life span	37
9	Analysis import	39
9.1	Supported search engines	40
9.2	Collecting search files	41
9.3	Import parameters (Mascot, Proteome Discoverer and X!Tandem)	42
9.4	Importing MaxQuant data	44
9.5	Import DIA data	44
9.5.1	From PeakView	44
9.5.2	From OpenSWATH	45
9.5.3	Running OpenSWATH quantification	45
9.5.4	From Spectronaut	46
9.6	Analysis summary	46
10	Analysis validation	49
10.1	Automated peptide/protein validation	49
10.1.1	FDR (False discovery rate) - based validation	49
10.1.2	Qualitative validation	50
10.1.3	Comparative validation	50
10.1.4	Validation templates	50
10.2	Manual peptide/protein validation	50
10.2.1	Peptides selection/exclusion	50
10.2.2	Protein exclusion and filtering	50
10.3	Lower-scoring peptides activation	50
10.4	Clear peptide/protein selections	50
10.5	Sequence modification validation	50
10.5.1	Phosphorylation sites validation with PhosphoRS	50
10.5.2	Manual validation of modifications	50
10.6	Validation traceability	50
10.7	Reporting	50
11	Biological Samples	51
11.1	Properties	51
11.2	Treatments	51
11.3	Recording a biological sample	51
11.4	Linking biological samples to MS Analyses	51
12	Validated proteins	53
12.1	Match groups and protein visibility	53
12.1.1	Top protein selection rules	53
12.1.2	Project-wide protein visibility rules	54
12.1.3	Checking for conflicting match groups	54

12.1.4	Displaying match group composition	54
12.1.5	Manual edition	55
12.1.6	Peptide distribution in match group	55
12.2	Identifier mapping	55
12.3	Single protein view	55
13	Protein lists	57
13.1	Project-based protein lists	57
13.2	User-defined protein lists	57
13.3	Export protein lists	57
14	Search for proteins	59
15	Compare project proteins	61
15.1	Full protein-level comparison	61
15.2	Pairwise protein-level comparison	61
15.3	Pairwise peptide-level comparison	61
15.4	Saving a comparison	61
16	Peptide Quantification	63
16.1	Data import from search results file	63
16.2	Data extraction from LC/MS file with MassChroQ	63
16.2.1	Managing mzXML files	63
16.2.2	Running XIC extraction	65
16.3	Virtual/Ghost peptides and proteins	68
16.4	Displaying peptide quantification data	69
17	Protein Quantification	71
17.1	Absolute abundance quantification	71
17.1.1	emPAI (label-free)	71
17.1.2	SIn (label-free)	71
17.1.3	MaxQuant: Intensity, LFQ, iBAQ	71
17.1.4	Proteomic Ruler	71
17.1.5	Displaying single abundance quantification data	75
17.2	Relative abundance quantification	75
17.2.1	Single-Analysis quantification (labeled)	75
17.2.2	Design-based quantifications	76
17.2.3	Displaying relative abundance quantification data	77
17.3	Label-free quantifications	77
17.3.1	TnPQ	77
17.4	Comparing multiple protein quantifications	78
17.5	Exporting multiple quantifications	78
18	PTMs quantification	81
18.1	Set PTMs relevance to project	81
18.2	Display modification sites distribution	81
18.3	Compare PTMs between projects	81
18.4	Quantify modification sites	81
19	Exploratory analysis	83
19.1	Launching exploratory analyses on protein quantification	83
19.2	Launching exploratory analyses on peptides count	85
19.3	Principal Component Analysis (PCA)	86
19.3.1	Summary / edit / delete	87
19.3.2	Displaying a PCA	88

19.4	2D-Clustering	89
19.4.1	Summary / edit / delete	89
19.4.2	Displaying clustering	90
20	Functional analysis	91
20.1	Gene Ontology	91
20.1.1	GO summary	91
20.1.2	GO enrichment analysis	92
20.1.3	Quantitative gene enrichment analysis	95
20.2	Pathway enrichment	98
20.2.1	Launch pathway analysis	99
20.2.2	Summary / edit / delete	99
20.2.3	Displaying pathway analysis	99
20.2.4	Export	104
20.3	Motif enrichment analysis	104
20.3.1	Launch analysis	105
20.3.2	Summary/editing/deleting	106
20.3.3	Displaying motif enrichment	107
20.3.4	Heatmap motif analysis	107
21	Indices and tables	111

CHAPTER 1

Git Repository

CHAPTER 2

Docker deployment

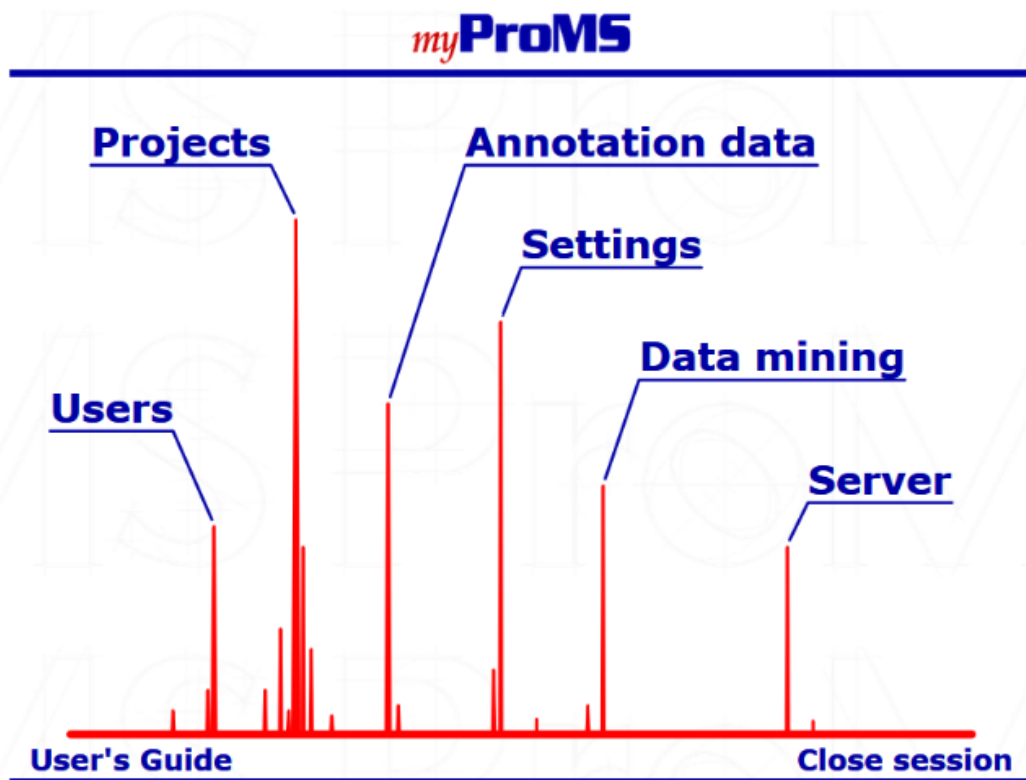
CHAPTER 3

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 Architecture

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 to users depending on their access privileges. Typically, most end users will have access only to their account and projects. See [User classes and access privileges](#) for more information.

5.1 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 as 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.

5.2 Account management

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

User information:

Login :

Password : Confirm :

Status :

Workgroup :

Mascot IDs : User: Group(s): (comma-separated list)

User name :

Laboratory :

Telephone :

E-mail :

Other information :

Preferences : ☒ Use interactive spectrum
☐ Set label vertical

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.

Note: A user cannot create an account class 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.

Annotation data management

6.1 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.

6.1.1 Databank types

Multiple databank types are available in *myProMS* depending on the proteomic resource used to download the fasta file, corresponding to **different fasta file entries**:

- **UniProt:** >sp|P15311|EZRI_HUMAN Ezrin OS=Homo sapiens (Human) GN=EZR Ezrin

	UniProt - ALL	UniProt - ACC / ACC#-n	UniProt - ID
Protein identifier	sp P15311 EZRI_HUMAN	P15311	EZRI_HUMAN
Description	Ezrin		
Species	Homo sapiens		

- **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:** >gil125987826|sp|P15311|EZRI_HUMAN Ezrin (p81) (Cytovillin) (Villin-2) [Homo sapiens (Human)]

	NCBI - ALL	NCBI - GI
Protein identifier	gil125987826 sp P15311 EZRI_HUMAN	gil125987826
Description	Ezrin (p81) (Cytovillin) (Villin-2)	
Species	Homo sapiens	

- **IPI:** >IPI:IPI00843975.1|SWISS-PROT:P15311| Tax_Id=9606 Gene_Symbol=EZR Ezrin‘

Protein identifier	IPI00843975
Description	Tax_Id=9606 Gene_Symbol=EZR Ezrin
Species	9606

Warning: 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

Protein identifier	Whole identifier block
Description	Everything else
Species	None recorded

Note: If you need you need to handle other entries type, please contact your local *myProMS* administrator or email to myproms@curie.fr.

6.1.2 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>

6.1.3 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.

6.1.4 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.

6.1.5 Deleting a databank

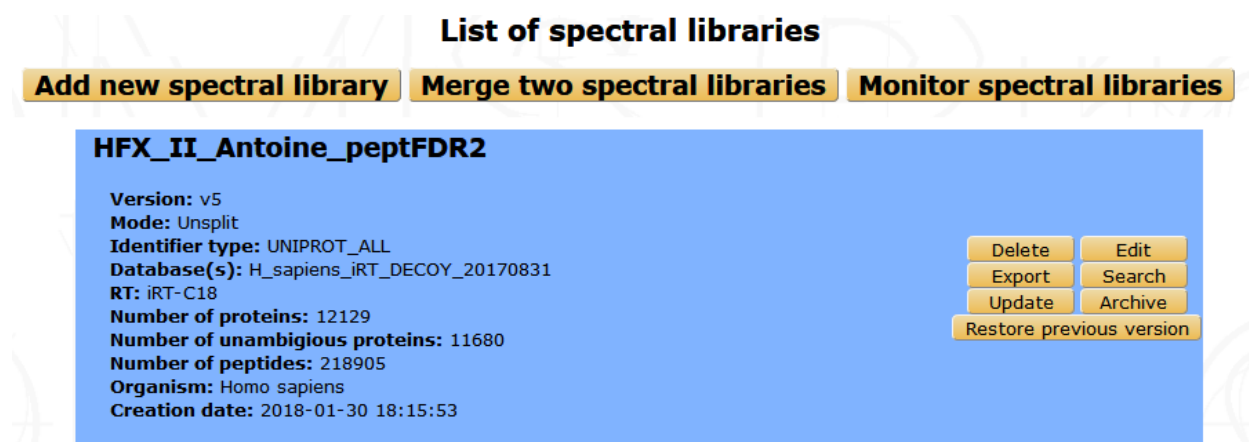
We recommend to delete any databank that will no longer be used to keep the list displayed as short as possible. Deletion of a databank has no effect on the traceability information of analyses using this databank. A databank can be deleted at any time except during import of analyses using this databank.

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.

6.2 Spectral (SWATH) libraries management

6.2.1 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.



6.2.2 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.

Adding new spectral library to Server

Task : ☐ Merge with an other library
☐ Create new library

Species :
(Update list of reference species if your species is not listed)

Consensus library options : ☐ Split ☐ Unsplit

Files : ☐ **Import files :**
 Aucun fichier sélectionné.
 Aucun fichier sélectionné.
 Aucun fichier sélectionné.
(.dat, .mzXML and .tandem.pep.xml)

☐ **Import from project :**

☐ **Import archive file :**
 Aucun fichier sélectionné.
(zip or gz archive)

☐ **Select directory from server :**

(only for bioinformaticien)

☐ **Shared data directory**

Fragmentation type :

Instrument :

Databank :

Mayu options: FDR estimation with Mayu software.
Missed cleavage :
FDR : Type :

RT file :

Description :

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.

Danger: Update with the new version

6.2.3 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.

Important: Only two libraries with the same iRT file, databank type and consensus library option (SPLIT or UNSPLIT) can be merged.

6.2.4 Editing a library

From the libraries list window, click on the `Edit` button on the right side of the library row. The following form will be displayed :

Edit Humain_SWATHAtlas

New name :

Description :

Note: Only the name and the description can be modified

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

6.2.5 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.

6.2.6 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.

6.2.7 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 on 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.

Search in Humain_SWATHAtlas

Entry :

Species name : ☒ *Homo sapiens*

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.

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	QSQHMTVVVR	-	165-174	607.8010	2+	-20	100	P04637
6	CSDSGLAPPQHLIR	Carbamidomethyl (C:1)	182-196	833.4043	2+	24.1	100	P04637
7	CSDSGLAPPQHLIR	Carbamidomethyl (C:1)	182-196	555.9386	3+	25.6	100	P04637
8	RPILTIITLEDSSGNLLGR	-	249-267	690.0635	3+	95.2	100	P04637
9	RTEENLR	-	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

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 AIY QSQHMT EVVRRCPPHE								
181 RCSDSGLAP PQHLIR VEGN LRVEYLDDRN TFRHSVVVPY EPPEVGSDCT TIHYNMNCNS								
241 SCMGGMNRRP ILTIITLED S SGNLLGR NSF EVRVCACPGR DRR RTEENLR KKGEPHEL P								
301 PGSTKR ALPN NTSSSPQPK K KPLDGEYFTL QIRGRER FEM FRELNEALEL KDAQAGKE PG								
361 GSRAHSSHLK SKKGQSTSRH KKLMEKTEGP DSD								

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

6.2.8 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.

Export Humain_SWATHAtlas

Export format :	- Select format -	
Mass range of fragment ions :	Min : 350	Max : 2000
Ion series and charge :	Ions : (separated by ',') for example : 'b,y'	
	Charge : 1,2	
Number of ions per peptide :	Min : 3	Max : 20
Files :	Windows SWATH file : Parcourir... Aucun fichier sélectionné. File with modifications delta mass : Parcourir... Aucun fichier sélectionné. Labelling file : Parcourir... Aucun fichier sélectionné. Fasta file : Parcourir... Aucun fichier sélectionné.	
Other options :	<input type="checkbox"/> Remove duplicate masses from labelling <input type="checkbox"/> Use theoretical mass Time scale : <input checked="" type="radio"/> seconds <input type="radio"/> minutes UIS order : 2 Maximum permissible error : 0.05 Allowed fragment mass modifications :	
Protein list :	Parcourir... Aucun fichier sélectionné. (List of desired proteins's accession numbers separated by ',;' or enter/space)	
<div> <input type="button" value="Submit"/> <input type="button" value="Clear"/> <input type="button" value="Cancel"/> </div>		

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.

6.2.9 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).

6.3 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*. From *myProMS* main window, select `Annotation data` and follow the `GO annotations` link to display the list of GO files recorded.

Note: Only bioinformaticians and authorized assistants/managers can manage GO files

6.3.1 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.

Note: 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.

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.

Warning: 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`.

6.3.2 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`:

Add a new Annotation File

Name :

Description :

Species : (Only reference species can be selected)

File : ☐ **Use a local file:**
 Aucun fichier sélectionné.

☐ **Use a remote file** (FTP/HTTP URL - .gz accepted):

Identifier used :

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.

Warning: **Identifier** must be set carefully to insure proper GO annotation mapping

6.4 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 assistants/managers can manually record or correct species information. By default, a list of 5 model organisms species data is provided with *myProMS* as reference.

6.4.1 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 in 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.

6.4.2 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.

6.4.3 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.

6.4.4 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.).

List of Modifications Add modifications

PSI-MS Name: A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z | *

<p>Acetyl (A)</p> <p>PSI-MS Name: Acetyl</p> <p>Interim Name: Acetyl</p> <p>Description: Acetylation</p> <p>Specificity: Protein N-term, Any N-term, C, H, K, S, T, Y</p>	<div>Edit</div> <div>Delete</div>
<p>Ammonia-loss</p> <p>PSI-MS Name: Ammonia-loss</p> <p>Interim Name: N-oxobutanoic</p> <p>Alternative Name(s): oxobutanoic acid from N term Thr, pyruvic acid from N-term ser</p> <p>Description: Loss of ammonia</p> <p>Specificity: C (Any N-term), N, S (Protein N-term), T (Protein N-term)</p>	<div>Edit</div> <div>Delete</div>
<p>Delta:H(2)C(2)</p> <p>PSI-MS Name: Delta:H(2)C(2)</p> <p>Interim Name: Acetald+26</p> <p>Description: Acetaldehyde +26</p> <p>Specificity: Protein N-term, Any N-term, H, K</p>	<div>Edit</div> <div>Delete</div>
<p>Propionamide</p> <p>PSI-MS Name: Propionamide</p> <p>Interim Name: Propionamide</p> <p>Alternative Name(s): Acrylamide</p> <p>Description: Acrylamide adduct</p> <p>Specificity: Any N-term, C, K</p>	<div>Edit</div> <div>Delete</div>

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).

6.5 Editing or merging PTMs

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

Editing modification **Acetyl**

PSI-MS Name :	Acetyl
Interim Name :	Acetyl
Alternative Name(s) :	
Description :	Acetylation
Monoisotopic :	42.0106
Average :	42.0367
Unimod Accession # :	1
Specificity :	Protein N-term, Any N-term, C, H, K, S, T, Y
Hide Specificity Editing	<input checked="" type="checkbox"/> Any N-term <input checked="" type="checkbox"/> Protein N-term <input type="checkbox"/> A <input type="checkbox"/> C <input type="checkbox"/> D <input type="checkbox"/> E <input type="checkbox"/> F <input type="checkbox"/> G <input type="checkbox"/> H <input type="checkbox"/> I <input type="checkbox"/> K <input type="checkbox"/> L <input type="checkbox"/> M <input type="checkbox"/> N <input type="checkbox"/> P <input type="checkbox"/> Q <input type="checkbox"/> R <input type="checkbox"/> S <input type="checkbox"/> T <input type="checkbox"/> V <input type="checkbox"/> W <input type="checkbox"/> Y <input type="checkbox"/> Protein C-term <input type="checkbox"/> Any C-term
Project display :	-Set code: <input type="text" value="A"/> -Choose color: <input type="text" value="00CC00"/> <input type="button" value="Reset color"/>
Is label :	<input type="radio"/> Yes <input checked="" type="radio"/> No
Is substitution :	<input type="radio"/> Yes <input checked="" type="radio"/> No
Merge with :	<input type="text" value="- Select -"/>
<input type="button" value="Save"/> <input type="button" value="Cancel changes"/> <input type="button" value="Cancel"/>	

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 selectable in every project you manage.

Here is a list of relevant PTMs and their associated code-color designation:

Relevant PTMs :	<input type="checkbox"/> Acetyl (A)	<input type="checkbox"/> Carbamidomethyl (C)	<input type="checkbox"/> Dimethyl (D)
	<input type="checkbox"/> Methyl (M)	<input type="checkbox"/> Oxidation (O)	<input type="checkbox"/> Propionyl (P)
	<input type="checkbox"/> Phospho (P)	<input type="checkbox"/> Label:13C(6) (Si)	<input type="checkbox"/> Trimethyl (T)
	<input type="checkbox"/> GlyGly (U)		

Note: For more information on that topic, please, see Project [Creation](#).

CHAPTER 7

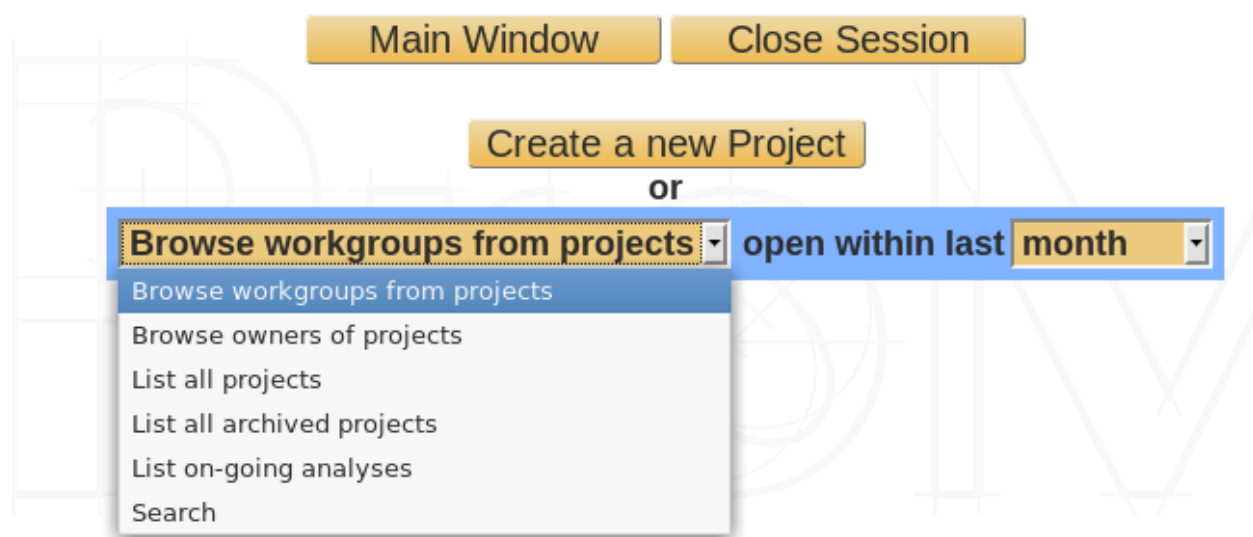
Server administration

Projects settings

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.

8.1 Selection

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



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:

Create a new Project

or

Search

Search for :

Match is **not** case sensitive.

Match : **all words** ▾

in :

- ☐ **Item name (Project, Experiment, ...)**
- ☐ **Data files (raw data or search results)**
- ☐ **Protein identifier**
- ☐ **Protein description**

Restrict to : **all** ▾ **projects.** ☐ **Include archived projects**

Search

Projects are then listed together with the items that were matched during the search. Once listed, click on the Open button to go to the selected project.

8.2 Creation

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:

Creating a new Project

Name :	<input type="text"/>
Description :	<input type="text"/>
Protein visibility :	<input checked="" type="radio"/> A protein is Visible only when Alias of a Match Group. <input type="radio"/> A protein is Visible everywhere if Alias of at least 1 Match Group. <input type="radio"/> A protein is Visible everywhere if Alias or made Visible in at least 1 Match Group.
Identifier conversion :	<input type="text" value="None"/>
Relevant PTMs :	<input type="checkbox"/> Acetyl (A) <input type="checkbox"/> Biotin (B) <input type="checkbox"/> ,Biotine-phenol, (B) <input type="checkbox"/> ,BG-PEG9-NHS, (BG) <input type="checkbox"/> Carbamidomethyl (C) <input type="checkbox"/> Dimethyl (D) <input type="checkbox"/> Methyl (M) <input type="checkbox"/> Oxidation (O) <input type="checkbox"/> Phospho (P) <input type="checkbox"/> Propionyl (P) <input type="checkbox"/> Sulfo (S) <input type="checkbox"/> ,Snap-Tag, (ST) <input type="checkbox"/> ,Snap-Tag oxyde, (STO) <input type="checkbox"/> Trimethyl (T) <input type="checkbox"/> GG (U)
Project owner :	<input type="text" value="Marine Le Picard"/>
Workgroup :	<input type="text" value="None"/>
Start date :	07/02/2018 18:06:01
Status :	Starting
Comments :	<input type="text"/>
<input type="button" value="Save"/> <input type="button" value="Clear"/> <input type="button" value="Cancel"/>	

- **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.
- **Identifier 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.

Note: Projects can be edited at any time to modify any of these settings.

8.3 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.

Accessibility to Project User1
No Workgroup assigned.

Users allowed to access this Project

User	Status	Workgroup	Access Right*
No users			

Allow **to access this Project.**

Access rights description:

Guest : Read access to validated data.

User : Read/Write access to validated data.

Administrator : **User** + Project access management.

Power (User/Administrator) : **User/Admin.** with additional read/write access to non-validated protein data.

Super (User/Administrator) : **User/Admin.** with full access rights on the current project.

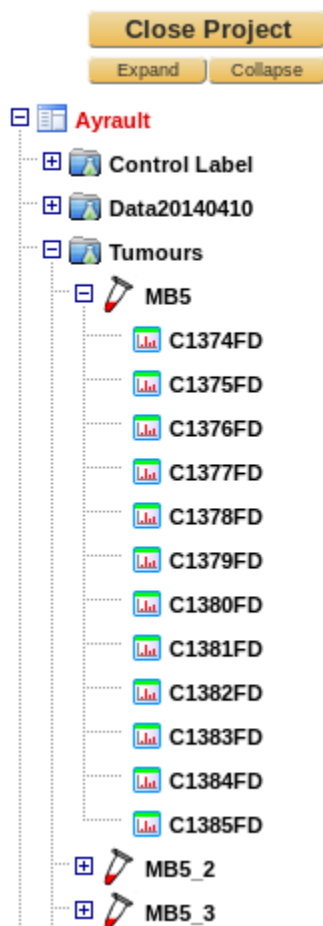
Manager : Full access rights on all projects of a workgroup.

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](#) for more information.

Click on the `Save` button to validate any changes.

8.4 Organization

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



8.4.1 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.

Adding new Experiment(s)

Name :	<input type="text"/>
Multiple entries labels :	<input type="text"/> Use ',' between single values and '-' for range (eg. 1,3,5-10).
Description :	<input type="text"/>
Start date :	07/02/2018 18:15:57
Preferred species :	-- Select --
Comments :	<input type="text"/>
<input type="button" value="Save"/> <input type="button" value="Clear"/> <input type="button" value="Cancel"/>	

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");

8.4.2 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 to 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.

Adding new Sample(s)

Name :	<input type="text"/>
Multiple entries labels :	<input type="text"/> Use ',' between single values and '-' for range (eg. 1,3,5-10).
Description :	<input type="text"/>
Start date :	07/02/2018 18:16:50
Comments :	<input type="text"/>
<input type="button" value="Save"/> <input type="button" value="Clear"/> <input type="button" value="Cancel"/>	

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

8.4.3 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.

8.5 Navigation

- Navigation frame
- Sub-navigation frame
- Option frame
- Results frame

Warning: A COMPLETER

8.6 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.

CHAPTER 9

Analysis import

The collection of spectra/peptides/proteins(/quantification) data contained in a search result file are imported into myProMS as an **Analysis**.

Note: Only bioinformaticians, massists and managers can import Analyses.

Select the **Experiment**, **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.

Process Multiple Analyses

Process type : **Analysis Management** ▼

Analysis Management:

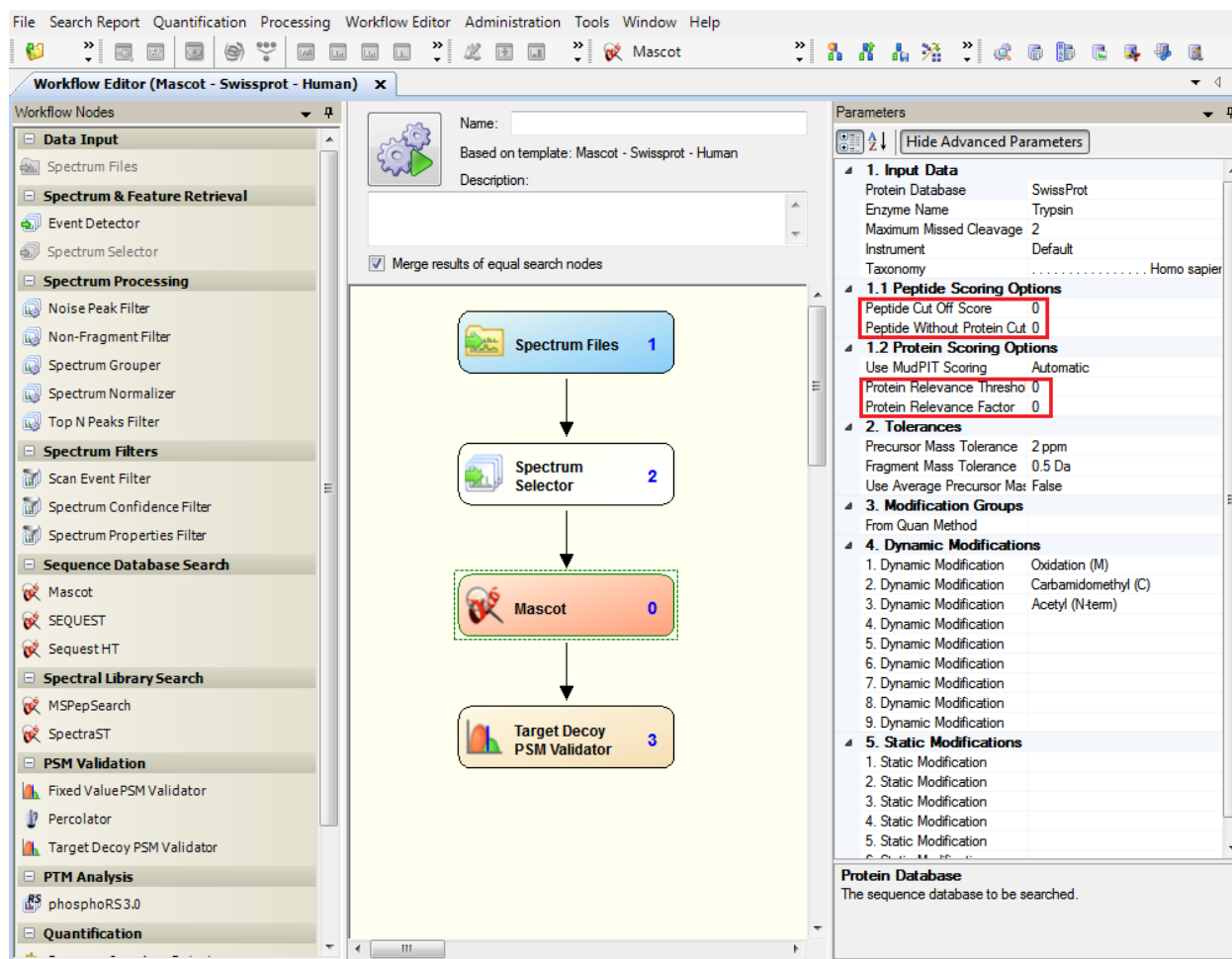
Proceed	Import multiple analyses
Proceed	Import decoy data into multiple analyses
Proceed	Import elution time into multiple analyses
Proceed	Delete multiple analyses
Proceed	Duplicate multiple analyses

9.1 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).

Attention: If you perform Mascot searches with Proteome Discoverer (PD) Software, make sure you do not use Protein and Peptide filters. In the **Workflow Editor**, click on Mascot node and then, set the four filters Peptide Cut 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.



9.2 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).

To select data files in **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](#)).

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 separate 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: 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.

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

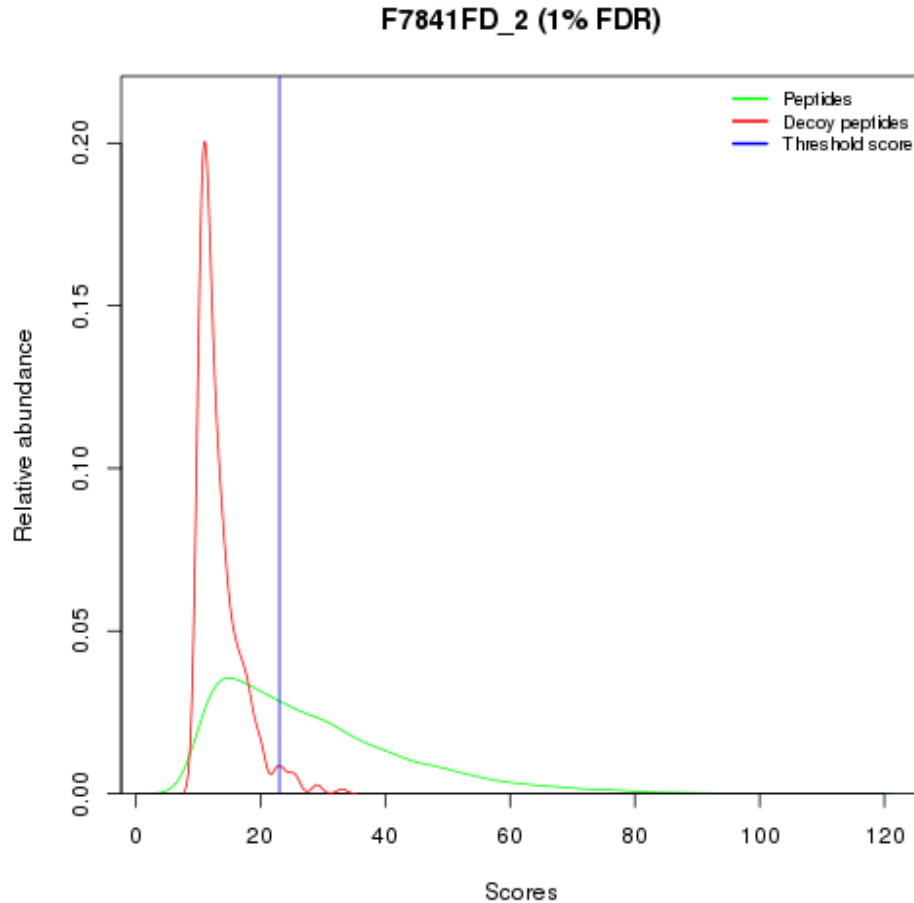
Files retrieved are listed in alphabetical order together with pertinent information about the search performed: the MS file, search type (MS2, MS1 or mix of both), databank(s) and taxonomy and search title used.

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 use the quality algorithm¹, 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.

¹ Quality algorithm : Käll et al. Bioinformatics 2008, 25

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

9.4 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 `Import 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`. Specify 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.

9.5 Import DIA 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.

9.5.1 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.

Select options to import SWATH data

File :	Parcourir...	Aucun fichier sélectionné.
Library name :	-= Select Library =-	
FDR :	<input type="text"/>	%
PeakView parameters :		
Peptide filter :		
Number of peptides per protein : <input type="text"/>		
Number of transitions per peptide : <input type="text"/>		
<input type="checkbox"/> Exclude modified peptides		
XIC options :		
XIC extraction window (min) : <input type="text"/>		
XIC width : <input type="text"/> Da		
<input type="button" value="Submit"/> <input type="button" value="Clear"/> <input type="button" value="Cancel"/>		

9.5.2 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.

Select options to import OpenSwath data

Result file :	Parcourir...	Aucun fichier sélectionné.
Library name :	-= Select Library =-	
Library export parameter file :	Parcourir...	Aucun fichier sélectionné.
Library export file :	Parcourir...	Aucun fichier sélectionné.
TRIC methode used:	LOCAL MST*	*Recommanded option
Software version :	<input type="text"/>	ex : 1.2, 2.1.3 ...
<input type="button" value="Submit"/> <input type="button" value="Clear"/> <input type="button" value="Cancel"/>		

9.5.3 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.

OpenSwath quantification

Library name :	<input type="text" value="-= Select Library =-"/>
Library export management :	<input type="radio"/> Use the selected library <input type="radio"/> Import your own library formatted for Openswath
OpenSwath parameters files :	iRT file : <input type="text" value="Parcourir..."/> Aucun fichier sélectionné. Windows file : <input type="text" value="Parcourir..."/> Aucun fichier sélectionné.
mzXML files :	<input type="radio"/> Upload multiple files <input type="text" value="Parcourir..."/> Aucun fichier sélectionné. <input type="radio"/> Import from shared data directory
OpenSwath workflow options:	mz_threshold : <input type="text" value="0.05"/>
Pyprophet options:	d_score.cutoff : <input type="text" value="1"/>
TRIC methode:	<input type="text" value="LOCAL MST*"/> *Recommended option
Merge with other experiment:	<input type="checkbox"/> OpenSwath_all_20transitions_90maxRTdiff <input type="checkbox"/> OpenSwath_all_6transitions_30maxRTdiff
<input type="button" value="Submit"/> <input type="button" value="Clear"/> <input type="button" value="Cancel"/>	

Danger: A MODIFIER AVEC LE NOUVEAU WORKFLOW

9.5.4 From Spectronaut

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

Select options to import Spectronaut data

Result file :	<input type="text" value="Parcourir..."/> Aucun fichier sélectionné.
Library name :	<input type="text" value="-= Select Library =-"/>
Library export parameter file :	<input type="text" value="Parcourir..."/> Aucun fichier sélectionné.
Library export file :	<input type="text" value="Parcourir..."/> Aucun fichier sélectionné.
Software version :	<input type="text" value=""/> ex : 1.2, 2.1.3 ...
<input type="button" value="Submit"/> <input type="button" value="Clear"/> <input type="button" value="Cancel"/>	

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).

9.6 Analysis summary

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

Analysis F7228FD

Name :	F7228FD
Description :	
Position :	1
Start date :	12/12/2012 14:29:05
MS type :	MS/MS Ions Search Export Elution Data
Search engine :	Mascot from Proteome Discoverer
Search file :	F7228FD_2.pdm Show Search Parameters
MS data file :	F7228FD.RAW
Databank(s) :	•SwisProt-Mascot Type: UniProt - ID
Taxonomy :	Homo sapiens (human)
Labeling :	None
Max. rank :	1
Min. score :	20.52 [FDR-based] Show Score Distribution
False discovery :	Targeted FDR: 1% - Observed FDR: [Data not reported: No peptides validated.] Decoy method: Automated decoy search.
Status :	Not validated. Show validation history
Comments :	

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.

CHAPTER 10

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

- : 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.

10.1 Automated peptide/protein validation

10.1.1 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 *Analysis import*).

10.1.2 Qualitative validation

10.1.3 Comparative validation

10.1.4 Validation templates

10.2 Manual peptide/protein validation

10.2.1 Peptides selection/exclusion

10.2.2 Protein exclusion and filtering

10.3 Lower-scoring peptides activation

10.4 Clear peptide/protein selections

10.5 Sequence modification validation

10.5.1 Phosphorylation sites validation with PhosphoRS

PhosphoRS¹ 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.

10.5.2 Manual validation of modifications

10.6 Validation traceability

10.7 Reporting

¹ PhosphoRS: T. Taus et al., J. Proteome Res., 2011

CHAPTER 11

Biological Samples

11.1 Properties

11.2 Treatments

11.3 Recording a biological sample

11.4 Linking biological samples to MS Analyses

12.1 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.

12.1.1 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:
 - a. SwissProt identifier.
 - b. trEMBL identifier.

- c. None of the following keywords found in the protein description.
 - d. The hypothetical keyword is found in the protein description.
 - e. The unknown or unnamed keywords are found in the protein description.
 - f. 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.*

12.1.2 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 [Saving a comparison](#)) can be modified by the resulting changes in protein visibility.

12.1.3 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 [Manual edition](#)).
- Display detailed information on a protein in the context of the analysis of your choice by directly clicking on that analysis' name.

12.1.4 Displaying match group composition

List Proteins at Analysis-level (see [Project-based protein lists](#)) 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>

12.1.5 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)* lets you 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.

12.1.6 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 [Single protein view](#) for more information on how to interpret the displayed data.

12.2 Identifier mapping

12.3 Single protein view

CHAPTER 13

Protein lists

13.1 Project-based protein lists

13.2 User-defined protein lists

13.3 Export protein lists

CHAPTER 14

Search for proteins

Compare project proteins

15.1 Full protein-level comparison

15.2 Pairwise protein-level comparison

15.3 Pairwise peptide-level comparison

15.4 Saving a comparison

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...).

16.1 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).

16.2 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**¹ to perform this task. However, the corresponding LC/MS file(s) must be provided in **mzXML** format.

16.2.1 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

¹ MassChroQ : Valot B et al, *Proteomics*, 2011

displayed, select **Analysis Quantification** to display the list of available options. Click on **Proceed** next to the **Manage mzXML files** process as shown below.

Process Multiple Analyses

Process type : Analysis Quantification ▼

Monitor on-going quantifications

Peptide Quantification:

Proceed	Manage mzXML files
Proceed	XIC extraction with MassChroQ

Protein Label-free Quantification:

Proceed	Import emPAI data
Proceed	SIN quantification

Protein Label Quantification:

Proceed	SILAC-based quantification
Proceed	iTRAQ-based quantification

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

Manage mzXML files

Upload a new file:

Files already imported:

☐ G130322_0174_c_ich.mzXML

☐ G130322_0175_c_ich.mzXML

☐ G130327_0204_c_ich.mzXML

☐ G130327_0205_c_ich.mzXML

☐ G130402_0216_c_ich_130402090421.mzXML

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/mzXML file matching in the quantification launch step.

16.2.2 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.

Select Analyses in Experiment Tg Experiment for Ext. ion chrom. Quantification

Name : Ext. ion chrom. extraction

Raw-data settings : Extraction type: Profile (for mzXML)

Alignment settings : Alignment algorithm: OBI-Warpl Reference: G130322_0175_c_ich

Align from 400 to 1200 m/z window

Peptide selection : ☐ Extract all charge states of the peptides (even if no MS/MS exists for it)

Quantification settings : Type of XIC: Sum

	Analysis	MS type & File	Labeling method	Instrument	Search file & Engine	Databank(s) Taxonomy	Min. score	Max. rank	Selected proteins
<input type="checkbox"/>	G130327_0202b_c_ich mzXML file: G130327_0202b_c_ich.mzXML	MS/MS G130327_0202b_c_ich.raw	None	ESI-FTICR	F058447.dat MASCOT	SwisProt-Mascot Rodentia	28.1422 1		585 (1348)
<input type="checkbox"/>	G130327_0203_c_ich_130329121732 mzXML file: G130327_0203_c_ich_130329121732.mzXML	MS/MS G130327_0203_c_ich_130329121732.raw	None	ESI-FTICR	F058448.dat MASCOT	SwisProt-Mascot Rodentia	29.753 1		350 (926)
<input type="checkbox"/>	G130327_0204_c_ich mzXML file: G130327_0204_c_ich.mzXML	MS/MS G130327_0204_c_ich.raw	None	ESI-FTICR	F058449.dat MASCOT	SwisProt-Mascot Rodentia	29.707 1		347 (915)
<input type="checkbox"/>	G130327_0205_c_ich mzXML file: G130327_0205_c_ich.mzXML	MS/MS G130327_0205_c_ich.raw	None	ESI-FTICR	F058450.dat MASCOT	SwisProt-Mascot Rodentia	30.8247 1		345 (911)
<input type="checkbox"/>	G130402_0216_c_ich_130402090421 mzXML file: G130402_0216_c_ich_130402090421.mzXML	MS/MS G130402_0216_c_ich_130402090421.raw	None	ESI-FTICR	F058451.dat MASCOT	SwisProt-Mascot Rodentia	28.584 1		528 (1205)
<input type="button" value="Launch Quantification"/> <input type="button" value="Cancel"/>									

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.

Important: 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.

Name : Ext. ion chrom. extraction

Raw-data settings : Extraction type: Profile (for mzXML)

Isotope labeling : SILAC

Channel1 name: Heavy

Quantification Label:

Label Name: 13C6-15N2

Modification target: Side chain

Modification: Label:13C(6)15N(2) / +8.0142 Da on K

Quantification Label:

Label Name: Arg10

Modification target: Side chain

Modification: Label:13C(6)15N(4) / +10.0083 Da on R

Add quantification label

Remove quantification label

Channel2 name: Light

Quantification Label:

Label Name: Light

Modification target: Side chain

Modification: Light / +0.0000 Da on

Add quantification label

Channel3 name:

Quantification Label:

Label Name:

Modification target: Side chain

Modification: -- Select -- on

Add quantification label

Alignment settings : Alignment algorithm: OBI-Warp Reference: -- Select --

Align from 400 to 1200 m/z window

Peptide selection : ☐ Extract all charge states of the peptides (even if no MS/MS exists for it)

Extract XIC traces : No ☒ Yes ☐

Quantification settings : Type of XIC: BasePeak XIC

More settings

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).

<Figure Monitor Quantification window>

Note: 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.

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)`.

In the second part of the form, you must select the analyses corresponding to the extraction and associate its mzXML file to each of them. If the mzXML file name matches the MS data file recorded for the Analysis, myProMS will do the job for you. (Check the [MassChroQ manual](#) for help on setting these parameters properly).

Note: 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.

16.3 Virtual/Ghost 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.

Note: *myProMS* solves this issue by adding these missing peptides to the list of validated peptides but with a special status: **virtual peptides** (also called **ghost peptide**).

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.

Important: 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 reextracted 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 **ghost protein(s)**. Ghost proteins appear in italics in most protein lists.

16.4 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 section). 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.

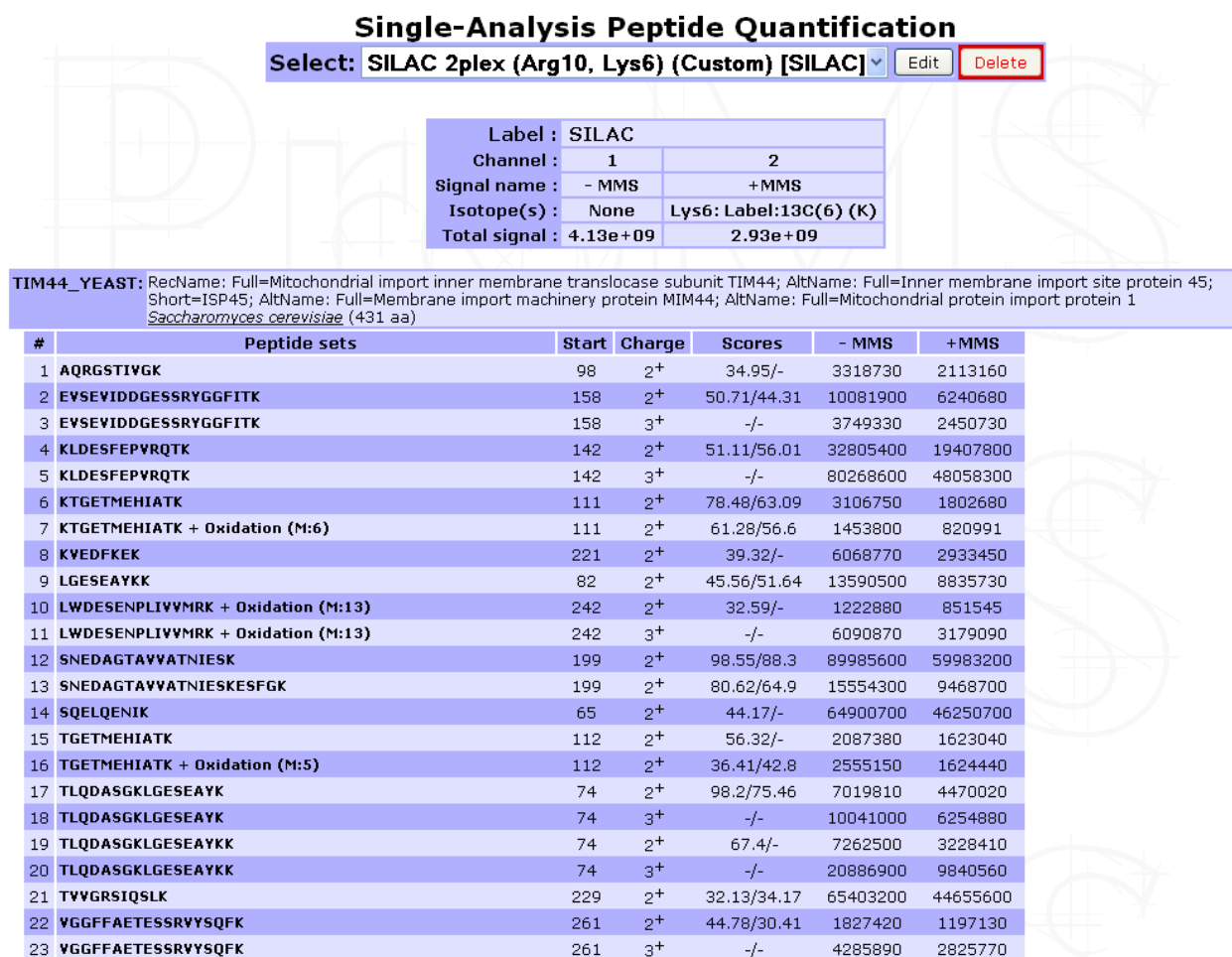


Fig. 1: Legend: Case of a direct extraction of SILAC-labeled peptide XIC from a search result file

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	

Fig. 2: Legend: Case of a MassChroQ extraction with alignment of multiple Analyses

17.1 Absolute abundance quantification

17.1.1 emPAI (label-free)

The **Exponentially Modified Protein Abundance Index** (emPAI) is a spectral-count method that estimates the relative quantitation of proteins in a complex mixture¹ 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.

17.1.2 SIn (label-free)

The **Spectral Index Normalized** (SIn) is a normalized label-free quantitative method which combines three abundance features : peptide and spectral count with fragment-ion (MS/MS) intensity² . 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.

17.1.3 MaxQuant: Intensity, LFQ, iBAQ

17.1.4 Proteomic Ruler

The Proteomic Ruler yields absolute quantification values such as protein copy number per cell, concentration or mass of protein per cell but it can also give abundance values or ranks of abundance in the sample. The idea is that the quantity of a specific protein compared to all the proteins in the sample is proportional to the protein intensity, measured by MS, compared to the total MS signal.

¹ emPAI : Hishima et al, Mol Cell Proteomics, 2005

² SIn : Griffin NM et al, Nat Biotechnol., 2009

Start Proteomic Ruler Quantification from Design MaxQuant [2019/06/20 17:55:52]

Name:	Name of quantification	
Proteomic Ruler:	-- Select --	
Normalization mode:	All samples separately	
MaxQuant metric:	-- Select --	
Select MaxQuant metric to display available quantifications		
Total cellular protein concentration (g/L):	200	
Protein selection:	Exclude	proteins from List: -- Select --
Organism name:	-- Select --	
Desired output(s):	Copy number per cell Concentration [nM] Mass per cell [pg] Mass Abundance (mass/total mass)	
<div>Launch Quantification</div> <div>Cancel</div>		

Thus, the computation of quantification values is based on previous measures of intensities, typically Intensity or LFQ values as output by MaxQuant (see *MaxQuant: Intensity, LFQ, iBAQ* above). For this reason, the Proteomic Ruler is available only from a MaxQuant design, with quantifications already computed.

test_dataset_LSMP_Mouse > MaxQuant [2019/06/20 17:55:52]

Summary	Delete	Add Quantification	Monitor Quantification(s)	Export Quantifications
Edit	Proteomic Ruler		Compare Quantifications	

Proteomic Ruler Type : You can choose between three different methods of Proteomic Ruler : the Total Protein Approach, the Histone Proteomic Ruler or the Custom Proteins Ruler.

- **Total Protein Approach** : This is based on the work of Wiśniewski *et al.*, 2012³. The copy numbers per cell are determined by comparison of the individual intensities with the sum of all the MS intensities, then dividing this ratio by the molecular weight and multiplying by the Avogadro constant and by the total protein content (mass) of a single cell. Therefore you must specify this total protein content per cell as input. The formula is then :

$$\text{Protein copy number per cell} = \text{Total amount of proteins per cell} \cdot \frac{N_A}{M} \cdot \frac{\text{Protein MS intensity}}{\text{Total MS intensity}}$$

- **Histone Proteomic Ruler** : Wiśniewski and colleagues published this as an extension of their *Total Protein Approach* in 2014⁴. Instead of relying on a user-given total amount of proteins per cell (most probably calculated from the total amount of proteins in the sample and the number of cells in the sample), the *Histone Proteomic Ruler* is based on the **Histone MS signal** and on the **ploidy** of the studied cells.

In their paper, Wiśniewski *et al.* show that the **total protein mass per cell** can be estimated from the **cellular DNA mass** and the ratio of the **Histone MS signal** to the **Total MS signal**, following the formula :

$$\text{Cellular protein mass} = \text{Cellular DNA mass} \cdot \frac{\text{Total MS signal}}{\text{Histone MS signal}}$$

For a specific cell type, we can obtain the **amount of DNA per cell** from the **genome size** and the **ploidy** of the cell type. For example, for a diploid human cell, considering a genome size of $3.2 \cdot 10^9 \text{bp}$ and an average base pair weight of 615.8771Da , the expected amount of DNA is approximatively 6.5pg/cell .

Once the **total protein mass per cell** has been computed, the estimation of absolute copy numbers per cell is then obtained in the same way as for the *Total Protein Approach* (see above).

³ Total Protein Approach : Wiśniewski *et al.*, Mol Syst Biol., 2012

⁴ Histone Proteomic Ruler : Wiśniewski *et al.*, Mol Cell Proteomics, 2014

Warning: Because this method relies heavily on the histone content of the sample(s), Wiśniewski *et al.* state in their paper that for accurate results, “the only prerequisite is a eukaryotic, whole-cell proteome dataset where the chromatin fraction is not over- or underrepresented as a result of sample handling” and that “a reasonable depth of proteomic analysis is needed to ensure a robust contribution of the histone MS signal” (depth of around 12 000 peptides).

- **Custom Proteins Ruler :** This method follows the same rationale as the two above, namely that the proportion of a given protein in the sample compared to the total protein content can be estimated from the proportion of its MS signal compared to the total MS signal (intensity). The only difference in this case is that instead of using the total protein content per cell or the cellular DNA mass/histone MS signal, the scaling is done with proteins of known quantity (hereafter referred to as **custom proteins**). You can use a spike-in or any other proteins of your sample as custom proteins, as long as you can provide the total quantity per cell of these proteins in each of your samples.

For example, let’s say you want to quantify two samples, 1 and 2, and your custom proteins are protein A, B and C. In each sample you know that you have :

- 100ng of protein A,
- 70ng of protein B,
- 30ng of protein C,

so a total of 200ng of custom proteins in each sample. However, you also have :

- 20 000 cells in sample 1,
- 25 000 cells in sample 2

You will have to provide the software with your custom proteins identifiers (either a list that you have already saved or a text file with one identifier per line) and with the values :

- $\frac{200 \cdot 10^{-9}}{20000} = 10pg$ for sample 1,
- $\frac{200 \cdot 10^{-9}}{25000} = 8pg$ for sample 2

Name:	test_proteomic_ruler	
Proteomic Ruler:	Custom Proteins Ruler	
Normalization mode:	All samples separately	
MaxQuant metric:	Intensity	
Parent quantification(s) (MaxQuant) to use for proteomic ruler	Custom proteins quantities per cell for each sample (pg)	
<input checked="" type="checkbox"/> 1_a [Intensities]	10	
<input checked="" type="checkbox"/> 2_a [Intensities]	8	
<input type="checkbox"/> 1_b [Intensities]		
<input type="checkbox"/> 2_b [Intensities]		
<input type="checkbox"/> 1_c [Intensities]		
<input type="checkbox"/> 2_c [Intensities]		
Custom proteins (Uniprot ACC):	Choose a List: mouse_histones or Upload a file: Choisir un fichier Aucun fichier choisi	
Total cellular protein concentration (g/L):	200	
Protein selection:	Exclude proteins from List: Contaminants LSMP Mouse 2	
Organism name:	Mus musculus	
Desired output(s):	<input type="checkbox"/> Copy number per cell <input type="checkbox"/> Concentration [nM] <input type="checkbox"/> Mass per cell [pg] <input type="checkbox"/> Mass Abundance (mass/total mass)	
<input type="button" value="Launch Quantification"/> <input type="button" value="Cancel"/>		

Normalization mode: In the default mode, each sample is scaled individually according to the method you chose. If you quantify more than one sample at a time, you may want to normalize them in another way. Your different options

are :

- **All samples separately** : The default mode. A normalization factor is computed for each sample and applied to it individually.
- **Same normalization for all** : A normalization factor is computed for each sample, but only the average of all these normalization factors is applied to scale the intensities of every sample.
- **Same normalization within groups** : In an experiment, some samples are often related to some others, like biological replicates for instance. In that case, you may want to group related samples together and scale them according to the group they belong to. A normalization factor is thus computed for every sample and the average normalization factor of each group is retained to scale the data of samples from this group.

This option requires that you later select the groups for each sample you want to quantify (see figure ...)

- **Average all samples** : This is a very particular case where the data (intensities) from all your samples is averaged (actually, the software takes the median intensities) for each protein before any other computation, which creates an virtual “average sample”. The normalization is then done on this average sample. This mode yields a unique quantification value per protein, corresponding to the average sample. It is meaningful only in some particular cases, for example if you quantify only some biological or technical replicates together and want to have a global view on them.

MaxQuant metric : This is the metric computed by MaxQuant on your samples and that you want to use as the basis for this quantification. You can choose between *Intensity* and *LFQ* metrics. The samples for which the chosen metric is available are displayed after selection of the metric.

Total cellular protein concentration : To compute concentration values for each protein, the software calculates the volume of the sample. To do so it needs the protein concentration of your sample in *g/L*.

Protein selection : Use this feature if you want to exclude one of your lists of proteins from the quantification before launching the computation, or, on the contrary, if you want to restrict the considered proteins to a specific subset. For example, you can exclude contaminants. You need to previously create your own list of proteins (see [User-defined protein lists](#)) to use this feature.

Warning: You should not exclude proteins that are not contaminants because the Proteomic Ruler relies on the total MS signal to quantify the proteins. If you exclude viable proteins that were in your samples, your results

will not be accurate and some features such as *Mass* or *Molecular abundance* will simply become meaningless. Do not exclude proteins only because you don't need to quantify them.

Organism name : Provide the name of the organism from which your proteins are from (used mostly with the Histone Proteomic Ruler method).

Desired output : Select at least one type of quantification value that you are interested in. You can select multiple outputs by maintaining the mouse button clicked during the selection or by clicking on multiple features while holding the Ctrl key on your keyboard. The quantification types available are Copy number per cell, Concentration, Mass per cell, Mass Abundance, Molecular Abundance, Copy number rank and Relative copy number rank.

Note: The *Copy number per cell* is the basis to compute all the other quantification values, so it will be computed anyway. We suggest that you select it even if that is not the main feature you are interested in.

17.1.5 Displaying single abundance quantification data

17.2 Relative abundance quantification

17.2.1 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.

Protein Quantification based on SILAC-labeled Peptides from Analyses in Sample **Detection2**

Name : SILAC-based protein ratios

***Labeled states :** #1: WT #2: Mutant

Peptide selection : Specificity: Proteotypic Missed cleav.: Allowed PTMs: Not allowed Charges: All °Sources: All

Quantification settings : -Bias correction: Scale normalization
☐ Avoid infinite ratios whenever possible (Always true if more than 2 states selected).
Advanced settings:
 -Variation coefficient threshold between replicates: Auto
 (Ignored if no replicates)
☒ FDR control to 5 % Method: Benjamini-Hochberg
 -p-value threshold for outlier detection: 0.05
 -Alternative hypothesis for comparison: Two-sided
 -Confidence interval on protein abundance: 0.95 (0-1)

*Each State will be used as reference for all following States

<input type="checkbox"/>	Analysis	MS type & File	Labeling method	Instrument	Search file & Engine	Databank(s) Taxonomy	Min. score	Max. rank	Selected proteins
<input checked="" type="checkbox"/>	F4628MT	MS/MS F4628MT.RAW	SILAC 2plex (Arg10, Lys8) (Custom)	ESI-TRAP	F4628MT_2.pdm MASCOT	NCBI-Mascot All entries	20	1	878 (1007)

Launch Quantification Cancel

- **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 applied 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 than 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.

17.2.2 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

17.2.3 Displaying relative abundance quantification data

17.3 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**.

17.3.1 TnPQ

Silva et al. showed in their work⁵ 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. Then, the **Top 3 Protein Quantification**⁶ 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.

**Select Analyses in Design Test_Design
for Protein-Ratio Quantification**

Name :

Algorithm :

***States :** #1: #2:

Labeling method :

Charge states : Specificity: Missed cleav.: PTMs: Charges: Sources:

Quantification settings :

•Advanced settings:

- Variation coefficient threshold between replicates: (Ignored if no replicates)
- ☒ FDR control to % Method:
- p-value threshold for outlier detection:
- Alternative hypothesis for comparison:
- Confidence interval on protein abundance: (0-1)

*Each State will be used as reference for all following States

Steps involved in TnPQ computation:

1. Retrieval of all available XICs (area) of each peptide of the protein for all conditions
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

Warning: 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

3. If a bias correction setting was selected (scale or reference protein normalization), a normalization step is introduced by computing bias estimates on unique peptides⁷. All XIC are divided by those bias factors.

⁵ TnPQ : Silva et al, Mol Cell Proteomics, 2006

⁶ T3PQ : Grossmann et al, J Proteomics, 2010

⁷ TnPQ bias correction (scale normalization part) : Yang et al. 2002

Note: If None was chosen, nothing is done to the data

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.
5. Compute for each protein the geometrical mean of peptide XICs
6. Quality control of the data (normality test on the data and variance sameness)
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
8. If chosen, adjust p-values to control FDR level

17.4 Comparing multiple protein quantifications

17.5 Exporting multiple quantifications

Export Multiple Quantifications From Current Experiment

Focus : Phospho-proteome

Features : Protein quantifications **Quantification type:** Protein ratio

Data transform : LOG2 **Gene Name:** ☒

Data filtering :

Abs. fold change \geq 1 in at least 1 quantification

Infinite ratios: Allow 25% per protein

p-value \leq 1 in at least 1 quantification (Does not apply to normalized ratios)

Peptides: All \geq 1

Missing values: Allow all per protein

Exclude ambiguous sites: ☒

Aggregate (by best modification sites): ☒

Protein filtering :

Custom list: Restrict to All proteins

☒ Keep only the: 200 most changing proteins between groups

☒ p-value \leq 0.05

Data selection : ☐ Auto-extend selection

Quantifications (7 selected)

Copy selection from: - Select -

Annotation for: MB Group

Design-based quantifications:

<input checked="" type="checkbox"/>	Gr1vsGr2vsGr3vsGr4 Phospho > 1x5v2x7v3x5v4x11-P : Gr1/MixSILAC	1
<input checked="" type="checkbox"/>	Gr2/MixSILAC	2
<input checked="" type="checkbox"/>	Gr3/MixSILAC	3
<input checked="" type="checkbox"/>	Gr4/MixSILAC	4
<input checked="" type="checkbox"/>	Gr2°/Gr1°	2
<input checked="" type="checkbox"/>	Gr3°/Gr1°	3
<input checked="" type="checkbox"/>	Gr4°/Gr1°	4
<input type="checkbox"/>	Gr3°/Gr2°	3
<input type="checkbox"/>	Gr4°/Gr2°	4

Export dataset

- For the explanation of all features see section “Exploratory analyses” except for the following item:
 - Gene Name : transform all protein name in gene name (<which?geneID/unigene>).

- Keep most changing proteins between:
 - * Sample: For a given isoform (or protein), a standard deviation is calculated for all selected sample and only the isoform (or protein) with the best value is kept. The standard deviation is a measure that is used to quantify the amount of variation or dispersion of a set of data values.
 - * Group: For a given isoform (or protein), an Anova is applied, a p-value is calculated between group. Analysis of variance (ANOVA) is a collection of statistical models and their associated procedures (such as “variation” among and between groups) used to analyze the differences among group.
- After clicking on `export dataset` button, a compressed directory is downloaded which contains following file :
 - `R_parameter.txt`: all the parameters used by the statistical analysis.

MODIF	QUANTIF	FAM	EXCLUDE	AMB	GENE_NAME	PROTEIN_SELECTION	AGREGATE	KEEP_PROT	KEEP_PROT_NB	P_VALUE_CHK	P_VALUE	MISSING_VALUE
0	RATIO		FALSE		FALSE	sample	FALSE	FALSE	FALSE	FALSE	FALSE	3

- `parameter.txt`: all the parameters used for data filtering.

FOCUS	Proteins
QUANTIF_METHOD	RATIO
DATA_TRANSFORM=LOG2	
%_MISSING_VALUES_ALLOWED	100
LIST_FILTERING	None
MIN_NUM_PEPTIDES	1
PEPTIDE_TYPE	All
PROTEIN_SELECTION	sample
NORMALIZED_RATIOS	No
MAX_ABS_RATIO	1
MIN_RATIO_OCCURENCE	1
MINUS_INFINITE_RATIO	-1000
PLUS_INFINITE_RATIO	1000
MAX_P_VALUE	1
MIN_P_VALUE_OCCURENCE	1
%_INFINITE_RATIOS_ALLOWED	NA

- `matrix_pvalue_processed.txt`:

	A	B	C
CNOT1_HUMAN	1.68310691294029e-10	0.000671000753688762	0.161346999591561
FTO_HUMAN	7.91721274188521e-14	4.84808016967787e-15	0.468731319378907
ANKY2_HUMAN	0.972626794388628	0.0692130183589147	0.38577078788442
NUFP2_HUMAN	1.4055851718406e-06	0.00490857128359717	0.187769563851311

- `matrix_pep_processed.txt`: peptide number

	A	B	C	D	E	F	G
K1211_HUMAN	48	48	48	48	50	48	82
PCTL_HUMAN	1	1	1	1	3	1	4
ANS1A_HUMAN	26	26	26	26	60	26	82
TPD52_HUMAN	40	40	40	40	57	40	89
DNJC2_HUMAN	32	32	32	32	55	32	92

- `matrix_log2ratio_processed.txt`:

	A	B	C	D
NFX1_HUMAN-150-154:2/3	-3.43329889213729	0.905873070751604	NA	NA
SRGP1_HUMAN-835-844:1/3	NA	-0.26539086979225	NA	-4.12062296229416
MYCPP_HUMAN-1093-1110:2/6	NA	NA	-1.97355334485319	-0.446925819092622
KAT7_HUMAN-45-57:1/7	0.0887578489641253	NA	NA	2.90026647997175
ZFAN5_HUMAN-48-65:1/9	NA	0.17152266067652	NA	-0.375015614652475
VCIP1_HUMAN-756-772:1/9	0.551334862744437	NA	-0.719128633768636	-1.11283949405069
RAB3A_HUMAN-S190	NA	1.55839106219827	NA	-2.0239219516308

– annotation_processed.txt:

PROTEIN	GENE	SYNONYMS	DESCRIPTION
1433B_HUMAN	YWHA8		14-3-3 protein beta/alpha
1433E_HUMAN	YWHA8		14-3-3 protein epsilon
1433F_HUMAN	YWHAH	YWHA1	14-3-3 protein eta
1433G_HUMAN	YWHA8		14-3-3 protein gamma
1433S_HUMAN	SFN	HME1	14-3-3 protein sigma
1433T_HUMAN	YWHAQ		14-3-3 protein theta
1433Z_HUMAN	YWHAZ		14-3-3 protein zeta/delta
1A02_HUMAN	HLA-A	HLAA	HLA class I histocompatibility antigen, A-2 alpha chain
1A24_HUMAN	HLA-A	HLAA	HLA class I histocompatibility antigen, A-24 alpha chain
1A69_HUMAN	HLA-A	HLAA	HLA class I histocompatibility antigen, A-69 alpha chain
1B57_HUMAN	HLA-B	HLAB	HLA class I histocompatibility antigen, B-57 alpha chain
2A5A_HUMAN	PPP2R5A		Serine/threonine-protein phosphatase 2A 56 kDa regulatory subunit alpha isoform
2A5D_HUMAN	PPP2R5D		Serine/threonine-protein phosphatase 2A 56 kDa regulatory subunit delta isoform
2A5E_HUMAN	PPP2R5E		Serine/threonine-protein phosphatase 2A 56 kDa regulatory subunit epsilon isoform
2A5G_HUMAN	PPP2R5C	KIAA0044	Serine/threonine-protein phosphatase 2A 56 kDa regulatory subunit gamma isoform
2AAA_HUMAN	PPP2R1A		Serine/threonine-protein phosphatase 2A 65 kDa regulatory subunit A alpha isoform
2AAB_HUMAN	PPP2R1B		Serine/threonine-protein phosphatase 2A 65 kDa regulatory subunit A beta isoform
2ABA_HUMAN	PPP2R2A		Serine/threonine-protein phosphatase 2A 55 kDa regulatory subunit B alpha isoform
2ABD_HUMAN	PPP2R2D	KIAA1541	Serine/threonine-protein phosphatase 2A 55 kDa regulatory subunit B delta isoform
2B11_HUMAN	HLA-DRB1		HLA class II histocompatibility antigen, DRB1-1 beta chain

– sd.txt: standard deviation value (between sample)

GENE	PROTEIN	SD
CCDC94	CCD94_HUMAN	0.00280133968209144
PGAM1	PGAM1_HUMAN	0.00836298300070294
RAP1A	RAP1A_HUMAN	0.00915012652433019
API5	API5_HUMAN	0.0104967492955313
NXF1	NXF1_HUMAN	0.0113288413640802
CALCOCO2	CACO2_HUMAN	0.0116156741249492
RAN	RAN_HUMAN	0.015262736317894
POLR2D	RPB4_HUMAN	0.0170700841516235
SPECC1L	CYTSA_HUMAN	0.0174741988686923
BOLA2	BOLA2_HUMAN	0.0177524842910831
VPRBP	VPRBP_HUMAN	0.0218941806096138
ARL2	ARL2_HUMAN	0.0223736408867747
YWHA8	1433B_HUMAN	0.0231212217113701
ARL5A	ARL5A_HUMAN	0.0243350931504144

CHAPTER 18

PTMs quantification

18.1 Set PTMs relevance to project

18.2 Display modification sites distribution

18.3 Compare PTMs between projects

18.4 Quantify modification sites

Exploratory analysis

In statistics, **exploratory data analysis (EDA)** is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Two types of exploratory analyses are available in *myProMS* : **Principal Component Analysis (PCA)** and **Clustering**. They can be launch on **protein quantification** or **peptide count/abundance** datasets.

19.1 Launching exploratory analyses on protein quantification

First select the Experiment containing the dataset to analyze. Then click on the `Start Exploratory Analyses` button. The following form will be displayed:

Start Principal Component and/or Clustering Analyses

☒ PCA name :
☒ Clustering name :
 Focus :
 Features : Quantification type:
 Data transform :
 Data filtering :
 Abs. fold change \geq in at least quantification
 Infinite ratios: per protein
 p-value \leq in at least quantification (Does not apply to normalized ratios)
 Peptides: \geq
 Missing values: per protein
 Custom list:
 Exclude ambiguous sites: ☐
 Protein selection:
 Clustering parameters : Method: Metric:
 Data selection : ☐ Auto-extend selection
 Quantifications (0 selected)
 Copy selection from:
 Design-based quantifications:
 ☐ Gr1vsGr2vsGr3vsGr4 Phospho > 1x5v2x7v3x5v4x11-P : Gr1/MixSILAC
 ☐ Gr2/MixSILAC
 ☐ Gr3/MixSILAC
 ☐ Gr4/MixSILAC
 ☐ Gr2°/Gr1°
 ☐ Gr3°/Gr1°
 ☐ Gr4°/Gr1°
 ☐ Gr3°/Gr2°
 ☐ Gr4°/Gr2°

- **Name:** Provide a name for the PCA and/or the clustering analysis. The analysis is saved and can be retrieved by this name in the Exploratory analyses tree displayed in the sub-navigation frame.
- **Focus:** 3 types of data can be analyze, Proteins, Proteins modifications (phosphorylation , acetylation...) and Normalized proteins modifications.
- **Features:** run exploratory analysis on Protein quantifications or Peptide count.
- **Quantification type:** Select one the protein quantification methods available as listed below:

Quantification type for protein quantification	
emPAI	emPAI
	emPAI (Mol %)
	emPAI (Mr %)
MaxQuant Intensities	Intensities
	iBAQ
	LFQ
	MS/MS count
Normalized spectral index	SIN
Protein ratio	Protein ratio

- **Data transform:** by default data are not logged, user can transform data in log2 or log10. This is critical, in particular for protein ratio in order to make ratio data symmetrical and centered on 1 (0 in log mode)
- **Data filtering:** - Abs. fold change in at least n quantifications (only for protein ratio): Only proteins with an absolute fold change (before log) greater than or equal to the value provided found in at least n quantifications

will be kept . - Infinite ratios (only for protein ratio): choose how to deal with infinite ratios. - p-value in at least n quantifications (only for protein ratio): Only proteins with p-value lesser than or equal to the value provided found in n quantifications will be kept. - Peptides: Select the counting method: All, Distinct, Razor, ... (the exact list of available options depends of the quantification method selected) and the minimum value allowed in a quantification to consider the protein to be present or absent. - Missing values: choose how to deal with missing values. Different options are available: None allowed, Allow x%, Allow all. Value is per individual proteins across dataset. Proteins with more missing values than allowed will be excluded. - Exclude ambiguous site (only for modifications proteins): exclude or not sites with ambiguous (uncertain) modification position (ambiguous site contains “~” eg, xxx-254~262:1/5). - Aggregate (by best modifications sites, only for modifications proteins): keep only 1 modification site per protein (gene). The site with the highest variance across the dataset will be kept as a proxy for the whole protein (gene).

- **Protein filtering:** - Custom list: by default all proteins that passed above filters are used for the analysis. It is also possible to focus only or exclude a set of proteins based on the custom list selected. - Keep n most changing proteins: Only the n proteins with highest variance across the dataset will be kept.
- **Clustering parameters:** choose the method and the metric parameters (default method is “Ward” and metric is “pearson”)
- **Data selection:** choose the quantifications to analyze. - Auto-extend selection: after selecting a quantification, all the following ones are automatically selected (or unselected). - Copy selection: copy the dataset used in a previously performed exploratory analysis.

19.2 Launching exploratory analyses on peptides count

PCA name :

Clustering name :

Features : Quantification type: Grouping method:

Missing values : per protein

Clustering parameters : Method: Metric:

Data selection :

Samples/Analyses (0 selected)	Group by: <input type="text" value="samples"/>
<input type="checkbox"/> MB5-Phospho > C1337FD	MB5-Phospho
<input type="checkbox"/> MB5_2-Phospho > C1489FD	MB5_2-Phospho
<input type="checkbox"/> C1490FD	MB5_2-Phospho
<input type="checkbox"/> C1491FD	MB5_2-Phospho
<input type="checkbox"/> C1492FD	MB5_2-Phospho
<input type="checkbox"/> C1493FD	MB5_2-Phospho
<input type="checkbox"/> C1494FD	MB5_2-Phospho
<input type="checkbox"/> C1495FD	MB5_2-Phospho
<input type="checkbox"/> C1496FD	MB5_2-Phospho
<input type="checkbox"/> C1497FD	MB5_2-Phospho
<input type="checkbox"/> C1498FD	MB5_2-Phospho

- **Name:** Provide a name for the PCA and/or the clustering analysis. The analysis is saved and can be retrieved by this name in the Exploratory analyses tree displayed in the sub-navigation frame.
- **Features:** run exploratory analysis on protein quantifications or peptide count
- **Quantification type:** different types of quantification are available.
- **Peptide ID:** <to be completed>

- **Peptide Count:** <to be completed>
- **Spectral Count:** <to be completed>
- **Peptide XIC:** <to be completed>
- **Grouping method:** 4 methods are available (except for peptide ID).
- **Min.** <to be completed>
- **Max.** <to be completed>
- **Mean** <to be completed>
- **Sum** <to be completed>
- **Missing values:** choose how to deal with missing values. Different options are available: “None allowed”, “Allow x%”, “Allow all”.
- **Clustering parameters:** choose the method and the metric parameter (by default, method is “Ward” and metric is “pearson”).

19.3 Principal Component Analysis (PCA)

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated **variables called principal components**.

The number of distinct principal components is equal to the smaller of the number of original variables or the number of observations minus one. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components.

The resulting vectors are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables.

19.3.1 Summary / edit / delete

PCA PCA1

PCA name :	PCA1
Description :	
Focus :	Normalized Phospho-Proteins
Feature :	Protein quantifications : RATIO [Protein ratio]
Data transform :	LOG2
Data filtering :	<ul style="list-style-type: none"> Abs. fold change ≥ 1 in at least 1 quantification(s) Infinite ratios are treated as missing values All peptides ≥ 3 34% missing values allowed / protein 0.58% missing values (58 values imputed) 500 proteins used
Status :	✓ Finished
Quantifications used :	<div>Hide list</div> <p>Normalization:</p> <ol style="list-style-type: none"> Gr1vsGr2vsGr3vsGr4 Phospho > 1x5v2x10v3x9v4x12-P : Gr1/MixSILAC - Gr1vsGr2vsGr3vsGr4 > 1x5v2x10v3x9v4x12 : Gr1/MixSILAC Gr1vsGr2vsGr3vsGr4 Phospho > 1x5v2x10v3x9v4x12-P : Gr4°/Gr3° - Gr1vsGr2vsGr3vsGr4 > 1x5v2x10v3x9v4x12 : Gr4°/Gr3° Gr1vsGr2vsGr3vsGr4 Phospho > 1x5v2x10v3x9v4x12-P : Gr2/MixSILAC - Gr1vsGr2vsGr3vsGr4 > 1x5v2x10v3x9v4x12 : Gr2/MixSILAC Gr1vsGr2vsGr3vsGr4 Phospho > 1x5v2x10v3x9v4x12-P : Gr3/MixSILAC - Gr1vsGr2vsGr3vsGr4 > 1x5v2x7v3x5v4x11 : Gr3/MixSILAC Gr1vsGr2vsGr3vsGr4 Phospho > 1x5v2x10v3x9v4x12-P : Gr4/MixSILAC - Gr1vsGr2vsGr3vsGr4 > 1x5v2x7v3x5v4x11 : Gr4/MixSILAC Gr1vsGr2vsGr3vsGr4 Phospho > 1x5v2x10v3x9v4x12-P : Gr2°/Gr1° - Gr1vsGr2vsGr3vsGr4 > 1x5v2x10v3x9v4x12 : Gr2°/Gr1° Gr1vsGr2vsGr3vsGr4 Phospho > 1x5v2x10v3x9v4x12-P : Gr3°/Gr1° - Gr1vsGr2vsGr3vsGr4 > 1x5v2x10v3x9v4x12 : Gr3°/Gr1° Gr1vsGr2vsGr3vsGr4 Phospho > 1x5v2x10v3x9v4x12-P : Gr4°/Gr1° - Gr1vsGr2vsGr3vsGr4 > 1x5v2x7v3x5v4x11 : Gr4°/Gr1° Gr1vsGr2vsGr3vsGr4 Phospho > 1x5v2x10v3x9v4x12-P : Gr3°/Gr2° - Gr1vsGr2vsGr3vsGr4 > 1x5v2x10v3x9v4x12 : Gr3°/Gr2° Gr1vsGr2vsGr3vsGr4 Phospho > 1x5v2x10v3x9v4x12-P : Gr4°/Gr2° - Gr1vsGr2vsGr3vsGr4 > 1x5v2x10v3x9v4x12 : Gr4°/Gr2° <p>Test:</p>

List of graphical highlights used:

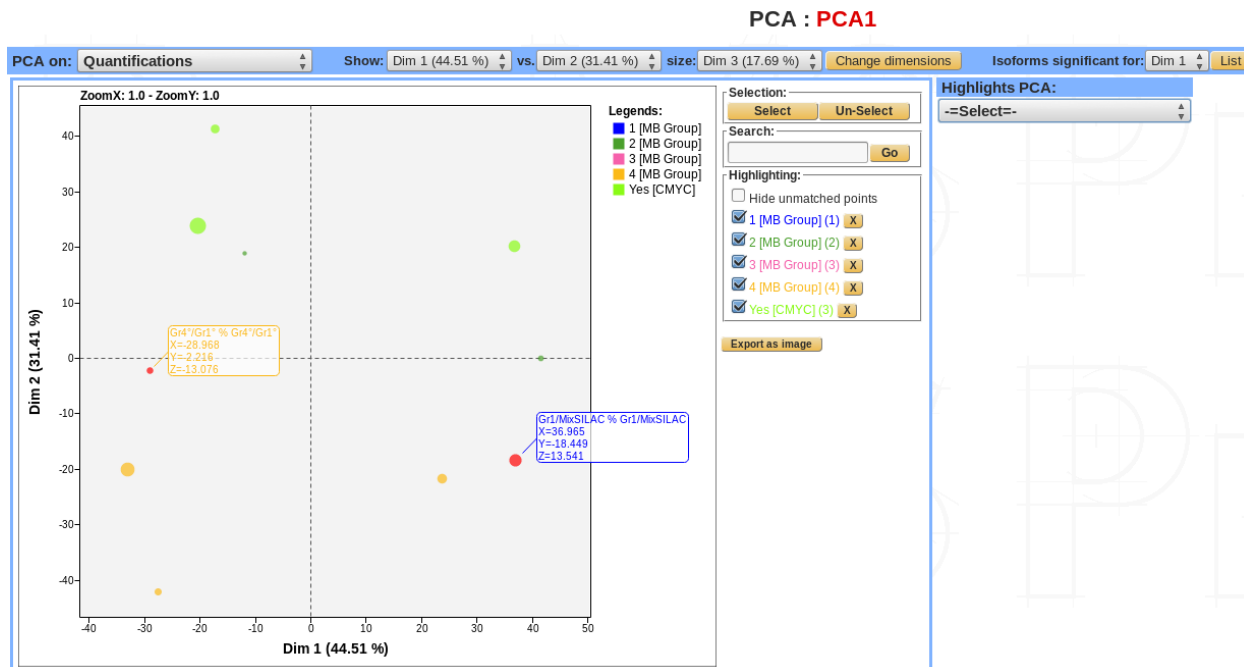
Biosample property highlights for quantifications:
MB Group = "1"
MB Group = "2"
MB Group = "3"
MB Group = "4"
CMYC = "Yes"

If a PCA analysis is selected, a summary of the information available for that analysis is displayed in the mainframe.

- **PCA name**
- **Description**
- **Focus**
- **Features**
- **Data transform**
- **Data Filtering**
- **Status:** There are 3 possible status: Ongoing, Finished or Error
- **Quantifications used:** list of quantifications used for the analysis (or name of the list)
- List of graphical highlight used (see below).

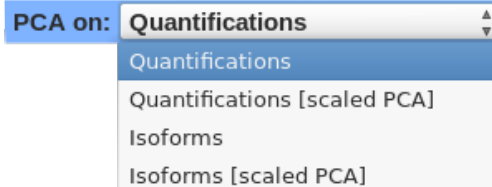
The PCA name and the description can be modify by clicking the **Edit** button. The analysis can be deleted by clicking the **Delete** button.

19.3.2 Displaying a PCA



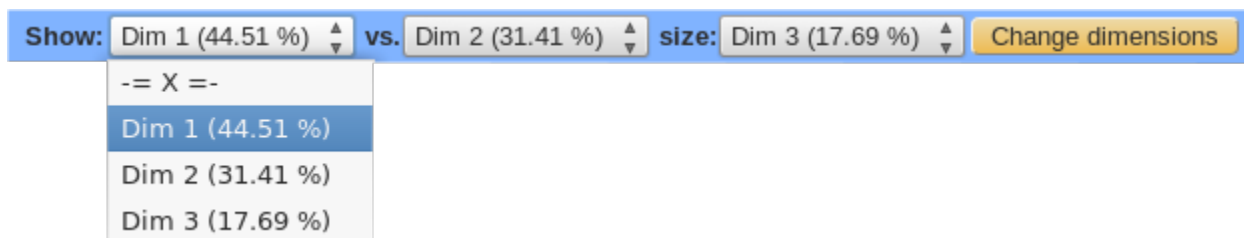
PCA can be viewed with 4 different options:

- Quantifications view.
- Quantifications scaled view.
- Protein (or isoforms for modification sites) view.
- Protein scaled (or isoforms scaled if modifications sites).



The PCA dimensions can be changed, a 2.1D view is available by selecting a third dimension. In this case, the points' size becomes proportional to the values of the selected dimension.

Danger: Add the new PCA 3D display



For each dimension the list of proteins (or isoforms) can be displayed.

List of significant isoforms for dimension 1

☒ Auto-extend selection

251 isoforms	Gene Name	Correlation	p-value	MW ^{kd}	Description - Species
<input type="checkbox"/> ARFP1_HUMAN-S132	ARFIP1	-0.995	3.6e-09	41.7	Arfaptin-1 <i>Homo sapiens</i>
<input type="checkbox"/> SAFB1_HUMAN-S601.S604	SAFB	-0.993	1.3e-08	102.6	Scaffold attachment factor B1 <i>Homo sapiens</i>
<input type="checkbox"/> CLAP2_HUMAN-S360	CLASP2	-0.992	2.0e-08	141.1	CLIP-associating protein 2 <i>Homo sapiens</i>
<input type="checkbox"/> ZMYM4_HUMAN-T107.S110	ZMYM4	-0.990	4.4e-08	172.8	Zinc finger MYM-type protein 4 <i>Homo sapiens</i>
<input type="checkbox"/> AMPD2_HUMAN-S190	AMPD2	0.988	9.1e-08	100.7	AMP deaminase 2 <i>Homo sapiens</i>
<input type="checkbox"/> NCAM1_HUMAN-S784	NCAM1	0.987	1.4e-07	94.6	Neural cell adhesion molecule 1 <i>Homo sapiens</i>
<input type="checkbox"/> MTAP2_HUMAN-S1347.S1353	MAP2	0.986	1.5e-07	199.5	Microtubule-associated protein 2 <i>Homo sapiens</i>
<input type="checkbox"/> ANM6_HUMAN-T21	PRMT6	-0.986	1.6e-07	41.9	Protein arginine N-methyltransferase 6 <i>Homo sapiens</i>
<input type="checkbox"/> KI67_HUMAN-T1327	MKI67	-0.979	7.6e-07	358.7	Antigen KI-67 <i>Homo sapiens</i>

Danger: <annotation highlight <to be completed>>

19.4 2D-Clustering

Cluster analysis or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).

19.4.1 Summary / edit / delete

The summary part of clustering is exactly the same than the PCA summary plus the information of clustering parameters.

20.1 Gene Ontology

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.

20.1.1 GO summary

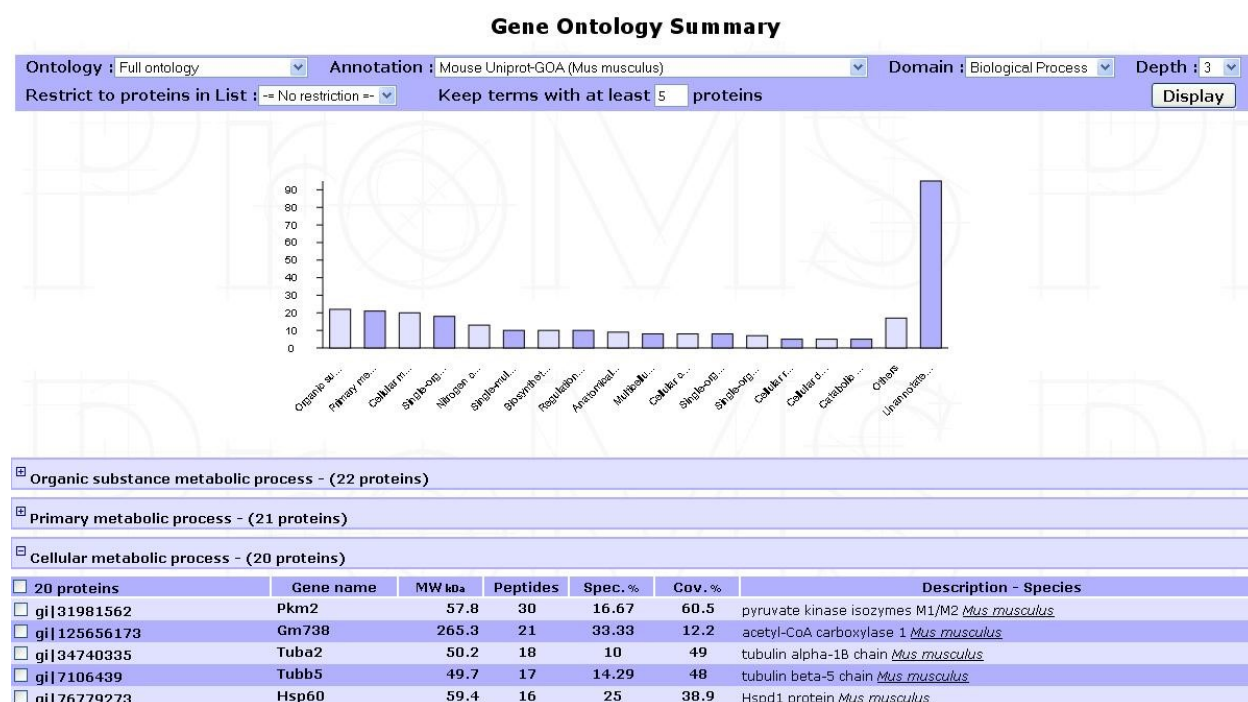
The GO summary tool can be used to simply regroup proteins sharing common GO terms. This tool can be run from the option frame on any project item, by clicking on the `Gene Ontology summary` button. The following form is then displayed:

Gene Ontology Summary

Ontology : -- Select an ontology file --	Annotation : -- Select an annotation file --	Domain : Biological Process	Depth : 2
Restrict to proteins in List : -- No restriction --	Keep terms with at least 5 proteins	<input type="button" value="Display"/>	

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 (Optional)	If a selected term contains less proteins than this specified value, this term will be ignored and the matching proteins will be added to the “Other” category

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.

20.1.2 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 perl¹. 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 clicking on an experiment and selecting the **Start GO Analysis** button in the option frame. The following form is then displayed:

¹ GO::TermFinder package : Boyle et al., Bioinformatics, 2004

Close Project

Expand Collapse

Demo project

- Protein Identification
- Protein Labeled Quantification**
- SILAC
- iTRAQ
- Protein Label-free Quantification
- Exp. with gels
- Peptide Mass Fingerprint

Sample(s)
Add Design
Process Analyses
List Proteins
Compare Project Items
Gene Ontology Summary
Start GO Analysis

Add 2D-Gel
Export Proteins
Compare Quantifications
Start Q. GO Analysis

Gene Ontology Enrichment Analysis

Name :

Description :

Ontology File : Full ontology (http://ftp.geneontology.org/pub/go/ontology/obo_format_1_2/gene_ontology.1_2.obo)

Annotation : Human Uniprot-GOA - Homo sapiens (Human)

Domain(s) : ☒ Biological Process ☒ Cellular Component ☒ Molecular Function

Advanced Parameters : Estimated number of proteins in organism: (default is number of annotated proteins in organism)

Background population: ☒ Select a List: -- Select -- ☐ Upload a local file: Aucun fichier sélectionné.

Statistical settings:

☒ Control FDR at % with method

☐ Use a p-value threshold: with ☒ Bonferroni correction

☐ Show non-significant terms in graphical view

☐ Include only proteins containing at least peptide(s)

Select a protein set from:

☐ a List

-- Select --

☒ Project items

Expand Collapse

Protein Labeled Quantification

View : -- Select --

view selected.

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	
Domain(s)	Select one or more domains to test	
Advanced parameters (Optional)	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.

Protein Identification > GO Enrichment analysis

Summary Biological Process **Cellular Component** Molecular Function

Cellular Component

Table View Graphical View

Cell • Cell part • Cortical actin cytoskeleton • Cytoplasm • Cytoplasmic part • Cytoskeletal part • Cytosol • Cytosolic large ribosomal subunit • Cytosolic part • Cytosolic ribosome • Cytosolic small ribosomal subunit • Intermediate filament • Intermediate filament cytoskeleton • Intracellular • Intracellular membrane-bounded organelle • Intracellular non-membrane-bounded organelle • Intracellular organelle • Intracellular organelle lumen • Intracellular organelle part • Intracellular part • Keratin filament • Large ribosomal subunit • Macromolecular complex • Membrane-bounded organelle • Membrane-enclosed lumen • Non-membrane-bounded organelle • Nuclear lumen • Nuclear part • Nucleolus • Nucleoplasm • Nucleus • Organelle • Organelle lumen • Organelle part • Prefoldin complex • Ribonucleoprotein complex • Ribosomal subunit • Ribosome • Signal recognition particle • Signal recognition particle, endoplasmic reticulum targeting • Small ribosomal subunit

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

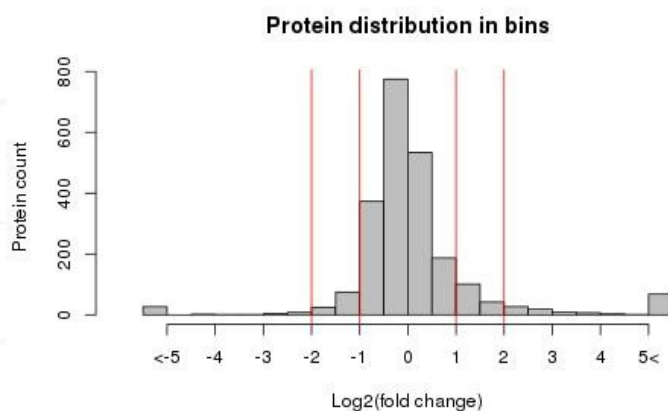
- **Cloud view:** Highly significant terms (low p-value) are represented with a large font, and less significant terms with a small font. The proteins mapped to a term can be listed by clicking on each term.
- **Table view:** More details can be viewed in table format which contains the p-value and enrichment 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.

20.1.3 Quantitative gene enrichment analysis

When a quantification is available, a quantitative gene enrichment analysis can be performed as it was originally done for SILAC experiments². 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:

² Quantitative GO SILAC : Pan C et al, MCP, 2009

Gene Ontology Enrichment Analysis On Quantification Data [?]



Bin proportions (%) 1.9 4.4 81.7 6.2 5.9
 Log ratio thresholds -2 -1 1 2 [Preview](#)

Name:

Description:

Quantification: SAK vs WT

Ratio: SAK-B/WT-B*

Peptides: ☐ Include only proteins containing at least quantified peptide(s)

Protein-ratio p-value threshold: (leave empty for no threshold)

Ontology file: Complete Gene Ontology (gene_ontology.1_2.obo) **Depth:** All terms

Annotation: GOA Human - Homo sapiens (Human)

Domain: ☒ Biological Process ☐ Cellular Component ☐ Molecular Function

Advanced parameters:

Background population: ☒ All **quantified** proteome
☐ All **annotated** proteome
☐ Select a List: -- Select --
☐ Upload a local file: Aucun fichier sélectionné.

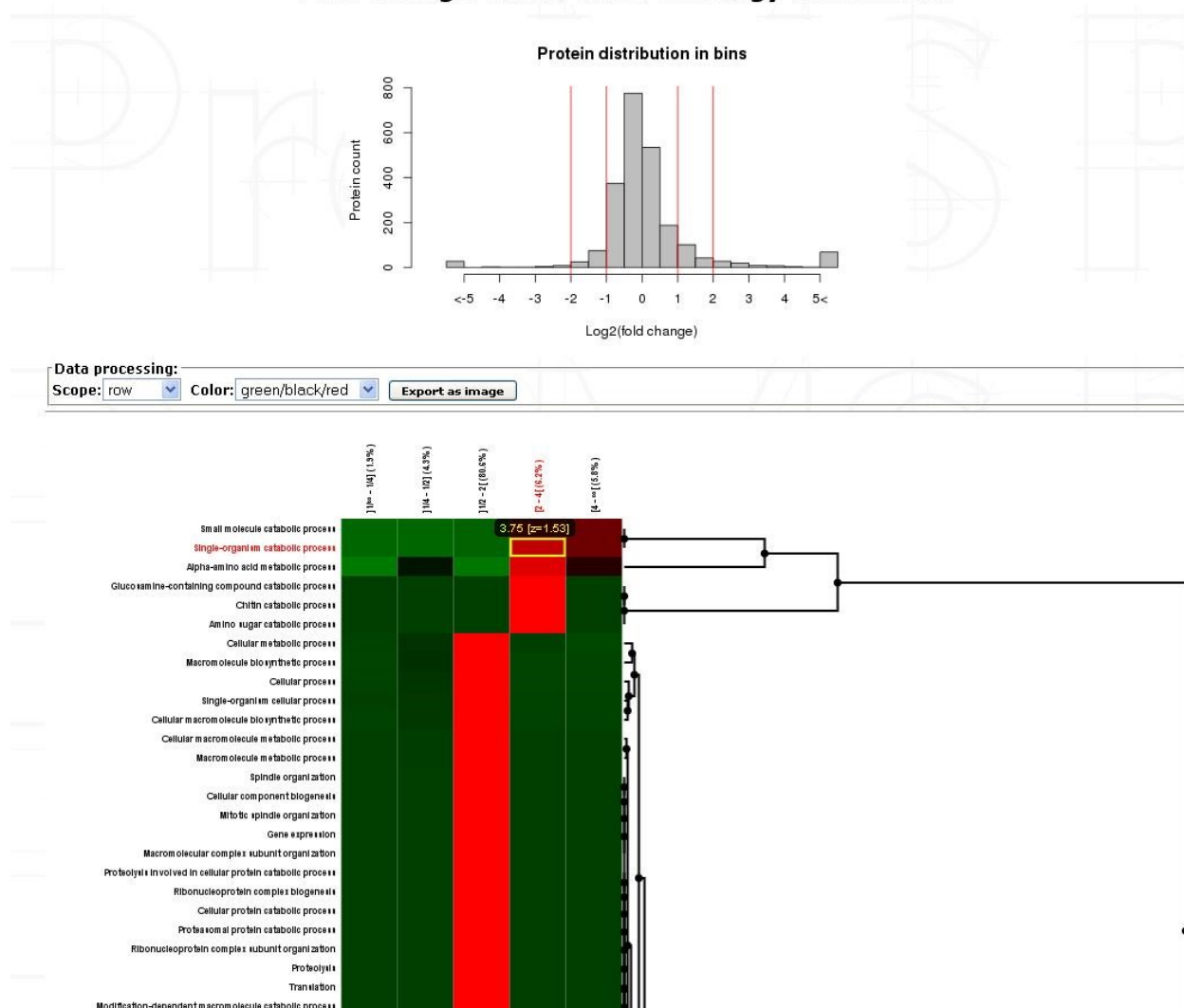
Enrichment test statistical settings:
☒ Control FDR at % with Benjamini & Hochberg method
☐ Use a p-value threshold: with ☒ Bonferroni correction

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 (Optional)	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 (Optional)	Background population	Select the population to which the tested protein set will be compared. See GO enrichment analysis section for custom list recommendations
	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:

Fold Change-based Gene Ontology Enrichment



Each row represent a GO-Term and each cell is the $-\log_{10}$ 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**).

Note: 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.

20.2 Pathway enrichment

Pathway is the term from molecular biology which depicts an artificial simplified model of a process within a cell or tissue. In bioinformatics research, **pathway analysis** is used to identify **related proteins** within a pathway. This is helpful when analyzing any omics dataset with a large number of proteins. By examining the changes in proteins in a pathway, its biological causes can be explored.

Typical pathway model starts with extracellular signaling molecule that activates a specific protein. Thus triggers a chain of protein-protein or protein-small molecule interactions. Pathway analysis helps to understand or interpret omics data from the point of view of canonical prior knowledge structured in the form of pathways diagrams. It allows finding distinct cell processes (cellular processes), diseases or signalling pathway that are statistically associated with selection of differentially expressed proteins between two samples. To do so, *myProms* uses [Reactome's](#) web-service.

20.2.1 Launch pathway analysis

Pathway Enrichment Analysis

Name :
Description :
Advanced Parameters : Statistical settings:
• Use a p-value threshold:

Select a protein set from:
☐ a List

☒ Project Items

☐ Phospho-Tumours

Name	Provide a name for the Pathway analysis. The analysis is saved and can be retrieved by this name in the Functional analyses tree displayed in the sub-navigation frame
Description (Optional)	Free text
Advanced parameters (Optional)	Set a p-value threshold
Select proteins	Select proteins to use in the analysis, from custom list or project item

20.2.2 Summary / edit / delete

Pathway Analysis : testPathway

Analysis name : testPathway
Description :
Feature selection : • FDR=1%
• p-value=0.000001
Status : ✓ Finished
Proteins used : List : Kinases

Analysis performed by Reactome (Milacic M. et al. Cancers 4, 2012, Croft D. et al. Nucleic Acid Research 42, 2013.)

If a pathway analysis is selected, a summary of the information available for that analysis is displayed in the mainframe.

20.2.3 Displaying pathway analysis

3 different views are available : **Cloud**, **graphical** and **table view** :

Cloud view :

Display Pathway Enrichment Analysis for **testPathway2**

Table View Graphical View Export

189/727 unannotated proteins (click for details)

Enrichment factor : < 2 < 4 < 6 < 8 ≥ 8

Signaling by Receptor Tyrosine Kinases . Negative regulation of the PI3K/AKT network . PI3K/AKT Signaling in Cancer . Intracellular signaling by second messengers . PIP3 activates AKT signaling . Regulation of TP53 Activity through Phosphorylation . PI5P, PP2A and IER3 Regulate PI3K/AKT Signaling . Diseases of signal transduction . Axon guidance . Regulation of TP53 Activity . RAF/MAP kinase cascade . VEGFA-VEGFR2 Pathway . Signaling by VEGF . MAPK1/MAPK3 signaling . Toll Like Receptor 4 (TLR4) Cascade . Toll Like Receptor 3 (TLR3) Cascade . TRIF(TICAM1)-mediated TLR4 signaling . MyD88-independent TLR4 cascade . MAPK family signaling cascades . Toll Like Receptor 5 (TLR5) Cascade . Toll Like Receptor 10 (TLR10) Cascade . MyD88 cascade initiated on plasma membrane . Signaling by SCF-KIT . MyD88:Mal cascade initiated on plasma membrane . Toll Like Receptor TLR6:TLR2 Cascade . Toll Like Receptor 2 (TLR2) Cascade . Toll Like Receptor TLR1:TLR2 Cascade . mTOR signalling . Toll-Like Receptors Cascades . TRAF6 mediated induction of NFkB and MAP kinases upon TLR7/8 or 9 activation . MAP kinase activation in TLR cascade . NGF signalling via TRKA from the plasma membrane . Toll Like Receptor 7/8 (TLR7/8) Cascade . MyD88 dependent cascade initiated on endosome . CD28 co-stimulation . Toll Like Receptor 9 (TLR9) Cascade . Constitutive Signaling by Aberrant PI3K in Cancer . EPH-ephrin mediated repulsion of cells . Interleukin-3, 5 and GM-CSF signaling . Interleukin-17 signaling . GPVI-mediated activation cascade . EPH-Ephrin signaling . Transcriptional Regulation by TP53 . Signalling by NGF . EPHA-mediated growth cone collapse . Signaling by ERBB2 . RET signaling . Regulation of KIT signaling

The displaying pathway is ordered by p-value. The color corresponds to the enrichment factor value.

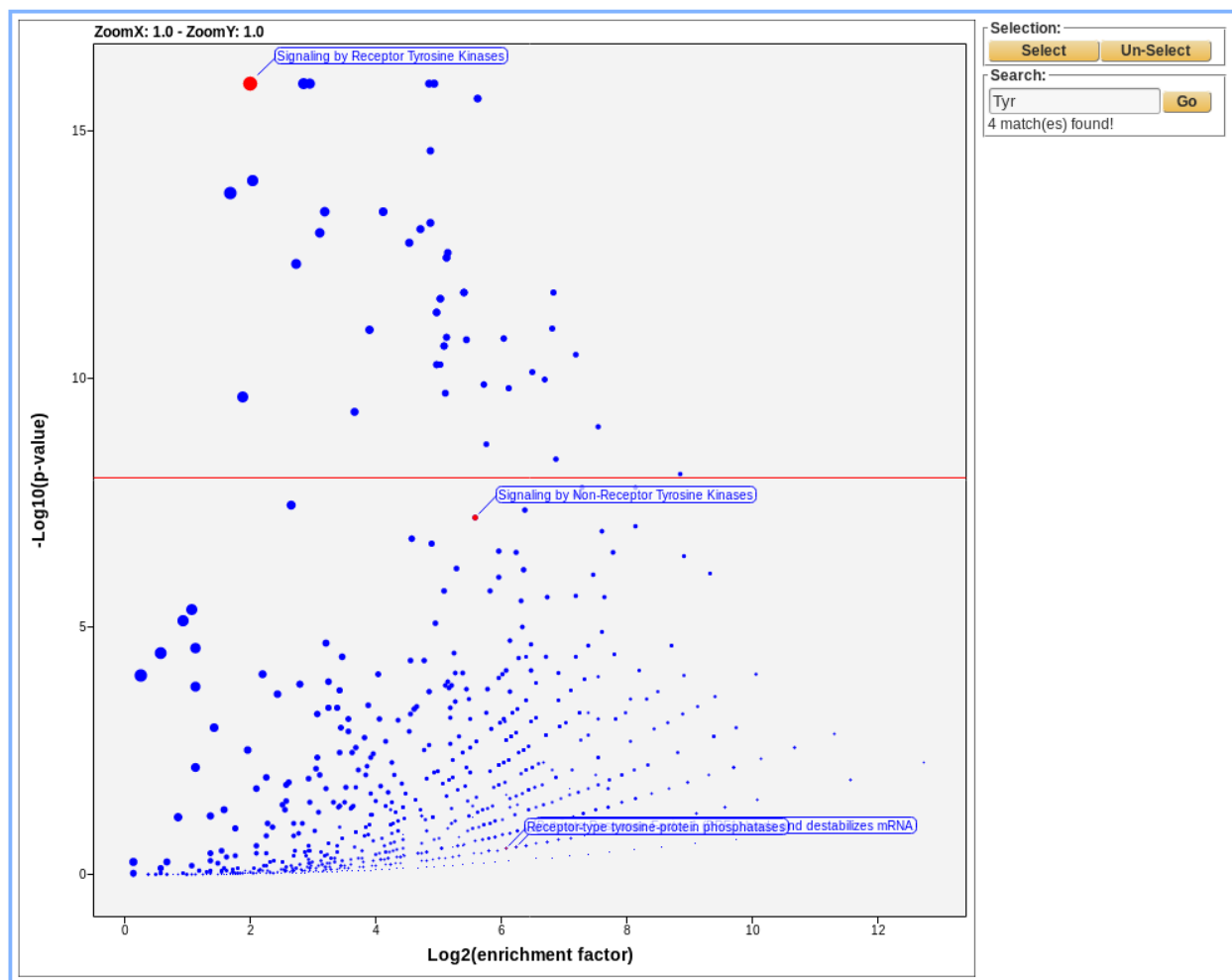
Users have access to the unannotated proteins, which does not match with the reactome database.

189/727 unannotated proteins (click for details)

Unannotated proteins

189 Proteins	Uniprot AC	Gene name	Peptides	Description - Species
PIPSL_HUMAN	A2A3N6	PIPSL	2	Putative PIP5K1A and PSMD4-like protein Homo sapiens
CHRD1_HUMAN	Q9UHD1	CHORDC1	1	Cysteine and histidine-rich domain-containing protein 1 Homo sapiens
DBF4B_HUMAN	Q8NFT6	DBF4B	1	Protein DBF4 homolog B Homo sapiens
4EBP2_HUMAN	Q13542	EIF4EBP2	1	Eukaryotic translation initiation factor 4E-binding protein 2 Homo sapiens
M3K6_HUMAN	O95382	MAP3K6	1	Mitogen-activated protein kinase kinase kinase 6 Homo sapiens
EPOR_HUMAN	P19235	EPOR	1	Erythropoietin receptor Homo sapiens
ANS1A_HUMAN	Q92625	ANKS1A	1	Ankyrin repeat and SAM domain-containing protein 1A Homo sapiens
RSPH3_HUMAN	Q86UC2	RSPH3	1	Radial spoke head protein 3 homolog Homo sapiens
RBM38_HUMAN	Q9H0Z9	RBM38	1	RNA-binding protein 38 Homo sapiens

Graphical view :



Export

Signaling by Receptor Tyrosine Kinases

Enrichment Factor: 4.0, p-value=1.11e-16

101 Proteins	Uniprot AC	Gene name	Peptides	Description - Species
KS6A2_HUMAN	Q15349	RPS6KA2	2	Ribosomal protein S6 kinase alpha-2 Homo sapiens
ARHG2_HUMAN	Q92974	ARHGEF2	1	Rho guanine nucleotide exchange factor 2 Homo sapiens
PGFRA_HUMAN	P16234	PDGFRA	1	Platelet-derived growth factor receptor alpha Homo sapiens
UBP8_HUMAN	P40818	USP8	1	Ubiquitin carboxyl-terminal hydrolase 8 Homo sapiens
JAK2_HUMAN	O60674	JAK2	1	Tyrosine-protein kinase JAK2 Homo sapiens
FGFR3_HUMAN	P22607	FGFR3	1	Fibroblast growth factor receptor 3 Homo sapiens
PP2AB_HUMAN	P62714	PPP2CB	1	Serine/threonine-protein phosphatase 2A catalytic subunit beta isoform Homo sapiens
TRIB3_HUMAN	Q96RU7	TRIB3	1	Tribbles homolog 3 Homo sapiens
RANB9_HUMAN	Q96S59	RANBP9	1	Ran-binding protein 9 Homo sapiens
CBL_HUMAN	P22681	CBL	1	E3 ubiquitin-protein ligase CBL Homo sapiens
IRS4_HUMAN	O14654	IRS4	1	Insulin receptor substrate 4 Homo sapiens
STAM2_HUMAN	O75886	STAM2	1	Signal transducing adapter molecule 2 Homo sapiens
KDIS_HUMAN	Q9ULH0	KIDINS220	1	Kinase D-interacting substrate of 220 kDa Homo sapiens
RHOA_HUMAN	P61586	RHOA	1	Transforming protein RhoA Homo sapiens
HGS_HUMAN	O14964	HGS	1	Hepatocyte growth factor-regulated tyrosine kinase substrate Homo sapiens
IGF1R_HUMAN	P08069	IGF1R	1	Insulin-like growth factor 1 receptor Homo sapiens
AP2M1_HUMAN	Q96CW1	AP2M1	1	AP-2 complex subunit mu Homo sapiens
YES_HUMAN	P07947	YES1	1	Tyrosine-protein kinase Yes Homo sapiens
ROCK2_HUMAN	O75116	ROCK2	1	Rho-associated protein kinase 2 Homo sapiens
RON_HUMAN	Q04912	MST1R	1	Macrophage-stimulating protein receptor Homo sapiens
RAC1_HUMAN	P63000	RAC1	1	Ras-related C3 botulinum toxin substrate 1 Homo sapiens
PP2AA_HUMAN	P67775	PPP2CA	1	Serine/threonine-protein phosphatase 2A catalytic subunit alpha isoform Homo sapiens
MAPK2_HUMAN	P49137	MAPKAPK2	1	MAP kinase-activated protein kinase 2 Homo sapiens
EPS15_HUMAN	P42566	EPS15	1	Epidermal growth factor receptor substrate 15 Homo sapiens
KS6A3_HUMAN	P51812	RPS6KA3	1	Ribosomal protein S6 kinase alpha-3 Homo sapiens
INSR_HUMAN	P06213	INSR	1	Insulin receptor Homo sapiens
KS6A5_HUMAN	O75582	RPS6KA5	1	Ribosomal protein S6 kinase alpha-5 Homo sapiens
PDPK1_HUMAN	O15530	PDPK1	1	3-phosphoinositide-dependent protein kinase 1 Homo sapiens
KPCI_HUMAN	P41743	PRKCI	1	Protein kinase C iota type Homo sapiens
PTK6_HUMAN	Q13882	PTK6	1	Protein-tyrosine kinase 6 Homo sapiens
FAK1_HUMAN	Q05397	PTK2	1	Focal adhesion kinase 1 Homo sapiens
SH3K1_HUMAN	Q96B97	SH3KBP1	1	SH3 domain-containing kinase-binding protein 1 Homo sapiens
NGF_HUMAN	P01138	NGF	1	Beta-nerve growth factor Homo sapiens
PK3CB_HUMAN	P42338	PIK3CB	1	Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit beta isoform Homo sapiens

Danger: A COMPLETER

Table view :

Pathway Term	Proteins			P-Value
	Enrichment factor	FDR (%)	List	
Signaling by Receptor Tyrosine Kinases	4.0	0.00	Details	1.11e-16
Negative regulation of the PI3K/AKT network	30.7	0.00	Details	1.11e-16
PI3K/AKT Signaling in Cancer	29.0	0.00	Details	1.11e-16
Intracellular signaling by second messengers	7.2	0.00	Details	1.11e-16
PIP3 activates AKT signaling	7.8	0.00	Details	1.11e-16
Regulation of TP53 Activity through Phosphorylation	49.6	0.00	Details	2.22e-16
PI5P, PP2A and IER3 Regulate PI3K/AKT Signaling	29.2	0.00	Details	2.55e-15
Diseases of signal transduction	4.1	0.00	Details	1.03e-14
Axon guidance	3.2	0.00	Details	1.79e-14
Regulation of TP53 Activity	17.5	0.00	Details	4.25e-14
RAF/MAP kinase cascade	9.1	0.00	Details	4.35e-14
VEGFA-VEGFR2 Pathway	29.5	0.00	Details	7.12e-14
Signaling by VEGF	26.4	0.00	Details	9.60e-14
MAPK1/MAPK3 signaling	8.7	0.00	Details	1.14e-13
Toll Like Receptor 4 (TLR4) Cascade	23.2	0.00	Details	1.83e-13
Toll Like Receptor 3 (TLR3) Cascade	35.4	0.00	Details	2.92e-13
TRIF(TICAM1)-mediated TLR4 signaling	34.7	0.00	Details	3.68e-13
MyD88-independent TLR4 cascade	34.7	0.00	Details	3.68e-13
MAPK family signaling cascades	6.7	0.00	Details	5.05e-13
Toll Like Receptor 5 (TLR5) Cascade	42.1	0.00	Details	1.81e-12
Toll Like Receptor 10 (TLR10) Cascade	42.1	0.00	Details	1.81e-12
MyD88 cascade initiated on plasma membrane	42.1	0.00	Details	1.81e-12
Signaling by SCF-KIT	113.9	0.00	Details	1.82e-12
MyD88:Mal cascade initiated on plasma membrane	33.0	0.00	Details	2.46e-12
Toll Like Receptor TLR6:TLR2 Cascade	33.0	0.00	Details	2.46e-12
Toll Like Receptor 2 (TLR2) Cascade	31.3	0.00	Details	4.69e-12
Toll Like Receptor TLR1:TLR2 Cascade	31.3	0.00	Details	4.69e-12
mTOR signalling	112.9	0.00	Details	1.01e-11
Toll-Like Receptors Cascades	15.0	0.00	Details	1.03e-11
TRAF6 mediated induction of NFkB and MAP kinases upon TLR7/8 or 9 activation	35.1	0.00	Details	1.44e-11
MAP kinase activation in TLR cascade	65.5	0.00	Details	1.59e-11
NGF signalling via TRKA from the plasma membrane	43.5	0.00	Details	1.71e-11
Toll Like Receptor 7/8 (TLR7/8) Cascade	33.8	0.00	Details	2.22e-11
MyD88 dependent cascade initiated on endosome	33.8	0.00	Details	2.22e-11
CD28 co-stimulation	145.1	0.00	Details	3.42e-11
Toll Like Receptor 9 (TLR9) Cascade	31.4	0.00	Details	5.10e-11
Constitutive Signaling by Aberrant PI3K in Cancer	33.0	0.00	Details	5.42e-11
EPH-ephrin mediated repulsion of cells	89.6	0.00	Details	7.51e-11
Interleukin-3, 5 and GM-CSF signaling	103.0	0.00	Details	1.08e-10
Interleukin-17 signaling	52.6	0.00	Details	1.32e-10
GPVI-mediated activation cascade	69.5	0.00	Details	1.60e-10
EPH-Ephrin signaling	34.5	0.00	Details	1.96e-10
Transcriptional Regulation by TP53	3.7	0.00	Details	2.30e-10
Signalling by NGF	12.7	0.00	Details	4.69e-10
EPHA-mediated growth cone collapse	186.7	0.00	Details	9.19e-10
Signaling by ERBB2	54.3	0.00	Details	2.12e-09
RET signaling	117.3	0.00	Details	4.23e-09
Regulation of KIT signaling	460.2	0.00	Details	8.25e-09

By default, pathways are ordered by p-value. They can be sorted by enrichment factor or FDR by clicking the corresponding name. A pathway can be display in Reactome by clicking on its name.

The screenshot displays the Reactome database interface for the 'Signaling by Receptor Tyrosine Kinases' pathway. The left sidebar shows a hierarchical tree of biological processes, with 'Signal Transduction' and 'Signaling by Receptor Tyrosine Kinases' highlighted. The main central area displays a grid of pathway boxes, each representing a specific signaling mechanism (e.g., Signaling by EGFR, Signaling by ERBB2, Signaling by MET, Signaling by FGFR, Signaling by ERBB4, Signaling by MST1, Signaling by NGF, Signaling by Insulin Receptor, Signaling by PDGF, Signaling by Type 1 Insulin-like Growth Factor 1 Receptor (IGF1R), Signaling by VEGF, Signaling by SCF-KIT). The bottom section includes a table with columns for Description, Molecules, Structures, Expression, Analysis, and Downloads, and a summary text box providing detailed information about the pathway.

20.2.4 Export

An excel file can be downloaded which correspond to the table view information.

20.3 Motif enrichment analysis

Motifs are recurring short sequence elements. Their over-representation usually implies some functional significance. *myProMS* uses *rmtotifx* R package for the enrichment analysis and *ggseqlogo* R package for drawing the motif.

20.3.1 Launch analysis

Motif Enrichment Analysis

Motif Analysis Name :

Modifications :

Foreground Selection :

Foreground filtering : Ratio ☒ **Exclude infinite ratios**
p-value ≤

Central character :

Width :

Min. Occurrence :

Significance :

Background Selection : ☒ **quantification selected**
☐ **Random** **sequences**

- **Name:** provide a name for the PCA and/or the clustering analysis. The analysis is saved and can be retrieved by this name in the Exploratory analyses tree displayed in the sub-navigation frame.
- **Modification:** choose the type of modifications (phosphorylation, acetylation, methylation...).
- **Foreground selection:** select a quantification or a previously saved list.
- **Foreground filtering:**
 - Ratio: all proteins with a ratio corresponding to the chosen value are kept.
 - Infinite ratio: exclude or not.
- **p-value:** all proteins with a p-value smaller than or equal to the chosen value are kept.
- **Central residue:** following the type of modifications, the drop down menu is automatically generated. For instance, phosphorylation modification generate 3 type of residue (S, T, Y).
- **Width:** the width is the number of total characters is the motif. It should be an odd number between 3 and 35. However, choosing a motif width that is too narrow can result in the exclusion of motifs with critical longer-range dependencies and choosing a motif width that is too wide (without adjusting the significance threshold accordingly) can yield spurious motif results.
- **Min. occurrence:** the occurrence threshold refers to the minimum number of times you wish each of your extracted motifs to occur in the data set. An occurrence threshold of 20 usually is appropriate, although this parameter may be adjusted to yield more specific or less specific motifs. This parameter can be used to tune the specificity of motifs since motifs with greater specificity (i.e., more “fixed” positions) are expected to occur less often than those motifs with lower specificity. Users that wish to extract a maximal number of motifs should set this parameter to a low value (for example, “5”) and rely solely on the significance parameter (see step 8) to extract motifs. On occasion it may be useful to set this parameter as a fractional percent of the total number of modification sites in order to compare motifs with similar specificities across data sets that vary in size (e.g., to

compare motifs from data sets of 300 and 3000 sites, one may opt to set the occurrences parameter to 5 and 50, respectively).

- **Significance:** The significance refers to the P-value threshold for the binomial probability. This is used for the selection of significant residue/position pairs in the motif. It is critical to note that this value does not take into account a correction for multiple hypotheses (such as the Bonferroni correction). On any given motif-x search step there are (number of possible characters at each position) * (number of non fixed positions) hypotheses being tested. For example, in an S-centered analysis of width 15, there would be $(20) * (14) = 280$ hypotheses tested. To ensure an alpha-value of at least 0.05 by the Bonferroni method, one would need to divide the desired alpha-value by the total number of hypotheses tested (i.e., $0.05/280 = 0.00018$). We suggest a threshold of 0.000001 to maintain a low false positive rate in standard protein motif analyses.
- **Background selection:** The background simply refers to the item from which the data set was taken. This is important for accurate statistical analysis. You may choose the selected quantification or generate a random background sequence based on the probability of each amino acid within the proteome.

20.3.2 Summary/editing/deleting

Motif Enrichment Analysis : test

Analysis name :	test
Quantification :	Gr1vsGr2vsGr3vsGr4 Phospho > 1x5v2x7v3x5v4x11-P : Gr1/MixSILAC
Description :	
Foreground selection :	<ul style="list-style-type: none"> • Central Residue : Serine • Min.seqs : 20 • Width : 11 • Significance : 0.000001
Background selection :	Type : selected quantification
Feature selection :	<ul style="list-style-type: none"> • infinite ratios excluded • FC : Up ≥ 2 or Down $\leq 1/2$ • p-value : 0.01
Motif :	4 motif(s) found
Status :	✓ Finished

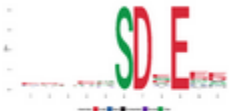



If a motif enrichment analysis is selected, a summary of the information available for that analysis is displayed in the mainframe.

The motif name and the description can be modified by clicking the `edit` button. The analysis can be deleted by clicking the `delete` button.

Important: If the motif analysis is involved in a heatmap, you have to delete the heatmap before deleting the motif analysis.

20.3.3 Displaying motif enrichment

Display Motif Enrichment Analysis for **test**

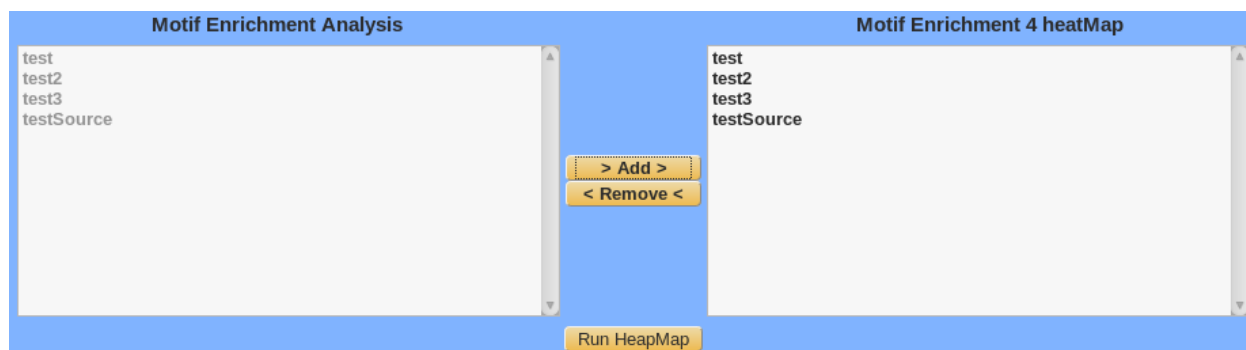
Motif	Fold Change	Score	Fg. Matches	Fg. Size	Bg. Matches	Bg. Size
	2.710	314.245	91	1639	643	31383
	1.449	8.039	231	1548	3165	30740
	2.672	14.617	61	1317	478	27575
	1.503	6.863	165	1256	2368	27097

- **Motif:** the logo picture made by ggseqlogo with the over-represented motif. Motif positions are labeled below the x-axis and residues are colored according to their chemical and physical properties.
- **Fold change:** is an indicator of the enrichment level of the extracted motifs. Specifically, it is calculated as $(\text{foreground matches}/\text{foreground size})/(\text{background matches}/\text{background size})$.
- **Score:** The “motif score” is calculated by taking the sum of the negative log probabilities used to fix each position of the motif. As such, higher motif scores typically correspond to motifs that are more statistically significant as well as more specific (i.e., greater number of fixed positions).
- **Fg. matches / Bg. matches:** indicate the number of peptides containing a given motif in those respective data sets following the removal of all peptides containing previously extracted motifs. Because of this iterative “set reduction” strategy, the “foreground matches” and “background matches” statistics may be less than or equal to the total number of instances of a given motif in the whole data set.
- **Fg. size / Bg. size:** indicate the total number of peptides contained in these data sets. The size of these data sets decreases as motifs are extracted (i.e., down a column) due to the fact that peptides are removed from both the foreground and background data sets following motif extraction. The total number of foreground peptides not falling into any extracted motif class can therefore be calculated as the difference between the “foreground size” and the “foreground matches” of the final motif class (e.g., $163 - 32 = 131$ unclassified peptides in Figure 5). The “fold increase” statistic is an indicator of the enrichment level of the extracted.

Clicking on either the links in the motif logo, brings users to the peptide data used to extract the given motifs. As such, the number of peptides found for each motif in this section corresponds exactly to the number of “foreground matches” for the motif in appropriate column of the results table.

20.3.4 Heatmap motif analysis

Launch HM motif analysis



Select the previously saved motif analysis in the left panel, click on the Add button and the selected analysis are moved in the right panel to be cluster.

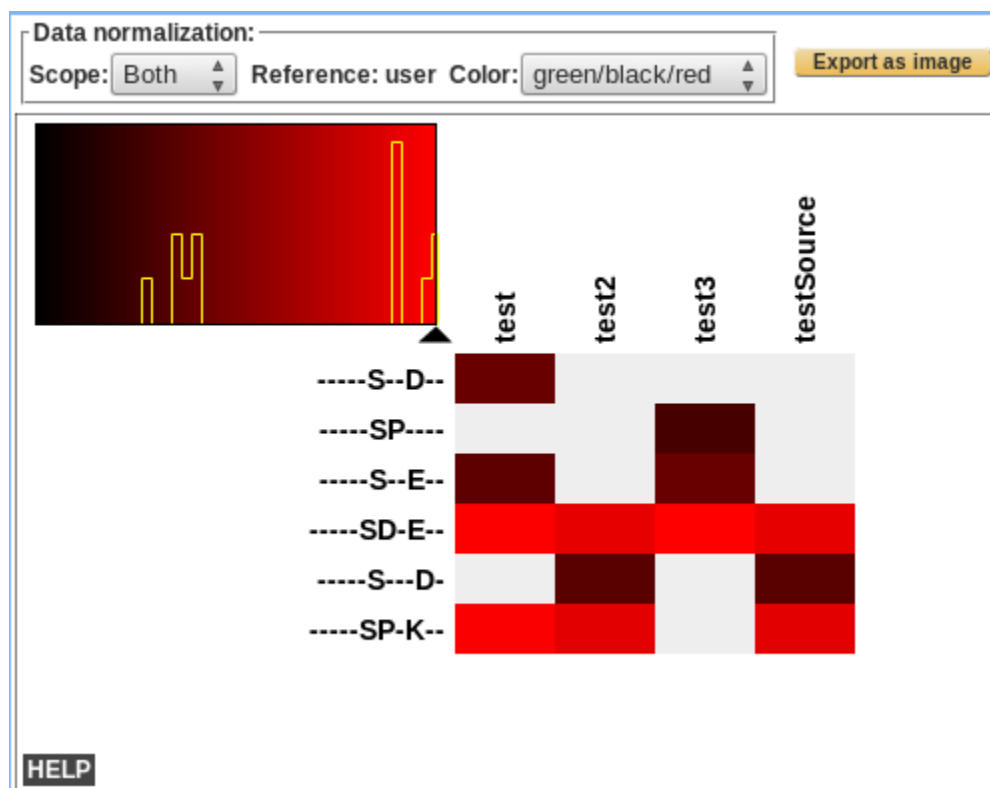
Summary / Edit / Delete

Heatmap Motif Enrichment Analysis : testHM

Analysis name :	testHM
Selection :	<ul style="list-style-type: none"> • test • test2 • test3 • testSource
Description :	

<to be completed>

Display Heatmap motif analysis



<to be completed>

CHAPTER 21

Indices and tables

- search