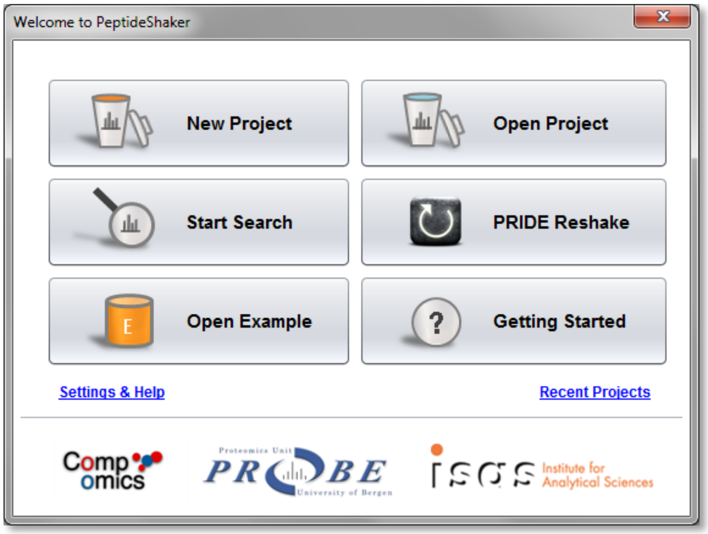
A. PSM, Peptide & Protein Visualization

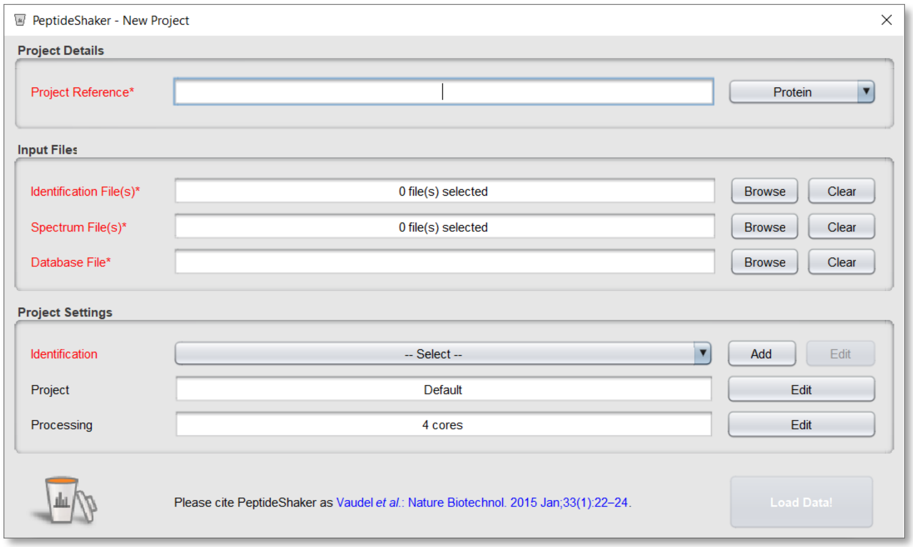
The search conducted in the "Peptide to Spectrum Matching" chapter generated two files containing the peptides matched by OMSSA and X! Tandem for each spectrum, so-called Peptide to Spectrum Matches (PSMs). From these we want to find the identified peptides and proteins. This is the task of PeptideShaker[1] (http://compomics.github.io/projects/peptide-shaker.html).

First, start PeptideShaker by double-clicking the file called PeptideShaker-X.Y.Z.jar in the PeptideShaker-X.Y.Z folder located in the software folder (X.Y.Z is the version number). You should see the following dialog:



From the Welcome dialog, you can create a new project, open a previously saved project, start a search with SearchGUI, Reshake (i.e., reanalyze) existing PRIDE[2] data, open an example dataset or navigate our 'Getting Started' presentation.

We are now going to create a new project with the previously generated OMSSA and X! Tandem files (also available in the resources folder). Click on ‘New Project’. You will see this screen:



PeptideShaker - New Project

Project Details

Project Reference*: Protein

Input Files

Identification File(s)*: 0 file(s) selected

Spectrum File(s)*: 0 file(s) selected

Database File*:

Project Settings

Identification: -- Select --

Project: Default

Processing: 4 cores

Please cite PeptideShaker as Vaudel *et al.*: Nature Biotechnol. 2015 Jan;33(1):22–24.

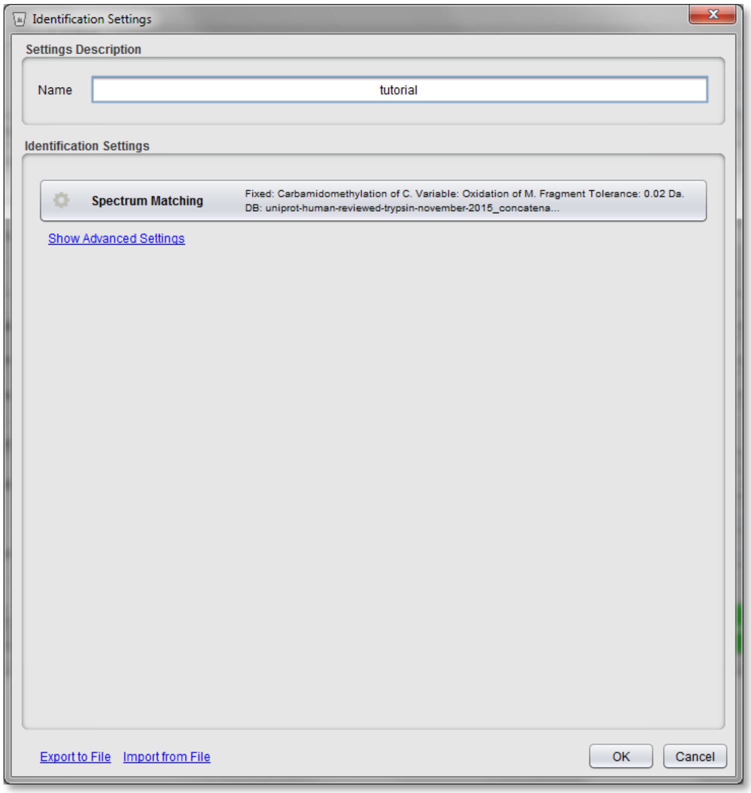
Start by giving your project a name at the top, and leave the project type as Protein.

The species of interest for this sample is human: this is in fact a measurement of a HeLa lysate, courtesy of the Leibniz-Institut für Analytische Wissenschaften - ISAS - e. V., Dortmund, Germany. Knowing the species type allows us to extract additional information about the genes the proteins come. For this via use Ensembl. So when you see gene information later on you now know where the information comes from. *What is Ensembl?* *[1.4a]*

Now we are ready to add the search result files. Click the 'Browse' button next to the 'Identification File(s)' text field. Navigate to the resources folder, select the zip file containing the OMSSA and X! Tandem result files (searchgui_out.zip) and click 'Add'.

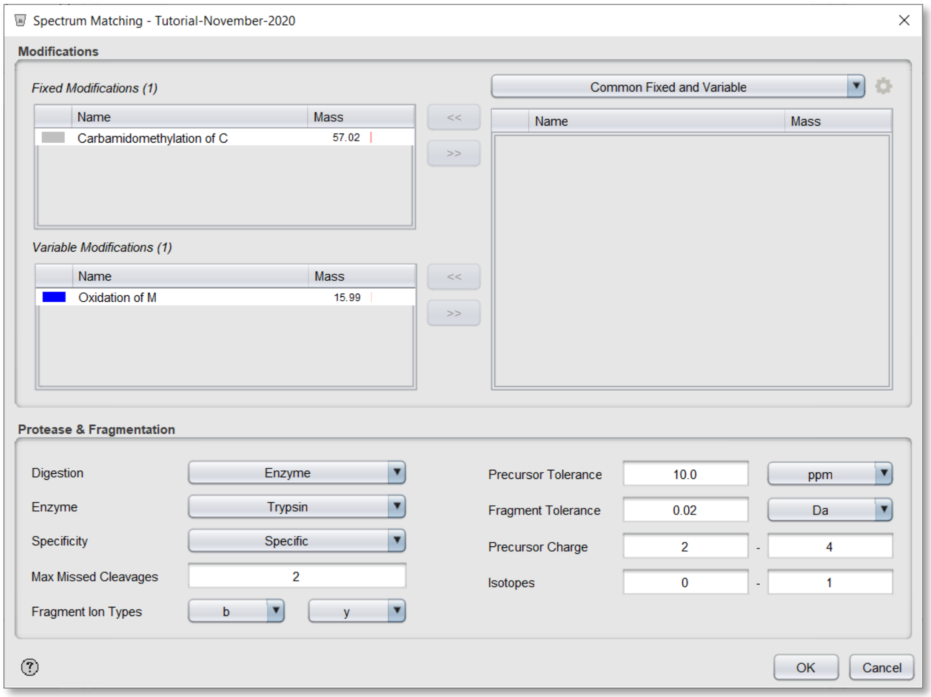
Note that the spectrum files and the database are automatically filled in when loading results from SearchGUI. Otherwise you have to load these manually (files are in the resources folder). Note also that the Identification Settings have changed. This information was extracted from the searchGUI_input.txt file and the parameters file created by SearchGUI along with the identification files. Note that if files have been moved in the interim, you can add them manually.

Once all the files have been loaded, click the 'Edit' button for the 'Identification Settings'.


Name: tutorial

Spectrum Matching: Fixed: Carbamidomethylation of C. Variable: Oxidation of M. Fragment Tolerance: 0.02 Da. DB: uniprot-human-reviewed-trypsin-november-2015_concatena...

By clicking the 'Spectrum Matching' option the dialog showing the used search settings will appear. They should be identical to the settings you provided in SearchGUI:



Spectrum Matching - Tutorial-November-2020

Modifications

Fixed Modifications (1)

Name	Mass
Carbamidomethylation of C	57.02

Variable Modifications (1)

Name	Mass
🟦 Oxidation of M	15.99

Common Fixed and Variable

Name	Mass

Protease & Fragmentation

Digestion: Enzyme
Enzyme: Trypsin
Specificity: Specific
Max Missed Cleavages: 2
Fragment Ion Types: b y

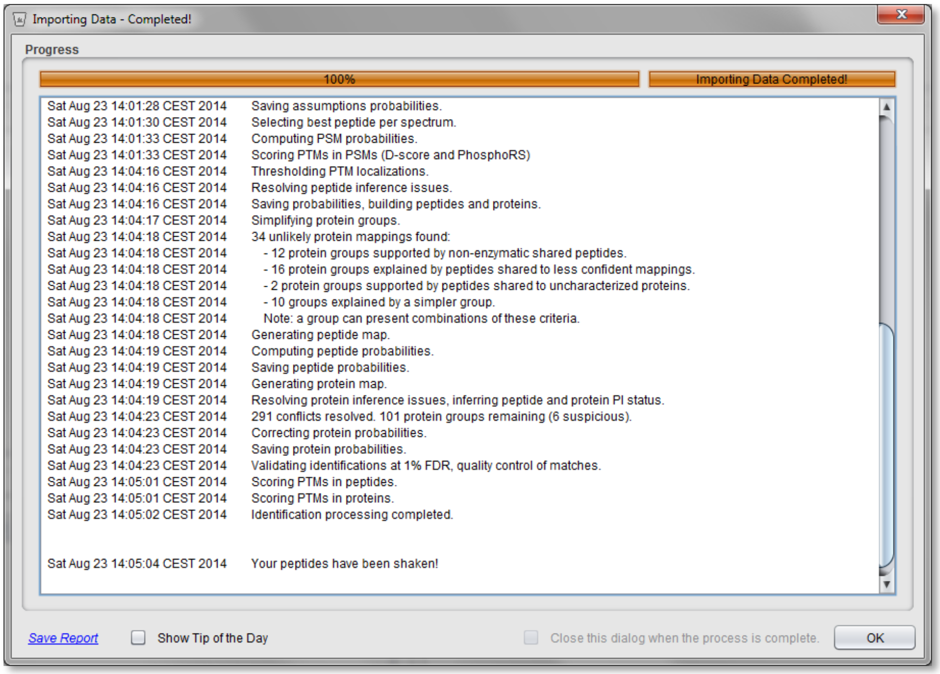
Precursor Tolerance: 10.0 ppm
Fragment Tolerance: 0.02 Da
Precursor Charge: 2 - 4
Isotopes: 0 - 1

In addition to the modifications listed here you may also see the following modifications in your search results later on: Acetylation of protein N-term, Pyrolidone from E, Pyrolidone from Q and Pyrolidone from carbamidomethylated C. *Where do these come from? [1.4b]*

Close the 'Search Settings' dialog to go back to the 'Identification Settings' overview dialog.

Feel free to also explore the 'Advanced Settings', but be sure to not alter them as the results may then not match the tutorial.

Click 'OK' to go back to the New Project dialog and then click 'Load Data!'. The selected files will now be processed and loaded into PeptideShaker, and you can follow the progress in the dialog while browsing some PeptideShaker usage tips in the lower right corner of the progress dialog.



```
Sat Aug 23 14:01:28 CEST 2014    Saving assumptions probabilities.
Sat Aug 23 14:01:30 CEST 2014    Selecting best peptide per spectrum.
Sat Aug 23 14:01:33 CEST 2014    Computing PSM probabilities.
Sat Aug 23 14:01:33 CEST 2014    Scoring PTMs in PSMs (D-score and PhosphoRS)
Sat Aug 23 14:04:16 CEST 2014    Thresholding PTM localizations.
Sat Aug 23 14:04:16 CEST 2014    Resolving peptide inference issues.
Sat Aug 23 14:04:16 CEST 2014    Saving probabilities, building peptides and proteins.
Sat Aug 23 14:04:17 CEST 2014    Simplifying protein groups.
Sat Aug 23 14:04:18 CEST 2014    34 unlikely protein mappings found:
Sat Aug 23 14:04:18 CEST 2014      - 12 protein groups supported by non-enzymatic shared peptides.
Sat Aug 23 14:04:18 CEST 2014      - 16 protein groups explained by peptides shared to less confident mappings.
Sat Aug 23 14:04:18 CEST 2014      - 2 protein groups supported by peptides shared to uncharacterized proteins.
Sat Aug 23 14:04:18 CEST 2014      - 10 groups explained by a simpler group.
Sat Aug 23 14:04:18 CEST 2014      Note: a group can present combinations of these criteria.
Sat Aug 23 14:04:18 CEST 2014    Generating peptide map.
Sat Aug 23 14:04:19 CEST 2014    Computing peptide probabilities.
Sat Aug 23 14:04:19 CEST 2014    Saving peptide probabilities.
Sat Aug 23 14:04:19 CEST 2014    Generating protein map.
Sat Aug 23 14:04:19 CEST 2014    Resolving protein inference issues, inferring peptide and protein PI status.
Sat Aug 23 14:04:23 CEST 2014    291 conflicts resolved. 101 protein groups remaining (6 suspicious).
Sat Aug 23 14:04:23 CEST 2014    Correcting protein probabilities.
Sat Aug 23 14:04:23 CEST 2014    Saving protein probabilities.
Sat Aug 23 14:04:23 CEST 2014    Validating identifications at 1% FDR, quality control of matches.
Sat Aug 23 14:05:01 CEST 2014    Scoring PTMs in peptides.
Sat Aug 23 14:05:01 CEST 2014    Scoring PTMs in proteins.
Sat Aug 23 14:05:02 CEST 2014    Identification processing completed.


Sat Aug 23 14:05:04 CEST 2014    Your peptides have been shaken!
```

Tip:

If you want to save time you can cancel the loading and open the PeptideShaker example dataset instead. This was created using the exact same input as above and is available in the Welcome Dialog.

Loading large dataset in PeptideShaker can be time consuming. The processing will be faster if increasing the memory allocated to PeptideShaker. You can do that *via* the *Edit > Java Settings* menu (or *via* the Welcome Dialog - *Settings & Help > Settings > Java Settings*). Note that large datasets (>100 000 spectra) or databases (>100 000 protein sequences) will require a lot of memory. For smaller datasets (<100 000 spectra and <100 000 protein sequences), a standard laptop (64 bits and 4 GB of RAM) is sufficient. The performance will also be improved if working with SSD discs and multiple CPUs.

If you do not want to wait for the data to load, you can always cancel the loading and when restarting PeptideShaker select the "Open Example" option in the Welcome Dialog.

Tip:

Do not put databases or mgf files on external drives, prefer a local SSD disc: the software needs to read these often, maximizing the reading speed will reduce the loading time!

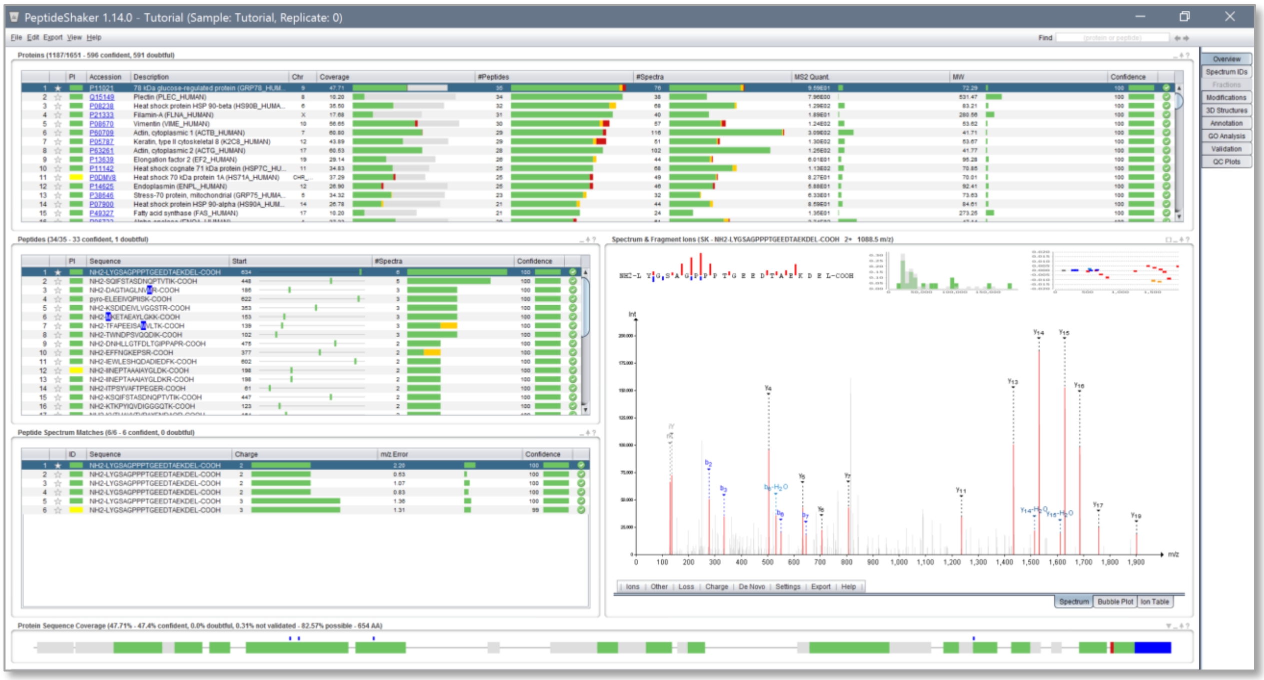
Tip:

Before starting an ambitious experiment, make sure that the software you need exists and that you have the computing power to analyze the data. Run smaller tests before starting the experiment!

Tip:

Before loading a large dataset consisting of multiple runs, load a single run first to verify that everything works fine!

When finished, the dialog closes and the identification results are displayed, you should see something like this:


PeptideShaker 1.14.0 - Tutorial (Sample: Tutorial, Replicate: 0)
File Edit Export View Help
Find (protein or peptide)
Overview
Spectrum IDs
Fractions
Modifications
3D Structures
Annotation
GO Analysis
Validation
QC Plots
NH2-L Y G S A G P P P T G E E D T A E K D E L-COOH
Int
m/z
0 100 200 300 400 500 600 700 800 900 1,000 1,100 1,200 1,300 1,400 1,500 1,600 1,700 1,800 1,900
b_2 b_3 b_6-H_2O b_6 b_7 y_4 y_5 y_6 y_7 y_{11} y_{13} y_{14} y_{15} y_{16} y_{14}-H_2O y_{15}-H_2O y_{17} y_{19}
Ions | Other | Loss | Charge | De Novo | Settings | Export | Help
Spectrum | Bubble Plot | Ion Table

Proteins (1187/1651 - 596 confident, 591 doubtful)

		PI	Accession	Description	Chr	Coverage	#Peptides	#Spectra	MS2 Quant.	MW	Confidence	
1	★	🟩	P11021	78 kDa glucose-regulated protein (GRP78_HUM...	9	47.71	35	76	9.59E01	72.29	100	✓
2	☆	🟩	Q15149	Plectin (PLEC_HUMAN)	8	10.20	34	38	7.96E00	531.47	100	✓
3	☆	🟩	P08238	Heat shock protein HSP 90-beta (HS90B_HUMA...	6	35.50	32	68	1.29E02	83.21	100	✓
4	☆	🟩	P21333	Filamin-A (FLNA_HUMAN)	X	17.68	31	40	1.89E01	280.56	100	✓
5	☆	🟩	P08670	Vimentin (VIME_HUMAN)	10	56.65	30	57	1.24E02	53.62	100	✓
6	☆	🟩	P60709	Actin, cytoplasmic 1 (ACTB_HUMAN)	7	60.80	29	116	3.09E02	41.71	100	✓
7	☆	🟩	P05787	Keratin, type II cytoskeletal 8 (K2C8_HUMAN)	12	43.89	29	51	1.30E02	53.67	100	✓
8	☆	🟩	P63261	Actin, cytoplasmic 2 (ACTG_HUMAN)	17	60.53	28	102	1.25E02	41.77	100	✓
9	☆	🟩	P13639	Elongation factor 2 (EF2_HUMAN)	19	29.14	26	44	6.01E01	95.28	100	✓
10	☆	🟩	P11142	Heat shock cognate 71 kDa protein (HSP7C_HU...	11	34.83	25	68	1.13E02	70.85	100	✓
11	☆	🟨	P0DMV8	Heat shock 70 kDa protein 1A (HS71A_HUMAN)	CHR_...	37.29	25	49	8.27E01	70.01	100	✓
12	☆	🟩	P14625	Endoplasmin (ENPL_HUMAN)	12	26.90	25	46	5.88E01	92.41	100	✓
13	☆	🟩	P38646	Stress-70 protein, mitochondrial (GRP75_HUMA...	5	34.32	23	32	6.33E01	73.63	100	✓
14	☆	🟩	P07900	Heat shock protein HSP 90-alpha (HS90A_HUM...	14	26.78	21	44	8.59E01	84.61	100	✓
15	☆	🟩	P49327	Fatty acid synthase (FAS_HUMAN)	17	10.20	21	24	1.35E01	273.25	100	✓

Peptides (34/35 - 33 confident, 1 doubtful)

		PI	Sequence	Start	#Spectra	Confidence	
1	★	🟩	NH2-LYGSAGPPPTGEEDTAEKDEL-COOH	634	6	100	✓
2	☆	🟩	NH2-SQIFSTASDNQPTVTIK-COOH	448	5	100	✓
3	☆	🟩	NH2-DAGTIAGLNVMR-COOH	186	3	100	✓
4	☆	🟩	pyro-ELEEIVQPIISK-COOH	622	3	100	✓
5	☆	🟩	NH2-KSDIDEIVLVGGSTR-COOH	353	3	100	✓
6	☆	🟩	NH2-MKETAEAYLGKK-COOH	153	3	100	✓
7	☆	🟩	NH2-TFAPEEISAMVLTK-COOH	139	3	100	✓
8	☆	🟩	NH2-TWNDPSVQQDIK-COOH	102	3	100	✓
9	☆	🟩	NH2-DNHLLGTFDLTGIPPAPR-COOH	475	2	100	✓
10	☆	🟩	NH2-EFFNGKEPSR-COOH	377	2	100	✓
11	☆	🟩	NH2-IEWLESHQDADIEDFK-COOH	602	2	100	✓
12	☆	🟨	NH2-IINEPTAAAIAYGLDK-COOH	198	2	100	✓
13	☆	🟩	NH2-IINEPTAAAIAYGLDKR-COOH	198	2	100	✓
14	☆	🟩	NH2-ITPSYVAFTPEGER-COOH	61	2	100	✓
15	☆	🟩	NH2-KSQIFSTASDNQPTVTIK-COOH	447	2	100	✓
16	☆	🟩	NH2-KTKPYIQVDIGGGQTK-COOH	123	2	100	✓

Peptide Spectrum Matches (6/6 - 6 confident, 0 doubtful)

		ID	Sequence	Charge	m/z Error	Confidence	
1	★	🟩	NH2-LYGSAGPPPTGEEDTAEKDEL-COOH	2	2.20	100	✓
2	☆	🟩	NH2-LYGSAGPPPTGEEDTAEKDEL-COOH	2	0.53	100	✓
3	☆	🟩	NH2-LYGSAGPPPTGEEDTAEKDEL-COOH	2	1.07	100	✓
4	☆	🟩	NH2-LYGSAGPPPTGEEDTAEKDEL-COOH	2	0.83	100	✓
5	☆	🟩	NH2-LYGSAGPPPTGEEDTAEKDEL-COOH	3	1.36	100	✓
6	☆	🟨	NH2-LYGSAGPPPTGEEDTAEKDEL-COOH	3	1.31	99	✓

Spectrum & Fragment Ions (SK - NH2-LYGSAGPPPTGEEDTAEKDEL-COOH 2+ 1088.5 m/z)

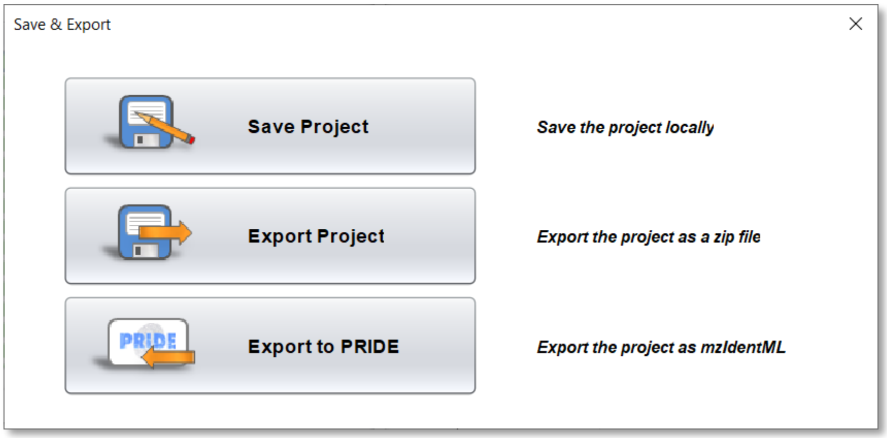
Protein Sequence Coverage (47.71% - 47.4% confident, 0.0% doubtful, 0.31% not validated - 82.57% possible - 654 AA)

This is the main screen of PeptideShaker and it gives you an overview of the results in your files. PeptideShaker is split into various tabs found in the upper right corner, each with a different function.

The Overview tab is split into five main sections: the protein table at the top; the peptide and peptide to spectrum match (PSM) tables, both in the middle left; the spectrum view in the middle right; and the sequence coverage panel at the bottom. All five sections are connected. This means that selecting a protein in the protein table updates the peptide table, the PSM table, the spectrum view and the sequence coverage panel. Try selecting some of the other proteins in the protein list and see how the other sections are updated accordingly.

At the top right of every section there are small icons allowing you to interact with the data displayed: a square to maximize the section, an underscore to minimize it, an arrow to export the content and a question mark to open contextual help.

Before doing anything else we will start by saving our PeptideShaker project. This way we do not have to reload all the data the next time we want to look at them, but can rather do a much faster load of the project directly. To save your project, go to the 'File' menu and click the 'Save' option. PeptideShaker offers you different save options:



Save the project locally

Export the project as a zip file

Export the project as mzIdentML

'Save Project' allows you to save the project in the PeptideShaker database format (.psdb files). 'Export Project' exports the project together with all related spectrum and database files as a single zip file. This option is particularly relevant when sharing your results: send the zip file to a colleague, and after unzipping your colleague will be able to open your project in PeptideShaker with just a few clicks and interact with the results. 'Export to PRIDE' allows you to save your project in the standard mzIdentML format, this will be the subject of the "Data Sharing" chapters.

For now, click on 'Save Project', choose a name for your project, and click 'Save'. Note that saving big projects can some take time: there is a lot of information to store!

A .psdb file should now have been created. If you want to open this project later simply open this file via the PeptideShaker Welcome Dialog. The project can also be accessed via the 'Open Project' and 'Open Recent Project' options in the 'File' menu after starting PeptideShaker.

As already mentioned, protein identification using mass spectral data always starts with the spectra. We will therefore start by looking at a particular spectrum in detail and see how well this corresponds to the identified peptide.

Start by making sure that the protein with accession number Q15149 with description Plectin (PLEC_HUMAN) (originating from Chromosome 8) is selected.

Looking at the row for this protein we see that we have covered 10.2 % of the protein's amino acid sequence, detecting a total of 34 peptides from 38 spectra. (Note: the color coding for the peptides and spectra will be explained in the "Peptide and Protein Validation" chapter.) *Why are there more spectra than peptides*? *How do you define a peptide?* *[1.4c]*

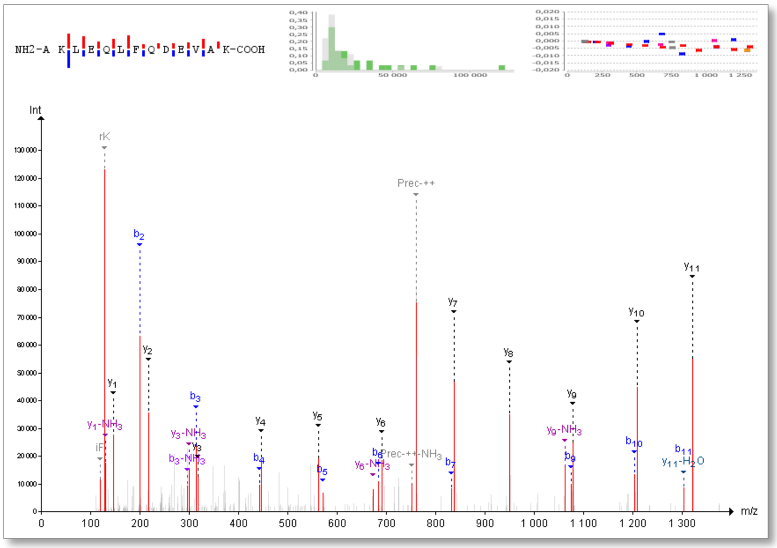
In the peptide table, you will see that some amino acids are colored. These residues were identified as carrying a post-translational modification (PTM). In fact, for an easy interpretation, PTMs are color coded everywhere in PeptideShaker. Holding the mouse over a colored residue will give you the PTM details. *How many modified peptides were identified for this protein? [1.4d]*

Tip:

By default only variable modifications are displayed. You can change this in the 'View' menu via the 'Fixed Modifications' option.

Make sure that the first peptide sequence AALAHSEEVTASQVAATK is selected. In the peptide to spectrum matches table you will see that this peptide is found two times. That means that two different spectra have been identified as the peptide in question. Notice that one of the spectra has a precursor charge of 2+, while other has a charge of 3+. *Why do not all precursors carry the same charge? [1.4e]*

Now move to the peptide sequence AKLEQLFQDEVAK. The spectrum used to identify the peptide is now shown to the right of the table, and should look like this:



By looking at the spectrum, is it obvious that this spectrum is identified as the sequence AKLEQLFQDEVAK? Which are the different ions displayed here? Are the same ions detected in all peptide to spectrum matches or peptides? What are the standard fragment ion types and how do they relate to the peptide sequence? [1.4f]

The menu bar below the spectrum allows you to fine tune the annotation settings and the spectrum display. It is additionally possible to export the spectrum in various formats. Note that only the highest peaks are annotated. You can change the annotation level by scrolling when over the spectrum. If you hold 'Ctrl' and scroll, the m/z tolerance will be changed.

Tip:

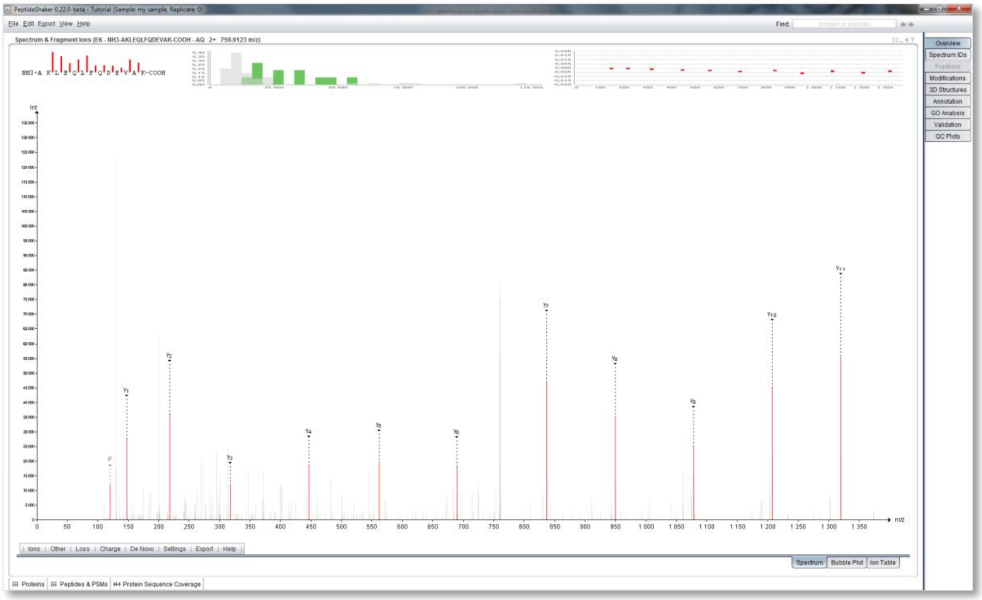
You can also change the annotation tolerances using sliders which can be enabled in the 'View' menu.

Before continuing, make sure that the intensity level is at 75% and the accuracy at 0.02 Da. (The values can also be set via *Edit > Spectrum Annotations*.)

The first thing we will do is to see how well the spectrum matches the sequence it has supposedly identified. To do this we will perform a bit of manual *de novo* sequencing, essentially checking whether the detected fragment ions match the peptide sequence. We will start with the y-ions, so disable the display of b-ions by clicking the 'Ions' menu below the spectrum and unselecting the ‘b’ ion in the list.

Neutral losses are selected automatically by PeptideShaker depending on the peptide sequence and modification status of the peptide. In order to hide the fragment ions with a neutral loss, unselect ‘Adapt’ in the ‘Loss’ menu, and then deselect the neutral losses one by one.

To make the sequencing a bit easier we will also hide the other sections of the Overview tab. Click on the square in the top right corner of the spectrum panel. Note that the other tabs have been minimized at the bottom left of the screen. The spectrum should now cover the whole screen:


PeptideShaker 0.22.0-beta - Tutorial (Sample: my sample, Replicate: 0)

Spectrum & Fragment Ions (EK - NH3-AKLEQLFQDEVAK-COOH - AQ 2+ 759.9123 m/z)

Next, to make the sequencing even easier, zoom in on the m/z range covered by the y-ions y_1 to y_{11}, simply by clicking and holding the left mouse button from just before the y_1-ion and (while holding the left mouse button pressed) dragging the mouse horizontally to the right until you have passed the y_{11}-ion. At that point just let go of the left mouse button to zoom in on the selected range. If you want to zoom in further, you can repeat the same procedure. If you at any point want to zoom out again, simply click the right mouse button to return to the full spectrum display.

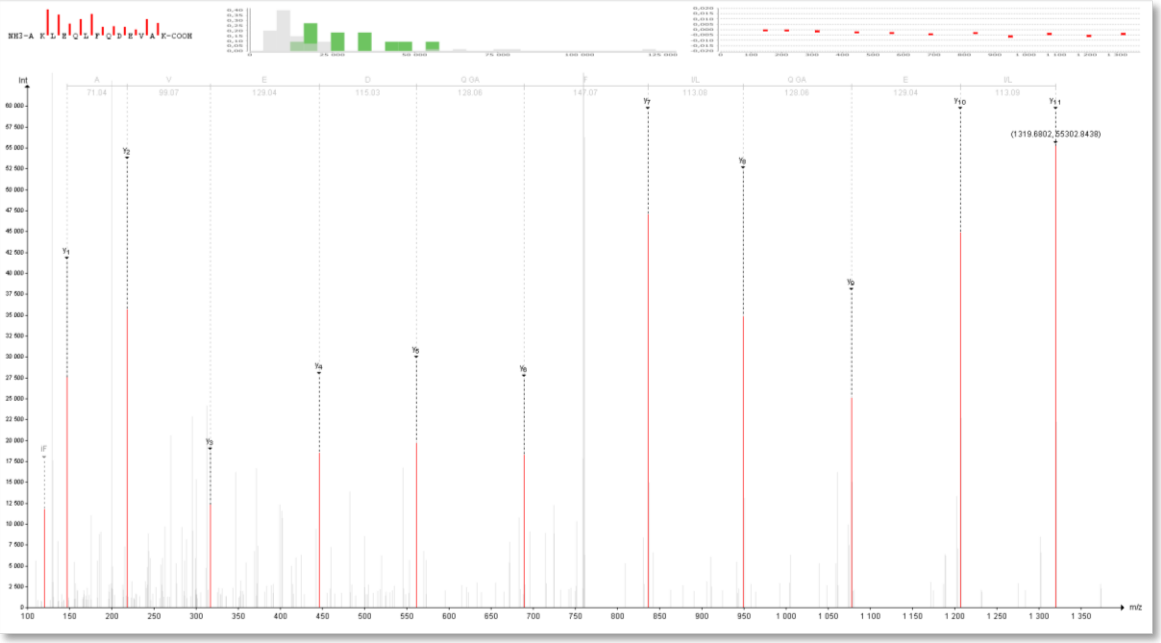
Now start the *de novo* sequencing by first clicking the y_1-ion. If you still find it difficult to click the correct peak, try zooming in just on the peak in question, select it, and then zoom back out again.

Tip:

In order to click a peak, it first needs to be highlighted (showing its exact m/z and intensity above the peak) and this highlighting only occurs if the mouse is held above the peak within 1.5 times the peak height.

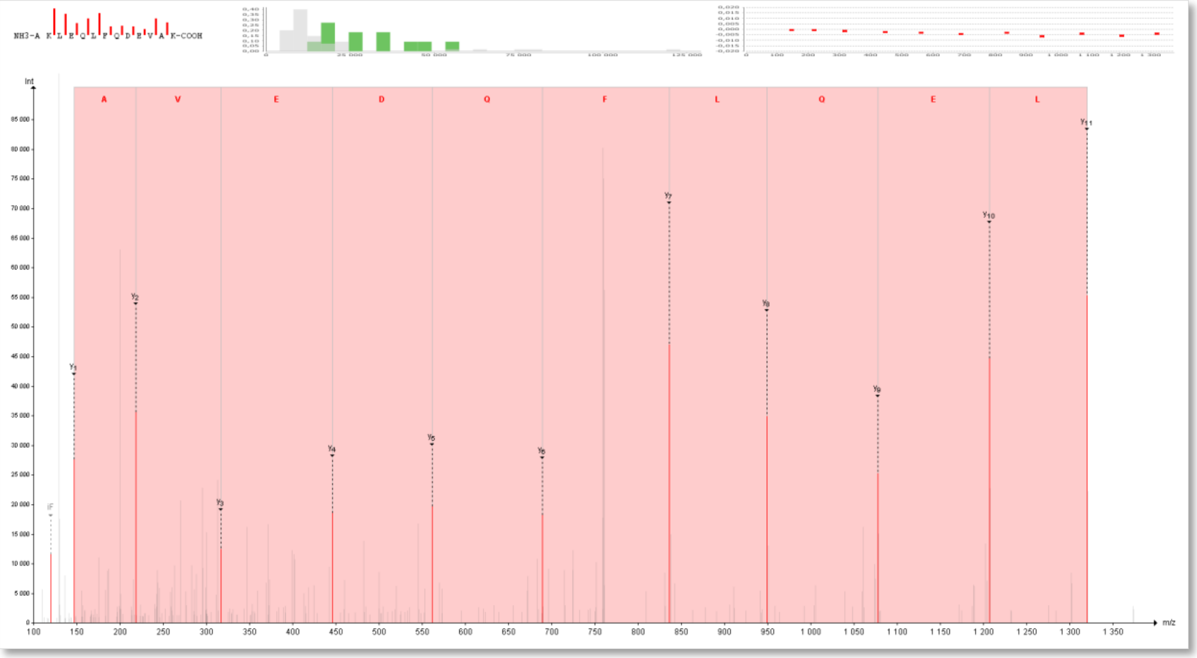
Continue the sequencing by clicking the y-ions in ascending order from left to right. If you click the wrong peak, the last selection can be removed by holding down the Shift key while clicking the peak again. Ctrl + Click will save the current sequence, and will allow you to start a second sequence.

When done, you should have a spectrum display that looks something like this:

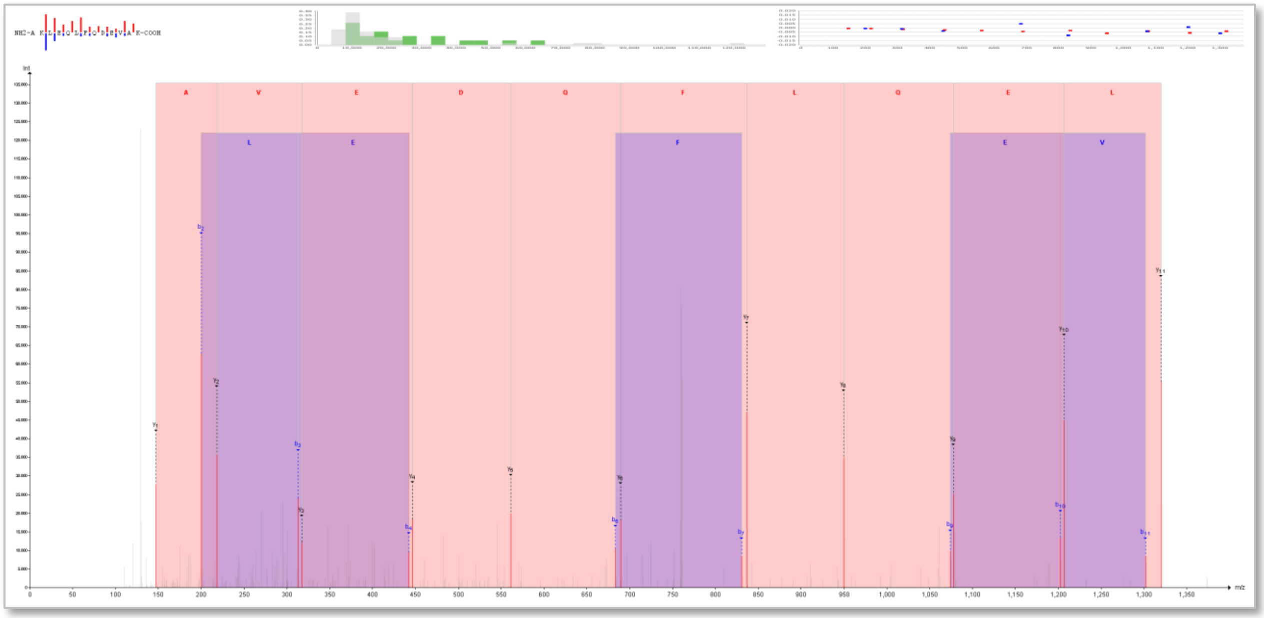


If you now read the sequence at the top you will find that it reads (from left to right): A, V, E D, Q GA, F, I/L, Q GA, E, I/L. Given that y-ions are indexed from right to left (carboxy-terminus to amino-terminus) relative to the sequence, we have to reverse the sequence to match it to the peptide sequence. Hence we get the sequence: I/L E Q/GA I/L F Q/GA D E V A. Comparing this to the peptide sequence we see that they are very similar, and if we resolve the ambiguous residues, we have a perfect match to the proposed peptide's sequence from residue 3 - 12. *Why do we not have complete coverage? And why is complete coverage in most cases not necessary? Where do the ambiguous residues come from? Can they impact the final result? What is the role of modifications in the ambiguity? Can you relate this to identification issues when many (variable) modifications need to be considered? What is an immonium ion, and how can these be used in the de novo sequencing?* [1.4g]

In the 'De Novo' menu, select 'y-ions', PeptideShaker then shows you the solution retained for this spectrum:



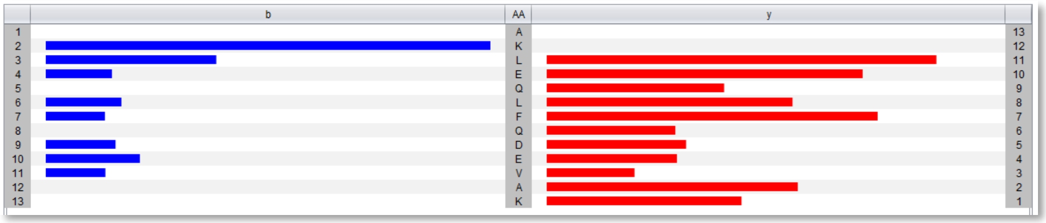
Now, enable the display of b-ions and select ‘b-ions’ in ‘De Novo’. You should see the following:



Note that for the b-ions there are less ions you can use. Remember that b-ions are indexed from the left to the right (amino-terminus to carboxy-terminus), so in this case you do not have to reverse the sequence. *Why do we have lower coverage and intensity for b-ions relative to the y-ions for the same peptide? Is this the same for all peptides, all instruments, all protocols? [1.4h]*

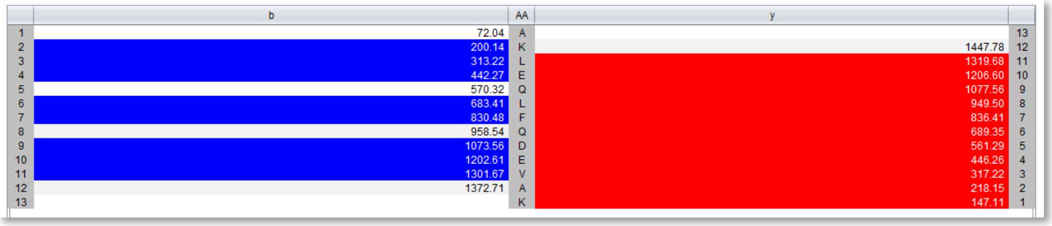
When you are done sequencing, make sure that both the b- and the y-ion types are selected in the menu below the spectrum, and click the ‘Ion Table’ tab in the lower right corner.

Here you will see an overview of the detected fragment ions and how they correspond to the sequence of the peptide. *How do the results here correspond to the results you found in your de novo sequencing? [1.4i]*



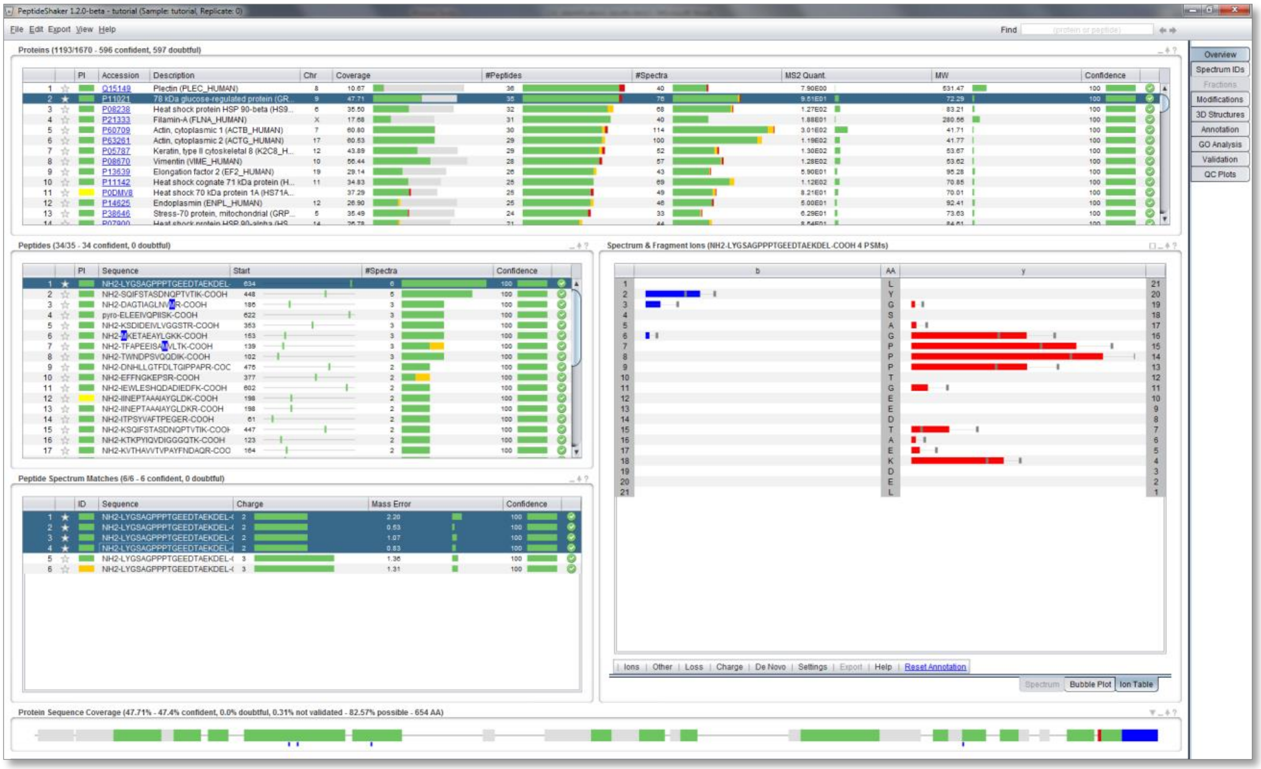
	b	AA	y	
1		A		13
2	▬	K		12
3	▬	L	▬	11
4	▬	E	▬	10
5		Q	▬	9
6	▬	L	▬	8
7	▬	F	▬	7
8		Q	▬	6
9	▬	D	▬	5
10	▬	E	▬	4
11	▬	V	▬	3
12		A	▬	2
13		K	▬	1

In the 'Settings' menu below the table you can also select m/z-based display instead of intensity:



	b	AA	y	
1	72.04	A		13
2	200.14	K	1447.78	12
3	313.22	L	1319.68	11
4	442.27	E	1206.60	10
5	570.32	Q	1077.56	9
6	683.41	L	949.50	8
7	830.48	F	836.41	7
8	958.54	Q	689.35	6
9	1073.56	D	561.29	5
10	1202.61	E	446.26	4
11	1301.67	V	317.22	3
12	1372.71	A	218.15	2
13		K	147.11	1

We will now move one step further by comparing multiple spectra identified as the same peptide. Bring the spectrum panel back to its normal size by clicking on the square and select the protein in the top of the protein list called P11021. You will see that the first peptide (LYGSAGPPPTGEEDTAEKDEL) has been identified by six spectra: four with a charge of 2+ and two with a charge of 3+. While holding down the Ctrl-key, select all spectra with a charge of 2+. You should see this:


PeptideShaker 1.2.0-beta - tutorial (Sample: tutorial, Replicate: 0)
File Edit Export View Help
Find (protein or peptide)
Overview
Spectrum IDs
Fractions
Modifications
3D Structures
Annotation
GO Analysis
Validation
QC Plots

Proteins (1193/1670 - 596 confident, 597 doubtful)

		PI	Accession	Description	Chr	Coverage	#Peptides	#Spectra	MS2 Quant.	MW	Confidence	
1	☆	🟩	Q15149	Plectin (PLEC_HUMAN)	8	10.67	36	40	7.90E00	531.47	100	✓
2	★	🟩	P11021	78 kDa glucose-regulated protein (GR...	9	47.71	35	76	9.51E01	72.29	100	✓
3	☆	🟩	P08238	Heat shock protein HSP 90-beta (HS9...	6	35.50	32	68	1.27E02	83.21	100	✓
4	☆	🟩	P21333	Filamin-A (FLNA_HUMAN)	X	17.68	31	40	1.88E01	280.66	100	✓
5	☆	🟩	P60709	Actin, cytoplasmic 1 (ACTB_HUMAN)	7	60.80	30	114	3.01E02	41.71	100	✓
6	☆	🟩	P63261	Actin, cytoplasmic 2 (ACTG_HUMAN)	17	60.53	29	100	1.19E02	41.77	100	✓
7	☆	🟩	P05787	Keratin, type II cytoskeletal 8 (K2C8_H...	12	43.89	29	52	1.30E02	53.67	100	✓
8	☆	🟩	P08670	Vimentin (VIME_HUMAN)	10	56.44	28	57	1.28E02	53.62	100	✓
9	☆	🟩	P13639	Elongation factor 2 (EF2_HUMAN)	19	29.14	26	43	5.90E01	95.28	100	✓
10	☆	🟩	P11142	Heat shock cognate 71 kDa protein (H...	11	34.83	25	69	1.12E02	70.85	100	✓
11	☆	🟨	P0DMV8	Heat shock 70 kDa protein 1A (HS71A...		37.29	25	49	8.21E01	70.01	100	✓
12	☆	🟩	P14625	Endoplasmin (ENPL_HUMAN)	12	26.90	25	48	5.00E01	92.41	100	✓
13	☆	🟩	P38646	Stress-70 protein, mitochondrial (GRP...	5	35.49	24	33	6.29E01	73.63	100	✓
14	☆	🟩	P07900	Heat shock protein HSP 90-alpha (HS	14	36.78	21	44	8.54E01	84.61	100	✓

Peptides (34/35 - 34 confident, 0 doubtful)

		PI	Sequence	Start	#Spectra	Confidence	
1	★	🟩	NH2-LYGSAGPPPTGEEDTAEKDEL-	634	6	100	✓
2	☆	🟩	NH2-SQIFSTASDNQPTVTIK-COOH	448	5	100	✓
3	☆	🟩	NH2-DAGTIAGLNVMR-COOH	186	3	100	✓
4	☆	🟩	pyro-ELEEIVQPIISK-COOH	622	3	100	✓
5	☆	🟩	NH2-KSDIDEIVLVGGSTR-COOH	353	3	100	✓
6	☆	🟩	NH2-MKETAEAYLGKK-COOH	153	3	100	✓
7	☆	🟩	NH2-TFAPEEISAMVLTK-COOH	139	3	100	✓
8	☆	🟩	NH2-TWNDPSVQQDIK-COOH	102	3	100	✓
9	☆	🟩	NH2-DNHLLGTFDLTGIPPAPR-COC	475	2	100	✓
10	☆	🟩	NH2-EFFNGKEPSR-COOH	377	2	100	✓
11	☆	🟩	NH2-IEWLESHQDADIEDFK-COOH	602	2	100	✓
12	☆	🟨	NH2-IINEPTAAAIAYGLDK-COOH	198	2	100	✓
13	☆	🟩	NH2-IINEPTAAAIAYGLDKR-COOH	198	2	100	✓
14	☆	🟩	NH2-ITPSYVAFTPEGER-COOH	61	2	100	✓
15	☆	🟩	NH2-KSQIFSTASDNQPTVTIK-COOH	447	2	100	✓
16	☆	🟩	NH2-KTKPYIQVDIGGGQTK-COOH	123	2	100	✓
17	☆	🟩	NH2-KVTHAVVTVPAYFNDAQR-COO	164	2	100	✓

Peptide Spectrum Matches (6/6 - 6 confident, 0 doubtful)

		ID	Sequence	Charge	Mass Error	Confidence	
1	★	🟩	NH2-LYGSAGPPPTGEEDTAEKDEL-(	2	2.20	100	✓
2	★	🟩	NH2-LYGSAGPPPTGEEDTAEKDEL-(	2	0.53	100	✓
3	★	🟩	NH2-LYGSAGPPPTGEEDTAEKDEL-(	2	1.07	100	✓
4	★	🟩	NH2-LYGSAGPPPTGEEDTAEKDEL-(	2	0.83	100	✓
5	☆	🟩	NH2-LYGSAGPPPTGEEDTAEKDEL-(	3	1.36	100	✓
6	☆	🟨	NH2-LYGSAGPPPTGEEDTAEKDEL-(	3	1.31	100	✓

Spectrum & Fragment Ions (NH2-LYGSAGPPPTGEEDTAEKDEL-COOH 4 PSMs)

	b	AA	y	
1		L		21
2		Y		20
3		G		19
4		S		18
5		A		17
6		G		16
7		P		15
8		P		14
9		P		13
10		T		12
11		G		11
12		E		10
13		E		9
14		D		8
15		T		7
16		A		6
17		E		5
18		K		4
19		D		3
20		E		2
21		L		1

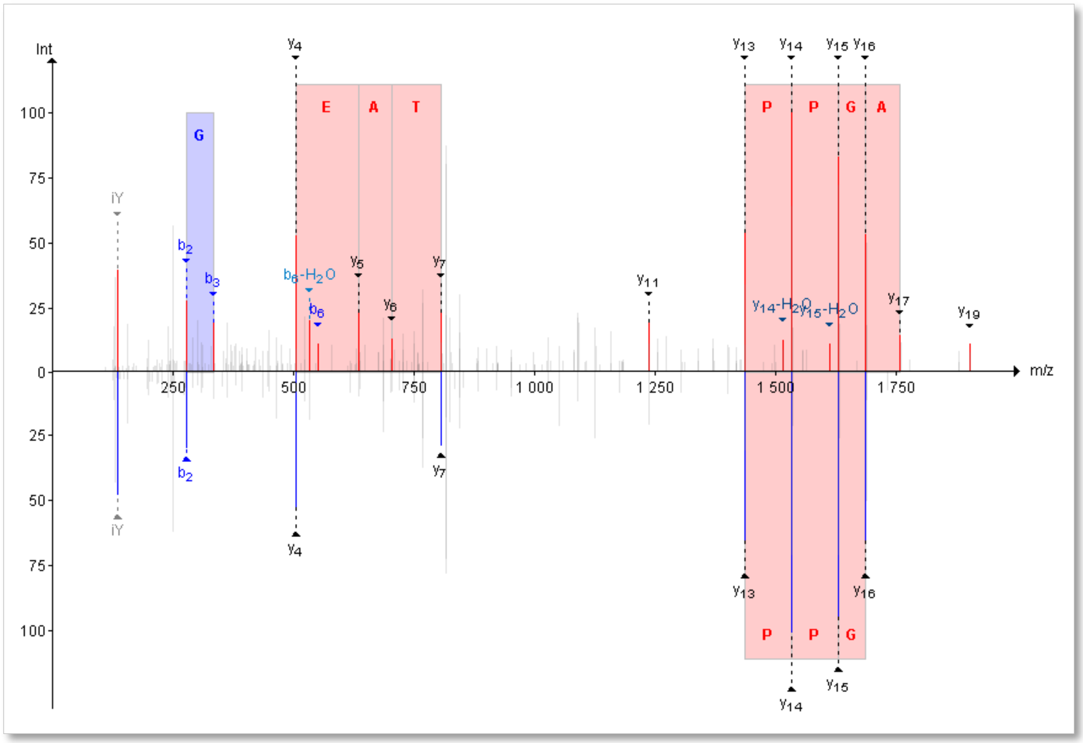
Ions | Other | Loss | Charge | De Novo | Settings | Export | Help | Reset Annotation
Spectrum
Bubble Plot
Ion Table

Protein Sequence Coverage (47.71% - 47.4% confident, 0.0% doubtful, 0.31% not validated - 82.57% possible - 654 AA)

The intensity difference between the spectra is displayed with error bars. *Do you expect reproducibility in intensities between spectra? What do you think about the comparison between 2+ and 3+ PSMs?* [1.4j]

In a standard visual inspection of both spectra (select them one by one in the PSM table and look at each spectrum) it can be difficult to tell how similar they are. Attempting to see which fragment ions are detected in all of the spectra, and at which intensities, is also very hard.

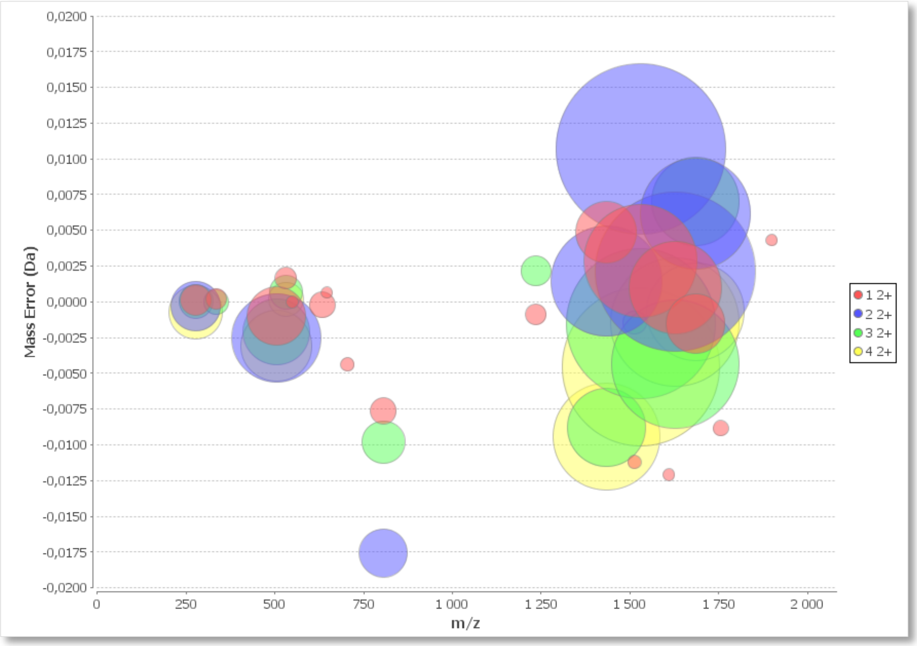
If comparing two spectra so-called mirrored spectra can be used, i.e., showing one spectrum above the x-axis and another spectrum mirrored below the x-axis. Select only the two first PSMs in the table and go back to the Spectrum tab. You should now see the following:



The use of mirrored spectra allows you to easily compare two spectra and highlight the differences between them. (Note that in order to make the intensities comparable the peaks are shown relative to the highest peak in each spectrum.)

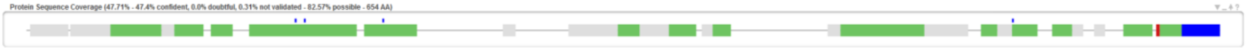
When comparing more than two spectra PeptideShaker uses a novel view, referred to as the bubble view or the *Planetary System View*. You will find this view by clicking the 'Bubble Plot' tab below the spectrum (or by simply selecting more than two rows in the spectrum table). Selecting/highlighting all the rows in the PSM table with a 2+ charge should give you something like the figure below.

In this view of the spectra we still have the m/z value on the x-axis, but on the y-axis we now have the mass error (the distance between the theoretical and the experimental mass of the fragment ions). The size of each bubble represents the (normalized) intensity of each fragment ion.



This way of looking at multiple spectra takes a bit of getting used to, but once one is comfortable with the setup, it becomes very easy to pick up details that would have been very difficult to see by a manual inspection of each spectrum individually. *Now, what do you think about the error distributions? [1.4k]*

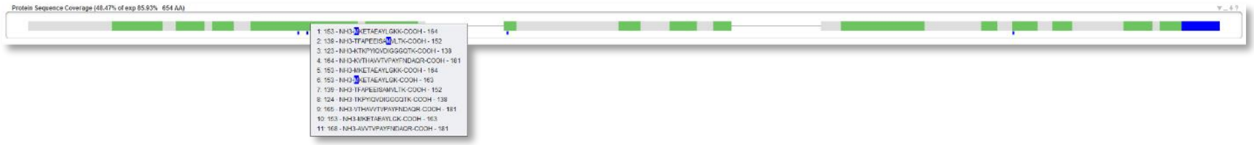
We will now study the sequence coverage of the selected protein – P11021. The sequence coverage is indicated in the protein table, 47.71% and illustrated at the bottom in the Protein Sequence Coverage panel:


Protein Sequence Coverage (47.71% - 47.4% confident, 0.0% doubtful, 0.31% not validated - 82.57% possible - 654 AA)

The 35 identified peptides are mapped onto the sequence in green, yellow and red (color coding will be explained later), whereas the non-covered parts are left in grey. Note that the parts of the sequence which do not generate observable peptides are thinner than the others. For this protein, the observable coverage was estimated to 69.88% - the observable coverage is also displayed in gray in the protein table.

If you hold the mouse over a part of the sequence, the observed peptides will be displayed. *The currently selected peptide is displayed in blue, what is the first peptide observed in the sequence?* [1.4l]

Note that this peptide contains a missed cleavage and that the cleaved version is also detected. Also, keen observers will have noticed that the modifications are readily mapped above the sequence. Clicking at the corresponding peptide section allows the selection of the oxidized peptide:


Protein Sequence Coverage (48.47% of exp 85.93% 654 AA)

```
1: 153 - NH3-MKETAEAYLGKK-COOH - 164
2: 139 - NH3-TFAPEEISAMVLTK-COOH - 152
3: 123 - NH3-KTKPYIQVDIGGGQTK-COOH - 138
4: 164 - NH3-KVTHAVVTVPAYFNDAQR-COOH - 181
5: 153 - NH3-MKETAEAYLGKK-COOH - 164
6: 153 - NH3-MKETAEAYLGK-COOH - 163
7: 139 - NH3-TFAPEEISAMVLTK-COOH - 152
8: 124 - NH3-TKPYIQVDIGGGQTK-COOH - 138
9: 165 - NH3-VTHAVVTVPAYFNDAQR-COOH - 181
10: 153 - NH3-MKETAEAYLGK-COOH - 163
11: 168 - NH3-AVVTVPAYFNDAQR-COOH - 181
```

How many modified peptides does this protein contain? How many modification sites? [1.4m]

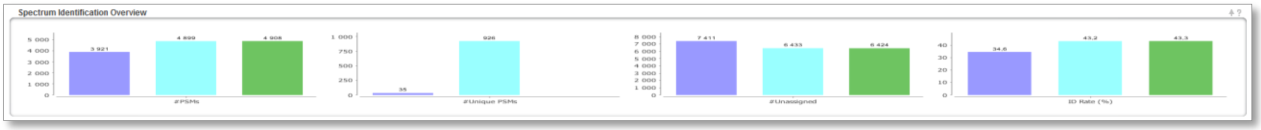
More details on PTM localization can be found in the “PTM Analysis” chapter.

B. Quality Control

Before further investigation, it is important to verify the quality of the search. We will start by verifying that all search engines worked correctly. As you remember, we loaded identification results from X! Tandem and OMSSA. Open the tab called ‘Spectrum IDs’ in the upper right corner, loading a tab can take a few seconds for big datasets.

> **Tip:**
>
> *When PeptideShaker is busy with a process its icon turns orange. When finished, it turns grey again to inform you that the coffee break is over!* ☺

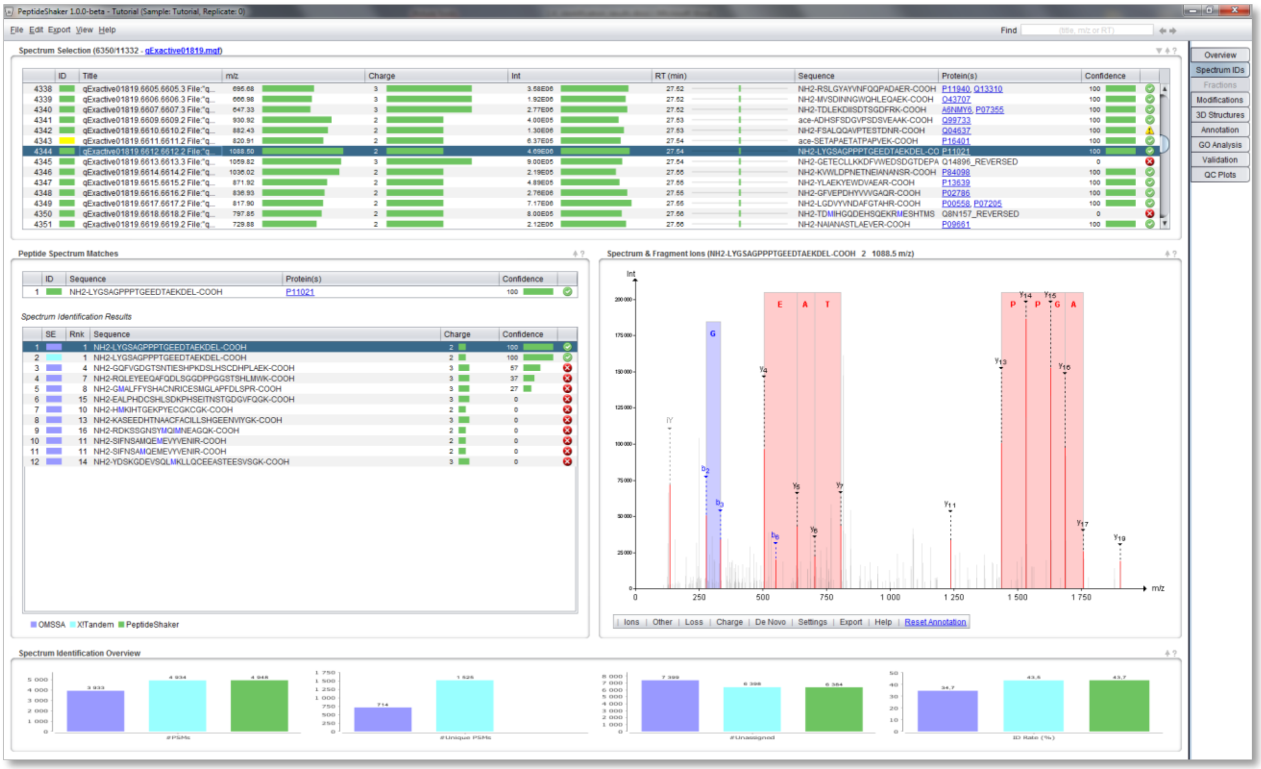
At the bottom of the tab you will see something like this:



Here you see how the search engine results compare at the PSM level. Bar charts that for each search engine shows the number of validated PSMs, unique PSMs, unassigned spectra and the ID rate.

Why are the search engine results different? Will the differences depend on the dataset used? Can we take advantage of these differences? *[1.4n]*

The rest of the 'Spectrum IDs' tab allows you to browse the search engine results and see how PeptideShaker combines the separate search engine results into a single result set. By clicking a spectrum in the spectrum table at top you will see the PSMs for the search engines used and how these were scored. The selected PSM will be used to annotate the given spectrum.



Spectrum Selection (6350/11332 - qExactive01819.mgf)

	ID	Title	m/z	Charge	Int	RT (min)	Sequence	Protein(s)	Confidence	
4338	🟩	qExactive01819.6605.6605.3 File:"q...	695.68	3	3.58E06	27.52	NH2-RSLGYAYVNFQQPADAER-COOH	P11940, Q13310	100	✓
4339	🟩	qExactive01819.6606.6606.3 File:"q...	666.98	3	1.92E06	27.52	NH2-MVSDINNGWQHLEQAEK-COOH	O43707	100	✓
4340	🟩	qExactive01819.6607.6607.3 File:"q...	647.33	3	2.77E06	27.52	NH2-TDLEKDIISDTSGDFRK-COOH	A6NMY6, P07355	100	✓
4341	🟩	qExactive01819.6609.6609.2 File:"q...	930.92	2	4.00E05	27.53	ace-ADHSFSDGVPSDSVEAAK-COOH	Q99733	100	✓
4342	🟩	qExactive01819.6610.6610.2 File:"q...	882.43	2	1.30E06	27.53	NH2-FSALQQAVPTESTDNR-COOH	Q04637	100	⚠
4343	🟨	qExactive01819.6611.6611.2 File:"q...	820.91	2	6.37E05	27.54	ace-SETAPAETATPAPVEK-COOH	P16401	100	✓
4344	🟩	qExactive01819.6612.6612.2 File:"q...	1088.50	2	4.69E06	27.54	NH2-LYGSAGPPPTGEEDTAEKDEL-CO	P11021	100	✓
4345	🟩	qExactive01819.6613.6613.3 File:"q...	1059.82	3	9.00E05	27.54	NH2-GETECLLKKDFVWEDSDGTDEPA	Q14896_REVERSED	0	✗
4346	🟩	qExactive01819.6614.6614.2 File:"q...	1036.02	2	2.19E05	27.56	NH2-KVWLDPNETNEIANANSR-COOH	P84098	100	✓
4347	🟩	qExactive01819.6615.6615.2 File:"q...	871.92	2	4.89E05	27.56	NH2-YLAEKYEWDVAEAR-COOH	P13639	100	✓
4348	🟩	qExactive01819.6616.6616.2 File:"q...	836.93	2	2.76E06	27.56	NH2-GFVEPDHYVVVGAQR-COOH	P02786	100	✓
4349	🟩	qExactive01819.6617.6617.2 File:"q...	817.90	2	7.17E06	27.55	NH2-LGDVYVNDAFGTAHR-COOH	P00558, P07205	100	✓
4350	🟩	qExactive01819.6618.6618.2 File:"q...	797.85	2	8.00E05	27.56	NH2-TDMIHGQDEHSQEKRMESHTMS	Q8N157_REVERSED	0	✗
4351	🟩	qExactive01819.6619.6619.2 File:"q...	729.88	2	2.12E06	27.56	NH2-NAIANASTLAEVER-COOH	P09661	100	✓

Peptide Spectrum Matches

	ID	Sequence	Protein(s)	Confidence	
1	🟩	NH2-LYGSAGPPPTGEEDTAEKDEL-COOH	P11021	100	✓

Spectrum Identification Results

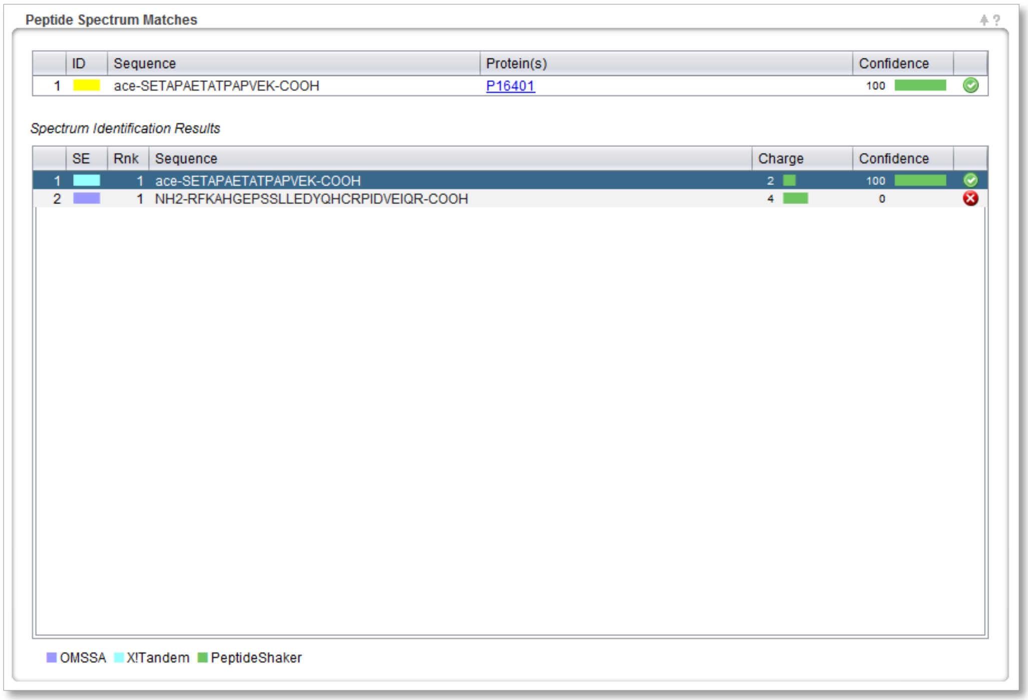
	SE	Rnk	Sequence	Charge	Confidence	
1	🟪	1	NH2-LYGSAGPPPTGEEDTAEKDEL-COOH	2	100	✓
2	🟦	1	NH2-LYGSAGPPPTGEEDTAEKDEL-COOH	2	100	✓
3	🟪	4	NH2-GQFVGDGTSNTIESHPKDSLHSCDHPLAEK-COOH	3	57	✗
4	🟪	7	NH2-RQLEYEEQAFQDLSGGDPPGGSTSHLMWK-COOH	3	37	✗
5	🟪	8	NH2-GMALFFYSHACNRICESMGLAPFDLSPR-COOH	3	27	✗
6	🟪	15	NH2-EALPHDCSHLSDKPHSEITNSTGDGVFQGK-COOH	3	0	✗
7	🟪	10	NH2-HMKIHTGEKPYECGKCGK-COOH	2	0	✗
8	🟪	13	NH2-KASEEDHTNAACFACILLSHGEENVIYGK-COOH	3	0	✗
9	🟪	16	NH2-RDKSSGNSYMQIMNEAGQK-COOH	2	0	✗
10	🟪	11	NH2-SIFNSAMQEMEVYVENIR-COOH	2	0	✗
11	🟪	11	NH2-SIFNSAMQEMEVYVENIR-COOH	2	0	✗
12	🟪	14	NH2-YDSKGDEVSQLMKLLQCEEASTEESVSGK-COOH	3	0	✗

🟪 OMSSA 🟦 X!Tandem 🟩 PeptideShaker

Tip:

If you are interested in non-assigned spectra, you can export them using the 'Follow Up Analysis' option in the 'Export' menu.

Note the ID column which indicates whether a spectrum was identified and if so, whether the search engines agreed on the identification. Select row 4343, spectrum qExactive01819.6611.6611.2 File:"qExactive01819.raw", NativeID:"controllerType=0 controllerNumber=1 scan=6611", with a yellow square. In the Peptide-Spectrum Matches section you should see this:



Peptide Spectrum Matches

	ID	Sequence	Protein(s)	Confidence	
1	🟨	ace-SETAPAETATPAPVEK-COOH	P16401	100 🟩	✓

Spectrum Identification Results

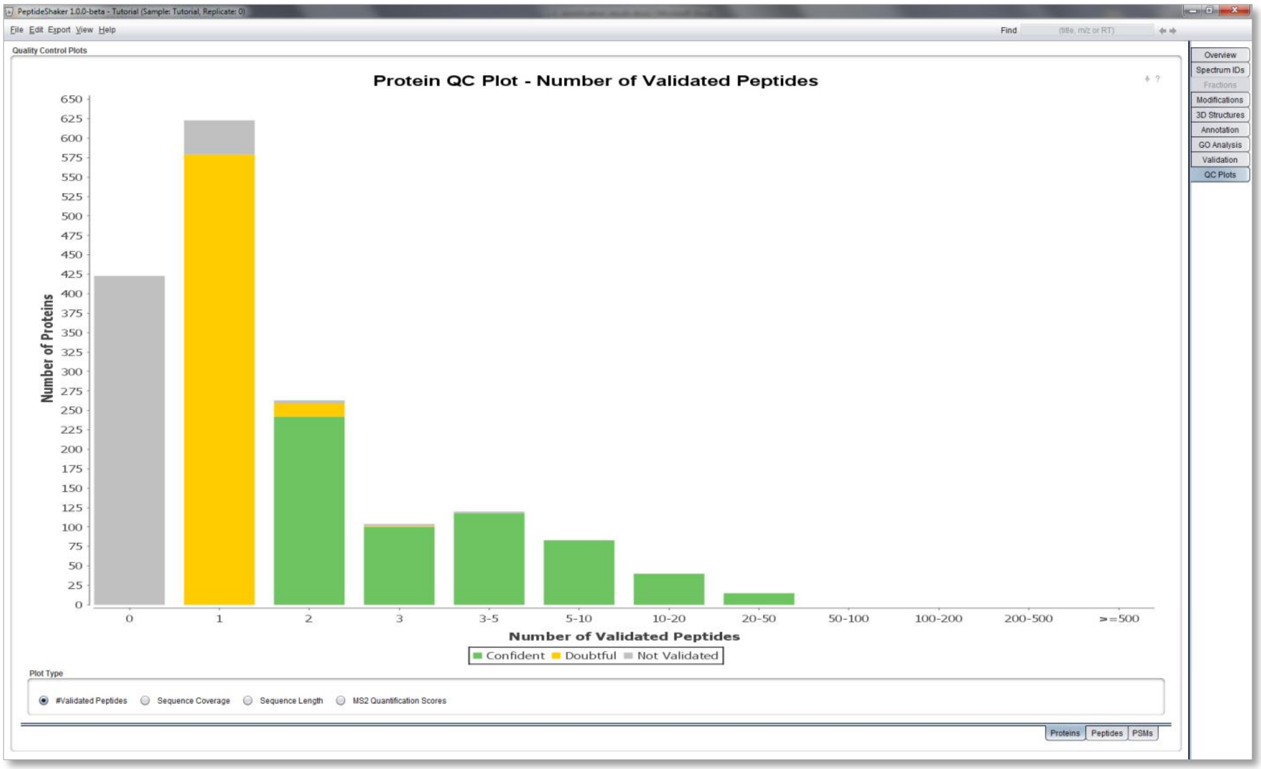
	SE	Rnk	Sequence	Charge	Confidence	
1	🟦	1	ace-SETAPAETATPAPVEK-COOH	2 🟩	100 🟩	✓
2	🟪	1	NH2-RFKAHGEPSSLLEDYQHCRPIDVEIQR-COOH	4 🟩	0	✗

🟪 OMSSA 🟦 X!Tandem 🟩 PeptideShaker

As you can see, a peptide found by X! Tandem took precedence over others suggested by OMSSA. *Which of the possibilities is the best match in your opinion? Would you trust peptides where the search engines disagree?* *[1.4o]*

Now, select the 'QC Plots' tab. In this tab, a list of Quality Control metrics is given for the proteins, peptides and PSMs. The category (Proteins, Peptides, PSMs) is selected at the bottom right corner of the screen.

The default view is the Protein option, and you should now see the distribution of proteins according to their number of identified peptides:



How many peptides would you require to trust a protein identification? [1.4p]

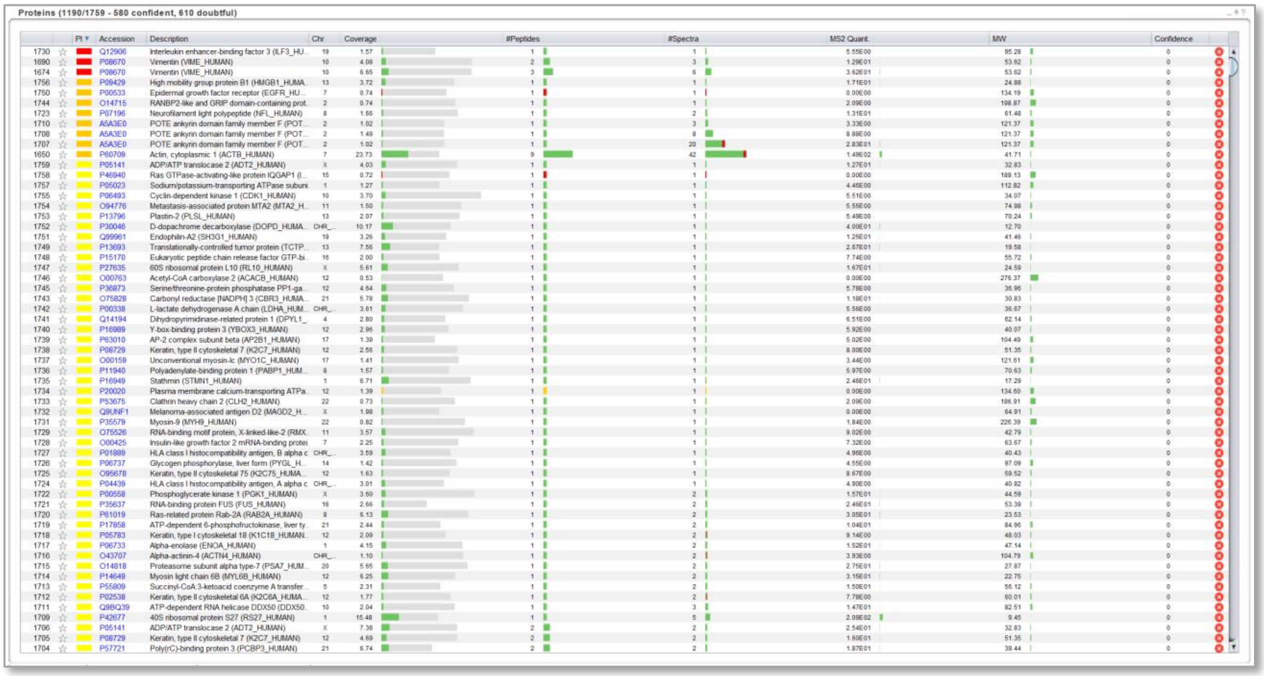
To save time, we will not go through all quality control plots but feel free to explore them by yourself. We welcome any questions/suggestions!

C. Protein Inference

When trying to figure out which protein a given set of peptides originated from, it is in some cases impossible to decide between two or more proteins that share the same peptide(s). This issue is known as the *protein inference problem*.[3]

Go to the 'Overview' tab and sort the proteins by the PI column by clicking the column header and click the header once more to sort in descending order. PI stands for Protein Inference, and the proteins are now sorted so that the protein groups with protein inference issues are at the top.

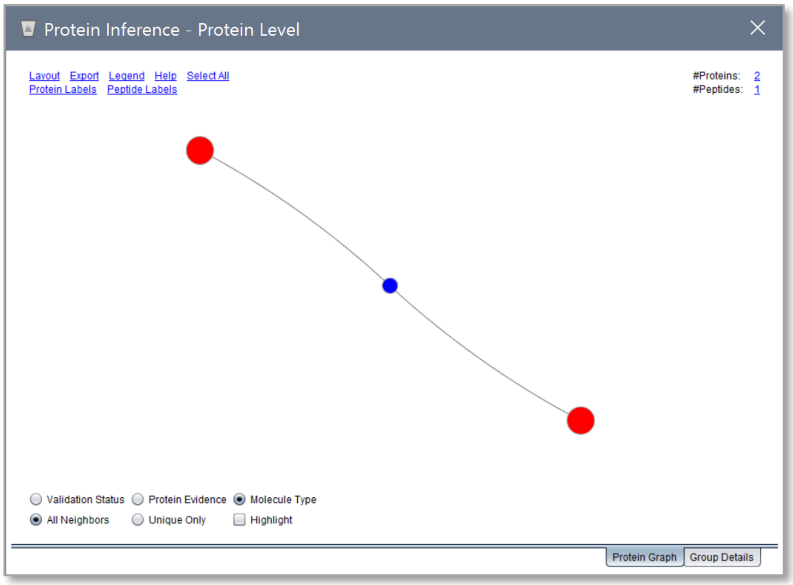
In PeptideShaker, proteins are flagged using four colors: green, yellow, orange and red. For the best overview, hide all sections but the Protein table:



Proteins (1190/1759 - 580 confident, 610 doubtful)

		PI ▼	Accession	Description	Chr	Coverage	#Peptides	#Spectra	MS2 Quant.	MW	Confidence	
1730	☆	🟥	Q12906	Interleukin enhancer-binding factor 3 (ILF3_HU...	19	1.57	1	1	5.55E00	95.28	0	⊗
1690	☆	🟥	P08670	Vimentin (VIME_HUMAN)	10	4.08	2	3	1.29E01	53.62	0	⊗
1674	☆	🟥	P08670	Vimentin (VIME_HUMAN)	10	6.65	3	6	3.62E01	53.62	0	⊗
1756	☆	🟧	P09429	High mobility group protein B1 (HMGB1_HUMA...	13	3.72	1	1	1.71E01	24.88	0	⊗
1750	☆	🟧	P00533	Epidermal growth factor receptor (EGFR_HU...	7	0.74	1	1	0.00E00	134.19	0	⊗
1744	☆	🟧	O14715	RANBP2-like and GRIP domain-containing prot...	2	0.74	1	1	2.09E00	198.87	0	⊗
1723	☆	🟧	P07196	Neurofilament light polypeptide (NFL_HUMAN)	8	1.66	1	2	1.31E01	61.48	0	⊗
1710	☆	🟧	A5A3E0	POTE ankyrin domain family member F (POT...	2	1.02	1	3	3.33E00	121.37	0	⊗
1708	☆	🟧	A5A3E0	POTE ankyrin domain family member F (POT...	2	1.49	1	8	8.88E00	121.37	0	⊗
1707	☆	🟧	A5A3E0	POTE ankyrin domain family member F (POT...	2	1.02	1	20	2.83E01	121.37	0	⊗
1650	☆	🟧	P60709	Actin, cytoplasmic 1 (ACTB_HUMAN)	7	23.73	9	42	1.49E02	41.71	0	⊗
1759	☆	🟨	P05141	ADP/ATP translocase 2 (ADT2_HUMAN)	X	4.03	1	1	1.27E01	32.83	0	⊗
1758	☆	🟨	P46940	Ras GTPase-activating-like protein IQGAP1 (I...	15	0.72	1	1	0.00E00	189.13	0	⊗
1757	☆	🟨	P05023	Sodium/potassium-transporting ATPase subuni...	1	1.27	1	1	4.46E00	112.82	0	⊗
1755	☆	🟨	P06493	Cyclin-dependent kinase 1 (CDK1_HUMAN)	10	3.70	1	1	5.51E00	34.07	0	⊗
1754	☆	🟨	O94776	Metastasis-associated protein MTA2 (MTA2_H...	11	1.50	1	1	5.55E00	74.98	0	⊗
1753	☆	🟨	P13796	Plastin-2 (PLSL_HUMAN)	13	2.07	1	1	5.49E00	70.24	0	⊗
1752	☆	🟨	P30046	D-dopachrome decarboxylase (DOPD_HUMA...	CHR_...	10.17	1	1	4.00E01	12.70	0	⊗
1751	☆	🟨	Q99961	Endophilin-A2 (SH3G1_HUMAN)	19	3.26	1	1	1.25E01	41.46	0	⊗
1749	☆	🟨	P13693	Translationally-controlled tumor protein (TCTP...	13	7.56	1	1	2.67E01	19.58	0	⊗
1748	☆	🟨	P15170	Eukaryotic peptide chain release factor GTP-bi...	16	2.00	1	1	7.74E00	55.72	0	⊗
1747	☆	🟨	P27635	60S ribosomal protein L10 (RL10_HUMAN)	X	5.61	1	1	1.67E01	24.59	0	⊗
1746	☆	🟨	O00763	Acetyl-CoA carboxylase 2 (ACACB_HUMAN)	12	0.53	1	1	0.00E00	276.37	0	⊗
1745	☆	🟨	P36873	Serine/threonine-protein phosphatase PP1-ga...	12	4.64	1	1	5.78E00	36.96	0	⊗
1743	☆	🟨	O75828	Carbonyl reductase [NADPH] 3 (CBR3_HUMA...	21	5.78	1	1	1.18E01	30.83	0	⊗
1742	☆	🟨	P00338	L-lactate dehydrogenase A chain (LDHA_HUM...	CHR_...	3.61	1	1	5.56E00	36.67	0	⊗
1741	☆	🟨	Q14194	Dihydropyrimidinase-related protein 1 (DPYL1_...	4	2.80	1	1	6.51E00	62.14	0	⊗
1740	☆	🟨	P16989	Y-box-binding protein 3 (YBOX3_HUMAN)	12	2.96	1	1	5.92E00	40.07	0	⊗
1739	☆	🟨	P63010	AP-2 complex subunit beta (AP2B1_HUMAN)	17	1.39	1	1	5.02E00	104.49	0	⊗
1738	☆	🟨	P08729	Keratin, type II cytoskeletal 7 (K2C7_HUMAN)	12	2.56	1	1	8.00E00	51.35	0	⊗
1737	☆	🟨	O00159	Unconventional myosin-Ic (MYO1C_HUMAN)	17	1.41	1	1	3.44E00	121.61	0	⊗
1736	☆	🟨	P11940	Polyadenylate-binding protein 1 (PABP1_HUM...	8	1.57	1	1	5.97E00	70.63	0	⊗
1735	☆	🟨	P16949	Stathmin (STMN1_HUMAN)	1	6.71	1	1	2.46E01	17.29	0	⊗
1734	☆	🟨	P20020	Plasma membrane calcium-transporting ATPa...	12	1.39	1	1	0.00E00	134.60	0	⊗
1733	☆	🟨	P53675	Clathrin heavy chain 2 (CLH2_HUMAN)	22	0.73	1	1	2.09E00	186.91	0	⊗
1732	☆	🟨	Q9UNF1	Melanoma-associated antigen D2 (MAGD2_H...	X	1.98	1	1	0.00E00	64.91	0	⊗
1731	☆	🟨	P35579	Myosin-9 (MYH9_HUMAN)	22	0.82	1	1	1.84E00	226.39	0	⊗
1729	☆	🟨	O75526	RNA-binding motif protein, X-linked-like-2 (RMX...	11	3.57	1	1	9.02E00	42.79	0	⊗
1728	☆	🟨	O00425	Insulin-like growth factor 2 mRNA-binding protei...	7	2.25	1	1	7.32E00	63.67	0	⊗
1727	☆	🟨	P01889	HLA class I histocompatibility antigen, B alpha c...	CHR_...	3.59	1	1	4.96E00	40.43	0	⊗
1726	☆	🟨	P06737	Glycogen phosphorylase, liver form (PYGL_H...	14	1.42	1	1	4.55E00	97.09	0	⊗
1725	☆	🟨	O95678	Keratin, type II cytoskeletal 75 (K2C75_HUMA...	12	1.63	1	1	8.67E00	59.52	0	⊗
1724	☆	🟨	P04439	HLA class I histocompatibility antigen, A alpha c...	CHR_...	3.01	1	1	4.90E00	40.82	0	⊗
1722	☆	🟨	P00558	Phosphoglycerate kinase 1 (PGK1_HUMAN)	X	3.60	1	2	1.57E01	44.59	0	⊗
1721	☆	🟨	P35637	RNA-binding protein FUS (FUS_HUMAN)	16	2.66	1	2	2.46E01	53.39	0	⊗
1720	☆	🟨	P61019	Ras-related protein Rab-2A (RAB2A_HUMAN)	8	6.13	1	2	3.05E01	23.53	0	⊗
1719	☆	🟨	P17858	ATP-dependent 6-phosphofructokinase, liver ty...	21	2.44	1	2	1.04E01	84.96	0	⊗
1718	☆	🟨	P05783	Keratin, type I cytoskeletal 18 (K1C18_HUMAN...	12	2.09	1	2	9.14E00	48.03	0	⊗
1717	☆	🟨	P06733	Alpha-enolase (ENOA_HUMAN)	1	4.15	1	2	1.52E01	47.14	0	⊗
1716	☆	🟨	O43707	Alpha-actinin-4 (ACTN4_HUMAN)	CHR_...	1.10	1	2	3.93E00	104.79	0	⊗
1715	☆	🟨	O14818	Proteasome subunit alpha type-7 (PSA7_HUM...	20	5.65	1	2	2.75E01	27.87	0	⊗
1714	☆	🟨	P14649	Myosin light chain 6B (MYL6B_HUMAN)	12	6.25	1	2	3.15E01	22.75	0	⊗
1713	☆	🟨	P55809	Succinyl-CoA:3-ketoacid coenzyme A transfer...	5	2.31	1	2	1.50E01	56.12	0	⊗
1712	☆	🟨	P02538	Keratin, type II cytoskeletal 6A (K2C6A_HUMA...	12	1.77	1	2	7.78E00	60.01	0	⊗
1711	☆	🟨	Q9BQ39	ATP-dependent RNA helicase DDX50 (DDX50...	10	2.04	1	3	1.47E01	82.51	0	⊗
1709	☆	🟨	P42677	40S ribosomal protein S27 (RS27_HUMAN)	1	15.48	1	5	2.09E02	9.45	0	⊗
1706	☆	🟨	P05141	ADP/ATP translocase 2 (ADT2_HUMAN)	X	7.38	2	2	2.54E01	32.83	0	⊗
1705	☆	🟨	P08729	Keratin, type II cytoskeletal 7 (K2C7_HUMAN)	12	4.69	2	2	1.60E01	51.35	0	⊗
1704	☆	🟨	P57721	Poly(rC)-binding protein 3 (PCBP3_HUMAN)	21	6.74	2	2	1.87E01	39.44	0	⊗

Click on the yellow box corresponding to protein accession number Q8WYB5, and a Protein Inference Dialog appears, displaying that this group consists of two proteins (in red) sharing a single peptide (in blue).



Protein Inference - Protein Level

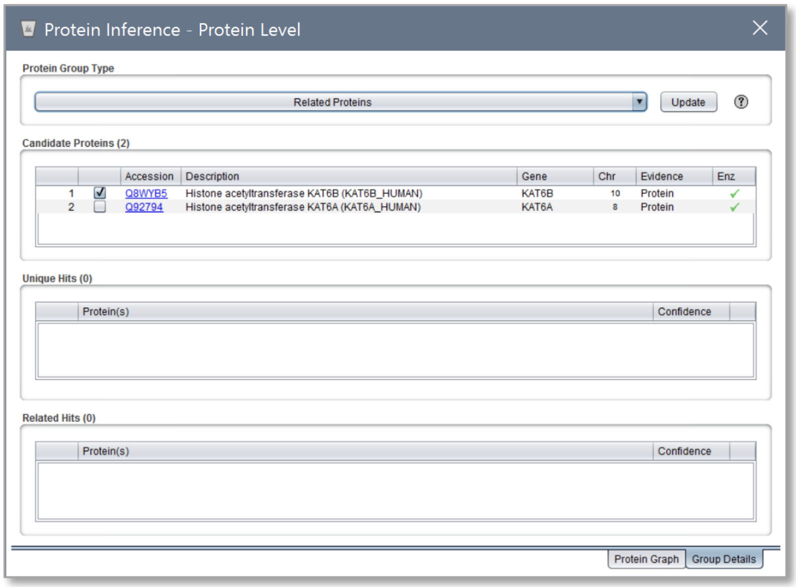
Layout Export Legend Help Select All
Protein Labels Peptide Labels

#Proteins: 2
#Peptides: 1

○ Validation Status ○ Protein Evidence ● Molecule Type
● All Neighbors ○ Unique Only □ Highlight

Protein Graph | Group Details

Select 'Group Details' option in the lower right corner to see the protein details:



Protein Inference - Protein Level

Protein Group Type

Related Proteins ▾ Update ?

Candidate Proteins (2)

		Accession	Description	Gene	Chr	Evidence	Enz
1	☑	Q8WYB5	Histone acetyltransferase KAT6B (KAT6B_HUMAN)	KAT6B	10	Protein	✓
2	□	Q92794	Histone acetyltransferase KAT6A (KAT6A_HUMAN)	KAT6A	8	Protein	✓

Unique Hits (0)

	Protein(s)	Confidence

Related Hits (0)

	Protein(s)	Confidence

Protein Graph | Group Details

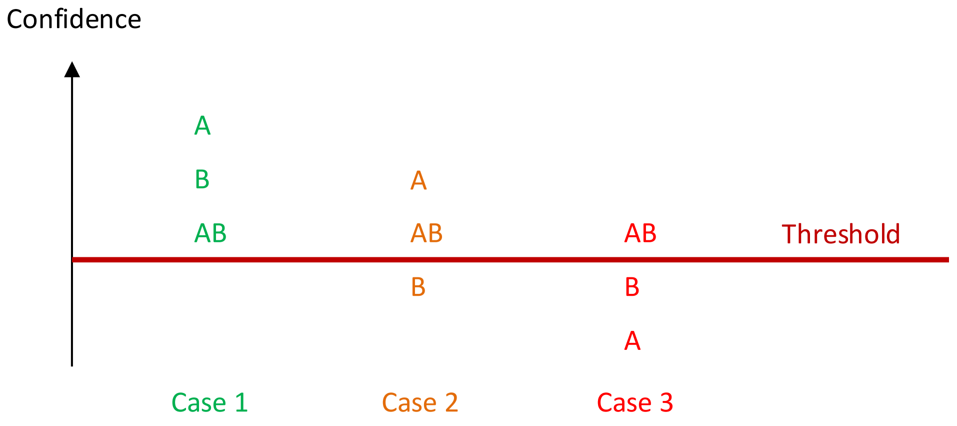
PeptideShaker could not decide whether Q8WYB5 or Q92794 was present in the sample as the identified peptide is shared by the both proteins. *According to the information provided, which protein suggestion would you trust? What do the Evidence and Enz columns represent, and how can this information be used? [1.4q]*

Note that proteins in a protein group can be attached to different genes or chromosomes. Choosing one or the other can hence strongly impact the biological interpretation of the results! PeptideShaker always chooses the protein with enzymatic peptides and with the best evidence according to UnitProt[4].

Tip:

In order to benefit from comprehensive protein annotation, we strongly recommend the use of UniProt databases.

Inferring proteins from peptides is not as simple as it may seem. Some peptides can map to different proteins, let's say protein A and protein B. If no unique peptide unambiguously indicates the presence of either protein A or B, PeptideShaker cannot say whether protein A or protein B is present.



In this simple example, how do you interpret the various cases? [1.4r]

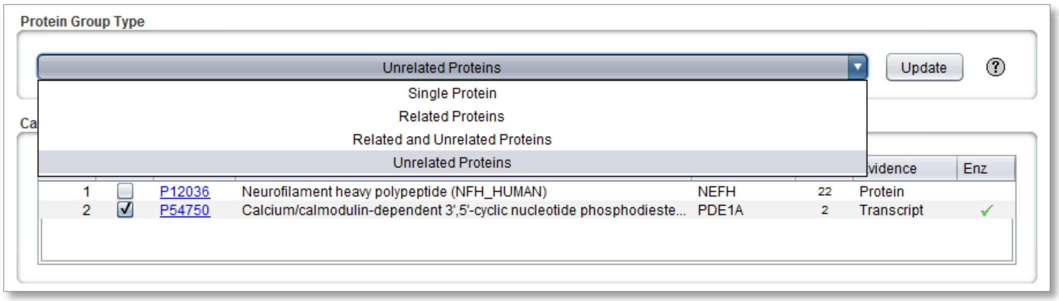
Often, the problem is a lot more complex and involves dozens of intricate peptide-to-protein mappings. PeptideShaker solves protein inferences of Case 1 and 2, as you may have noticed when loading the identification files. However, Case 3 requires additional information about the sample to find out, if possible, which protein was actually present. These types of conflicts are in the PI column colored in yellow, orange and red depending on the properties of the conflict. By clicking on the corresponding box (as we have done above), you can manually select the protein you consider as the correct one.

In most cases, the protein groups consist of related proteins. Based on the protein gene name and the protein descriptions, PeptideShaker sorts the protein inference conflicts into four categories:

Green:	No Conflict, i.e., a Single Protein unambiguously identified
Yellow:	Group of related proteins
Orange:	Group of Related and Unrelated Proteins
Red:	Group of Unrelated Proteins

How accurate is this type of sorting? [1.4s]

The conflict type can be changed in the Protein Inference dialog:



Protein Group Type

Unrelated Proteins ▾ Update ?

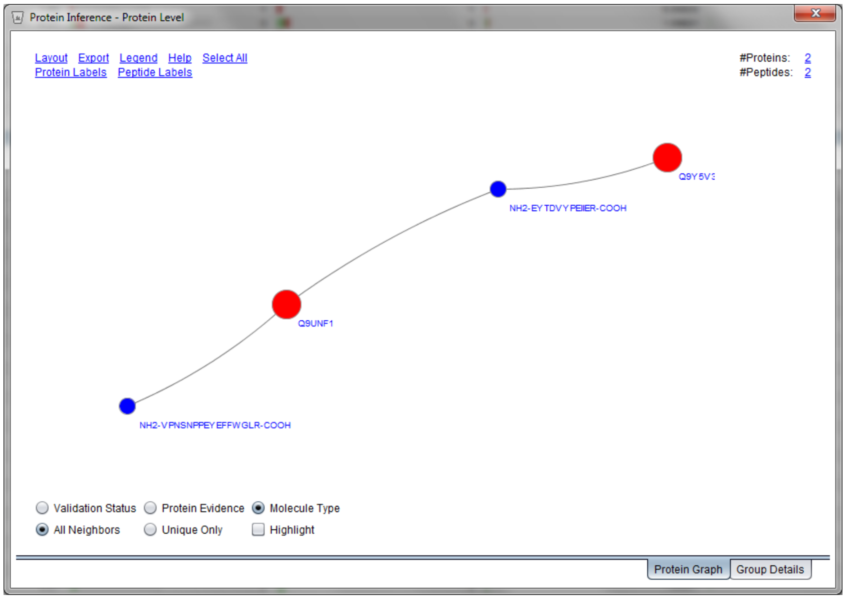
- Single Protein
- Related Proteins
- Related and Unrelated Proteins
- Unrelated Proteins

Ca

						vidence	Enz
1	☐	P12036	Neurofilament heavy polypeptide (NFH_HUMAN)	NEFH	22	Protein	
2	☑	P54750	Calcium/calmodulin-dependent 3',5'-cyclic nucleotide phosphodieste...	PDE1A	2	Transcript	✓

If the type is changed, the color for the group is updated in the PI column of the protein table.

Now look for protein Q9UNF1, and click on the yellow box in the PI column. The Protein Inference dialog below will appear. Turn on both the Protein and Peptide Labels in the upper left corner:



Protein Inference - Protein Level

Layout Export Legend Help Select All
Protein Labels Peptide Labels

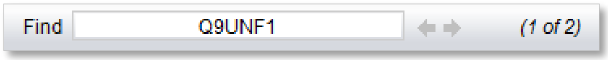
#Proteins: 2
#Peptides: 2

○ Validation Status ○ Protein Evidence ● Molecule Type
● All Neighbors ○ Unique Only ☐ Highlight

Protein Graph | Group Details

Here, a peptide belonging to Q9UNF1 was found shared with Q9Y5V3, and the group was detected with a high confidence (100%). Additionally, the dialog also shows that Q9UNF1 was also identified by unique peptides but with lower confidence. As a result, Q9UNF1 will appear twice in the protein table: once as a group (Q9UNF1 or Q9Y5V3) and once as unique hit (Q9UNF1).

In order to navigate such hits, we will use the Find feature of PeptideShaker: type Q9UNF1 in the top right corner of the interface.



Find Q9UNF1 (1 of 2)

Using the two arrows, you can navigate the two groups related to this accession. *Select the unique protein hit, what is the unique peptide for Q9UNF1 and what do you think of the quality of the peptide to spectrum match? [1.4t]*

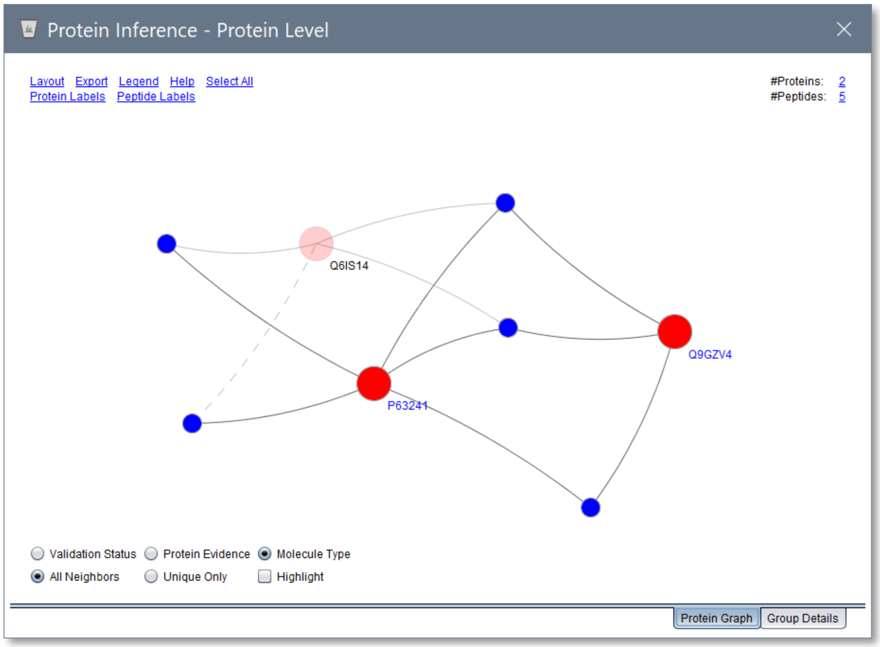
Note that you can change the retained accession for any group by selecting the protein in the first column in the Protein Inference dialog.

Can we delete a useless group? [1.4u]

Tip:

The Find function accepts protein accessions, descriptions and even peptide sequences!

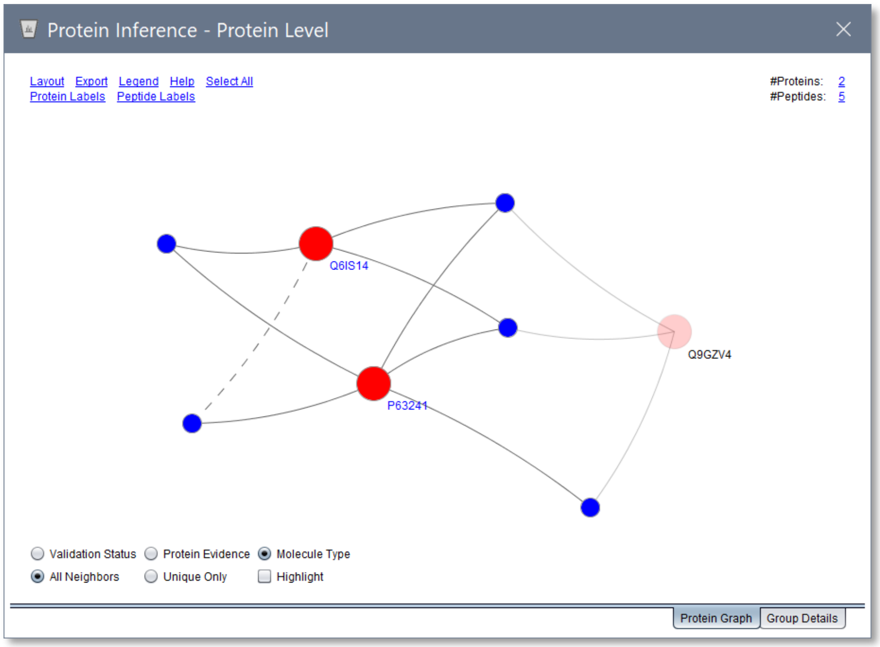
Similarly, if peptides shared with one of these proteins were found in another group, the group is reported in the 'Related Hits'. See for example the group "P63241 or Q9GZV4", were P63241 is shared with the group "P63241 or Q6IS14":



Protein Inference - Protein Level

#Proteins: 2
#Peptides: 5

○ Validation Status ○ Protein Evidence ● Molecule Type
● All Neighbors ○ Unique Only □ Highlight

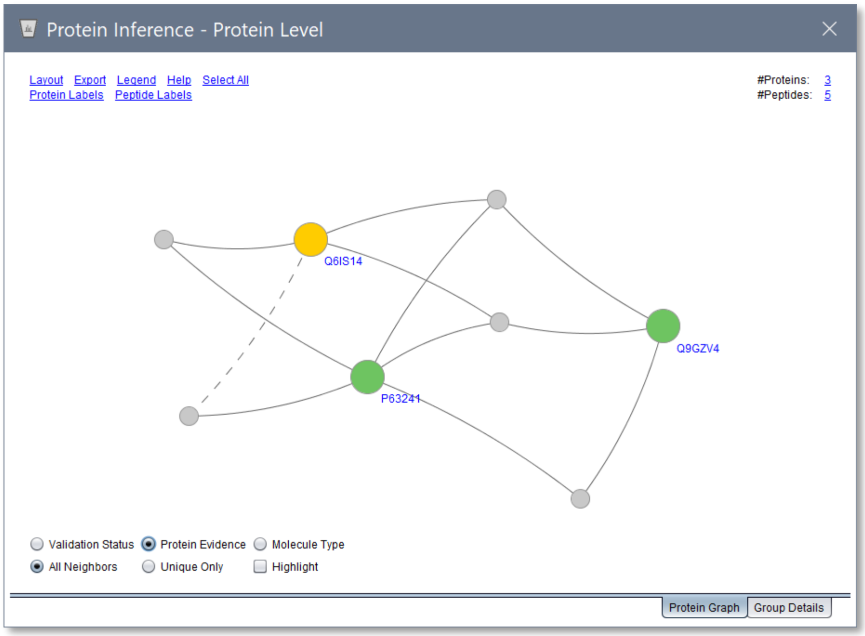


Protein Inference - Protein Level

#Proteins: 2
#Peptides: 5

○ Validation Status ○ Protein Evidence ● Molecule Type
● All Neighbors ○ Unique Only □ Highlight

In cases such as this we might include additional evidence to try to decide between the groups. As an example, click the Protein Evidence option below the graph:


Protein Inference - Protein Level
Layout Export Legend Help Select All
Protein Labels Peptide Labels
Q6IS14
Q9GZV4
P63241
Protein Graph
Group Details

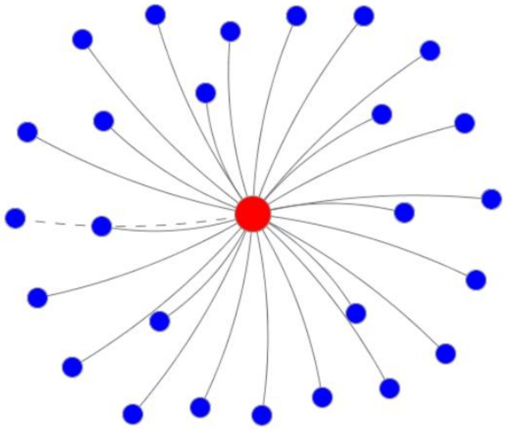
#Proteins: 3
#Peptides: 5

○ Validation Status ● Protein Evidence ○ Molecule Type
● All Neighbors ○ Unique Only □ Highlight

Now the proteins are color coded based on the protein evidence level from UniProt: green meaning that there is evidence at the protein level, while yellow means that the evidence for the protein is doubtful. In this case one could perhaps therefore assume that Q6IS14 is less likely to be in our sample. However, without further follow-up studies we cannot rule this protein out or separate the two other proteins in the group, although if we remove Q6IS14, P63241 will then have a unique peptide.

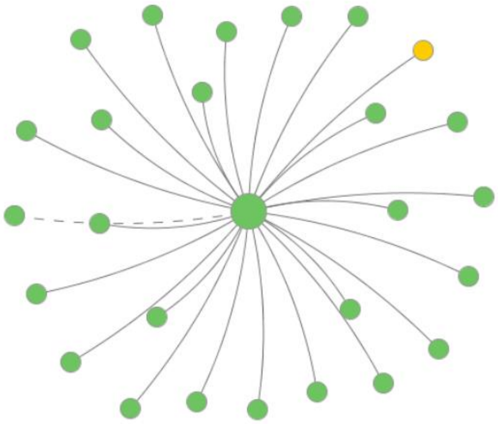
We will now look at some more examples of protein inference and see how graphs such as the ones above can help better understand the complexity of protein inference.

First, sort the proteins on the index column (the first column in the protein table) in ascending order by clicking the column header. Now select the protein P13639 – Elongation factor 2 (EF2_HUMAN) and click the green square in the PI column. This will show the Protein Inference dialog where you will see the following protein inference graph:

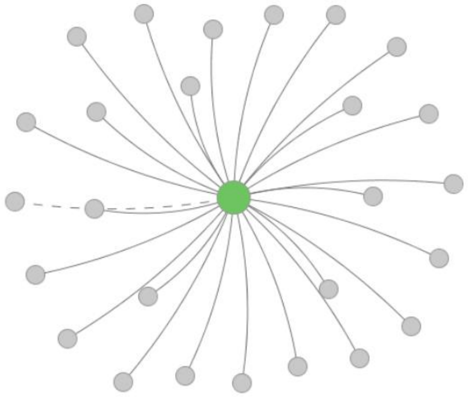


As before, the big red circle represents the protein while the blue smaller circles represent the peptides. A line is drawn between a peptide and a protein if the sequence of the peptide can be located in the sequence of the protein. A fully-drawn line means that the peptide sequence is enzymatic in the given protein, while a dotted line means that the peptide sequence is not adhering to all the cleavage rules of the enzyme used. *How many non-enzymatic peptides have been detected for the protein above?* [1.4v]

Now we will add some additional annotation on top of the graph. Select the Validation Status option below the graph. This will color the proteins and peptides depending on their validation status (from Confident (green), *via* Doubtful (yellow) to Not Validated (red). (The validation details will be further explained in the next chapter.) *How many doubtful and non-validated peptides have been detected for the given protein? [1.4w]*

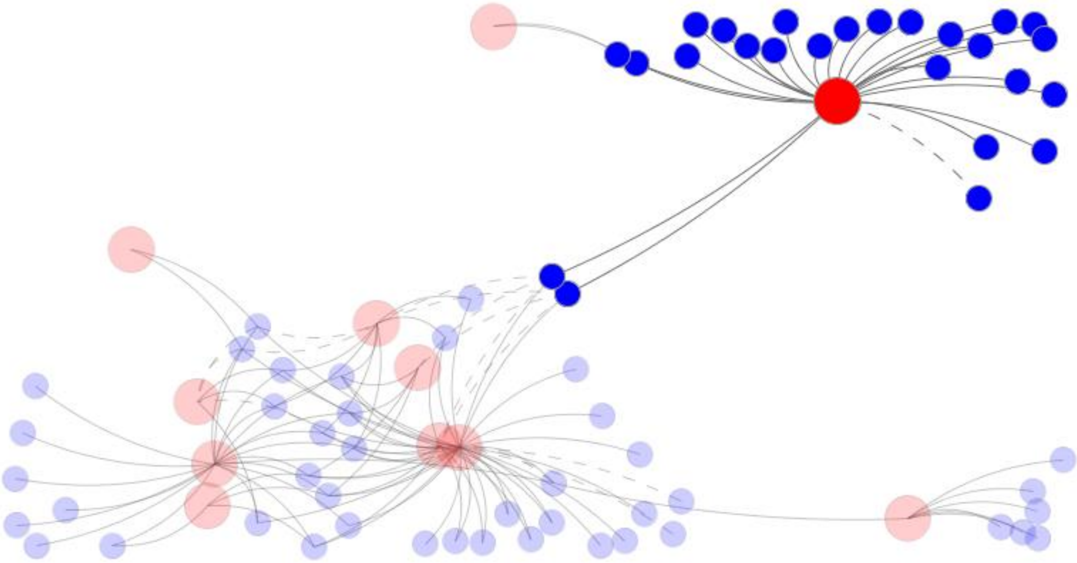


If a UniProt database is used it is also possible to color code the proteins based on the protein evidence available in UniProt (from Protein Evidence (green) to Uncertain (red)):



For an overview of the color coding options, click the Legend option at the top of the graph.

Now we will look at a slightly more complex example. Open the Protein Inference dialog for the protein P14625 - Endoplasmin (ENPL_HUMAN).



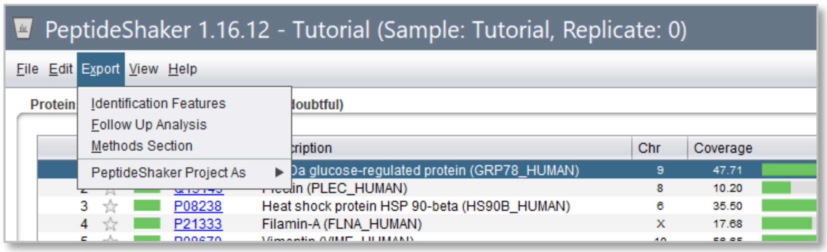
The highlighted protein in the upper right corner represents Endoplasmin (as you can confirm by holding the mouse over it), and the highlighted blue peptides represent the 25 peptides that could be explained by Endoplasmin being in our sample.

However, four of the peptides that could be explained by Endoplasmin could also come from other proteins. *By using the graph options, can you locate these four peptides? And how likely do you think that it is that each of them originates from Endoplasmin? [1.4x]*

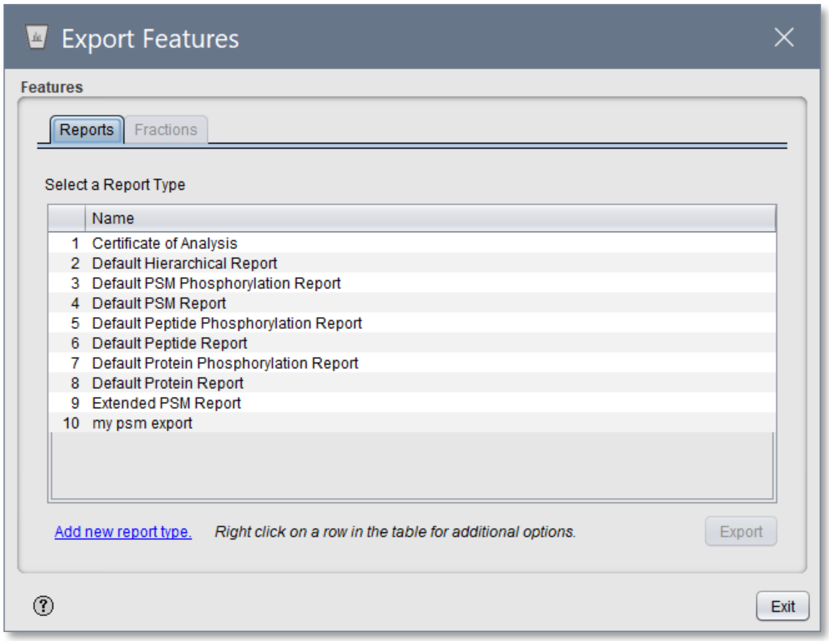
If you have time, explore the protein inference graphs for other protein groups. *Can we avoid protein inference issues in shotgun proteomics? [1.4y]*

D. Exporting Data

There are several ways to export data from PeptideShaker. In this section we will cover the four most important: (i) Identification Features, (ii) Follow Up Analysis, (iii) Methods Section, and (iv) Complete project export (either as mzIdentML[5], e.g. for submission to public data repositories such as PRIDE[6], or as a single zip file for sharing with other PeptideShaker users). All of these options are available via the Export menu.


PeptideShaker 1.16.12 - Tutorial (Sample: Tutorial, Replicate: 0)
File Edit Export View Help
Identification Features
Follow Up Analysis
Methods Section
PeptideShaker Project As
Protein
loubtful)
ription
Chr
Coverage
Da glucose-regulated protein (GRP78_HUMAN) 9 47.71
ectin (PLEC_HUMAN) 8 10.20
3 P08238 Heat shock protein HSP 90-beta (HS90B_HUMAN) 6 35.50
4 P21333 Filamin-A (FLNA_HUMAN) X 17.68

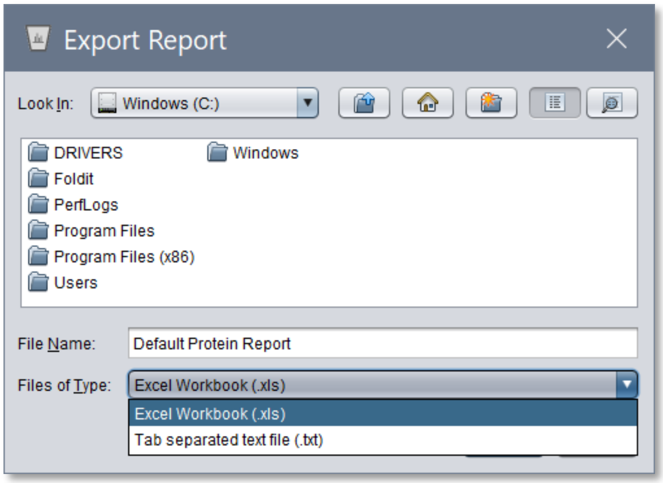
The Identification Features export allows for the export of specific identification features in tab separated file or Excel Workbooks. A default set of reports are available at the protein, peptide and PSM level, but you can also set up your own custom reports by clicking the 'Add new report type' option in the lower left corner, here indicated by the report called “my psm export”.



	Name
1	Certificate of Analysis
2	Default Hierarchical Report
3	Default PSM Phosphorylation Report
4	Default PSM Report
5	Default Peptide Phosphorylation Report
6	Default Peptide Report
7	Default Protein Phosphorylation Report
8	Default Protein Report
9	Extended PSM Report
10	my psm export

Right click on a row in the table for additional options.

The report file type, i.e. tab separated text file or Excel Workbook, is chosen as part of selecting the name and location of the file to export to:



Look In: Windows (C:)

File Name: Default Protein Report

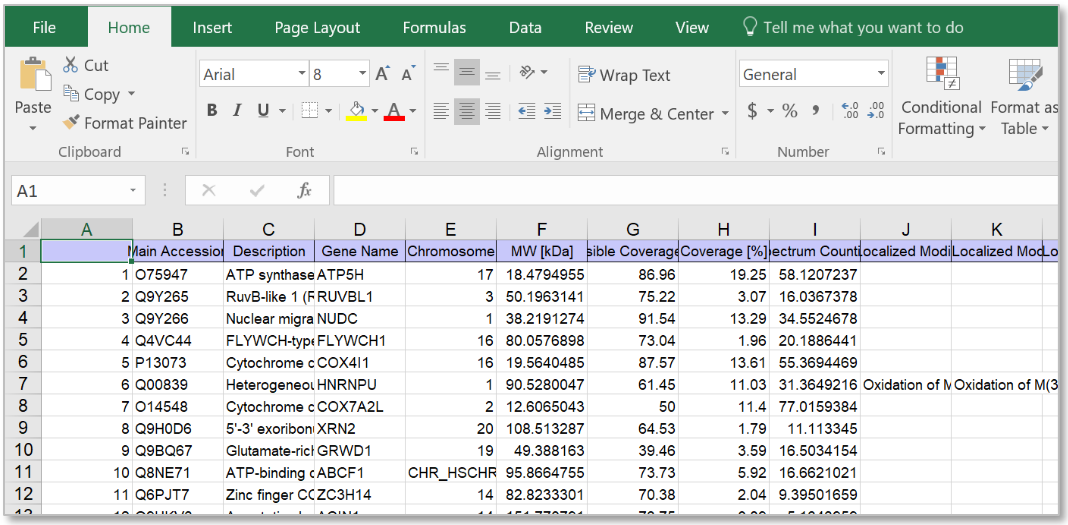
Files of Type: Excel Workbook (.xls)

Excel Workbook (.xls)
Tab separated text file (.txt)

Tip:

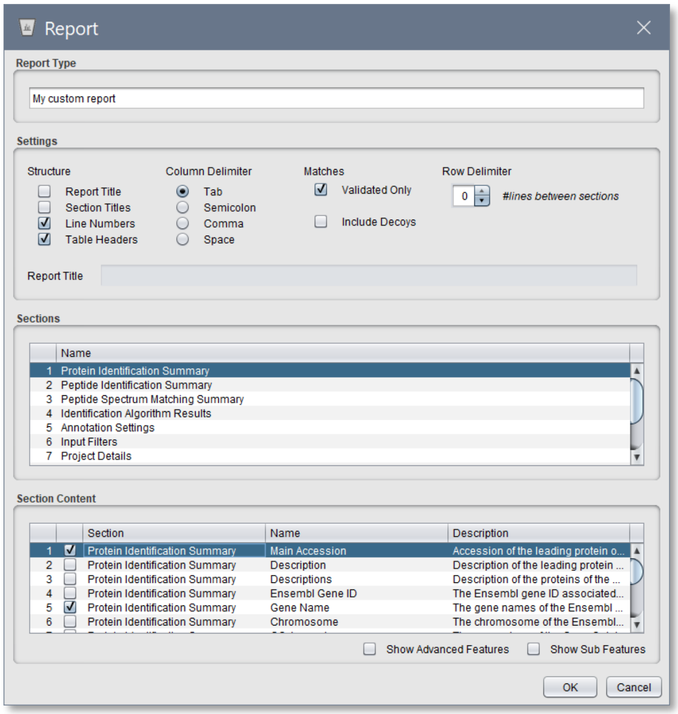
When exporting large amounts of data, it is recommended to choose the tab separated text file format as Excel Workbooks have a maximum limit for the number of rows it can contain.

Default Protein Report example opened in Excel:



	A	B	C	D	E	F	G	H	I	J	K	
1		Main Accessio	Description	Gene Name	Chromosome	MW [kDa]	sible Coverag	Coverage [%]	ectrum Count	ocalized Modi	Localized Mod	Lo
2	1	O75947	ATP synthase	ATP5H	17	18.4794955	86.96	19.25	58.1207237			
3	2	Q9Y265	RuvB-like 1 (F	RUVBL1	3	50.1963141	75.22	3.07	16.0367378			
4	3	Q9Y266	Nuclear migra	NUDC	1	38.2191274	91.54	13.29	34.5524678			
5	4	Q4VC44	FLYWCH-type	FLYWCH1	16	80.0576898	73.04	1.96	20.1886441			
6	5	P13073	Cytochrome c	COX4I1	16	19.5640485	87.57	13.61	55.3694469			
7	6	Q00839	Heterogeneou	HNRNPU	1	90.5280047	61.45	11.03	31.3649216	Oxidation of M	Oxidation of M(3	
8	7	O14548	Cytochrome c	COX7A2L	2	12.6065043	50	11.4	77.0159384			
9	8	Q9H0D6	5'-3' exoribon	XRN2	20	108.513287	64.53	1.79	11.113345			
10	9	Q9BQ67	Glutamate-rich	GRWD1	19	49.388163	39.46	3.59	16.5034154			
11	10	Q8NE71	ATP-binding c	ABCF1	CHR_HSCHR	95.8664755	73.73	5.92	16.6621021			
12	11	Q6PJT7	Zinc finger CC	ZC3H14	14	82.8233301	70.38	2.04	9.39501659			

Example of how to set up a custom report:



Report

Report Type

My custom report

Settings

Structure	Column Delimiter	Matches	Row Delimiter
□ Report Title	● Tab	☑ Validated Only	0 #lines between sections
□ Section Titles	○ Semicolon		
☑ Line Numbers	○ Comma	□ Include Decoys	
☑ Table Headers	○ Space		

Report Title

Sections

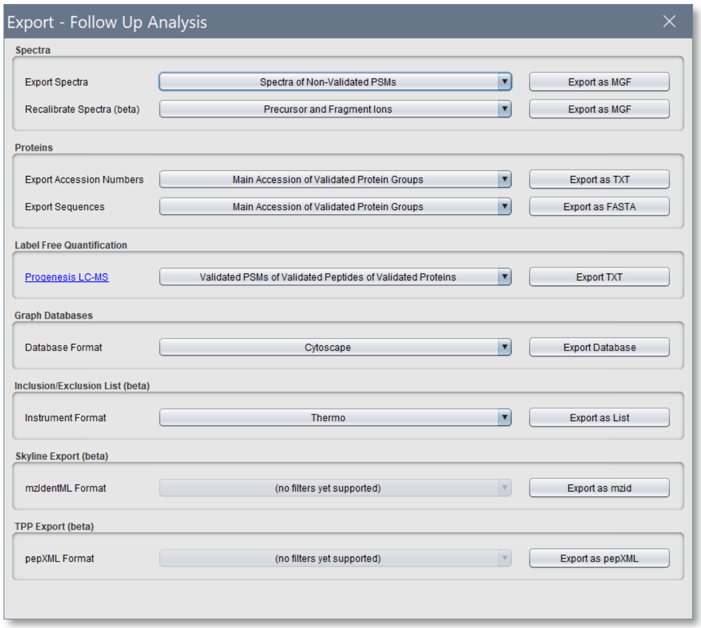
	Name
1	Protein Identification Summary
2	Peptide Identification Summary
3	Peptide Spectrum Matching Summary
4	Identification Algorithm Results
5	Annotation Settings
6	Input Filters
7	Project Details

Section Content

		Section	Name	Description
1	☑	Protein Identification Summary	Main Accession	Accession of the leading protein o...
2	□	Protein Identification Summary	Description	Description of the leading protein ...
3	□	Protein Identification Summary	Descriptions	Description of the proteins of the ...
4	□	Protein Identification Summary	Ensembl Gene ID	The Ensembl gene ID associated...
5	☑	Protein Identification Summary	Gene Name	The gene names of the Ensembl ...
6	□	Protein Identification Summary	Chromosome	The chromosome of the Ensembl...

□ Show Advanced Features □ Show Sub Features

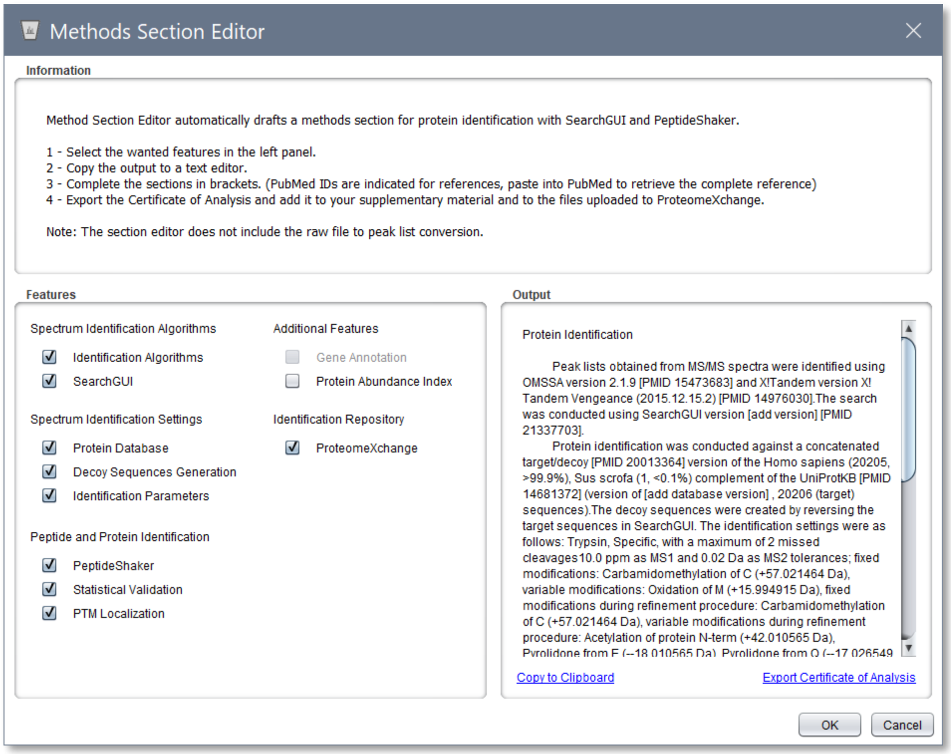
Next, the Follow Up Analysis option makes it straightforward to export data for various follow up analyses, for example exporting all the unidentified spectra for special processing or exporting the list of protein sequences for all the validated proteins.



Export - Follow Up Analysis

Spectra		
Export Spectra	Spectra of Non-Validated PSMs	Export as MGF
Recalibrate Spectra (beta)	Precursor and Fragment Ions	Export as MGF
Proteins		
Export Accession Numbers	Main Accession of Validated Protein Groups	Export as TXT
Export Sequences	Main Accession of Validated Protein Groups	Export as FASTA
Label Free Quantification		
Progenesis LC-MS	Validated PSMs of Validated Peptides of Validated Proteins	Export TXT
Graph Databases		
Database Format	Cytoscape	Export Database
Inclusion/Exclusion List (beta)		
Instrument Format	Thermo	Export as List
Skyline Export (beta)		
mzIdentML Format	(no filters yet supported)	Export as mzid
TPP Export (beta)		
pepXML Format	(no filters yet supported)	Export as pepXML

The Methods Section export option makes it easy to collect all the details required for the Methods section of your publication. Simply select the features you want to include, and a Methods section draft is created for you. All you have to do is replace the PubMed identifiers using your reference manager of choice. This greatly simplifies the extraction of the numerous settings used as part of the identification procedure and ensures that no details are missed.



Methods Section Editor

Information

Method Section Editor automatically drafts a methods section for protein identification with SearchGUI and PeptideShaker.

1 - Select the wanted features in the left panel.
2 - Copy the output to a text editor.
3 - Complete the sections in brackets. (PubMed IDs are indicated for references, paste into PubMed to retrieve the complete reference)
4 - Export the Certificate of Analysis and add it to your supplementary material and to the files uploaded to ProteomeXchange.

Note: The section editor does not include the raw file to peak list conversion.

Features

Spectrum Identification Algorithms
- ☑ Identification Algorithms
- ☑ SearchGUI

Spectrum Identification Settings
- ☑ Protein Database
- ☑ Decoy Sequences Generation
- ☑ Identification Parameters

Peptide and Protein Identification
- ☑ PeptideShaker
- ☑ Statistical Validation
- ☑ PTM Localization

Additional Features
- ☐ Gene Annotation
- ☐ Protein Abundance Index

Identification Repository
- ☑ ProteomeXchange

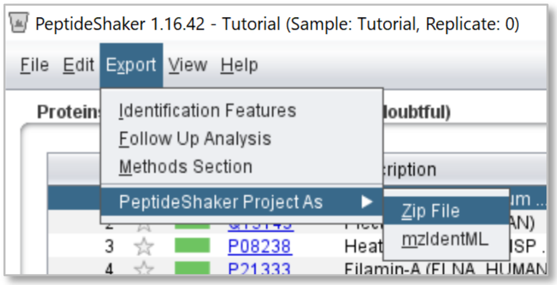
Output

Protein Identification

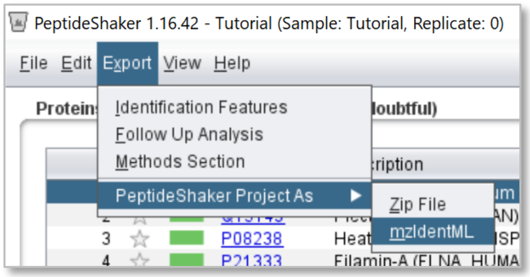
Peak lists obtained from MS/MS spectra were identified using OMSSA version 2.1.9 [PMID 15473683] and X!Tandem version X! Tandem Vengeance (2015.12.15.2) [PMID 14976030].The search was conducted using SearchGUI version [add version] [PMID 21337703].

Protein identification was conducted against a concatenated target/decoy [PMID 20013364] version of the Homo sapiens (20205, >99.9%), Sus scrofa (1, <0.1%) complement of the UniProtKB [PMID 14681372] (version of [add database version] , 20206 (target) sequences).The decoy sequences were created by reversing the target sequences in SearchGUI. The identification settings were as follows: Trypsin, Specific, with a maximum of 2 missed cleavages10.0 ppm as MS1 and 0.02 Da as MS2 tolerances; fixed modifications: Carbamidomethylation of C (+57.021464 Da), variable modifications: Oxidation of M (+15.994915 Da), fixed modifications during refinement procedure: Carbamidomethylation of C (+57.021464 Da), variable modifications during refinement procedure: Acetylation of protein N-term (+42.010565 Da), Pyrolidone from E (--18.010565 Da), Pyrolidone from Q (--17.026549

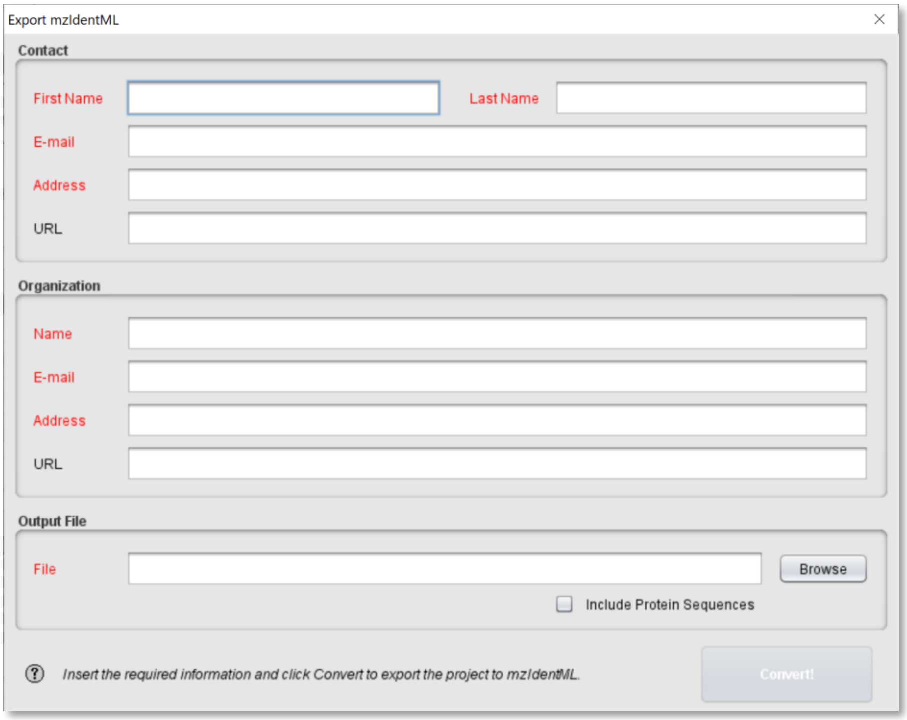
If you want to share your PeptideShaker project with a colleague or collaborator, the best option is to export the complete project as a single zip file. This ensures that all the required files are included and that the exported project can easily be opened on any computer with PeptideShaker installed.



Finally, it is also straightforward to export all of the combined identification results in the standard mzIdentML data format:



Simply fill in the required details regarding contact and organization, choose the output file, and click the “Convert” button.



Insert the required information and click Convert to export the project to mzIdentML.

For further details on how to submit the generated mzIdentML file to the PRIDE data repository, please see Chapter 3 – Data Sharing.

References

1. Vaudel, M. et al. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat Biotech* **33**, 22-24 (2015).
2. Martens, L. et al. PRIDE: the proteomics identifications database. *Proteomics* **5**, 3537-3545 (2005).
3. Nesvizhskii, A.I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* **4**, 1419-1440 (2005).
4. Magrane, M. & Consortium, U. UniProt Knowledgebase: a hub of integrated protein data. *Database : the journal of biological databases and curation* **2011**, bar009 (2011).
5. Jones, A.R. et al. The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol Cell Proteomics* **11**, M111 014381 (2012).
6. Vizcaino, J.A. et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res* **44**, D447-456 (2016).