1. PSM, Peptide & Protein Visualization

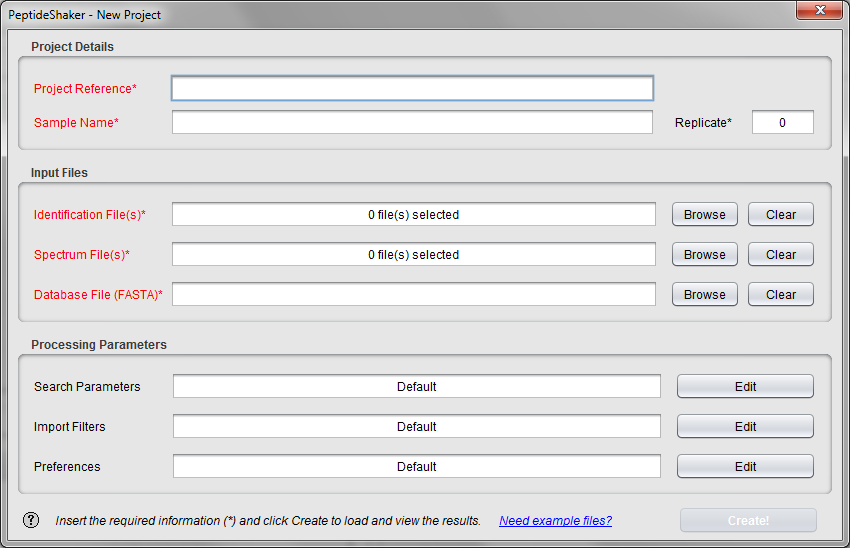
The search conducted in the previous part generated two files containing the peptides matched by OMSSA and X!Tandem for each spectrum, so-called Peptide to Spectrum Matches (PSMs). From these we need to find the identified peptides and proteins. This is the task of PeptideShaker ([http://peptide-shaker.googlecode.com](http://peptide-shaker.googlecode.com/)). PeptideShaker is freely available and is still under development so any comments or suggestions will be greatly appreciated!

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 PRIDE1 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 should see this screen:



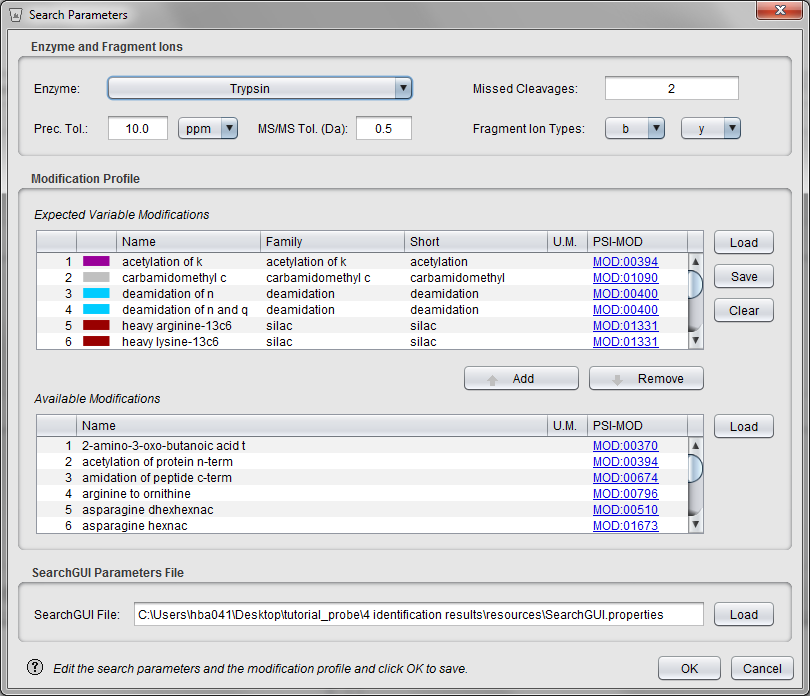
Note the question mark at the bottom left. These are present throughout PeptideShaker and will open contextual help. Additional example files are also available on our website.

Start by giving your project and sample a name at the top, and leave replicate number at 0. We will not take advantage of these details in this exercise, but such annotation is crucial for later (re-)use for real projects!

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 OMSSA and X!Tandem result files (Velos005137.omx and Velos005137\_[date]\_[time].t.xml) and click ‘Add’.

Notice that the spectrum files and the database are automatically filled in when using results from SearchGUI. Otherwise you have to load these manually (files are in the resources folder). Note also that the descriptions of the processing parameters have changed. This information was extracted from the searchGUI\_input.txt file and the parameters file created by SearchGUIalong with the identification files. Note that if files have been moved in the interim, they have to be added manually. PeptideShaker requires every mgf header to contain the spectrum title, the precursor charge and retention time in seconds. Note that the spectrum title must be unique for each file.

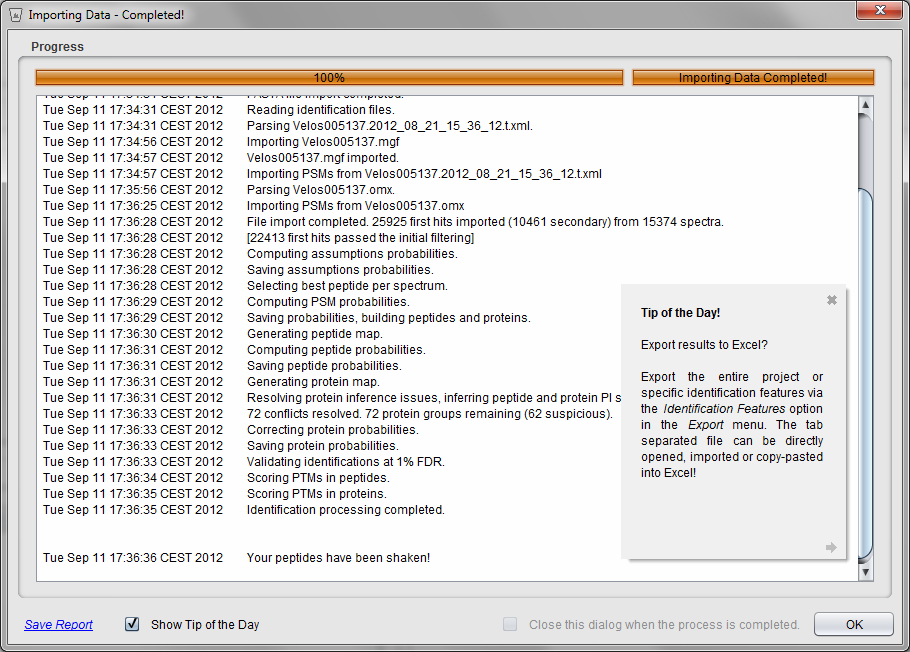
Once all the files have been loaded, click the ‘Edit’ button for the ‘Search Parameters’. You should see this dialog appearing:



This lists all the parameters used during the search. Note that when loading data from SearchGUI, all of them will be completed automatically. You should pay careful attention to the modification profile. Only the modifications listed on the table ‘Expected Variable Modifications’ will be recognized by PeptideShaker. It is fine if additional modifications are present (like iTRAQ in our case), you do not have to remove them.

In this table, you can group modifications by family, note for instance that all phosphorylations are grouped together here. *What is the point in grouping modifications? What kind of modifications shall we group into a family?* Note that once the project is created, you cannot edit the modification families anymore. Also note that adding a modification here has no effect if the modification was not used in the original search.

Click ‘OK’ to go back to the New Project dialog and then click ‘Create!’ to load the data. The selected files will now be loaded into PeptideShaker, and you can follow the progress in the dialog while browsing some PeptideShaker usage tips. When finished, the ‘Cancel' button in the lower right of this dialog will change into an ‘OK’ button and the dialog should look like this:



Click ‘OK’ to open the loaded files. After a few seconds your files will be loaded and you should see something like this:



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 As’ item (or type Ctrl + S). Choose a name for your data file, and click ‘Save’. Note that saving can take time: there is a lot of information to store!

A .cps file should now have been created. If you want to open this project later on you simply have to open this cps file in PeptideShaker. Also note that the project can be accessed via ‘Open Recent Project’ on the 'File' menu.

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.

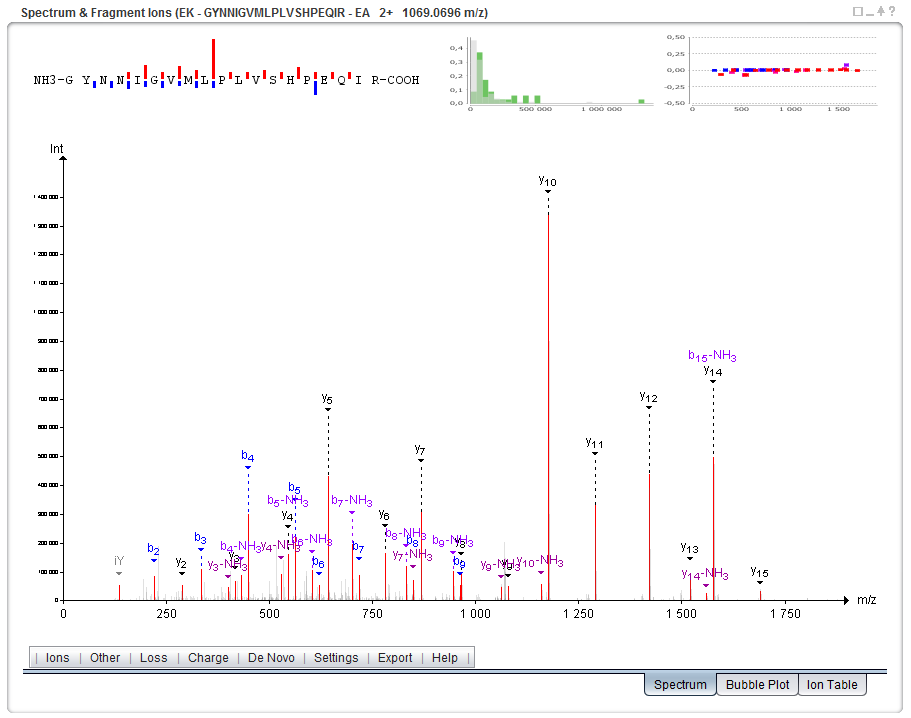
Start by selecting the first protein in the protein table, the one with the protein accession number P42850 and with the descriptionPPSA\_PYRFU Phosphoenolpyruvate synthase.

Looking at the row for this protein we see that we have covered 72.34 % of the protein's amino acid sequence, detecting a total of 88 peptides from 317 spectra. *Why are there more spectra than peptides*? *How do you define a peptide?*

We will now look at one of these peptides. In the peptide table, look for the peptide sequence GYNNIGVMLPLVSHPEQIR.

In the peptide to spectrum matches table you will see that this peptide is found five times. That means that five different spectra have been identified as the peptide in question. Also notice that some of these have a precursor charge of 2+, while others have a charge of 3+. *Why do not all precursors carry the same charge?*

Make sure that the peptide with a precursor charge of 2+ and a mass error of 2.86 is selected, it should be the second in the table. 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 GYNNIGVMLPLVSHPEQIR?* *Which are the different ions displayed here? Are the same ions detected in all peptide to spectrum matches or peptides?*

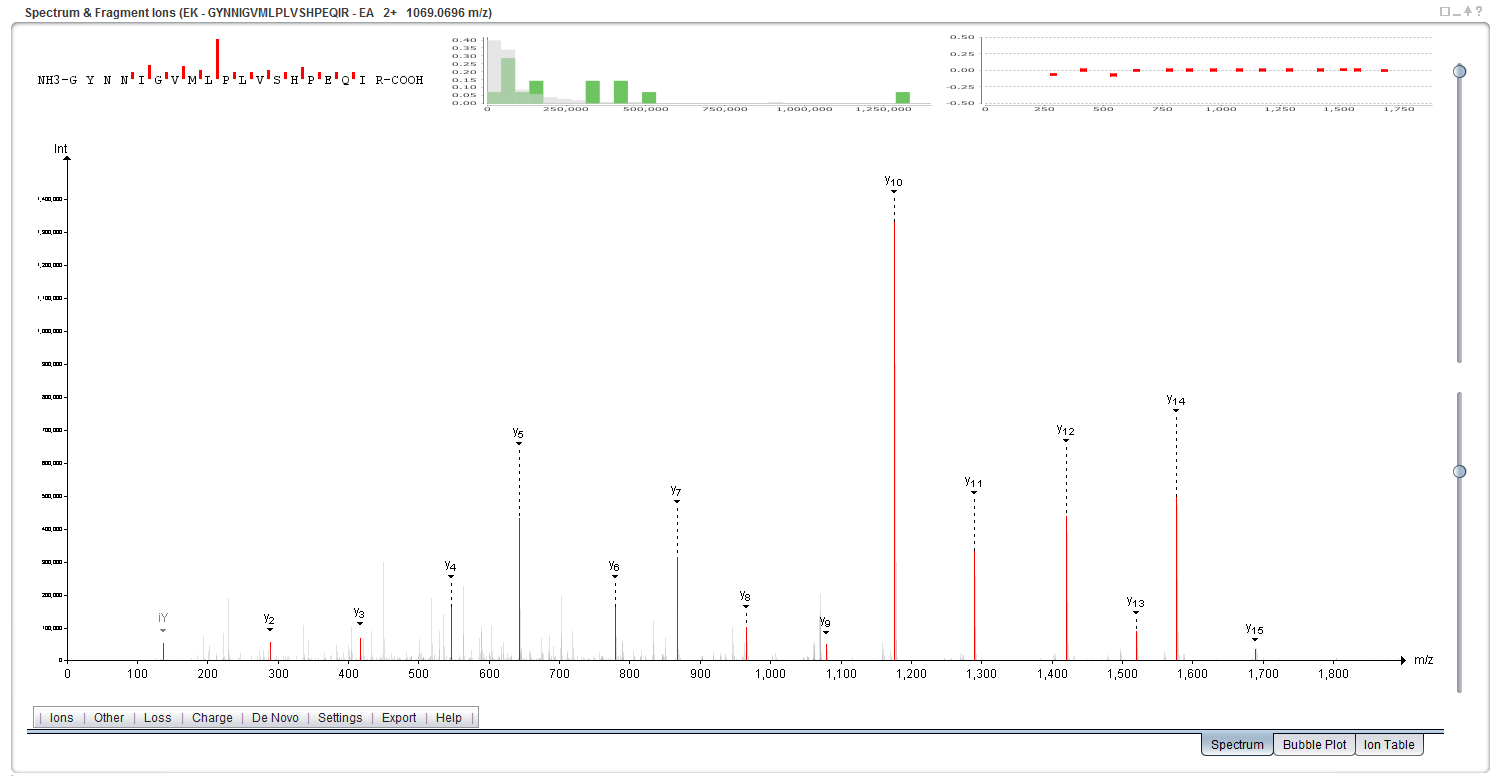
The menu bar below the spectrum allows you to fine tune the annotation settings and the spectrum display. It is moreover possible to export the spectrum in various formats. Note that only the highest peaks are annotated. You can change the annotation level by scrolling when over the spectrum. If you hold ‘Ctrl’ and scroll, the m/z tolerance will be changed. *Tip: you can also change these settings using sliders which can be enabled in the ‘View’ menu.*

Before continuing, makes sure that the intensity level is at 75% and the accuracy at 0.5 Da.

The first thing we will do is 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 on the top right of the spectrum panel. Note that other tabs have been minimized at the bottom left of the screen. The spectrum should now cover the whole screen:

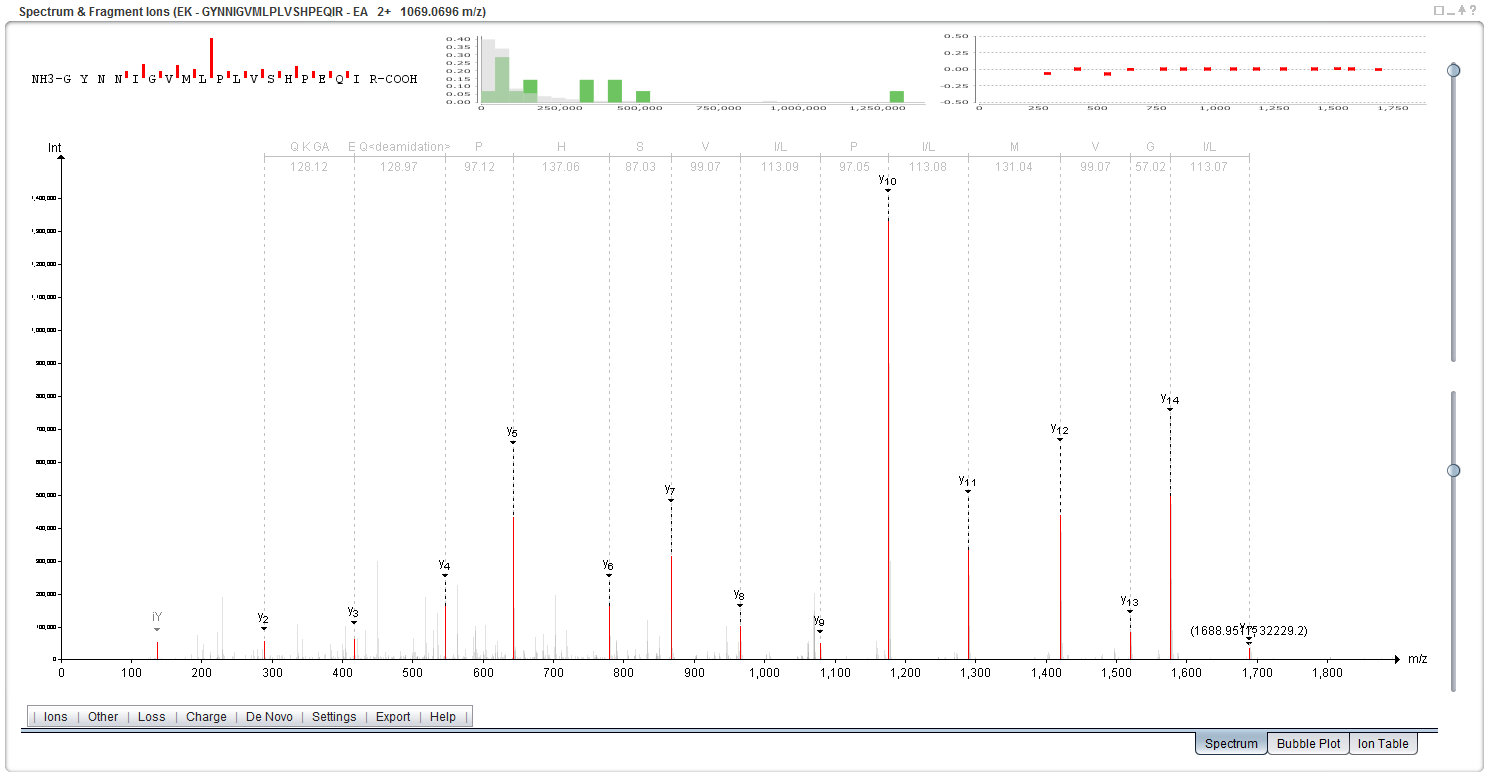


Next, to make the sequencing even easier, zoom in on the m/z range covered by the y-ions y2 to y15, simply by clicking and holding the left mouse button from just before the y2-ion and (while holding the left mouse button pressed) dragging the mouse horizontally to the right until you have passed the y15-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 y2-ion. *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.* 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.

Continue the sequencing by clicking the y-ions in ascending order from left to right. *Tip: 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.*

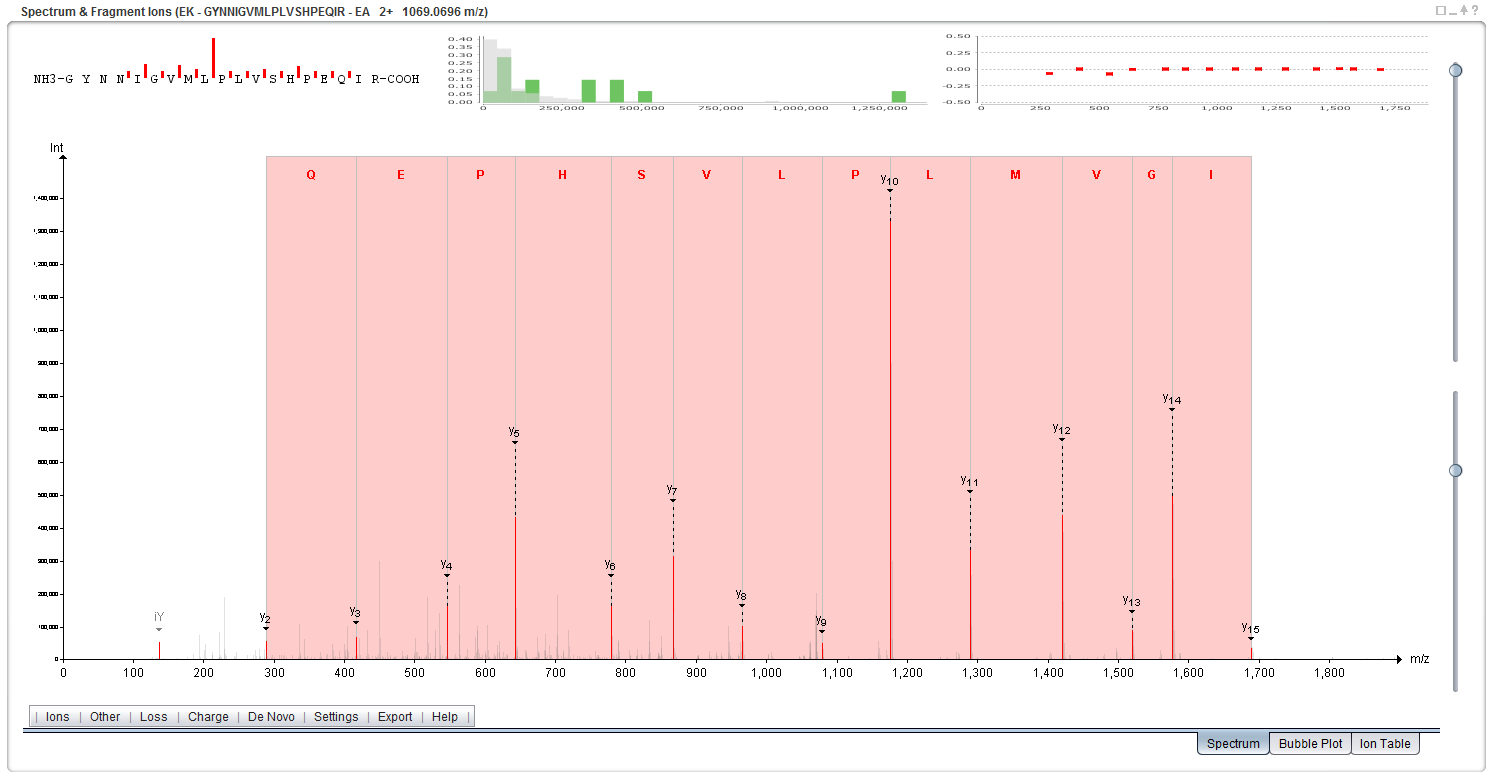
When done, you should have a spectrum display that looks something like this:



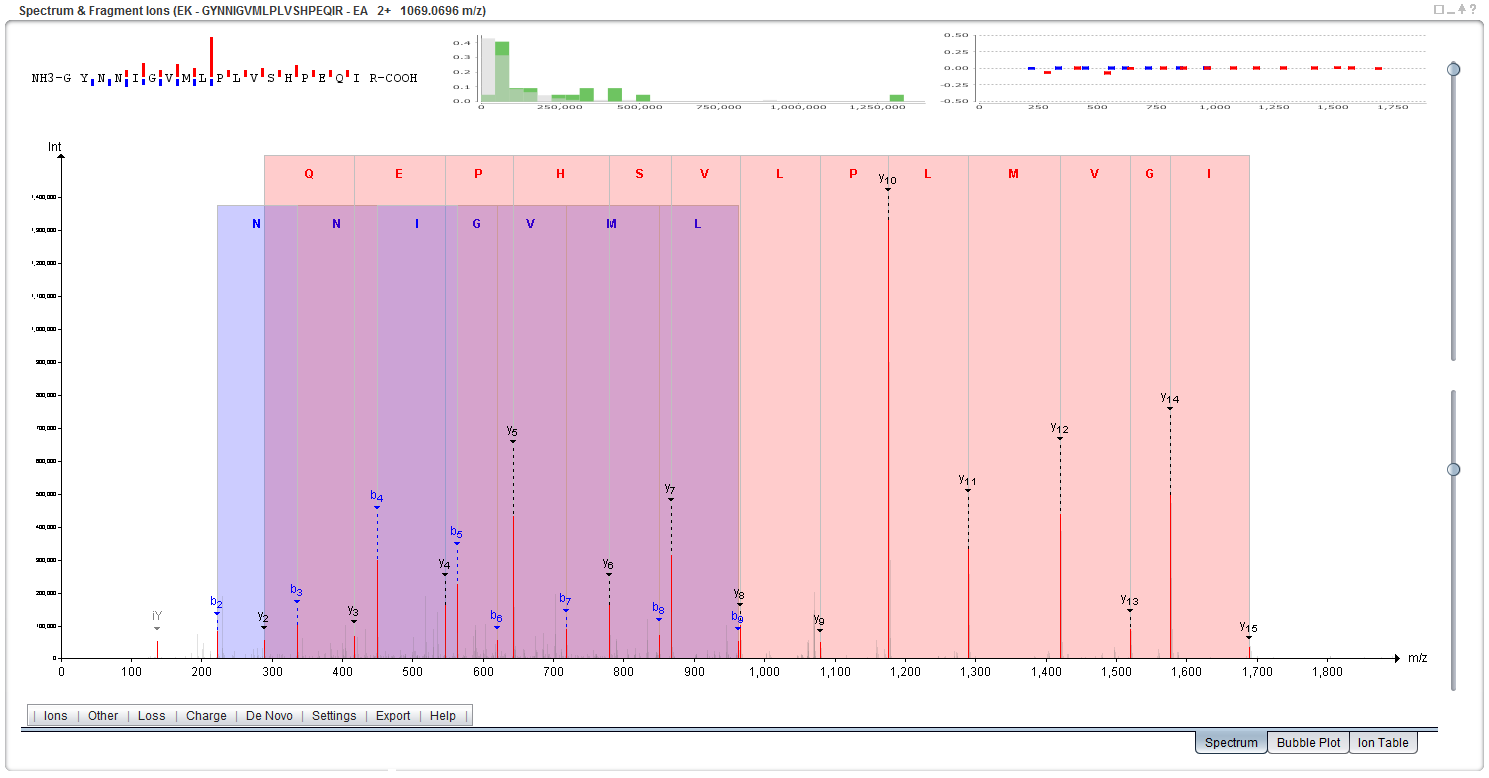
If you now read the sequence at the top you will find that it reads (from left to right): Q K GA, E Q<deamidation>, P, H, S, V, I/L, P, I/L, M, V, G, I/L. Given that y-ions are indexed from right to left (carboxy-terminus to amino-terminus) relative to the sequence, we have to reverse the sequence to match it to the peptide sequence.

Reversing and keeping only the unmodified options we get the sequence: (Q/K/GA) (E/Q<deamidation>) P H S I/L M V G I/L. 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 - 15. *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?*

In the ‘De Novo’ menu, select ‘y-ions’, PeptideShaker then shows you the solution retained for this spectrum.



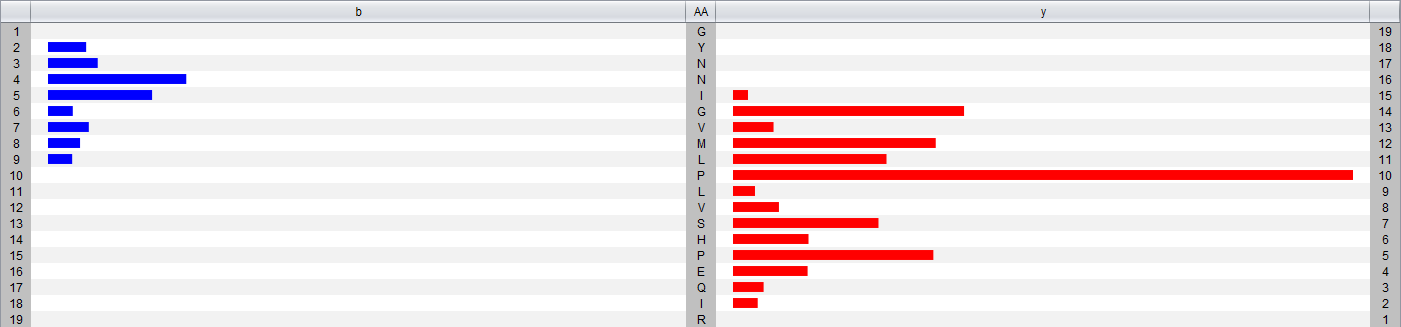
Now, enable the display of b-ions and select ‘b-ions’ in ‘De Novo’. You should see the following screen:



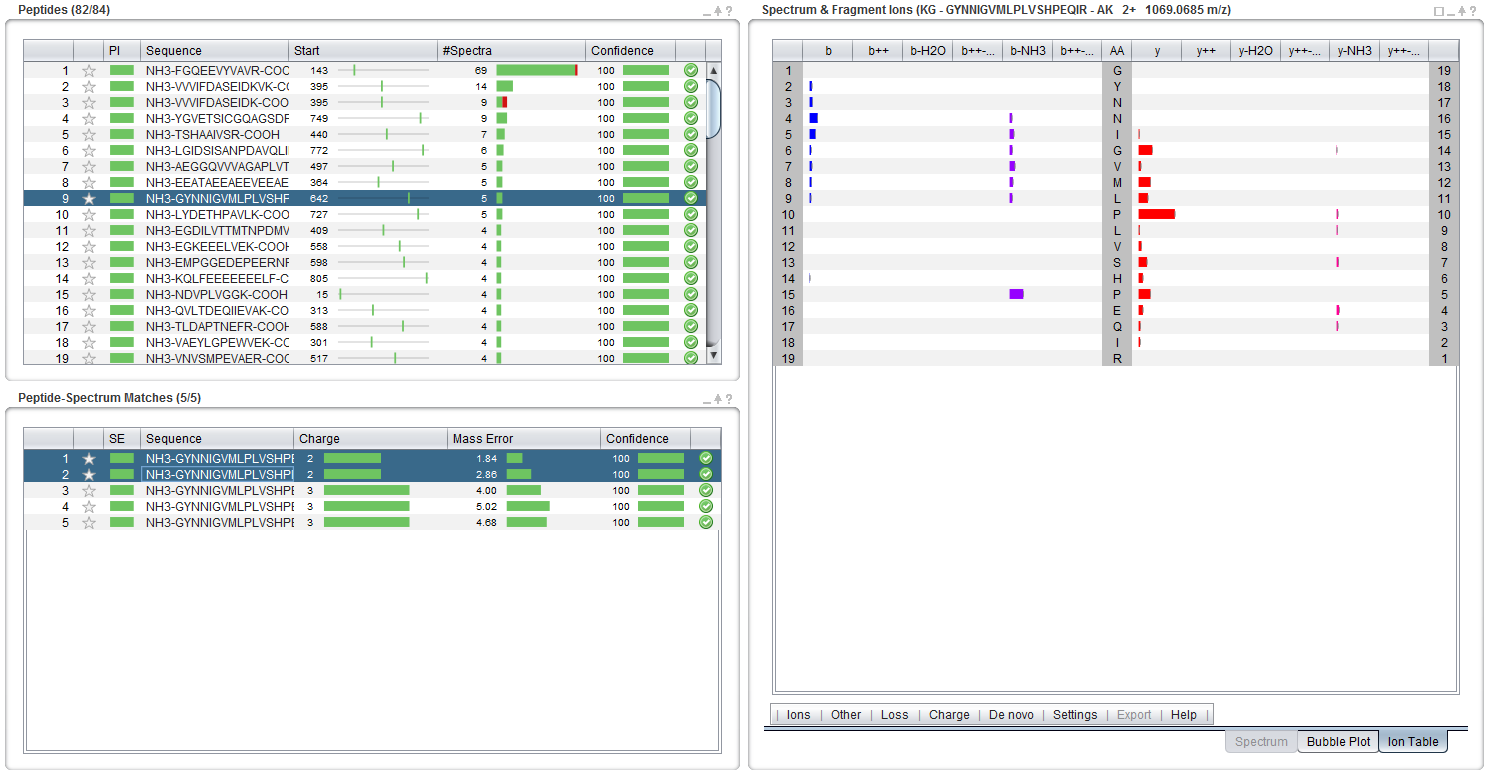
Note that in this case you can only use b2 to b9. Remember that b-ions are indexed from the left to the right (amino-terminus to carboxy-terminus), so in this case you don't have to reverse the sequence. *Why do we have lower coverage and intensity for b-ions relative to the y-ions for the same peptide? Is this the same for all peptides, all instruments, all protocols?*

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. *Tip: m/z or intensity display can be selected via the Settings menu below the spectrum.* *How do the results here correspond to the results you found in your de novo sequencing?*



We will now move one step further by comparing multiple spectra identified as the same peptide. Click on ‘Peptides & PSMs’ at the bottom of the screen to display the corresponding tables. While holding down the Ctrl-key, select the second spectrum with a charge of 2+. You should see this:

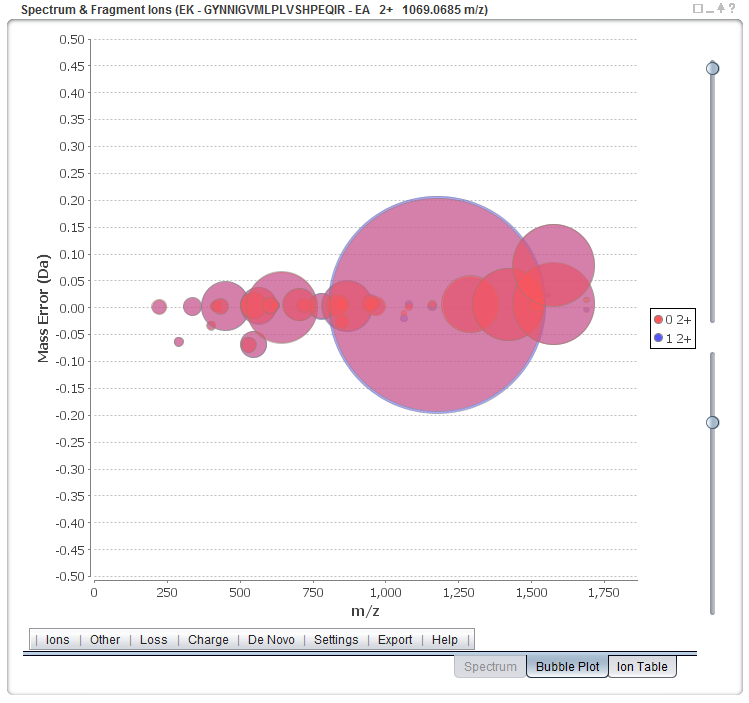


The intensity difference between the two spectra are displayed with error bars. *Do you expect reproducibility in intensities between spectra? What do you think about the comparison between 2+ and 3+ PSMs?*

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.

For this purpose a novel view has been developed, 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 one row in the spectrum table). Selecting/highlighting all the rows in the PSM table with the same 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.

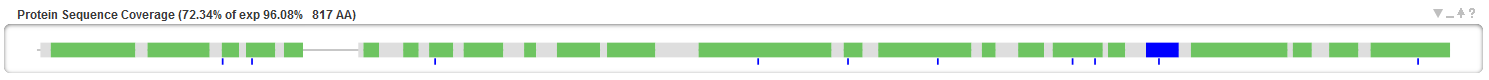


This way of looking at multiple spectra takes a bit of getting used to, but once one is comfortable with the setup, it becomes very easy to pick up details that would have been very difficult to see by a manual inspection of each spectrum individually.

*Now, what do you think about the comparison between 2+ and 3+ PSMs?*

Restore the standard view by clicking on the ‘Proteins’ and ‘Protein Sequence Coverage’ buttons at the bottom left.

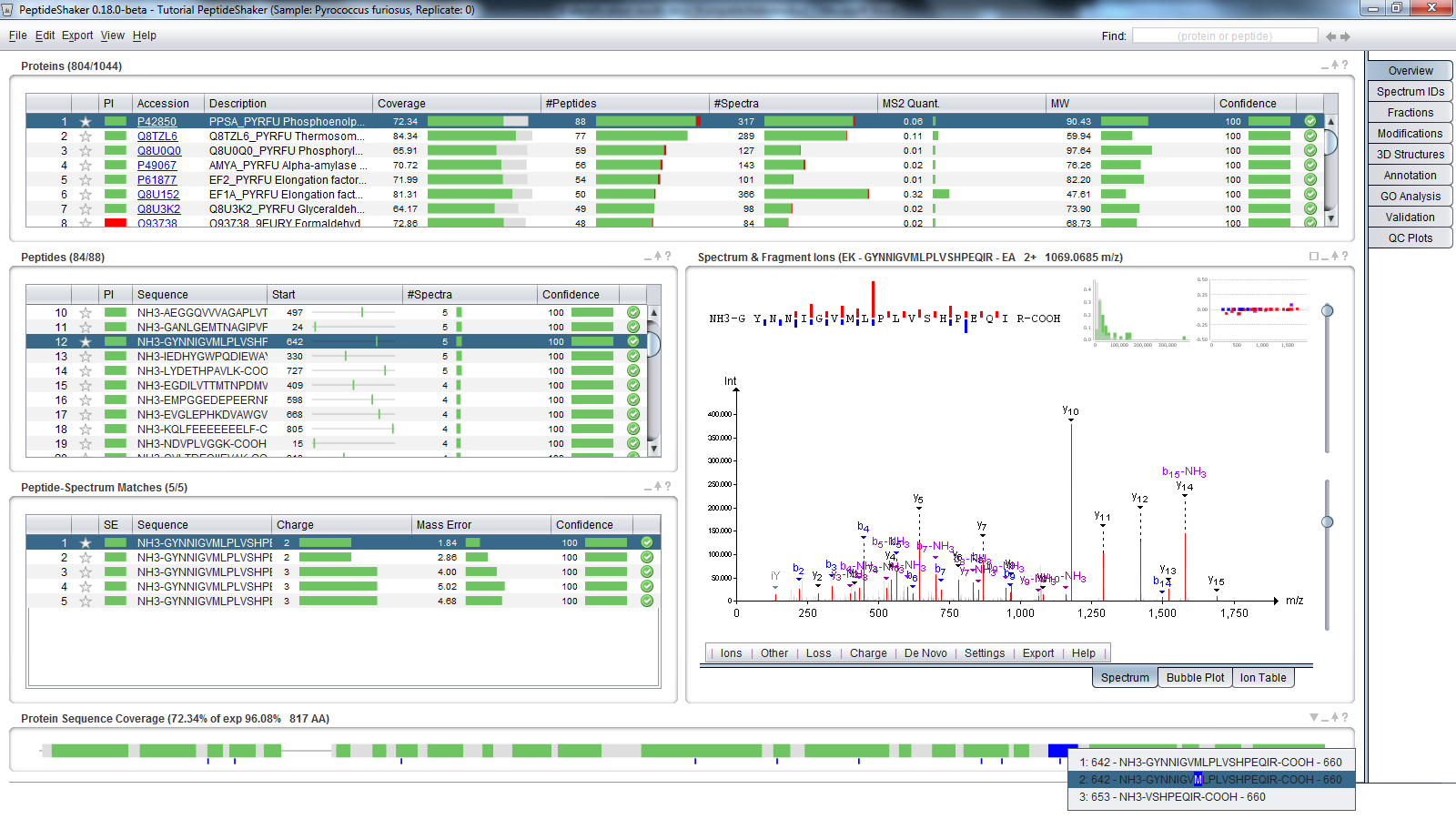
We will now study the sequence coverage of the selected protein – P42850. The sequence coverage is indicated in the protein table, 72.34% and illustrated at the bottom in the Protein Sequence Coverage panel:



The 84 identified peptides are mapped onto the sequence in green 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 96.08%, actually, some might have observed that the observable coverage is readily displayed 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 are the possible peptides at this position of the sequence?*

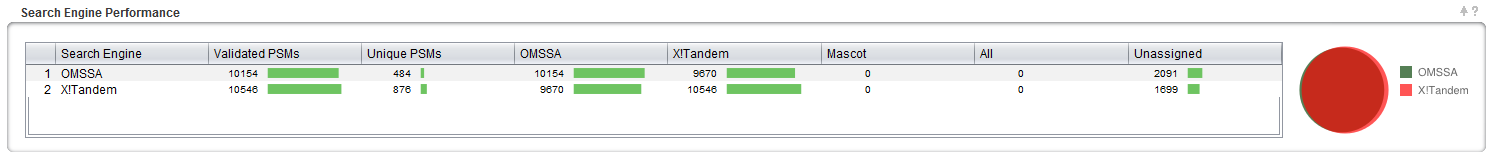
You will see that the peptide GYNNIGVMLPLVSHPEQIR was also identified with an oxidized methionine. In fact, keen observers will have noticed that the oxidation sites are readily mapped under the sequence. Clicking on the blue peptide allows the selection of the oxidized version:



1. Quality Control

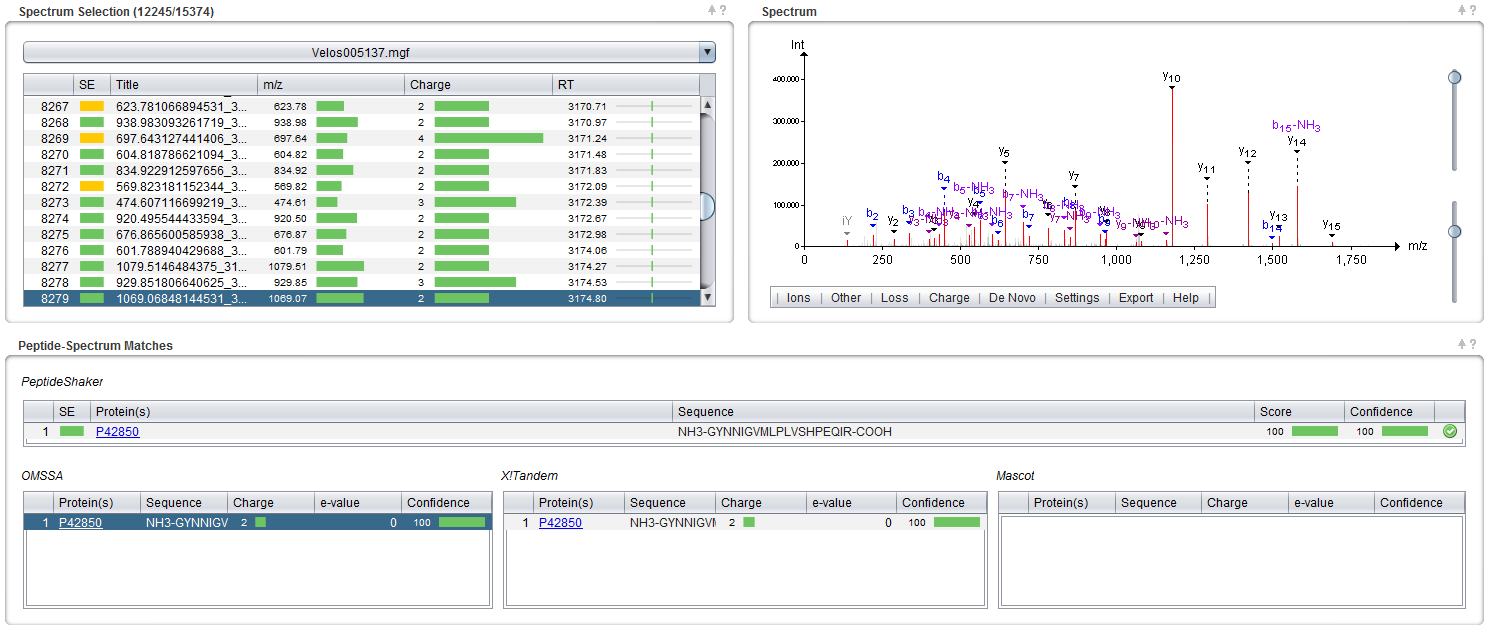
Before further investigation, it is important to verify the quality of the search. We will start by verifying that all search engines worked correctly. As you remember, we loaded identification results from OMSSA and X!Tandem. Note that it is also possible to load Mascot results. Open the tab called ‘Spectrum IDs’ in the upper right corner, loading a tab takes a few seconds. 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 top you should now see something like this:



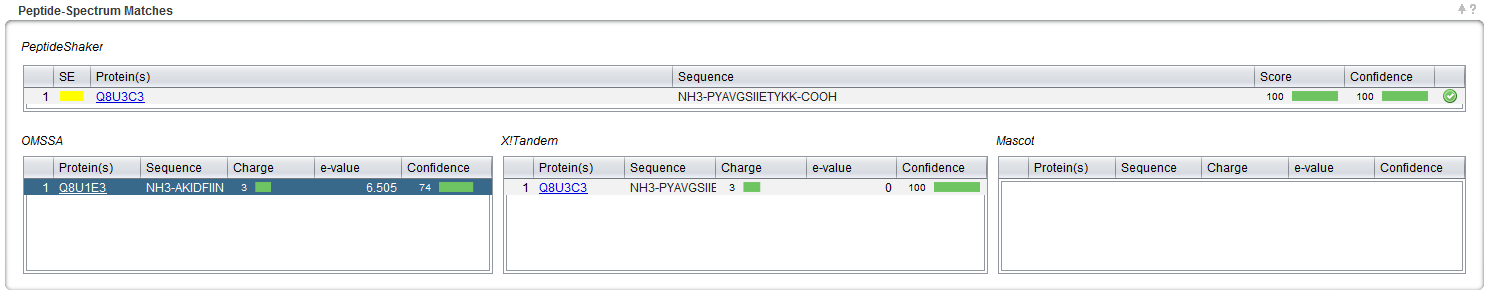
This view shows you how well the search engine results overlap at the PSM level. If the bubble representing a search engine is significantly smaller compared to the other(s), it is worth verifying that the search did not encounter any unexpected issues and that the same search settings were used for all search engines. As you can see here they are pretty close but not identical. *Why are the search engine results different? Will the differences depend on the dataset used? Can we take advantage of these differences?*

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 on the left you will see the PSM for the two search engines and how these were scored. The selected PSM will be used to annotate the given spectrum.



*Tip: if you are interested in non-assigned spectra, you can export them using the ‘Follow Up Analysis’ option on the ‘Export’ menu.*

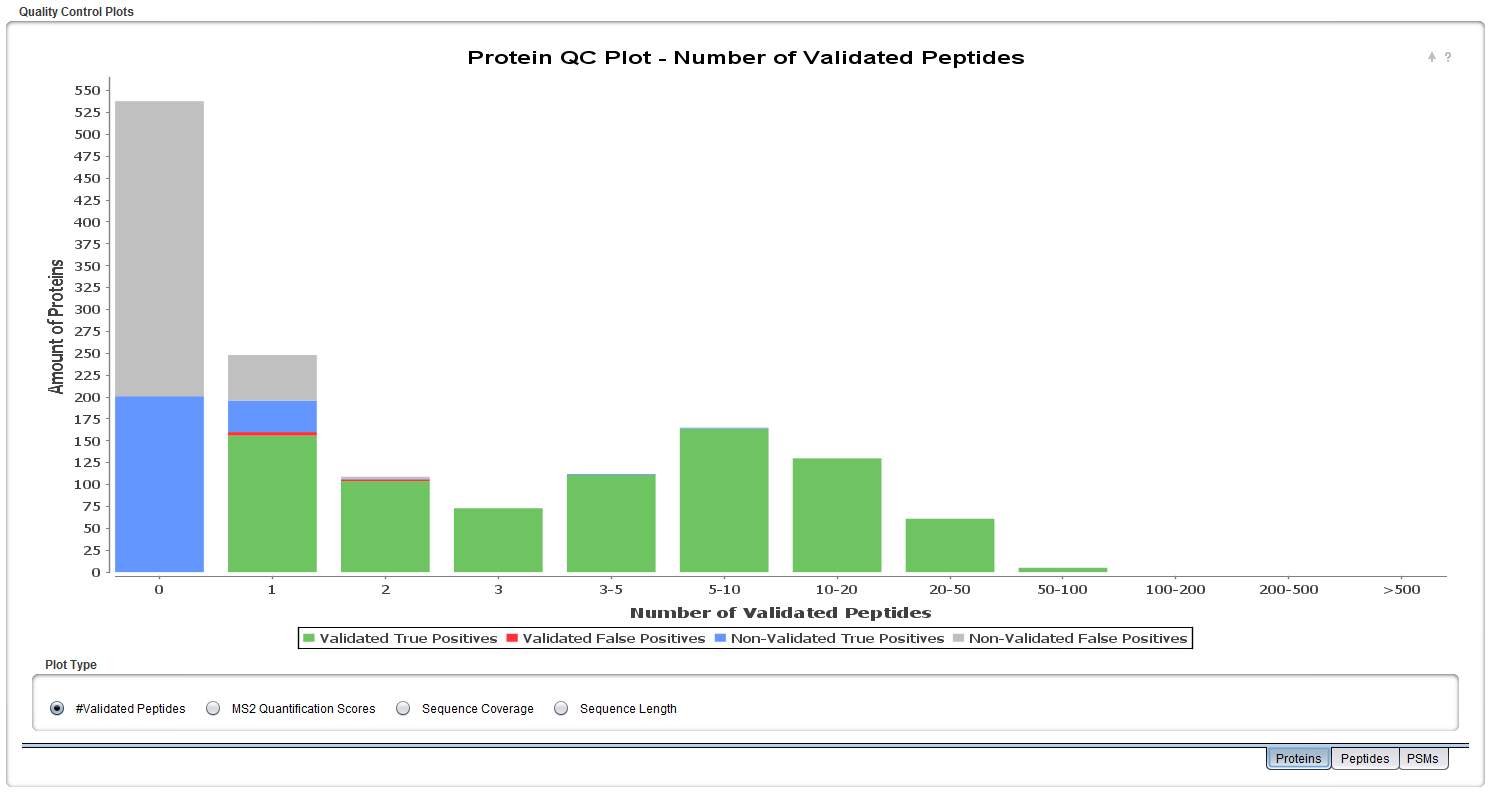
Note the SE (Search Engine) column which indicates whether a spectrum was identified and if so, whether the search engines agreed on the identification. Click on the row 5658, spectrum 490.272308349609\_2286.9417, with a yellow square. You should see this:



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

Now, select the ‘QC Plots’ tab. In this tab, some Quality Control metrics are 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 amount of identified peptides:



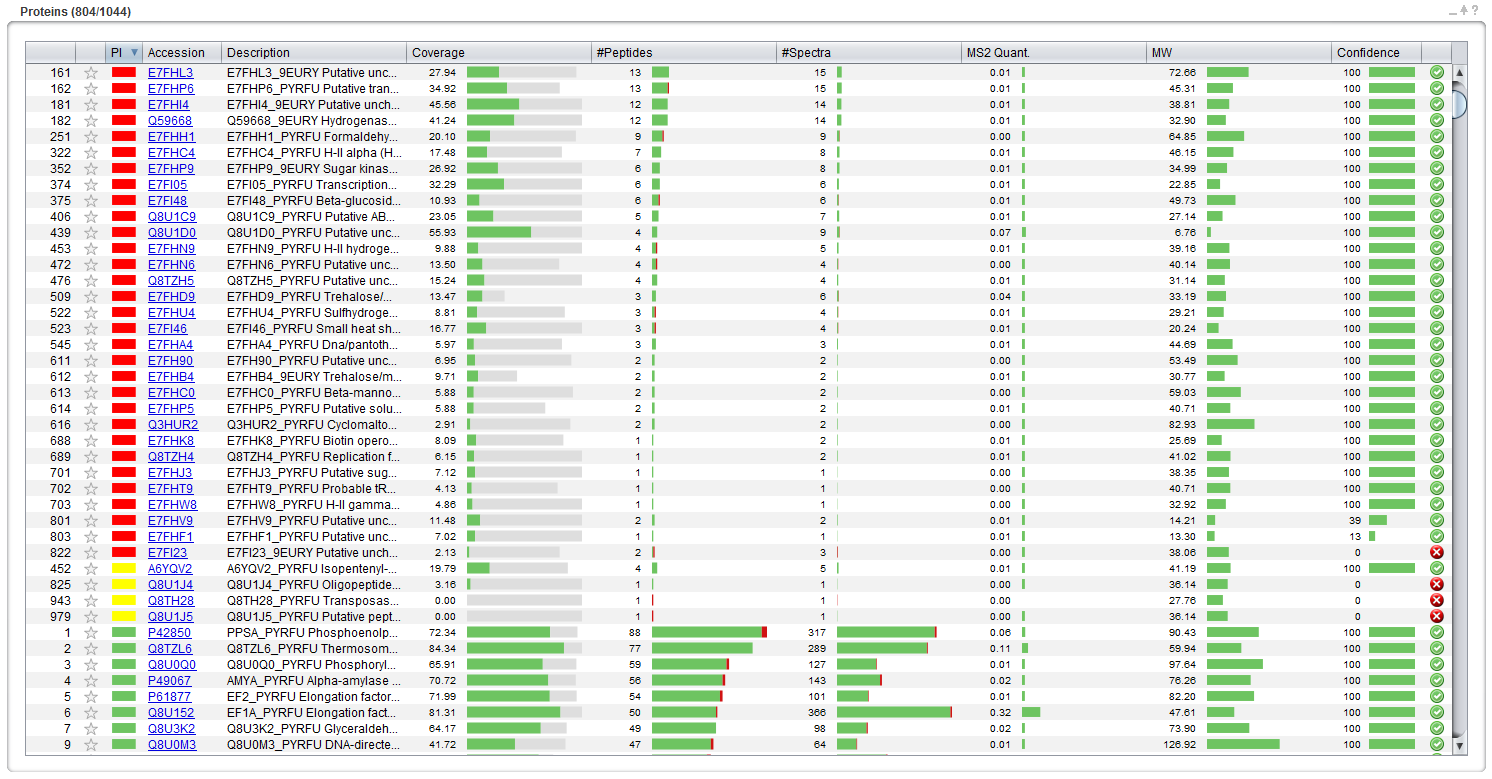
*How many peptides would you require to trust a protein identification?*

For the sake of brevity, we will not go through all quality control plots but feel free to explore them by yourself. We welcome any questions/suggestions!

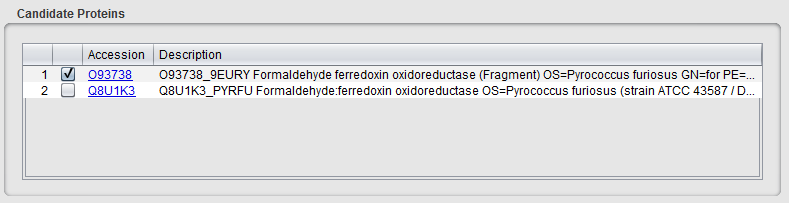
1. Protein Inference

When building proteins from peptides, conflicts occur and these can make it impossible to decide between two proteins that share the same peptide(s). This issue is known as the protein inference problem[2](#_ENREF_1). Go back to the ‘Overview' tab and sort the proteins by the PI column by clicking the column header twice. PI stands for Protein Inference, and the proteins are now sorted so that the proteins with protein inference issues are at the top.

In PeptideShaker, proteins are flagged using four colors: green, yellow, blue and red; as you see, here we do not have an example of a blue category protein. For the best overview, hide all sections but the Protein table:



Click on the red box corresponding to protein accession number O93738, and a Protein Inference Dialog appears, displaying two proteins. PeptideShaker could not decide whether Q8U1K3 or O93738 was present in the sample as the identified peptides are all shared by the two proteins. *According to the description, are these two proteins very different?*



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.

Confidence

A

B

AB

B

AB

A

A

B

AB

Case 3

Threshold

Case 1

Case 2

*In this simple example, how do you interpret the various cases?*

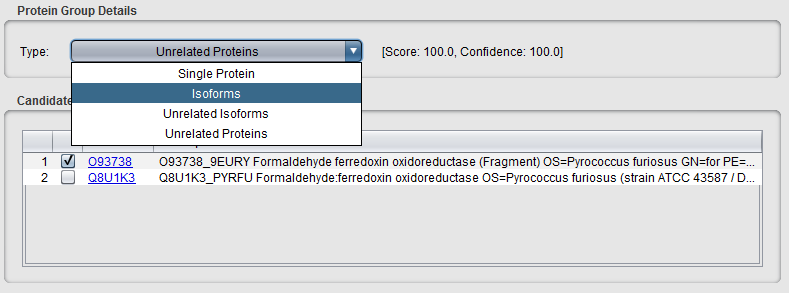
Often, the problem is a lot more complex and involves dozens of intricate peptide to protein mappings. PeptideShaker solves protein inferences of Case 1 and 2, as you may have noticed when loading the identification files. However, Case 3 requires additional information about the sample to find out, if possible, which protein was actually present. These types of conflicts are in the PI column colored in yellow, blue 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 are isoforms of the same protein. Based on the protein descriptions, PeptideShaker sorts the protein inference conflicts into four categories:

* Green: No Conflict, i.e., a Single Protein unambiguously identified
* Yellow: Isoforms identified
* Blue: Unrelated Isoforms identified
* Red: Unrelated Proteins identified

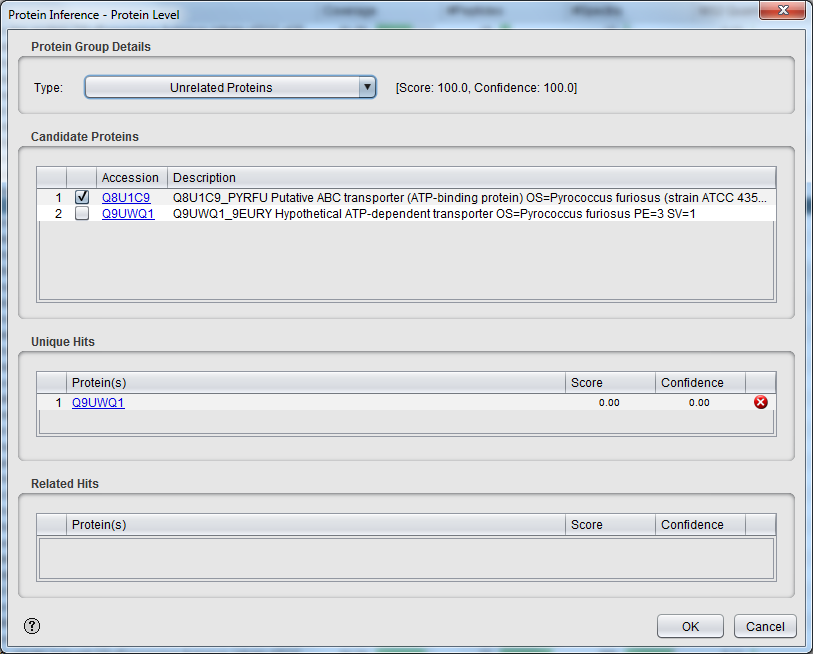
*Based on the O93738* *example, how accurate is the sorting?*

The conflict type can be edited in the Protein Inference Dialog.



Q8U1K3 and O93738 are very similar, and this is therefore not too problematic a conflict; you can consider flagging it as Isoforms. Note that the change is updated in the protein table.

Now look for protein Q8U1C9 (row index 406), and click on the red box in the PI column. A dialog similar to the following should appear:



Here, peptides belonging to Q8U1C9 were found shared with Q9UWQ1 with a high confidence (100%). PeptideShaker however indicates that unique peptides from Q8U1C9 were also found at a lower confidence (0.00%). Similarly, if peptides shared with one of these proteins were found in another group, the group would be related in the 'Related Hits'. The results thus consist of a protein group (Q8U1C9 or Q9UWQ1) which passed the FDR threshold and a single hit (Q9UWQ1) which did not make it.

In order to navigate these hits, we will use the Find feature of PeptideShaker: type Q9UWQ1 on the top right corner of the interface.

C:\Users\vaudel\Documents\rendus\tutorials\tutorials protein identification\4 identification results\illustrations\ps 23.png

Using the two arrows, you can navigate the two groups related to this accession. *Tip: Note that the Find feature accepts protein accessions, descriptions and even peptide sequences.* *Select the unique protein hit, can you find the unique peptide?*

*Now, in this example, what is the protein most likely to be present in the sample? How can experimental knowledge help in this case? How does the database choice influence the protein inference?*

Note that you can change the retained accession for any group.

*Can we delete a useless group?*

References

1. Martens, L. et al., *Proteomics* **5**, 3537-3545 (2005).

2. Nesvizhskii, A.I. & Aebersold, R., *Mol Cell Proteomics* **4**, 1419-1440 (2005).