Protein Quantification Using Reporters

Identifying the proteins in a sample(s) is in many cases only the first step in the data analysis. Usually one also wants to find the abundance/amount of the detected proteins. There are various methods that can be used for this purpose, often divided into label-based and label-free methods.

In label-based quantification one typically examines two or more protein samples simultaneously. To distinguish the variants of the peptides, the peptides are tagged with different labels, where each sample gets a unique label. After labeling, the samples are combined, providing a single mixture for analysis. In some labeling methods one of the samples can simply remain unlabeled, allowing *n + 1* samples to be combined, with *n* being the number of labels.

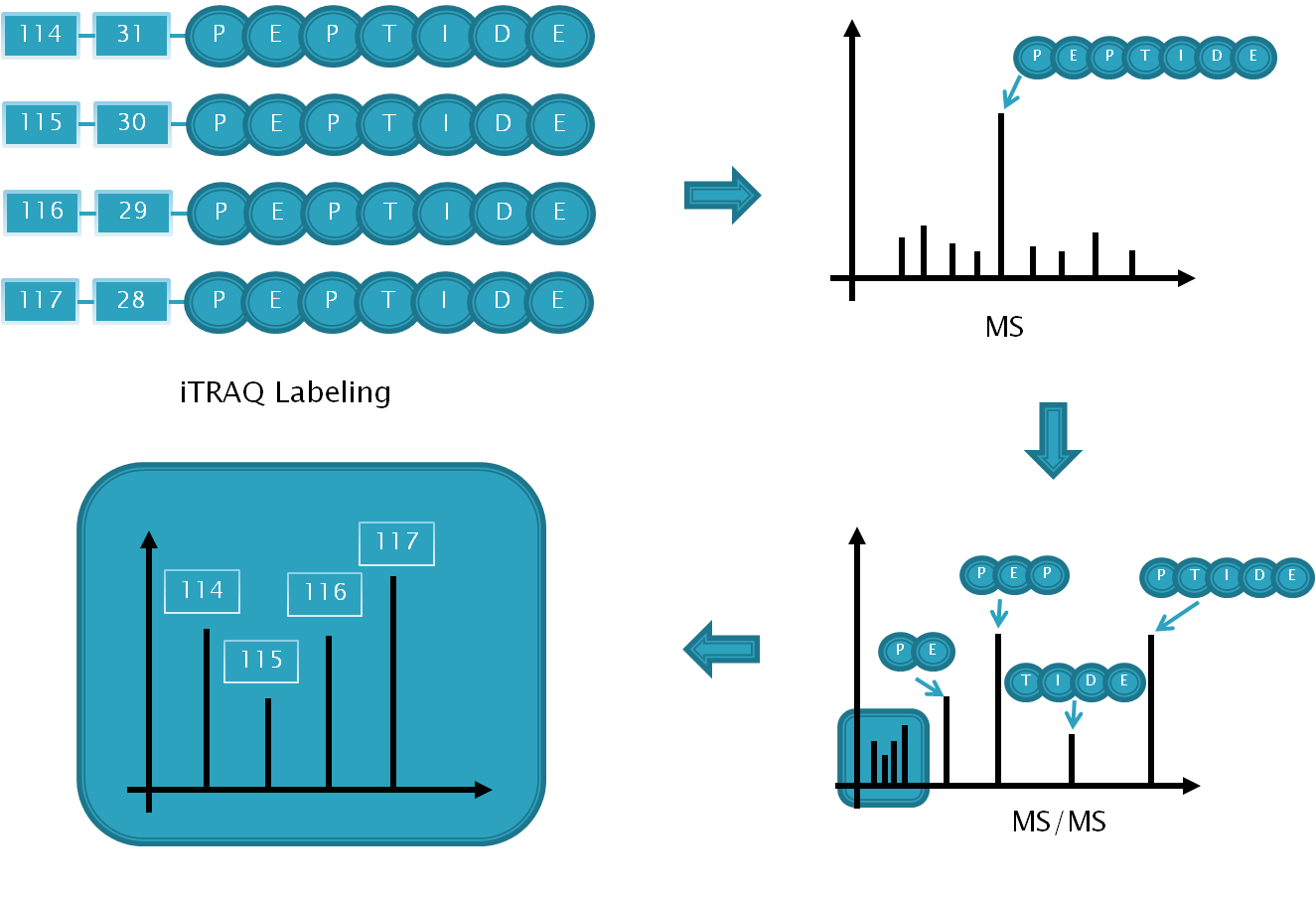
Here we will look at one group of label-based methods referred to as reporter based quantification, where the differentiation of the (labeled) peptide variants in an MS/MS spectrum is based on the dissociation of the labels in the fragmentation process, with each label producing a so-called reporter ion appearing as a separate peak at a fixed m/z-position. The reporter peaks are then used for  
   
quantification, and multiplexing is usually employed since the different reporter ions are chosen to be clearly distinguishable based on their unique m/z in the MS/MS spectrum.

Figure : Overview of the iTRAQ process.

To ensure easy detection of the reporters they should occur in areas of the spectrum where there are few or (preferably) no other peaks. This reduces the chance of the reporter ions interfering with the rest of the spectrum (or vice versa), allowing the reporters to be quantified accurately, and the spectrum to be identified in the standard way, see Figure 2 below.

The two common reporter based techniques are iTRAQ and TMT (Tandem Mass Tag), where iTRAQ supports 4 or 8 samples, and TMT 2 or 6 samples. In the following we will use iTRAQ, but the same procedure can be used for TMT.

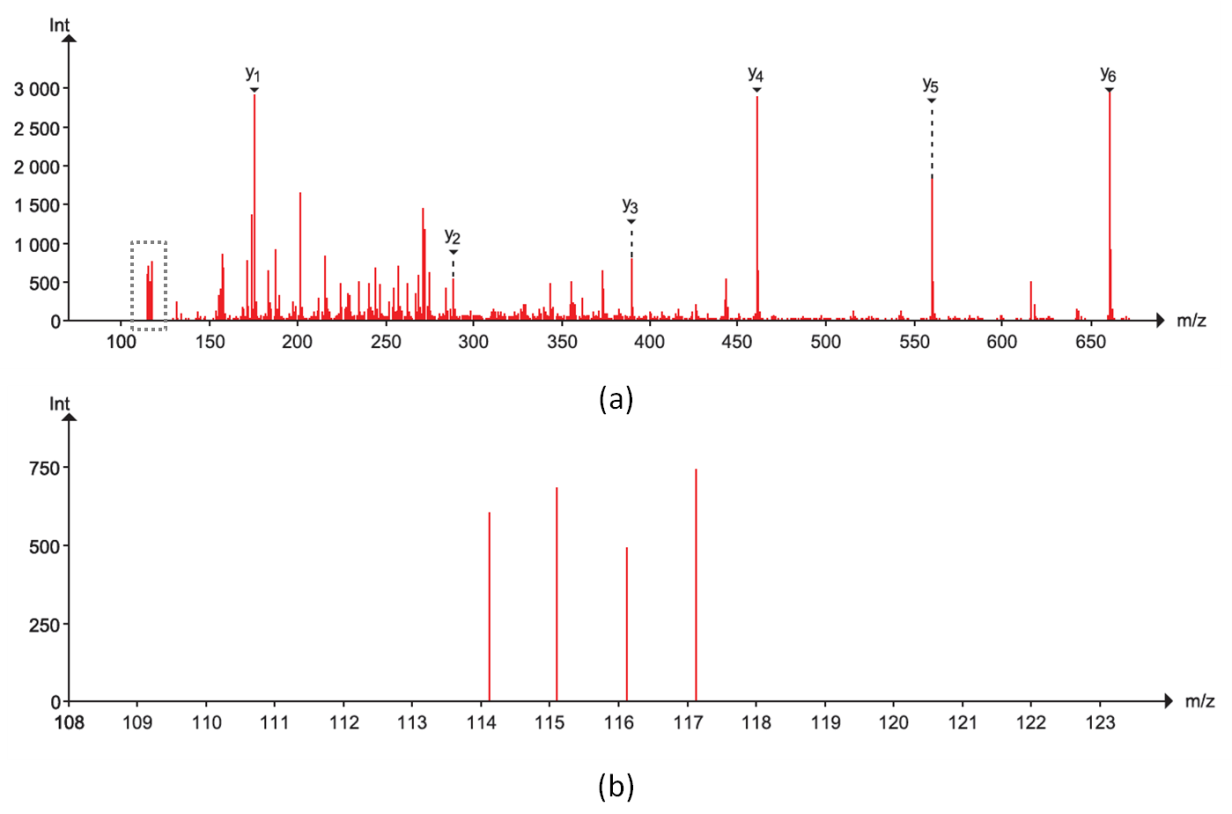


Figure : (a) shows a full MS/MS spectrum with the y-ions annotated, and the region for the reporter peaks indicated. (b) zooms in on the reporter region, highlighting the reporter peaks at positions 114.1 to 117.1.

In the resources folder you will find two mgf files: iTRAQ\_yellow.mgf and iTRAQ\_red.mgf. Both have been created using so-called 4plex iTRAQ, meaning that four labels have been used, most commonly referred to by the (integer) masses of the labels: 114, 115, 116 and 117. (For 8plex we add 113, 118, 119 and 121.)

The experimental setup is as follows: Three non-human proteins have been spiked into four samples of human cerebrospinal fluid (CSF). The amount spiked in varies for the three proteins and varies across the samples. In the yellow experiment the CSF used is identical across all labels, i.e., we should have "identical" background, while in the red experiment the CSF samples come from three different patients and the background cannot be assumed identical.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 114 | 115 | 116 | 117 |
| Yellow | CSF Average | CSF Average | CSF Average | CSF Average |
| Red | CSF Average | CSF 1 | CSF 2 | CSF 3 |

Table : The experimental setup.

The goal is now to try to find the three spiked in proteins, and the amounts spiked in!

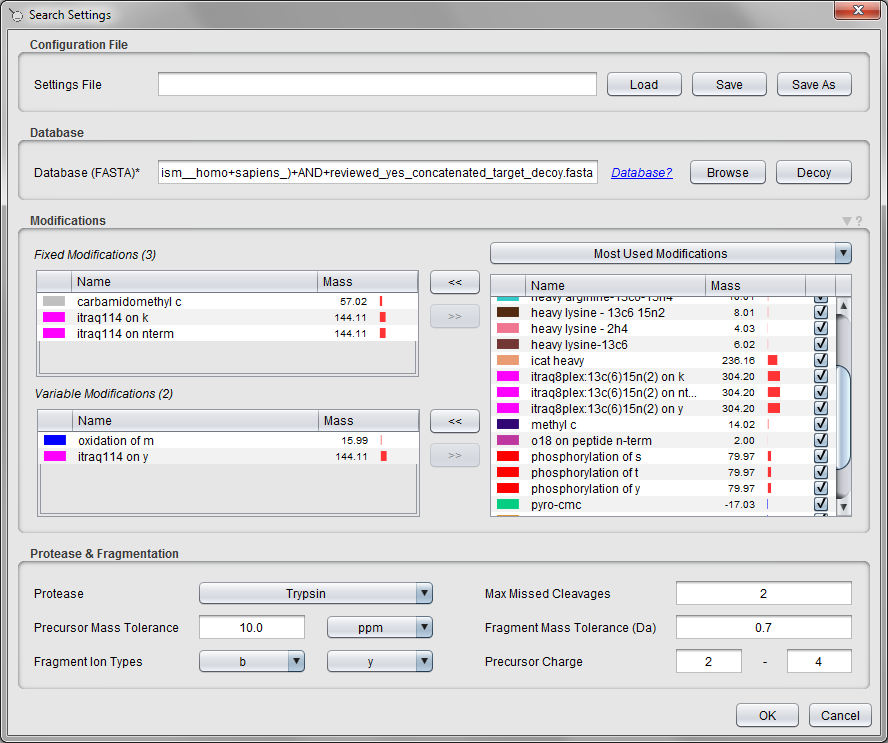
We will start with the simpler yellow dataset.

Open SearchGUI and select the file iTRAQ\_yellow.mgf as the spectrum file. Click 'Edit' behind the 'Settings File' to open the Search Settings dialog, and for the database choose the file uniprot-(organism\_\_homo+sapiens\_)+AND+reviewed\_yes+3\_spike\_in.fasta (also located in the resources folder). Click 'Yes' to the question of if you want to add decoy sequence. This database contains all the standard reviewed human protein sequences plus the sequences of the three spiked in proteins at the end. *Why do we need to add the three spiked in proteins to the database? [2.2a]*

Now select the following modifications: Fixed: carbamidomethyl c, itraq114 on k and itraq114 on nterm; Variable: oxidation of m and itraq114 on y. *Why is it ok to only select the 114 variants of the iTRAQ modifications even though we have used all four labels in our experiment? And why should iTRAQ 114 be variable on Y, but fixed on n-term and K? [2.2b]*

Also change the Fragment Ion Tolerance to 0.7 Da.

You should now have the following settings:



Save the search settings at the top, and click 'OK' to close the dialog.

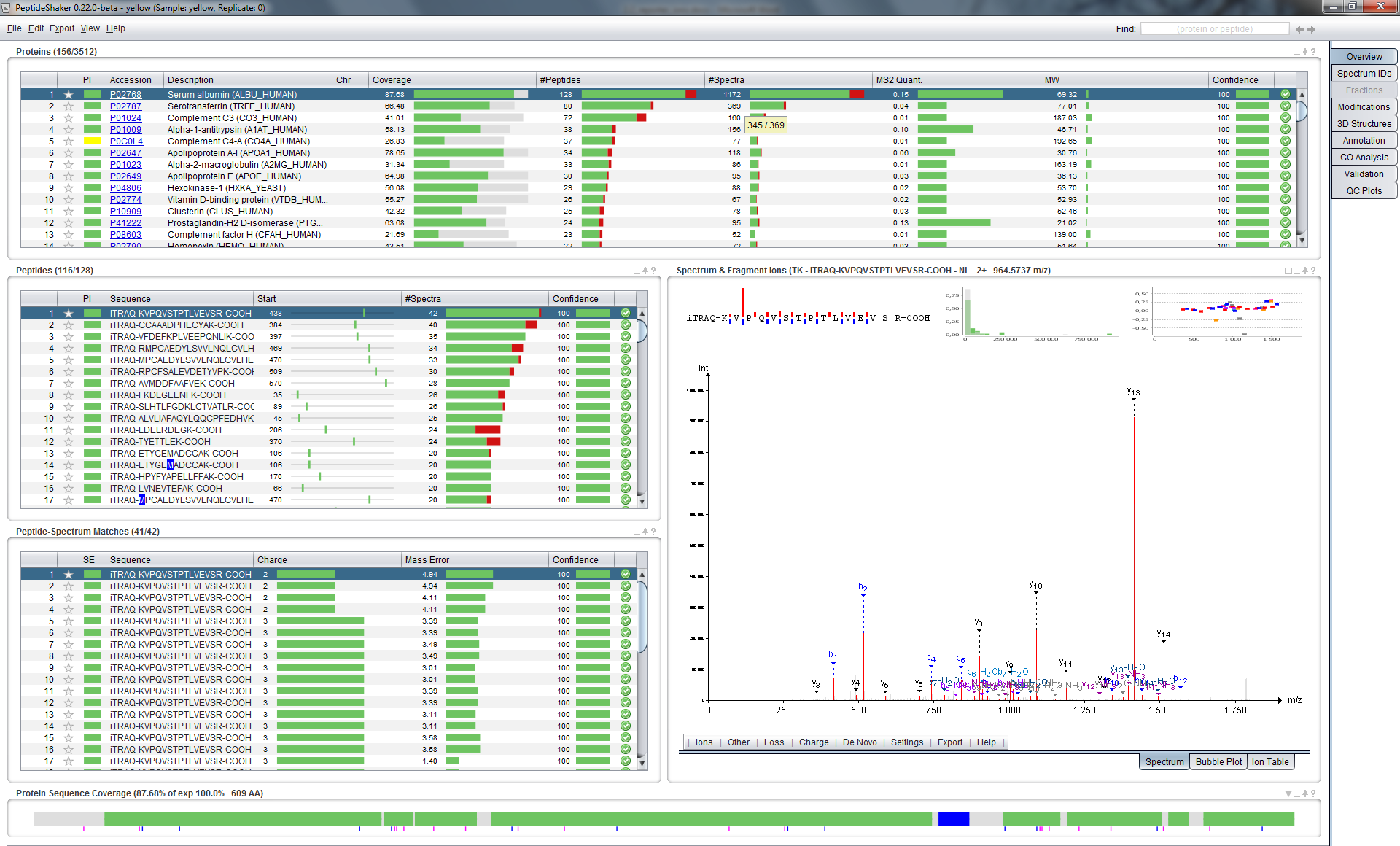
Set up post-processing in PeptideShaker choosing 'Yellow' as the Project Reference, the Sample Name and the name of the output file.

Finally, select an output folder for the search in the main SearchGUI dialog, and click the 'Start the Search!' button.

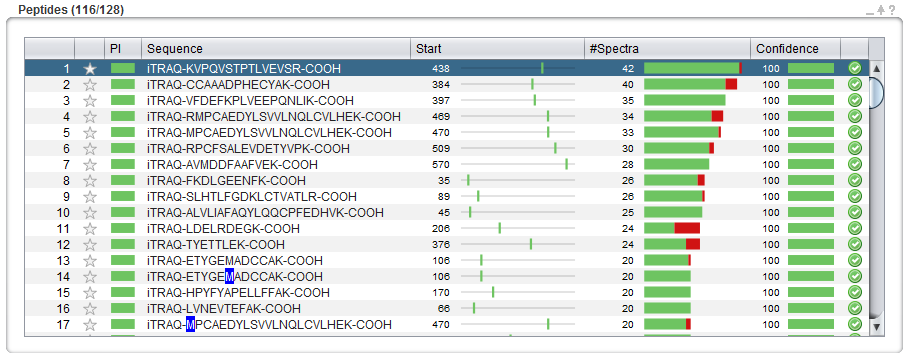
**Tip:**  
*It is strongly recommended to always select an empty folder for the SearchGUI output. This makes it a lot simpler to load the results in PeptideShaker later, and reduced the chance of errors occurring.*

***The search should only take a couple of minutes.***

You should now have the search results displayed in PeptideShaker:



Note how all the peptides have iTRAQ annotated at the n-terminal:

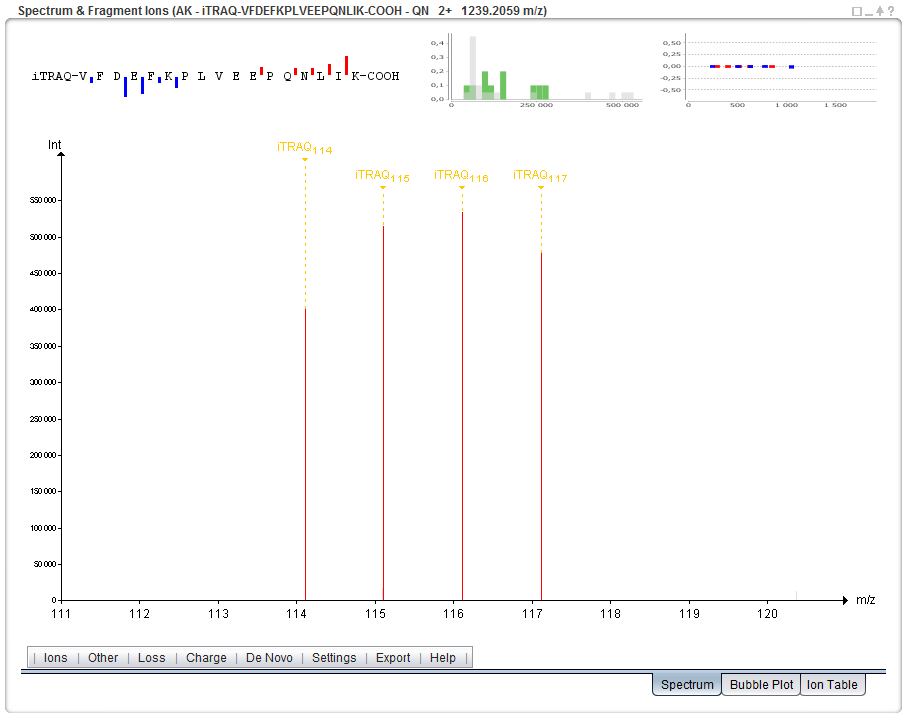


For the top protein (Serum albumin), select the peptide (iTRAQ-VFDEFKPLVEEPQNLIK-COOH).

If you look at the peptide-spectrum-matches (PSMs) for this peptide you will see that half of them have iTRAQ reporter ions annotated in the spectrum, and half not. The reason is that we for this experiment set up the instrument to iterate between two different fragmentation types: CID (Collision Induced Dissociation) and HCD (High-energy C-Trap Dissociation). CID usually results in good (peptide) sequence coverage, which is important for peptide identification, but have a cutoff in the low mass region. While HCD (traditionally) results in reduced sequence coverage, but on the other hand has a better resolution in the low mass region where the interesting reporter ions are located. Using a combination of CID and HCD was therefore previously used to provide the best of both worlds: good sequence coverage ***and*** high resolution in the low mass region resulting in better detection of the reporter ions.

In the recent years the HCD fragmentation has however been greatly improved, to the extent which CID is no longer needed when doing reporter based quantification. (CID is still the preferred option when maximum peptide sequence coverage is desired. In addition, CID is quicker, and thus more peptides can be sequenced.)

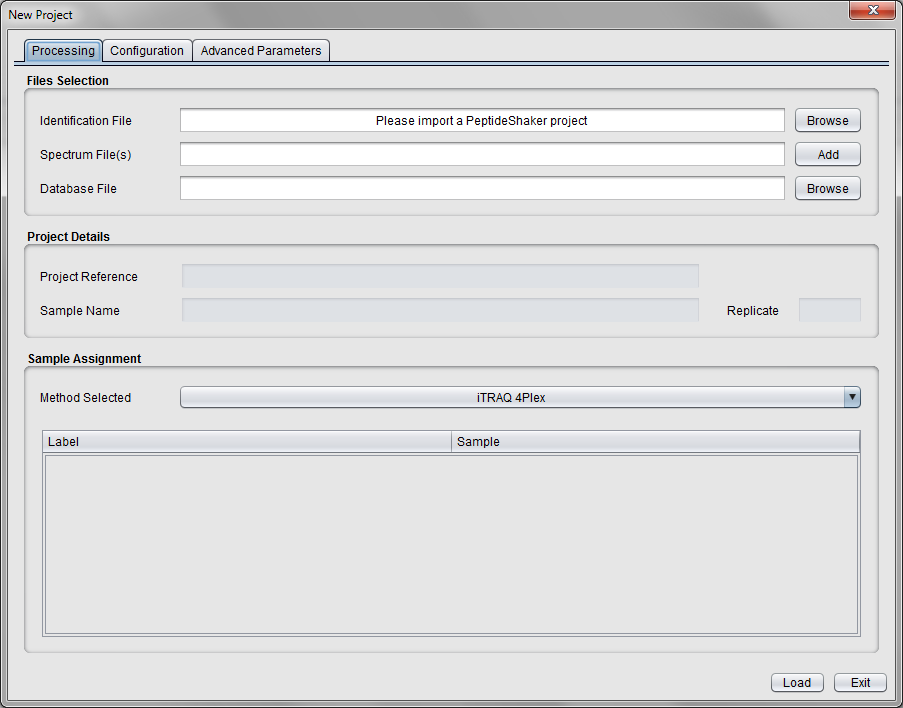
Still using the iTRAQ-VFDEFKPLVEEPQNLIK-COOH peptide as the example, zoom in on the iTRAQ reporter range and see if you can see any differences between the height (intensity) of the peaks. *Do you see the same pattern when looking at other PSMs from the same peptide? Why should the pattern be similar across the PSMs? What about across different peptides from the same protein? How would shared peptides affect the results? [2.2c]*

**

So far we have only looked at the iTRAQ intensities at the spectrum level. But before we can look at the same information at the peptide and protein level we need to do some calculations. For this we will use a combination of a PeptideShaker add-on called Reporter, and Excel. The Excel part is needed as Reporter is not yet completed. In the near future the semi-manual Excel work below will instead be carried out using Reporter.

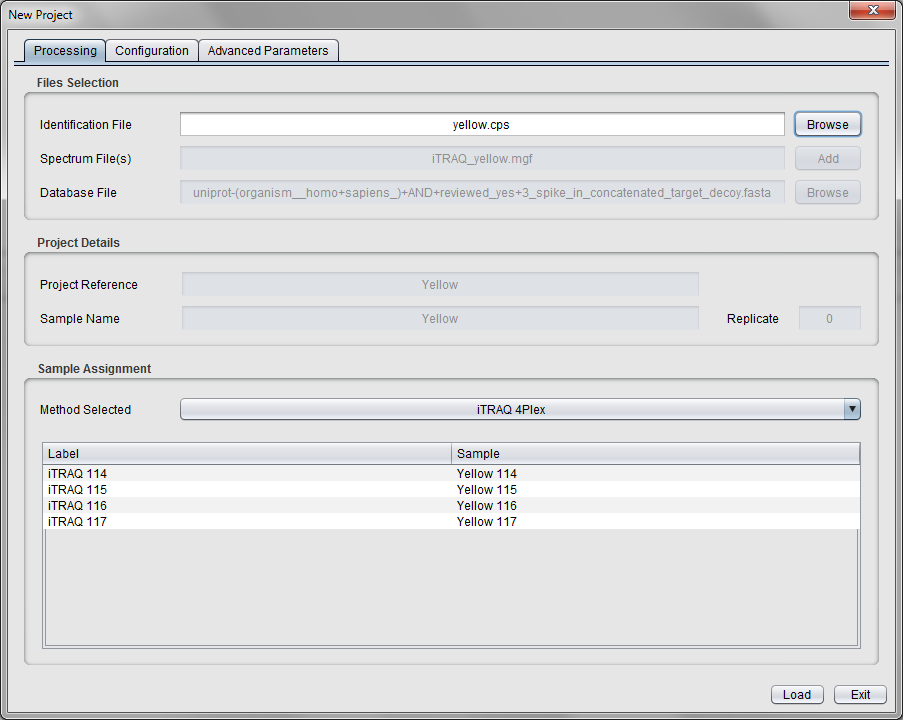
Your PeptideShaker project should already have been saved, if not save it now (File > Save As...) and use the file name 'Yellow'.

Now open Reporter. You should see this dialog:



Click the 'Browse' button and select the just saved PeptideShaker project called 'Yellow'.

Note that the rest of the information is filled in and that iTRAQ 4plex is chosen as the detected method.



The 'Configuration' tab provides more information on the expected masses of the reporter ions and isotope correction factors provided by the iTRAQ vendor. *Why is isotope correction required? [2.2d]*

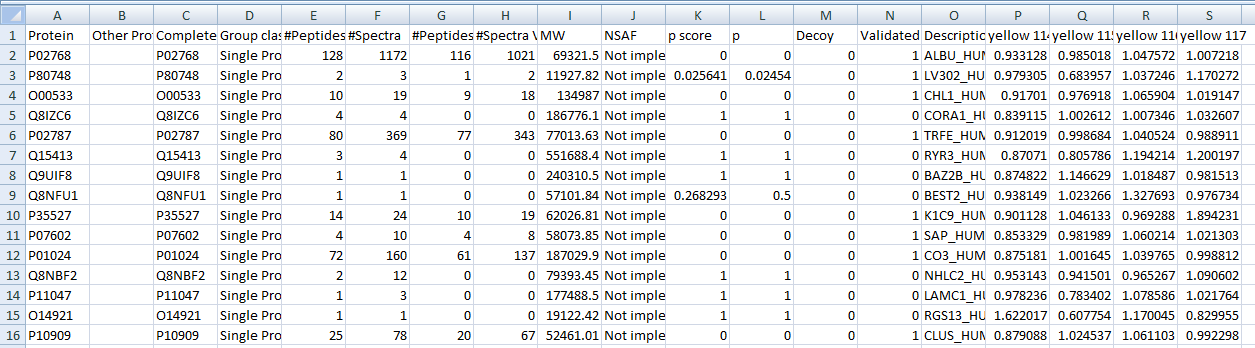
Now open the 'Advanced Parameters' tab and under 'Link to Quantification' on the right select the 'Precursor Matching' option. Note that this option is only needed because the spectra are split into HCD and CID spectra. When using the now more common HCD only setup, use the default 'Same Spectra' option. The other default values are fine.

Click the 'Load' button in the lower right corner to start calculating the PSM, peptide and protein quantifications, and wait for the process to complete.

***At this stage we need to move to Excel to continue the analysis...***

To get the data out of Reporter and into Excel, go to the 'Export' menu in Reporter. Note that you can here export the quantification results are three different levels: protein, peptide and PSMs. Choose the 'Protein' option, select a file to save the results to and click 'OK'.

When the export is done, open the file in Excel. The top part should look something like this:

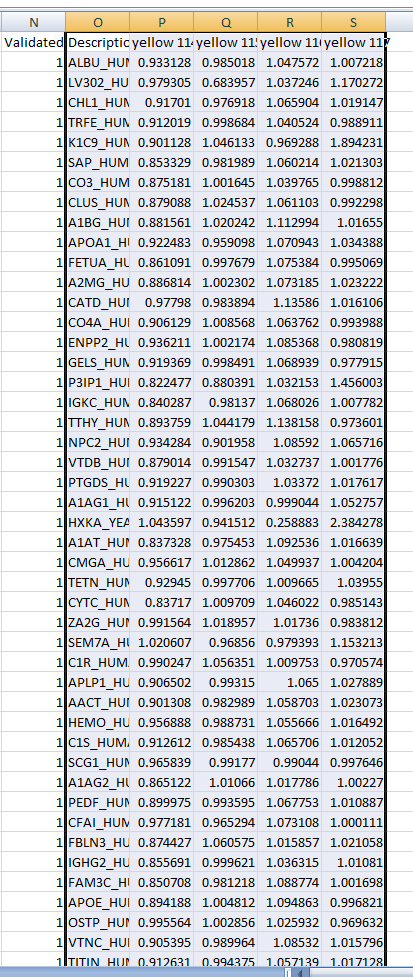


Now we will try to locate the spiked in proteins. First sort the list of proteins such that the validated proteins (the ones with a 1 in the column called 'Validated)' are at the top.

Before plotting the data it is also recommended to remove all the 'NaN' values with. *Hint: Click Ctrl + H for 'Find and Replace'.*

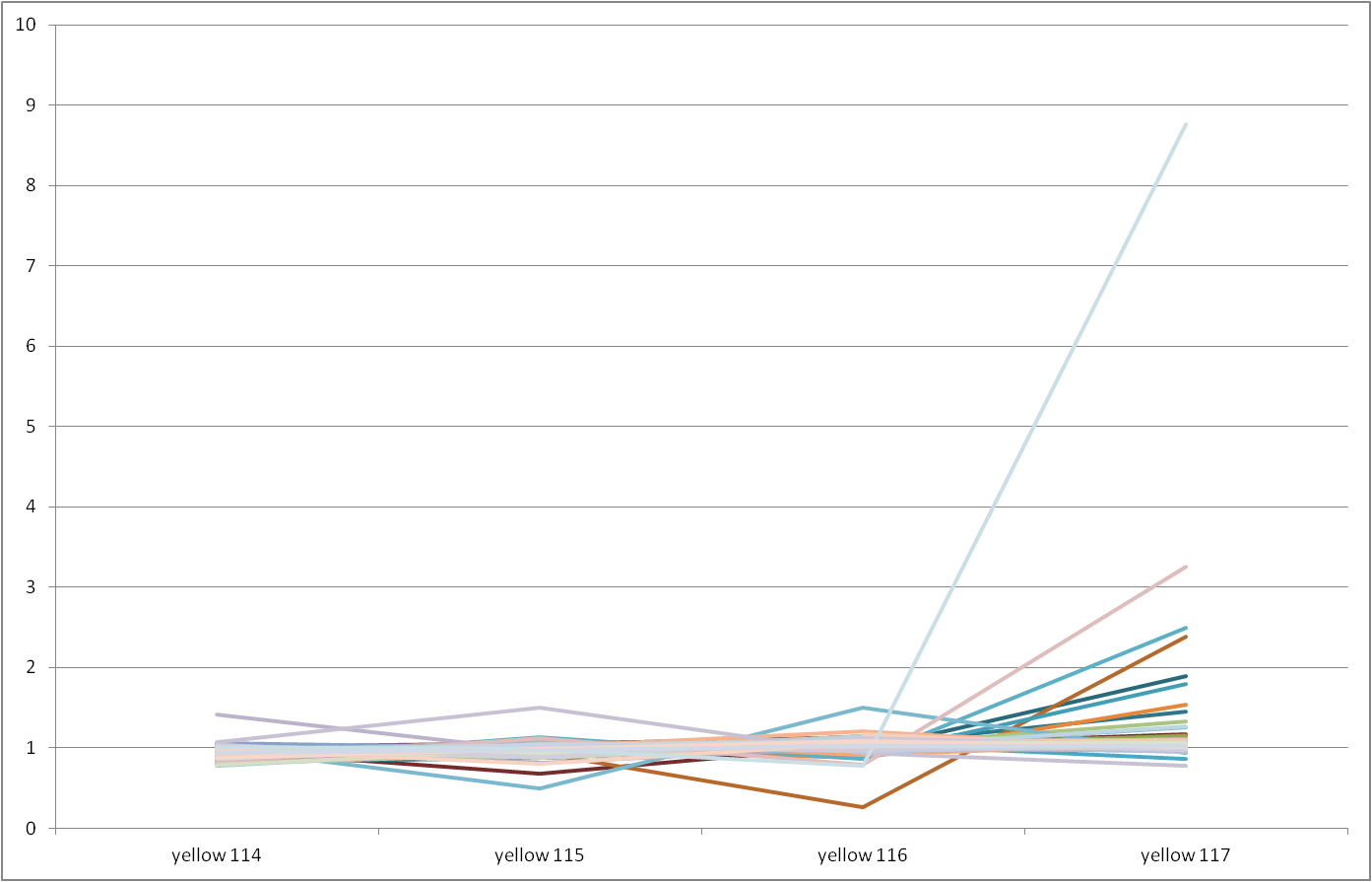
Select all the validated proteins in the last five columns (Description, Yellow 114, Yellow 115, Yellow 116 and Yellow 117).

Example:



Create a line plot based on the selected cells. (You may have to switch the rows/columns so that you get the Yellow 114, Yellow 115 etc on the x-axis.)

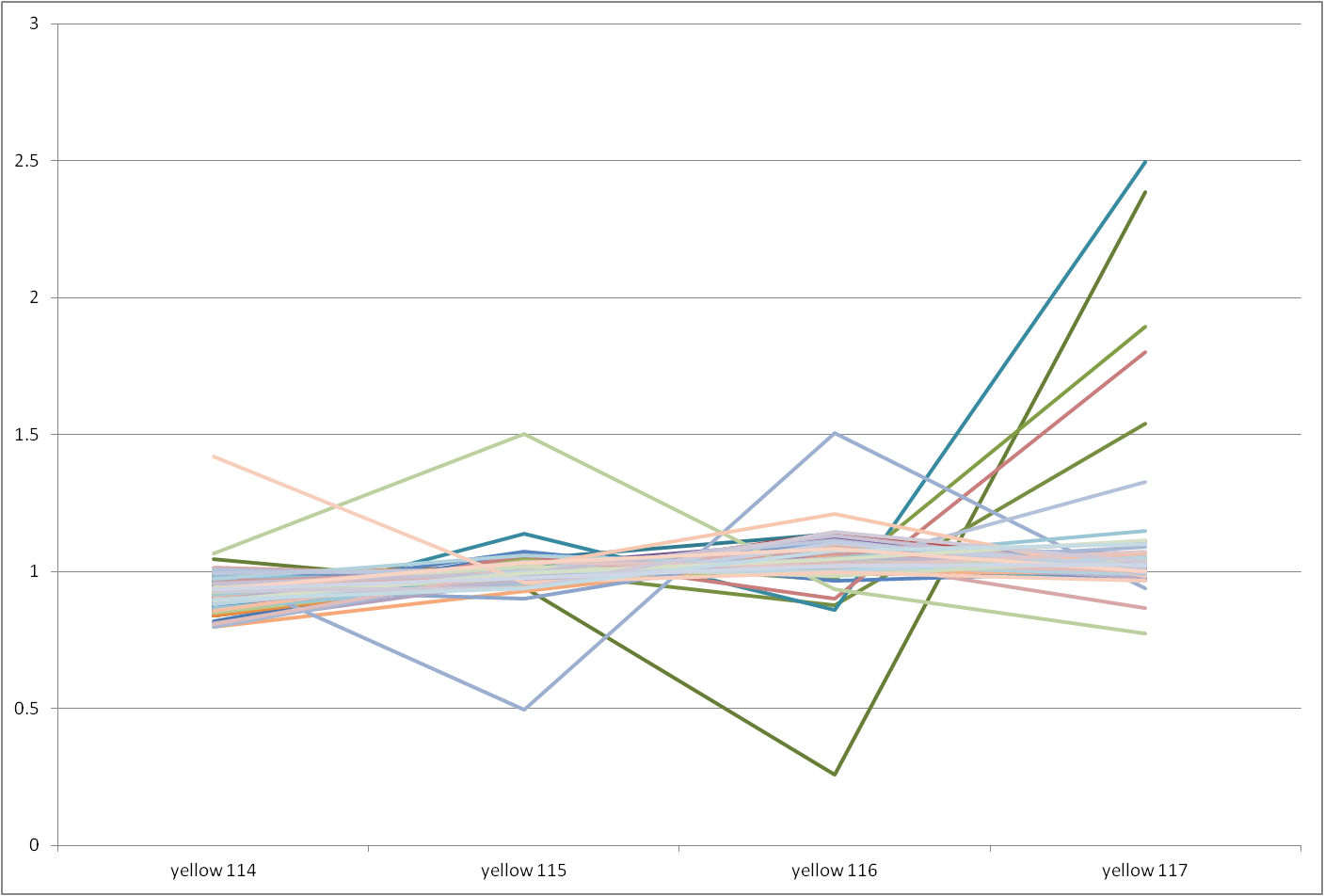
The plot should look like this:



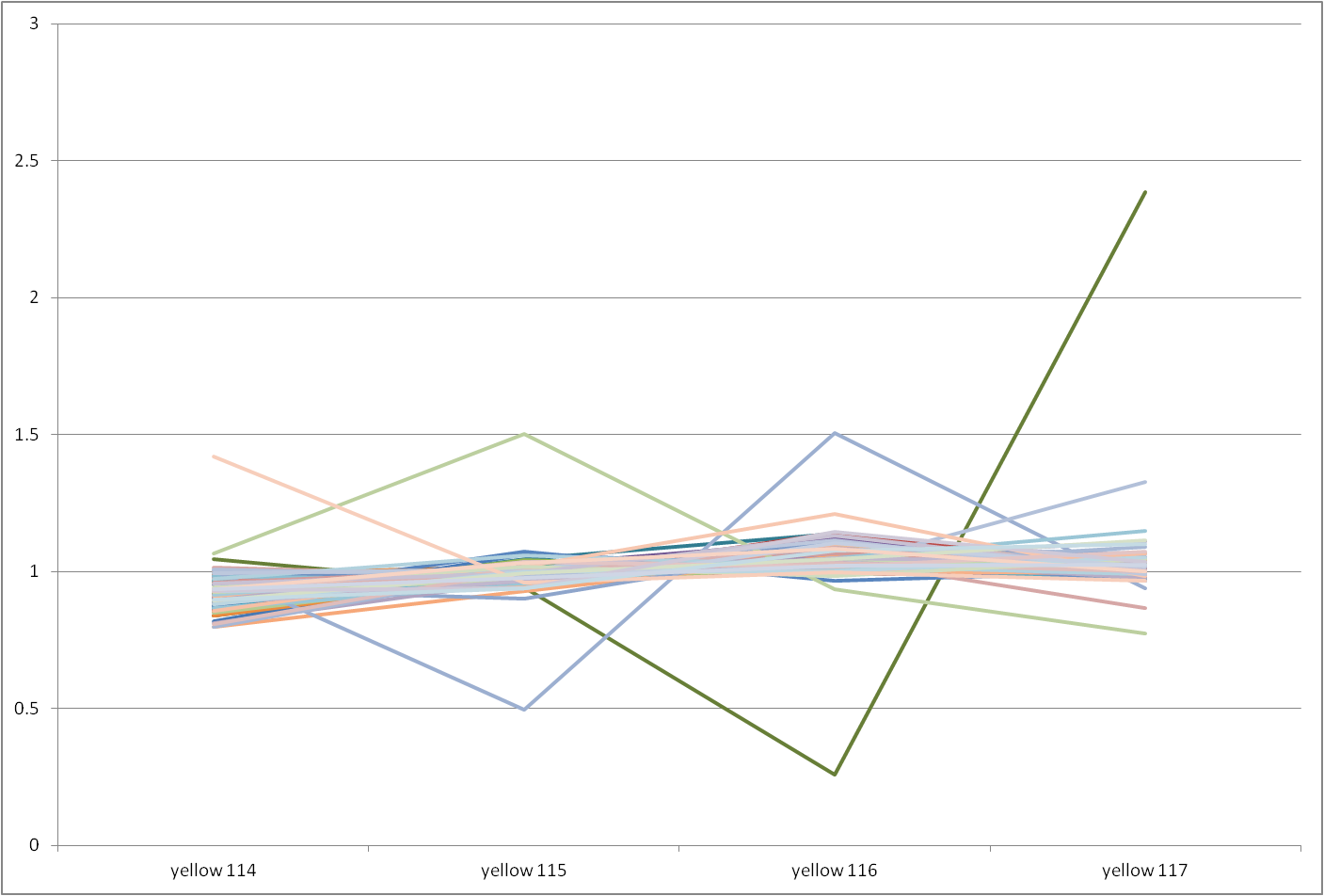
There is an obvious outlier in Yellow 117. If you hold the mouse over the line you will see that this is the protein 'TPM3\_HUMAN Tropomyosin alpha-3 chain'. However, if look at the rest of the row for this protein you will see that it only has one validated peptide and two validated spectra. *Can we trust a quantification with just one validated peptide and two validated spectra? [2.2e]*

Redo the plot, but this time only include the proteins with at least two validated peptides.

The updated plot should look like this:



However, if you hold the mouse over the high values on the right, you will notice that most of these are keratin proteins, in other words contaminants, and should be removed. Remove the keratin proteins by selecting and deleting the lines in the plot.



*Are you now able to pick out the three spiked in proteins?* *[2.2f]*

This was the simple dataset with identical background in all samples. Now redo the same analysis for the more complicated red dataset.

***If you do not want to repeat all the steps above, the Reporter Excel protein export can be found in the resources folder.***

*Are you able to pick up the same proteins in the red dataset? Are they as easy to detect as in the yellow dataset? If not, why? [2.2g]*

Finally, remember that this tutorial was carried out on a rather simple dataset where we knew what we were looking for. This is of course not the case for real experiments. Which makes protein quantification one of the most challenging tasks to do correctly in proteomics today.

References

(to be added...)