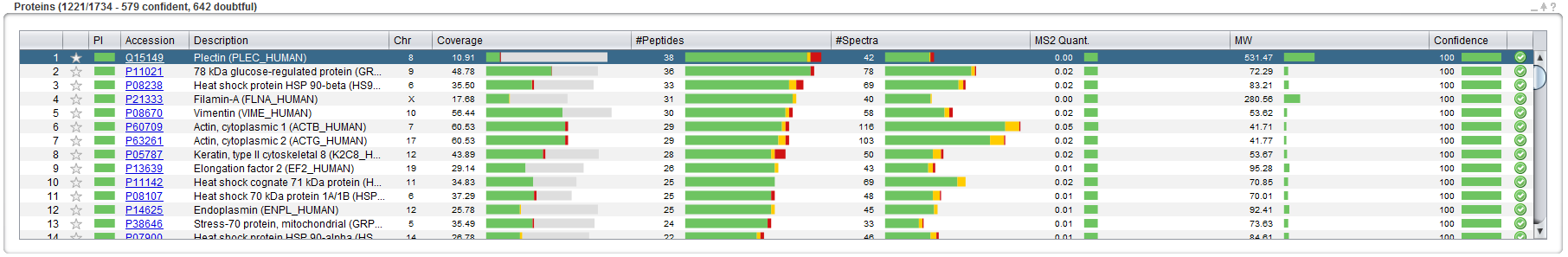
Peptide and Protein Validation

Start by loading the PeptideShaker example project (that same project that has been used in the previous chapters). The project is easily available from the PeptideShaker Welcome Dialog. We are now looking at the protein table:



If you scroll down the protein list you will note that some of the proteins further down the list are supported by very few identified peptides that in total do not cover much of the corresponding protein sequences. Some of these low quality hits are likely to be false identifications introduced by errors of the search engines. We are now going to validate the good quality hits, filtering out (most of) the false positives, and keeping the true positives.

|  |  |  |
| --- | --- | --- |
|  | Validated Proteins | Rejected Proteins |
| Proteins Actually  in the Sample | True Positive | False Negative |
| False Identifications | False Positive | True Negative |

As seen in the table above, depending on your validation threshold, protein hits can be sorted into four classes. *Which population do we want to retain? To control?* *[1.5a]*

Note that PeptideShaker provides a confidence for every protein, peptide and peptide to spectrum match (PSM). These metrics provide an unbiased estimation of the quality of the hits, independent of the sample, the mass spectrometer and the search engine. How is this possible? When using SearchGUI earlier in the tutorial, we actually appended sequences of non-existing proteins (so-called decoy sequences) to the protein database. In fact, these fake sequences are the reversed versions of the actual sequences.

Here is an example from our database (the FASTA file):

>sw|Q8TCZ7|CU074\_HUMAN Putative uncharacterized protein encoded by LINC00308 OS=Homo sapiens GN=LINC00308 PE=5 SV=2

MAYVFNLSCLGSQVERLLEARSSRPTWIIQPSPKKAPEACFSFHSSYERNWA

>sw|Q8TCZ7\_REVERSED|CU074\_HUMAN Putative uncharacterized protein encoded by LINC00308 OS=Homo sapiens GN=LINC00308 PE=5 SV=2-REVERSED

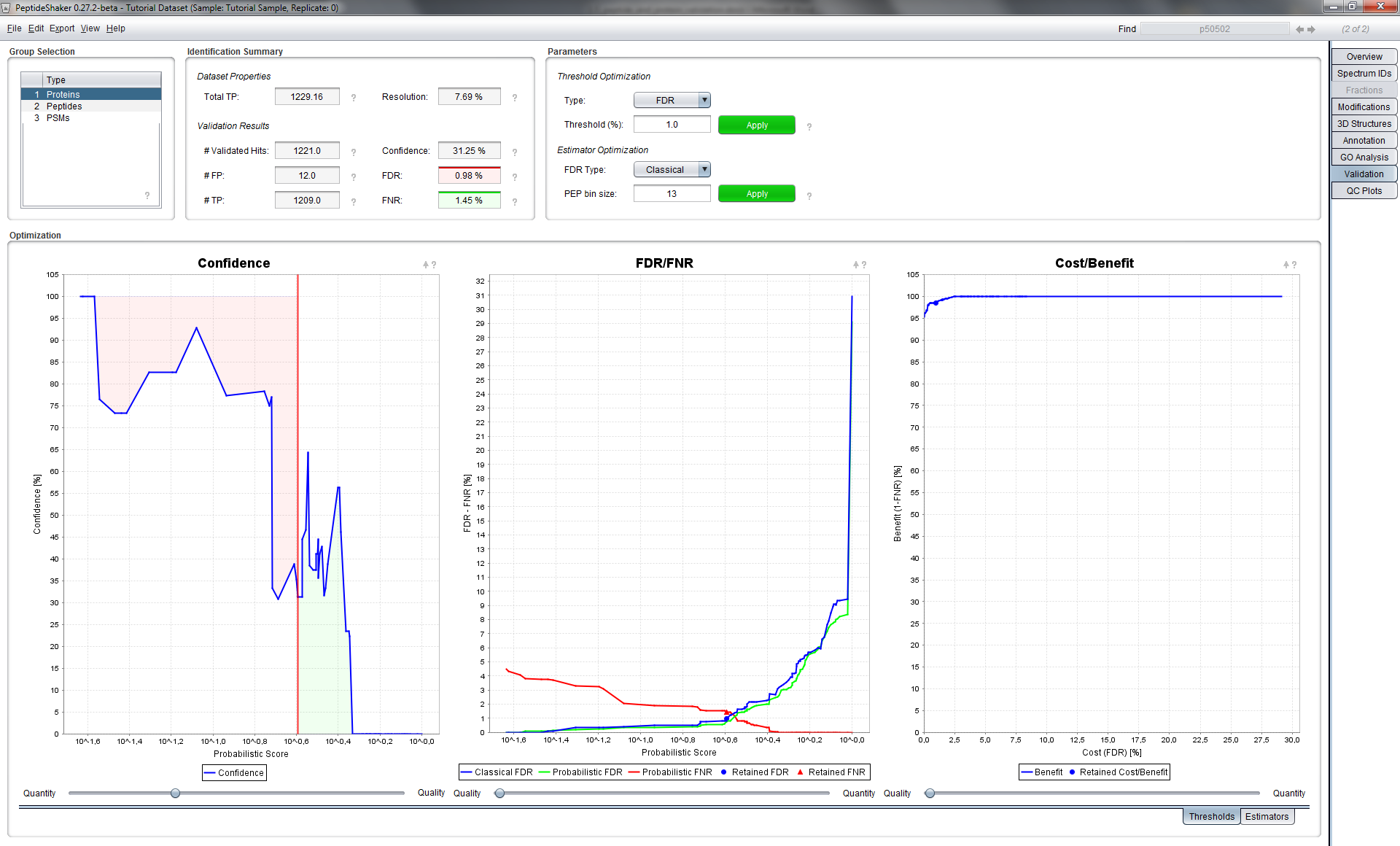
AWNREYSSHFSFCAEPAKKPSPQIIWTPRSSRAELLREVQSGLCSLNFVYAM

Thus, whenever a mistake is made, it is as likely to happen in the real database (called the target database) as it is in the artificial database (called the decoy database).[2](#_ENREF_2) When a decoy hit is found among five target hits

We hence assume that there is one false positive among the target hits (20% error). *Do we know which one? Are there other ways to create a decoy database? Which one is the best?* *[1.5b]*

The decoy hits allow us to estimate the number of false positives in a result set. There are two main ways to control the amount of false positives in the validated protein set. First of all, we can set a confidence threshold; typically we would validate protein hits in which we are more than 95% confident. However, scientists usually prefer to control the False Discovery Rate (FDR), the total error share in the result set; typically we allow 1% FDR, meaning that 1% of the retained proteins are expected to be false positives. PeptideShaker already did this standard validation for you. Some may have noticed the symbols at the far right of each row. These indicate that the corresponding protein, peptide or PSM passed a 1% FDR threshold. *On the top of the table you can see that* *1221 proteins were validated here out of 1734, how many false positives do we expect?* *[1.5c]*

The validation threshold can be optimized in the ‘Validation' tab of PeptideShaker. Opening the 'Validation' tab you should see this:



We will now change the validation criteria for our peptides and proteins. The group selected in the top-left box should be ‘Proteins’. The 'Identification Summary' section provides results from our 1% FDR validation. The 'Parameters' section to the right allows us to customize the estimation, and plots below visualize the results and control their quality. We will now only focus on the main settings.

**Tip:**  
*Note the question marks present everywhere to guide you through all the parameters.*

Two metrics can be defined to evaluate the validation procedure: (A) the False Discovery Rate (FDR) indicating the share of retained false positives; (B) False Negative Rate (FNR) indicating the share of false negatives:

|  |  |  |
| --- | --- | --- |
|  | Number of  Validated Proteins | Number of  Rejected Proteins |
| Proteins Actually  in the Sample | nTP | nFN |
| False Identifications | nFP | nTN |

The identification summary indicates that 1221 proteins were validated including 12 false positives. PeptideShaker estimates that a maximum of 1229 true positive proteins could be found in the data set: we are thus including almost all of them.

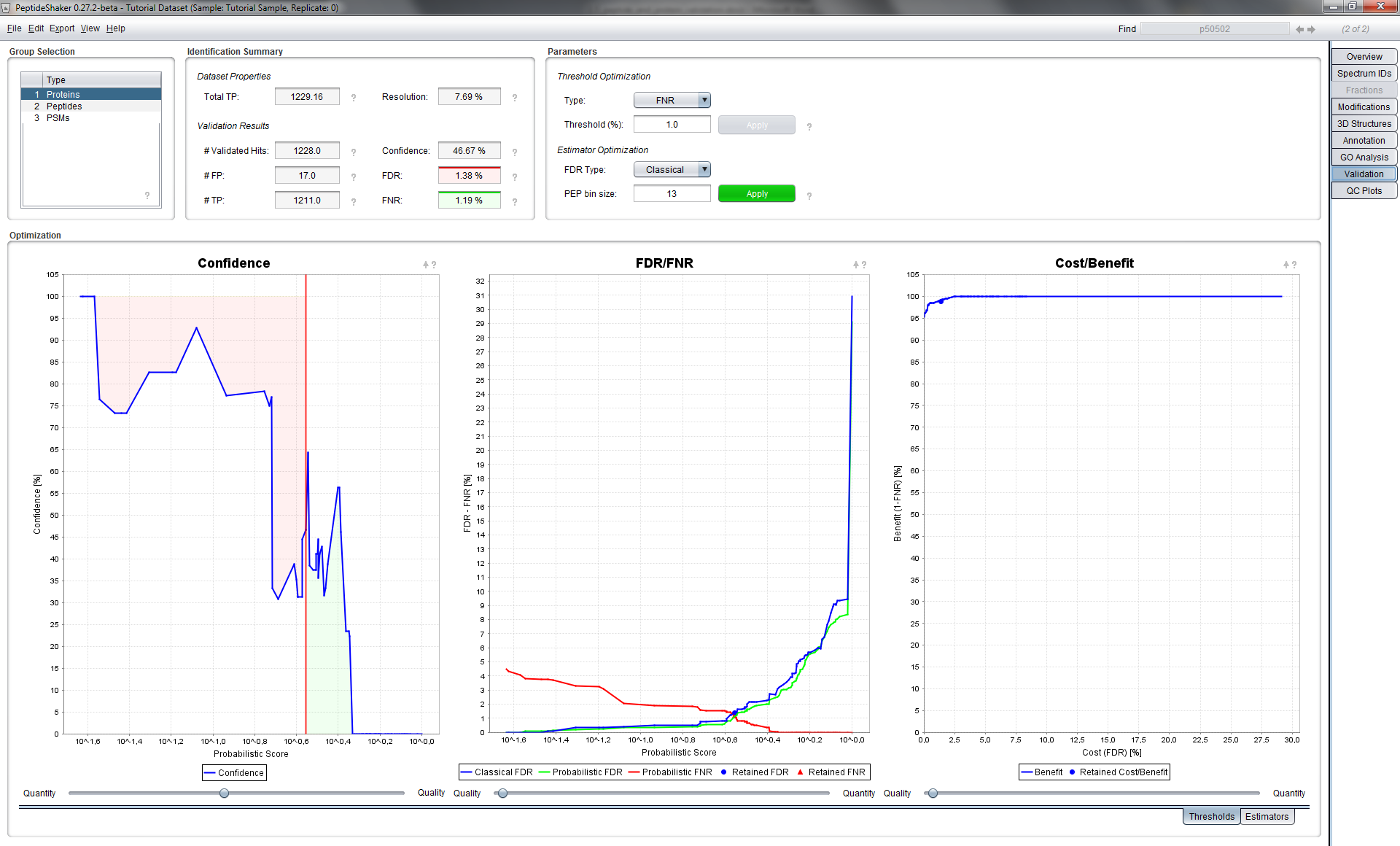
The ‘Validation Results’ show that the FDR limit used is actually 0.98%. *Why is it not 1%? [1.5d]*

The three plots at the bottom display the current threshold settings. The Confidence Plot shows the variation of the protein confidence against the score and the chosen threshold in red. Note that the green and red areas in this plot represent the number of false positives and false negatives, respectively. These are used to estimate the FDR and FNR values displayed in the FDR/FNR Plot and in the corresponding green and red boxes. These metrics allow the drawing of a Cost/Benefit Curve (also called a receiver operating characteristic or ROC curve) which allows you to optimize your threshold. Note that the current setting is represented by a point on the curves. Note that the point can deviate from the curve: this is a direct illustration of the confidence estimation imprecision.

As you can see from the confidence plot, our threshold (red line) is set in an area where the confidence is around 31%. *How accurate is the confidence estimation in this case? If we include hundred proteins at 95% confidence, how many false positives do we expect?* *[1.5e]*

Note that it is possible to set a threshold at a stringent confidence level. However, we are now going to ask PeptideShaker to focus on quantity and set a False Negative Rate (FNR) of 1%. Select FNR as the threshold type, type 1 and hit Enter.

You should see these results:



Note that the red line illustrating the threshold and the points indicating the FDR and FNR have moved to the right in the plots. *What are the new FDR and FNR values? What are the values using a 95% confidence threshold? In your opinion, what is the best threshold?* *[1.5f]*

**Tip:**  
*Use the ‘Apply’ button only when you are happy with the threshold.*

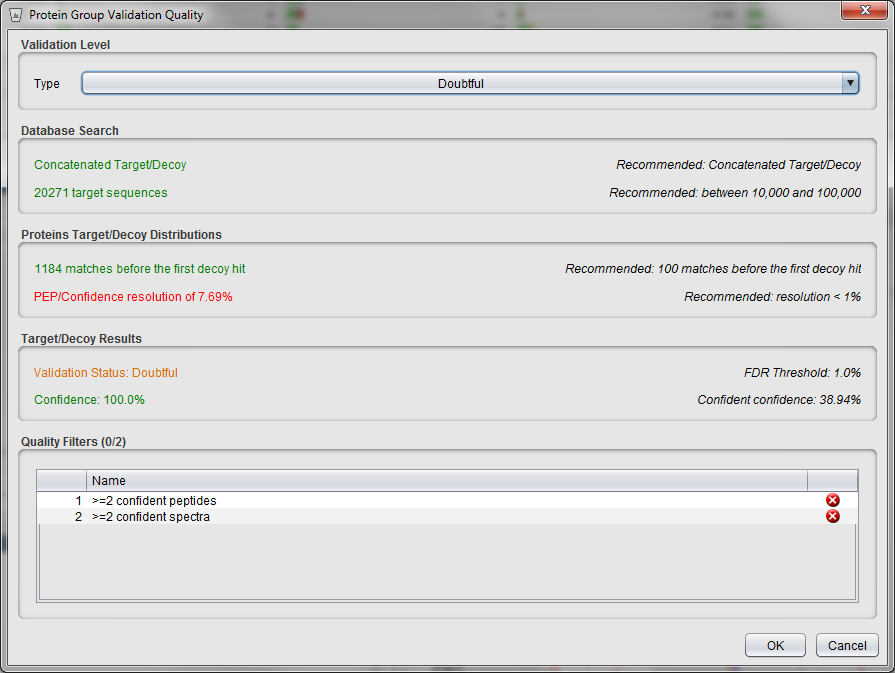
The same operations can also be conducted on peptides and PSMs when changing the selected population in the top left section. *How do the validation metrics for peptides and PSM compare to the protein level?* *[1.5g]*

If you want to apply new validation settings to the actual dataset, click on the green 'Apply' button in the Parameters section. If you go back to the 'Overview' tab, the green and red symbols indicating protein validation will reflect the new validation settings. Note that without clicking the 'Apply' button the new thresholds will be ignored!

**Tip:**  
*When statistical significance is ensured, PSMs are grouped according to their charges and peptides according to their modification status in order to maximize the identification yield.*[*1*](#_ENREF_1)

Now go back to the Overview tab, on top of every table, you will see that PeptideShaker marked some of the validated matches as confident and some as doubtful. At 1% FDR 579 protein matches were marked as confident and 642 as doubtful, out of a total of 1221 statistically validated hits. *Why are protein matches doubtful although validated? [1.5h]*

If you scroll down to the 272, you will see the first group marked as doubtful: Q92598. If you click on the yellow warning icon, a diagnostic dialog will open showing why this protein validation was marked as doubtful:

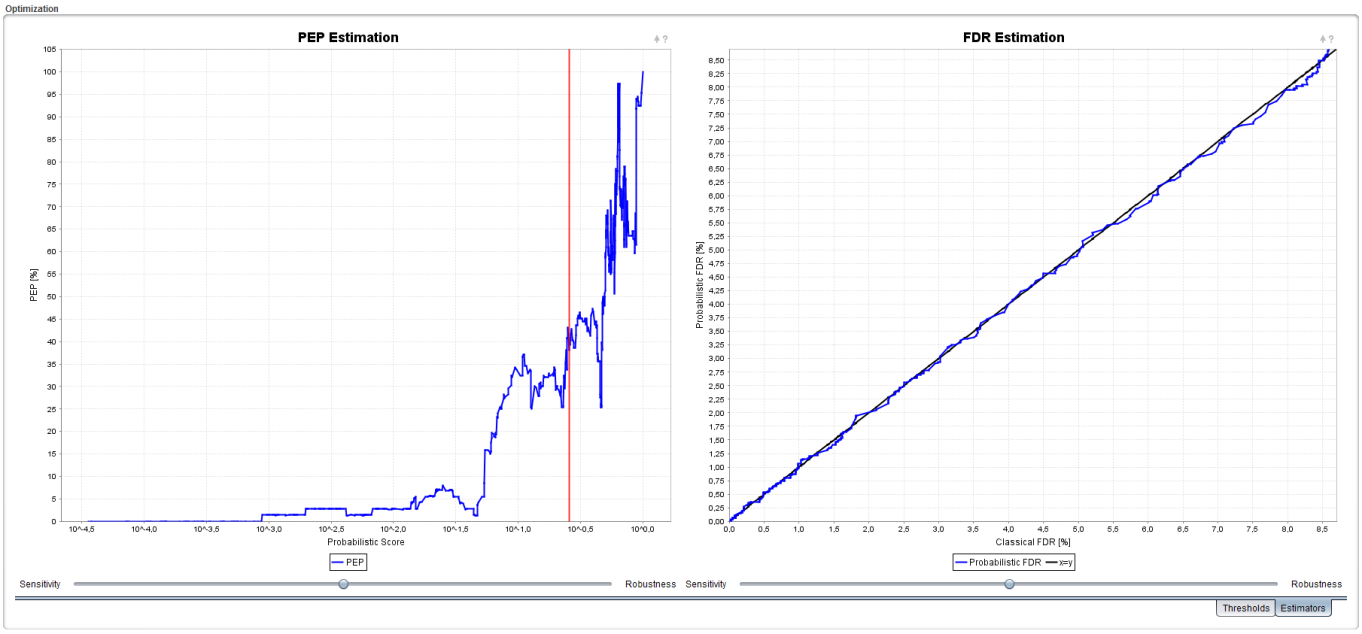


As you can see here, the protein was marked as doubtful because its identification is supported by only one confident peptide and one confident spectrum. This does not mean that the protein identification is wrong, but it should be considered with care.

Similar filters are also used at the peptide and PSM level, and the following colors are always used: green (validated and confident), yellow (validated, but doubtful) and red (not validated).

Advanced

In this advanced section, we will inspect the quality of the validation metrics. Select the ‘PSMs’ category and go to the ‘Estimators’ tab at the bottom right of the screen. You should see the two following plots:



On the left you have the evolution of the Posterior Error Probability (PEP). Keen observers will have noticed that PEP = 100 - confidence (when confidence is given as 0-100%). This plot is thus similar to the confidence plot. On the right, the Probabilistic FDR – which is the FDR estimated thanks to the PEP – against the Classical FDR – estimated thanks to the decoy hits. When the classical FDR is valid (this is questionable with X!Tandem results[3](#_ENREF_7)), this plot is thus a straightforward measure of the quality of the confidence estimation. If the blue line closely follows the black line, the confidence is well estimated. If the blue line deviates from the black line, inaccuracies will occur in the confidence estimation.

*What is your interpretation of this curve for the proteins?* *[1.5i]*

References

1. Vaudel, M., Burkhart, J.M., Sickmann, A., Martens, L. & Zahedi, R.P. Peptide identification quality control. *Proteomics* **11**, 2105-2114 (2011).

2. Elias, J.E. & Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **4**, 207-214 (2007).

3. Everett, L.J., Bierl, C. & Master, S.R. Unbiased statistical analysis for multi-stage proteomic search strategies. *J Proteome Res* **9**, 700-707 (2010).