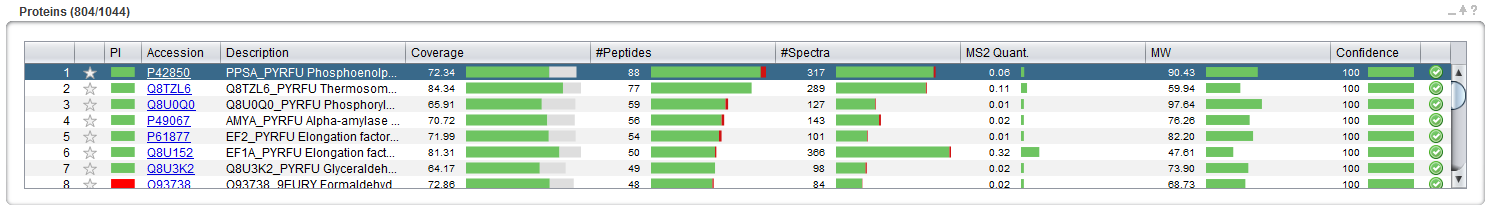
Peptide and Protein Validation

Load the *Pyrococcus furiosus* project which has been used in the previous sections into PeptideShaker (the corresponding cps file is in the resources folder). We are now looking at the protein table:



Note that some of the proteins 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 random matches 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, protein hits can be sorted into four classes. *Which population do we want to retain? To control?*

Note that PeptideShaker provides a score and a confidence for every protein, peptide and peptide to spectrum match (PSM). We will now use the protein scores and confidence to perform protein validation.

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):

>tr|Q8U2V9|Q8U2V9\_PYRFU Putative uncharacterized protein OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=PF0719 PE=4 SV=1

MEDEYVIKDLDQFVELWTSIYNTGGKPDWSHILPYYSENIHFRDSIQEIHGIEEFKKMVERLTKRSKELKFVIK

>tr|Q8U2V9\_REVERSED|Q8U2V9\_PYRFU Putative uncharacterized protein OS=Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1) GN=PF0719 PE=4 SV=1-REVERSED

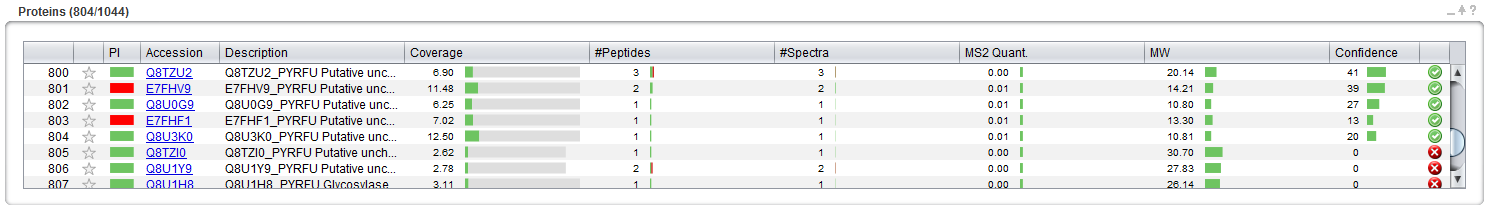
KIVFKLEKSRKTLREVMKKFEEIGHIEQISDRFHINESYYPLIHSWDPKGGTNYISTWLEVFQDLDKIVYEDEM

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[1](#_ENREF_1) (called the decoy database).

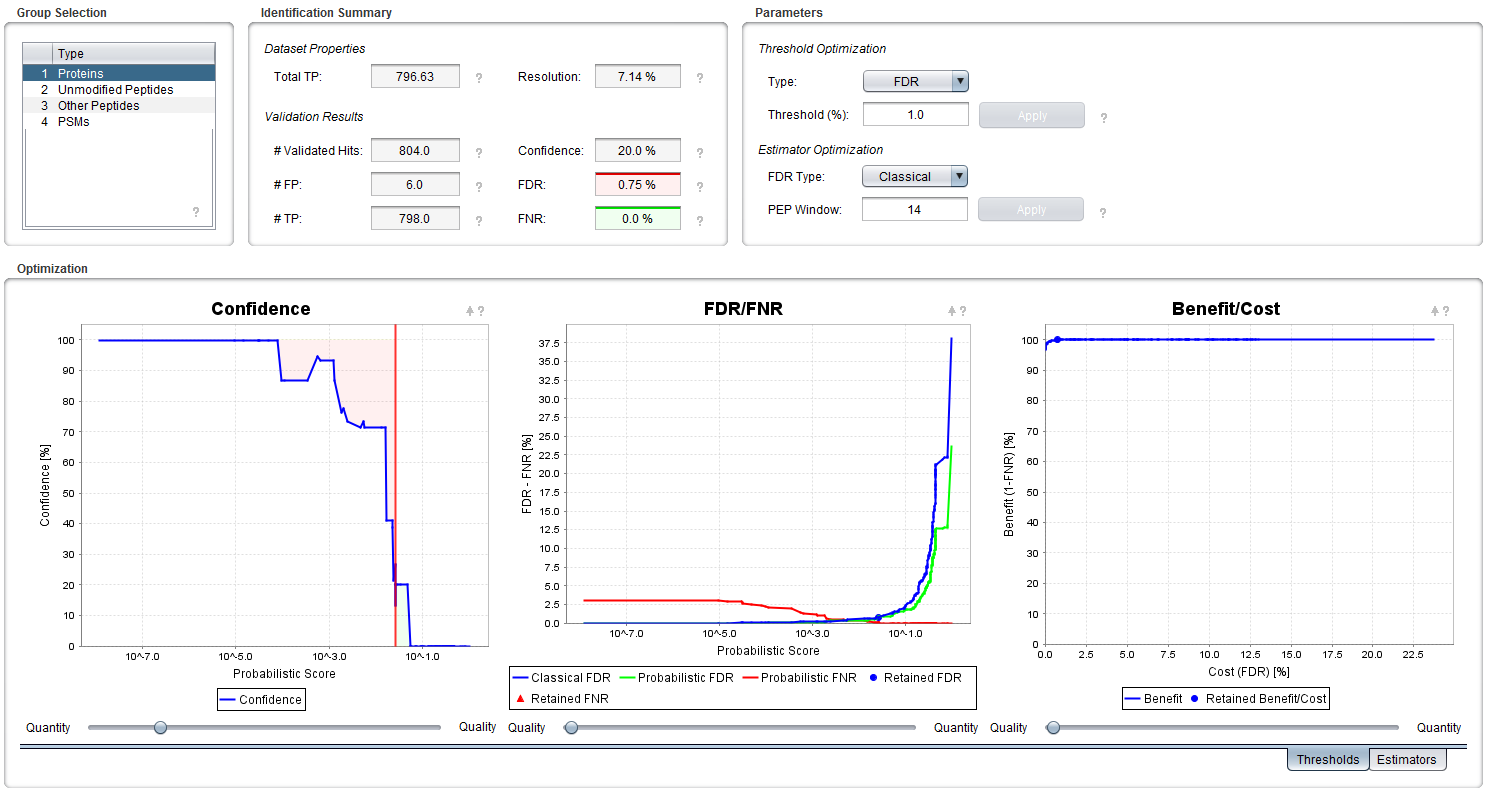
There is one false positive among the target hits. *Do we know which one? Are there other ways to create a decoy database? Which one is the best?*

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 green and red symbols at the far right of each row (see figure below). These indicate whether the corresponding protein, peptide or PSM passed a 1% FDR threshold. *804 proteins were validated here out of 1044, how many false positives do we expect?*



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 below plots are provided in order to visualize the results and control their quality. We will now only focus on the main parameters. Note that question marks are 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 amount of retained false positives; (B) False Negative Rate (FNR) indicating the amount of false negatives:

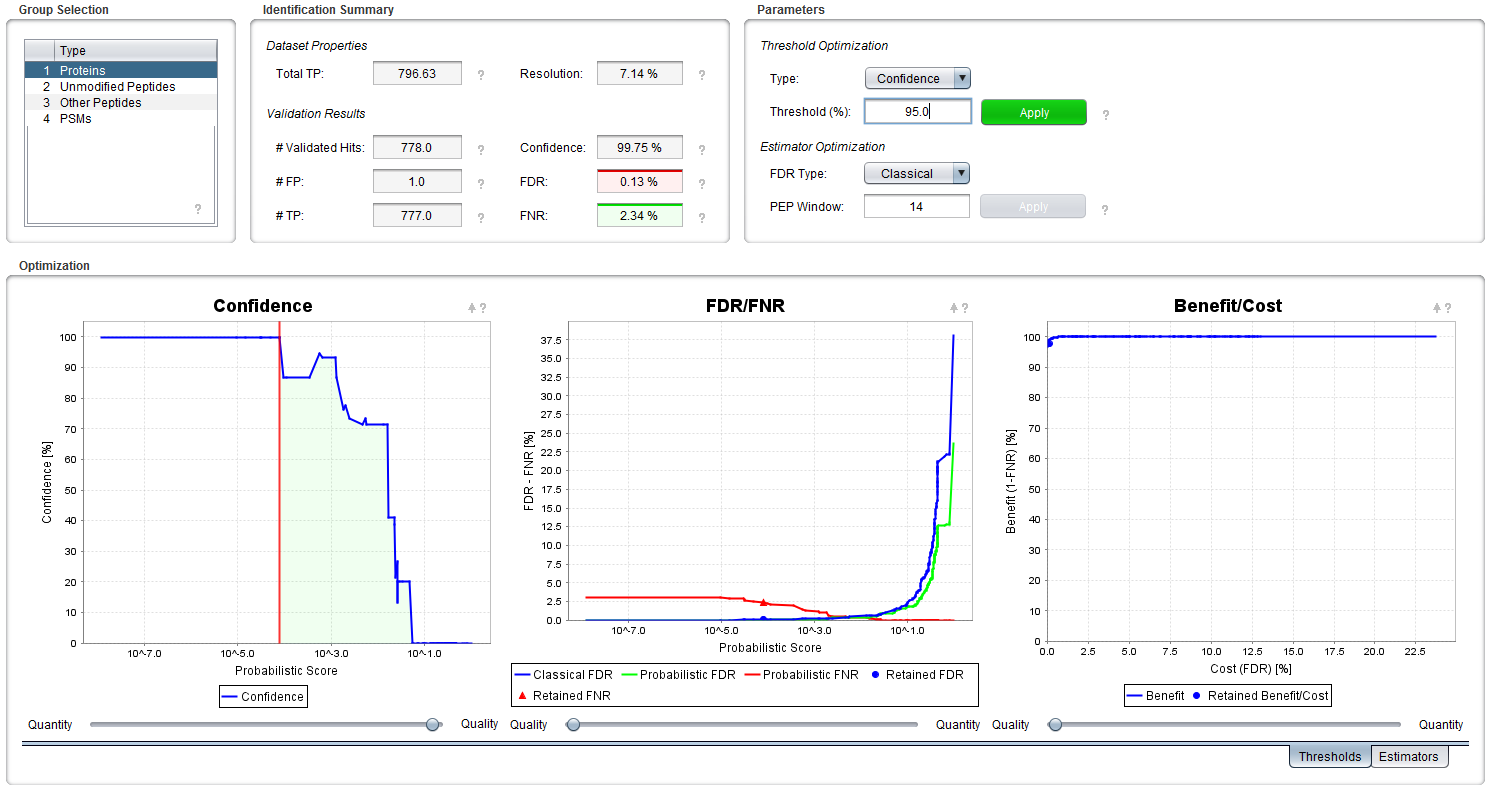
|  |  |  |
| --- | --- | --- |
|  | Number of  validated proteins | Number of  rejected proteins |
| Proteins actually  in the sample | nTP | nFN |
| False identifications | nFP | nTN |

The identification summary indicates that 804 proteins were validated including 6 false positives. PeptideShaker estimates that a maximum of 796.63 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 0.75%. *Why is it not 1%?*

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 amount 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 Benefit/Cost Curve (also called a ROC curve) which allows you to optimize your threshold. Note that the current setting is represented by a point on the curves. As you can see from the confidence plot, our threshold (red line) is set in an area where the confidence is fluctuating between around 20%. *How accurate is the confidence estimation in this case? If we include twenty proteins with such a confidence, how many false positives do we include?*

We are now going to ask PeptideShaker for a higher quality and set a confidence threshold of 95%. All the validated proteins will hence have a minimal confidence of 95%. Select Confidence as the threshold type, type 95 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 left in the plots. *What are the new FDR and FNR values? Are our new settings better than the previous ones?*

The identification summary now indicates that 778 proteins were validated, including 1 false positive. *How accurate is this estimation? The estimated minimal confidence is now 99.75%.* *How accurate is this estimation?*

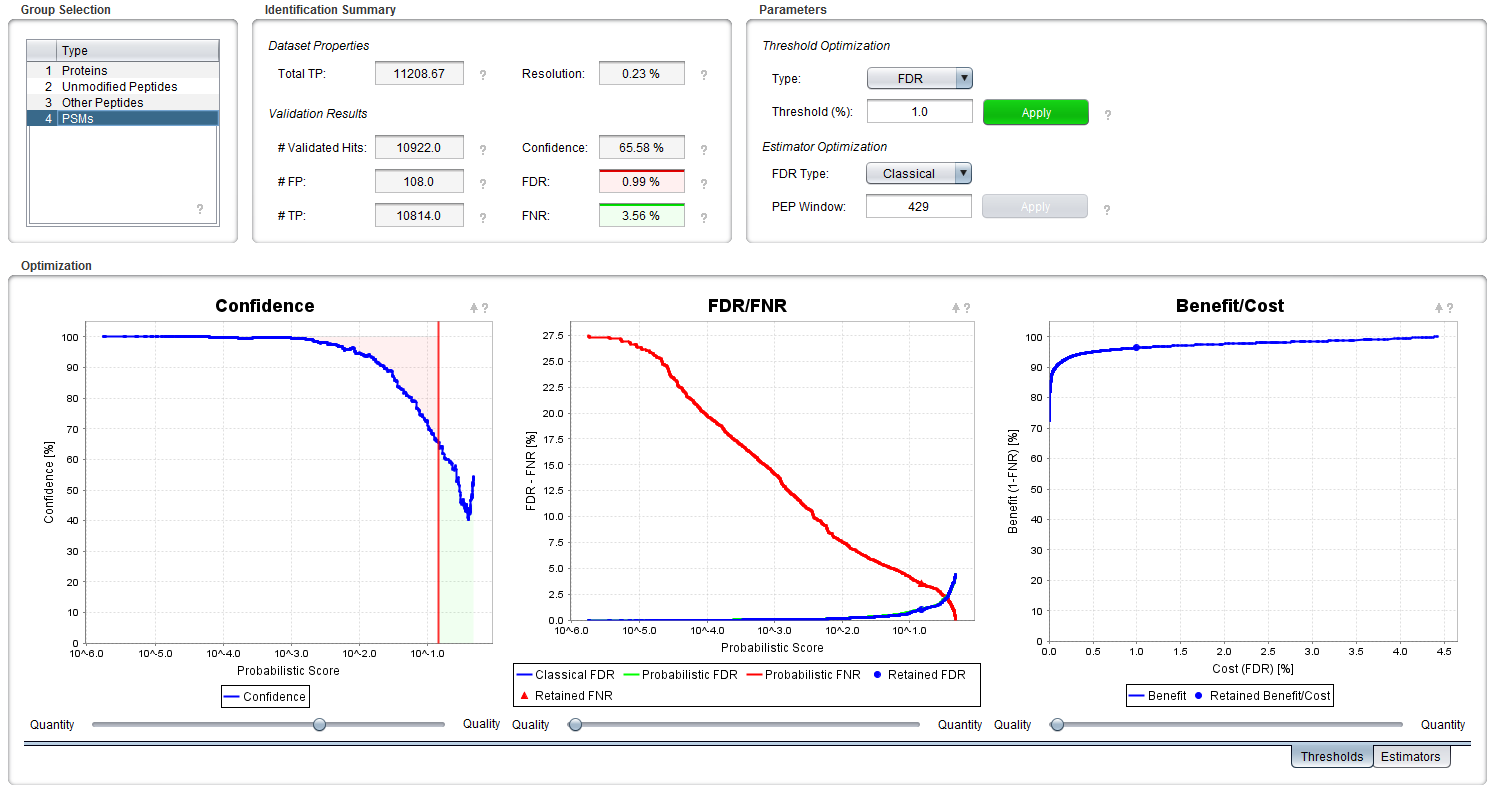
This operation can also be conducted on peptides and PSMs when changing the selected population in the top left section. You will however note that PeptideShaker separates modified and unmodified peptides. *Why are peptides separated into groups?*

In fact, when statistical significance is ensured, PSMs are grouped according to their charges and peptides according to their modification status in order to maximize the amount of identifications[2](#_ENREF_2). If phosphorylated peptides are present in a sufficient amount for instance, it will be possible to conduct the validation process only on this population! Here there were not enough phosphorylated peptides so all modified peptides are grouped together in the ‘Other Peptides’ category.

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!

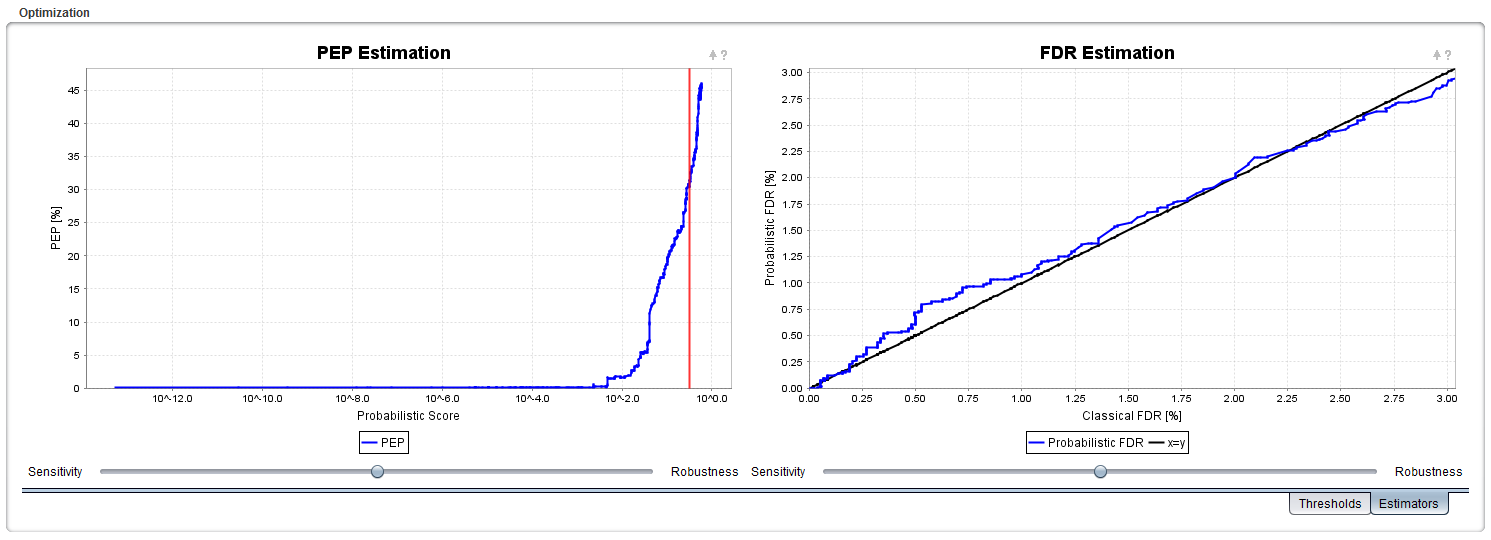
Advanced

In this advanced section, we will inspect the quality of the validation metrics. Select the ‘PSMs’ category. You will notice in the confidence plot that the confidence never reaches 0.



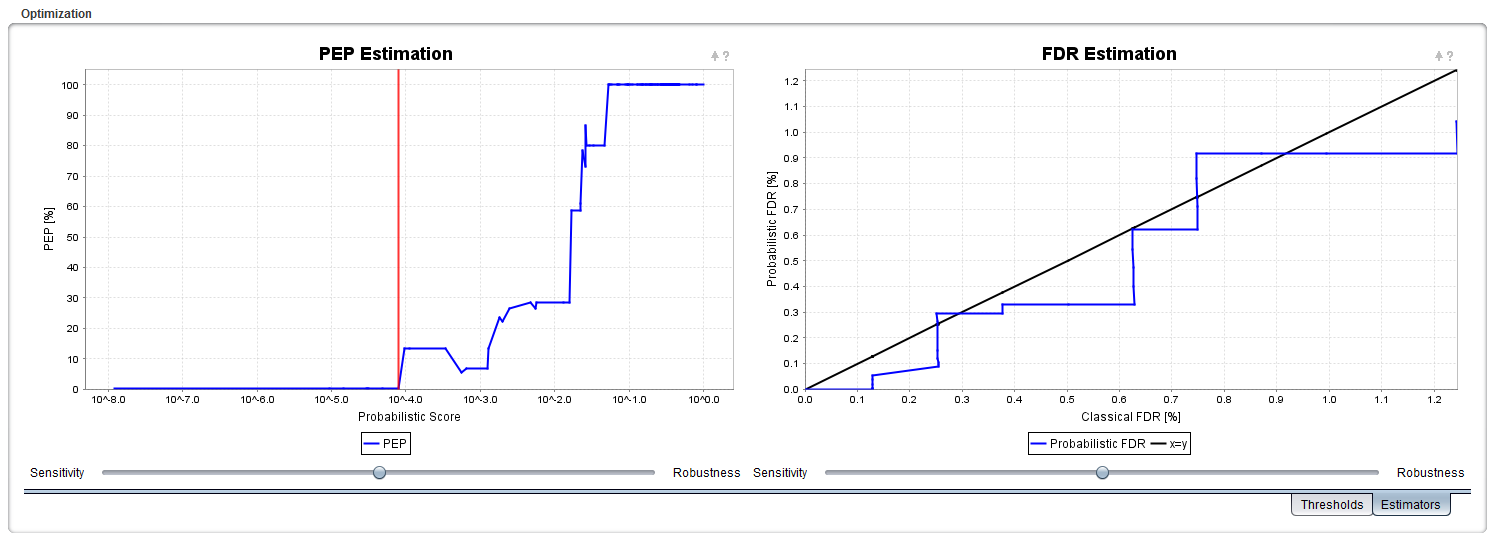
*What will be the effect on identification? How can we avoid this issue?*

Select unmodified peptides 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 Probabililty (PEP). Keen observers will have noticed that the PEP is nothing but 1 - confidence. 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_3)), 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.

Now select the ‘Protein’ category again:



You notice here that the quality is much lower (the blue line deviates from the black line). This can also be noticed from the fluctuations in the PEP Estimation plot. The reason for this is simply that we have fewer proteins than peptides. The lack in statistical confidence is shown by the Resolution (found in the 'Identification Summary' section at the top): 7.14% for proteins, 0.2% for unmodified peptides.

References

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

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

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