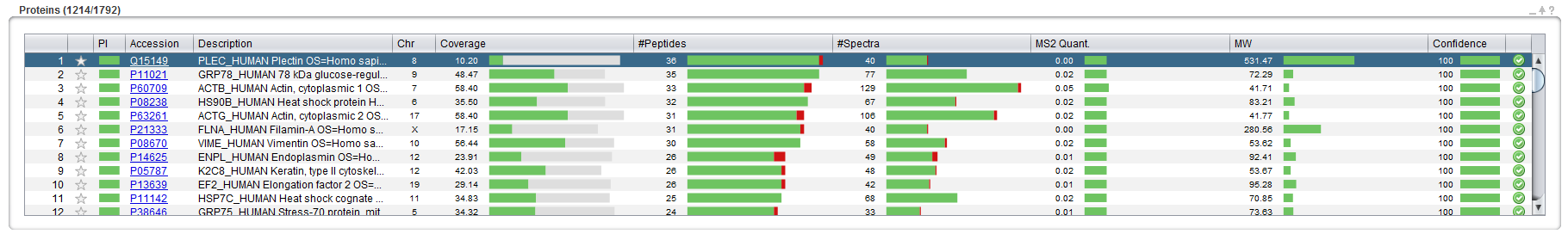
Peptide and Protein Validation

Load the HeLaproject which has been used in the previous chapters 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 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, protein hits can be sorted into four classes. *Which population do we want to retain? To control?*

*In order to maximize our proteome coverage, we will try to maximize the number of true positives while controlling our error rate: the share of false positives.*

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):

>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

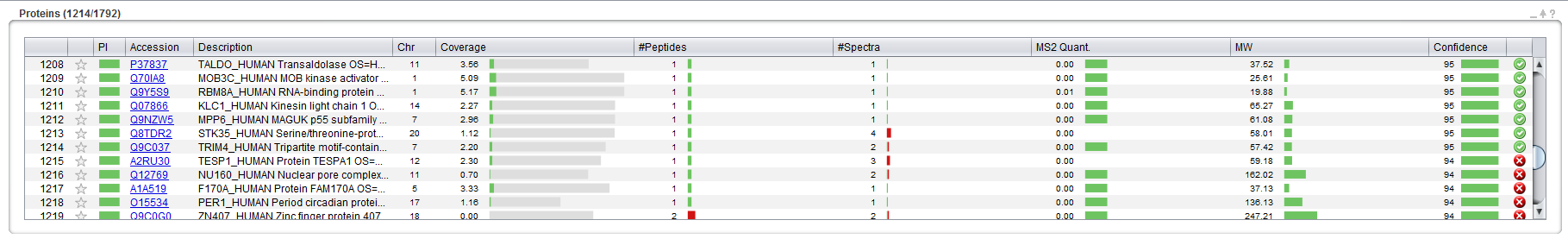
Thus, whenever a mistake is made, it is as likely to happen in the real database (called the target database) as it is in the artificial database (called the decoy database).[2](#_ENREF_2) When a decoy hit is found among five target hits

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?*

*The decoy hits only indicate the propensity for the search engine to introduce random matches at a given score. In no way they indicate which target hit is the wrong one.*

*It is also possible to create decoy databases by randomizing amino acids. This is particularly easy with dbtoolkit.*[*3*](#_ENREF_3) *Both reverse and random decoy sequences were shown to perform equally well.*[*4*](#_ENREF_4)*,* [*5*](#_ENREF_5) *The random approaches present the advantage to allow the creation of different versions.*

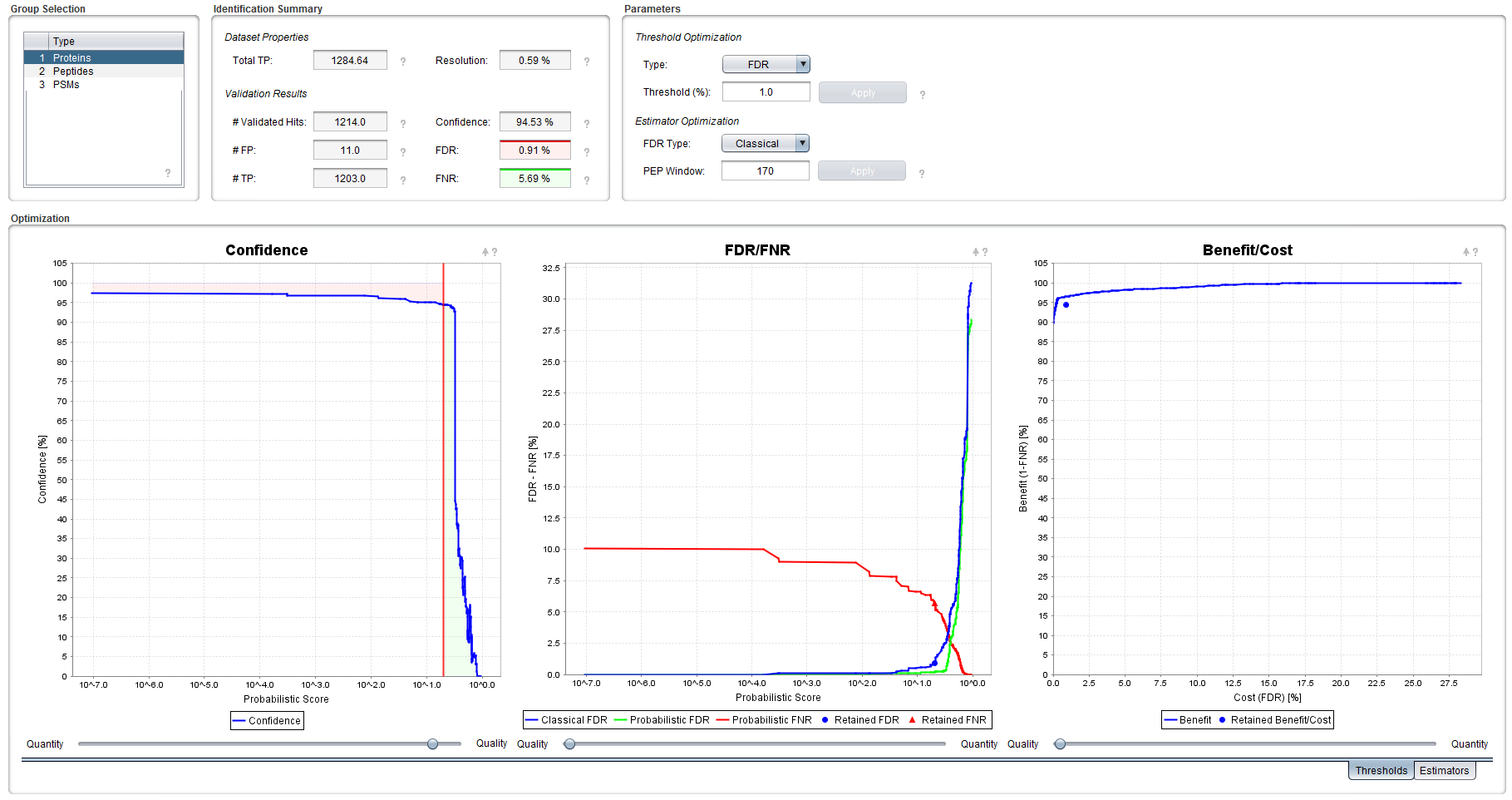
The decoy hits allow us to estimate the number of false positives in a result set. There are two main ways to control the amount of false positives in the validated protein set. First of all, we can set a confidence threshold; typically we would validate protein hits in which we are more than 95% confident. 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.



These indicate whether the corresponding protein, peptide or PSM passed a 1% FDR threshold. *1,214 proteins were validated here out of 1,792, how many false positives do we expect?*

*We expect a maximum of 12 false positives: 1% of 1214.*

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 settings. 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 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 1,214 proteins were validated including 11 false positives. PeptideShaker estimates that a maximum of 1284.64 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%?*

*This value was the best below 1%. Including more proteins would have in all cases implied FDR > 1%. PeptideShaker hence stopped at 0.75% this is called a q-value.*[*6*](#_ENREF_6)

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 Benefit/Cost 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. You might notice that the point is not exactly following 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 95%. *How accurate is the confidence estimation in this case? If we include hundred proteins with such a confidence, how many false positives do we expect?*

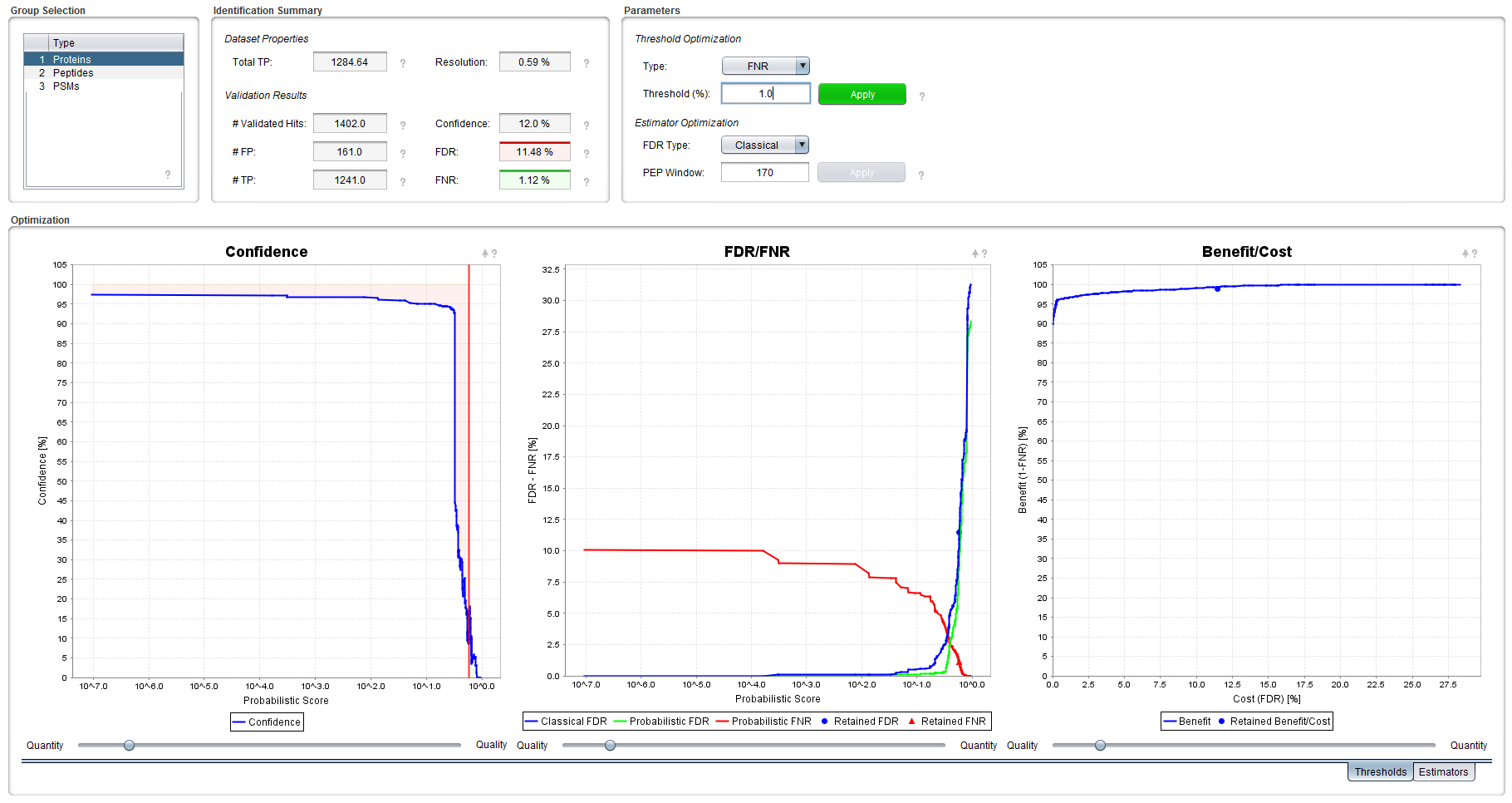
*As one can see on the right of the plot, the confidence can fluctuate at a given score. This shows that our estimation is not an exact estimation. In fact, PeptideShaker tells you that it estimates its resolution to 0.59 percentage points (pp). One can hence expect our confidence estimation to be percentage point accurate.*

*Including hundred hits at 95% confidence, we expect 95 true positives, hence 5 false positives. The complement of the confidence is named Posterior Error Probability (PEP): .*[*6*](#_ENREF_6)

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.

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

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? Are our new settings better than the previous ones?*

*The new estimated FDR value is 11.48%, corresponding to an estimated FNR of 1.12%. We have hence included 150 false positives to rescue 38 true positives. The interest of this quantity-driven threshold is obviously disputable. However, there is no perfect threshold, it is up to the scientist to draw the line based on his experiment.*

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. *What is the minimal peptide confidence at 1% FDR? What is the FDR when thresholding at 95% confidence?*

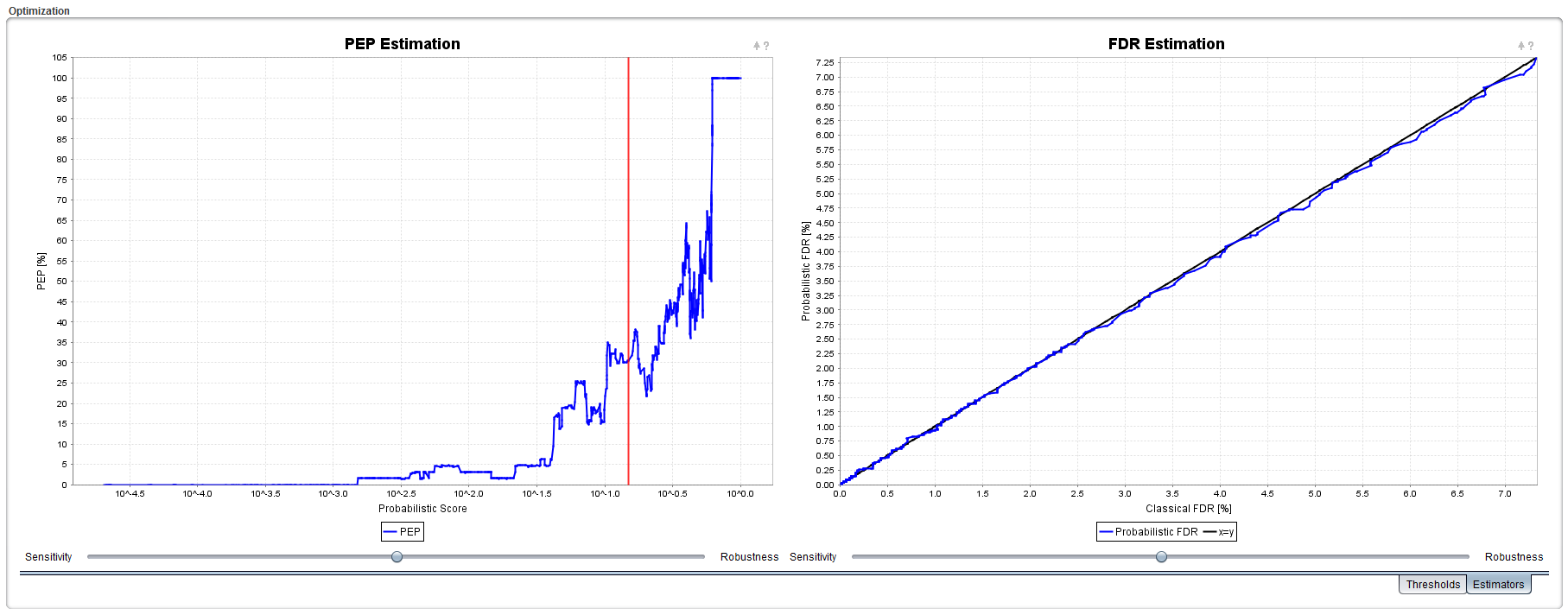
*At 1% FDR, the lowest confidence retained is 63% estimated at an accuracy of approximately 1.5 percentage points. When thresholding at a minimal confidence of 95%, we obtain an estimated FDR of 0.06%.*

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)

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 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[7](#_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?*

*For the proteins, the blue line clearly deviates from the black line. This is simply due to the fact that there are fewer proteins than spectra: the statistical estimation is hence less accurate. This deviation is directly linked to the deviation of the operating point of the ROC curve.*

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. Martens, L., Vandekerckhove, J. & Gevaert, K. DBToolkit: processing protein databases for peptide-centric proteomics. *Bioinformatics* **21**, 3584-3585 (2005).

4. Wang, G., Wu, W.W., Zhang, Z., Masilamani, S. & Shen, R.F. Decoy methods for assessing false positives and false discovery rates in shotgun proteomics. *Analytical chemistry* **81**, 146-159 (2009).

5. Colaert, N., Degroeve, S., Helsens, K. & Martens, L. Analysis of the resolution limitations of peptide identification algorithms. *Journal of proteome research* **10**, 5555-5561 (2011).

6. Kall, L., Storey, J.D. & Noble, W.S. QVALITY: non-parametric estimation of q-values and posterior error probabilities. *Bioinformatics* **25**, 964-966 (2009).

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