The Good, the Bad and the Ugly: when statistics tells you to throw away peptide IDs.

Siegfried Gessulat[1], Tobias Schmidt[2], Michael Graber[1], Florian Seefried[1], Dave Horn[3], Christoph Henrich[4], Bernhard Delanghe[4], Daniel Zolg[2], Mathias Wilhelm[2], Bernhard Kuster[2], Martin Frejno[1]
[1]msAId GmbH, Garching, Germany; [2]Technical University of Munich, Freising, Germany; [3]Thermo Fisher, San Jose, CA; [4]Thermo Fisher Scientific (Bremen) GmbH, Bremen, Germany

ABSTRACT

Purpose: Identify the optimal PSM validation method available.

Methods: Comparison of different PSM validation methods and their performance in separating targets from decoys.

Results: Semi-supervised machine learning using multiple scores can separate targets from decoys better than classical approaches based on a single score but there is room for improvement.

INTRODUCTION

Current instrumentation and data analysis workflows allow for the identification of thousands of proteins per hour and the resulting data are reported as tabular output or directly visualized. Few people still take the time to investigate the underlying spectral data once they were processed. However, it is important to bear in mind what effects a widely accepted false discovery rate of 1% has, what it means for individual peptide-spectrum matches (PSMs) in terms of local FDR and how many low-quality spectra will remain in a dataset. Here, we exemplify such effects and visualize the corresponding spectra to raise awareness for data quality.

MATERIALS AND METHODS

Sample Preparation and Data Acquisition

Example data: 200ng HeLa protein digest (Pierce) was loaded onto Thermo Scientific™ EASY-Spray™ PepMap™ RSLC C18, 25 cm C18 column and separated with a 60 min gradient (8-30 % B [80% ACN in 0.1% FA] in 60 min, 5min to 50 %, another 5 min to 90 %B, 8 min at 90 % B). The eluting peptides were analyzed on the change to: Thermo Scientific™ Orbitrap Exploris™ 480 mass spectrometer. The system was operated in a data dependent mode, selecting as many precursors as possible in 1 second cycle time.

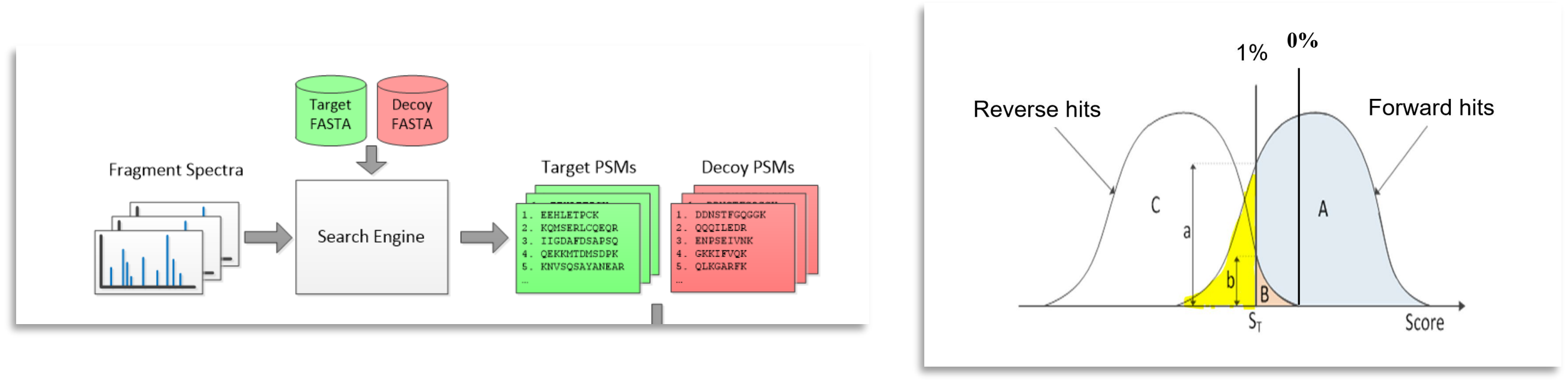
Test Method

For several years, target-decoy approaches have been the standard method to calculate the False Discovery Rate (FDR) in bottom-up proteomics experiments. Machine learning methods such as Percolator are commonly used to separate incorrect from correct matches. FDR cut-offs are then adjusted on PSM, peptide or protein level to allow for a maximum of 1% decoy hits in the resulting dataset.

Data Analysis

Data analysis was performed using a beta version of Thermo Scientific™ Proteome Discoverer™ 2.5 software.

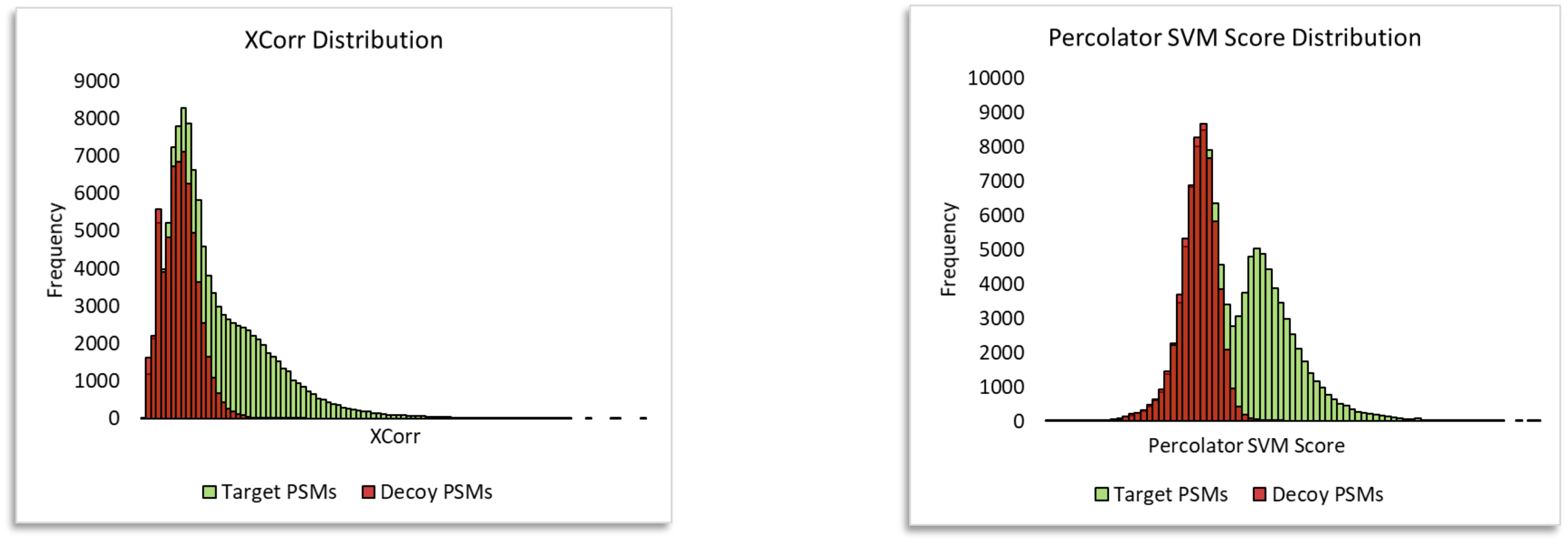
Figure 1. Target Decoy PSM Validator with separate target and decoy searches. The FDR is based on single search engine score.



$$\mathrm{FDR}=\frac{B}{A+B}$$

How the FDR is calculated is explained in Figure 1. The spectra are searched against the target FASTA and against the decoy FASTA (obtained by reversing the protein sequences) resulting in target PSMs and decoy PSMs. The desired FDR can then be obtained by selecting the score threshold in such way that the number decoys above the score threshold divided by the number of targets and decoys above the score threshold. The target PSMs below that threshold are thrown away. An alternative method consist of concatenating the target and decoy FASTA and perform only one search. Each spectra is identified with either a target or decoy PSM.

Figure 2. Score distribution of Target and Decoy PSMs using a Concatenated Target Decoy strategy (left) and using Percolator in concatenated mode (right).



In 2007 a paper was published by L. Käll et al. (1) describing a semi-supervised machine learning approach, Percolator, that improves the separation between reverse and forward hits, outperforming the classical Target Decoy methods (see Figure 2.). A new industry standard was born. Percolator is combining multiple features into one score.

RESULTS

An overview of the identifications on Protein Group, Peptide and PSM level of the different Target Decoy strategies is giving in Figure 3. Percolator clearly outperforms the classical Target Decoy approach. The Concatenated mode performs slightly better than the Separated mode.

Nevertheless all the target PSMs below the threshold are still discarded, all the ugly spectra are ignored.

Figure 3. Comparison at the level of identifications for the example Hela raw file.

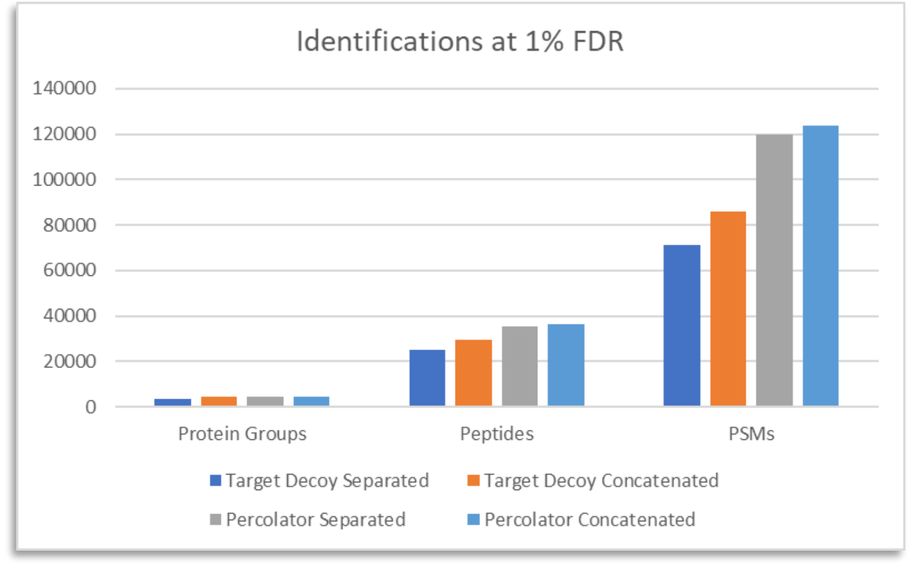
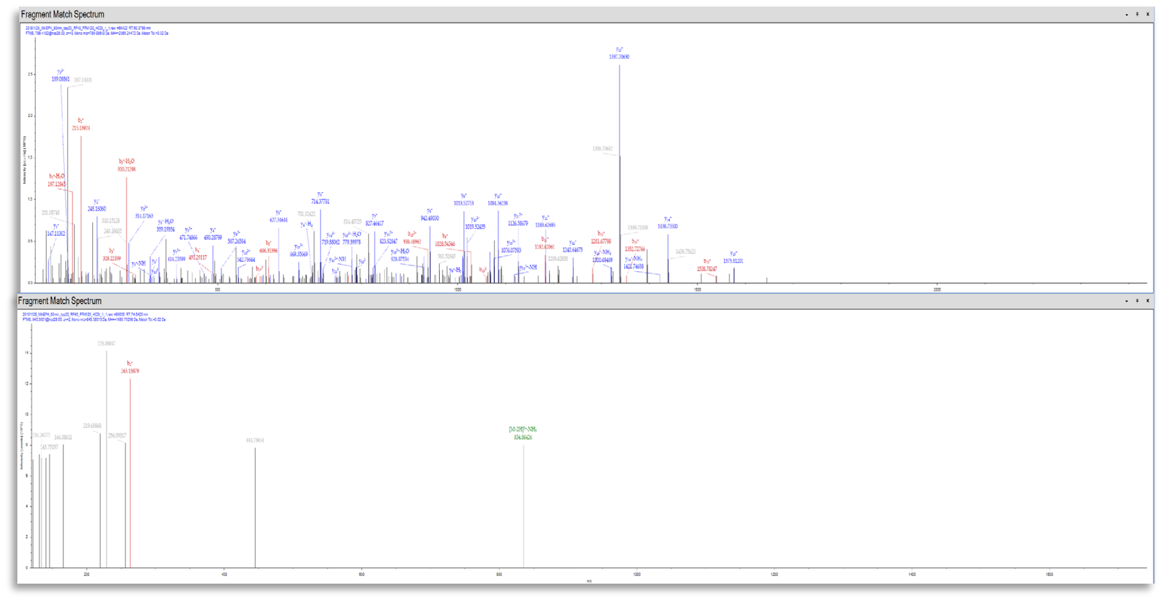


Figure 4. Example good and bad spectrum.



Good and bad spectra are easy to identify and true and false identifications of those are easy to validate statistically (Figure 4.). For ugly spectra, this is still difficult. Figure 5. shows a spectrum that was not identified with SequestHT (neither using Percolator nor Target Decoy), most probably due to the high isolation interference (95%). However, this identification was recovered using intensity and fragment information from a spectral library using MSPepSearch.

Figure 5. Example of an ugly spectrum No. 1

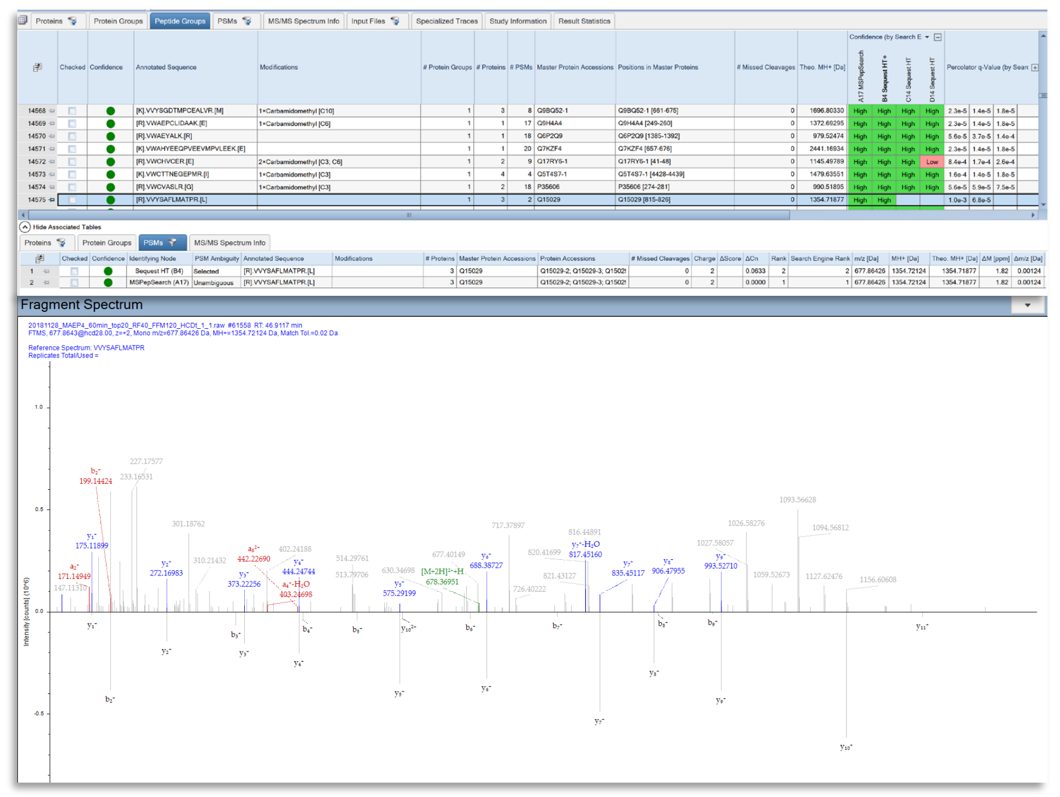
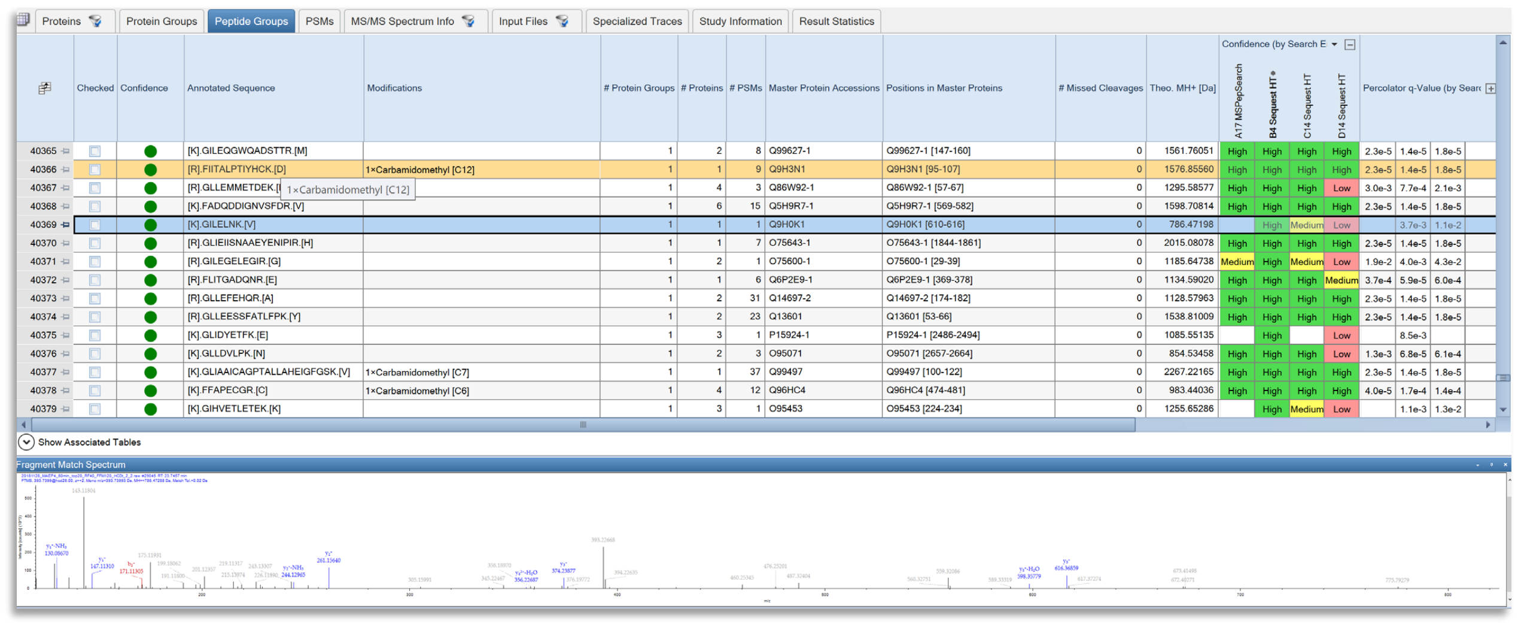


Figure 6. shows another example of an ugly spectrum, this time SequestHT almost identified it (medium confidence) and MSPepSearch didn't. Although almost all (except 1) y-series fragments are present, this PSM is discarded as many more in this dataset.

Figure 6. Example of an ugly spectrum No. 2



CONCLUSIONS

- There are still many good peptide identifications hiding in ugly spectra.
- This is especially problematic for data sets containing many low-abundant, chimeric or very similar spectra (e.g. HLA peptides, proteogenomic data sets).
- The problem gets exacerbated by using big search spaces (e.g. HLA, Metaproteomics).
- More information how this will be solved in Proteome Discoverer 2.5 is presented in poster “Separating the wheat from the chaff: Prediction-assisted rescoring of peptidic fragment ion spectra”.

REFERENCES

1. L. Käll, et al. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods, 2007, 4, 923-925

TRADEMARKS/LICENSING