thermoscientific

Separating the Wheat from the Chaff: Prediction-Assisted Rescoring of Peptidic Fragment Ion Spectra

Siegfried Gessulat[1], Tobias Schmidt[2], Michael Graber[1], Florian Seefried[1], Carmen Paschke[3], Kai Fritzemeier[3], Dave Horn[4], Bernard Delanghe[3], 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 Scientific (Bremen) GmbH, Bremen, Germany; [4]Thermo Fisher, San Jose, CA

ABSTRACT

Purpose: Improve the separation of target and decoy identifications in proteomics data sets in order to boost the confidence in search results.

Methods: Several datasets were analysed using a beta version of Thermo Scientific™ Proteome Discoverer™ 2.5 software with SequestHT and the new Prosit-derived (1) Rescoring node by MSAID.

Results: Deep-learning-based prediction of fragment ion intensities enables the addition of intensity-based scores to identification workflows with SequestHT, which increase the confidence in search results.

INTRODUCTION

Most database search algorithms compare experimental spectra of peptides with theoretical fragment masses to calculate similarity measures while largely disregarding the intensity dimension. Automatically matching peptides to spectra in this way will yield false identifications of low-quality spectra or misrepresent their confidence. The standard method to control for erroneous matching of such spectra is the target-decoy approach that estimates the False Discovery Rate (FDR) in bottom-up proteomics experiments. Machine learning methods such as Percolator (2) are commonly used to separate incorrect from correct matches, but their performance heavily depends on the calculated scores. Here, we show how intensity-based scores circumvent common issues and challenges in peptide identification.

MATERIALS AND METHODS

Samples

The following datasets were analysed: Thermo Scientific™ Pierce™ HeLa protein digest, HLA (3) and metaproteomic datasets (4) were obtained from the ProteomeXchange consortium.

Databases

The HeLa protein digest was searched against SwissProt including isoforms with tryptic digestion, the HLA data set was searched against SwissProt including isoforms with unspecific digestion and the Metaproteomic data set was searched against the Integrated Reference Catalogue of the human gut microbiome (IGC, (5)) with tryptic digestion. Peptide length was restricted from 7-30 amino acids and charge states were restricted from 2-6.

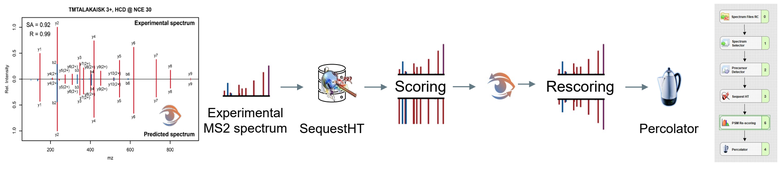
Test Method

The new rescoring node is accurately predicting fragments and intensities at the specific Collision Energy (CE) for each PSM that SequestHT is proposing. This spectrum is then compared to the measured spectrum and additional scores, including spectral angel, are calculated and included in the features set for Percolator (Figure 1.).

Data Analysis

Data analysis was performed using a beta version of Proteome Discoverer 2.5 software and the new Rescoring node by MSAID.

Figure 1. Intensity based rescoring workflow.



RESULTS

Figure 2. displays the results for â single Hela raw file. 10% increase in identifications on PSM, 8% on peptide and 4% on protein level is observed using the new rescoring node, compared to the same workflow without rescoring.

Looking at results from larger and more complex datasets, e.g. a Metaproteomics study (8 raw files (4)) the improvements are even bigger: 13% increase in identifications on PSM, 11% on peptide and 10% on protein level (Figure 3).

Figure 2. Comparison at the level of identifications for the Hela sample.

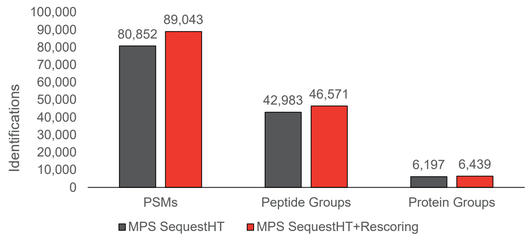


Figure 3. Comparison at the level of identifications for the Metaproteomics – Human stool samples.

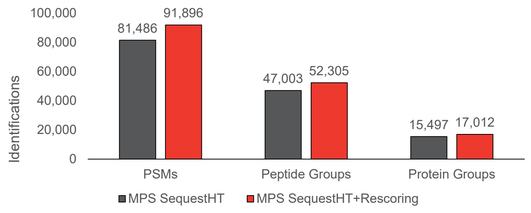


Figure 4. Comparison at the level of identifications for the Immunopeptidomics data set.

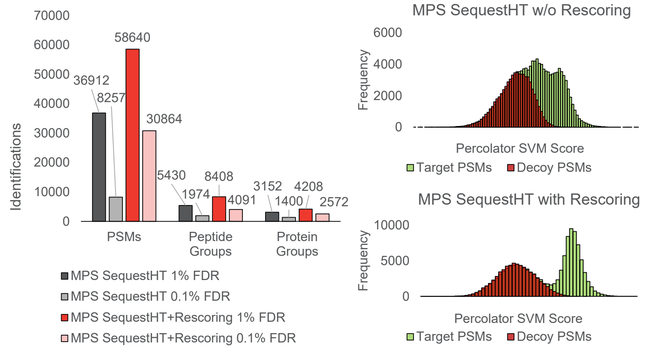
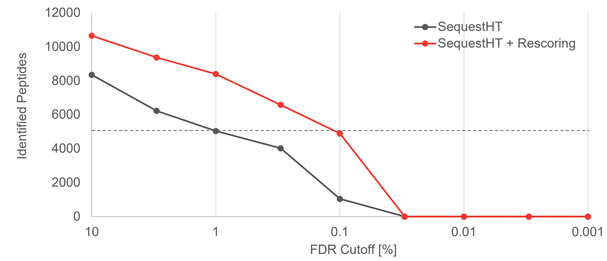


Figure 5. Comparison of the FDR rates with or without rescoring for the example Immunopeptidomics data set.



The power of rescoring is fully exploited when used for the most difficult searches (no enzyme searches), when the search space is huge and when peptides with very similar properties are expected. Figure 4. shows the result of an HLA Class 1 peptide data set (Patient derived melanoma cell line, 8 raw files (3)). An increase of 59% identifications on PSM, 55% on peptide and 34% on protein level is obtained using the rescoring node. The rescoring greatly improves the separation of targets and decoys PSM scores.

Figure 5. shows another advantage of rescoring: 10-times lower FDR cutoffs can be used while maintaining a similar number of identifications on peptide level. Increasing the confidence in results is especially crucial in the field of immunopeptidomics.

CONCLUSIONS

- Comparing experimental and predicted spectra unlocks the intensity dimension in peptide fragment ion spectra, increases confidence in search results and improves target-decoy separation.
- Rescoring
  - boosts PSM, peptide and protein identification rates in classical tryptic datasets
  - improves new applications like immunopeptidomics, where 55% more peptides are identified
  - enables more stringent FDR cutoffs, promoting a paradigm shift towards increased confidence in results

REFERENCES

1. S. Gessulat, T. Schmidt, et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nat Methods, 2019, 16(6), 509-518
2. L. Käll, et al. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods, 2007, 4, 923-925
3. C. Chong, et al. Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. Nat Commun., 2020, 11(1):1293
4. J. Rechenberger, et al. Challenges in Clinical Metaproteomics Highlighted by the Analysis of Acute Leukemia Patients with Gut Colonization by Multidrug-Resistant Enterobacteriaceae. Proteomes, 2019, 7(1)
5. J. Li etc al. An integrated catalog of reference genes in the human gut microbiome. Nature Biotech. 2014, 32, 834-814

TRADEMARKS/LICENSING

PO65834-EN 0422S

MSAID

ThermoFisher SCIENTIFIC