Improved Deep Learning-based Rescoring for Immunopeptide Identification

Mick Greer[1] ; David M. Horn[2]; Daniel Hermanson[2]; Martin Frejno[3]; Daniel P Zolg[3]; Tobias Schmidt[3]; Bernard Delanghe[4]
[1]Thermo Fisher Scientific, Austin, TX; [2]Thermo Fisher Scientific, San Jose, CA; [3]MSAID GmbH, Garching b.München, Germany; [4]Thermo Fisher Scientific (Bremen) GmbH, Bremen, Germany

ABSTRACT

Purpose: Evaluate the utility of deep-learning based spectral prediction for immunopeptidomics

Methods: HLA enriched peptides searched with and without INFERYS™ spectral re-scoring

Results: Improvements in deep learning-based prediction increase the performance of immunopeptidomics workflows.

INTRODUCTION

Identification of peptides in LC-/MS/MS-based immunopeptidomic data is challenging due to the large search space, their nonstandard fragmentation, and the presence of peptides with similar physicochemical properties. Most search engines were optimized to identify tryptic peptides, but only some of the HLA alleles produce peptides with basic amino acids at or near the C-terminus. As a result, such HLA peptides can exhibit low fragment coverage, making them difficult to differentiate from random database matches. Recently, deep learning prediction-based rescoring has been shown to increase the number of immunopeptide IDs by increasing confidence in correctly matched peptides. Here, we describe the application of an updated INFERYS™ deep learning-based rescoring workflow to improve the performance of Sequest HT for HLA peptide analysis.

MATERIALS AND METHODS

Sample Preparation

A full description of sample preparation can be found in reference 1. Briefly; data analyzed in this study were from patient derived primary tumors. HLA-peptide complexes were immunoprecipitated and enriched prior to LC/MS analysis.

LC/MS Data Acquisition

Peptides were eluted with a linear gradient. MS/MS were acquired in a data-dependent acquisition on a Thermo Scientific™ Orbitrap Exploris™ 480 equipped with a FAIMS Pro interface. For HLA-I peptides, up to five precursors of charge 1+ between 800 and 1700 m/z or 10 precursors of charge 2 to 4+ were subjected to MS/MS acquisition. Precursors were isolated with a 1.1 m/z window, and 120 ms maximum injection time, fragmented at 30% higher energy collisional dissociation (HCD), and acquired at 15,000 resolution.

The FAIMS Pro interface was operated at a spray voltage of 1900 V set to standard resolution. FAIMS compensation voltages (CVs) were set to -50 and -70 with a cycle time of 1.5 s per FAIMS experiment. MS2 fill time was set to 100 ms. A complete description of data acquisition can be found in *Klaeger et al. MCP (20)2021*

Figure 1. Thermo Scientific Orbitrap Exploris 480 with FAIMS Pro Interface

Data Analysis

HLA class I datasets were downloaded from PRIDE (PXD027165). All data were processed in Thermo Scientific™ Proteome Discoverer™ software using Sequest HT alone, Sequest HT with the original INFERYS Rescoring algorithm, and Sequest HT with the INFERYS 2.0-based rescoring algorithm released in Proteome Discoverer 3.0 software (**Figure 2**).

INFERYS Rescoring (**Figure 3**) assigns a spectral angle by comparing a predicted fragment ion spectra and its intensities to the collected spectra. The greater the similarity between predicted and observed ion intensity the higher the spectral angle score[2]. The spectral angle is measured from 0 to 1.These spectral angle scores can then be used by Percolator (as XCorr and many other PSM metrics are) to estimate the FDR threshold.

Figure 2. Immunopeptide processing workflow including INFERYS™ Rescoring node

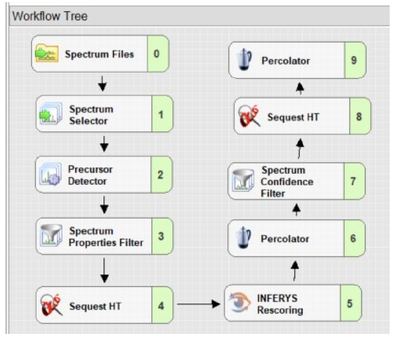
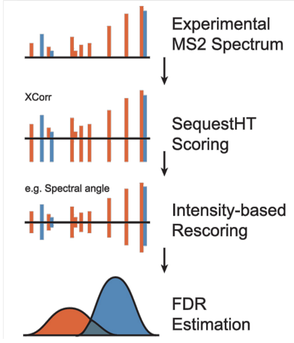


Figure 3. Each PSM receives Spectral Angle from INFERYS™ Rescoring node.

Adapted from Ref. 2



Additional analysis was performed on the resulting peptides at 1% FDR. Peptides were further filtered within the Proteome Discoverer software to remove tryptic like peptides (i.e. peptides containing C-terminal R and K residues were filtered out) and focus on the known HLA-type I peptides sequence lengths(i.e. 7-14 residues.)

From this filtered data set all 9-mer peptides were submitted to the NetMHCpan 4.1 server (https://services.healthtech.dtu.dk) for binding predictions to known MHC molecules.

RESULTS

Increase in HLA type I peptides

The data was searched using the template workflows available in the Proteome Discoverer software. A two-stage SequestHT search with INFERYS rescoring was selected for this analysis. In this strategy the first round of SequestHT search generates compatible PSMs for rescoring (e.g unmodified and carbamidomethylated) and the second round of search includes variable modifications for common sample preparation artifacts. The processing workflows (**Figure 2**) were identical apart from inclusion or exclusion of the INFERYS Rescoring node following the first SequestHT search node. The three processing workflows compared in this study will be referred to as follows:

1.) SEQUEST HT

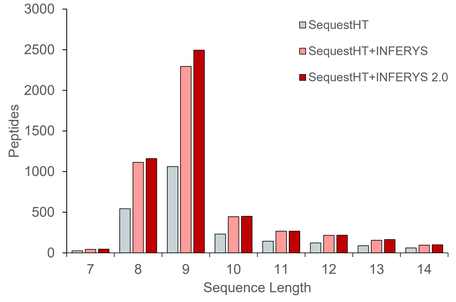
2.) SEQUEST HT + INFERYS

3.) SEQUEST HT + INFERYS 2.0.

Note “INFERYS” refers to the INFERYS Rescoring node found in version 2.5 of the Proteome Discoverer software and “INERYS 2.0” refers to the INFERYS Rescoring node found in version 3.0 of the Proteome Discoverer software

Results from these searches were further filtered to remove tryptic peptides and peptides outside the expected sequence length of HLA class I type peptides. **Figure 4** shows the increase of identified immunopeptides. A significant increase is achieved by inclusion of INFERYS, while a more modest increase is noted when comparing INFERYS to “INFERYS 2.0”

Figure 4. Increased MHC class I type peptides at 1%FDR

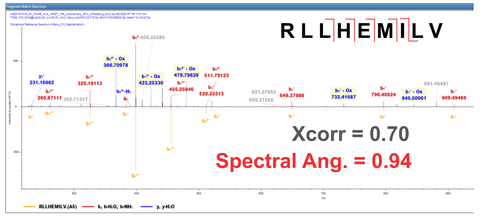


This result suggest that traditional search engines alone such as SequestHT, which were developed for analysis of tryptic peptides and do not consider fragment ion intensity, are insufficient for the analysis of immunopeptides which are endogenously expressed and do not undergo traditional enzymatic digestion into peptides.

Figure 5 shows an example of a strong HLA binding peptide (predicted) which was only found in the results from the INFERYS Rescoring workflows.

This spectra received a low cross correlation score (Xcorr) from SequestHT of 0.7 one of the metrics used by Percolator when estimating the 1% FDR threshold estimated. Xcorr is a correlation that considers the number of matched fragments, and in **Figure 5** most of the y-ion series are absent resulting in a low score. The low Xcorr score may be a contributing factor for this PSM not being in the results of the Sequest HT workflow, in the absence of INFERYS Rescoring. However, this spectra received a high spectral angle score (0.94) from the INFERYS Rescoring node due to the excellent agreement between the predicted fragment ion intensities and the observed ion intensities. When INFERYS is present in a workflow the spectral angle score is utilized by Percolator, as well as other metrics from the database search engine. Including the spectral angle score in the Percolator FDR estimation improved this PSM’s rank compared to the workflow without INFERYS rescoring.

Figure 5. Representative PSM resurfaced by INFERYS Rescoring



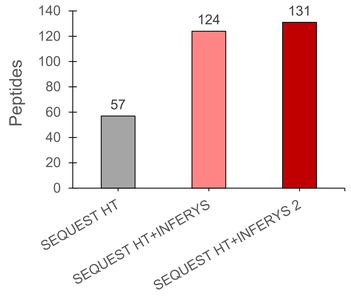
The peptide sequence shown in **Figure 5** is representative of MHC class I peptide. It has a sequence length of 9 residues, contains leucine and valine at positions 2 and 9 and does not contain a basic residue at its C-terminus. A similar trend was observed for the global set of identified 9-mer HLA-I peptides.

Increase in predicted strong binding peptides

In order to determine if these peptides resurfaced by INFERYS Rescoring are of biological interest they were submitted to the NetMHCpan 4.1 server to predict their binding potential. Only the 9-mer peptides were submitted since they were the most abundant species. The HLA B13-02 allele had the largest number of predicted strong binders from the list of 9-mer peptides identified in this study.

Figure 6 shows the increased number of identified 9-mer HLA-1 peptides which were predicted as strong binder by

Figure 6. Increase in NetMHCPan 4.1 predicted strong binders (HLA-B13:02)



CONCLUSIONS

Deep learning-based prediction and rescoring increase the performance of immunopeptidomic workflows of MHC class I peptides

- Database search engines trained to identify enzymatic peptides are not the best choice for immunopeptide identification
- Inclusion of fragment intensity metrics in FDR estimation increases immunopeptide identification

REFERENCES

1. Klaeger, S, Apffel, A , Clauser, K, et al. Optimized Liquid and Gas Phase Fractionation Increases HLA-Peptidome Coverage for Primary Cell and Tissue Samples.Mol Cell Proteomics (2021) 20 100133
2. Zolg, DP, Gessulat, S, Paschke, C, et al. INFERYS rescoring: Boosting peptide identifications and scoring confidence of database search results. Rapid Commun Mass Spectrom. 2021;e9128. https://doi.org/10.1002/rcm.9128

TRADEMARKS/LICENSING

PO66155-EN0422S