CCS-centric protein identification using IM Score

HUPO 2021

Robin Park[1], Patrick Garrett[3], Sven Brehmer[2], Titus Jung[3], Vijayaraja Gnanasambandan[1], Marc-Antoine Beauvais[1], Sebastian Wehner[2], Lennard Pfennig[2], Hyunsoo Kim[3], Christopher Adams[1], Dennis Trede[2], John R. Yates, III[3], Rohan Thakur[1]

[1]Bruker Daltonics Inc., USA
[2]Bruker Daltonik GmbH, Bremen, Germany
[3]Scripps Research, San Diego, USA

Introduction

The database search algorithms to identify proteins have advanced since SEQUEST was first introduced in 1994. The search engines calculate scores by comparing experimental spectra from MS/MS data against in-silico spectra derived from protein databases. Even though a lot of efforts have been invested in improving protein identification, still a large number of spectra cannot be identified. One of the obstacles is spectrum ambiguity. For example, more than one peptide candidate from the protein database can often have the same precursor mass and similar fragment ion patterns generating competing PSM search scores.

A recent breakthrough of 4D proteomics by the timsTOF Pro adds an extra dimension of separation, providing a solution to accurately clarifying a significant number of spectra with CCS (collisional cross-section) values. To address the spectrum ambiguity, we have developed a deep learning model to predict retention time, fragment ion intensities using bidirectional long-short term memory (LSTM) recurrent neural networks, and CCS values using PepMDN (peptide mixture density network) model. IP2GPU search engine calculates an ion mobility-based search score called IMScores, p-values derived the CCS distribution based on peptide candidates. Next, we added IMScores as an additional metric for DTASelect to classify decoy and target peptides and calculate FDR using quadratic discriminant analysis. The unique CCS prediction model and IMScore we developed can be used to improve both protein identification and quantification.

Methods

A nanoElute (Bruker Daltonics) high pressure nanoflow system was connected to a high-resolution TIMS-QTOF (timsTOF Pro, Bruker Daltonics). Whole protein extracts of human cervical cancer cells (HeLa) were used for the experiment. Peptides were separated on the analytical column using a 120 min gradient. LC-MS/MS data were acquired using the PASEF method with a total cycle time of 1.1 s, including 1 TIMS MS scan and 10 PASEF MS/MS scans. The data analysis was performaed using IP2-GPU search engine.

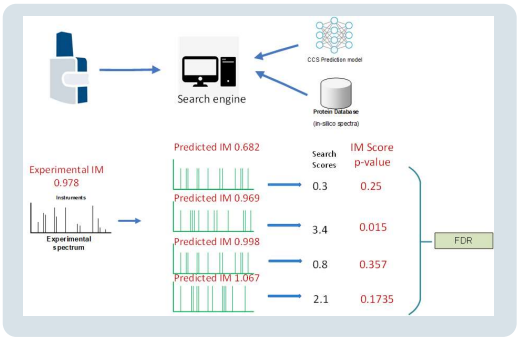

Figure 1: Overview of CCS-centric database search
As the top peptide candidates for a given spectrum can similar search scores, we applied a CCS-centric search algorithm to minimize the peptide candidate ambiguity problem. When the search engine compares experimental and theoretical spectra, it calculates IM Score by comparing ion mobility values. The IM Score is a p-value showing how likely a peptide candidate can be true.

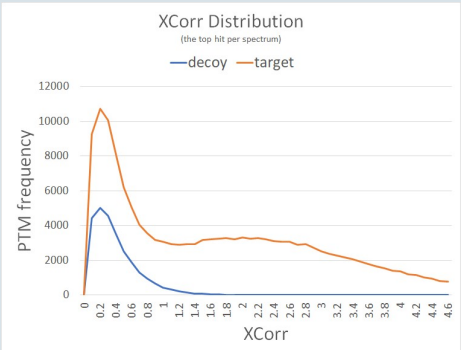

Figure 2: XCorr distribution
The search results shows typical distribution of target and decoy PSMs. Both target and decoy peptides show well separation in high search score region

Results

Search score like XCorr is an important metric to classify targe and decoy PSMs. Overall, we can see similar pattern with IMScore displaying good separation in high score region. We expect adding IMScore to the FDR calculation algorithm as an additional parameter can imporve the search results. The IMScore also can be used to clarify when PTM peptides have ambiguous localization scores (modificaiton site is unclear).

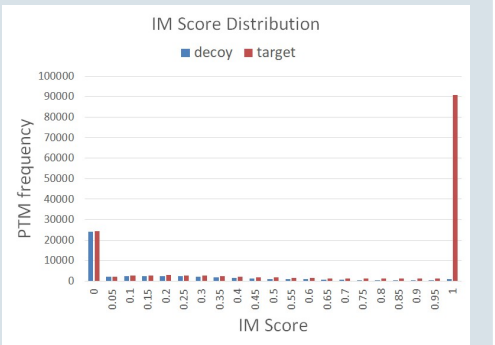

Figure 3: IMScore distribution
IMScore is a statistical score comparing the difference between experimental and predicted ion mobility values. High IMScore means the peptide candidate can be likely true. The IMScore distribution shows target and decopy peptide candidates are well separated as IMScores go higher having similar pattern with XCorr distribution. It proves that IMScore can be an important metric in the data analysis

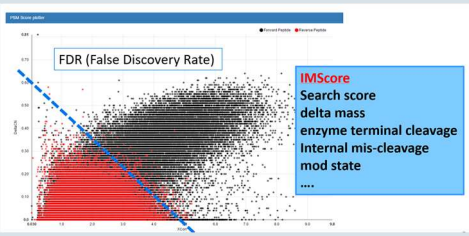

Figure 4: FDR Calculation with IM Score.
FDR calculation uses search score, deltaCN, delta mass distribution, enzyme cleavage stat, etc. IMScore can be added to the list of metrics to boost the number of identification under the same FDR threshold

Conclusions

- This study proves IMScore based on ion mobility values can play important role in the data anlaysis
- We are continuously integrating IMScore to the FDR calculation to boost confidence in the search results
- We will apply IMScore to a smart acquisition from PaSER real-time search platform, PTM analysis.

timsTOF Pro