The Benefit of Peptide CCS Value Prediction and Experimental Determination

ASMS 2019, WP 492

Sebastian Wehner[1]; Favio Salinas[2]; Stuart Pengelley[1]; Heiko Neuweger[1]; Heiner Koch[1]; Anjali Alving[3]; Na Parra[3]; Greig Michael[3]; Juergen Cox[2]; Detlev Suckau[1]

[1] Bruker Corporation, Bremen, Germany
[2] Max Planck Institute of Biochemistry, Martinsried, Germany
[3] Bruker Corporation, Billerica, MA, USA

Overview

- Peptide spectra with low MS/MS information content are in need of methods to enable precise and confident identification.
- These comprise singly charged, non-tryptic or lower abundant peptides, such as immunopeptides (MHC) or host cell protein (HCP) derived peptides.
- Modern ion mobility mass spectrometers can reproducibly determine **collision cross sections (CCS)** of peptides with an accuracy in the 0.2 % range, which are specific for a given peptide sequence.
- We developed a **CCS value prediction algorithm** based on the plain sequence string of the peptides in question, which was evaluated against a very feature-rich proteomic HeLa dataset comprising > 80,000 identified peptides.

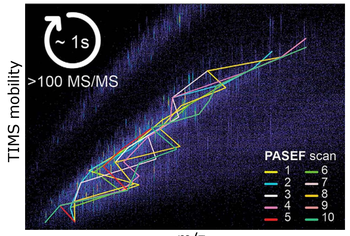

Fig. 1 PASEF Experimental data was generated using a standard 1.1 sec PASEF acquisition cycle. In this method, parallel accumulation and serial fragmentation results in increased sensitivity due to mobility focusing of the ions which are sequentially fragmented at > 100 Hz. (Figure adapted from Meier et al., MCP, 2018).

Methods

Experimental data were generated using tryptic digests of human cancer cell lysates (HeLa), which were separated by nano-flow LC fitted with a 25 cm pulled emitter C18 column (IonOpticks, Australia), applying a linear gradient of 5-30% buffer B (100% ACN and 0.1% FA) at a flow rate of 400 nL/min. MS, MS/MS and CCS data were obtained using the PASEF method on a timsTOF Pro IMS-QTOF mass spectrometer (Bruker, **Fig.1**).

Data analysis was performed using MaxQuant (v1.6.4.0) to obtain a data set consisting of >80.000 unmodified peptides over five charge states.

For CCS prediction two deep learning models were applied: a bidirectional long-short-term-memory (LSTM, Jürgen Cox, MPI) and a branched convolutional neural network (CNN, Bruker inhouse). Both models use the high-level API of Keras (v2.2.4) with Tensorflow (v1.8.0) as backend.

Input data preparation for the models relies only on character-wise embedding of the textual amino acid sequences, for training as well as prediction.

Results

The Parallel Acquisition Serial Fragmentation (PASEF) method implemented on the trapped ion mobility spectrometry time of flight mass spectrometer (timsTOF Pro) described previously (Meier et al., JPR, 2015) separates ions according to CCS to enable the generation of non-chimeric MS/MS spectra for co-eluting peptides. The recorded CCS values are reproducible (average absolute deviation of 0.2% over 4 runs, Meier et al., MCP, 2018), sequence specific and allow differentiation of peptides with the same or similar precursor m/z.

The analysis focused on the measured HeLa doubly-charged ions which typically had CCS values in the range 300-550 Å² . At any given m/z value there is a wide distribution of CCS values, exemplified in **Fig.2** which shows the recorded CCS values for peptides within a 0.07 m/z window. The broad distribution of CCS values for similar sized peptides provides the rational for using prediction and measurement to distinguish isobaric peptides even in the case of failed MS/MS identifications.

The amino acid sequences of 80% of the more than 67.000 unique, identified HeLa peptides (z=2) were used as input data for the training of the deep learning models. The remaining 20% were used for validation, resulting in a mean error of -0.43% (-2.63 Å²)/0.18% (1.12 Å²) and a standard deviation of 1.29%/1.43% for the CNN and LSTM models respectively (**Fig.3**).

The Pearson correlation between the experimental and predicted CCS values results in R^2= 0.9652 for the CNN and R^2=0.95 for the LSTM network..

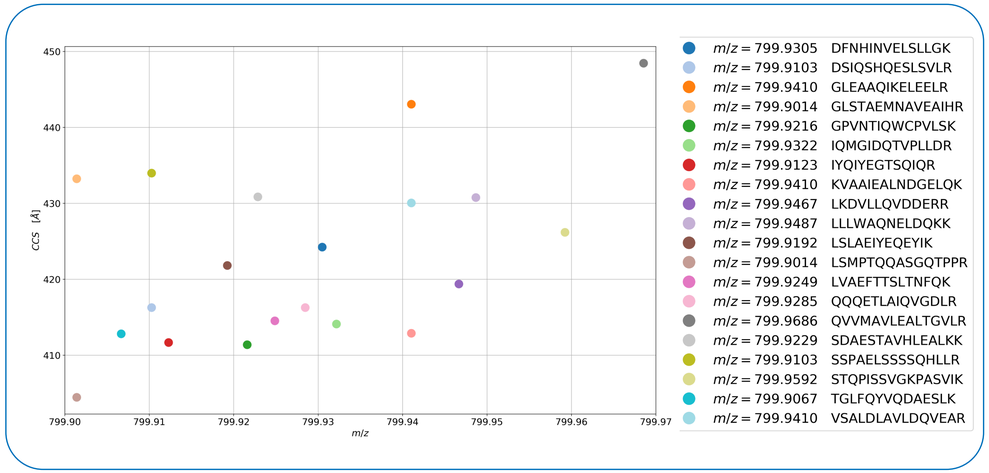

Fig. 2 Measured CCS values for twenty near-isobaric doubly-charged HeLa peptides.

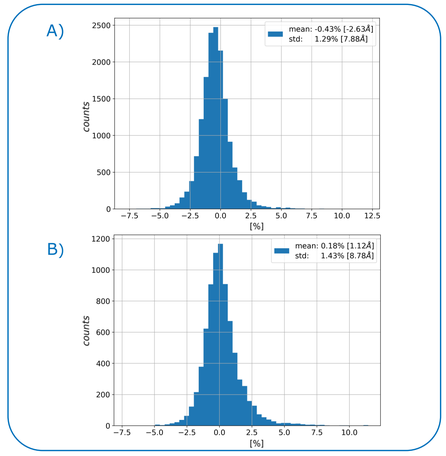

Fig. 3 Distribution of validation prediction accuracy of CCS values for HeLa doubly charged peptides using A) CNN and B) LSTM deep learning models.

Outlook and Future Developments

Possible applications of CCS value prediction of peptides and their measurements are multifold. For instance, the dynamic range of Host Cell Protein (HCP) analysis in biopharmaceutical development can possibly be extended beyond the sensitivity for MS/MS based peptide identification. Preliminary calculations also indicate that proteolytic peptides derived from the complementarity determining regions (CDRs) in monoclonal antibodies can provide different CCS values based on isobaric Leu vs. Ile exchanges. This may accelerate the *de novo* sequencing of antibodies in the future as Leu and Ile cannot be distinguished by low energy CID methods.

Conclusions

- timsTOF Pro with PASEF provides reproducible CCS values (< 0.2 % RMS) and non-chimeric MS/MS identification.
- Near isobaric peptides can in most cases be distinguished by CCS values alone.
- Identification is feasible by high accuracy of mass (< 5 ppm) and CCS value determination alone without MS/MS
- A sequence text based prediction algorithm for CCS values provided prediction errors < 2%, indicating its future use for peptide identification based on accurate mass and mobility data alone.
- This approach could be applied to the identification of low abundant, non tryptic or single charged peptides

BioPharma