HUPO 2023 – PP05-611

Precise and accurate real-time *de novo* sequencing of timsTOF data with the Novor algorithm on the Bruker ProteoScape™ platform

Rui Zhang[1]; Qixin Liu[1]; Mingjie Xie[1]; Dennis Trede[2]; Tharan Srikumar[3]; Jonathan Krieger[3]; Bin Ma[1]; George Rosenberger[4]

[1]Rapid Novor Inc., Kitchener, ON; [2]Bruker Daltonics GmbH & Co. KG, Bremen, Germany; [3]Bruker Ltd., Milton, ON; [4]Bruker Switzerland AG, Faellanden, Switzerland

Introduction

Bruker ProteoScape™ (BPS) was developed to provide state-of-the-art, real-time database searching for the timsTOF instrument family. Since its initial conception, BPS has been transforming into a comprehensive proteomics data analysis platform that can integrate third-party tools while utilizing the concept of data streaming to realize fully customizable real-time processing workflows including on-the fly decision making based on the data generated. However, database searching is only the preferred solution when canonical proteins are being investigated. To expand the capabilities of the BPS platform for immunopeptidomic, metaproteomic, and other applications, we developed and integrated a newly timsTOF optimized *de novo* sequencing engine from Rapid Novor Inc., called BPS Novor.

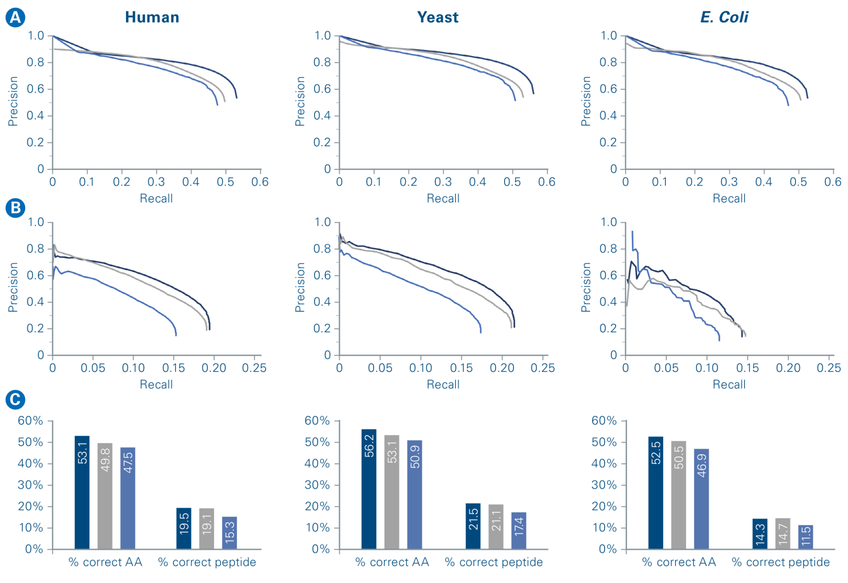

Fig. 1: Amino acid (A) and peptide (B) precision recall graphs for a trypsin digested mix of Human, Yeast and E.coli sample. The percent of correct amino acids and peptides assigned by each algorithm is also shown (C) , where all results are segmented by their species as well.

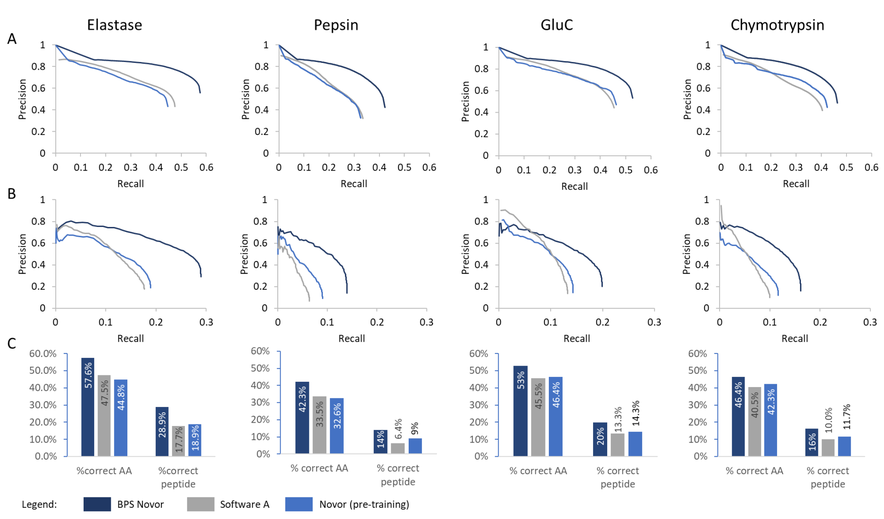

Fig. 2: Amino acid (A) and peptide (B) precision recall graphs for Human lysate digested with one of 4 enzymes, Elastase, Pepsin, GluC or Chymotrypsin. The percent of correct amino acids and peptides assigned by each algorithm is also shown (C).

Results

To validate the performance of BPS Novor, we analyzed a mixed species run (Prianichnikov *et al.*, 2020, PXD014777) that was not utilized in the prior training. This sample consisted of a mix of trypsin digested Human (65%), Yeast (15%) and E. coli (20%), evaluating the effects species specificity might have on BPS Novor performance.

Figure 1A depicts the amino acid precision-recall curves, Figure 1B shows the peptide precision-recall curves and Figure 1C shows the percentage of correct amino acid or peptide assignments by the three de novo algorithms BPS Novor, Software A and Novor (pre-training). In all cases, the dataset was split by species.

BPS Novor showed on average an increase of 5% in correct amino acid identifications vs. software A and 11% vs. Novor algorithm pre-trained. As expected, there were no observable differences between spectra originating from different species. Training Novor on non-tryptic digests allowed it to learn a generalized model, particularly suited for sequencing of non-enzymatically digested peptides (see Figure 2).

The timsTOF platform is capable of >150Hz scan speeds. To realize on-the fly real-time de novo sequencing, the algorithm needs to be capable of processing speed greater than the scan speed. We first evaluated the processing speed of Novor across 8 datasets, where each dataset consisted of >137,000 MS/MS spectra (Figure 3). Average processing time ranged from 86-199 seconds. This translated to an average processing speed of 1338±226 spectra/second. We utilized the BPS Novor module to produce de novo peptide sequences from real time acquired MS/MS spectra.

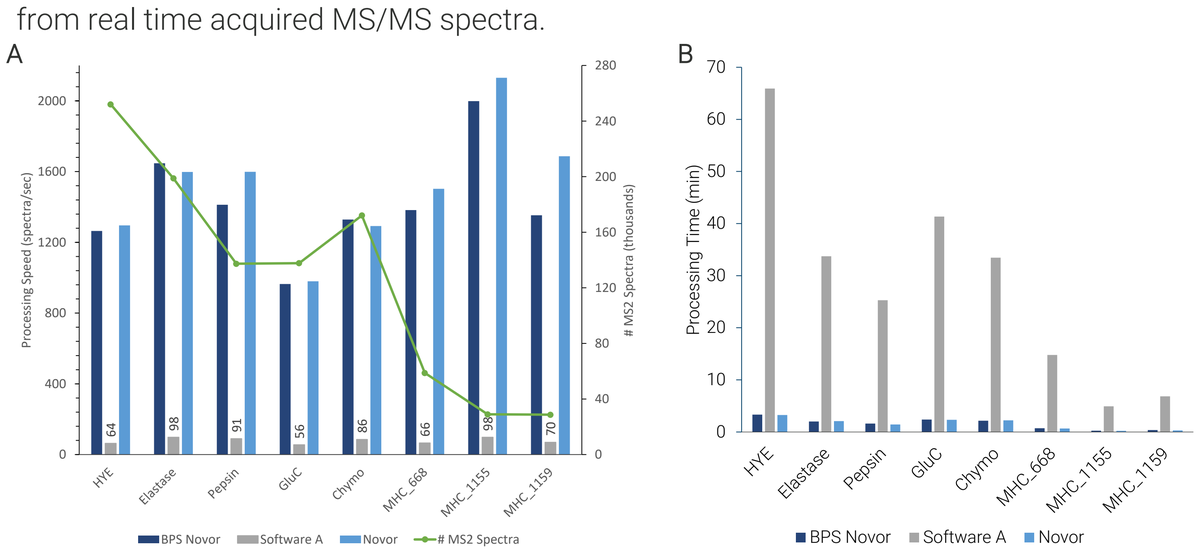

Fig. 3: Processing speed (A) and processing time (B) for 8 datasets that were benchmarked with BPS Novor, Software A and classical Novor. The datasets varied in number of spectra from >240,000 to 38,000.

Methods

BPS Novor was trained on a variety of timsTOF acquired data, where ground truth was taken from ProLuCID-GPU (Xu *et al.*, 2015) database search results filtered to 1% PSM FDR with DTASelect (Tabb *et al.*, 2002). The data included experiments with fixed collision energy measurements of deeply fractionated, GluC, Pepsin, Elastase, Chymotrypsin and Trypsin digested K562 lysates. Collectively, >1,780,000 PSMs were part of the training dataset utilized to optimize BPS Novor's decision tree-based scoring functions (Ma, 2015). Training BPS Novor on non-tryptic digests allowed learning of a generalized model, particularly suited for sequencing of non-enzymatically digested peptides.

We utilized offline functionality to compare BPS Novor against other *de novo* tools across multiple datasets as well as Novor (version 3.0; pre-timsTOF-training). MGF files were utilized to remove the confounding effects of any pre-processing and to allow direct comparison between algorithms based on scan number matching. Amino acid level and peptide level precision and recall for each dataset were computed as in (Ma, 2015).

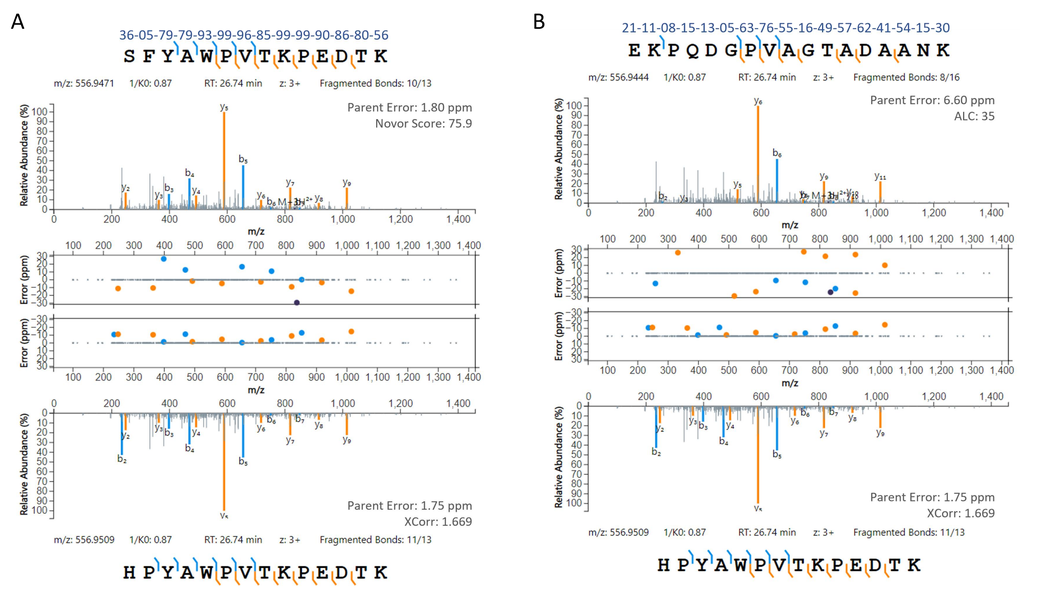

Fig. 4: Two mirror plots depicting the same MS2 spectrum, and the sequence provided by BPS Novor (A upper) and Software A (B upper) vs. the ground truth as identified by ProLuCID (A & B lower)

Small differences observed in summarized amino acid-level algorithm accuracy can have a more profound impact on the sequencing performance for individual spectra.

One such difference is highlighted in Figure 4. Here, an example of BPS Novor's (Figure 2A) and Software A's (Figure 4B) assignments vs. ProLuCID's for the same MS2 spectrum is shown. While both algorithms failed to completely sequence the complex spectra, BPS Novor correctly identified 12/14 amino acids, whereas Software A only correctly identified 3/14 amino acids.

To demonstrate the performance improvements for immunopeptidomic applications, we compared the sequencing results of BPS Novor, Novor (pre-training) and Software A against the ProLuCID-GPU ground truth using a published MHC class-I dataset acquired on the timsTOF platform (Feola *et al.*., 2021) (Figure 5).

The results suggest even greater performance improvements of BPS Novor for challenging immunopeptidomic applications, where typically a lower fraction of peptides can be confidently identified or sequenced.

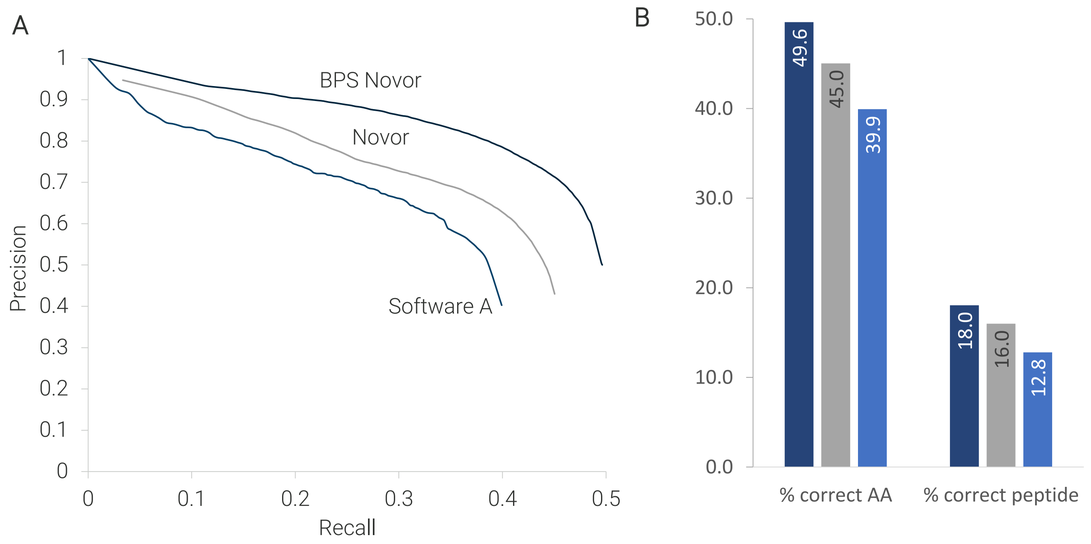

Fig. 5: Reanalysis of MHC class-I data (Feola *et al.*, 2021) shows that BPS Novor can provide accuracy greater than other solutions.

Conclusion

- A fast, accurate and precise peptide de novo sequencing algorithm has been integrated into BPS, providing Run & Done capabilities to additional 4D-Proteomics applications.
- BPS Novor does not show a noticeable bias for digestion specificity or species and is 20x faster than competing products.
- Combined with PASEF technology on the timsTOF platform, BPS Novor provides enhanced sensitivity for real-time de novo sequencing for a variety of applications including immunopeptidomics.

Technology