# Identifying More Peptides at a Lower False Discovery Rate with PEAKS DB Software

Bioinformatics Solutions Inc.

University of Waterloo

Jing Zhang[1], Baozhen Shan[1], Lei Xin[1], Bin Ma[2]
[1] Bioinformatics Solutions Inc, Waterloo, ON
[2] David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON

## Background

In mass spectrometry based proteomics, researchers frequently face the dilemma between keeping more identified peptides and maintaining a lower false discovery rate (FDR). Such situation can only be improved with new analytical software that uses more accurate scoring functions to better separate the true and false identifications. In this abstract we present a new software tool, PEAKS DB, for identifying significantly more peptides with lower FDR than other software commonly in use.

## Method

The algorithm uses an innovative way to combine the de novo sequencing and the database search results. De novo sequencing is traditionally used only when there is no protein database. However, our study showed a surprising fact that today's de novo sequencing software (such as PEAKS) could usually compute a correct long sequence tags for peptides that can be confidently identified with database search. Thus, the similarity between the de novo sequence and the database peptide is used as an important feature in PEAKS DB's scoring function (See Figure 1).


Figure 1. A significant match between de novo sequencing result and database search result indicates the database search result is likely correct.

PEAKS DB additionally outputs a list of "de novo only" peptides, which are found by de novo sequencing with high confidence but do not have any significant match in the protein database. These are potentially novel or modified peptides in the sample that deserve further examination. Figure 2 shows the workflow used by PEAKS DB.

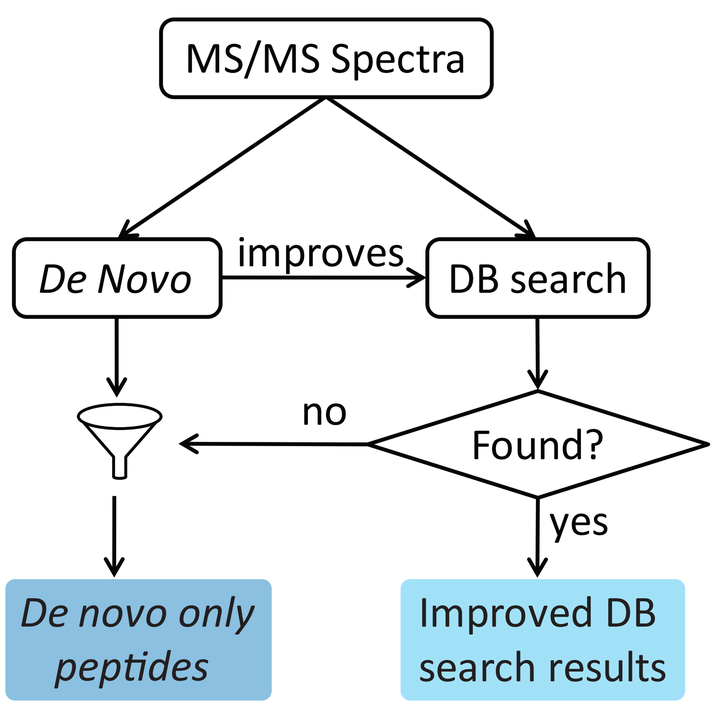


Figure 2. The workflow of PEAKS DB. De novo sequencing is utilized to not only improve the database search, but also report the potentially novel peptides in the sample.

## Results

The performance of PEAKS DB was studied in a CID and an ETD MS/MS datasets, respectively. The CID dataset was from the trypsin digest of Pseudomonas aeruginosa and downloaded from http://www.marcottelab.org/MSdata/Data_12/DATA/20090115_SMPA14_2.RAW.gz. The ETD dataset was from obtained from the Lys-C digest of a yeast lysate and used as the standard dataset by ABRF/iPRG in their 2011 study of the peptide identification software from ETD MS/MS. Both were collected with a Thermo Orbitrap instrument.

PEAKS DB (in PEAKS Studio 5.3) was compared with Mascot 2.3 on both datasets. The same protein databases and search parameters were used. An enhanced target-decoy method (called decoy-fusion) was used to measure the FDR of both methods. At 1% FDR, the peptide-spectrum matches (PSMs) identified by the two engines are shown in Figure 3.

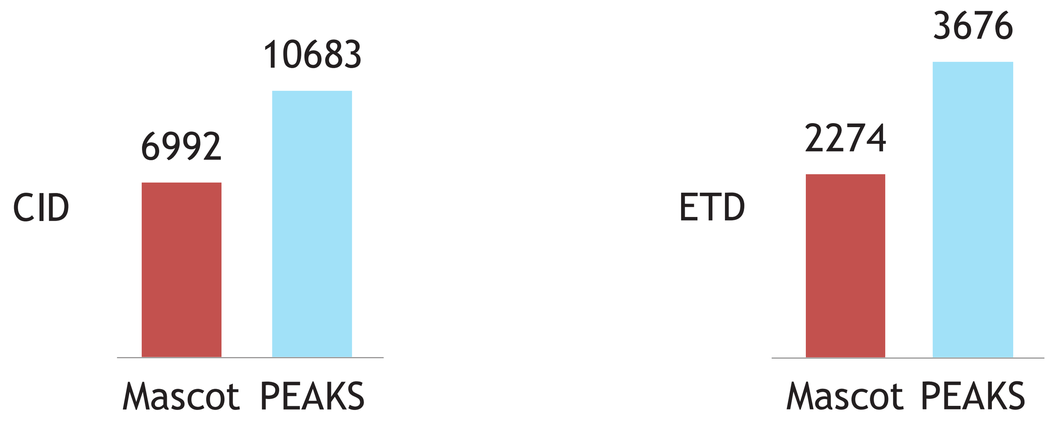


Figure 3. The number of peptide-spectrum matches (PSMs) identified by PEAKS DB and Mascot at 1% false discovery rate (FDR). The numbers here includes only the database peptides but not the "de novo" only peptides.

Not only PEAKS DB reported more database peptides, it additionally reported many "de novo only" peptides from the MS/MS spectra that do not find significant matches in the protein database. Figure 4 shows the number of significant "de novo only" peptides found by PEAKS DB.

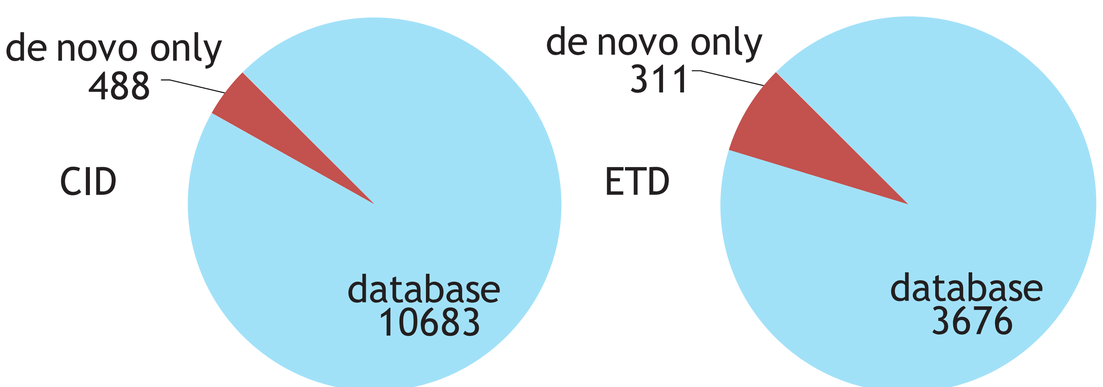


Figure 4. The number of significant "de novo only" peptides, in relative to the number of database peptides, found by PEAKS DB on the two datasets.

## Conclusion

By combining de novo sequencing and database search, PEAKS DB software significantly improves peptide identification performance on both CID and ETD MS/MS data.