De Novo Sequencing vs Database Searching

Bioinformatics Solutions Inc.

Jing Zhang[1], Bin Ma[2]
[1] Bioinformatics Solutions Inc, Waterloo, ON
[2] University of Waterloo, Waterloo, ON

Novel Aspect

De novo sequencing performs surprisingly well and de novo sequence tag search outperforms database search with un-interpreted spectra.

Introduction

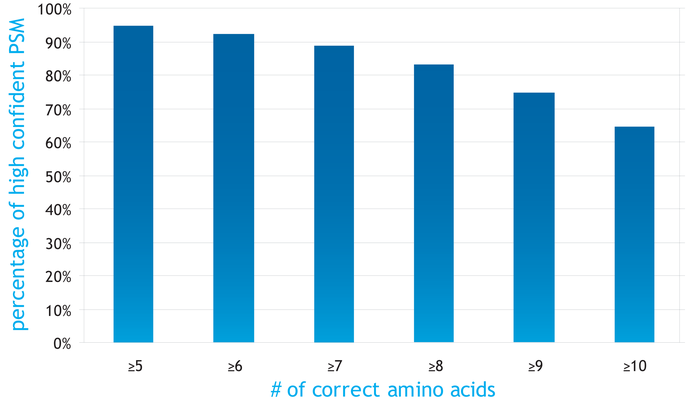
De novo sequencing and database search with un-interpreted spectra are widely known as the two different approaches for peptide identification from MS/MS. De novo sequencing is the only choice for novel peptide identification. However, for peptides in a sequence database, researchers would immediately assume that the database search approach provides better performance. However, in this study we show the opposite with today's standard software, PEAKS de novo sequencing and Mascot database search, respectively. We first demonstrate that de novo sequencing performs remarkably well in terms of determining a significantly long sequence tag. Then we show that an approximate de novo sequence tag search outperforms the conventional database search approach even when the target peptides are in a known sequence database.

Methods

First, the latest versions of the standard de novo sequencing and database software, PEAKS 5.3 and Mascot 2.3, were used to conduct de novo sequencing and database search, respectively. The de novo sequencing results are evaluated in terms of the number of correctly identified amino acids in each reported peptide.

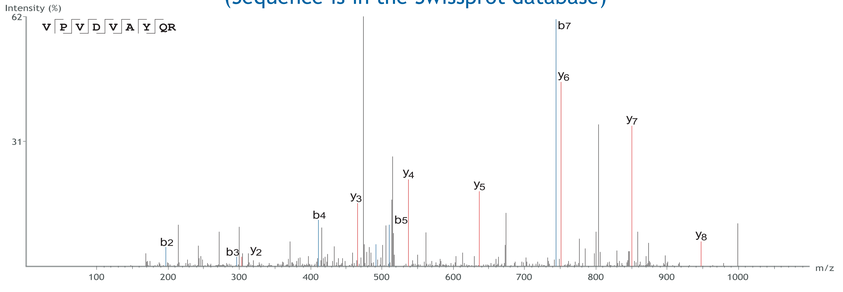
The LC-MS/MS test data was collected with a Thermo LTQ-Orbitrap CID instrument on a fraction of Trypsin digest of Human tumour cell. The dataset consists of approximately 9000 MS/MS spectra. Mascot is used to search the Swissprot database and the top-scoring peptide for each spectrum is kept. A random decoy database is searched together to determine the false discovery rate (FDR). At 1% FDR, Mascot identified 2942 peptide-spectrum matches (PSM). Figure 1 shows the performance of the de novo sequencing on these 2942 PSMs. For over 95% of the PSMs that were reported by Mascot, PEAKS de novo sequencing found 5 or more correct amino acids. This indicates that today's automated de novo sequencing software has a satisfactory performance for computing sequence tags.

Figure 1. The Percentage of Mascot High-Confident PSMs that De Novo Sequencing Computes ≥X Correct Amino Acids



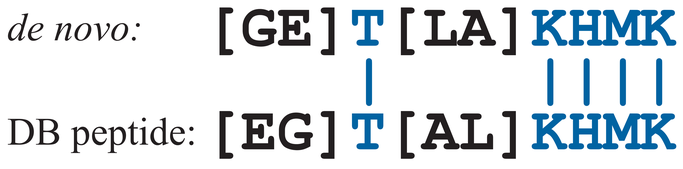
Surprisingly, there are many high confident de novo sequencing results that are not reported by Mascot database search as the top-scoring peptide of a spectrum above the 1% FDR threshold. Figure 2 illustrates an example. The actual performance of de novo sequencing is better than shown in Figure 1.

Figure 2. Peptide Computed by De Novo Sequencing Correctly, But Not Found by Mascot (Sequence is in the Swissprot database)



The above results inspired us to develop a simple sequence tag search tool to find approximate matches of the de novo sequences from the Swissprot database. The following score ranks the approximate matches: number of matched amino acids + 0.1 × number of matched mass blocks + 0.01 × number of maximum consecutive matches. If there is a tie, we report the match whose matched amino acids have lower frequencies in the database.

Figure 3. De Novo vs Database Peptide
The number of matched amino acids is 5; the number of matched mass blocks is 2; the maximum consecutive match is 4; so the score is equal to 5.24.

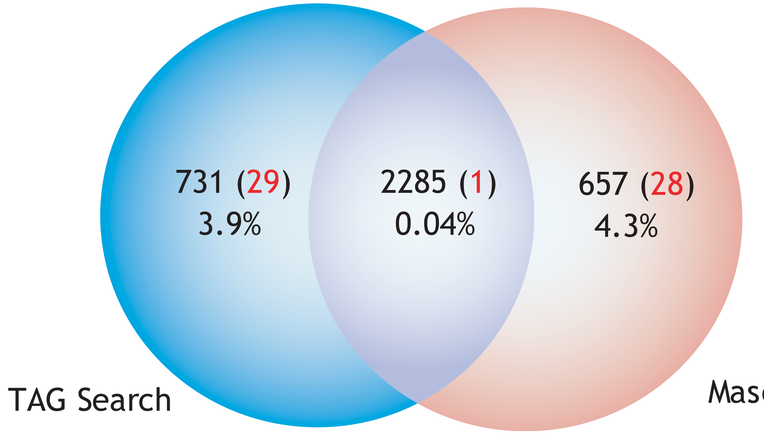


Secondly, the de novo sequencing results were used to perform a simple sequence tag search in the protein database to identify peptides. The matching score between the de novo and database sequences is measured by the number of matched amino acid and the number of consecutively matched amino acids. This simple tag search method is compared against the conventional database search tool Mascot.

Results

This simple tag searching strategy by de novo sequencing surpassed the Mascot search. At 1% FDR (estimated by target-decoy), the tag search identified 3016 PSMs, while Mascot identified only 2942. Figure 4 shows further details of the comparison.

Figure 4. The Venn Diagram of Identified PSMs by Tag Search and Mascot Search at 1% FDR



The numbers in red are the decoy matches and the percentiles are the corresponding FDR.

Seeing these results, we advocate the use of de novo sequencing in every proteomics data analysis, not only for the purpose of finding novel peptides, but also to complement the database search approach for identifying more peptides with higher confidence. In fact, this strategy is used in the PEAKS DB algorithm (a module of PEAKS Studio 5.3) to significantly improve the database search accuracy and sensitivity (see ASMS 2011 Poster MP457).

Conclusion

1. De novo sequencing performs very well in finding long sequence tags.
2. De novo sequencing tag search outperforms traditional database search with un-interpreted spectra.