Integrating de novo sequencing and database search to improve peptide identification

Lei Xin, Zefeng Zhang, Lian Yang, Baozhen Shan

1 Bioinformatics Solutions Inc, Waterloo, ON, Canada; 2 University of Waterloo, Waterloo, ON, Canada

Overview
Purpose: To improve peptide identification
Methods: Integrating de novo sequencing and database search approaches
Results: 6% more peptide identification

Introduction
A key step in shotgun proteomics is peptide identification. There are two complementary approaches for the analysis of LC-MS/MS spectra: database search and de novo sequencing. A protein sequence database search is prioritized for prioritized for database peptides and modified peptides, when a database is available. De novo sequencing is the only option for novel or homolog peptides which are not in the database. Unlike target-decoy approach for database search, there lacks a validation approach for peptide de novo sequencing. Here we describe a workflow integrating database search and de novo sequencing, in which database peptides are used to validate de novo peptides. Thus, the accuracy of de novo peptides can be estimated. The workflow maximized the peptide identification.

Method
A local confidence score was assigned to each residue of the de novo peptide to indicate how likely a residue is correctly sequenced.

Let T1 be the set of M/S/MS spectra. Perform de novo sequencing and database search with T1.

Let T2 be the set of the spectra identified by database search with 1% of FDR. A de novo peptide in T2 was validated with the database peptide at residue level. The local confidence score distributions were plotted for de novo residues that agree/disagree with database residues. For the de novo peptides in T3 = T1–T2, their score distributions of correct and incorrect residues were estimated with validated distributions in Step 3.

Results
The LC-MS/MS data set from yeast on an Orbitrap instrument [1] was used to demonstrate the workflow. It contains 35821 MS1 scans and 66479 MS/MS scans.

1. Perform database search with PEAKS 7 against SWISS-PROT. 43349 of 66479 spectra (65.2%) were identified with 1% false discovery rate. The target hits and decoy hits were shown in Figure 2.

2. Local confidence score was used to estimate the correctness of de novo sequence at residue level. Confident database hit was used to validate the de novo sequence of the same spectrum as shown in Figure 3. The assignment of amino acid in de novo sequence is correct if consistent with the one in database peptide, incorrect otherwise.

3. A local confidence score threshold can be determined to filter low confidence residues. Average local confidence (65%) was used to filter de novo sequences with less than 10% of residue error rate. 4428 spectra (6.7%) were identified with de novo sequences. The confident de novo peptides were exported along with confident peptides by database search.

4. With protein BLAST, 90% of the exported de novo peptides are significant (Table 1).

Table 1. Protein BLAST results of de novo peptides (part of the list)

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Protein Name</th>
<th>E-value</th>
<th>De novo Sequence</th>
<th>ALC ppm</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEAFGSGTAAVVSPIK</td>
<td>P33853</td>
<td>5.40E-06</td>
<td>HCI-ST.ID.ID.ID</td>
<td>MS2 spectra remain unmatched (28.1%)</td>
</tr>
<tr>
<td>VT2551</td>
<td>ZEFENG.ZHANG</td>
<td>8.20E-06</td>
<td>DEIEIEIEIEI</td>
<td>MS2 spectra with database hits (65.2%)</td>
</tr>
<tr>
<td>DEIEIEIEIEI</td>
<td>P33853</td>
<td>8.20E-06</td>
<td>DEIEIEIEIEI</td>
<td>MS2 spectra with de novo hits (6.7%)</td>
</tr>
</tbody>
</table>

Conclusions
A workflow to improve peptide identification.

References