In Depth Analysis of the Spectra Unassigned by Database Search

Bioinformatics Solutions Inc.

Lian Yang[1], Baozhen Shan[2], Zefeng Zhang[2], Weiwu Chen[2], Bin Ma[1]
University of Waterloo, Waterloo, ON [1]
Bioinformatics Solutions Inc, Waterloo, ON [2]

Introduction

Due to unexpected PTMs, mutations, contaminants and novel peptides, nearly every proteomics mass spectrometry (MS) experiment produces a large amount of high-quality spectra not matched by any database peptides. The confident identification of these "non-database" peptides are valuable for all proteomics research and particularly important to such applications as protein sequencing, antibody confirmation, and biomarker discovery.

We propose an automated workflow to identify both the database and non-database peptides to maximize protein coverage. The workflow combines de novo sequencing, database search, unspecified PTM search, and homology search together (Figure 1). Moreover, it supports the use of multiple enzymes for digesting the protein to maximize the sequence coverage[1].

Method

The mass spectrometry data was collected with an Orbitrap instrument on a standard protein (ALBU_BOVIN) sample ordered from Sigma. To maximize the coverage, three enzymes, Trypsin, LysC, and GluC, are used to digest the protein, respectively. The MS/MS spectra are analyzed by the following algorithmic workflow:

1. All MS/MS spectra are analyzed by both de novo sequencing and database search. A list of highly confident proteins are identified. The spectra with highly confident de novo sequence tags but no significant database peptide matches are selected for further analysis.
2. Each peptide from the highly confident proteins are “modified” in-silico by trying all
possible PTMs in the Unimod database. These theoretically modified peptides are compared with the selected spectra to identify modified peptides. De novo sequence tag matches are used to speed up the search.
3. The remaining highly confident de novo sequence tags are used for a homology search to identify mutated peptides. The SPIDER algorithm is used to reconstruct the most likely peptides from a database peptide approximately matched by the de novo sequence tags.
4. The de novo sequences tags that are unassigned in any of the above steps are reported as possible novel peptides.

Results

The workflow was implemented in the PEAKS 6 software. The number of PSMs found in each step is shown in Figure 1. To generate the numbers, 1% FDR was used to filter results of the database, PTM, and homology searches; and PEAKS ALC>70% was used to filter the results of the confident de novo tags.

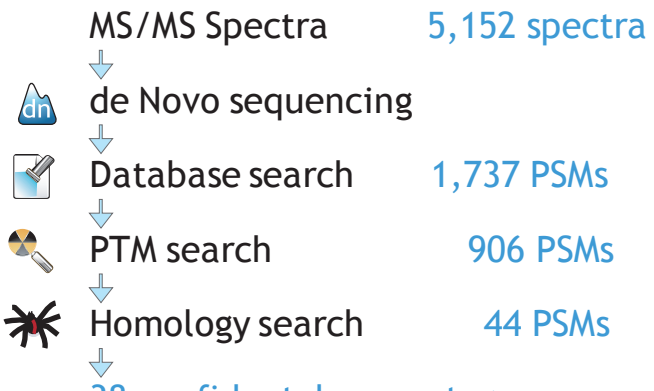

Figure 1. Each additional step of the workflow identified more peptides than using database search alone.

The software confidently identified the ALBU_BOVIN protein with almost full coverage (Figure 2). Additionally, it also reported several contaminant proteins, including human keratin proteins (K2C1_HUMAN and K1C9_HUMAN), bacteria protein (SSPA_STAAR), and trypsin (TRY1_BOVIN). Each of these contaminants has at least 2 unique confident peptides identified and their peptide-spectrum matches were all manually examined for correctness.

Figure 2. The Protein Coverage Outline on the ALBU_BOVIN Protein.

In the database search step, three expected variable PTMs: carbamidomethylation (C), oxidation (M), deamidation (NQ), were specified manually. Additionally, the PTM search step discovered many other PTMs not specified by the user. The most frequently ones are: carbamidomethylation on other amino acids, dehydration, methyl ester, carbamylation, dethiomethyl, ammonia loss, acetylation, formylation, hexose, sodium and pyro-glu from Q.

The homology search step also reported a few suspicious mutation sites. The most likely one is on position 214. Figure 3 shows that the software identified several PSMs with the same mutation on that site, indicating this is likely correct. The peptide-spectrum annotation is examined by clicking on one of these peptides (Figure 4); and indeed, the mutated peptide is supported by highly confident peaks. A literature search also confirmed that the mutation at this site was previously reported in (Brown 1975, Fed. Proc. 34:591).

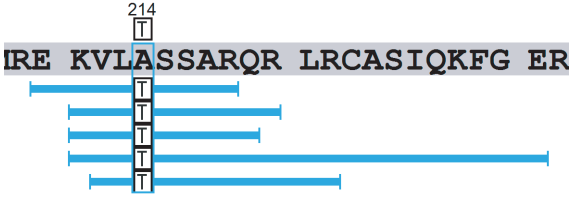

Figure 3. PEAKS software reported several PSMs with the same mutation, indicating a highly confident mutation site.

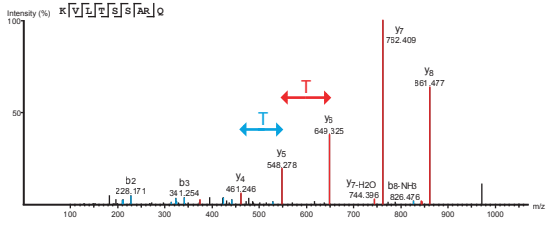

Figure 4. The peptide-spectrum annotation reports strong evidence peaks for the mutation at site.

We have not tried to interpret all the de novo only peptides. Potentially, these peptides can be from endogenous peptides, cross-linked peptides, contaminants, or peptides with more complex PTMs such as glycosylation.

Discussion

The new automated workflow identifies both the database and non-database peptides from MS/MS. It supports the “blind” search of PTMs by trying all PTMs in Unimod database, searches for mutations, and reports the “de novo only” peptides. The workflow has been implemented in the PEAKS 6 software.

Reference

X. Liu et al. Automated protein (re)sequencing with MS/MS and a homologous database yields almost full coverage and accuracy. Bioinformatics 25(17):2174-2180. 2009.