Automated Multiple Round Searches to Increase Coverage of Peptide/Protein Identification

Baozhen Shan[1]; Lei Xin[1]; Weijie Yang[1]; Gilles Lajoie[2]; Bin Ma[3]

[1]Bioinformatics Solutions Inc., Waterloo, ON. [2]University of Western Ontario, London, ON. [3]University of Waterloo, Waterloo, ON.

Bioinformatics Solutions Inc.

University of Waterloo

Western

Overview

Purpose: To maximize the protein coverage in MS/MS proteomic data analysis.

Methods: (1) PEAKS™ de novo sequencing results are used to automatically discover variable PTMs. (2) Protein coverage is increased with the second and third round searches with the discovered PTMs.

Introduction

One of the challenges researchers face in mass spectrometry-based proteomics investigations is that there are often a significant amount of high-quality spectra remaining un-interpreted due to PTMs and errors in MS/MS data and protein sequence databases. Specifying many variable PTMs in the protein identification software can increase the coverage, but also drastically slow down the searching speed. This dilemma can be partially solved with a two-round search approach: the first round searches a large database with only a few PTMs, followed by a second round on only the identified proteins but with many variable PTMs specified. However, this still requires a human's knowledge about the variable PTMs in the sample, in order to specify them correctly in the second round search. We propose to use PEAKS™ de novo sequencing [1] results to automatically discover the variable PTMs existing in the sample. In addition, we propose a workflow for multi-round searches which results in higher protein coverage.

Methods

The work flow of the multiple round search is shown in Figure 1.

Round 1. This is typical PEAKS™ database search. Only a few most common PTMs are specified for the search. PEAKS™ database search requires the de novo sequencing results as sequence tags, and therefore is performed after the de novo sequencing.

Round 2. In this round, variable PTMs are first discovered by comparing the de novo sequencing results with the peptides identified in the first round. We observe that for a peptide with a variable PTM, very often both the unmodified sequence (S) and the modified sequence (S') produce spectra in the data. The first round search may identify S but not S'.

However, the de novo sequencing result on spectrum of S' usually shares a long sequence tag with S if the spectrum is of high quality. Therefore, by comparing the de novo sequencing results with the first round database search results, many variable PTMs can be discovered.

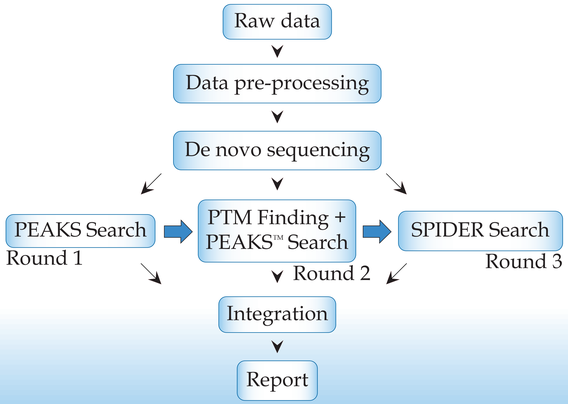
Then, another PEAKS™ database is done for the spectra unidentified in the first round. However, all the newly discovered variable PTMs are turned on, and the search is limited to only the proteins identified in the first round.

Round 3. The de novo sequences of the spectra not identified in the first two rounds are used to perform a homology search against the identified proteins. The SPIDER module [2] of the PEAKS™ Studio software is used for the search. This round will pick up more peptides that were missing in the first two rounds due to reasons such as precursor mass error, database sequence error, and point mutations between the studied protein and the database protein.

Results

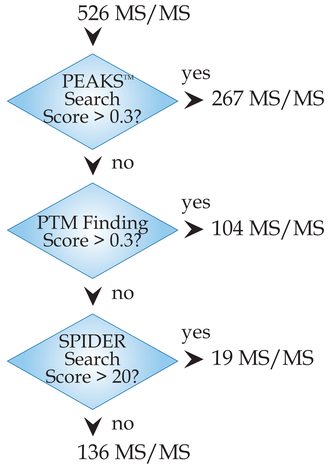
The method was tested with MS/MS data sets from six protein mixture (LC Packings) digested with trypsin and alkylated with iodoacetic acid. All spectra were acquired on a LTQ Orbitrap XL. 900 MS/MS spectra were collected along with 2076 survey scans.

Figure 1. Work flow of multiple round search



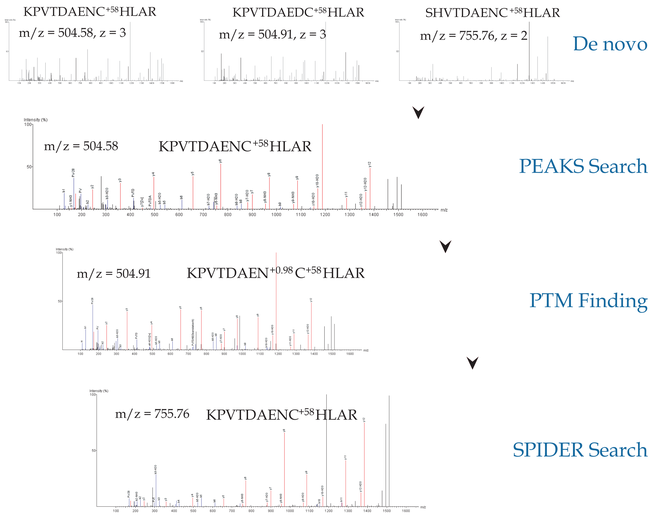
After removing spectra with poor quality by data pre-processing, 526 MS/MS spectra were searched against SWISS-PROT database with PTMs of Carboxymethyl on Cys. The result is shown in Figure 2. 267 spectra were identified in the first round. The coverage is 50.8%. In second round, the PTMs of oxidation on Met and deamidation on Asn and Gln were automatically selected. 104 more spectra were identified. With SPIDER search, 19 more spectra were identified, and the total coverage was 74.1%.

Figure 2. Outcome for each round of the searches



An example of the multi-round search result for three spectra were shown in Figure 3. The first round search successfully identified the sequence KPVTDAENC^{+58}HLAR for the spectrum m/z = 504.58 but failed with the other two spectra. The second round search further identified KPVTDAEN$^{+0.98}$C^{+58}HLAR for the spectrum m/z = 504.91 after deamidation on Asn was added as a variable PTM. The third round search further explained the spectrum m/z = 755.76. It comes from the same peptide as the spectrum with m/z = 504.58 but was not identified in the first two rounds due to a data error in the precursor mass.

Figure 3. An example multiple round search



Conclusion

The experiment results showed that our multiple round search approach increased the coverage of protein/peptide identification. The identification process is fully automated.

Reference

B Ma, et al., Rapid Communications in Mass Spectrometry, 17(20):2337-2342. 2003

Y. Han, et al., Journal of Bioinformatics and Computational Biology 3:697-716. 2005.