New Computational Method for Identifying Peptides with Unspecified Post-Translational Modifications

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

Overview

Purpose

To identify modified peptides from tandem mass spectrometry (MS/MS), without the need to specify the expected post-translational modifications (PTM) types. Instead, all known PTM types from the Unimod database are used.

Methods

1. PeaksPTM utilizes a two-pass search approach and a new scoring function to identify modified peptides by turning on all modifications from Unimod database.
2. A simple but effective strategy is used to combine multiple PTM search engines' results to further improve the identification rates.

Results

1. This consensus strategy helps to identify more modified peptides with high confidence.
2. PeaksPTM contributes more identifications than other three widely used search engines.

Introduction

Efficient and accurate identification of peptides with modifications plays an important role in the proteomics analysis. Most current database search software tools can not turn on all the known modifications since it will lead to a combinatorial explosion of the search space. Therefore, researchers have to specify a few possible PTMs using their best judgment. In such cases, the peptides with unspecified modifications would be lost in the final identification. Many researchers believe this is an important reason for the low characterization rate in the MS/MS data interpretation, especially for the data with many unassigned high quality MS/MS spectra using traditional database search engines.

To overcome the limitation of the database search approach, several software tools have been previously developed to provide the blind or unspecified search functions, considering either all possible mass shifts or hundreds of known PTMs. In this poster, we present a new database search engine, PeaksPTM [1], which outperforms several other software tools. We also demonstrate that a simple combination of the results of multiple tools can achieve a better performance.

Methods

PeaksPTM utilizes a two-pass search approach. The first pass uses PEAKS software for protein identification with only a few common PTMs. The second pass searches for modified peptides of those identified proteins, while considering all the PTM types from Unimod library.

For any peptide candidate with a modification, three features are used to determine its correctness:

1) the PEAKS LDF score
2) the rareness of the modification
3) the peptide pair

To estimate the false discovery rate (FDR), we follow the modified target-decoy method proposed in [2], which is specifically designed for the two-round search approach. The first pass identifies some target proteins (P1) and decoy proteins (P2). In the second pass, the decoy database consists of both the decoy proteins P2 and the random shuffles of proteins in P1. This is more strict than the traditional target-decoy approach and ensures that the FDR is not underestimated.

A consensus strategy is applied to combine the identifications from four search engines: PeaksPTM, Mascot (error-tolerant mode), Paragon™ and InsPecT. A peptide-spectrum match (PSM) is identified by more than one search engine or by only one search engine with FDR less than 0.8% is considered as a confident identification.

Data Set

The experimental data set was collected using an LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific™, Bremen, Germany) from human heart tissue sample, consisting of 11,207 MS spectra and 15,030 MS/MS spectra.

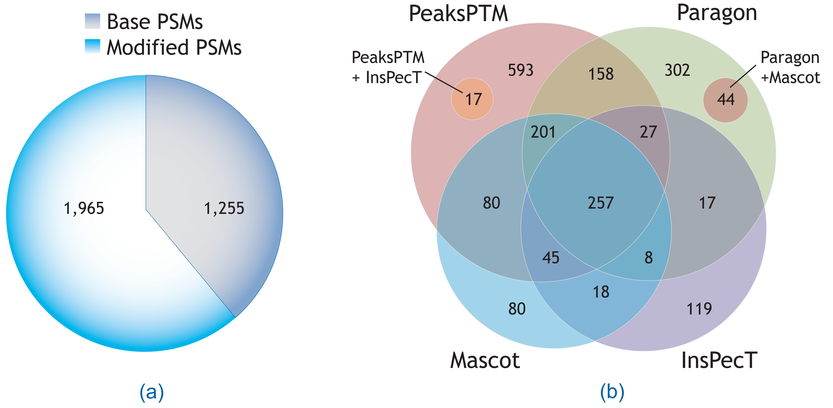
Results

The performance of PeaksPTM was compared with other three widely used search engines, Mascot (error tolerant mode), Paragon and InsPecT with FDR 1% (shown in Table 1). As we do not have access of the source code of other engines, their FDR was estimated by the traditional target-decoy approach, which may underestimate the FDR. Although using a more strict FDR control method, PeaksPTM still performances the best among the four search engines.
Refer to **Table 1**.

Table 1. Results of Four Search Engines

Search Engines		Mascot (Error Tolerant Search Mode)	Paragon	InsPecT	PeaksPTM
Identifications with FDR≤1%	Base PSM Number	612	943	612	**1,016**
	Modified PSM Number	743	1,029	521	**1,394**

Using the aforementioned consensus strategy, 3,220 PSMs were reported, 1,965 of which were modified PSMs and 1,255 base PSMs (see **Figure 1(a)**). This consensus strategy can increase at least 40% high confident identifications of modified PSMs than any single search engine. **Figure 1(b)** illustrates the composition of 1,965 modified PSMs contributed by four search engines.



(a) (b)

Figure 1.
(a) Consensus result of identified PSMs by using consensus strategy on human heart dataset.
(b) The composition of modified PSMs contributed by four search engines.

Conclusions

PeaksPTM does not only find more modified peptides than other search engines, but also contributes to more identifications when used in the consensus analysis with other engines.

References

1. X. Han, L. He, L. Xin, B. Shan and B. Ma. PeaksPTM: Mass Spectrometry Based Identification of Peptides with Unspecified Modifications. *Journal of Proteome Research*. Accepted for publication.
2. M. Bern, B. S. Phinney, D. Goldberg. Reanalysis of Tyrannosaurus Rex Mass Spectra. *Journal of Proteome Research*. **2009**, 8, 4328-4332.

Availability

The PeaksPTM software is freely accessible at http://bioinfor.net/ptm.

Acknowledgement

This work is supported by NSERC (RGPIN 238748-2006), two MITACS Accelerate Fellowships, and Bioinformatics Solutions Inc.