Integrating *de novo* Sequencing and Database Search for Peptide Identification

Bioinformatics Solutions Inc.

Baozhen[1] Shan, Lei Xin[1], Bin Ma[2]

[1] Bioinformatics Solutions Inc, Waterloo, ON

[2] University of Waterloo, Waterloo, ON

Overview

Purpose: To improve peptide identification for maximum proteome coverage.

Methods: Integrating database search and de novo sequencing approaches.

Results: With high resolution MS data, the integration significantly increases proteome coverage.

Introduction

Peptide identification with high sensitivity and accuracy is vital in mass spectrometry-based proteomics. Database searching is the primary method for identifying tandem mass spectra. Unfortunately, standard database searching is limited to the identification of spectra for which peptides are present in the database, preventing the identification of peptides from mutated or alternatively sliced sequences. De novo sequencing has the ability to provide alternative peptide identifications, as it does not require a protein database; however, identification without a database potentially reduces the accuracy. One approach to increase confidence of peptide identification is through high resolution tandem mass spectrometry on both precursor and fragment steps. A workflow is presented to combine de novo and database search for peptide identification on high resolution data.

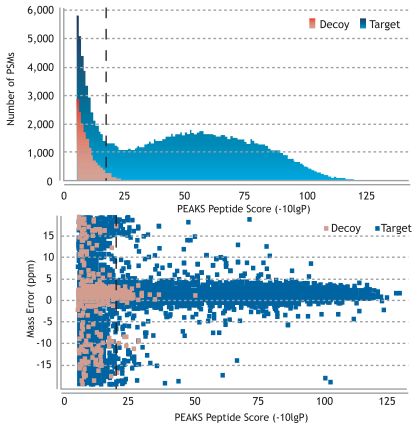
Methods

The workflow integrates de novo sequencing and database searching for peptide identification.

It contains 3 steps:

1. Perform database search with all MS/MS spectra against protein sequence database. Database peptides were selected with 1% false discovery rate (FDR) at the level of peptide-spectrum match (PSM). To estimate FDR accurately, an improved target-decoy method, decoy fusion, was used, in which the target and decoy sequences of the same protein were concatenated together as a single entry of the database [1].
2. For unidentified spectra in step 1, use PEAKS PTM algorithm [2] to perform a database search with the spectra of highly confident de novo sequence tags and by turning on all modifications in Unimod database. Peptides with unsuspected modifications were selected with 1% FDR.
3. Select the spectra with high confident de novo sequences but not identified in above steps.

Figure 1: The Distribution of PEAKS Peptide Score (-10lgP)



Results

The workflow was implemented in PEAKS. A high resolution MS dataset published by D.S. Kelkar [3] was used to test the algorithm. Twenty-one raw data files were downloaded from the ProteomeCommons.org Tranche Network, containing 208 844 MS/MS spectra obtained from cell lysates of Mycobacterium tuberculosis with strong caution exchange chromatography on LTQ-Orbitrap Velos.

De novo assisted database search was performed with PEAKS DB [1] against a protein database of M. tuberculosis H37Rv, containing 3 989 protein entries. Search parameters were as follows: (a) Trypsin as an enzyme allowing semi-tryptic cleavage and up to 2 missed cleavages; (b) Precursor mass error tolerance of 20 ppm; (c) Fragment mass error tolerance of 0.1 Da; (d) Fixed modification was carbamidomethylation of cycteine. Variable modifications were oxidation of methionine, acetylation of peptide N-terminus and formylation of methionine. FDR was estimated using target-decoy approach. The distributions of PSM scores (-10logP) in target and decoy were shown in Figure 1.

PSMs were filtered based on the score threshold for 1% FDR (corresponding peptide score of -10logP = 17.5). 116 192 PSMs were identified in step 1. The total number of unique peptides was 20 643, with 5 ppm precursor mass errors and 15 ppm of fragment mass errors, as presented in Figures 2 & 3 respectively.

29 048 of 116 192 spectra have de novo confidence scores (ALC) great than 80%. Compared with database peptides, the percentage of consistent amino acids for de novo sequences was 91%, where I = L, GG = N and M(+15.99) = F. 6 318 of 29 048 PSMs were exactly same between database peptides and de novo sequences.

Figure 2: The Precursor Mass Error Distribution

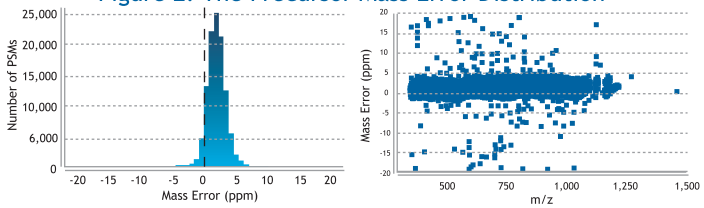
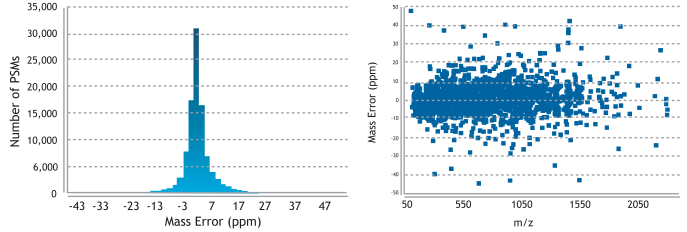
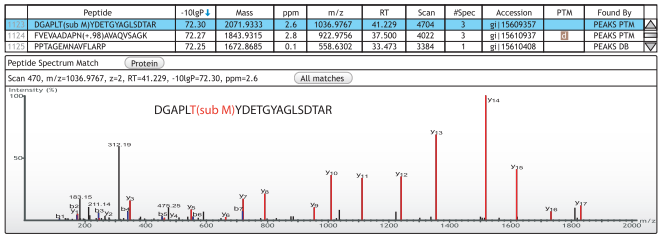


Figure 3: The Fragment Mass Error Distribution



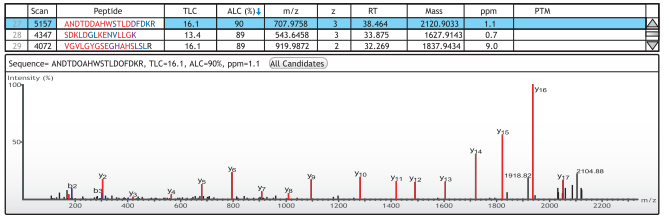
65 805 MS/MS spectra provided very good de novo sequences (ALC > 80), but the database peptide scores were far below the threshold (-10logP < 10). One major reason was the limited number of PTMs specified for the database search. Traditional database search software will run into speed problem if too many PTMs are specified. To solve this problem, the new PEAKS PTM module of PEAKS 5.4 was used to perform modified peptide search. PEAKS PTM allows users to turn on all modifications when searching the Unimod database, including single amino substitutions. By doing this, 14 181 additional PSMs were identified. Combined with the PSMs identified by PEAKS DB, 130 373 PSMs were identified with 26274 unique peptides. One example is presented in Figure 4.

Figure 4: An Example of Modified Peptide



In addition, 51 624 spectra were selected in step 3 with ALC great than 80%. Such high confidence peptide identification provides a direct evidence of translational potential of a genomic region. More than 13 novel ORFs were identified with those de novo-only peptides [1], as shown in Figure 5.

Figure 5: Identification of a Peptide from the Novel ORF (Genome: 2402507-2402722)



Conclusions

Integrating de novo sequencing and database search significantly improves peptide identification.

References

[1] Han X. et al. PEAKS PTM: Mass spectrometry-based identification of peptides with unspecified modifications. J. Proteome. Res. 10:2930-36(2011).

[2] Zhang J. et al. PEAKS DB: De Novo sequencing assisted database search for sensitive and accurate peptide identification, MCP.M111.010587 (2011).

[3] Kelkar D. S. et al. Proteogenomic analysis of Mycobacterium tuberculosis by high resolution mass spectrometry, MCP.M111.011627(2011).