Integrating de novo Sequencing and Database Search for Monoclonal Antibody Sequencing

Bioinformatics Solutions Inc.

Baozhen Shan, Lei Xin
Bioinformatics Solutions Inc, Waterloo, ON

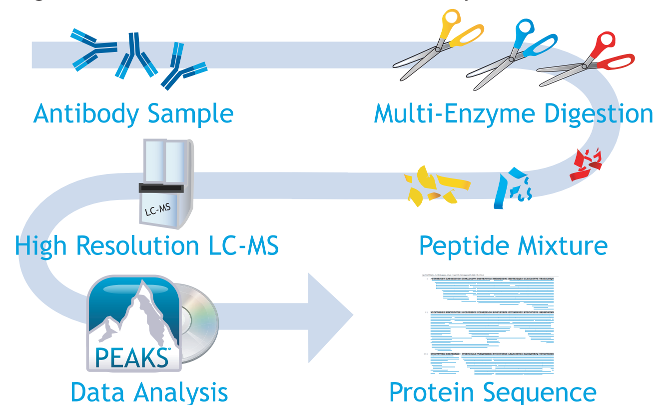
Overview

Purpose: To improve monoclonal antibody characterization
Methods: Integrating database search and de novo sequencing approaches
Results: An alternative workflow for sequencing a monoclonal antibody

Introduction

Tandem mass spectrometry has become an important method for identifying protein mixtures. However, it is still an area of active research to develop workflows to determine the primary sequence of antibodies. Software tools solely relying on database search are insufficient delineating the sequences of monoclonal antibodies because the complete sequences of antibodies are not contained in any database. Advancement in mass spectrometry enables more accurate de novo sequencing for MS/MS. A workflow, Figure 1, will be presented that combines multi-enzyme digestion, acquisition of high-resolution on both precursor and fragment steps, and integration of database search and de novo sequencing for characterizing the monoclonal antibodies.

Figure 1. Workflow of Monoclonal Antibody Characterization

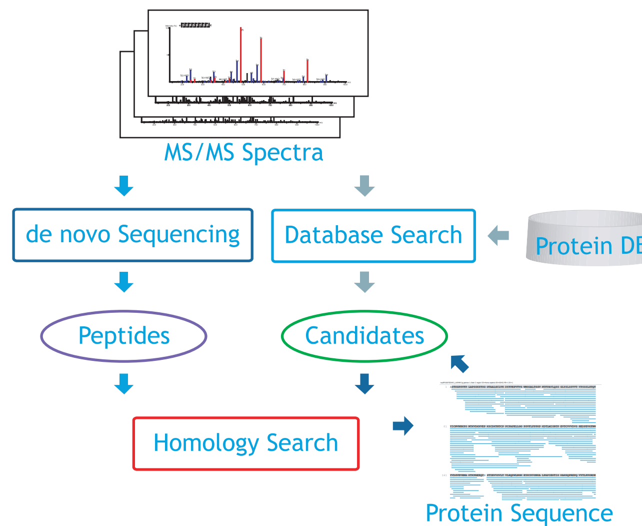


Methods

The data analysis in the workflow integrates de novo sequencing and database searching for sequencing monoclonal antibodies. It contains 3 steps presented in Figure 2.

1. Perform database search and PTM search by turning on all modifications in Unimod database with all MS/MS spectra. Select the top ranking proteins as candidates.
2. Perform de novo sequencing with the unidentified spectra in Step 1. The good de novo sequences are selected to perform homology matching with candidate proteins. Find the best homology matches, and replace the segments of a protein sequence with the de novo sequences. The new candidates are generated.
3. Repeat Step 2 until no better protein candidates could be found.

Figure 2. Diagram of Data Analysis

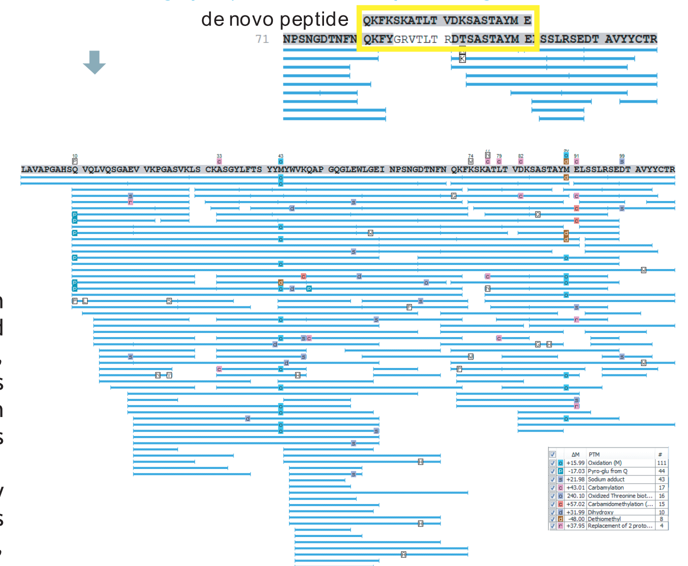


Results

This workflow was implemented in the PEAKS software. A human monoclonal antibody sample was reduced with DTT, alkylated with iodoacetamide, and digested with different enzymes (AspN, chymotrypsin, GluC, LysC, pepsin and trypsin). Peptide mixtures were analyzed using nana-LC-MS on LTQ-Ortritrap at high resolution in the Orbitrap for MS1 and MS2. Six raw data files containing 14483 MS/MS spectra were analyzed with PEAKS.

All MS/MS spectra were searched against a public antibody database. The mass accuracy was set 10 ppm for precursor ions and to 0.05 Da for fragment ions. The proteins (IGHG1, HV102, IGKC, KV304, etc.) were selected as the candidate for further analysis. At 0.5% of false discovery rate (FDR) at peptide-spectrum match (PSM) level, majority of the constant region sequences of heavy chain Ig gamma and light chain Ig kappa can be found in current database with the coverage of 99% and 100% respectively. However, the variable domains containing CDRs are not complete in the database, with the coverage of 40% for the heavy chain and 36% for the light chain. By finding homology matches between de novo sequences and database sequences, full sequences of the antibodies including both constant regions and variable regions were determined with 0.5% of FDR at PSM level. For example, the sequence of the variable region of the heavy chain was determined by the homology match of KFKSKATLTVDKSASTAYME from de novo sequence to KFYGRVTLTRDTSASTAYME from the database (in Figure 3).

Figure 3. An Example of Antibody Sequencing with PEAKS



Conclusions

A workflow enables an alternative approach for sequencing a monoclonal antibody.

Reference

[1] X. Liu, Y. Han, D. Yuen, and B. Ma. Automated Protein (Re)Sequencing with MS/MS and a Homologous Database Yields Almost Full Coverage and Accuracy. Bioinformatics 25(17):2174-2180. 2009.