Algorithm for Identification and Sequencing of Protein Variants Using Top-Down MS Data

Christian Heckendorf, Roger Theberge, Jean L. Spencer, Catherine E. Costello, Mark E. McComb.
Cardiovascular Proteomics Center, Boston University School of Medicine, Boston, MA.

Overview

- Top-down proteomics has emerged as a technique that preserves labile post-translational modifications and offers full protein sequence coverage.
- One of the major problems facing topdown MS/MS is the assignment of peaks due to the possibility of a large number of fragments from an intact protein.
- Now in development is a web-based extension of BUPID-Top-Down (Boston University Protein Identifier Top-Down) used to identify protein variants in top-down MS/MS spectra.
- This program will facilitate the penetration of top-down techniques into a greater number of mass spectrometry laboratories.

Introduction to Top Down MS/MS

Top down proteomics involves introducing intact protein ions into the mass spectrometer and fragmenting them using ion-activation methods such as CID, ECD, and ETD. This has the potential for complete protein sequence and PTM identification without having to spend time digesting the protein. Making use of top-down data is very computationally taxing and the availability of software that can do this effectively is limited.

Here, we describe the development of a web-based continuation of BUPID-Top-Down designed from the ground up for protein variant identification.

Algorithm for Variant Identification

BUPID Top-Down uses a sequence tag approach to identify proteins in a protein sequence database. Once the wild type sequence is known, this algorithm can be applied to identify variants.

The first step is to find the expressed protein sequence.

- Sequence tags are identified for the given peak list and sequence. These are used as guides for where the protein is known to be fragmented.
- Starting from each end, we truncate the protein by one amino acid and note which truncation yields the most fragment assignments within the sequence tag.

Identification of the peak representing the intact variant protein allows us to reduce the search space as well as the number of false positives. This is compared with the theoretical intact wild type protein mass to identify the variant mass shift.

- If the user has provided the software with the precursor mass, we simply use that.
- If there are sequence tags, we look at the peaks used to generate those tags. If there is a difference between the mass of the peak and the theoretical mass of the fragment that generated that peak, we add it to the list of mass shifts to look at.
- If there are no usable sequence tags, we scan the peak list for the highest intensity peak within the maximum variant mass shift of the theoretical wild type precursor.

Finally, potential variants are selected and measured.

- The peak list is recalibrated according to assignments made against the wild type sequence.
- We iterate through the list of possible variants and do the following for each one that's within a tolerance of one of the shifts:
  - Assign fragments to the peaks accoding to the modified sequence.
  - Count the total number of fragments and the total number of fragments that are large enough to contain the variant amino acid.
  - Find the Mahalanobis and Euclidean distance between the mass errors of wild type and variant assignments

Results

BUPID Top-Down Analysis of MS/MS of Hb SS ($E_6 \rightarrow V$)

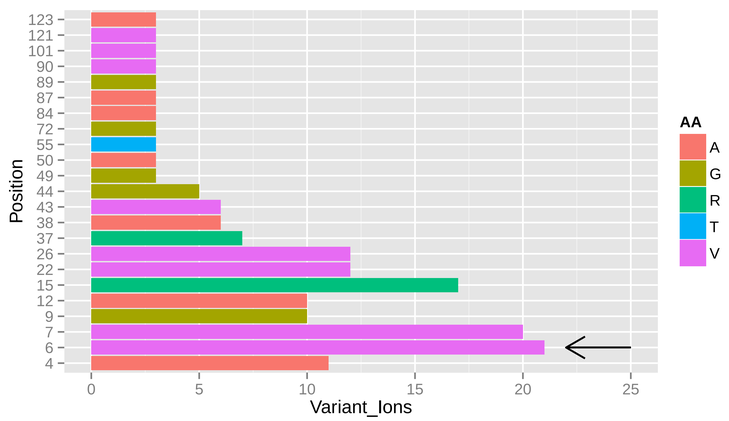

Figure 1, A plot of the total number of ions assigned that contained the variant amino acid organized by candidate amino acid and position.

Figure 2, A heat map representing the calculated mass error Mahalanobis distances organized by candidate amino acid and position.

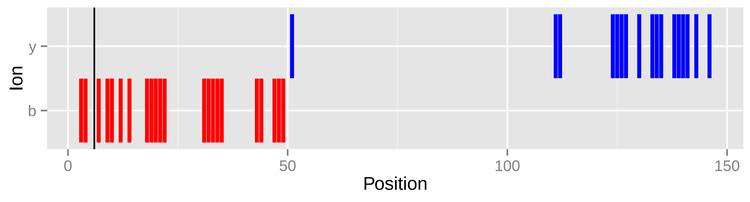

Figure 3, A graphical respresentation of the peaks assigned by BUPID Top-Down.

- These figures show the results of variant identification by BUPID Top-Down for a sample of Hb SS.
- Analysis of the BUPID Top-Down results show that the $E_6 \rightarrow V$ variant assigns the most fragment ions.
- The fragment ions produced by that same variant were assigned with mass errors closest to the fragment ions assigned for the wild type protein.

Example of Peak Assignment Using BUPID Top-Down

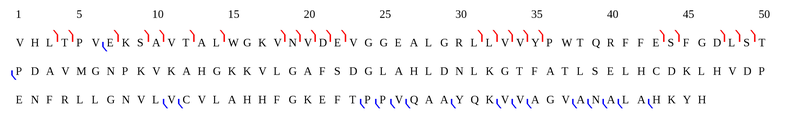

Figure 4, The Hb SS protein sequence with peaks assigned by BUPID Top-Down identified.

Discussion

Mass Shift Identification

- Using sequence tags to identify the mass shift works well when there are tags available. This is rarely the case.
- Deconvolution algorithms aren't perfect. Many times the wrong monoisotopic peak is selected, creating a false 1Da shift. Checking for these errors adds to the complexity of the project by:
  - Taking more time to compute
  - Introducing false positives

Variant Identification

- Given sufficient fragmentation of the protein, it becomes easy to identify the variant by maximizing the ion count while minimizing the error distance.
- There are often not enough fragments available to make a clear identification. This could be related to either the position of the variant or the fragmentation method.
- Distance methods can help but may misdirect the results. The distance for a candidate variant can be higher or lower than normal if there aren't enough observations.

Implementation

- This algorithm was implemented as a standalone program written in C with searches submitted through a web interface.
- Our interface allows users to connect to an open access server and easily submit searches as well as review the results independent of platform.

Conclusions

- BUPID Top-Down allowed us to easily identify the correct protein variant in the sample of Hb SS.
- A web based implementation of this algorithm on an open access server will enable many users to quickly analyze data.
- Use of BUPID Top-Down will enable facile and accurate data analysis thus expanding this approach to many users in the field.

Future Work

- Expansion to PTM identification.
- Expand web interface to allow more control over search parameters.

Acknowledgements

This project was funded by NIH grants R21 HL107993, P41 RR010888/GM104603, S10 OD010724, S10 RR020946, and S10 RR025082 and NIH-NHLBI contract HHSN268201000031C.
Special thanks to the members of the CBMS laboratories for their help.