Development of a Web-based Top-Down Protein Identification Tool

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

BOSTON UNIVERSITY

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 proteins 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 proteins into the mass spectrometer and fragmenting them using methods such as CID, ECD, ETD, etc. 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.

There are few tools available for top-down proteomics data analysis that support protein identification; these include:

- Mascot Top-Down (http://www.matrixscience.com): commercial, license required.
- ProsightPTM (http://prosightptm2.northwestern.edu; Nucleic Acids Research 2007; doi: 10.1093/nar/gkm371): free, web based.

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

Implementation

There are three major parts of the implementation of this program:

- The web front-end responsible for accepting search information and displaying the results of the searches.
- The back end where the searches are performed.
- The database generator where XML or FASTA databases are converted into a binary format which is easier to work with.

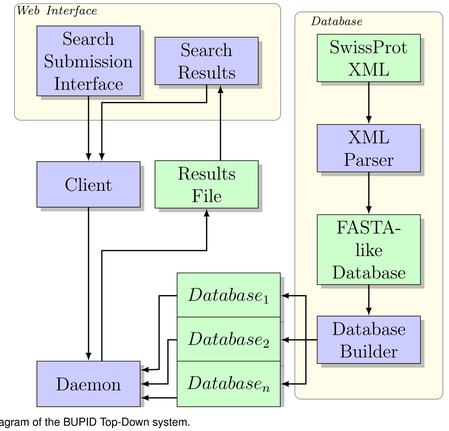

Figure 1, Diagram of the BUPID Top-Down system.

Challenges

The previously developed version of BUPID Top-Down relied on the user to provide the protein sequence. For unidentified proteins, this requirement is impossible to fulfill. This version eliminates that requirement by analyzing the peak list and selecting the best protein matches. In order to identify the correct protein, some obstacles must be taken into consideration:

- Experimental protein mass may not be equal to theoretical protein mass due to PTMs, sequence variants, etc.
- PTMs may limit the software's ability to generate good quality sequence tags.
- In most protein sequence databases, only the wild type is listed.

Algorithm for Protein Identification

BUPID Top-Down uses a sequence tag approach to identify proteins in a protein sequence database.

- Sequence tags are generated by comparing the mass difference between each pair of experimental masses.
- Each peak is used as a start point and masses following it are compared with it to find differences that are close to an amino acid mass. This method can be used independent of fragmentation type because two fragments of the same type will have a relative offset of zero.
- Mass differences that are within a set tolerance of an amino acid mass are added to the sequence tag graph.
- Sequence tags are generated by traversing the graph.
- Tags that meet the minimum length requirement are searched against the protein database. Since both C-term and N-term ions are produced, the tag must be searched in reverse order as well.
- Each protein is assigned a score based on the sequence coverage of the tags and proteins that exceed the score threshold are included in the results.

Protein Scoring Function

$$ProteinScore = \sum_{i=0}^{n} \frac{t_i^2}{s}$$

Where n is the number of tags, t is the length of the tag, and s is the length of the protein. This function gives sequence tags with a greater length a higher score. Since it is common to have several tags present, we sum the scores of each tag to get the total protein score.

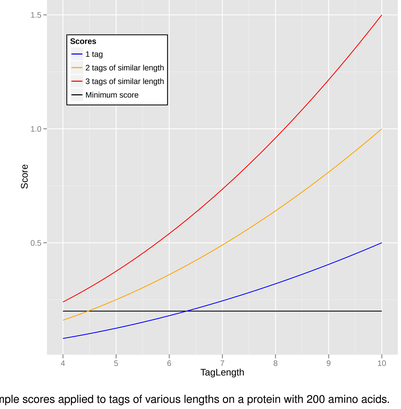

Figure 2, Example scores applied to tags of various lengths on a protein with 200 amino acids.

Results

Example of Peak Assignment Using BUPID Top-Down

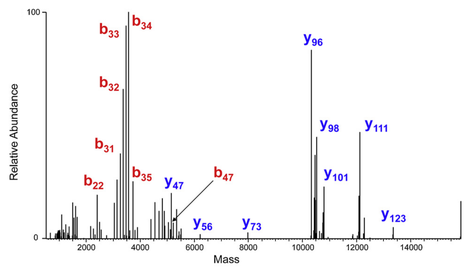

Figure 3, Source CID MS/MS of the beta chain of human hemoglobin: b and y ions shown. (R. Théberge, G. Infusini, W. Tong, M. E. McComb, C. E. Costello, Int. J. Mass Spectrom. 300, 2-3 (2010); doi:10.1016/j.ijms.2010.08.012)

Comparison with Mascot

Protein searches made on BUPID Top-Down produced comparable results to those made on Mascot Top Down.

- The BUPID Top-Down results ranked higher than the correct protein were typically the right protein with the wrong organism.
- Altering the scores of the top results based on the identifications made by the BUPID Top-Down peak fitting algorithm* allowed us to greatly improve the results.
- By filtering the results based on a user specified taxonomy, the few remaining inconsistencies and ambiguities were removed.

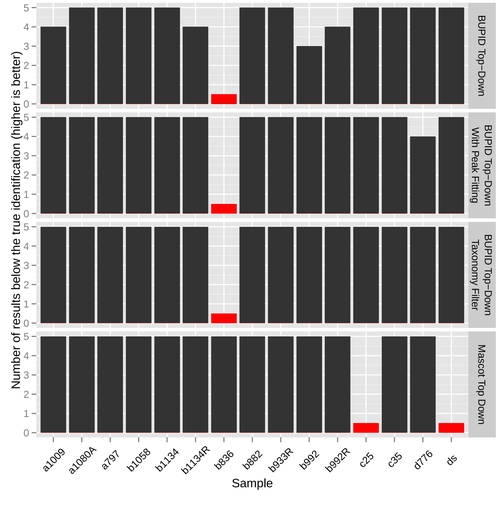

Figure 5, Comparison of result rankings between BUPID-Top-Down and Mascot Top Down. Results with the same score are merged into one rank. The top 5 results for each sample were taken and the number of results below (and including) the true protein were counted. Sample *b836*, which failed to match in any BUPID Top-Down run, had a number of great peaks but they were spaced such that no sequence tag was formed. Sample *ds* was matched in Mascot but failed to achieve a score high enough to be significant.

-* Tong, W., Théberge, R., Infusini, G., Cui, W., Perlman, D.H., Lin, C., McComb, M.E., Costello, C.E., BUPID-Top-Down: Database Search and Assignment of Top-Down MS-MS Data , Proceedings of the 57th ASMS Conference on Mass Spectrometry and Allied Topics, Philadelphia, PA, May 31-June 4, 2009.

Analysis of Hbβ by Top Down MS/MS

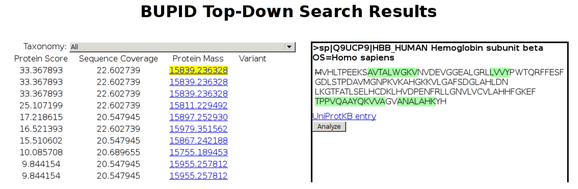

Figure 4, BUPID-Top-Down analysis of MS/MS of Hbβ. Sequence tags identified for the top result are shown as hilighted segments of the sequence.

Discussion

- Using a sequence tag approach combined with a peak fitting filter allows BUPID Top-Down to correctly identify the protein in most samples.
- BUPID Top-Down searches could make incorrect identifications for a few reasons:
  - Incorrect proteins may have been ranked higher by chance. The impact of this was reduced by introducing a peak fitting step which greatly boosts the score of the correct protein while generally providing no significant boost to incorrect proteins.
  - Good peaks may exist but are positioned such that no tags are formed. In order for a sequence tag to be considered, it must contain at least 4 amino acids. If the matched peaks are too scattered, tags of the required length will not be found.
- By designing the program with a client-server model in mind, we were able to overcome some of the design challenges. With only a single instance of the server, the database can be loaded into memory so that it will only need to be read from the disk once and subsequent requests can be made from the much faster main memory. This model also allows us to dedicate the total system resources to a single search rather than dividing them and risking overloading the system.

Conclusions

- Searches produce results comparable to Mascot.
- Use of BUPID-Top-Down will enable facile and accurate data analysis thus expanding this approach to many users in the field.
- Using sequence tags to search the database enables us to quickly and accurately identify the protein.
- Sequence tag searching combined with a peak fitting step improves the accuracy of the results but will not help if there are no tags to begin with.

Future Work

- Improved accuracy of sequence tag identification module.
- User configuration of currently hardcoded search parameters.

Acknowledgements

This project was funded by NIH/NHLBI N01 HV-28178, NO1 HV-00239, NIH/NCRR P41 RR10888/GM104603, S10 RR020946 and S10 RR025082. Special thanks to the members of the CBMS laboratories for their help.