Automated Protein Identification and Sequencing 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

BOSTON UNIVERSITY

Overview

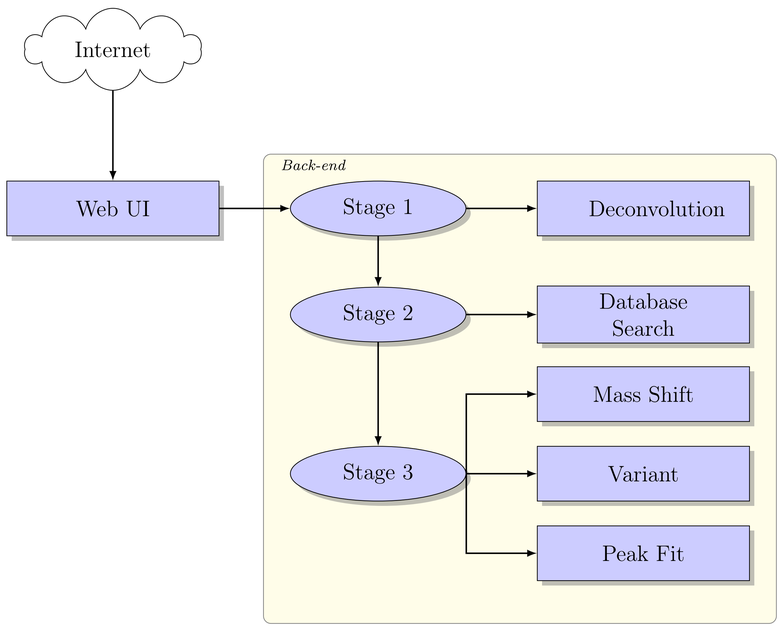
- Top-down proteomics has emerged as a technique that preserves labile post-translational modifications and offers full protein sequence coverage.
- One of the major problems facing topdown MS/MS is the assignment of peaks due to the possibility of a large number of fragments from an intact protein.
- Now in developpent is a web-based pipeline for BUPDID Top-Down, allowing for simplified access to a number of analysis tools for 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 analysis.

Architecture



Methods

- Each tool was redesigned to read from and write to a common results file format.
- Analysis can start from any of the three stages of processing depending on input:
  - Preprocessing - deconvolution
  - Identification - database search
  - Characterization - fragment, sequence variant, mass shift analysis
- Tools are run sequentially, managed by a shell script, but many of the tools are MPI/OpenMP enabled.

Results

Overview

Each module was tested with a series of simulated peak lists. Each peak list corresponds with a protein fragmented at 40% sequence coverage. The identification and fragment assignment modules were tested with the same five unmodified proteins. The variant analysis module was tested with the beta chain of human hemoglobin using several different fictitious mutations. The mass shift analysis module was tested using the original five proteins with the addition of three PTMs on each.

Identification

Protein	Rank	Lead
CYC	1	22.94
HBA	1	10.60
HBB	1	5.18
MYO	1	40.35
TTR	1	31.57

Lead shows the score ratio between this protein and the next-ranked one. HBB was ranked ahead of the very similar delta chain. HBA was ranked ahead of the somewhat similar zeta chain.

Fragment Assignment

Protein	TP	FP	Total
CYC	83	0	84
HBA	112	0	112
HBB	116	1	116
MYO	121	0	122
TTR	101	0	102

Variant Analysis

Variant	Rank	Total	Ambiguity
A140T	2	56	N
E6V	1	23	N
G25M	1	13	N
K120E	1	38	N
*L96Y	9	30	N
*T50C	4	58	N

*Human interpretation of the L96Y and T50C results allows for clear identification of the variants despite the low ranks.

Mass Shift Analysis

Protein	Mods Visible	Tag	Peak
CYC	3/3	3	3
HBA	3/3	2	3
HBB	3/3	2	3
MYO	3/3	2	3
TTR	3/3	2	3

Tools Involved

- Deconvolution
  - Uses the MasSPIKE algorithm at its core.
  - Extended for parallel processing, LC-MS, mzML.
  - Kaur, Parminder, & O'Connor, Peter B. 2006. Algorithms for automatic interpretation of high resolution mass spectra. *Journal of the American Society for Mass Spectrometry*, 17(3), 459–468.
- Database Search
  - Identifies candidate proteins using sequence tags.
  - Refines matches by assigning fragments.
  - Heckendorf, C., Théberge, R., Costello, C.E., & M.E., McComb. 2012. Development of a Web-based Top-Down Protein Identification Tool. *Proceedings of the 60th ASMS Conference on Mass Spectrometry and Allied Topics.*
- Fragment Ion Assignment
  - Identifies modified and unmodified fragments from a peak list.
  - Recalibrates the peak list based on assignments to correct for drifting accuracy.
  - Tong, W., Théberge, R., Infusini, G., Cui, W., Perlman, D.H., Lin, C., McComb, M.E., & Costello, C.E. 2009. 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.*
- Sequence Variant Analysis
  - Determines candidate amino acid substitutions and their positions.
  - Measures distance of assignment mass errors from wild type and number of fragments assigned.
  - Heckendorf, C., Théberge, R., Spencer, J.L., Costello, C.E., & M.E., McComb. 2013. Algorithm for Identification and Sequencing of Protein Variants Using Top-Down MS Data. *Proceedings of the 61st ASMS Conference on Mass Spectrometry and Allied Topics.*
- Mass Shift Analysis
  - Determines possible mass shifs present in the data at positions in a given protein sequence.
  - Leverages sequence tag information.
  - Calculates repeated deviations from theoretical fragment masses at various peaks.

Conclusions

- The individual modules in this pipeline offer a variety of information that build off each other.
- When the modules are combined into a pipeline, it can provide a fast high level overview of the data or be tuned for more detailed analysis.
- Use of BUPID Top-Down will enable facile and accurate data analysis thus expanding this approach to many users in the field.

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.