Automated Protein Identification and Sequencing Using Top-Down MS Data

View Presentation Detail

Authors
Christian Heckendorf; Roger Theberge; Jean Spencer; Catherine E. Costello; Mark E. Mccomb

Institutes
Boston University School of Medicine, Boston, MA

Introduction
Top-down mass spectrometry is a fast and efficient method to analyze proteins for identification, sequencing, variant determination and characterization of post-translational modifications. The progress of the technique is hampered by the lack of availability of readily accessible data interpretation tools. Here we describe the continued development of a web-based open-access search engine for top-down proteomics: BUPID Top-Down (Boston University Protein Identifier Top-Down). The software can now be used as an automated pipeline to analyze spectra obtained with various top-down fragmentation methods including CID, ECD and ETD. The development of an open-access top-down data interpretation tool via a web interface will facilitate the penetration of top-down techniques in a greater number of mass spectrometry laboratories.

Methods
The BUPID Top-Down software suite now consists of several tools for the different stages of data analysis. This software has also been expanded to allow for use as an automated pipeline, in addition to being used for more detailed study as standalone modules. By submitting profile data as mzML or a tab separated value file containing peak lists, and providing a series of analysis parameters, the data will be processed using the appropriate analysis modules and the results of each will be returned. This software is written using a combination of POSIX shell scripts and C programs designed to run on Linux/Unix servers. BUPID Top-Down is used with a web front-end which allows access through standard http web browsing.

Preliminary Results/Abstract
When the input provided is in mzML format, it is first processed using a deconvolution module before analysis. The protein identification module uses a sequence tag approach combined with a database search to select the candidate protein. Obtaining the correct protein identification is facilitated by requiring the user to input additional information such as the organism or precursor mass which will be used to further filter the results of the database search. Once the protein sequence has been identified, secondary processing can be used to identify features of interest in the given protein. Modules at this stage of analysis include assignment of fragment ions, identification of sequence variants, and mass-shift detection which can be used to characterize PTMs. Example results are presented on data obtained using this software by analyzing the top-down CID and ECD spectra of proteins pertinent to our center's research aims. These include variants of human hemoglobins, TTR, and PTMs, including glycosylation. Initial tests show that the protein could reliably be identified; the data were further processed to determine candidates for the variant, glycan, or other PTM. An overview of the software and representative results will be presented.

Acknowledgements: This project was funded by NIH-NHLBI contract HHSN268201000031C and NIH grants P41 GM104603, R21 HL107993, S10 RR020946, S10 OD010724, and S10 RR025082.

Novel Aspect
Novel software for protein identification and sequencing of top-down MS/MS data obtained from CID, ETD, and ECD.