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

Christian F. Heckendorf, Roger Théberge, Catherine E. Costello, Mark E. McComb.

Boston University School of Medicine, Boston, MA.

Novel Aspects (20):
Web-based software tool with a novel algorithm for identifying proteins in top-down MS/MS data using CID, ETD and ECD.

Introduction (120):
Top-down proteomics has emerged as a technique that preserves labile post-translational modifications and offers full protein sequence coverage. However, the progress of the technique is hampered by the lack of availability of readily accessible data interpretation tools. Here we describe the development of a web-based open access extension of BUPID Top-Down (Boston University Protein Identifier Top-Down) to identify proteins in top-down MS/MS spectra. The software can be used 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 (120):
BUPID Top-Down uses a sequence tag approach to identifying proteins in a protein sequence database. Sequence tags are generated by comparing the difference between each pair of experimental masses. Mass differences that are within a specific tolerance of an amino acid mass are added to the sequence tag tree. The tree is then traversed and tags that meet a minimum length requirement are matched against proteins in the database. This algorithm is implemented as a standalone program written in C designed to be run on Linux/Unix servers. The standalone program is used with a web front-end which allows it to be accessible through standard HTTP web browsing.

Preliminary Data (300):
This web-based protein identification version of BUPID Top-Down was designed to take advantage of server grade hardware to efficiently process data. A client/server model was chosen for the back-end to reduce overhead and system strain associated with simultaneous searches. The server maintains a search queue so that each search request can be quickly processed using the maximum available resources without the risk of overwhelming the system. Additionally, the server is able to convert the FASTA databases into a binary format and store them in memory allowing them to be used by all searches.

Example results were obtained using BUPID Top-Down to analyze the CID spectrum of a beta chain of hemoglobin. Initial tests show that the protein was correctly identified in the top results for all samples. Additional tests run using the ECD spectrum of beta chain of hemoglobin showed similar results. An overview of the algorithm, representative results and a comparison to other tools will be presented. The ultimate goal for this work is to develop a flexible top-down protein identification tool that can correctly identify proteins regardless of the presence of post-translational modifications.
Acknowledgements: This project was funded by NIH/NHLBI NO1 HV-00239, NIH/NCRR P41 RR10888 and S10 RR020946.