# Top-down MSMS Data Analysis: Software and Algorithms for Protein Identification and Fragment Ion Assignment

Christian Heckendorf, Roger Theberge, Deborah R. Leon, Kshitij Khatri, Jean L. Spencer, Catherine E. Costello, Mark E. McComb
Center for Biomedical Mass Spectrometry, 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.
- We previously developed BUPID Top-Down, a web-based analysis pipeline for processing top-down proteomics data, assisting with the assignment of fragments from potentially unknown proteins, and BTDR, an R package to visualize results.
- We redesigned the deconvolution algorithm to improve computational performance and enhance the accuracy of future identifications.
- We modified the protein identification score to take sequence tag frequency into account.

## 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 obtaining complete protein sequence and PTM identification without spending time digesting the protein. Processing top-down data is computationally taxing and the availability of software that can do this effectively is limited.

Recent development of BUPID Top-Down was aimed at improving the accuracy of results and reducing the computational resources needed to process the data. This change was prompted by the need to quickly process long LC/MS-MS runs and confidently assign ions in complex spectra.

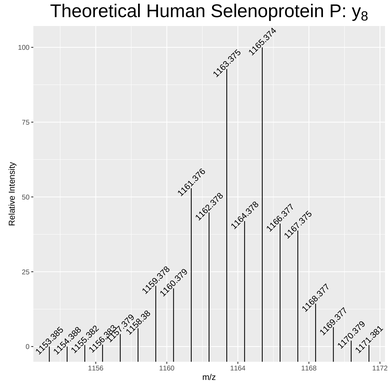
## Configurable Isotopic Distribution Models

Proteins containing unusual isotopic distributions may be processed using:

- Custom averagine
- Dynamically generated compositions based on protein sequence
- Specialized library called on demand

Using an appropriate isotopic distribution model is critical to choosing the correct monoisotopic mass, particularly when it's not visible in the data. Even in the low mass fragment illustrated here, 6 of the lightest isotopic peaks will likely be below the noise floor due to the presence of selenium.

Theoretical Human Selenoprotein P: $y_8$

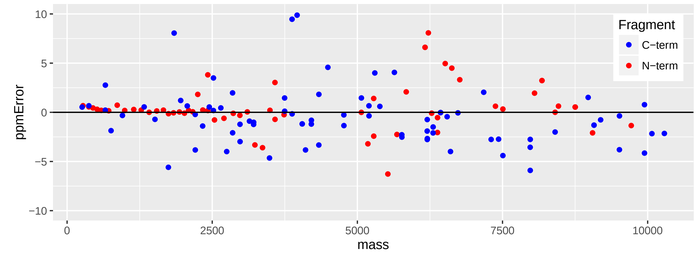



## Deconvolution Accuracy and Efficiency

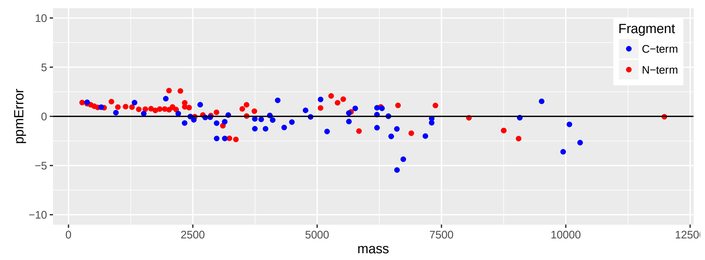
The deconvolution software was redesigned to reduce computational requirements, helpful for processing large LC-MS/MS results. The new version also tends to return less false positive matches.

The selected data is the ECD spectrum of the glycoprotein ribonuclease B. The figures were generated by importing the BUPID Top-Down results into the BTDR package.

```
library(BTDR)
fragment.mass.error(read.bupid("RNaseB-decon2015.results"))
```




```
fragment.mass.error(read.bupid("RNaseB-decon2016.results"))
```




Reducing the false positives returned in the peak list result in fewer incorrect matches during fragment assignment. This is particularly true when assigning modified fragments due to the larger search space.

Improvements to the deconvolution results are achieved using several approaches:

- Combining the benefits of profile and centroid mode data. Profile mode is used to sum spectra. The merged spectrum is then centroided internally for faster and more consistent processing
- Isotopic cluster detection was redesigned using a peak picking method rather than rely on *m/z* ranges. This allows for more complete results due to reduced interference from unrelated peaks.
- Scoring isotopic cluster assignments is now done with an approach similar to the one used by MS-Deconv, which has proven to give more conservative results.

## Search Scoring

BUPID Top-Down uses a sequence tag approach for identifying proteins during a database search. Tag matches are scored based on the relative rarity of the tag in the set of candidate proteins. This serves to filter out weak matches containing only ambiguous tags.

… V **H L T** P **V V K** S … **Theoretical Protein A**

… V **H L T** P V E K S … **Theoretical Protein B**

… M **H L T** P V E K S … **Theoretical Protein C**

The tag *HLT* will not add as much to the overall score as *VVK* due to how frequently each sequence of amino acids is seen in the database. Tags with amino acid sequences that occur in many proteins will contribute less to the overall score than sequences with fewer candidates.

## Conclusions

- Theoretical isotopic distribution models can be configured to more closely match those present in the sample.
- The deconvolution algorithm used in BUPID Top-Down now generates peak lists more efficiently and with fewer false positives, reducing the possibility for error in further processing.
- Database search using sequence tags now calculates scores based on tag frequency, reducing weak matches in results.
- BTDR, the R package used for reviewing the BUPID Top-Down results and generating figures, can be installed using the devtools package:

```
devtools::install_github("heckendorfc/BTDR")
```

- The BUPID Top-Down web service can be freely accessed at: http://bupid.bumc.bu.edu/

## Acknowledgements

This project was funded by:

NIH-NHLBI contract HHSN268201000031C and NIH grants P41 RR010888/GM104603, R21 HL107993, S10 RR020946, S10 OD010724, and S10 RR025082.

Special thanks to the members of the CBMS laboratory for their help.