BUPID-Top-Down: Database Search and Assignment of Top-Down MS/MS Data

Weiwei Tong, Roger Théberge, Giuseppe Infusini, David H. Perlman, Catherine E. Costello, Mark E. McComb
Cardiovascular Proteomics Center, Boston University School of Medicine, Boston, MA.

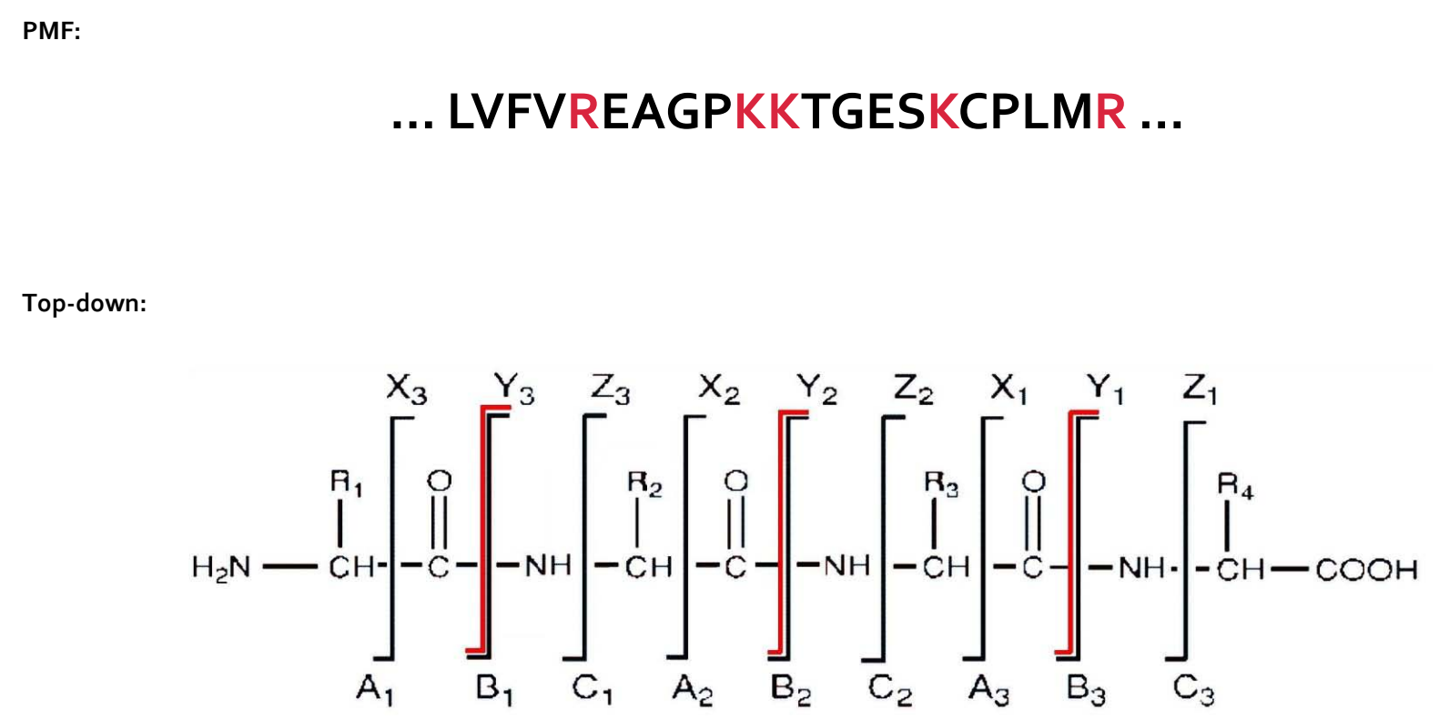
Overview

A major goal of proteomics is to identify and characterize all proteins expressed in cells under various conditions. Mass spectrometry (MS) has become popular for identification of proteins in high-throughput proteomics research. We previously reported a software application, BUPID, of database searching using peptide mass fingerprinting (PMF) data. The algorithm utilizes a log-likelihood ratio score to discriminate correctly assigned peaks from incorrectly assigned ones.

In this poster, we describe a new algorithm/application derived from BUPID to assign product ions in top-down MS/MS spectra. The software can be used to analyze spectra obtained with various fragmentation methods including CID, IRMPD, ECD, ETD and EDD. It also identifies internal fragments, side-chain losses, neutral-losses and post-translation modifications (PTM). For unknown proteins, **BUPID-Top-Down** searches through a protein sequence database for the best match and uses a heuristic model to expedite the calculation.

Top-down proteomics has emerged as a technique that preserves labile post-translational modifications and offers full protein sequence coverage. Originally developed on Fourier-transform ion cyclotron resonance (FT-ICR) instruments, top-down experiments now can be carried out on many newer and cheaper instruments as the sensitivity of mass analyzers improved drastically.

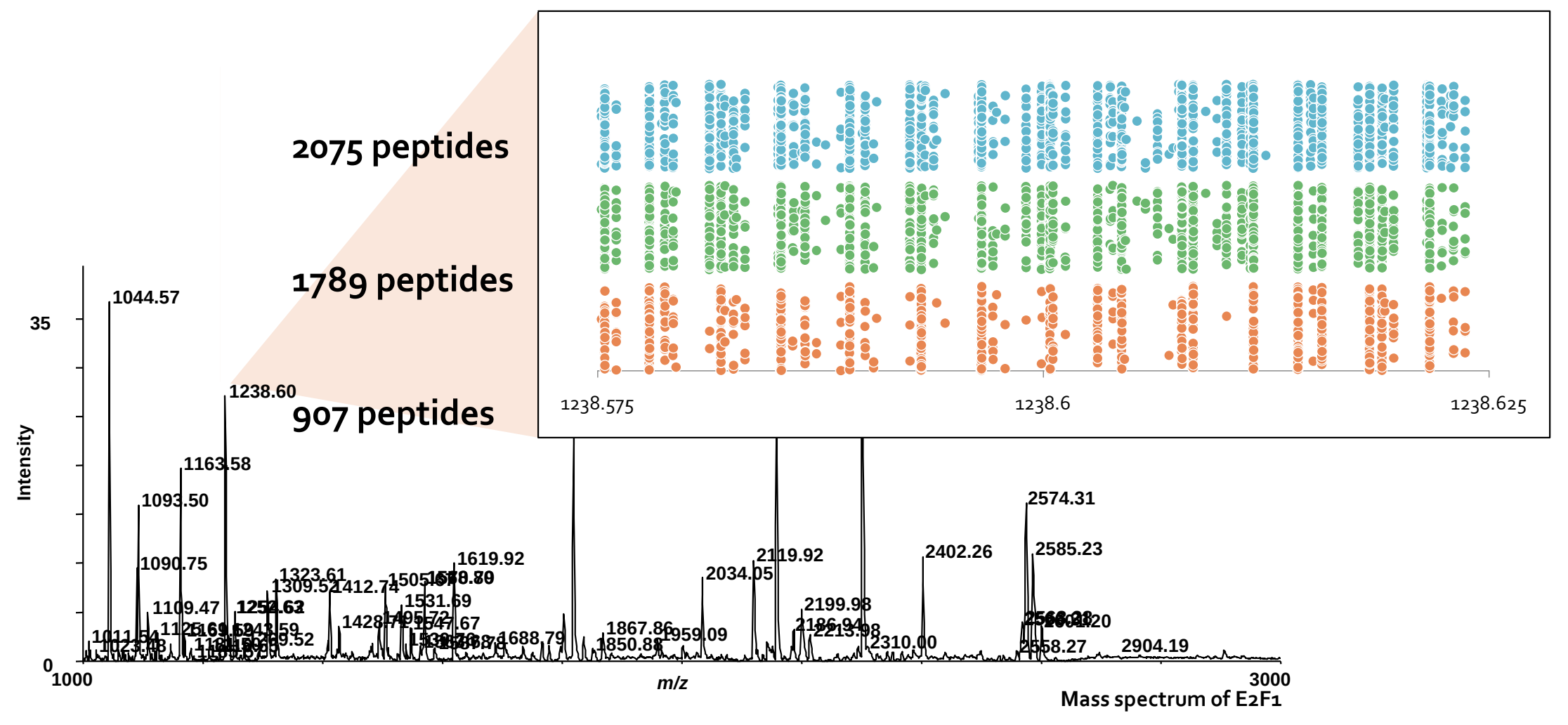
Technically, top-down is MS/MS carried out on the intact protein. However the principle of top-down data interpretation is more similar to that of PMF than MS/MS.


PMF:
… LVFVREAGPKKTGESKCPLMR …
Top-down:
X3 Y3 Z3 X2 Y2 Z2 X1 Y1 Z1
R1 R2 R3 R4
O O O
H2N CH C NH CH C NH CH C NH CH COOH
A1 B1 C1 A2 B2 C2 A3 B3 C3

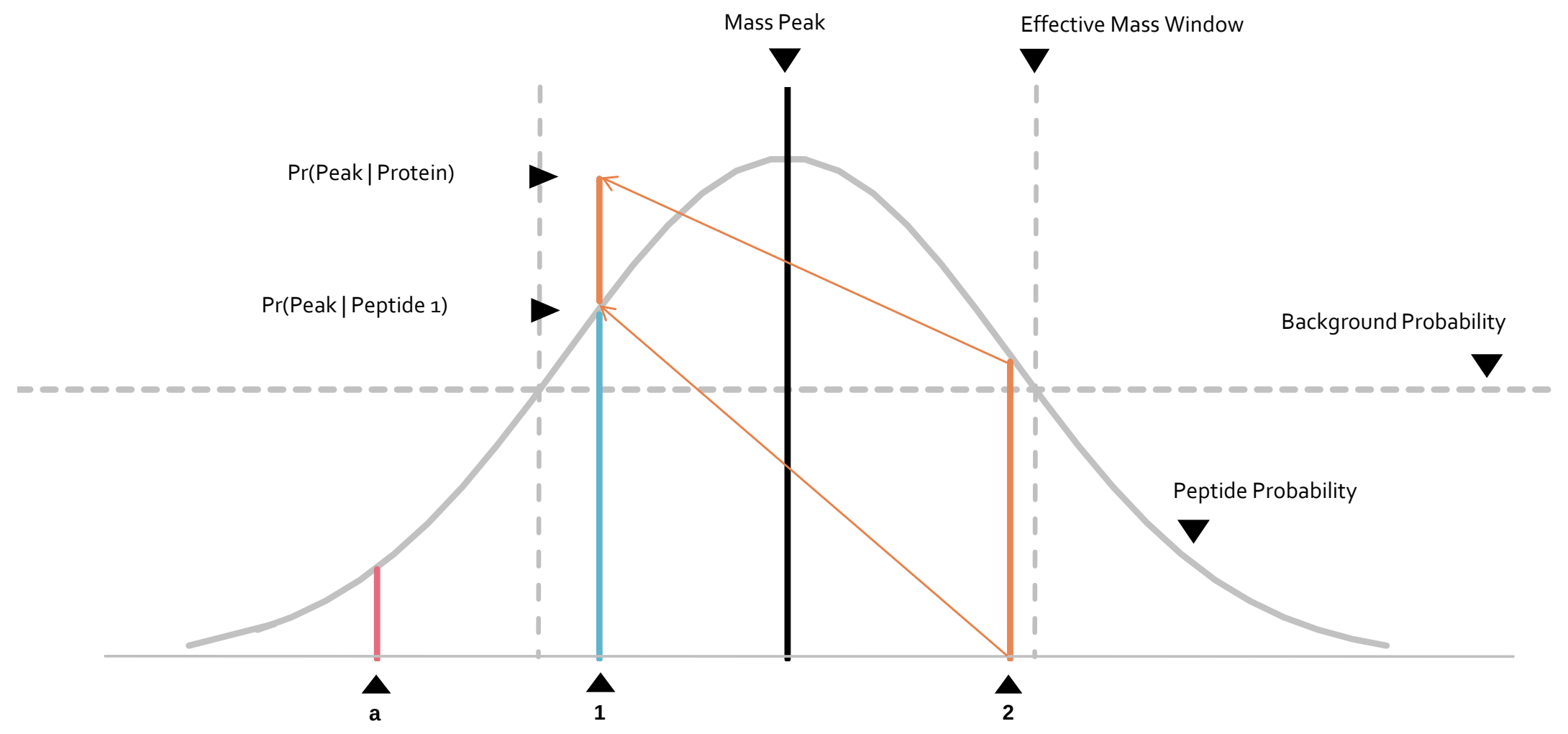
PMF database search engines (e.g. Mascot or BUPID) take the peptide masses measured by mass spectrometers and query them against masses of hypothetical peptides derived from a sequence database (e.g. SwissProt). Theoretical peptides are assigned to peaks based on mass similarity. In top-down data analysis, peptides are substituted with ions; and everything else can be borrowed from PMF search engines directly.

Algorithm

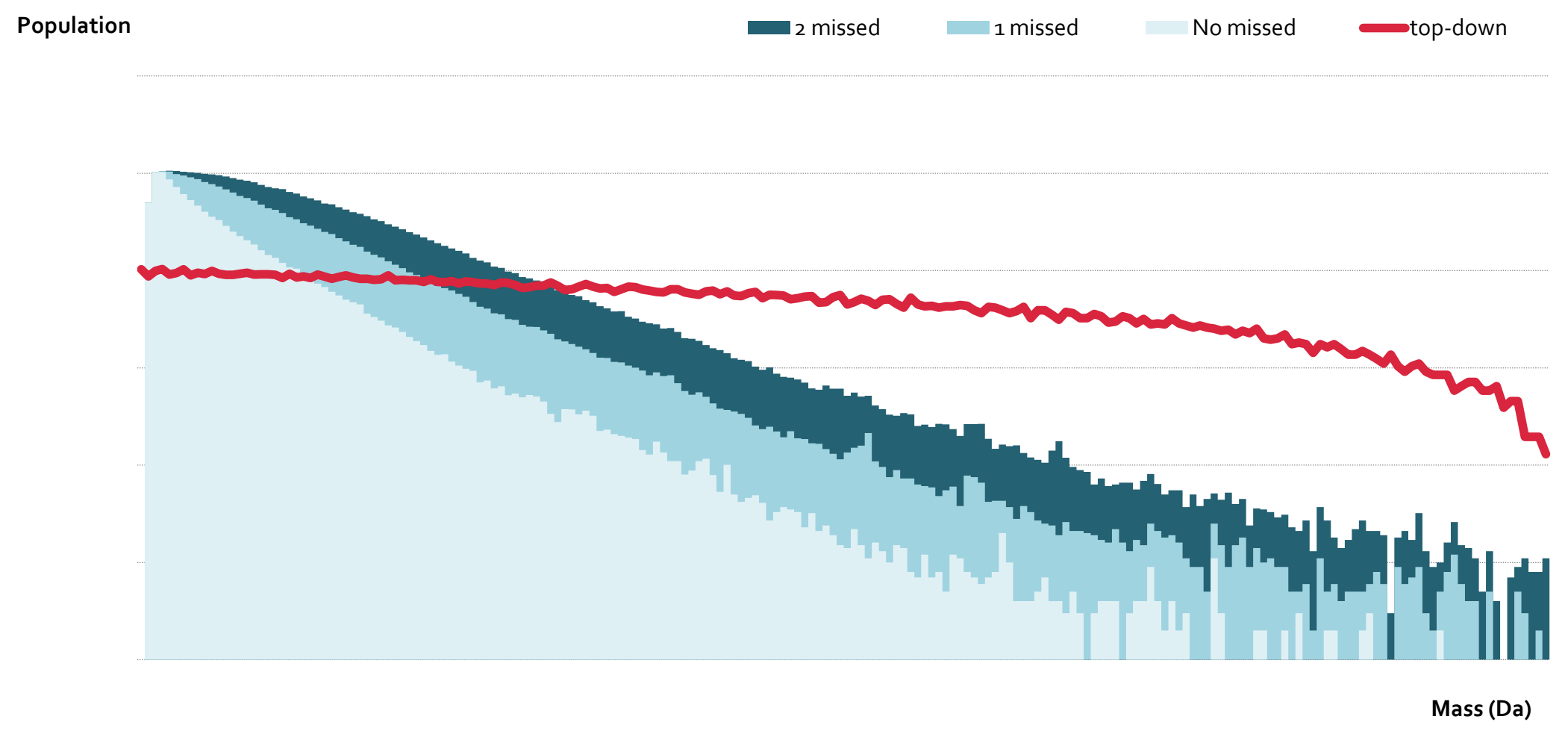
The major challenge in PMF search or top-down search is to decide which peak matches with which peptide or ion from which protein. Without a proper scoring function, the match process (by eyeballing, for example) could be labor intensive.



In BUPID, the background probability is used as the yardstick to assess the goodness-of-fit between the peak and the protein. If the mass difference between the peak and the ion is less than the mass tolerance decided by the background, the algorithm marks the pair as a match.



Note that the mass distribution of tryptic peptides and the mass distribution of product ions in top-down experiments are very different. Therefore the background noise level is different, too.



In the first step of the search process, BUPID-Top-Down constructs a library of ions. The library contains all daughter ions that can be derived from the original protein sequence under specific cleavage rules. For example, ECD and ETD induce primarily *c*/*z* fragmentation; CID induces *b*/*y* fragmentation; IRMPD induces *b*/*y* and *a* fragmentations. Correspondingly, ions with the original protein N-terminus (*a*•, *b*, *c*-ions) or C-terminus (*x*, *y*, *z*•-ions) are generated with one cleavage on the protein backbone. Secondary dissociations (and higher-order dissociations, for that matter) are also taken into consideration. These "Internal fragments" are constructed by making two cleavages on the protein backbone. For instance, ion *b*12 can decay into a shorter b-ion b4 and an internal fragment from residue 5 to 12, denoted as *i*5-12. BUPID-Top-Down searches for both primary fragmentation ions and internal fragments.

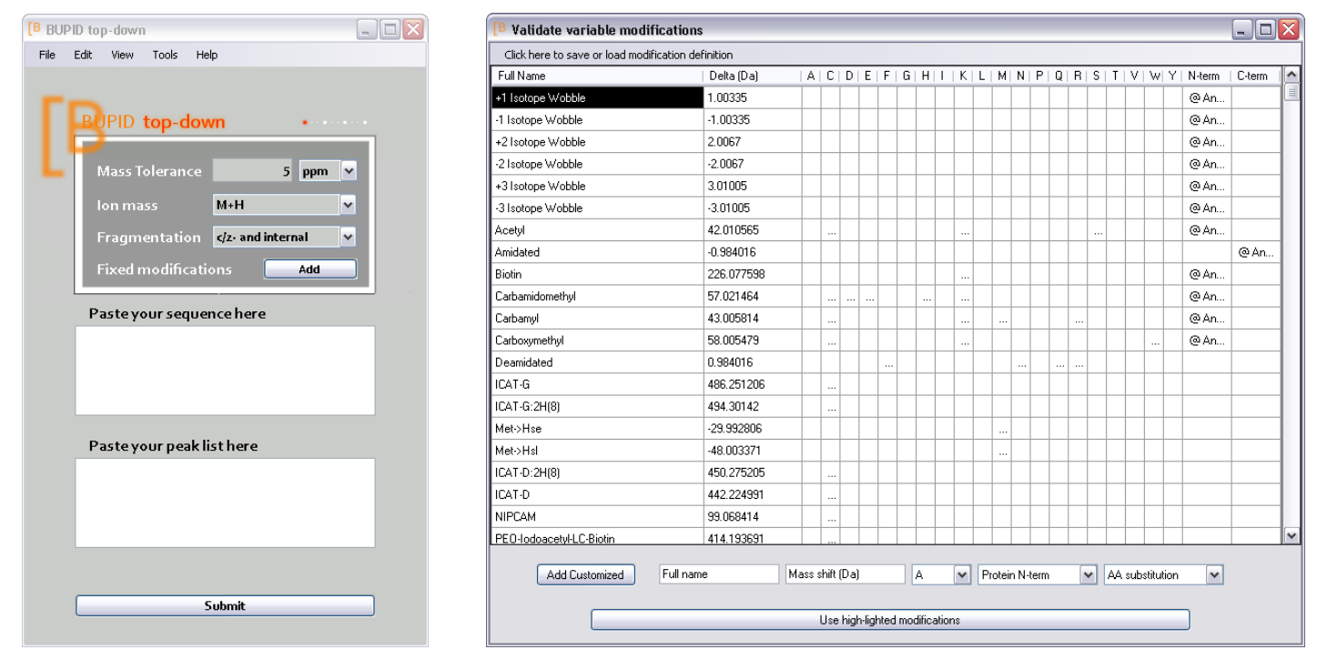
Since the random background is calculated separately for primary fragments, internal fragments and fragments with modifications. It follows that if a *b*-ion and an *i*-ion share the same mass (one instance is when the *b*-ion and the *i*-ion are anagrams), the *i*-ion assignment will have a larger p-value (less significant) than the *b*-ion assignment, since there may be more *i*-ions than *b*/*y*-ions. Specifically, the number of primary fragments is O(n); and the number of high-order fragments is O(n^2), where n is the length of the protein sequence. P-values are calculated using the modeled histogram.

BUPID handles two types of PTMs. Fixed modifications can be assigned to residues on which the presence is known based on a priori information (e.g. biotinylation). Adding fixed modifications to a residue is equivalent to using a different mass for the residue in successive searches. Variable modifications are PTMs that are suspected to be present in the sample. The program searches for all possible combinations of modifications that can be applied to each ion. Definitions of variations and modifications can be loaded from Unimod XML files. In addition, customized modifications can be defined on the run by providing BUPID with the mass shift and residue and/or terminus specificities.

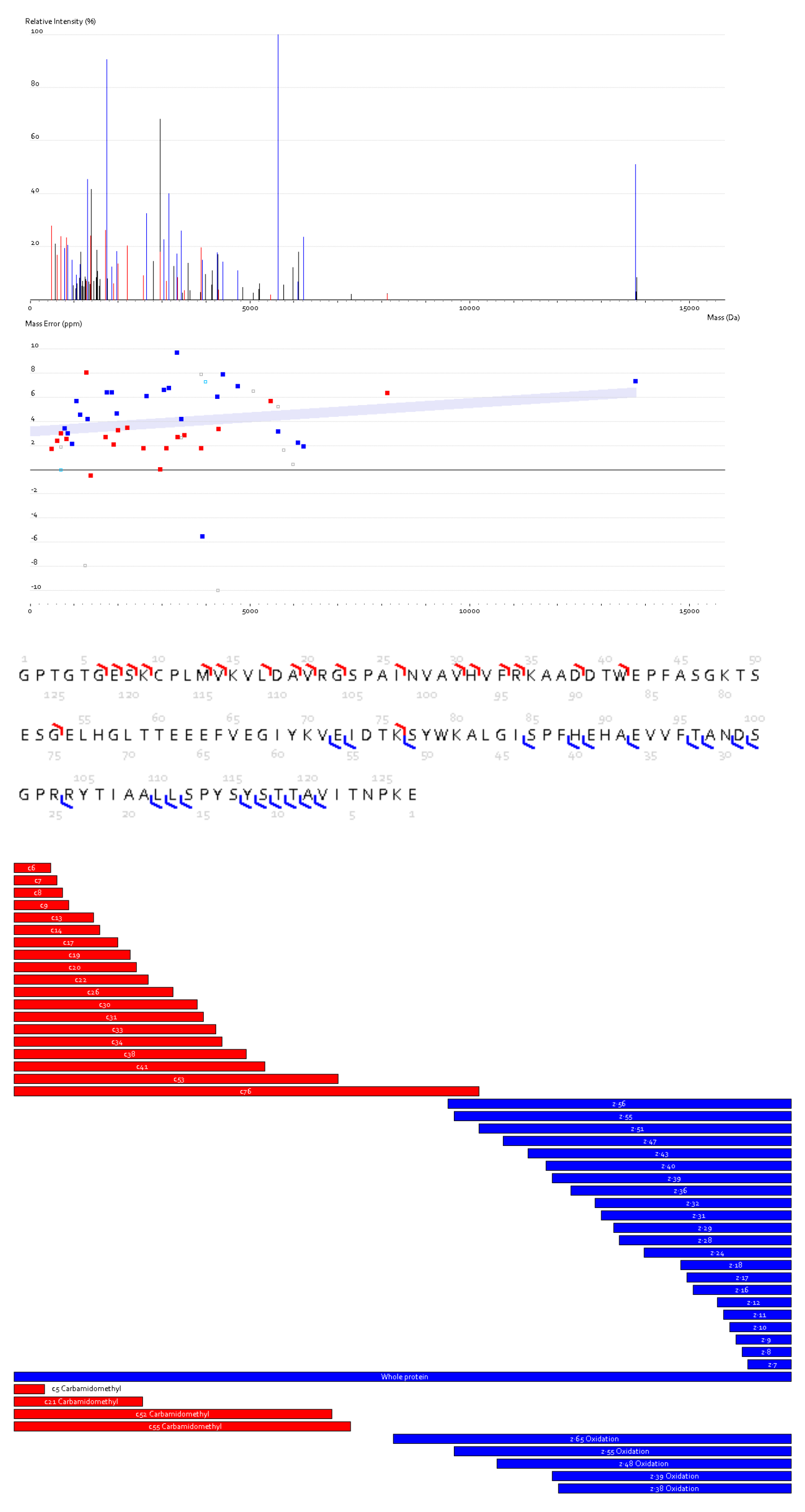
Results

We analyzed the results for example proteins analyzed with top-down MS/MS using CID, IRMPD and ECD obtained under optimal conditions. Top-down experiments provided 100% sequence coverage. Limited sidechain-losses and neutral-losses were observed in CID and IRMPD experiments. Identification of primary fragment ions (*b*,*y*,*c*,*z*•) and sidechain-losses/neutral-losses were essentially identical with results from other search engines. We also discovered multiple PTMs and identified abundant internal fragments (*i*-ions). CID results obtained for a target protein showed that mass errors of 85% of *b*-ions and *y*-ions were normally distributed within ±2.5 ppm (after re-calibration within BUPID). In comparison, 54% of assigned *i*-ions fell within the same range, the rest uniformly distributed within the mass error search tolerance of ±20 ppm (potentially false positives); 80% of *i*-ions within ±2.5 ppm shared one or two termini with identified *b*-ions or *y*-ions. These *b*/*y* ions were thus recognized as more reliable identifications than other *i*-ions. Additional CID and IRMPD data confirmed this trend.

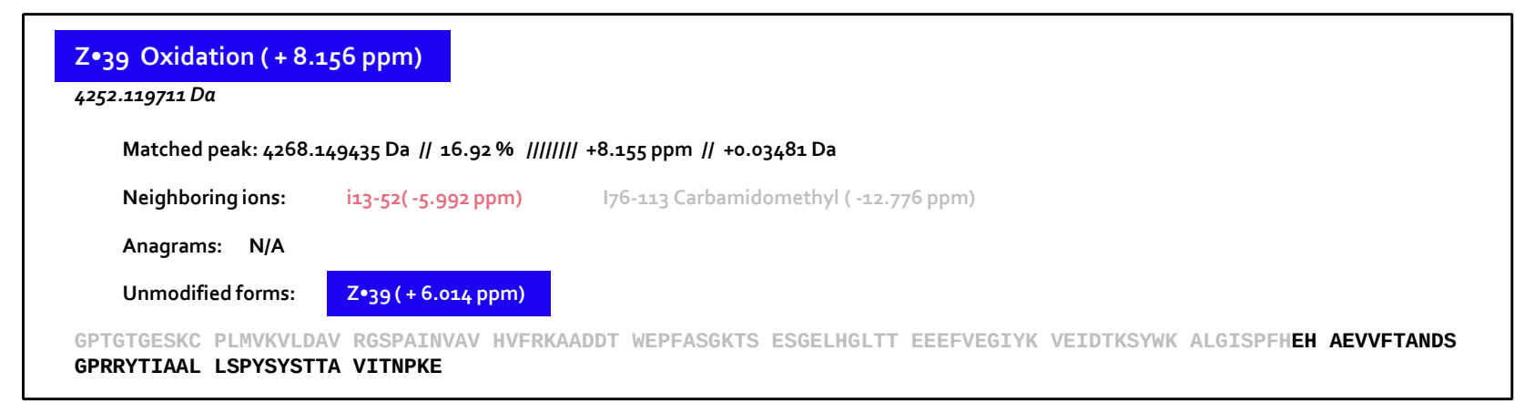
Graphic user interface of the Windows program:



Graphic display of search results



Report card of *z*•39 ion with oxidation:



Z•39 Oxidation (+ 8.156 ppm)
4252.119711 Da

Matched peak: 4268.149435 Da // 16.92 % //////// +8.155 ppm // +0.03481 Da

Neighboring ions: i13-52(-5.992 ppm) I76-113 Carbamidomethyl (-12.776 ppm)

Anagrams: N/A

Unmodified forms: Z•39 (+ 6.014 ppm)

GPTGTGESKC PLMVKVLDAV RGSPAINVAV HVFRKAADDT WEPFASGKTS ESGELHGLTT EEEFVEGIYK VEIDTKSYWK ALGISPFHEH AEVVFTANDS
GPRRYTIAAL LSPYSYSTTA VITNPKE

This project was funded by NIH/NHLBI N01 HV-28178 and NIH/NCRR P41 RR10888 and S10 RR020946. We thank the members of the CBMS for testing BUPID-Top-Down.