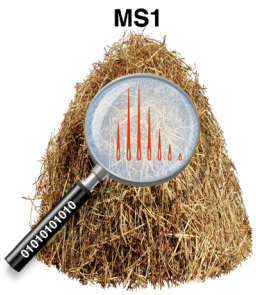
J. Am. Soc. Mass Spectrom. (2015) 26:2133–2140
DOI: 10.1007/s13361-015-1229-4

RESEARCH ARTICLE

Software Analysis of Uncorrelated MS1 Peaks for Discovery of Post-Translational Modifications

Bruce D. Pascal,[1] Graham M. West,[2] Catherina Scharager-Tapia,[2] Ricardo Flefil,[2] Tina Moroni,[2] Pablo Martinez-Acedo,[2] Patrick R. Griffin,[3] Anthony C. Carvalloza[1]

[1]Informatics Core, The Scripps Research Institute, Jupiter, FL 33458, USA
[2]Proteomics Core, The Scripps Research Institute, Jupiter, FL 33458, USA
[3]Department of Molecular Therapeutics, The Scripps Research Institute, Jupiter, FL 33458, USA



Abstract. The goal in proteomics to identify all peptides in a complex mixture has been largely addressed using various LC MS/MS approaches, such as data dependent acquisition, SRM/MRM, and data independent acquisition instrumentation. Despite these developments, many peptides remain unsequenced, often due to low abundance, poor fragmentation patterns, or data analysis difficulties. Many of the unidentified peptides exhibit strong evidence in high resolution MS^1 data and are frequently post-translationally modified, playing a significant role in biological processes. Proteomics Workbench (PWB) software was developed to automate the detection and visualization of all possible peptides in MS^1 data, reveal candidate peptides not initially identified, and build inclusion lists for subsequent MS^2 analysis to uncover new identifications. We used this software on existing data on the autophagy regulating kinase Ulk1 as a proof of concept for this method, as we had already manually identified a number of phosphorylation sites Dorsey, F. C. et al (J. Proteome. Res. **8**(11), 5253–5263 (2009)). PWB found all previously identified sites of phosphorylation. The software has been made freely available at http://www.proteomicsworkbench.com.

Keywords: Post-translational modifications, MS1 data analysis, Peptide identification, Mass spectrometry, Computational proteomics, PTMs

Received: 16 April 2015/Revised: 29 June 2015/Accepted: 30 June 2015/Published Online: 12 August 2015

Introduction

A primary goal of proteomics is to confidently identify, sequence, and/or quantify all peptides from a complex mixture in a high throughput manner. This task is conventionally addressed with a "bottom up" strategy using liquid chromatography tandem mass spectrometry (LC-MS/MS). However, this strategy often misses many of the peptides that hold important biological relevance. Post-translational modifications (PTMs) of proteins are often not found, yet known to regulate a myriad of cellular mechanisms, and identification of these modifications leads to a better understanding of the components of the signaling networks that modify proteins. Signaling networks are found to be increasingly complex, as evidence mounts that signaling events require combinatorial modifications on a single protein working together to modulate the protein's function. Mass spectrometry-based proteomics has become a fundamental tool in the identification of proteins and PTMs from both complex cellular systems and simple protein mixtures [2–5]; however, identification of combinatorial modifications remains an extremely important and challenging problem. Strong evidence of PTM peptides is readily available in high-resolution MS^1 data and can be mined to substantially support the identification process. We believe this approach has been underutilized and describe a computational method to complement existing techniques in this important task.

The conventional technique used for the identification of peptides and proteins has been the "shotgun proteomics" approach, in which the protein(s) of interest is denatured and digested (usually with the enzyme trypsin), and the resulting peptides are separated using liquid chromatography. This is followed by tandem mass spectrometry (MS^2) using data-dependent acquisition (DDA), in which a survey scan is initially run and then ions are selected for fragmentation based on highest abundance and/or other predefined criteria. MS/MS spectra are generated and search engine software is used to

Correspondence to: Bruce Pascal; *e-mail:* bpascal@scripps.edu

identify the peptides/proteins by comparing the observed spectrum to a theoretical spectrum for a given peptide sequence [6–26]. The operation of this method is straightforward, as it is not necessary to know which proteins are being targeted in advance, and performs well in common proteomic studies looking specifically at which proteins are present.

Despite the benefits, there are limitations to DDA, including limited dynamic range, reproducibility, and a bias toward abundant peptides. Relying solely on MS^2 data from DDA for the identification of peptides is often problematic, especially those that are post-translationally modified. Successful identification by tandem mass spectroscopy is only possible if a peptide is selected for fragmentation and PTMs typically exist on only a portion of a protein species, making them substoichiometric. In many cases, phosphorylation has been detected at less than 1% of total protein concentration [27]. Typical experiments rely on peak intensity to select ions for fragmentation, so it is routine for modified species to be missed using this approach [24]. Furthermore, it is common to use collision-induced dissociation (CID) to achieve peptide sequencing and fragmentation related nuances, which result in MS^2 spectra that are difficult to decipher because of the peptide bond break localization [15, 17].

Several bioinformatic approaches have been developed to address the difficulty of PTM identification. The vast majority of these strategies are based on analysis of MS^2 data using a peptide spectrum match (PSM) algorithmic approach, in which the observed spectrum is compared with a theoretical spectrum for a given peptide sequence [6–26]. De novo sequencing is another method for identification of PTMs using MS^2 data, in which the mass distances between peaks in the mass spectrum are matched to known residue masses, allowing the sequence to be derived solely from the spectral data. This method has been previously described [28–35], and while it is somewhat limited to high quality data and has lower throughput, it has less of a computational limitation compared with the database search approach. These strategies have been used almost exclusively, as it is data provided by the fragmentation of the MS^2 spectra and allows the determination of the site of modification and the sequence of the modified peptide. As a result, the development of software tools for the identification of PTMs from MS^2 data has been an extremely active area of research.

More recently, investigators have adopted unrestrictive search approaches [36–39], where all possible modifications are searched at once. When combined with conventional protein database search strategies, these methods are limited by the number of simultaneous variable peptide modifications that can be searched. This is due to the exponential nature of the problem and the restrictions of existing computational capacities. Despite the doubling of computational advances approximately every two years for more than half a century [40, 41], large numbers of modifications in combination result in a large number of permutations. This leads to a problem that is often prohibitively large or NP-complete [42], especially in cases where unrestrictive searches are run.

The application of filtering to reduce the amount of data that needs to be considered has become a central approach for addressing this issue. This seems to be a reasonably effective strategy for the analysis of MS^2 data and several variations on this theme have been described [43–48]. Some have tackled this problem by using parallelization strategies [49]. In addition, unrestrictive searches are prone to high false discovery rates, requiring each search result to be examined. Tanner et al. [50], addressed this issue with PTMFinder. Database search approaches have been described that iterate the search multiple times to significantly reduce the number of possible sequences considered. The result is that searches are accomplished more quickly, with fewer computational resources and with more stringent parameters [51, 52]. The large amount of research in this field outlines the importance of the problem and although significant advances have been made, there is still much to be accomplished to address the challenges in this domain.

Recently, instruments have been able to function in data independent acquisition (DIA) mode in which MS/MS data are generated from all of the sample precursors. There are several different methods, but they are all essentially accomplished by fragmenting all ions within preset *m/z* and retention time windows. In this way, the precursor ion selection is not biased and low abundant peptides will be fragmented, increasing the number of identifications and sequence coverage. A downside to this method is that isolation windows can often result in co-fragmentation of precursor ions producing complex multiplexed spectra that are difficult to interpret. Several approaches have partially addressed this data analysis step [53, 54]; however, it remains difficult.

Another method has been described in which software builds a database of algorithmically selected peaks and directly interfaces with the instrument acquisition process. In this manner, it is able to generate inclusion lists using different peaks from previous runs, effectively increasing MS/MS sequence coverage and the number of proteins identified, while accomplishing this in a fully automated manner [55]. Despite improvements in instrumentation, techniques, and efforts in data analysis, we are still falling short of identifying all peptides and their isoforms in a sample, particularly in complex mixtures. This continues to be highlighted by the fact that many biologically important PTM identifications are missed. In light of this, methods to improve this area are needed. We believe that many of these issues may be addressed by complementing MS^2 identification with the analysis of MS^1 data, which is now commonly available with high mass accuracy, particularly in support of PTM identifications.

Some efforts have been made in this area. Multiple LC-MS runs have been aligned and quantitative information has been measured for the purpose of targeting post-translationally modified peptides for fragmentation [56–58]. Other tools compare the results from a database search with MS^1 data to correlated unmatched spectra [59, 60] or to improve the scoring of search results [61, 62]. Although these efforts do not focus solely on the PTM identification problem and do not use the MS^1 interrogation as a central and initial method, they do highlight the

value of the spectral data available in MS^1. To date the only work to specifically search for PTM peptides using the information present in MS^1 data was presented recently [63, 64].

Using a conventional DDA approach, although it may not be possible to precisely determine the presence of a particular modification site through examination of MS^1 data, it is possible to quickly and conclusively determine its absence. For every confirmed MS^2 identification there will always be evidence of the peptide at the MS^1 level. More importantly, if a modified peptide is present, this strategy quickly provides a short list of possible permutations associated with the isotopic signature. In complex multiple modified examples, there are often permutations within the same peptide that share a common chemical formula. If searched with MS^1, only the unique combinations of formula and charge state need be considered, as they can be expected to result in the same isotopic distribution and monoisotopic *m/z* within the parts-per-million (ppm) error of the instrument. High scoring results can either be selected for subsequent targeted MS^2 experiments or confirmed if fragmentation spectra are available. The smaller subset of permutations that need to be considered in MS^1 data will theoretically lead to a reduction in computational expense and faster search times. Additional benefits of this strategy may include increased sequence coverage over MS^2 and that MS^1 data is readily available, as it is a prerequisite to MS^2 data, allowing this approach to be applicable to older data sets.

We propose a computational strategy for the identification of protein post-translational modifications, which initially interrogates MS^1 data and compares the theoretical isotopic distributions against the experimental data to generate a list of possible matches for examination. A goal of this strategy is to initially reduce the number of peptide sequences taken into account, in order to expedite subsequent MS^2 analysis. In this manner, it contains some similarity to previously described iteratively refined MS^2-based search methods [51, 52], but instead uses MS^1 interrogation as the first step. This approach is amenable for an automated computation and has been implemented in a software framework to resolve some of the key limitations of the MS^2 based methods: examination of non-fragmented peaks, identification of protein modifications that are associated with poor fragmentation patterns and computational algorithmic enhancements.

Methods

The workflow for these methods was implemented in the Java programming language and took the following as input: Thermo Scientific raw data files, the sequence of the protein in FASTA format [65], a list of modifications, and a list of parameters described, which include enzyme, number of missed cleavages, as well as limits related to retention time, peptide length and ppm tolerance. An in-silico enzymatic digest of the protein was performed using variables specified in the input, to create the initial list of possible peptides from the protein. All possible modified permutations for each peptide were determined and each was searched against a distinct combination of sequence and charge states. This usually results in a much smaller number of permutations searched than would be needed for an MS^2 search. For example, a search at the MS^2 level for a singly modified phosphorylation site on the RORg peptide LISSIFD would need to search LI**pS**SIFD and LIS**pS**IFD, both of which are comprised of the same chemical formula as they both contain one modified serine residue. When using the MS^1 approach, only one of these forms would be preserved for the subsequent search.

Importantly, mass spectrometric data does not display the mass of the peptide but, instead, the mass-to-charge ratio of the peptides. Therefore, each of the calculated permutation masses was converted to *m/z* ratios within the range of expected charges. A user-defined *m/z* range filter was then applied, typically between 350 and 2000 *m/z* units. The resultant *m/z* that fell between +7 to –7 ppm (as defined by user input) were used to create single ion chromatograms (SIC) from the mass spectral data. The chromatographic peaks were then used to hone in on the possible retention times of the peptide and to determine the presence of the peptide and best retention time ranges in the MS^1 data using methods that have been described previously [63, 66, 67].

Each peptide is assigned a score that is used to rank the confidence of the match between the expected and observed spectra. The score is obtained by using primarily a least squares approach combined with mass error and works as follows: the retention time ranges provided by the SIC are used in conjunction with a sliding window strategy utilizing a preset expected elution range. The averaged mass spectrum for retention time range is generated in profile mode and undergoes a peak picking step in which all spectral peaks not consistent with peptides are removed. The software then employs a custom version of the software Qmass [68, 69] to generate theoretical isotopic distributions for the peptide ion of interest, and each theoretical peak is assigned to a corresponding experimental peak so long as the intensity exceeds 4% of the most abundant peak. The score is determined using a least squares approach comparing the theoretical peaks to the observed and is calculated as follows:

$$\sqrt{\frac{\sum_i^N (T_i - A_i)^2}{N}}$$

where T_i is the relative abundance of the *i*th peak of the theoretical distribution, A_i is the relative abundance of the *i*th peak of the observed distribution, and N is the number of peaks. The lowest score is assumed to be the best match for the peptide ion of interest, and results with a reported score exceeding the user-defined input threshold score and meeting the criteria of the input parameters are preserved for interrogation. Once the processing is complete, the results presented in the software interface are output in comma separated values (CSV) format for straightforward review. Once the list of candidate peptides has been validated, they can be compared against the MS^2

identifications. Masses of peptides not previously identified are then used to build inclusion lists for subsequent MS^2 analysis, often producing new identifications and increasing sequence coverage. An overview of this software approach is further described in Figure 1.

We tested the software using the data-set from a previous publication [1], in which several PTMs were discovered by calculating the theoretical peptide mass and manually extracting and validating the peaks using the Thermo Scientific Qual Browser. The sample preparation and MS^2 analysis was accomplished as follows: the kinase Ulk1 was prepared as an NTAP tagged construct and expressed in HEK293T cells; the resultant protein was purified using streptavidin binding resin and eluted to over 80% purity; the protein was digested, as previously described, and subjected to LC-MS/MS analysis using a Thermo Scientific Orbitrap.

To further validate and compare the software with MS^2 identifications from another software, the protein bovine serum albumin (BSA) was searched with both Mascot 2.3 (Matrix Science) and Proteomics Workbench. The BSA sample was prepared such that all cysteines were carbamidomethylated (+57), digested using trypsin, and fragments +2 and above were targeted for fragmentation. Data-dependent selection of the 10 most abundant ions was used for HCD. The resulting raw file was searched using Mascot with the following parameters: cysteine carbamidomethyl fixed modification, two missed cleavages, and digestion enzyme trypsin. Mass tolerance for the precursor ion was set to 50 ppm; mass tolerance for the fragment ions was 0.5 Da. To ensure that no identifications were precluded, the subsequent search was conducted with a decoy database, and identifications were considered valid if the results were above the determined FDR and had an ion score cutoff >10. The search using PWB used the following parameters: max peptide length 20, mass tolerance 10 ppm, fixed modification of carbamidomethyl cysteine, and two missed cleavages.

Results and Discussion

The 16 sites of phosphorylation that were previously identified in Ulk1 were all detected using the proposed computational workflow. A list of the peptides that were found and the associated scores are listed in Table 1. The manual detection of the PTMs in this study was accomplished in 3 wk, as the goal was to discover and verify new sites that needed to be searched manually, by generating selected ion chromatograms (SICs) for each peptide in both modified and unmodified form, and manually using the native instrument software (Thermo Scientific Qual Browser) to validate correlating spectra. Using the software approach, automated detection was completed in 3 h using a desktop PC with an Intel i7 processor with 16GB of RAM.

Concerning the BSA search, Mascot identified 56 non-duplicate peptides with sequence coverage of 56%. PWB scored 68 peptides with a sequence coverage of 66%. PWB found all of the peptides identified in Mascot, plus an additional 12. Of these 12, four were +1 peptides, which, as expected, were not found in the Mascot results because +1 ions, generally

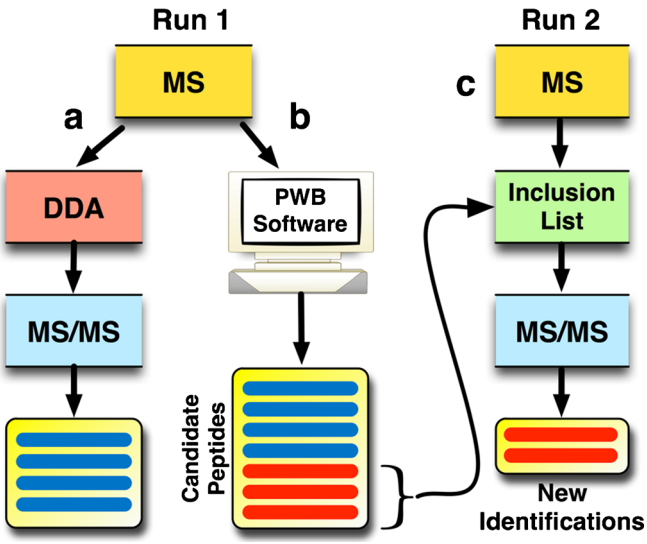


Figure 1. Software overview. **(a)** The conventional DDA approach selects a subset of peptides for fragmentation based on abundance. The resulting MS/MS are then subjected to search algorithm to identify peptides. **(b)** The same raw file is searched at the MS level using Proteomics Workbench software to find all possible candidate peptides. The peptides that were not identified using the initial approach are used to create an inclusion list. **(c)** In a subsequent run, the inclusion list masses are fragmented and the resulting MS/MS is searched, often resulting in new identifications

contaminants, were not fragmented (PWB only has a maximum charge parameter and cannot disregard +1 peptides). That left eight additional peptides that were found using the software. An inclusion list was made with these eight peptides and we targeted them for MS^2 fragmentation. Of these eight peptides, one additional peptide was confirmed at the MS^2 level by Mascot with the same parameters as previously, increasing sequence coverage by 4% (Figure 2). There results show that the software is comparable with existing established MS^2 software, and is able to search for modifications. This also shows that using this approach with targeted inclusion lists can increase the number of peptides over a conventional single runge.

To illustrate the difference between the number of permutations considered between MS^1 and MS^2 searches, the sequences of MKK5, with a sequence length of 448 amino acids, and a larger protein Ulk1, with a sequence of 1050 amino acids, were examined with conventional search parameters. The MS^2

Table 1. Number of Permutation Considered MS^1 Versus MS^2 for Proteins MKK5 and Ulk1. Using Conventional Search Parameters, the MS^1 Approach Becomes More Essential as the Number of Modifications Searched Increases

	MKK5 (448 aas)		Ulk1 (1050 aas)	
	MS1 permutations	MS2 permutations	MS1 permutations	MS1 permutations
2 Missed cleavages	1037	1037	772	772
Oxidized M	1550	2051	1114	6468
Phospho STY	30,674	47,355	34,198	374,550
GIcNAc ST	74,651	179,712	86,264	1,467,878

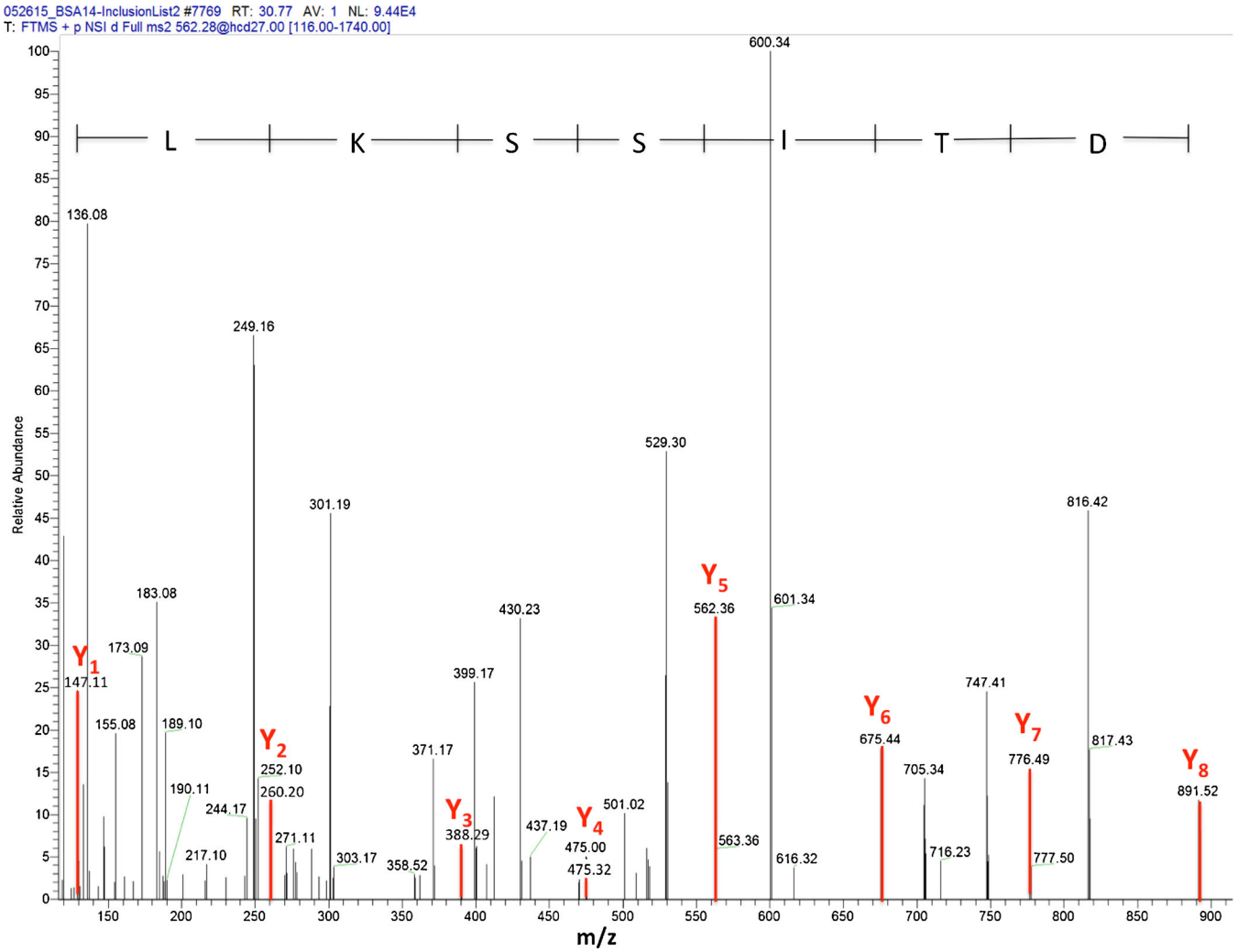


Figure 2. MS^2 spectra of inclusion list peptide: MS^2 spectra for the bovine serum albumin peptide YICDNQDTISSKLK (+3) are shown above. Using the conventional DDA approach, this peptide was not identified in the initial MS^2 data. Using software analysis of the MS1 data, strong evidence of this peptide was discovered, and an inclusion list was generated for directed MS. The peptide was identified in the subsequent MS^2 analysis using the same settings, yielding increased sequence coverage

Table 2. Ulk1 Peptides. A comprehensive list of all the identified phosphorylated peptides from Ulk1 identified previously, along with the corresponding theoretical masses, observed masses, and associated ppm errors [3]. The peptides listed above were confirmed manually by calculating the exact mass and "chro-ing" out the peaks using the Thermo Scientific Qual browser over a period of three weeks. Using the software, we were able to find and validate all of these peptides in three hours. Phosphorylation sites are indicated in lower case

Sequence	Monoisotopic mass	Observed *m/z*	Charge	ppm Error
HENIVALYDFQEMANsWLVMEYCNGGDLADYLHTMR	4403.9037	1101.9873	4	3.72
YMAPEVIMsQHY	1547.619	774.8167	2	0.13
APFQAsSPQDLR	1395.6184	698.8179	2	2
TLTSPADAAGFLQGsR	1670.7665	836.3912	2	0.84
IEQNLQsPTQQQTAR	1820.8418	911.4273	2	0.99
SGsTsPLGFGR	1224.4577	613.2377	2	2.61
ASPsPPSHTDGAM(ox)LAR	1689.7182	845.8644	2	2.36
ASPSPPsHTDGAM(ox)LAR	1689.7182	845.8644	2	2.36
VPsPQGADVR	1104.4965	553.2563	2	1.45
SPLPPILGsPTK	1285.6683	643.8423	2	1.4
GGGASSPAPWFTVGsPPSGATPPQSTR	2645.2486	882.7586	3	2.04
GsASEAAGGPEYQLQESWADQISQLSR	2956.3451	1479.1755	2	2.91
VAELLSsGLQTAIDQIR	1892.9608	947.4891	2	1.48
RLsALLSGVYA	1228.6217	615.3173	2	1.3
RLSALLsGVYA	1228.6217	615.3174	2	1.1
RLsALLsGVYA	1308.588	655.3012	2	0.2
LsALLSGWA	1072.5206	1073.53	1	1.96

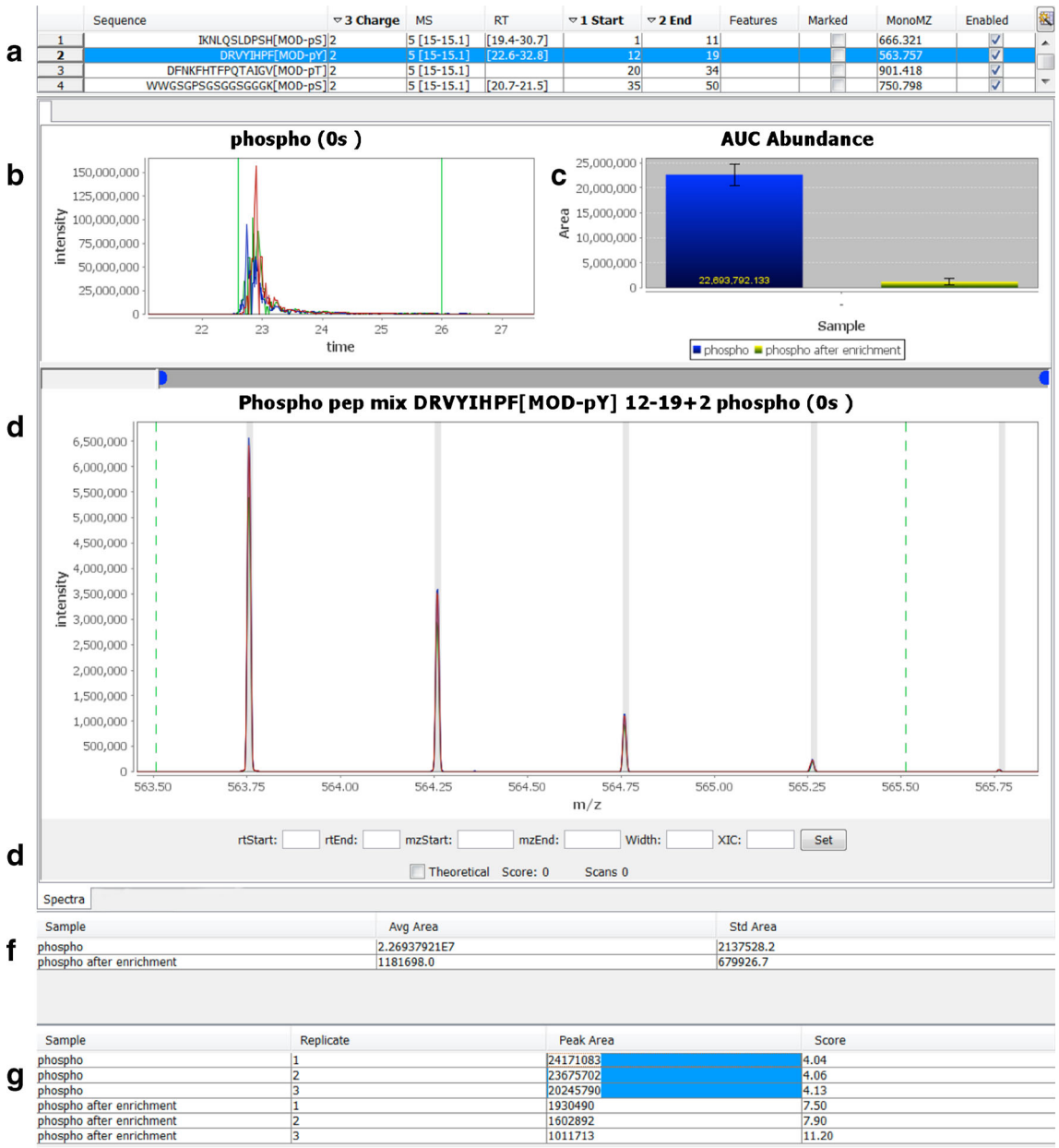


Figure 3. The main interface overview. This feature allows for integrating several graphic tools for editing and reviewing data. **(a)** The top table shows the selected peptide (highlighted in blue) has a phosphorylated tyrosine. Other columns include charge, theoretical *m/z*, and positional information. Peptide selection will launch results in the subsequent tools. **(b)** Extracted ion chromatogram (XIC) for one or more peptide replicates. **(c)** Bar charts present average area under the curve (AUC) from the XIC. **(d)** The spectral pane displays the averaged mass spectral data for one or several replicates. **(e)** Information Toolbar allows recalculation of input values. Tables **(f)** and **(g)** allow users to load from one to many replicates in the spectral and XIC panes for review. The replicate table **(g)** displays AUC results for each sample peptide replicate; **(b)**, **(c),** and **(d)** support zoom in/out functionality

search on the protein Ulk1 would need to consider 17 times or 1.4 million more permutations than an MS^1 approach (Table 2).

While the workflow described is suitable for the analysis of complex mixtures against large protein databases, Proteomics Workbench has been initially designed to interrogate single proteins, providing graphical tools to for quick validation and analysis. The software interface provides a central feature set for the integration of several graphic tools for editing and reviewing of data. Multiple peptides can be selected and the resulting spectra and chromatographic data are rendered in the software (Figure 3). Proteomics Workbench provides graphical tools in which a user can validate the peptide by visually examining the spectra, which is displayed in the same way as the native instrument software such as the Qual Browser (Thermo Scientific, San Jose, Ca, USA). The averaged mass spectrum for the peptide is rendered in the software interface and grey bars outline where each peak should reside theoretically based on peak width and charge state. The extracted ion

chromatogram (XIC) view is calculated based on all peak mass ranges within the user defined *m/z* start and *m/z* end, and presents the chromatographic peak and the retention time range used to generate the peptide's peaks. In the event that there are multiple peaks in the XIC, the user can select a different chromatographic peak and examine the resulting spectra by adjusting the retention time range. The theoretical isotopic distribution can also be overlaid onto the observed. Using these tools, the user can quickly and confidently validate peptide assignments.

Area under the curve (AUC) data are represented in bar charts and all results are exportable from the software. Another available feature can search all peptides within the experiment and flag those that are isobaric or have mass conflicts based on shared mass. The software supports both peptide-specific and global modifications. Peptide-specific modifications are set in the protein editor interface in the peptide set. Global modifications selected in the initial detect job are applied to all peptides as variable. All predefined modifications can be selected prior to the search. User-defined modifications are accessible through an exposed XML file and formatted in a manner similar to UNIMOD. It should be pointed out that in most cases the software can be used in place of the native instrument software and does not need to be tied to this specific workflow.

Conclusions

Identification of post-translational modifications on proteins is an important part of the experimentation performed in mass spectrometry labs worldwide. However, limited tools allow for the identification of PTMs and the tools that exist often do not take into account the combinatorial nature of modifications. This software makes the interrogation of high-resolution MS data to find mass signatures related to putatively modified peptides possible, which then can to be validated by MS/MS spectra. Overall, this software lets the investigator dig deeper into collected data and could provide additional information about modified or unmodified peptides that are present in the analysis. Many who continue to manually search for peptides in MS[1] would benefit from the automation and validation tools this software provides. This information will be of great value to investigators, as they continue to determine the biological significance and regulation that these modifications control. Moreover, when this experimental process is automated, it can be used for rapid screening of sample data and utilized as an iterative approach through continual development of peptide inclusion lists in each subsequent analysis encompassing all of these putative masses.

The software is freely available at http://www.proteomicsworkbench.com.

Acknowledgments

The authors declare that they have no competing financial interests.

References

1. Dorsey, F.C., Rose, K.L., Coenen, S., Prater, S.M., Cavett, V., Cleveland, J.L., Caldwell-Busby, J.: Mapping the phosphorylation sites of Ulk1. J. Proteome Res. **8**(11), 5253–5263 (2009)
2. Jensen, O.N.: Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. Curr. Opin. Chem. Biol. **8**(1), 33–41 (2004)
3. Mann, M., Jensen, O.N.: Proteomic analysis of post-translational modifications. Nat. Biotechnol. **21**(3), 255–261 (2003)
4. Pang, C.N., Hayen, A., Wilkins, M.R.: Surface accessibility of protein post-translational modifications. J. Proteome Res. **6**(5), 1833–1845 (2007)
5. Witze, E.S., Old, W.M., Resing, K.A., Ahn, N.G.: Mapping protein post-translational modifications with mass spectrometry. Nat. Methods **4**(10), 798–806 (2007)
6. Wan, Y., Yang, A., Chen, T.: PepHMM: a hidden Markov model based scoring function for mass spectrometry database search. Ann. Chem. **78**(2), 432–437 (2006)
7. Sadygov, R.G., Yates III, J.R.: A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. Anal. Chem. **75**(15), 3792–3798 (2003)
8. Sadygov, R.G., Liu, H., Yates, J.R.: Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. Anal. Chem. **76**(6), 1664–1671 (2004)
9. Nesvizhskii, A.I., Roos, F.F., Grossmann, J., Vogelzang, M., Eddes, J.S., Gruissem, W., Baginsky, S., Aebersold, R.: Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. Mol. Cell. Proteomics **5**(4), 652–670 (2006)
10. Nesvizhskii, A.I., Keller, A., Kolker, E., Aebersold, R.: A statistical model for identifying proteins by tandem mass spectrometry. Anal. Chem. **75**(17), 4646–4658 (2003)
11. Nesvizhskii, A.I., Aebersold, R.: Interpretation of shotgun proteomic data: the protein inference problem. Mol. Cell. Proteomics **4**(10), 1419–1440 (2005)
12. Matthiesen, R., Trelle, M.B., Hojrup, P., Bunkenborg, J., Jensen, O.N.: VEMS 3.0: algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins. J. Proteome Res. **4**(6), 2338–2347 (2005)
13. MacCoss, M.J., Wu, C.C., Liu, H., Sadygov, R., Yates III, J.R.: A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. Anal. Chem. **75**(24), 6912–6921 (2003)
14. MacCoss, M.J., McDonald, W.H., Saraf, A., Sadygov, R., Clark, J.M., Tasto, J.J., Gould, K.L., Wolters, D., Washburn, M., Weiss, A., Clark, J., Yates III, J.R.: Shotgun identification of protein modifications from protein complexes and lens tissue. Proc. Natl. Acad. Sci. U. S. A. **99**(12), 7900–7905 (2002)
15. Leitner, A., Foettinger, A., Lindner, W.: Improving fragmentation of poorly fragmenting peptides and phosphopeptides during collision-induced dissociation by malondialdehyde modification of arginine residues. J. Mass Spectrom. **42**(7), 950–959 (2007)
16. Keller, A., Nesvizhskii, A., Kolker, E., Aebersold, R.: Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal. Chem. **74**(20), 383–5392 (2002)
17. Ghesquiere, B., Van Damme, J., Martens, L., Vandekerckhove, J., Gevaert, K.: Proteome-wide characterization of *N*-glycosylation events by diagonal chromatography. J. Proteome Res. **5**(9), 2438–2447 (2006)
18. Geer, L.Y., Markey, S.P., Kowalak, J.A., Wagner, L., Xu, M., Maynard, D.M., Yang, X., Shi, W., Bryant, S.H.: Open mass spectrometry search algorithm. J. Proteome Res. **3**(5), 958–964 (2004)
19. Garavelli, J.S.: The RESID Database of Protein Modifications as a Resource and Annotation Tool. Proteomics **4**(6), 1527–1533 (2004)
20. Field, H.I., Fenyo, D., Beavis, R.C.: RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimizes protein identification, and archives data in a relational database. Proteomics **2**, 36–47 (2002)
21. Eng, J.K., McCormack, A.L., Yates, J.R.: An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. **5**, 976–989 (1994)
22. Creasy, D.M., Cottrell, J.S.: UNIMOD: protein modifications for mass spectrometry. Proteomics **4**(6), 1534–1536 (2004)
23. Craig, R., Beavis, R.C.: TANDEM: Matching Proteins with Tandem Mass Spectra. Bioinformatics **20**(9), 1466–1467 (2004)
24. Corthals, G.L., Wasinger, V.C., Hochstrasser, D.F., Sanchez, J.C.: The dynamic range of protein expression: a challenge for proteomic research. Electrophoresis **21**(6), 1104–1115 (2000)

25. Colinge, J., Masselot, A., Giron, M., Dessingy, T., Magnin, J.: OLAV: towards high-throughput tandem mass spectrometry data identification. Proteomics **3**(8), 1454–1463 (2003)
26. Colinge, J., Masselot, A., Cusin, I., Mahe, E., Niknejad, A., Argoud-Puy, G., Reffas, S., Bederr, N., Gleizes, A., Rey, P.A., Bougueleret, L.: high-performance peptide identification by tandem mass spectrometry allows reliable automatic data processing in proteomics. Proteomics **4**(7), 1977–1984 (2004)
27. Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Mazumder, R., O'Donovan, C., Redaschi, N., Suzek, B.: The universal protein resource (UNIPROT): an expanding universe of protein information. Nucleic Acids Res. **34**(Database issue), D187–191 (2006)
28. Zhang, B.M.K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., Lajoie, G.: PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun. Mass Spectrom. **17**(20), 2337–2342 (2003)
29. Taylor, J.A., Johnson, R.S.: Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. Anal. Chem. **73**(11), 2594–2604 (2001)
30. Taylor, J.A., Johnson, R.S.: Sequence Database Searches Via de Novo Peptide Sequencing by tandem mass spectrometry. Rapid Commun. Mass Spectrom. **11**(9), 1067–1075 (1997)
31. Frank, A., Pevzner, P.: PepNovo: de novo peptide sequencing via probabilistic network modeling. Anal. Chem. **77**(4), 964–973 (2005)
32. Dancik, V., Addona, T.A., Clauser, K.R., Vath, J.E., Pevzner, P.A.: De novo peptide sequencing via tandem mass spectrometry. J. Comput. Biol. **6**(3/4), 327–342 (1999)
33. Chen, T., Kao, M.Y., Tepel, M., Rush, J., Church, G.M.: A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. J. Comput. Biol. **8**(3), 325–337 (2001)
34. Bafna, V., Edwards, N.: On de-novo interpretation of tandem mass spectra for peptide identification. Proceedings of the 7th Annual International Conference on Computational Molecular Biology, Berlin, Germany — April 10 - 13 (2003)
35. Ning, K., Ng, H.K., Leong, H.W.: An accurate and efficient algorithm for peptide and PTM identification by tandem mass spectrometry. Genome Inform. **19**, 119–130 (2007)
36. Ahrne, E.M.M., Lisacek, F.: Unrestricted identification of modified proteins using MS/MS. Proteomics **10**, 671–686 (2010)
37. Zhang, K., Chen, Y., Zhang, Z., Tao, S., Zhu, H., Zhao, Y.: Unrestrictive identification of non-phosphorylation PTMs in yeast kinases by MS and PTMap. Proteomics **10**(5), 896–903 (2010)
38. Tsur, D., Tanner, S., Zandi, E., Bafna, V., Pevzner, P.A.: Identification of post-translational modifications by blind search of mass spectra. Nat. Biotechnol. **23**, 1562–1567 (2005)
39. Na, S., Jeong, J., Park, H., Lee, K.J., Paek, E.: Unrestrictive identification of multiple post-translational modifications from tandem mass spectrometry using an error-tolerant algorithm based on an extended sequence tag approach. Mol. Cell. Proteomics **7**(12), 2452–2463 (2008)
40. Moore, G.: Cramming more components onto integrated circuits. Electronics. 38(8) (1965) Available at: http://www.cs.utexas.edu/users/fussell/courses/cs352h/papers/moore.pdf. Accessed 24 July 2015
41. Moore, G.: Excerpts from a conversation with Gordon Moore: Moore's Law. Intel Video Transcript (2005) Available at: http://large.stanford.edu/courses/2012/ph250/lee1/docs/Excepts_A_Conversation_with_Gordon_Moore.pdf. Accessed 24 July 2015
42. NP-Complete. Available at: http://en.wikipedia.org/wiki/NP-complete. Accessed 24 July 2015
43. Tanner, S., Shu, H., Frank, A., Wang, L., Zandi, E., Mumby, M., Pevzner, P.A., Bafna, V.: InsPecT: fast and accurate identification of post-translationally modified peptides from tandem mass spectra. Anal. Chem. **77**(14), 4626–4639 (2005)
44. Tabb, D.L., Saraf, A., Yates III, J.R.: GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. Anal. Chem. **75**(23), 6415–6421 (2003)
45. Sunyaev, S., Liska, A.J., Golod, A., Shevchenko, A., Shevchenko, A.: MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. Anal. Chem. **75**(6), 1307–1315 (2003)
46. Searle, B.C., Dasari, S., Turner, M., Reddy, A.P., Choi, D., Wilmarth, P.A., McCormack, A.L., David, L.L., Nagalla, S.R.: High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS de novo sequencing results. Anal. Chem. **76**(8), 2220–2230 (2004)
47. Mann, M., Wilm, M.: Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal. Chem. **66**(24), 390–4399 (1994)
48. Day, R.M., Borziak, A., Gorin, A.: PPM-Chain – de novo peptide identification program comparable in performance to Sequest. Proceedings of 2004 I.E.E.E. Computational Systems Bioinformatics Conference (CSB 2004), 505–508, 16–19 Aug 2004
49. Wang, L., Wang, W., Chi, H., Wu, Y., Li, Y., Fu, Y., Zhou, C., Sun, R., Wang, H., Liu, C., Yuan, Z., Xiu, L., He, S.M.: An efficient parallelization of phosphorylated peptide and protein identification. Rapid Commun. Mass Spectrom. **24**(12), 1791–1798 (2010)
50. Tanner, S., Payne, S.H., Dasari, S., Shen, Z., Wilmarth, P.A., David, L.L., Loomis, W.F., Briggs, S.P., Bafna, V.: Accurate annotation of peptide modifications through unrestrictive database search. J. Proteome Res. **7**(1), 170–181 (2008)
51. Creasy, D.M., Cottrell, J.S.: Error tolerant searching of uninterpreted tandem mass spectrometry data. Proteomics **2**(10), 426–34 (2002)
52. Craig, R., Beavis, R.C.: A method for reducing the time required to match protein sequences with tandem mass spectra. Rapid Commun. Mass Spectrom. **17**(20), 2310–2316 (2003)
53. Egertson, J.D., Kuehn, A., Merrihew, G.E., Bateman, N.W., MacLean, B.X., Ting, Y.S., Canterbury, J.D., Marsh, D.M., Kellmann, M., Zabrouskov, V., Wu, C.C., MacCoss, M.J.: Multiplexed MS/MS for improved data-independent acquisition. Nat. Methods **10**(8), 744–746 (2013)
54. Rost, H.L., Rosenberger, G., Navarro, P., Gillet, L., Miladinovic, S.M., Schubert, O.T., Wolski, W., Collins, B.C., Malmstrom, J., Malmstrom, L., Aebersold, R.: OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat. Biotechnol. **32**(3), 219–223 (2014)
55. Hoopmann, M.R., Merrihew, G.E., Von Haller, P.D., MacCoss, M.J.: Post-analysis data acquisition for the iterative MS/MS sampling of proteomics mixtures. J. Proteome Res. **8**(4), 1870–1875 (2009)
56. Arntzen, M.O., Osland, C.L., Raa, C.R., Kopperud, R., Doskeland, S.O., Lewis, A.E., D'Santos, C.S.: POSTMan (post-translational modification analysis), a software application for PTM discovery. Proteomics **9**(5), 1400–1406 (2009)
57. Rinner, O., Mueller, L.N., Hubalek, M., Muller, M., Gstaiger, M., Aebersold, R.: An integrated mass spectrometric and computational framework for the analysis of protein interaction networks. Nat. Biotechnol. **25**(3), 345–352 (2007)
58. Chen, Y., Chen, W., Cobb, M.H., Zhao, Y.: PTMap—a sequence alignment software for unrestricted, accurate, and full-spectrum identification of post-translational modification sites. Proc. Natl. Acad. Sci. U. S. A. **106**(3), 761–766 (2009)
59. Pappin, D.J., Hojrup, P., Bleasby, A.J.: Rapid identification of proteins by peptide-mass fingerprinting. Curr. Biol. **3**(6), 327–32 (1993)
60. Perkins, D., Pappin, D.J.C., Creasy, D., Cottrell, J.: Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis **20**(18), 551–567 (1999)
61. Lu, B., Motoyama, A., Ruse, C., Venable, J., Yates III, J.R.: Improving protein identification sensitivity by combining MS and MS/MS information for shotgun proteomics using LTQ-Orbitrap high mass accuracy data. Anal. Chem. **80**(6), 2018–2025 (2008)
62. He, Z., Yu, W.: Improving peptide identification with single-stage mass spectrum peaks. Bioinformatics **25**(22), 2969–2974 (2009)
63. Pascal, B.D., Willis, S., Lauer, J.L., Landgraf, R.R., West, G.M., Marciano, D., Novick, S., Goswami, D., Chalmers, M.J., Griffin, P.R.: HDX Workbench: software for the analysis of H/D exchange MS data. J. Am. Soc. Mass Spectrom. **23**(9), 1512–1521 (2012)
64. Rhoads, T.W., Williams, J.R., Lopez, N.I., Morre, J.T., Bradford, C.S., Beckman, J.S.: Using theoretical protein isotopic distributions to parse small-mass-difference post-translational modifications via mass spectrometry. J. Am. Soc. Mass Spectrom. **24**(1), 115–124 (2013)
65. Available at: Fasta Format.http://en.wikipedia.org/wiki/Fasta_format. Accessed 24 July 2015
66. Pascal, B.D., Chalmers, M.J., Busby, S.A., Griffin, P.R.: HD Desktop: an integrated platform for the analysis and visualization of H/D exchange data. J. Am. Soc. Mass Spectrom. **20**(4), 601–610 (2009)
67. Pascal, B.D., Chalmers, M.J., Busby, S.A., Mader, C.C., Southern, M.R., Tsinoremas, N.F., Griffin, P.R.: The Deuterator: software for the determination of backbone amide deuterium levels from H/D exchange MS data. BMC Bioinformatics **8**, 156 (2007)
68. Rockwood, A.L., Haimi, P.: Efficient calculation of accurate masses of isotopic peaks. J. Am. Soc. Mass Spectrom. **17**(3), 415–419 (2006)
69. Rockwood, A.L., Van Orman, J.R., Dearden, D.V.: Isotopic compositions and accurate masses of single isotopic peaks. J. Am. Soc. Mass Spectrom. **15**(1), 12–21 (2004)