Research Article

Rapid Communications in Mass Spectrometry

Received: 31 May 2011 Revised: 10 August 2011 Accepted: 16 September 2011 Published online in Wiley Online Library

Rapid Commun. Mass Spectrom. **2011**, *25*, 3617–3626
(wileyonlinelibrary.com) DOI: 10.1002/rcm.5257

An algorithm for identifying multiply modified endogenous proteins using both full-scan and high-resolution tandem mass spectrometric data

Matthew T. Mazur[1†,‡] and Ray Fyhr[2*,‡]

[1]Department of Proteomics, Merck & Co., Inc., 126 E. Lincoln Avenue, P.O. Box 2000, Rahway, NJ 07065, USA
[2]Department of Discovery & Pre Clinical Information Technology, Merck & Co., Inc., 126 E. Lincoln Avenue, P.O. Box 2000, Rahway, NJ 07065, USA

Mass spectrometry based proteomic experiments have advanced considerably over the past decade with high-resolution and mass accuracy tandem mass spectrometry (MS/MS) capabilities now allowing routine interrogation of large peptides and proteins. Often a major bottleneck to 'top-down' proteomics, however, is the ability to identify and characterize the complex peptides or proteins based on the acquired high-resolution MS/MS spectra. For biological samples containing proteins with multiple unpredicted processing events, unsupervised identifications can be particularly challenging. Described here is a newly created search algorithm (MAR) designed for the identification of experimentally detected peptides or proteins. This algorithm relies only on predefined list of 'differential' modifications (e.g. phosphorylation) and a FASTA-formatted protein database, and is not constrained to full-length proteins for identification. The algorithm is further powered by the ability to leverage identified mass differences between chromatographically separated ions within full-scan MS spectra to automatically generate a list of likely 'differential' modifications to be searched. The utility of the algorithm is demonstrated with the identification of 54 unique polypeptides from human apolipoprotein enriched from the high-density lipoprotein particle (HDL), and searching time benchmarks demonstrate scalability (12 high-resolution MS/MS scans searched per minute with modifications considered). This parallelizable algorithm provides an additional solution for converting high-quality MS/MS data of multiply processed proteins into reliable identifications. Copyright © 2011 John Wiley & Sons, Ltd.

Mammalian systems are comprised of considerable protein complexities arising from the naturally circuitous translation of gene to protein. Protein posttranslational processing events, such as signal peptide truncation, alternate gene splicing, and protein modifications, greatly challenge the 'one gene, one protein' hypothesis and exponentially expand the number of individual protein forms possible in higher organisms.[1] This is crucial when considering that even seemingly insignificant modifications to proteins can have huge implications in both folding and function.[2,3] Luckily, many of these processing events are partially predictable, such as *N*-linked glycosylations, allowing protein databases to be highly annotated for structure and experimental function.[4,5] Unfortunately, however, not all protein forms are predictable and many are often heavily processed post-translationally, such as seen in neuropeptides.[6] The goal of 'top-down' proteomics approaches is to not only detect and identify intact protein mixtures, but also to fully characterize each protein form present.[7] Unlike 'bottom-up' experiments that rely on sample proteolysis prior to mass spectrometric detection, top-down experiments tend to provide higher individual protein information at the cost of proteome coverage.[8] Recent 'middle-down' approaches attempt to capitalize on the advantages of both approaches by analyzing large peptide (typically 3–20 kDa) mixtures.[9]

The latest advancements in the commercialization of high-performance mass spectrometers have ushered in an increase in top- and middle-down proteomics efforts. Improvements including routine high mass accuracy (<2 parts per million, ppm) and resolution (>1 million), as well as new tandem mass spectrometry techniques (e.g., electron-transfer dissociation, ETD), allow detection (i.e., MS) and characterization (i.e., MS/MS) of protein analytes.[10,11] Exploiting both high mass accuracy full-scan and tandem mass spectrometry (MS/MS) measurements greatly diminish the number of potential protein candidates from a database search.[12] Additionally, the commercial availability of high-resolution ETD, which tends to provide complementary and more random backbone fragmentation compared to threshold methods (e.g., collisionally induced dissociation, CID, or infrared multiphoton dissociation, IRMPD), greatly expands the ability to characterize large, multiply charged peptides

* *Correspondence to:* R. Fyhr, Merck Research Labs, RY32-209, 126 E. Lincoln Avenue, P.O. Box 2000, Rahway, NJ 07065, USA. E-mail: ray_fyhr@merck.com

† Present address: BioAnalytical Sciences, Imclone Systems, a wholly owned subsidiary of Eli Lilly & Co., 22 Imclone Drive, Branchburg, NJ 08876, USA.

‡ These authors contributed equally to this work.

and proteins.[13] Using LTQ-Obritrap-ETD instruments, top-down and middle-down experiments have been successful in cataloguing human cerebral spinal fluid proteins,[14] interrogating human histones,[15] and quantifying significant abundance differences in complex, unchanging high-density lipoprotein (HDL) backgrounds.[16]

As instrument technologies improve, the need for database searching algorithms tailored toward top-down and middle-down datasets becomes increasingly apparent. Established spectral matching software tools, such as SEQUEST and Mascot, can analyze high-mass and high-resolution MS/MS data even though such algorithms were originally developed for relatively low-performance MS/MS experiments.[17–19] Relying predominately on predicted fragmentation spectra, and not neutral fragment masses, spectral matching algorithms begin failing at highly charged (typically >5+) precursor ions. Considerable improvements have been made to incorporate high mass accuracy MS/MS data; however, these tools restrict searches to <16 kDa precursor mass.[20] Additionally, experimental mass accuracy of both precursor and fragment masses are not considered when scoring, effectively rendering high and low mass accuracy experiments equal in information.

Software improvements have been made to better incorporate intact protein MS/MS searches often used for top-down proteomics experiments. Two commonly used search engines, Open Mass Spectrometry Search Algorithm (OMSSA)[21] and Top-down Mascot (aka, 'BIG Mascot'),[20] attempted to incorporate top-down MS/MS functionality into their algorithms by allowing intact protein (i.e., 'noncutting enzymes') searches. These searches were fundamentally similar to bottom-up approaches in the need to pre-select possible protein modifications and had a limited ability to identify unknown modifications. Other more advanced top-down algorithms, such as MS-Deconv[22] and Precursor Ion Independent Top-down Algorithm (PIITA),[23] have allowed searches of proteins containing modifications of unknown mass. These algorithms are both highly selective for the identification of intact proteins containing unknown modifications, and increasingly accurate for the localization of that modification. Unfortunately, these algorithms restrict searches to intact proteins, potentially missing complexities of real protein degradations, either through inefficient sample processing or complex *in vivo* post-translational protein processing (e.g., neuropeptides, neuroproteins).

Software tools that fully exploit the information obtained from high-performance instruments are required for characterizing large peptides or proteins from complex samples. In 2001, Kelleher and coworkers spotlighted the advantages of using high mass accuracy fragment ions to unambiguously identify proteins from a database and derived a probability-based scoring method for protein identification.[12] Unlike spectral matching tools, protein identification scores rely on the numbers of fragments observed and matching and the experimental mass accuracy. Thus it stands to reason that as instrument specifications improve the ability to convert high mass accuracy MS/MS data into highly probable protein matches should increase. ProsightPTM emerged as the first database searching algorithm specifically designed for top-down protein identifications.[24,25] Leveraging the posttranslational modification information contained within RESID protein database, the functionality of ProsightPTM was extended to 'middle-down' peptide mixtures.[9,26] Further, 'shotgun' annotations of protein databases provided even a higher specificity for characterizing individual protein forms.[27,28] Together, ProsightPTM has become a powerful tool for analyzing nearly any precursor mass between 1–100 kDa from high-resolution MS/MS data. However, as proteins become multiply processed (e.g., signal peptide truncation, post-translational modification, and proteolytic trimming), or if databases are insufficiently annotated, such searching tools may miss authentic matches from even high-quality spectra.

Searching algorithms that function without the need for highly annotated, and sometimes highly customized, peptide or protein databases can provide an opportunity when *a priori* protein information is unknown or unreliable. Algorithms, including SEQUEST and Mascot, that use databases of raw protein sequences can be quickly searched for peptide or protein sequences that match an experimentally determined mass, within a defined tolerance. Incorporating user-predefined amino acid modifications can circumvent the need for highly annotated databases for the consideration of posttranslational modifications. The simple and efficient FASTA format allows simple additions or edits without the need for lengthy database indexing. Additionally, the database simplicity presents the opportunity of parallelizing the searches, and greatly reducing the searching times required.

The mass information contained within the full-scan (i.e., MS only) spectra can provide valuable insight into unknown ion species, and is a key piece of information that is typically overlooked during protein identification algorithms. Although online reversed-phase chromatography is generally applied to simplify complex proteomic mixtures prior to mass spectrometric analysis, the chromatographic elution of protein species differing in relatively minor mass and structural changes is often not dramatically different. For example, under generally shallow, linear gradients a methionine-oxidized protein (Δ mass +15.99 Da in protein of ~16 kDa, or 0.1% mass change) typically chromatographically elutes from a reversed-phase separation within a few minutes of the unoxidized protein form. This chromatographic 'disadvantage' of co- or nearby-elution of similar protein species can be exploited for identifying modified protein species. To achieve this, flanking full-scan spectra of ions targeted by MS/MS can be interrogated for mass differences matching a list of known protein modifications (i.e., Unimod database).[29] These potential modifications can then be used to automatically generate a list of 'differential' modifications used during protein identification searches. Therefore, if multiple protein isoforms exist at detectable levels, even at modestly separated chromatographic times, key information from full-scan spectra can be extracted and utilized in determining possible identifications. PTMcRAWler, a functionality contained within ProsightPTM, is one of the first top-down software tools to begin to implement full-scan mass data for protein identifications.[30] Although PTMcRAWler does not consider chromatographic distance as an input parameter, identifications using full-scan mass differences have been shown for both phosphoproteins and highly modified histone proteins containing up to four acetylations.[30]

The ideal database searching algorithm would incorporate the advantages of each of the existing software tools in order to exploit the total of information contained in high

Rapid Communications in Mass Spectrometry

performance mass spectrometry data for efficient protein identifications. Described here is a custom developed software program that effectively assigns precursor masses, de-isotopes corresponding MS/MS data, identifies potential differential modifications from full-scan MS spectra, performs a database search for any peptide string contained with a protein database (i.e., not constrained by full-length protein restrictions), and scores the resulting matches from entire LC/MS/MS data files. Manually generated user-predefined differential modifications address the possibility of posttranslational modifications not identified from full-scan spectra. Any precursor mass is searchable, including peptides (<1 kDa) and very large proteins (>100 kDa), and conceptually there is no limit to the size or number of differential modifications. The architecture is such as to provide computational parallelization, significantly shortening search times. This provides the opportunity for iterative searching using prioritized differential modification lists of any number. Proof of functionality is demonstrated by identifying several apolipoproteins from an undigested human HDL sample, including a multiply modified apolipoprotein C-III protein containing an *O*-glycosylation and an oxidized apolipoprotein C-I species identified from a full-scan mass difference. This algorithm is compatible with data from both threshold (b/y type ions) and nonergodic ECD/ETD (c/z-type ions) fragmentation techniques. Search time benchmarks reveal the ability to search a high-resolution raw file containing 334 MS/MS events in just 31 min while considering any one of 351 possible post-translational modifications.

EXPERIMENTAL

HDL samples

Purified human HDL (d = 1.063–1.210 g/mL) was purchased (Millipore, Billerica, MA, USA), desalted by ZipTip C4 (Millipore, Billerica, MA, USA), lyophilized to dryness and resuspended in 0.1 M acetic acid (solvent A) containing 10 mM tris(2-carboxyethyl)phosphine (TCEP, to prevent protein disulfide bond formation). Samples were reconstituted to a total protein concentration of 0.35 mg/mL and used directly without further processing.

Mass spectrometric analysis

HDL samples were analyzed by reversed-phase nano-HPLC coupled to a LTQ-Orbitrap Velos-ETD hybrid mass spectrometer (Thermo Electron, San Jose, CA, USA). For online LC/MS analysis, 1 μL of sample was loop injected (Agilent Series 1100) and gradient eluted (Eksigent Technologies nanoLC Ultra 2D) using a linear increase from 2% to 50% solvent B (0.1 M acetic acid in acetonitrile) over 50 min. Intact proteins eluting from a ChromXP LC column (150 μm × 10 cm, 3 μm particle size; Eksigent Technologies) at 400 nL/min were introduced into the mass spectrometer by electrospray ionization using a 3 kV needle voltage and a heated metal capillary temperature of 275 °C. Each full-scan FTMS scan was followed by two high-resolution (30 000 resolving power) tandem MS/MS scans of the most abundant two ions. ETD MS/MS was performed on assigned charge states >2+ using an isolation width 10 *m/z* and charge-state-dependent activation time (default 70 ms, 8+ charge state). The exclusion list function was employed as to not target the same precursor ion (within 10 ppm mass tolerance) for a duration of 30 s; parameters include list size 500, repeat count 1, exclusion duration 60 s. Ion injection times were adjusted by the instrument automatic gain control (AGC, 1×10^5 a.u. setting) with a maximum accumulation time not to exceed 1 s; 1 and 5 μscans were collected for full- and high-resolution MS/MS scans, respectively.

LC/MS spectral de-isotoping

Full-scan MS and MS/MS spectra contained within acquired .raw files were converted into lists (.csv format) of monoisotopic, neutral masses using the Horn transformation function of the publically available program Decon2LS[31] running on a dual-core 2.5 GHz cpu PC. Parameters used for full-scan de-isotoping included: maximum mass 50 000, maximum charge 50, allowable shoulders 1, area peak determination, peptide background ratio 4.5, max fit <0.3, threshold intensity for deletion 3, and threshold intensity of score 3. Parameters used for MS/MS scan de-isotoping included: maximum mass 50 000, maximum charge 50, allowable shoulders 1, CHISQ determination, peptide background ratio 1, max fit <0.5, threshold intensity for deletion 1, and threshold intensity of score 1. The full-scan and MS/MS Decon2LS output isos.csv files were merged into a single .csv file using a custom batch file and used by the MAR search algorithm (below).

MAR algorithm: software and structure

The overall architecture of MAR consists of five custom-written code components (MAR_index, MAR_PTMdiff, MAR_NoE, MAR_ions, MAR_score), a FASTA-formatted protein database (.dat file), a user-defined and prioritized list of differential amino acid modifications, and a spectral de-isotoping and precursor ion mass determination program (Decon2LS).[31] The overall architecture is illustrated in Fig. 2. Each MAR custom module is written in standard C language. The entire MAR search engine is powered by the parallelization of computing processes originating at MAR_index, with an equal distribution of database protein sequences being allocated over as many CPUs (or ecores) that are made available. Routinely, parallelization scales linearly (i.e., searches halved when ecores are doubled). Scalability is easily achieved with executable MAR_split, which uses a simple technique of splitting the desired protein data file into N number of files where N is the number of ecores on a Linux multimode Beowulf style cluster. The development environment consists of eight nodes described below. Computer Make/Model, HP DL585 G5; four Dual-Core AMD Opteron(tm) Processor 8220; Clock Rate, 2.8 GHz; RAM per node, 64 GB; MFLOPS, ~800; Backbone Cluster high speed 1GByte LAN; All share drives are BlueArc shares within the local LAN; OS: Red Hat 3.4; Compiler, gcc version 3.4.6 20060404; No additional utilities beyond base Linux.

RESULTS AND DISCUSSION

The MAR protein search algorithm was developed to handle the very large number of candidate polypeptides that result when termini restrictions are eliminated and differential amino acid modifications are allowed. For example, within the Uniprot_Human database[32] (release 05/03/2011) comprised of 20 238 proteins with an average protein length of 558 residues (Fig. 1(A)), the number of unmodified peptide candidates 5 amino acids or longer in length is 6.8 billion (sum-total, Fig. 1(B)). Imposing the trypsin restriction requiring cleavage at Lys and Arg residues, that number is reduced by approximately 2050-fold to 3.3 million. As differential amino acid modifications are considered, the number of candidate peptides begin to expand factorially.[33] Such large candidate search spaces, and their computational time required, are a primary reason why enzyme-unrestricted searches containing multiple differential modifications can be very time- and resource-consuming.

The individual modules which make up the MAR algorithm were designed in an environment to best handle the large number of potential candidates in a reasonable computational time. The architecture (Fig. 2) is comprised of five modules to handle the input of de-isotoped full-scan and fragment masses (Decon2LS), collecting candidate polypeptides from precursor masses (MAR_NoE), identifying potential differential modifications based on adjacent full-scan spectra (MAR_PTMdiff), matching the fragment masses to candidate polypeptides (MAR_ions), and scoring of resulting matches. The 3500 total lines of code were written entirely in C language, with the expectation of parallelization on Beowulf-type clusters.

The MAR_index component is the first stage in addressing the computational requirements by preparing the protein database and appropriate MS/MS data for search querying. This program parses a .dat formatted protein database for protein ID (i.e., name) and sequence, and calculates the theoretical molecular weight of each protein. As this is being performed, a file byte position offset records the position of the protein ID from the beginning of the .dat file. The program writes a sequential array of the double-floating point value of each residue into a .bin file with a −1.0 terminator to indicate the C-terminus. The file offset within the binary file of the starting residue for each protein is also saved and becomes part of the .idx file. Using the Uniprot protein database[32] (release 05/03/2011) of 20 238 proteins, the MAR_index runs in 3.7 s and produces output files .idx and .bin with sizes of around 1 Mbyte and 80 Mbytes, respectively. These are loaded into RAM at the beginning of the execution of both MAR_NoE and MAR_ions programs. Of course when N ecores are used, the size of the files are 1/Nth as large (with 60 ecores this is 17 Kbytes and 1.3 Mbytes, respectively).

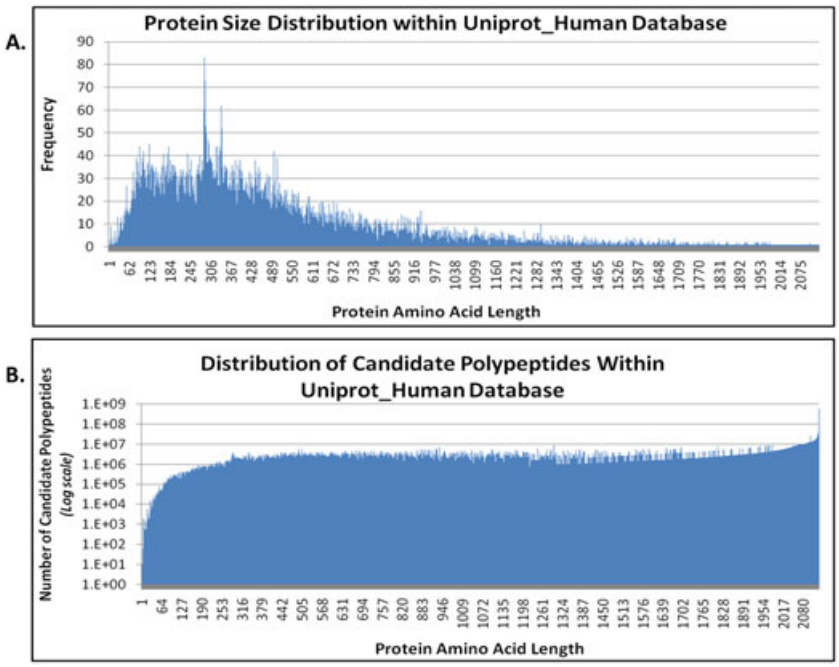


Figure 1. Protein identification search times are proportional to the size of the protein/peptide search space considered. (A) Protein size distribution of entries contained within the Uniprot_Human database[32] (release 05/03/2011) as a function of amino acid length; mean protein length is 558 amino acids. (B) Number of peptide candidates 5 amino acids in length or longer as a function of protein amino acid length for entries contained within the Uniprot_Human database. Proteins less than 559 kDa (5030 amino acids) total 79% of the total search space (5.4/6.8 billion, not considering modifications) despite representing 99.9% of all total proteins (20208/20238) contained within the same database.

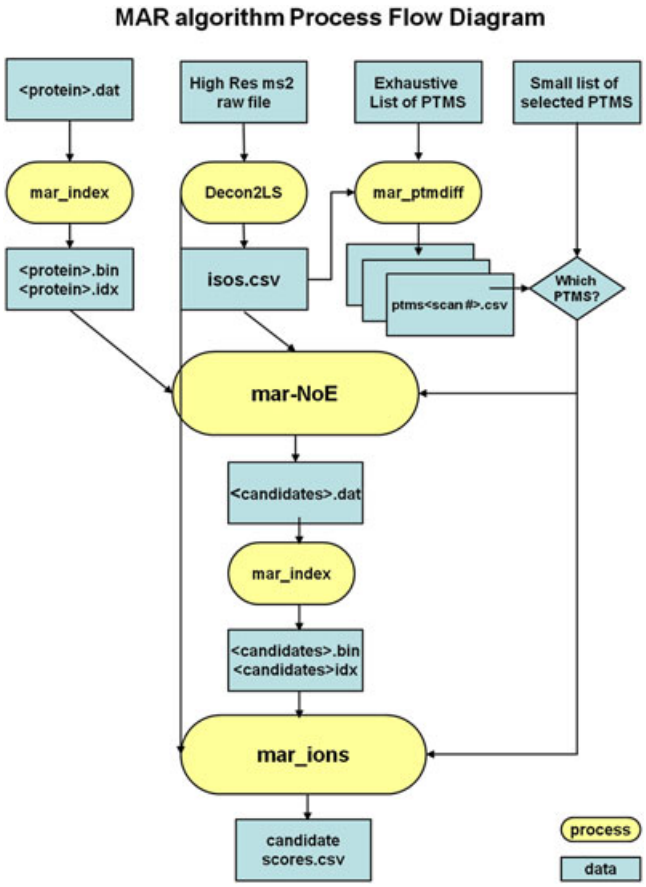


Figure 2. The MAR schema and architecture. The module MAR_index creates from a protein database two additional files (.idx and .bin) to enable an indexed sequential access method (ISAM) file system with binary data containing the double floating point representation of the amino acid sequences contained within the protein database. The MAR_NoE module searches and collects any string of amino acid sequence within the double floating point indexed database (.idx) that matches the input generated precursor ion mass within a given predefined tolerance, with consideration for both modified and unmodified amino acids. *In silico* generated fragments of candidate sequences are created by the MAR_ions module and compared against experimental fragment ions obtained from de-isotoped high-resolution MS/MS. Every candidate fragment is first assigned an internal score based on the sum of the number of ions matched between *in silico* and experimental molecular weights within the tolerance the user requested (30 ppm). This sum is then divided by the variance of the ion matches giving an internal scoring mechanism that can be calculated quickly. Upon completion of the MAR_ions module, the highest 200 candidate scores are then assigned P-scores.

The computational core of the MAR algorithm is the MAR_NoE component, which determines the precursor mass of the fragmented ion and extracts all polypeptide sequences that match this value (within a given mass tolerance) from the indexed protein database. The precursor mass is determined by comparing the instrument targeted *m/z* (contained within raw file) with the corresponding Decon2LS Horn transformation output, allowing the monoisotopic mass to be obtained. The second input parameter is the list of differential amino acid modifications, which is either user-generated or automatically compiled from full-scan spectral data (for automation, see MAR_PTMdiff, below). This .csv formatted file (Table 1) contains the amino acid specification, the mass of the modification, the maximum number of possible modifications considered for any single match, and the modification name. It should be noted that the maximum number of modifications is defined for each modification separately, and no total number of modifications value is imposed. This provides greater search flexibility. A user defined search tolerance for both precursor and fragment masses (in ppm) is included. The matching of experimental precursor mass to the database-contained polypeptide sequence proceeds through the 'inch worm' algorithm with MAR_NoE.

The 'inch worm' algorithm starts with the first residue's molecular weight of the first protein and adds all subsequent residues until the summed mass plus the mass of all differential modifications is 200 Da less than the experimental intact polypeptide. At this point the C-terminus of the peptide being observed is expanded by one amino acid at a time until the mass exceeds the experimental molecular weight being searched. All combinations of considered differential modifications (see below) are applied to each subject polypeptide in order to collect all candidate polypeptide matches. If no candidates match within the given tolerance, the algorithm moves the N-terminus in the C-terminal direction to the next amino acid and the process repeated until the protein terminates within the database. The key computational advantage of the differential modification consideration is the multiplicative (not factorial) expansion of possible candidate peptides:

$$\text{Maximum multiplying factor} = (\text{Mod}_1 + 1) \times (\text{Mod}_2 + 1) \times (\text{Mod}_3 + 1) \ldots \times \ldots (\text{Mod}_n + 1)$$

where n is the number of different modifications (Mod) considered. The position of applied modifications within the polypeptide is not considered until the end of the search, at which point the modification may be localized. Each candidate polypeptide, with or without modifications, whose mass falls within the tolerance of the experimental molecular weight is output to a file in .dat file format, with the protein ID and any modifications appended with the text containing the start and stop residue number of the peptide.

The MAR_NoE algorithm was designed to use a list of differential amino acid modifications used for selecting preliminary candidates from a given protein database (.dat format). Typically, these modifications are chosen up-front by the user based on sample preparation and/or laboratory experience. The MAR_PTMdiff module employs a new strategy to automate this selection by leveraging the ion information contained within the full-scan spectra acquired adjacent to the MS/MS experiment. Calculated mass differences between the MS/MS precursor mass and all other ions observed within ±5 min retention (or some other user-pre-defined time for longer gradients) are compared to a list of known protein modifications (e.g., Unimod database). Observed differences that match within a user-defined tolerance are automatically inserted into the differential modification .csv file (see Table 1) to be used for searching. Unfortunately, the presence of both (or multiple) protein forms within the analyzed spectrum range (i.e., ±5 min) is

Table 1. Typical structure of user-defined list of 'differential' modification input file (.csv format). Required items include the amino acid location, the mass of the modification, and the maximum number of possible modifications considered for any single match; the inclusion of a modification name is optional

Use	Residue	Delta mass (Da)	Max. frequency	Name
Y	M	15.99492	2	oxidation
N	S	79.96633	2	phosphorylation
Y	Q	−17.02650	1	pyroglutamic acid
Y	N	0.98402	1	deamidation
N	T	162.05280	1	glycation
N	K	14.01565	2	methylation

required for detecting differences used for identifications. To address this, the algorithm accommodates the more common user-pre-defined modification list for searching.

The MAR_ions component of the algorithm is tasked with matching the experimental fragment ions with *in silico* generated theoretical fragments of the candidates compiled from the MAR_NoE. Because each experimental ion can be either a b- or y-type fragment (or c- and z-type, in the case of ECD or ETD), a reverse component calculates the theoretical molecular weights from both the C- and N-terminus. For an unmodified peptide containing n amino acids, 2n fragments are generated; for candidate peptides containing potential modifications, the number of theoretical fragments generated *in silico* is determined by both the maximum number of allowed modifications (for each individual modification) and the number of potentially modified residues in each candidate peptide. Clearly, this number of theoretical fragments increases with increasing differential modifications considered [2^(n + 1) possible fragments for n number of modifications considered]. Theoretical fragment lists of each candidate peptide match are sorted using the standard Linux quick sort and each experimental fragment is compared to theoretical. The total number of fragment ions matching each candidate peptide, along with the mass differences of each fragment match, is output to a .bin file for subsequent scoring.

The scoring process uses the user-defined mass tolerance and the number of matching fragment ions to assign a 'random-match' probability to all matching candidate peptides. Simply, using the equation defined by Meng *et al.*:[12]

$$P = \left((xf)^n xe^{-xf}\right)/n!$$

a P-value is calculated for each candidate polypeptide and the top 5 best results are outputted to a .csv formatted file for review. For the experiments described here a mass tolerance of 0.5 Da is employed and only one modification considered (x = 0.036). Candidates are ordered by lowest to highest P-value. Protein database size is determined by the number of possible modifications considered (see Supporting Information). The location of the modification is determined using the MAR_ions algorithm, which compares the 'differential' modification residue specificity with the number of matching fragment ions for the matching candidate protein. Protein forms with the highest number of fragment ions are generally considered correct; however, manual inspection of the PTM location is currently required. Later versions of the MAR algorithm will better address the specific localization of detected PTMs.

The overall throughput gained by the general architecture and parallelization of the MAR algorithm can be seen when analyzing a representative number of fragmentation spectra that would be present in a typical LC/MS/MS experiment. As can be seen in Table 2, searching a data file containing 334 MS/MS spectra requires approximately 23 min of computational time in the absence of differential modifications. As the number of considered modifications increases, the search time increases. The searching of all 334 MS/MS spectra performed while considering any one of 351 Unimod 'differential' modifications (i.e., no MAR_PTMdiff full-scan spectra screening) requires just under 7 hours. As the automated selection of candidate modifications of the MAR_PTMdiff component is employed, search times reduced

Table 2. Search time benchmarks of the MAR algorithm when searching a typical raw LC/MS data file containing 334 FT-MS/MS spectra and considering a variety of modifications. When 0, 10, 100, or all 351 modifications are considered for each protein candidate exhaustively, search times expand from minutes (23 min, no modifications) to hours (7 hours, 351 modifications). Employing the MAR_PTMdiff to scan a 150 (~ ±1.25 min) and 300 spectra (~ ±2.5 min) LC time window for mass differences in full-scan data, search times relax to about 30 min. Note that increasing the LC time-scale from 150 to 300 spectra to identify potential mass differences has a very minor effect on the overall search performance (increase from 27 to 31 min)

raw file containing 5640 total scans, 334 ms2 scans, 30 ppm precursor mass tolerance, Uniprot human database of 20238 proteins (60 ecore cluster)	No PTMs	10 PTMs	100 PTMs	351 PTMs[a]	MAR_ptmdiff[b] (±1.25 min window)	MAR_ptmdiff[b] (±2.5 min window)
Elapsed time in minutes	23	29	118	413	27	31
MS^2 scans per minute	15	12	3	0.8	12	11
Permutations tested	1.0E + 11	1.1E + 12	1.0E + 13	3.6E + 13	1.3E + 12	1.9E + 12
Candidates found	1.2E + 07	1.3E + 08	1.1E + 09	3.7E + 09	1.4E + 08	2.1E + 08

[a]Calculated as 100 PTMs × 3.5.
[b]Δmass tolerance = 50 ppm.

to approximately 30 min. An increase in width of the MAR_PTMdiff chromatographic window for identifying mass differences from 2.5 to 5.0 min has a modest increase in overall search time (27 to 31 min, respectively). Interestingly, although the 30 highest molecular weight proteins represent a disproportionate amount of computational search space (~21%, see Fig. 1(B)), removal of these proteins from the database provides only a marginal improvement in search times (~2–5%). The parallelization of processes and the relatively large computational overhead inherent in 'housekeeping' tasks of the MAR code (~50%) are a primary reason for the lack of improvement in search performance.

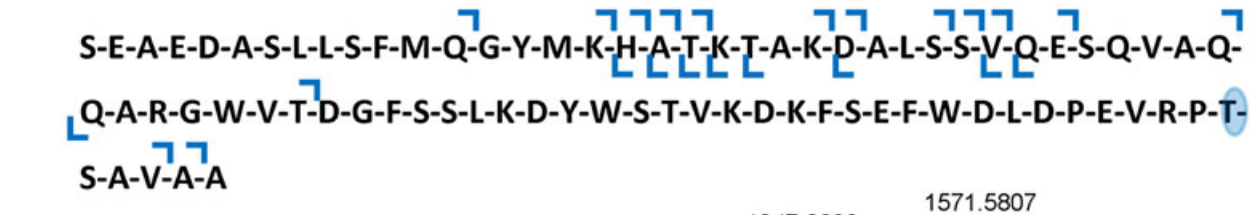
The selectivity of the MAR algorithm to correctly match an experimental MS/MS spectrum to a protein was tested using a relatively simple, undigested mixture of human HDL proteins. The searching algorithm, combined with the MAR_PTMdiff PTM finder function, was successful at identifying 54 unique polypeptide and protein species, several of which contained multiple processing events (peptide chain truncation and modification, Supplementary Table 1, see Supporting Information). These searches were performed using a relatively conservative intact and fragment tolerance of 30 ppm and a MAR_PTMdiff PTM finder window of 300 full-scan spectra (~ ±2.5 min). The number of modifications was restricted to 1 for any candidate polypeptide, effectively focusing the identifications to truncated polypeptide species with a single PTM. It is important to note that only two of the identifications were intact, mature proteins (apolipoprotein C-I and C-III). It was not clear whether this is biology relevant, or alternatively an indication of sample degradation during the processing.

The full capability of the MAR algorithm to identify multiply modified species from a simple protein database without predefined differential modifications was demonstrated with the identification of apolipoprotein C-III (accession P02656). This *O*-glycosylated human protein apolipoprotein C-III, whose mature form cleaves a 20-residue signal peptide, was fragmented by electron-transfer dissociation during LC/MS analysis (Fig. 3, bottom). Full-scan spectra detected the presence of two 'unknown' protein ions (8759.22 and 9415.46 Da), separated by a mass difference of +656.22 Da (no retention difference). This mass difference was within the pre-defined 50 ppm tolerance of a potential *O*-glycosylation modification listed in the Unimod database (Unimod Accession No. 149), and was added (unsupervised) to list of differential modifications to be search by the MAR algorithm. The searching function of MAR was able to correctly identify this protein, with 41 matching c/z ions (P-score 2e-13), as a mature, *O*-glycosylated form of the protein apolipoprotein C-III (Accession P02656, Fig. 3, middle and top). Interestingly,

Search Results:

P-score	c ions	z ions	Scan Number	Protein Name	PTM id	PTM mass	Modification Name
2.0E-13	23	18	1979	APOC3_HUMAN_NoE_21_99	10025	656.2393	T O-glycosylation
2.0E-13	23	18	1979	APOC3_HUMAN_NoE_21_99	10015	656.2393	S Hex1HexNAc1NeuAc1
2.0E-13	23	18	1979	APOC3_HUMAN_NoE_21_99	10016	656.2393	T Hex1HexNAc1NeuAc1
0.6	10	17	1979	ACRBP_HUMAN_NoE_241_322			
0.6	12	15	1979	FOXP4_HUMAN_NoE_29_120			

Fragmentation Map:



ETD FT-MS/MS:

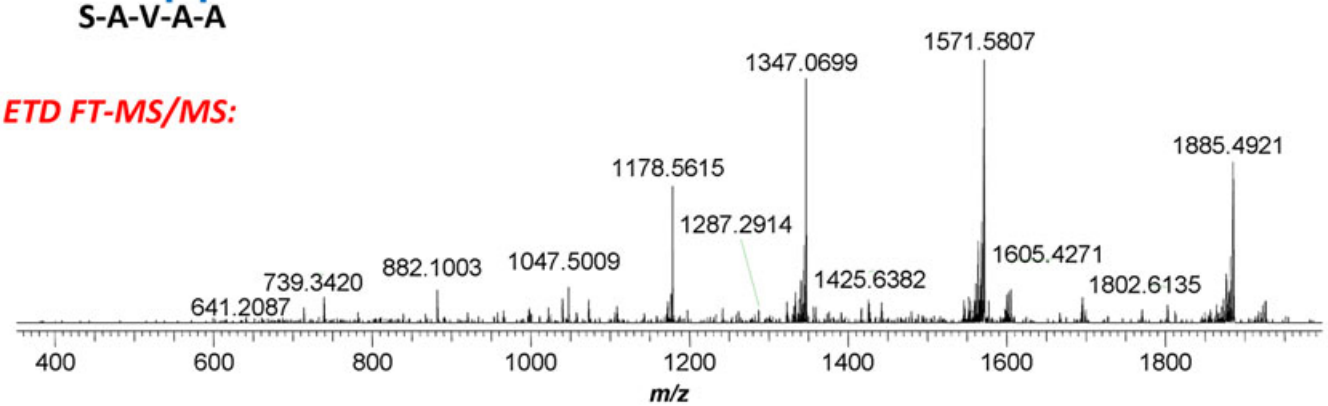


Figure 3. High-resolution ETD FT-MS/MS of apolipoproptein C-III (Accession P02656) and top 5 protein search results from the MAR algorithm. Single-scan ETD fragmentation spectrum of 9+ charge ion at 1047.84 *m/z* (36.3 min, bottom), with matching fragment ions (fragmentation map, middle). Thr^{94} (highlighted) is the site of *O*-glycosylation. MAR search results identify *O*-glycosylated apolipoprotein C-III (Ser^{21}-Ala^{99}, Thr^{94} modified by addition of 656.228 Da, precursor mass 9415.457 Da) as the top hit, with 23 and 18 matching c and z.ions (top).

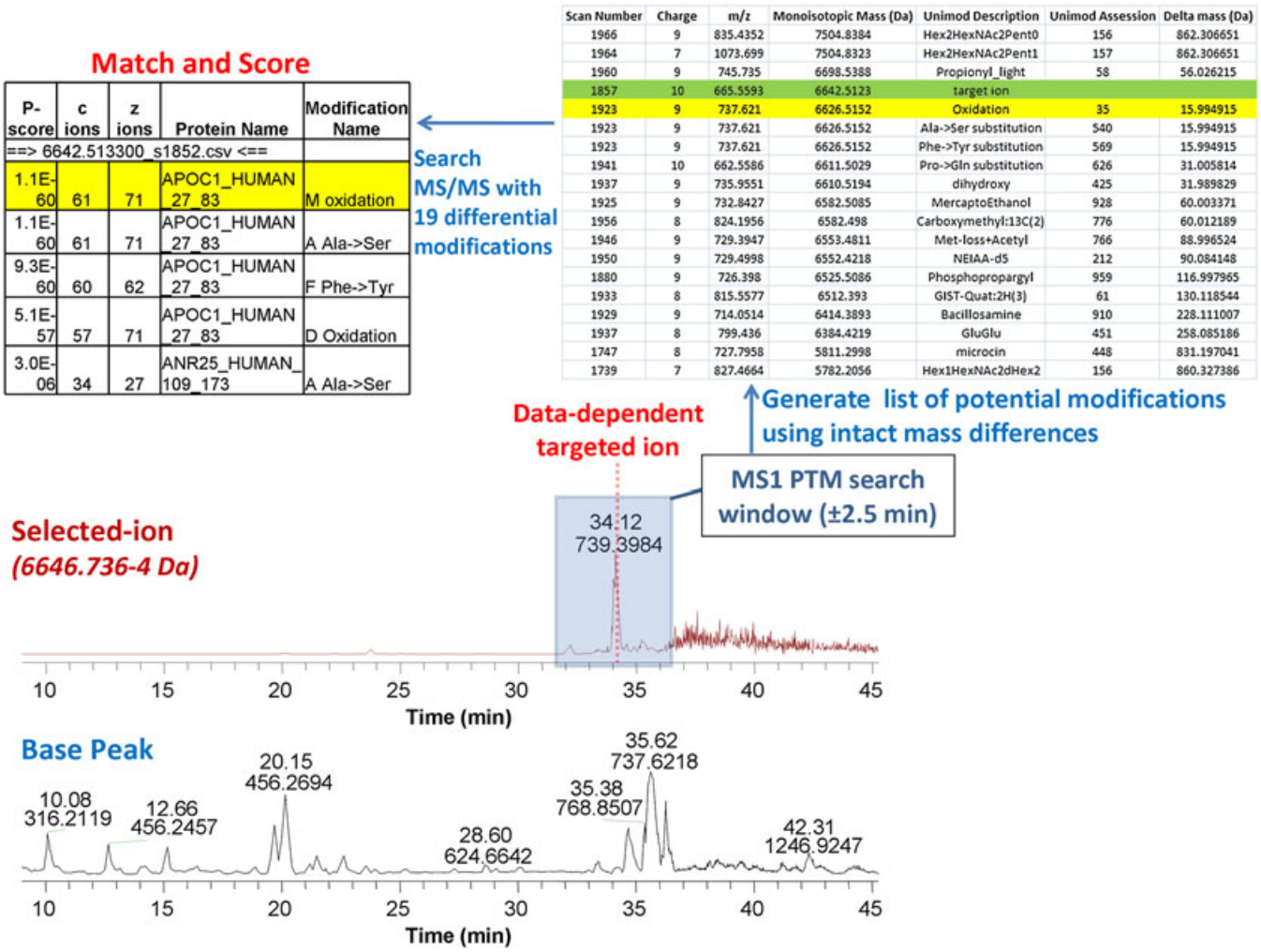


Figure 4. Identification of oxidized apolipoprotein C-I (Accession P02654) using MAR_PTMdiff. Intact masses contained within a 5 min retention time window of an ETD MS/MS targeted ion (739.3984 *m/z*, 9+ charge at 34.12 min) are extracted from full-scan raw data. Mass differences matching Unimod database entries to within 50 ppm are used to populate a list of differential modification used for protein searching (19 total). Search result of the MAR algorithm returns the correct protein assignment, oxidized apolipoprotein C-I (Thr^{27}-Ser^{83}, Met^{64} oxidized), as the top match, with 61 and 71 matching c- and z-type ions. Note, however, that the equal number of matching fragment ions (132 total) prevents the MAR_ions module from differentiating the Met^{64}-oxidized and Ala-to-Ser substitution apolipoprotein C-I forms, and further manual interrogation is required.

although this protein is known to be *O*-glycosylated, the RESID database of modification does not specifically identify the experimentally verified modification (GalNAc-Gal-NANA, +656.24 Da), making it difficult to identify this protein if relying solely on the pre-annotation of the database being searched.

The MAR_PTMdiff functionality was demonstrated by the identification of an oxidized form of apolipoprotein C-I, separated in mass by +15.99 Da and retention time by 1.5 min (Fig. 4). Using the full-scan MS data, MAR_ PTMdiff identified two ion species, 6626.52 Da (35.62 min) and 6642.51 Da (34.12 min) in mass, the second targeted for ETD-MS/MS fragmentation. This targeted ion was automatically searched using an oxidation modification (+15.994915 Da, Unimod Accession No. 35). MAR correctly identified this protein as mature apolipoprotein C-I (Thr^{27}-Ser^{83}, Accession P02654), matching 132 total ions (61 c- and 71 z-type ions) and having a P-score of 1e-60. This protein form was present without the signal peptide (Met^{1}-Gly^{27}) and oxidized at Met^{64}. This example demonstrates that co-elution of related ions is not required for identifying potentially modified proteins.

The specificity of the MAR algorithm is controlled by the quality of the input data required for protein identification. Specifically, several of the MAR-identified proteins, containing possible modifications (Supplementary Table 1, highlighted; see Supporting Information), are considered false-positive identifications because of the incorrect assignment of precursor mass and/or contamination of precursor ions in the MS/MS event. For example, oxidized apolipoprotein C-I (Thr^{27}-Ser^{83}, 6641.51 Da, Supplementary Table 1, row 56; see Supporting Information) was originally identified as containing a Leu to Gln modification (15.0006 Da), although manual inspection reveals that the de-isotoping program incorrectly assigned the monoisotopic mass. The correct assignment of 6642.51 Da is consistent with an oxidation modification of 15.9949 Da (previously identified, Supplementary Table 1, row 54–55; see Supporting Information). Given that the MAR algorithm requires only a .csv formatted input

Rapid Communications in Mass Spectrometry

list of protein and fragment masses, adopting any publically available de-isotoping algorithms better equipped at assigning monoisotopic peaks may provide an improvement in the number of identified spectra. These dependencies highlight current limitations to protein identification that are separate from the searching algorithm.

Although this work demonstrates an approach towards protein identification that is independent of well-annotated protein databases or *a priori* knowledge of posttranslational protein processing, the current MAR algorithm has limitations. The presence of both a modified and unmodified protein form is required for the efficient assignment of potential PTM differences. Clearly, there are many examples of singly modified protein forms being present exclusively. The MAR algorithm attempts to answer this with both the ability to pre-define differential modifications and overall computational speed. The parallelization of the algorithm allows raw file searches in minutes, thereby allowing multiple iterative searches of prioritized modification lists. As unknowns become identified, they can be removed from subsequent searches for even faster computational searches. Separately, the MAR algorithm is also not currently powered with the ability to identify multiple modifications from the MAR_PTMdiff functionality. Later versions are expected to combine the 'PTM finder' functionality with the consideration of 'exhaustive' differential modification searches. That is, searches would combine potential modifications generated from full-scan mass differences together with searches that consider any of the 351 Unimod modifications on any amino acid in a candidate polypeptide. Provided tandem mass spectrometry experiments produce the extensive fragment ions required, it is believed that authentic protein identifications via low P-score values are still highly possible.

The MAR algorithm was designed with the intention of providing an additional solution to the protein identification bottle-neck faced by current proteomics experiments. Allowing searches based on simple FASTA databases enables flexibility, and using full-scan ion information addresses the challenge of pre-defining modifications for consideration. The fresh approach to protein identification provided by the MAR algorithm aims to complement existing protein identification software tools and generally improve the efficiency of each proteomic experiment.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article.

Acknowledgements

The authors thank Ekaterina Deyanova and Helene Cardasis for the collection of high-resolution MS/MS data. The authors thank Fanyu Meng, Laurance Lee, and Cloud Paweletz for their helpful discussions and careful review of this manuscript. The authors also thank Ronald Hendrickson, Nathan Yates, and Martin Leach for their support. Decon2LS was graciously provided free of charge by the Pacific Northwest National Laboratory.

REFERENCES

[1] G. W. Beadle, E. L. Tatum. Genetic control of biochemical reactions in neurospora. *Proc. Natl. Acad. Sci. USA* **1941**, *27*, 499.
[2] Y. Ge, I. N. Rybakova, Q. Xu, R. L. Moss. Top-down high-resolution mass spectrometry of cardiac myosin binding protein C revealed that truncation alters protein phosphorylation state. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 12658.
[3] J. R. Lee, J. K. Kim, S. J. Lee, K. P. Kim. Role of protein tyrosine nitration in neurodegenerative diseases and atherosclerosis. *Arch. Pharmaceut. Res.* **2010**, *32*, 1109.
[4] E. Bause. Structural requirements of N-glycosylation of proteins. Studies with proline peptides as conformational probes. *Biochem. J.* **2010**, *209*, 331.
[5] B. Eisenhaber, F. Eisenhaber. Prediction of posttranslational modification of proteins from their amino acid sequence. *Methods Mol. Biol.* **2010**, *609*, 365.
[6] L. D. Fricker. Analysis of mouse brain peptides using mass spectrometry-based peptidomics: implications for novel functions ranging from non-classical neuropeptides to microproteins. *Mol. Biosyst.* **2010**, *6*, 1355.
[7] N. L. Kelleher. Top-down proteomics. *Anal. Chem.* **2004**, *76*, 197A.
[8] X. Han, A. Aslanian, J. R. Yates III. Mass spectrometry for proteomics. *Curr. Opin. Chem. Biol.* **2008**, *12*, 483.
[9] M. T. Boyne, B. A. Garcia, M. Li, L. Zamborg, C. D. Wenger, S. Babai, N. L. Kelleher. Tandem mass spectrometry with ultrahigh mass accuracy clarifies peptide identification by database retrieval. *J. Proteome Res.* **2009**, *8*, 374.
[10] J. E. Syka, J. J. Coon, M. J. Schroeder, J. Shabanowitz, D. F. Hunt. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 9528.
[11] M. Mann, N. L. Kelleher. Precision proteomics: the case for high resolution and high mass accuracy. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 18132.
[12] F. Meng, B. J. Cargile, L. M. Miller, A. J. Forbes, J. R. Johnson, N. L. Kelleher. Informatics and multiplexing of intact protein identification in bacteria and the archaea. *Nat. Biotechnol.* **2001**, *19*, 952.
[13] G. C. McAlister, D. Phanstiel, D. M. Good, W. T. Berggren, J. J. Coon. Implementation of electron-transfer dissociation on a hybrid linear ion trap-orbitrap mass spectrometer. *Anal. Chem.* **2007**, *79*, 3525.
[14] A. F. Altelaar, S. Mohammed, M. A. Brans, R. A. Adan, A. J. Heck. Improved identification of endogenous peptides from murine nervous tissue by multiplexed peptide extraction methods and multiplexed mass spectrometric analysis. *J. Proteome Res.* **2008**, *8*, 870.
[15] H. R. Jung, D. Pasini, K. Helin, O. N. Jensen. Quantitative mass spectrometry of histones H3.2 and H3.3 in Suz12-deficient mouse embryonic stem cells reveals distinct, dynamic post-translational modifications at Lys-27 and Lys-36. *Mol. Cell. Proteomics* **2010**, *9*, 838.
[16] M. T. Mazur, H. L. Cardasis, D. S. Spellman, A. Liaw, N. A. Yates, R. C. Hendrickson. Quantitative analysis of intact apolipoproteins in human HDL by top-down differential mass spectrometry. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 7728.
[17] J. K. Eng, A. L. McCormack, J. R. Yates III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976.
[18] D. N. Perkins, D. J. Pappin, D. M. Creasy, J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551.

[19] J. R. Yates III, J. K. Eng. Use of mass spectrometry fragmentation patterns of peptides to identify amino acid sequences in databases. US Patent 5,538,897, **1996**.
[20] N. M. Karabacak, L. Li, A. Tiwari, L. J. Hayward, P. Hong, M. L. Easterling, J. N. Agar. Sensitive and specific identification of wild type and variant proteins from 8 to 669 kDa using top-down mass spectrometry. *Mol. Cell. Proteomics* **2009**, *8*, 846.
[21] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, S. H. Bryant. Open mass spectrometry search algorithm. *J. Proteome Res.* **2004**, *3*, 958.
[22] X. Liu, Y. Inbar, P. Dorrestein, C. Wynne, N. Edwards, P. Souda, J. Whitelegge, V. Bafna, P. Pevzner. Deconvolution and database search of complex tandem mass spectra of intact proteins. *Mol. Cell. Proteomics* **2010**, *9*, 2772.
[23] Y. Tsai, A. Scherl, J. Shaw, C. MacKay, S. Shaffer, P. Langridge-Smith, D. Goodlett. Precursor ion independent algorithm for top-down shotgun proteomics. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 2154.
[24] G. K. Taylor, Y. B. Kim, A. J. Forbes, F. Meng, R. McCarthy, N. L. Kelleher. Web and database software for identification of intact proteins using "top down" mass spectrometry. *Anal. Chem.* **2003**, *75*, 4081.
[25] R. D. LeDuc, G. K. Taylor, Y. B. Kim, T. E. Januszyk, L. H. Bynum, J. V. Sola, J. S. Garavelli, N. L. Kelleher. ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry. *Nucleic Acids Res.* **2004**, *32*, W340.
[26] J. S. Garavelli. The RESID Database of protein structure modifications. *Nucleic Acids Res.* **1999**, *27*, 198.
[27] J. J. Pesavento, Y. B. Kim, G. K. Taylor, N. L. Kelleher. Shotgun annotation of histone modifications: a new approach for streamlined characterization of proteins by top down mass spectrometry. *J. Am. Chem. Soc.* **2004**, *126*, 3386.
[28] R. D. LeDuc, N. L. Kelleher. Using ProSight PTM and related tools for targeted protein identification and characterization with high mass accuracy tandem MS data. *Curr. Protocols Bioinformatics* **2007**, 19: 13.6.1–13.6.28.
[29] D. M. Creasy, J. S. Cottrell. Unimod: Protein modifications for mass spectrometry. *Proteomics* **2004**, *4*, 1534.
[30] K. R. Durbin, J. C. Tran, L. Zamdborg, S. M. M. Sweet, A. D. Catherman, J. E. Lee, M. Li, J. F. Kellie, N. L. Kelleher. Intact mass detection, interpretation, and visualization to automate Top-Down proteomics on a large scale. *Proteomics* **2010**, *10*, 3589.
[31] N. Jaitly, A. Mayampurath, K. Littlefield, J. N. Adkins, G. A. Anderson, R. D. Smith. Decon2LS: An open-source software package for automated processing and visualization of high resolution mass spectrometry data. *BMC Bioinforma.* **2009**, *10*, 87.
[32] Available: ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions/.
[33] Y. Li, H. Chi, L. H. Wang, H. P. Wang, Y. Fu, Z. F. Yuan, S. J. Li, Y. S. Liu, R. X. Sun, R. Zeng, S. M. He. Speeding up tandem mass spectrometry based database searching by peptide and spectrum indexing. *Rapid Commun. Mass Spectrom.* **2010**, *24*, 807.

