Precursor Ion Independent Algorithm for Top-Down Shotgun Proteomics

Yihsuan S. Tsai,[a] Alexander Scherl,[a] Jason L. Shaw,[a] C. Logan MacKay,[b] Scott A. Shaffer,[a] Patrick R. R. Langridge-Smith,[b] and David R. Goodlett[a]

[a] Department of Medicinal Chemistry, University of Washington, Seattle, Washington, USA
[b] Department of Chemistry, University of Edinburgh, Edinburgh, United Kingdom

We present a precursor ion independent top-down algorithm (PIITA) for use in automated assignment of protein identifications from tandem mass spectra of whole proteins. To acquire the data, we utilize data-dependent acquisition to select protein precursor ions eluting from a C4-based HPLC column for collision induced dissociation in the linear ion trap of an LTQ-Orbitrap mass spectrometer. Gas-phase fractionation is used to increase the number of acquired tandem mass spectra, all of which are recorded in the Orbitrap mass analyzer. To identify proteins, the PIITA algorithm compares deconvoluted, deisotoped, observed tandem mass spectra to all possible theoretical tandem mass spectra for each protein in a genomic sequence database without regard for measured parent ion mass. Only after a protein is identified, is any difference in measured and theoretical precursor mass used to identify and locate post-translation modifications. We demonstrate the application of PIITA to data generated via our wet-lab approach on a *Salmonella typhimurium* outer membrane extract and compare these results to bottom-up analysis. From these data, we identify 154 proteins by top-down analysis, 73 of which were not identified in a parallel bottom-up analysis. We also identify 201 unique isoforms of these 154 proteins at a false discovery rate (FDR) of <1%. (J Am Soc Mass Spectrom 2009, 20, 2154–2166) © 2009 Published by Elsevier Inc. on behalf of American Society for Mass Spectrometry

In the past decade, shotgun proteomics has been one of the mainstays of proteomics. There are many variations on the theme, but most involve protease digestion of a complex protein sample to make peptides that are in turn analyzed by tandem mass spectrometry to identify the proteins from which they were derived. This peptide based approach circumvents the fundamental decrease in fragmentation efficiency that accompanies increasing molecular weight of proteins, which is one reason that so-called bottom-up approaches have proliferated throughout the field of proteomics [1–3]. Although "bottom-up" strategies are able to provide a great amount of information on the proteins present in a sample and their quantities relative to other samples or standards, limitations remain. One important limitation of standard bottom-up shotgun methods is attributable to requisite proteolysis of proteins to peptides, only some of which are detected in the mass spectrometer while most are never detected. This loss of information means that "bottom-up" shotgun proteomic experiments typically produce very low protein sequence coverage, which precludes mapping all post-translation modifications (PTMs) as well as detection of genetic insertion/deletion events. Top-down proteomic methods seek to circumvent these deficiencies by analyzing whole proteins instead of peptides [4, 5]. One significant advantage, then, of a perfected top-down proteomic pipeline that parallels a standard bottom-up workflow would be mass spectrometric detection of all protein isoforms, much as is the case for two-dimensional electrophoresis (2DE), but with the additional caveat of also characterizing the isoforms simultaneous to detection. Thus, there has been great interest in top-down proteomics from laboratories that specialize in bottom-up proteomics, at least as a complimentary tool, to characterize proteomes.

At the moment, many restrictions remain in top-down proteomics. These include limited sensitivity for detection of high molecular weight proteins, complications in data analysis of electrospray ionization (ESI) [6] generated high charge-state precursor and fragment ion spectra, and separation of proteins in a capillary format as efficiently as peptides may currently be separated [7–9]. On the other hand, with instruments capable of delivering high measured resolution and mass accuracy [10–12], accurate monoisotopic mass assignment has been made easier, and even complex mass spectra containing many high charge state ions may be efficiently deisotoped and deconvoluted by publicly available software and algorithms [13–15]. For the moment, high-throughput top-down proteomics remains a useful tool for a few specialty laboratories, but new soft-

Address reprint requests to Dr. D. R. Goodlett, Department of Medicinal Chemistry, University of Washington, Box 357610, Seattle, WA 98195-7610, USA. E-mail: goodlett@u.washington.edu

1044-0305/09/$32.00
doi:10.1016/j.jasms.2009.07.024

Published online August 13, 2009
Received March 21, 2009
Revised July 29, 2009
Accepted July 29, 2009

ware and hardware solutions will facilitate dissemination and use as a companion tool to bottom-up methods [16–18].

In most top-down proteomics studies, electron-transfer dissociation (ETD) [19, 20] or electron capture dissociation (ECD) [21, 22] have been used to conduct tandem mass spectrometry. For several reasons, these two techniques are extremely attractive fragmentation methods used for top-down proteomics. First, these dissociation methods (i.e., ETD/ECD) are particularly adapted to higher charge state precursors such as the ones obtained by ESI of undigested proteins. Second, with ETD/ECD, fragmentation occurs predominantly on the peptide backbone, which allows covalent modifications to amino acids to be detected and located more easily than by collision induced dissociation (CID) [21] where fragmentation energetics result in facile loss of PTMs from amino acid side chains. Third, while ETD may be used on inexpensive ion traps, these mass analyzers do not offer the resolution required for interpretation of protein tandem mass spectra, and conversely ECD is mainly used in combination with expensive Fourier-transform ion cyclotron resonance (FT-ICR) mass analyzers that do provide this capability at a fairly high investment. All this considered, it remains the case that CID provides better sensitivity than ECD/ETD, is more widely available and is also implemented on high mass accuracy and resolution analyzers [10]. Recently, top-down protein identification was demonstrated using a hybrid linear ion trap-Orbitrap instrument with direct sample infusion [23], pointing the way for more studies using of CID for top-down proteomics.

To increase the number of proteins that may be analyzed in a single top-down proteomic experiment, sample fractionation and/or enrichment is often used to simplify protein mixtures before mass spectrometric analysis. For example, Sharma et al. identified 81 *Shewanella oneidensis* proteins by a combination of on-line reversed-phase liquid chromatography (WAX-RPLC) separation into six fractions and subsequent direct measurement of intact protein molecular weights [24]. Parks et al. identified 39 yeast proteins also using a WAX-RPLC separation, but collected 45 fractions before use of a data-dependent LC-MS/MS strategy to acquire tandem mass spectra via collision induced dissociation (CID) of the whole proteins [25]. Bunger et al. identified 174 *Escherichia coli* proteins using strong anion exchange fractionation to collect 36 fractions followed by on-line reversed-phase liquid chromatography (SAX-RPLC) coupled with data-dependent LC-MS/MS [26] via ECD. In our study, rather than use prior liquid-phase fractionation as other groups have done, we choose a more direct approach utilizing a common bottom-up proteomic method, gas-phase fractionation (GPF), to increase proteome coverage [27–29] and which directly parallels our bottom-up workflow that also eschews sample prefractionation.

To date, there are two basic approaches described in the literature to identify and characterize proteins from their tandem mass spectra; ProSight PTM [30, 31], which was the first software package made available primarily for top-down protein characterization of PTMs that uses a precursor ion based search with sequence tags, and MS-TopDown [32], which uses a spectral alignment algorithm [33, 34] to interpret protein tandem mass spectra. Additionally, a recent article describes a third informatics approach using BigMascot [35], described in part as a modification of Mascot that increases the upper molecular weight range limit available and allows for removal of the N-terminal methionine. Here, we present a fourth algorithmic approach to interpret protein tandem mass spectra that we combined with a direct sample analysis method that minimizes sample preparation and maximizes identifications. Specifically, our strategy uses C4-based HPLC separations into an LTQOT mass spectrometer with GPF that parallels our bottom-up strategy [27–29]. In our top-down strategy, protein identifications are made directly from deconvoluted, deisotoped tandem mass spectra of proteins acquired in the Orbitrap mass analyzer via data-dependent precursor-ion selection with CID carried out in the LTQ ion trap mass spectrometer. Notably different from the other top-down algorithms is that PIITA does not use precursor ion information in the initial protein identification stage. Rather, PIITA uses the measured precursor ion mass only after a gene match is made, and then only to map additions or deletions of mass from the genetically predicted protein molecular weight. The primary benefit of this approach is that PIITA has neither a prior expectation of PTM mass nor number of PTMs. This relaxation of the search process then allows PTMs and genetic insertions/deletions to be detected directly. By combining these concepts, we demonstrate that we can identify 154 unique proteins from a *Salmonella typhimurium* outer membrane extract at a very low false discovery rate (FDR) of <1%. We also demonstrate that PIITA can detect 210 isoforms of these 154 proteins and more importantly 73 small proteins are identified that were not identified in a parallel bottom-up analysis of the same sample.

Materials and Methods

Sample Preparation

There are three different samples used in this study: (1) a *Salmonella Typhimurium* outer membrane protein mixture, (2) a six-protein standard mixture, and (3) a purified histone H4 protein prepared and acquired previously [36], from which select data were provided to PIITA as a peak list. The first two datasets were generated by CID of proteins selected by data-dependent acquisition on HPLC time scales and the third by ECD of a single protein during infusion.

Salmonella Typhimurium were grown to an optical density of 0.6 in LB. Bacteria were centrifuged to form a pellet, which was washed twice with PBS buffer. Cells in the pellet were suspended 5 min in a solution containing 70% acetonitrile and 0.1% trifluoroacetic acid. Suspended cells were centrifuged and the supernatant partially evaporated to remove most of the acetonitrile, after which supernatant protein concentration was determined using a Coomassie-based protein assay (Pierce/Thermo Fisher, San Jose, CA, USA). Extracted proteins were reduced with dithiothreitol and alkylated with iodoacetamide. Denatured proteins were then desalted on a C4 spin column (The Nest Group, Southborough, MS, USA) according to the manufacturer's instructions and this mixture was then ready for separation online with a NanoAquity HPLC system (Milford, MA, USA). Proteins were "trapped" on a home-made 100 μm i.d. × 18 mm long precolumn packed with 300 Å (5 μm Magic C4 particles from Michrom SA, Auburn, CA, USA) that was in line with a home-made gravity-pulled 75 μm i.d. × 150 mm long analytical column packed with the same magic C4 material. This trap-column combination was interfaced to the mass spectrometer as per our standard bottom-up workflow [29]. Alternatively, for bottom up analysis, proteins were digested into peptides using the endopeptidase trypsin, desalted using a C18 microspin column, and separated on a C18 precolumn, and column as previously reported [29].

Six protein standard mixture was prepared from purified proteins purchased from Sigma (St. Louis, MO, USA). An equimolar mixture of the six proteins was constructed by combining bovine insulin (MW 5808), bovine β-lactoglobulin (MW 18,400), bovine α-casein (MW 24,000), bovine β-casein (MW 25,000), chicken egg lysozyme (MW 16,000), and bovine α lactalbumin (MW 16,200) each at 1 μM.

Mass Spectrometry

For *S. Typhimurium* and six standard protein mixture top-down LC-MS/MS analysis, an estimated amount of 2 μg of protein (0.5 μg/μL) was loaded on the precolumn at 4 μL/min in water/acetonitrile (95/5) with 0.1% (vol/vol) formic acid. Proteins were eluted using an acetonitrile gradient flowing at 250 nL/min using mobile phase consisting of: A, water, 0.1% formic acid; B, acetonitrile, 0.1% formic acid. The gradient program was 0 min: A (95%), B (5%), 55 min: A (58%), B (42%), 60 min: A (15%), B (85%), 65 min: A (85%), B (15%), 75–90 min: A (95%), B (5%). The ESI voltage was applied via a liquid junction using a gold wire inserted into micro-tee union (Upchurch Scientific, Oak Harbor, WA, USA) located between the precolumn and analytical column. Ion source conditions were optimized using the tuning and calibration solution recommended by the instrument provider. Injection waveforms for the linear ion trap-Orbitrap (LTQOT; ThermoFisher Scientific, San Jose, CA, USA) were kept on for all acquisitions. For precursor ion survey scans, Orbitrap resolution was set to 60,000 (at 400 Th) and Orbitrap ion populations were held at 5 × 10^5 through use of automatic gain control (AGC). All precursor ion survey scans were performed from 400 to 2000 u. CID was performed in the linear ion trap. Subsequently, tandem MS (i.e. MS/MS) were acquired in the Orbitrap, with a target ion population of 2 × 10^5, precursor isolation width of 5 Th, collision energy of 35% and resolution of 30,000. For each tandem mass spectrum, three microscans were summed. Data were acquired using data-dependent ion selection as is done for bottom-up shotgun proteomic protocols where the most abundant precursor ion at a given moment in chromatographic time was selected for CID. In a first analysis, precursor ions were selected over the entire range of 400–2000 u. Then, the experiment was repeated using GPF; i.e. precursor ions were selected over the following seven ranges: 500–605, 600–705, 700–805, 800–905, 900–1005, 1000–1205, and 1200–2000 u [27]. For *S. Typhimurium* bottom up analysis, peptides were analyzed and identified as previously reported [29]. Finally, the HPLC purified histone H4 protein was infused into a 12 T FT-ICR mass spectrometer and the [M + 13H + 2Methyl + Acetyl]$^{13+}$ ion subjected to ECD tandem mass spectrometry as previously reported [36]. The tandem mass spectrum of the [M + 13H + 2Methyl + Acetyl]$^{13+}$ ion was analyzed by PIITA.

Data Preprocessing and Database Construction

The general workflow is represented in Figure 1. Raw data from MS acquisitions were converted to MS1 and MS2 data format using the MakeMS2 software available on the internet at http://proteome.gs.washington.edu/software/makems2/MakeMS2.zip. Next, multiple charge state ions were deconvoluted into singly-charged monoisotopic *m/z* values using Hardklör [13]. For precursor ion scans, the maximum charge state was set to 50 and the Hardklör correlation threshold to accept a predicted isotope distribution to 0.5. For tandem mass spectra, the maximum charge state was set to 40 and the correlation threshold to accept a predicted isotope distribution to 0.9. Only tandem mass spectra with more than 10 deconvoluted fragment ion masses were used for database search.

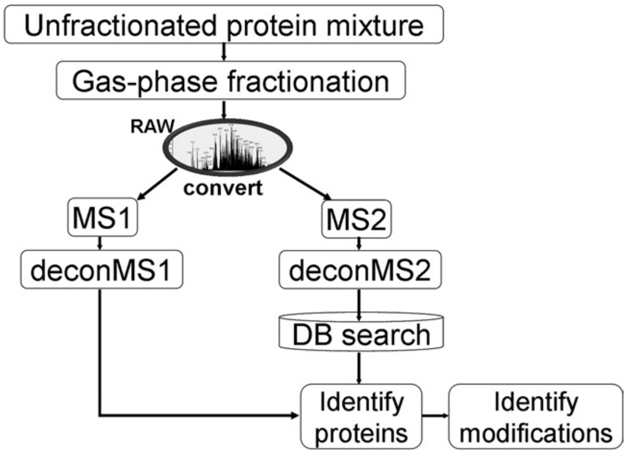


Figure 1. Precursor ion independent top-down algorithm (PIITA) workflow.

To minimize PIITA search time, all FASTA sequence files were preprocessed into files containing only fragment ion masses using the InSilicoSpectro library [37]. For CID data analysis, theoretical fragment ion masses included six ion types: b, b-H_2O, b-NH_3, y, y-H_2O, and y-NH_3, and for ECD: c, c-H_2O, c-NH_3, z, z-H_2O, and z-NH_3. Theoretical protein masses were calculated under the assumption that all cysteine residues were alkylated with iodoacetamide, but users may select any cysteine protecting group mass or no modification. The theoretical MS2 spectra were created by assuming the protein is not N-terminally processed, but PIITA also checks for presence/absence of the initial methionine. Protein sequences of *S. typhimurium LT2* were downloaded from NCBI NC_003197 *Salmonella typhimurium LT2* complete genome, which contains 4527 protein sequences. For the standard protein mixture data search, we added all *S. typhimurium* protein sequences as a background for a total of 4533 protein sequences in the considered database. For the PIITA versus ProSight PTM search time comparison, the *Bos taurus* protein sequence database of 15,410 protein sequences was downloaded (from http://www.uniprot.org/) and used as is.

PIITA Scoring

For the purpose of identifying proteins each observed tandem mass spectrum, processed as described above, was compared to all theoretical tandem mass spectra in the sequence database. To determine the best fit in a sequence database, a fragment ion match (FIM) score and a delta score (ΔSc) were assigned to each observed tandem mass spectrum by comparison to each of the theoretical tandem mass spectra (Figure 2). The FIM score is calculated as follows for each observed protein tandem mass spectrum. The number of all deconvoluted, monoisotopic masses in the observed tandem mass spectrum calculated to be within ±15 ppm mass error of a theoretical fragment ion mass is the FIM score. Additionally, each observed tandem mass spectrum receives a fragment ion total (FIT) score, which is the number of all available deconvoluted, monoisotopic masses. A ΔSc score, similar to the ΔCn of SEQUEST [38], calculated as a function of the difference in FIM scores between the highest score (FIM_{1st}) and the second highest score (FIM_{2nd}) is found by dividing this difference value by the FIT score. Thus, the numerical value of ΔSc is:

$$\Delta Sc = \frac{(FIM_{1st} - FIM_{2nd})}{FIT}$$

When this calculation is made for all theoretical tandem mass spectra a range of ΔSc scores from 0 to 1 results, where $\Delta Sc = 1$ is the best match and $\Delta Sc = 0$ the worst. As with the ΔCn score of SEQUEST, the greater the difference between ΔSc for the best and second best matches, the higher the confidence in the best match. The best matching theoretical tandem mass spectrum (and thus protein identity) is determined based on a

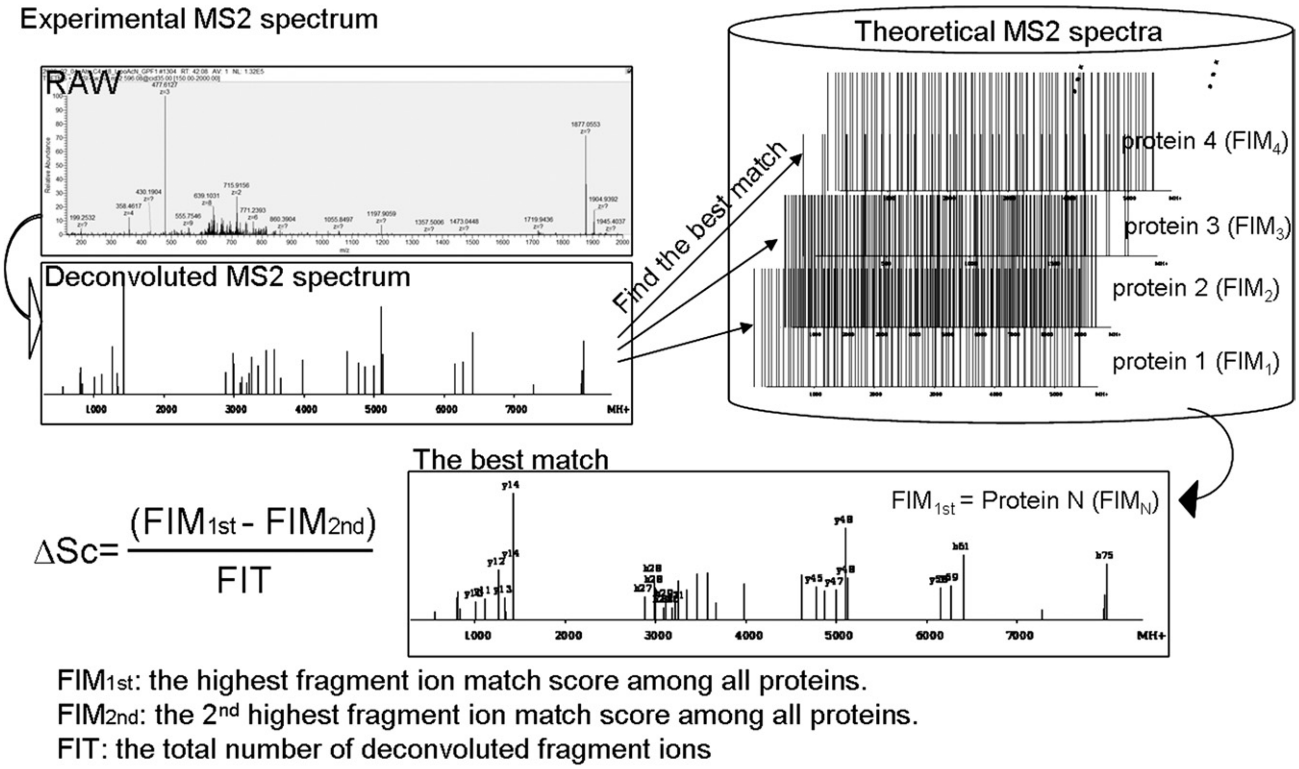

Experimental MS2 spectrum
RAW
Deconvoluted MS2 spectrum
1000
2000
3000
4000
5000
6000
7000
MH+
Find the best match
Theoretical MS2 spectra
protein 4 (FIM_4)
protein 3 (FIM_3)
protein 2 (FIM_2)
protein 1 (FIM_1)
The best match
FIM_{1st} = Protein N (FIM_N)
y14
y12
b28
y48
b61
b75
$\Delta Sc = \frac{(FIM_{1st} - FIM_{2nd})}{FIT}$
FIM_{1st}: the highest fragment ion match score among all proteins.
FIM_{2nd}: the 2nd highest fragment ion match score among all proteins.
FIT: the total number of deconvoluted fragment ions

Figure 2. Schematic of precursor ion independent top-down (PIITA) score calculations. Example of fragment ion match (FIM) score and Δscore (ΔSc) calculations shown.

rank order of calculated FIM scores and ΔSc values. Finally, given that PIITA's scoring scheme is not probabilistic, inter-laboratory comparisons of scores will be moot. To reconcile this problem, future versions of PIITA will incorporate a probabilistic scoring scheme similar to the E-value currently used in SEQUEST [39].

False Discovery Rate (FDR)

To estimate PIITA's ability to find the correct protein in the standard database, all data were also searched against a scrambled database of the same genome. The scrambled database was constructed by shuffling two amino acid pairs at a time until all amino acids in all protein sequences in the genome had been shuffled. Specificity and sensitivity of the database search was derived from receiver-operating characteristic-like (ROC-like) curves using searches in the real and scrambled database. Sensitivity is defined as the number of true positives divided by the sum of true positives and false negatives and specificity as the number of true negatives divided by the sum of true negatives and false positives. The true positives are approximated as the number of identifications in the forward database above the considered ΔSc while the true negatives as the number of identifications in the scrambled database below the ΔSc. False positives are approximated as the number of identifications in the scramble database above the ΔSc and false negatives as the number of identifications in the forward database below the ΔSc. FDR was calculated as false positive/true positive. From ROC-like curves, false positive ratios (FPR) corresponding to a given ΔSc were derived. The specificity, sensitivity, and FDR for a given ΔSc threshold were visualized with ROC-like curves, as in Figure 3 (see Unknown Protein Identification via PIITA, in Results and Discussion).

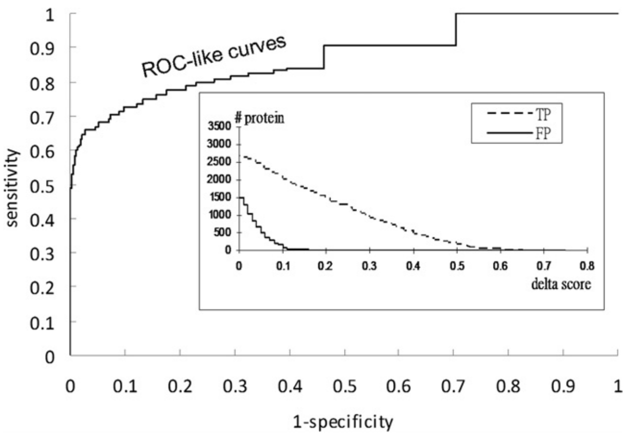


Figure 3. Receiver-operating characteristic-like (ROC-like) curves for *S. Typhimuirium* data. Sensitivity plotted as a function of 1-specificity where sensitivity is defined as number of true positives (TP) divided by number of true positives plus false negatives (FN) and where specificity is defined as number of true negatives (TN) divided by number of true negatives plus false positives (FP).

Covalent Modifications

PIITA's primary goal is to identify the protein giving rise to a particular tandem mass spectrum without reference to the measured precursor ion mass. However, by incorporating the measured precursor ion mass, PIITA may also be used to determine what, if any, modifications are present, and at what amino acids they occur. Once a protein from the database is matched to the spectrum, the ΔM difference in mass is calculated between the measured precursor mass and the theoretical mass of the protein. This value is computed directly from the gene sequence in the database accounting for removal of N-terminal methionine. If $|\Delta M|$ is < 2 Da, the protein is considered unmodified and full-length. We set this cutoff for $|\Delta M|$ because during initial data analysis we detected many 1 and 2 Da mass differences between the theoretical and observed precursor ion masses. Apparently, these result from limitations of defining protein precursor ion masses accurately in the Orbitrap mass analyzer during data-dependent operation. We note that this cutoff will preclude detection of obvious modifications like deamidation, but there is no inherent limitation in the algorithm that precludes detecting these differences when precursor ion mass is recorded accurately. Finally, if the difference between measured and theoretical precursor ion mass is >2 Da, PIITA considers the protein modified and attempts to locate the modification site(s).

To locate the site of modification PIITA begins by assuming the protein has a single modification of mass ΔM on the N-terminal amino acid. It generates the theoretical fragmentation spectrum for this modified protein and calculates its FIM score. The modification is moved, one amino acid at a time, until it reaches the C-terminus of the protein, calculating after each move the theoretical spectrum and FIM score associated with that putative modification site. The putative modification site which yields the highest FIM score is considered most likely to be the true modification site (see Supplementary Figure 1A, which can be found in the electronic version of this article). If there is no maximum FIM score but, instead, a group of consecutive amino acids A_i–A_j that share the maximum score (see Supplementary Figure 1B), PIITA assumes there is more than one modification and proceeds as described next and as shown in Supplementary Figure 1C–E.

This second round requires ΔM to be split into two modification masses: $\Delta M = \Delta M_1 + \Delta M_2$, where ΔM_1 and ΔM_2 are positive integers. For every possible pair of values (ΔM_1, ΔM_2), ΔM_1 is placed at amino acid A_i, a theoretical spectrum is generated, and a FIM score calculated. When all possible fragment ions arising from fragmentation between A_i and A_j are present (i.e., for high quality data covering all possible fragment ions), PIITA concludes that the value of ΔM_1, which yields the highest FIM score is the modification mass at site A_i, and the corresponding ΔM_2 is the modification mass at site A_j. When some fragment ions are missing (e.g., when data quality is poor and there are gaps in sequence coverage), PIITA terminates in one of two ways. First, if a maximum FIM score cannot be found, the algorithm concludes that a total modification of mass ΔM occurs between A_i and A_j, with the specific modification site(s) reported as unknown. Second, if a maximum FIM score is found, ΔM_1 is placed at site A_i and ΔM_2 is moved, one amino acid at a time, from A_{i+1} to A_j, each time calculating the theoretical spectrum and FIM score. The putative site for ΔM_2 which yields the highest FIM score is considered most likely to be the true site for ΔM_2, and the algorithm terminates with this as the reported location. If there is no maximum FIM score, but instead another group of consecutive amino acids sharing the maximum FIM score, ΔM_2 represents more than one modification, and is handled in the same manner as ΔM (see Supplementary Figure 1B, E) by beginning the process again.

Results and Discussion

Description of the Analytical Approach

In these studies, denatured proteins were separated by C4 reverse-phase chromatography and directly injected into the LTQ-Orbitrap mass spectrometer according to standard data-dependent precursor ion selection methods as used in shotgun proteomic analyses [29]. Specifically, precursor ions were selected over the entire available range of 400–2000 u. To compensate for the lack of prior liquid phase separations, GPF was employed to increase identifications from complex mixtures. After precursor ion selection, CID of the whole protein was carried out in the LTQ ion trap with subsequent acquisition of tandem mass spectra in the Orbitrap mass analyzer because high-resolution tandem mass spectra were required to accurately deconvolute and deisotope the tandem mass spectra.

Description of PIITA

During bottom-up proteomics generally the two most important parameters recorded for each peptide are (1) a precursor ion mass and (2) a collection of fragment ion masses from this precursor ion. Together, these two parameters are used to match a peptide sequence in a database of sequence to the tandem mass spectrum and thus identify the protein from which the peptide was derived. The precursor ion mass is used by the typical search engine to generate a list of putative peptide matches in the sequence database that are compared to the observed tandem mass spectrum. The theoretical peptide tandem mass spectrum with the best fit is reported as the identified peptide and by inference the parent protein. While this is an efficient means of making matches to peptide sequence in a database, the major pitfall is that there are generally many unexpected modifications to the peptide sequence in the database that, if unknown at the time of the search, result in an under reporting of the protein sequence variations in the mixture. Therefore, because of this and the fact that two-dimensional gel electrophoresis analysis of protein mixtures reports large numbers of isoforms for each gene, we chose to implement a precursor-ion independent search approach for analysis of top-down data. This precursor ion independent top-down (PIITA) approach allows for identification of unexpected modifications; i.e., both additions and subtractions of mass from the reported gene sequence.

PIITA is based on a database search that compares experimental CID (or other, e.g. ECD) fragment ion masses with theoretical masses of each protein in the database (Figure 2). Because the sequence of an intact protein is much longer than a digested peptide, the chance of the fragment ions matching the wrong protein is very low [40]. First, using each protein sequence in the database, a FIM score is derived for the observed tandem mass spectrum and then a best matched protein is derived using an additional ΔSc score defined above.

PIITA Scoring

After matching the protein tandem mass spectrum to a specific gene and thus parent protein name, the observed precursor mass of this spectrum is compared to the theoretical mass of the best scoring protein sequence match. Second, from this comparison, any difference in ΔM value greater or lesser than the identified gene sequence identifies changes to the gene sequence and/or covalent modifications to amino acids. Thus, compared to the general top-down algorithmic approaches that exist, PIITA skips the initial precursor ion mass directed search in favor of locating first the parent gene by comparison of each observed fragment ion mass to the list of theoretical ion masses after which observed change between precursor and theoretical protein mass is used to identify the addition of PTMs to the protein (Figure 1). While none of the available top-down routines appears to take this exact approach, there is one bottom-up approach, ModifiComb [41] that appears to use the mass difference between the precursor ion and the unmodified peptide to discover unknown modifications on peptides. In brief, the ModifiComb approach assumes that both the unmodified and modified form of a peptide exist in the data under examination using confidently identified unmodified peptides to hunt for modified forms of the same peptide. While the early released top-down algorithms, like ProSight PTM, primarily provided a means to thoroughly map known modifications, newer developments in top-down search algorithms will likely include more robust means to discover unexpected PTMs such as carried out by PIITA.

Testing PIITA

Before using PIITA to characterize previously unidentified proteins in the *Salmonella typhimurium* outer membrane extract, PIITA was tested on a contrived sample containing six proteins of known sequence. This six protein mixture was denatured, separated by C4 reverse-phase chromatography, and directly injected into the LTQ-Orbitrap mass spectrometer, where ions were selected for CID according to standard data-dependent precursor ion selection methods without use of GPF. After converting RAW files into a singly-charged monoisotopic *m/z* format, PIITA was used to identify proteins from a database containing the six standard proteins and, to provide background search "noise", all protein sequences of *S. typhimurium* genome were included. All six proteins known to be added to the sample were identified by PIITA at 0% FDR using a scrambled database with a $\Delta Sc > 0.2$. Additional confidence was gained by virtue of the fact that no *S. typhimurium* proteins were identified during this search. Interestingly, there were no hits with a $\Delta Sc > 0.2$ even when searching with a scramble database that contained 4535 possible proteins demonstrating a high level of confidence in these PIITA generated results. Additionally, we identified 61 total isoforms of these six proteins (Supplementary Table 1). Many (85%) of these protein isoforms were due to removal of a stretch of amino acid sequence from the N-terminus which may be due to degradation of the purchased standard proteins in the sample before analysis. For example, in a separate analysis β-casein analyzed by this method was found to be completely degraded on receipt from the producer (A. Scherl, unpublished result).

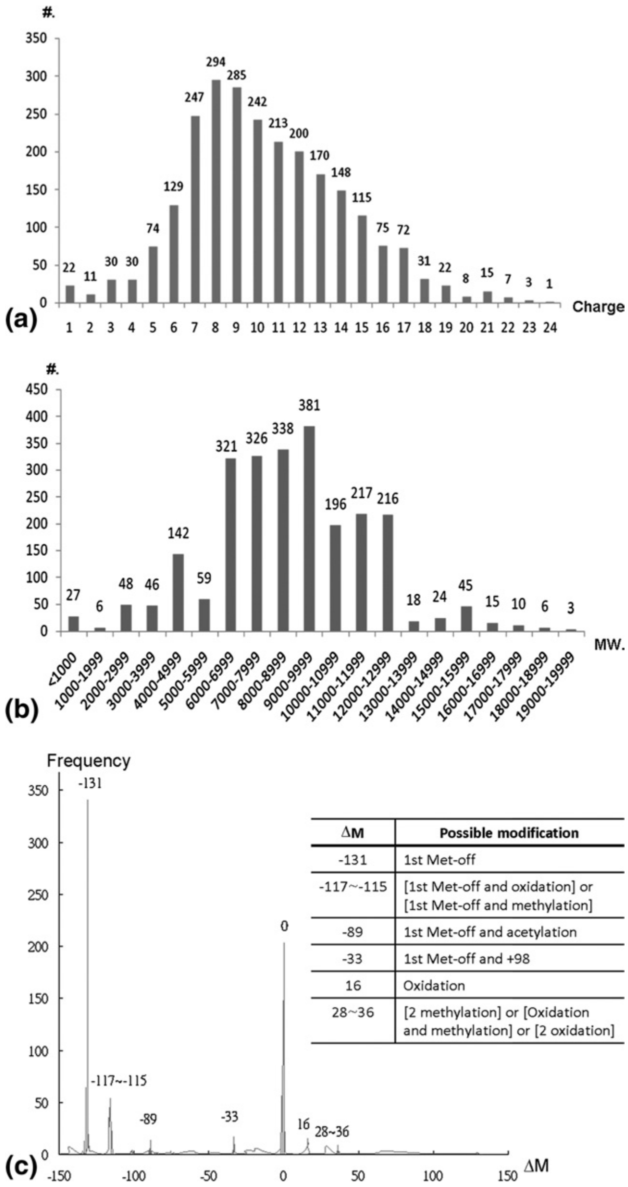
As mentioned, in addition to protein identification, PIITA was also designed for unknown protein modification discovery. To verify this ability, we obtained traditional ECD generated top-down data and analyzed this with PIITA. The single tandem mass spectrum was from a purified histone H4 protein preparation fragmented by ECD in a 12 T FTICR MS after which PTMs were mapped. Analysis of the ECD tandem mass spectrum of this protein by PIITA correctly identified the protein as histone H4 and also two known modifications that resulted from cellular treatment with Trichostatin A before protein purification [36]. To locate PTMs within a protein of known sequence PIITA uses an iterative approach where in a first pass the total identified mass modification to the gene sequence is noted and tentatively assigned to an amino acid region. For example, in the case of PIITA analysis of the histone H4 ECD tandem mass spectrum, an initial FIM score of 19 was produced with the score increasing to 30 when the ΔM value of +70 u was added within the amino acid sequence Ser1 to Arg23. Next, in a second pass analysis PIITA further refined the location(s) of the modification(s) by trying all possible combinations of mass modifications and amino acids within the identified initial stretch of peptide sequence. In the case of the histone H4 tandem mass spectrum, a best FIM sore of 34 was assigned when the +70 u was explained as resulting from two modifications with +42 u in the region between Ser1 to Gly13 and +28 u between His18 to Arg23. Further iterations did not increase the FIM score and it was noted that these two modifications correspond to previously reported modifications where Ser1 is acetylated and Lys20 is dimethylated.

Unknown Protein Identification via PIITA

Finally, an outer membrane protein extract from *Salmonella typhimurium* was analyzed using the same approach as for the six known proteins and histone H4 data. However, given the expected increase in complexity and to compensate for the lack of prior liquid phase separations, GPF was employed covering the following ranges: 500–605, 600–705, 700–805, 800–905, 900–1005, 1000–1205, and 1200–2000 u. This division required seven LC-MS/MS experiments from which proteins were identified. To identify proteins from acquired tandem mass spectra, a database search was performed on all 3510 obtained tandem mass spectra using the PIITA algorithm. Results from this process are shown in Figure 3 in the form of a ROC-like curve for all acquired tandem mass spectra. From the ROC-like curve, a false positive ratio (FPR) corresponding to a given ΔSc can be derived. To interpret the results with a FPR < 1%, we used a $\Delta Sc = 0.2$. This provided, from the combined tandem mass spectral dataset from seven GPF analyses, 154 proteins identified from 154 unique genes. A list of these proteins, identified with an average of 22% fragment ions, may be found in Supplementary Table 2. The identified proteins range from a molecular weight high for inorganic pyrophosphatase protein (NP_463275.1) which has a monoisotopic mass of 19,647 Da to a low of 4416 Da for 50S ribosomal subunit protein X. The average molecular weight for identified proteins was 11,280 and the average isoelectric point was 7.33. The molecular weight range for identified proteins is lower than observed in other reports as may be expected for proteins extracted off the surface of an organism without intentional cell lysis, but this may also be a result of the limit of the LTQOT to efficiently fragment protein ions and then resolve their fragment ions. The identified proteins had an average 95% sequence coverage without use of measured precursor ion mass.

Observed Distribution of Charge-State and Molecular Weight

From the 3510 acquired tandem mass spectra, 2444 were triggered from precursor ions where charge state deconvolution and monoisotopic precursor detection could be performed with high confidence as judged by a Hardklör dot product score ≥ 0.8 [13]. The Hardklör dot product score reflects the correlation between the experimental isotopic distribution and a predicted one which provides a means to evaluate high versus low quality mass spectral assignments. Figure 4a shows the distribution of precursor-ion charge states, and Figure 4b shows the distribution of their monoisotopic masses.



ΔM	Possible modification
-131	1st Met-off
-117~-115	[1st Met-off and oxidation] or [1st Met-off and methylation]
-89	1st Met-off and acetylation
-33	1st Met-off and +98
16	Oxidation
28~36	[2 methylation] or [Oxidation and methylation] or [2 oxidation]

Figure 4. Observed distributions for *S. Typhimuirium* data on protein charge state, monoisotopic mass, and most common ΔM modification. Distribution of (**a**) precursor ion charge state, (**b**) deconvoluted monoisotopic mass range, and (**c**) most common PTMs determined as differences between theoretical and observed protein molecular weights.

Most precursors (65%) were observed between z = 5 to 17 with corresponding molecular masses between 4000 to 13,000 Da. However, precursor ions up to z = 20 were observed and their tandem mass spectra successfully matched to sequence by database search as described above. The remaining tandem mass spectra (30%) corresponded to precursor ions with an unassigned monoisotopic precursor ion mass which were associated with a Hardklör dot product score < 0.8. We note a general trend toward lower Hardklör dot product scores being often assigned with very high charge states, e.g., z = 24, 25, or 26. Also, we note that even low abundant precursor ions gave rise to high quality tandem mass spectra as judged by number of fragment ions resulting ultimately in high confidence protein identification, an observation that points to data-independent top-down experiments [42].

Identification of Protein Isoforms and Chemical Modifications

The fact that PIITA makes an initial identification only using a precursor independent approach allows identification of proteins from the same gene that are present in many different isoforms. For example, proteins with post-translational N-terminal cleavage are identified with PIITA as long as there are enough C-terminal ions in the tandem mass spectrum to match a stretch of gene sequence or vice versa. This observation, then, may be used to locate the unknown modification (i.e., N-terminal cleavage). Conversely, proteins with a modification near the C-terminus are identified as long as there are enough N-terminal ions observed in the tandem mass spectrum to match a gene sequence. Additionally, proteins with PTMs located anywhere in the protein's amino acid sequence may be identified by a combination of N- and C-terminal ions detected before the site of modification. As an example of this, Figure 5 shows data from identification of a putative cytoplasmic protein in three different isoforms. A comprehensive list of all identified proteins and their isoforms is shown in Supplementary Table 2. However, there are a couple of scenarios where use of PIITA requires attention. First, when a protein exhibits predominantly b- or y-ion fragments and a modification on the same terminus, PIITA will fail to identify it. Second, when a protein is modified on both the N- and C-termini, it will not be identified directly because of the requirement for at least a b- or y-ion series of fragment ions. However, in the second case, where such protein forms are suspected, they may be detected by assuming the protein has at least two modifications. This will trigger the covalent modification location process (described in Materials and Methods, Covalent Modifications section) with a ΔM value split into two masses that examines all proteins in the database.

Among all the 154 identified proteins in the *S. Typhimurium* extract, 28 gene products are observed with more than one unique isoform from which 84 different isoforms are observed. Among the isoforms observed, N-terminal cleavage is the most common. Indeed, 71 proteins were present without the N-terminal methionine and an additional 30 proteins had a polypeptide varying from 2 to 261 amino acids missing. In addition to these cleaved protein forms, other modifications such as methylation, acetylation, and oxidation were commonly identified. Among the most frequent modifications (Figure 4c) are PTMs of ribosomal proteins that have been previously reported [43]. These include acetylation of 50S ribosomal subunit protein L7/L12, 30S ribosomal subunit protein Ser5 and 30S

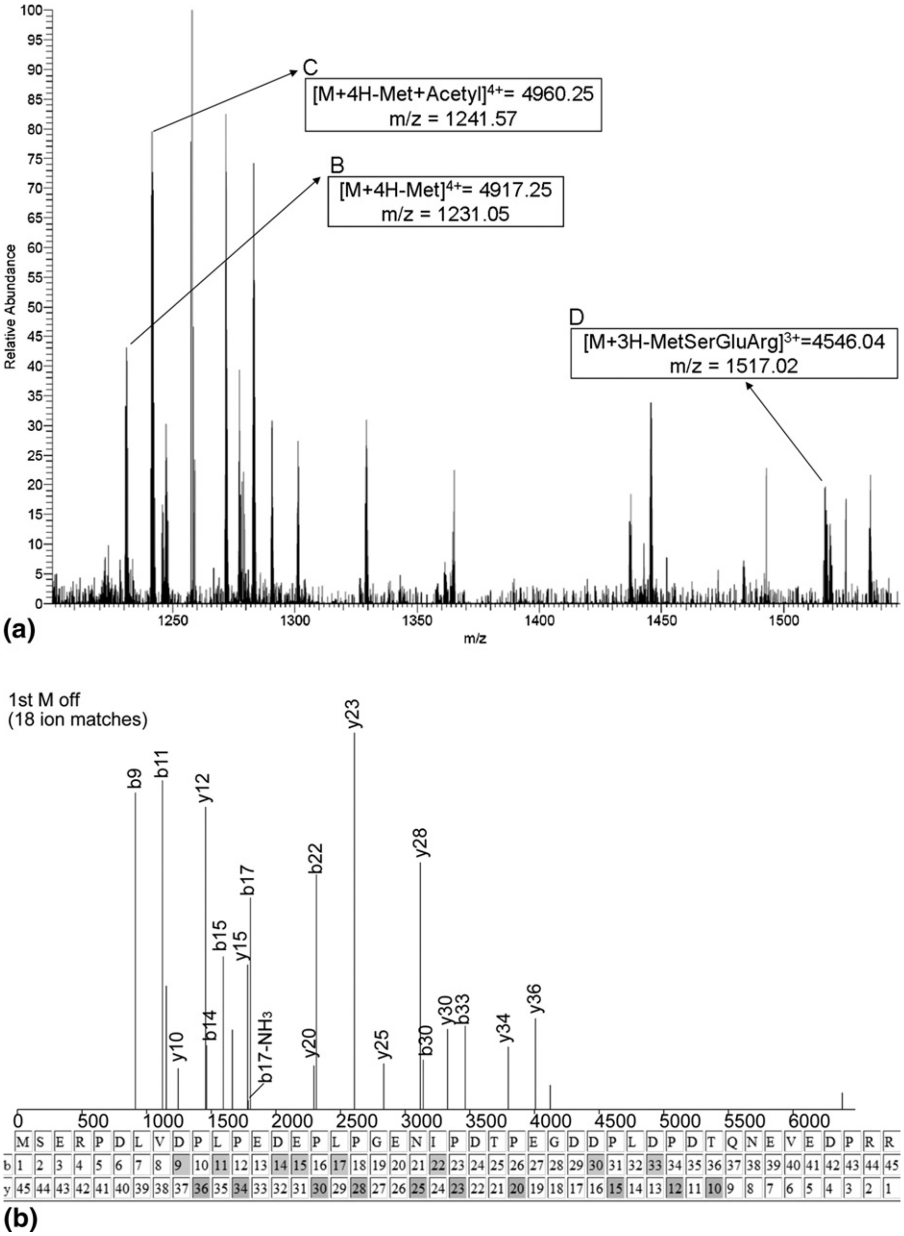


Figure 5. Data from three different protein isoforms of putative cytoplasmic protein. Averaged precursor ion scan (**a**) containing three identified ions identified as isoforms of putative cytoplasmic protein (NP_460766.1). Protein isoform with N-terminal methionine removal (**b**), protein with N-terminal methionine removal plus acetylation (**c**), and protein with N-terminal 4 amino acids truncated (**d**).

ribosomal subunit protein Ser18, methylation of 50S ribosomal subunit protein Lys33, and methylthiolation of 30S ribosomal subunit protein Ser12 (c.f. Supplementary Table 2). Finally, for all 30 proteins with an N-terminal peptide removed, SignalP3.0 (http://www.cbs.Dtu.dk/services/SignalP) [44] was used to verify the presence of a signal peptide. Indeed, 18 proteins are predicted to contain the same signal peptide as found by PIITA (Table 1). For the remaining 12 proteins, SignalP3.0 predicted either no signal peptides (nine proteins) or a different signal peptide position (three proteins). We also identified four proteins with a truncated C-terminus (c.f. Supplementary Table 2).

Comparison to Other Top-Down Software

Compared with the number of protein identification software tools designed to identify proteins from peptide tandem mass spectra, there are very few that utilize protein tandem mass spectra. ProSight PTM, the first to be released, is the most commonly used top-down protein analysis program. In contrast to PIITA's goal to identify as many proteins as possible, ProSight PTM was designed to map thoroughly as many PTMs from a single protein tandem mass spectrum as possible [31, 45, 46]. However, ProSight PTM does offer the ability to query a comprehensive human protein database that

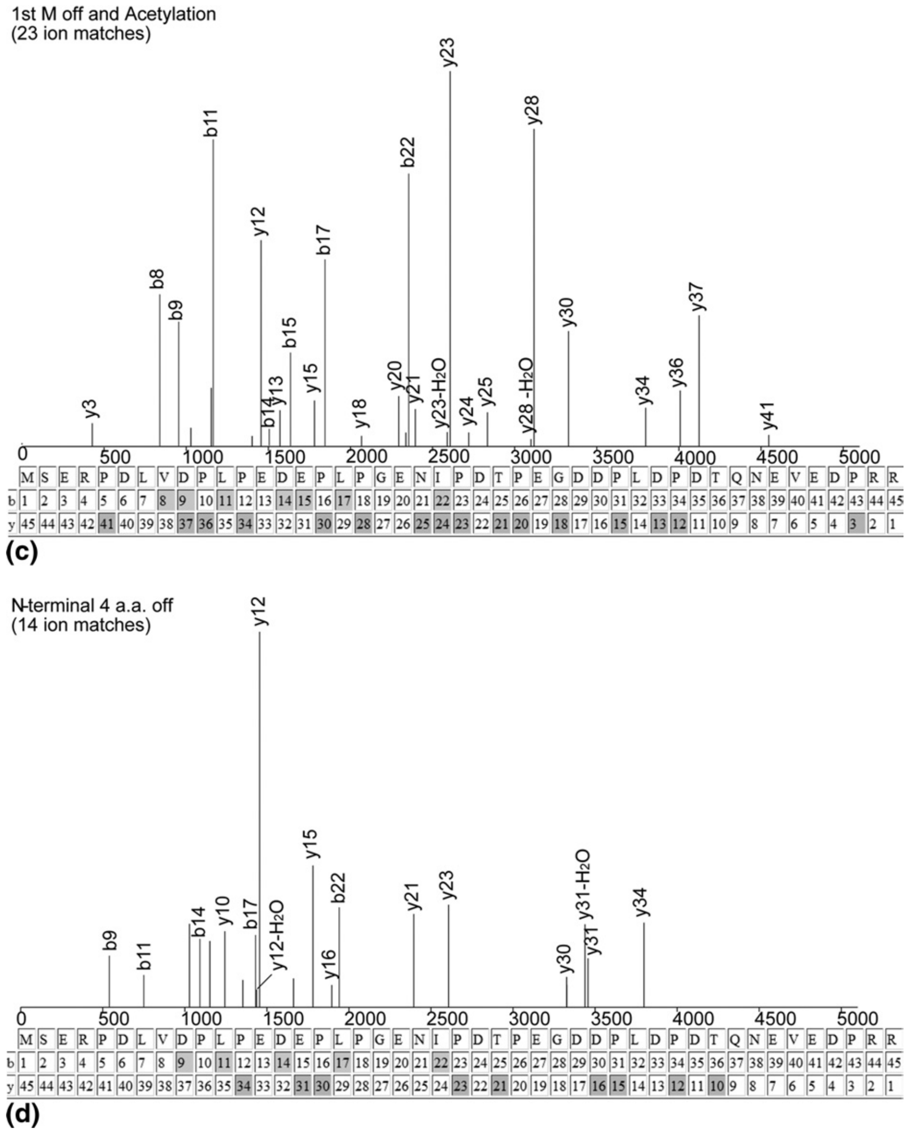


Figure 5. Continued.

includes 38,594 primary protein sequences, as well as a secondary list of 601,997 protein forms. One other notable difference between the two algorithms is the use by ProSight PTM of precursor ion mass as a primary database search screen. While PIITA search times are independent of precursor ion mass, ProSight PTM search time is directly proportional to precursor ion mass search window. Thus, the larger the precursor ion mass search window, the longer the search time for ProSight PTM. This presents obvious advantages to PIITA when trying to identify as many proteins as possible from as many data files as possible because less time will be required for the database search; see Supplementary Figure 2 for the results of a direct comparison of search time between PIITA and ProSight PTM. Another obvious difference between the two routines, which may provide some advantage to PIITA, is the computation and use of neutral losses of H_2O and NH_3 that allows for an increase in the number of matching fragment ions and is a common practice in bottom-up search routines.

The other top-down sequence assignment program, fully described in the literature, is MS-TopDown, which uses spectral alignment to find known/unknown protein modifications in an intact protein. This is done by considering all possible combinations of modified amino acids, but requires that an expected number of known and unknown PTMs be specified to initiate the search. Both acquired precursor ion mass and fragment ion masses are used by MS-TopDown to make an identification of the correct protein form. As is the case with PIITA, MS-TopDown uses precursor ion mass to calculate the difference in mass between observed and theoretical. Unlike ProSight PTM, both PIITA and MS-TopDown have no requirement to input suspected PTM mass before analysis. This approach provides for a wider search space, which is obviously useful in discovery experiments. One current limitation of the pub-

Table 1. Proteins predicted by SignalP 3.0 to lose a signal peptide identified by PIITA

	Protein and accession number	Mthe*	ΔM†	Signal peptide	SignalP-NN result
1	NP_459810.1 outer membrane protease precursor	18539.8984	−2245	1–23	C-score = 0.994
2	NP_459230.1 outer membrane protein H precursor	17894.3486	−2041	1–20	C-score = 1
3	NP_460463.1 putative periplasmic protein	13009.3997	−3043	1–28	C-score = 0.906
4	NP_463300.1 cytochrome b562	13903.3262	−2261	1–22	C-score = 0.932
5	NP_461092.1 putative periplasmic protein	11884.952	−2102	1–19	C-score = 0.964
6	NP_459379.1 phosphate starvation-inducible protein	11830.0402	−2201	1–22	C-score = 0.35
7	NP_459361.1 putative periplasmic protein	9913.14704	−2257	1–21	C-score = 0.977
8	NP_462272.1 putative periplasmic protein	9528.86712	−2264	1–22	C-score = 1
9	NP_460184.1 putative outer membrane protein	8794.56477	−2250	1–22	C-score = 0.817
10	NP_460216.1 putative periplasmic protein	12681.3018	−2592	1–24	C-score = 0.76
11	NP_460266.1 putative periplasmic protein	10441.2651	−2026	1–19	C-score = 0.622
12	NP_460460.1 putative outer membrane protein	10682.2153	−2154	1–22	C-score = 0.839
13	NP_460475.1 putative periplasmic protein	14376.4427	−2108	1–19	C-score = 1
14	NP_460832.1 putative copper resistance protein	13312.9803	−2654	1–26	C-score = 0.588
15	NP_462271.1 putative outer membrane protein	9141.55608	−2279	1–22	C-score = 1
16	NP_459087.1 putative secreted protein	10296.4327	−2157	1–21	C-score = 0.732
17	NP_462453.1 putative outer membrane protein	15289.6222	−1864	1–17	C-score = 0.915
18	NP_459739.1 hypothetical protein STM0759	13436.2137	−2371	1–24	C-score = 0.963

*Mthe = theoretical mass of precursor ion.
†ΔM = mass difference between theoretical mass and expected precursor mass.

licly available version of MS-TopDown (v1.0) is that only one protein characterization may be processed at a time and, for now, it works only with ECD data. Not unexpectedly, all three programs, ProSight PTM, MS-TopDown, and PIITA, were able to confidently identify the known PTM sites in the histone H4 ECD dataset. Unfortunately, at the time of writing this manuscript, there is no access to BigMascot at the Matrix Science website, which precludes further comparisons. However, from the limited description of how BigMascot functions [35], there do appear to be some similarities to PIITA, the most obvious being acceptance of CID data and a precursor ion independent option. This modification to Mascot suggests that other bottom-up software pipelines that incorporate accurate mass of fragment ions could also be modified to accept top-down data, but scoring routines may have to be adjusted accordingly. As with the comparisons between bottom-up algorithms that demonstrate different results on the same data [47], we expect that differences among top-down routines will provide select advantages/disadvantages to the user making their use in combination appealing.

Comparison to Bottom-Up Protein Identification

Finally, our PIITA results on top-down data were compared with a traditional "bottom-up" analysis carried out on the same extract from *S. Typhimurium*. Standard bottom up shotgun proteomics of this extract identified 300 proteins with high probability [29] of which 81 were identified by PIITA top-down analysis, the remaining 219 being identified by bottom-up only (Figure 6). Interestingly, we note that of the 154 proteins identified by PIITA, 73 proteins were not identified by bottom-up analysis. These top-down-only identified proteins have a median molecular weight of 10,600. This contrasts to a median molecular weight for the 219 proteins only identified by the bottom-up method of 27,848. The most obvious explanations for this disparity in molecular weight of proteins identified between the two methods are: (1) the bottom up approach is biased toward proteins of larger molecular weight which produce many more peptides from proteolysis of their parent proteins than smaller molecular weight proteins, and (2) in the top-down approach, proteins > 28 kDa were not readily observed because tandem mass spectra of proteins that are larger than 28 kDa exceed the LTQOT resolution used to acquire this data, making deconvolution inadequate to resolve enough fragment ions to match a theoretical tandem mass spectrum or they are of poor quality due to an insufficient number of ions in the ion trap. Thus, our top-down approach as carried out using CID in the Orbitrap was limited to small molecular weight proteins, which may be an advantage in the hunt for novel proteins. Additionally, as demonstrated with the ECD data of histone H4, there is no fundamental reason why the PIITA algorithm may not be used with traditional top-down data and data from larger proteins.

Conclusions

In summary, we developed and validated a precursor ion independent top-down algorithm (PIITA) for use in automated assignment of protein identifications from CID tandem mass spectra of whole proteins. From a standard solution containing six known proteins, we identified those six proteins as well as 61 isoforms at a 0% FDR. We also demonstrated PIITA is able to process standard ECD top-down data of a single protein and locate PTMs. Additionally, we used PIITA to characterize the proteins present in a previously uncharacterized

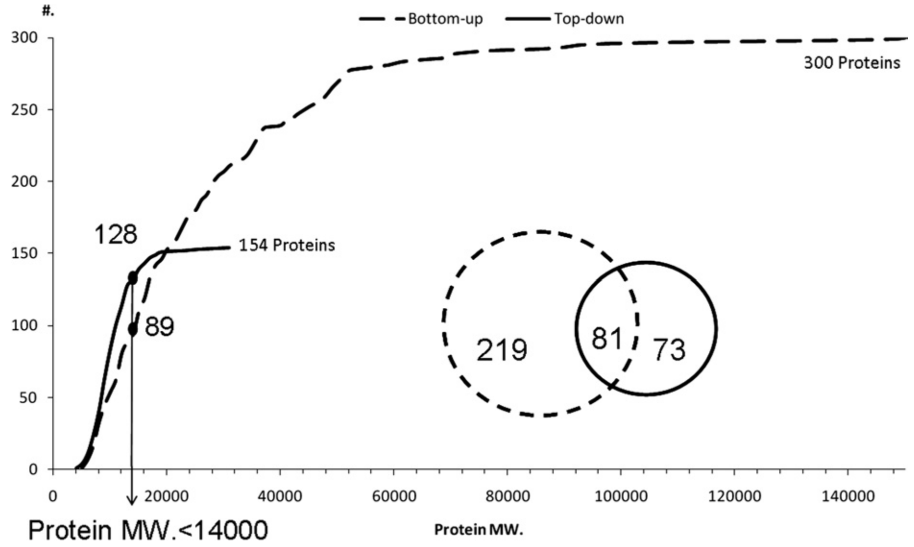


Figure 6. Cumulative distribution of protein molecular weight. Top-down identification (solid line) and bottom-up identification (dotted line) where insert shows total identified proteins of top-down (solid line) and bottom-up (dotted line) and the identified proteins in common (overlap).

Salmonella typhimurium outer membrane extract. Here, we used GPF to circumvent traditional solution-phase separation of proteins making this top-down process complementary to our standard bottom-up proteomic pipeline where we minimize sample preparation before MS analysis by genome-based GPF [29]. From this top-down approach, which differs from our bottom-up strategy in that we use C4 (instead of C18) packing material and do not digest the proteins, we identified 154 proteins from the *Salmonella typhimurium* outer membrane extract. Out of these 154 proteins, 73 were not identified in a parallel bottom-up analysis and 201 isoforms of the 154 proteins were identified at a FDR of < 1%. We believe this difference results in large part from the fact that the bottom-up process is biased towards proteins of higher molecular weight that produce more peptides, and a bias of our top-down process to proteins of low molecular weight because of limitations in the LTQ-Orbitrap.

Acknowledgments

D.R.G. thanks National Institutes of Health grants 5 U54 AI057141, 1S10RR023044-01, R33CA099139-01, and the University of Washington's School of Pharmacy Mass Spectrometry Center for support. The authors thank Ms. Carol Holman for providing the *Salmonella typhimurium* outer membrane extract. LM and PRRLS thank the BBSRC/EPSRC RASOR consortium for financial support.

Appendix A
Supplementary Material

Supplementary material associated with this article may be found in the online version at doi:10.1016/j.jasms.2009.07.024.

References

1. Yates, J. R., 3rd. Mass Spectral Analysis in Proteomics. *Annu. Rev. Biophys. Biomol. Struct.* **2004,** *33,* 297–316.
2. Wu, C. C.; MacCoss, M. J.; Howell, K. E.; Yates, J. R., 3rd. A Method for the Comprehensive Proteomic Analysis of Membrane Proteins. *Nat. Biotechnol.* **2003,** *21*(5), 532–538.
3. Aebersold, R.; Mann, M. Mass Spectrometry-Based Proteomics. *Nature* **2003,** *422*(6928), 198–207.
4. Siuti, N.; Kelleher, N. L. Decoding Protein Modifications Using Top-Down Mass Spectrometry. *Nat. Methods* **2007,** *4*(10), 817–821.
5. Sze, S. K.; Ge, Y.; Oh, H.; McLafferty, F. W. Top-Down Mass Spectrometry of a 29-kDa Protein for Characterization of any Posttranslational Modification to Within One Residue. *Proc. Natl. Acad. Sci. U.S.A.* **2002,** *99*(4), 1774–1779.
6. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science* **1989,** *246*(4926), 64–71.
7. Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. Evaluation of Multidimensional Chromatography Coupled with Tandem Mass Spectrometry (LC/LC-MS/MS) for Large-Scale Protein Analysis: The Yeast Proteome. *J. Proteome Res.* **2003,** *2*(1), 43–50.
8. McLuckey, S. A.; Stephenson, J. L., Jr. Ion/Ion Chemistry of High-Mass Multiply Charged Ions. *Mass Spectrom. Rev.* **1998,** *17*(6), 369–407.
9. Pyrek, J. S. Mass Spectrometry at Low and High Mass. *Curr. Opin. Chem. Biol.* **1997,** *1*(3), 399–409.
10. Scherl, A.; Shaffer, S. A.; Taylor, G. K.; Hernandez, P.; Appel, R. D.; Binz, P. A.; Goodlett, D. R. On the Benefits of Acquiring Peptide Fragment Ions at High Measured Mass Accuracy. *J. Am. Soc. Mass Spectrom.* **2008,** *19*(6), 891–901.
11. Hu, Q.; Noll, R. J.; Li, H.; Makarov, A.; Hardman, M.; Graham Cooks, R. The Orbitrap: A New Mass Spectrometer. *J. Mass Spectrom.* **2005,** *40*(4), 430–443.
12. Syka, J. E.; Marto, J. A.; Bai, D. L.; Horning, S.; Senko, M. W.; Schwartz, J. C.; Ueberheide, B.; Garcia, B.; Busby, S.; Muratore, T.; Shabanowitz, J.; Hunt, D. F. Novel Linear Quadrupole Ion Trap/FT Mass Spectrometer: Performance Characterization and Use in the Comparative Analysis of Histone H3 Post-Translational Modifications. *J. Proteome Res.* **2004,** *3*(3), 621–626.
13. Hoopmann, M. R.; Finney, G. L.; MacCoss, M. J. High-Speed Data Reduction, Feature Detection, and MS/MS Spectrum Quality Assessment of Shotgun Proteomics Data Sets Using High-Resolution Mass Spectrometry. *Anal. Chem.* **2007,** *79*(15), 5620–5632.
14. Tabb, D. L.; Shah, M. B.; Strader, M. B.; Connelly, H. M.; Hettich, R. L.; Hurst, G. B. Determination of Peptide and Protein Ion Charge States by Fourier Transformation of Isotope-Resolved Mass Spectra. *J. Am. Soc. Mass Spectrom.* **2006,** *17*(7), 903–915.
15. Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. Automated Reduction and Interpretation of High Resolution Electrospray Mass Spectra of Large Molecules. *J. Am. Soc. Mass Spectrom.* **2000,** *11*(4), 320–332.
16. Bogdanov, B.; Smith, R. D. Proteomics by FTICR Mass Spectrometry: Top Down and Bottom Up. *Mass Spectrom. Rev.* **2005,** *24*(2), 168–200.
17. Du, Y.; Meng, F.; Patrie, S. M.; Miller, L. M.; Kelleher, N. L. Improved Molecular Weight-Based Processing of Intact Proteins for Interroga-

tion by Quadrupole-Enhanced FT MS/MS. *J. Proteome Res.* **2004,** *3*(4), 801–806.

18. Henry, K. D.; Williams, E. R.; Wang, B. H.; McLafferty, F. W.; Shabanowitz, J.; Hunt, D. F. Fourier-Transform Mass Spectrometry of Large Molecules by Electrospray Ionization. *Proc. Natl. Acad. Sci. U.S.A.* **1989,** *86*(23), 9075–9078.
19. Mikesh, L. M.; Ueberheide, B.; Chi, A.; Coon, J. J.; Syka, J. E.; Shabanowitz, J.; Hunt, D. F. The Utility of ETD Mass Spectrometry in Proteomic Analysis. *Biochim. Biophys. Acta* **2006,** *1764*(12), 1811–1822.
20. Syka, J. E.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. Peptide and Protein Sequence Analysis by Electron Transfer Dissociation Mass Spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **2004,** *101*(26), 9528–9533.
21. Zubarev, R. A.; Horn, D. M.; Fridriksson, E. K.; Kelleher, N. L.; Kruger, N. A.; Lewis, M. A.; Carpenter, B. K.; McLafferty, F. W. Electron Capture Dissociation for Structural Characterization of Multiply Charged Protein Cations. *Anal. Chem.* **2000,** *72*(3), 563–573.
22. Axelsson, J.; Palmblad, M.; Hakansson, K.; Hakansson, P. Electron Capture Dissociation of Substance P Using a Commercially Available Fourier Transform Ion Cyclotron Resonance Mass Spectrometer. *Rapid Commun. Mass Spectrom.* **1999,** *13*(6), 474–477.
23. Macek, B.; Waanders, L. F.; Olsen, J. V.; Mann, M. Top-Down Protein Sequencing and MS3 on a Hybrid Linear Quadrupole Ion Trap-Orbitrap Mass Spectrometer. *Mol. Cell. Proteomi.* **2006,** *5*(5), 949–958.
24. Sharma, S.; Simpson, D. C.; Tolic, N.; Jaitly, N.; Mayampurath, A. M.; Smith, R. D.; Pasa-Tolic, L. Proteomic Profiling of Intact Proteins Using WAX-RPLC 2-D Separations and FTICR Mass Spectrometry. *J. Proteome Res.* **2007,** *6*(2), 602–610.
25. Parks, B. A.; Jiang, L.; Thomas, P. M.; Wenger, C. D.; Roth, M. J.; Boyne, M. T., 2nd, Burke, P. V.; Kwast, K. E.; Kelleher, N. L. Top-Down Proteomics on a Chromatographic Time Scale Using Linear Ion Trap Fourier Transform Hybrid Mass Spectrometers. *Anal. Chem.* **2007,** *79*(21), 7984–7991.
26. Bunger, M. K.; Cargile, B. J.; Ngunjiri, A.; Bundy, J. L.; Stephenson, J. L., Jr. Automated Proteomics of E. Coli via Top-Down Electron-Transfer Dissociation Mass Spectrometry. *Anal. Chem.* **2008,** *80*(5), 1459–1467.
27. Spahr, C. S.; Davis, M. T.; McGinley, M. D.; Robinson, J. H.; Bures, E. J.; Beierle, J.; Mort, J.; Courchesne, P. L.; Chen, K.; Wahl, R. C.; Yu, W.; Luethy, R.; Patterson, S. D. Towards Defining the Urinary Proteome Using Liquid Chromatography-Tandem Mass Spectrometry. I. Profiling an Unfractionated Tryptic Digest. *Proteomics* **2001,** *1*(1), 93–107.
28. Yi, E. C.; Marelli, M.; Lee, H.; Purvine, S. O.; Aebersold, R.; Aitchison, J. D.; Goodlett, D. R. Approaching Complete Peroxisome Characterization by Gas-Phase Fractionation. *Electrophoresis* **2002,** *23*(18), 3205–3216.
29. Scherl, A.; Shaffer, S. A.; Taylor, G. K.; Kulasekara, H. D.; Miller, S. I.; Goodlett, D. R. Genome-Specific Gas-Phase Fractionation Strategy for Improved Shotgun Proteomic Profiling of Proteotypic Peptides. *Anal. Chem.* **2008,** *80*(4), 1182–1191.
30. Zamdborg, L.; LeDuc, R. D.; Glowacz, K. J.; Kim, Y. B.; Viswanathan, V.; Spaulding, I. T.; Early, B. P.; Bluhm, E. J.; Babai, S.; Kelleher, N. L. ProSight PTM 2.0: Improved Protein Identification and Characterization for Top Down Mass Spectrometry. *Nucleic Acids Res.* **2007,** *35*(Web Server issue), W701–W706.
31. LeDuc, R. D.; Taylor, G. K.; Kim, Y. B.; Januszyk, T. E.; Bynum, L. H.; Sola, J. V.; Garavelli, J. S.; Kelleher, N. L. ProSight PTM: An Integrated Environment for Protein Identification and Characterization by Top-Down Mass Spectrometry. *Nucleic Acids Res.* **2004,** *32*(Web Server issue), W340–W345.
32. Frank, A. M.; Pesavento, J. J.; Mizzen, C. A.; Kelleher, N. L.; Pevzner, P. A. Interpreting Top-Down Mass Spectra Using Spectral Alignment. *Anal. Chem.* **2008,** *80*(7), 2499–2505.
33. Pevzner, P. A.; Dancik, V.; Tang, C. L. Mutation-Tolerant Protein Identification by Mass Spectrometry. *J. Comput. Biol.* **2000,** *7*(6), 777–787.
34. Pevzner, P. A.; Mulyukov, Z.; Dancik, V.; Tang, C. L. Efficiency of Database Search for Identification of Mutated and Modified Proteins via Mass Spectrometry. *Genome Res.* **2001,** *11*(2), 290–299.
35. Karabacak, N. M.; Li, L.; Tiwari, A.; Hayward, L. J.; Hong, P.; Easterling, M. L.; Agar, J. N. Sensitive and Specific Identification of Wild Type and Variant Proteins from 8 to 669 kDa Using Top-Down Mass Spectrometry. *Mol. Cell Proteom.* **2009,** *8*(4), 846–856.
36. Mackay, C. L.; Ramsahoye, B.; Burgess, K.; Cook, K.; Weidt, S.; Creanor, J.; Harrison, D.; Langridge-Smith, P.; Hupp, T.; Hayward, L. Sensitive, Specific, and Quantitative FTICR Mass Spectrometry of Combinatorial Post-Translational Modifications in Intact Histone H4. *Anal. Chem.* **2008,** *80*(11), 4147–4153.
37. Colinge, J.; Masselot, A.; Carbonell, P.; Appel, R. D. InSilicoSpectro: An Open-Source Proteomics Library. *J. Proteome Res.* **2006,** *5*(3), 619–624.
38. Eng, J. K.; McCormack, A. L.; Yates, J. R., 3rd. 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*, 976–989.
39. Eng, J. K.; Fischer, B.; Grossmann, J.; Maccoss, M. J. A Fast SEQUEST Cross Correlation Algorithm. *J. Proteome Res.* **2008,** *7*(10), 4598–4602.
40. Strittmatter, E. F.; Ferguson, P. L.; Tang, K.; Smith, R. D. Proteome Analyses Using Accurate Mass and Elution Time Peptide Tags with Capillary LC Time-of-Flight Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2003,** *14*(9), 980–991.
41. Savitski, M. M.; Nielsen, M. L.; Zubarev, R. A. ModifiComb, A New Proteomic Tool for Mapping Sub-Stoichiometric Post-Translational Modifications, Finding Novel Types of Modifications, and Fingerprinting Complex Protein Mixtures. *Mol. Cell. Proteom.* **2006,** *5*(5), 935–948.
42. Panchaud, A.; Scherl, A. S. S.; von Haller, P. D.; Kulasekara, H. D.; Miller, S. I.; Goodlett, D. R. PAcIFIC: How to Dive Deeper Into the Proteomics Ocean. *Anal. Chem.* **2009,** press.
43. Arnold, R. J.; Reilly, J. P. Observation of *Escherichia Coli* Ribosomal Proteins and Their Post-Translational Modifications by Mass Spectrometry. *Anal. Biochem.* **1999,** *269*(1), 105–112.
44. Bendtsen, J. D.; Nielsen, H.; von Heijne, G.; Brunak, S. Improved Prediction of Signal Peptides: SignalP 3.0. *J. Mol. Biol.* **2004,** *340*(4), 783–795.
45. Taylor, G. K.; Kim, Y. B.; Forbes, A. J.; Meng, F.; McCarthy, R.; Kelleher, N. L. Web and Database Software for Identification of Intact Proteins Using "Top Down" Mass Spectrometry. *Anal. Chem.* **2003,** *75*(16), 4081–4086.
46. Mann, M.; Wilm, M. Error-Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags. *Anal. Chem.* **1994,** *66*(24), 4390–4399.
47. Kapp, E. A.; Schutz, F.; Connolly, L. M.; Chakel, J. A.; Meza, J. E.; Miller, C. A.; Fenyo, D.; Eng, J. K.; Adkins, J. N.; Omenn, G. S.; Simpson, R. J. An Evaluation, Comparison, and Accurate Benchmarking of Several Publicly Available MS/MS Search Algorithms: Sensitivity and Specificity Analysis. *Proteomics* **2005,** *5*(*13*), 3475–3490.