research articles Journal of proteome research

Enhanced Peptide Mass Fingerprinting through High Mass Accuracy: Exclusion of Non-Peptide Signals Based on Residual Mass

Eric D. Dodds,[†] Hyun Joo An,[†] Paul J. Hagerman,[‡] and Carlito B. Lebrilla*[,†,‡]

Department of Chemistry, and Department of Biochemistry and Molecular Medicine, School of Medicine, University of California Davis, One Shields Avenue, Davis, California 95616

Received December 27, 2005

Peptide mass fingerprinting (PMF) is among the principle methods of contemporary proteomic analysis. While PMF is routinely practiced in many laboratories, the complexity of protein tryptic digests is such that PMF based on unrefined mass spectrometric peak lists is often inconclusive. A number of data processing strategies have thus been designed to improve the quality of PMF peak lists, and the development of increasingly elaborate tools for PMF data reduction remains an active area of research. In this report, a novel and direct means of PMF peak list enhancement is suggested. Since the monoisotopic mass of a peptide must fall within a predictable range of residual values, PMF peak lists can in principle be relieved of many non-peptide signals solely on the basis of accurately determined monoisotopic mass. The calculations involved are relatively simple, making implementation of this scheme computationally facile. When this procedure for peak list processing was used, the large number of unassigned masses typical of PMF peak lists was considerably attenuated. As a result, protein identifications could be made with greater confidence and improved discrimination as compared to PMF queries submitted with raw peak lists. Importantly, this scheme for removal of non-peptide masses was found to conserve peptides bearing various post-translational and artificial modifications. All PMF experiments discussed here were performed using Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS), which provided the high mass resolution and high mass accuracy essential for this application. Previously reported equations relating the nominal peptide mass to the permissible range of fractional peptide masses were slightly modified for this application, and these adjustments have been illustrated in detail. The role of mass accuracy in application of this scheme has also been explored.

Keywords: Peptide mass fingerprinting • peptide residual mass • matrix-assisted laser desorption/ionization • Fourier transform ion cyclotron resonance mass spectrometry

Introduction

Peptide mass fingerprinting (PMF) is one of the core protein identification technologies of modern proteomics.[1−4] In a typical PMF experiment, a target protein is digested with trypsin, and the resulting peptide mixture is analyzed by mass spectrometry (MS). Mass analysis for PMF is most commonly performed using matrix-assisted laser desorption/ionization (MALDI) coupled to a time-of-flight (TOF) mass analyzer. Using a proteomics search engine, the experimentally observed peptide masses are compared with predicted in silico digests of proteins cataloged in the queried database. Proteins potentially matching the experimental peptide mass fingerprint are reported, along with some numeric expression of the match quality.

Perhaps the most widely used search engine for PMF is the Mascot platform (www.matrixscience.com). Potential protein matches reported by Mascot are accompanied by probabilistic molecular weight search (Mowse) scores.[5] In this scoring scheme, each putative protein match is assigned an estimated absolute probability of the match occurring at random. For simplicity and clarity, the Mowse score S_M is reported as a logarithmic transform of the random match probability P

$$S_M = -10 \log(P) \quad (1)$$

Generally, a Mowse score is considered indicative of a significant protein match if the corresponding value of P is expected to occur by chance with a frequency of less than 5%. A statistically relevant protein identification can, in principle, be made based on the 2-fold specificity that PMF derives from combining mass measurement of tryptic peptides with the high fidelity of trypsin for cleavage at specific sites.[6,7]

Many factors affect the quality of a PMF search outcome. These factors include a number of analyst-defined search

* To whom correspondence should be addressed. E-mail: cblebrilla@ucdavis.edu. Telephone: 1-530-752-6364. Fax: 1-530-754-5609.
[†] Department of Chemistry, University of California Davis.
[‡] Department of Biochemistry and Molecular Medicine, School of Medicine, University of California Davis.

10.1021/pr050486o CCC: $33.50 © 2006 American Chemical Society

Published on Web 04/19/2006

parameters, such as inclusion of variable amino acid modifications, tolerance of missed tryptic cleavage sites, and mass tolerance. Other important considerations include the number of protein entries in the queried database, the number of matching peptide masses and their relative frequency of occurrence in the database, and the number of unassigned peaks in a submitted query. All of these factors can have a profound influence on whether a peptide mass fingerprint can be associated with a protein correctly and with statistical significance. Even when appropriate search parameters are specified, obtaining a meaningful protein match by PMF alone often remains a challenging task, as potential protein identities returned by PMF searches may be ambiguous or entirely inconclusive for a number of reasons. There are two principal culprits to which these indecisive outcomes may be ascribed: partial protein sequence coverage and the presence of unassigned masses.

Sequence coverage in PMF is often well under 50%, even under the most favorable circumstances. Peptides representing the remainder of the protein may be missed due to unknown post-translational modifications (PTMs), unintended modifications related to sample preparation, mutation or splice variation of the protein, or errors in the cataloged genome from which predicted peptides are derived. Differences in ionization efficiency among peptides and the biases of a given ionization source can also result in missed peptides.[8,9] In addition, peptide mass fingerprints are known to contain a large number of peaks not attributable to tryptic cleavage of the target protein.[10] These unassigned masses may arise from some causes of missed coverage cited above (e.g., unexpected modifications and database errors). Additional sources of extraneous masses may include contamination with human keratin peptides, trypsin autolysis peptides, or peptides from incompletely resolved proteins contained in a single gel spot following electrophoresis. Non-peptide adulterants may include protein stains, alkali-matrix clusters produced during MALDI, and other contaminants occurring as a consequence of biological sample complexity and the elaborate sample preparation involved.

Because of these challenges, the development of data processing strategies for improving the quality of PMF data sets has remained an active subject of research.[11] For example, several algorithms for refining PMF peak lists make use of isotopomer distribution fitting based on average amino acid elemental composition (the aptly named "averagine" residue).[12−14] Another method incorporates the use of Poisson statistics to extract monoisotopic peptide masses.[15] An iterative approach involving successive rounds of peak list refinement to optimize the outcome of PMF searches has also been recently suggested.[16] Other workers have developed a scheme in which multiple peak lists are combined in order to arrive at a more representative list.[17] Several of these algorithms are accompanied by postprocessing tools for removal of spurious peaks based on user-defined lists of contaminant masses.

Since the presence of unassigned peaks detracts from the significance of probability-based Mowse scores, one obvious means of improving the quality of PMF peak lists is to reduce the number of non-peptide masses in the query. Since the residual mass ranges inherent to peptides have been well-characterized, simple mass measurement-based criteria for exclusion of many non-peptide masses would seem feasible. The monoisotopic masses of peptides are known to be normally distributed about specific fractional values for a given nominal mass. Conveniently, the centroid and width of this distribution can be predicted based on the nominal mass alone. For a peptide of nominal mass M_n (i.e., the lower integer of the monoisotopic mass), the centroid mass M_c is given by

$$M_c = M_n + 0.00048M_n \quad (2)$$

The width about the centroid, W_c, encompassing 95% of all possible peptides can be obtained by

$$W_c = 0.19 + 0.0001M_n \quad (3)$$

These relationships were first proposed by Mann,[18] and their implications were further explored by Zubarev et al.[19]

A number of researchers have developed applications based on peptide mass distributions calculated in this manner. For example, the use of peptide fractional masses as a means of internal mass calibration for MALDI-TOF-MS has been discussed.[13,20] Some authors have suggested the use of predicted peptide residual masses as justification for removal of alkali-matrix clusters produced by MALDI.[10,21] Predicted peptide mass distributions were also recently used to exclude non-peptides and thus simplify spectral interpretation in stable isotope labeling proteomics experiments, although these studies were not concerned with PMF.[22] Interestingly, the natural distribution of peptide masses has also been viewed as an adverse factor in the analysis of tryptic digests. The use of oxidative treatment to expand the distribution of peptide masses has thus been suggested as a means of providing more unique peptide mass signatures.[23] While the broader spectrum of fractional masses among oxidized peptides was intended to provide more distinct peptide masses, it could be contrarily argued that the natural distribution of peptide masses is itself unique and useful as a distinguishing characteristic.

Surprisingly, there has been no example of peptide fractional mass distribution applied as a means of excluding non-peptide masses from PMF peak lists. This is most likely because many MS instruments used in proteomic research do not provide sufficient mass resolution and mass accuracy to allow confident rejection of non-peptides based on residual mass alone. Certainly, Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS) is able to meet the performance demands of such an application. FTICR-MS allows mass-to-charge ratios (m/z) to be determined with errors in the low parts per million (ppm) range, and provides mass resolving power on the order of 10^5 ($m/\Delta m_{FWHM}$) in broadband detection mode.[24] These qualities render FTICR-MS an exceptionally capable tool for PMF.[25−28]

Here, we describe the implementation of a new and relatively simple data processing workflow that includes the rigorous application of residual mass-based criteria for mitigating the number of non-peptide signals in PMF peak lists. When FTICR is used as the MS platform, the exclusion via fractional mass approach is simple and elegant, requiring only the accurate monoisotopic mass and trivial calculations to determine whether a spectral peak can potentially be attributed to a peptide. Protein tryptic digests were analyzed by MALDI−FTICR-MS, and PMF queries using raw and refined peak lists were compared. Slight modifications required to generalize eqs 2 and 3 for this purpose have also been implemented. In addition, the theoretical masses of peptides bearing various PTMs and artificial modifications have been subjected to this data-processing strategy in order to explore what peptide modifications are spared or rejected. This is an important consideration, given that the mathematical relationships describing peptide

mass distributions were developed with consideration of unmodified peptides only. Finally, the mass accuracy requirement for successful application of this scheme has been explored.

Experimental Section

Tryptic Digestion. Bovine serum albumin (BSA) and human apo-transferrin (HAT) were obtained from Sigma (St. Louis, MO). Human plasma fibrinogen (HPF) was purchased from Calbiochem (LaJolla, CA). Sequencing grade modified trypsin was obtained from Promega (Madison, WI).

Stock solutions of BSA, HAT, and HPF were prepared at 1 μg/μL in 8 M urea and 200 mM total tris (pH = 7.8). To prepare a stock tryptic digest for each protein, approximately 1 μg of protein (1 μL of the stock solution) was combined with 40 μL 8M urea/200 mM tris. The solution was treated with 10 μL 450 mM dithiothreitol in 50 mM NH_4HCO_3, and reduction was carried out by incubating at 55 °C for 1 h. For alkylation, 10 μL 500 mM iodoacetamide in 50 mM NH_4HCO_3 was added to the reduced protein solution. The mixture was then held in the dark at ambient temperature for 30 min. Following alkylation, each sample was diluted with 150 μL of deionized H_2O to bring the urea concentration to <2 M. Each sample was then treated with trypsin (1 μL of a 0.05 μg/μL solution in 50 mM NH_4HCO_3). For tryptic digestion, samples were incubated at 37 °C for approximately 8−10 h. A 10 μL aliquot of each tryptic digest was desalted by solid-phase extraction with C18 ZipTips (Millipore, Billerica, MA), and the purified tryptic peptides were eluted in 10 μL 50% acetonitrile (ACN) with 0.1% trifluoroacetic acid.

Mass Spectrometry. Purified tryptic digests were prepared for MALDI by spotting 1 μL (corresponding to approximately 100 fmol digested protein) on a stainless steel sample probe. Matrix solution (1 μL) was combined with the analyte solution on probe and allowed to dry. The applied matrix solution was 50 μg/μL 2,5-dihydroxybenzoic acid (DHB) in 50% ACN.

All MS analyses were performed using an IonSpec Corporation (Lake Forest, CA) FTICR-MS instrument equipped with a 7.0 T actively shielded superconducting magnet and an external MALDI source fitted with a frequency-tripled Nd:YAG laser (355 nm, 5 ns pulse width). In-cell accumulation of ions produced by a variable number of MALDI laser pulses was used to obtain optimum total ion intensity for each sample spot analyzed (typically, one to five pulses).

To achieve maximum mass measurement accuracy, mass spectra were internally calibrated using the molecular ion and *y* series fragments of $P_{14}R$, a synthetic peptide so named for its sequence of 14 proline residues followed by a C-terminal arginine residue (Sigma). Conveniently, $P_{14}R$ provides several useful calibrant masses through in-source decay. This is due to the tendency of peptides to fragment at proline residues (the "proline effect").[29] Rather than adding the internal standard to the purified peptide solutions, a technique for combining analyte and standard ions in the gas phase was applied.[30,31] Calibrant spots (1 μL of 1 μM $P_{14}R$ combined with 1 μL of the DHB solution) were spotted adjacent to analyte spots on the MALDI target. Analyte ions produced by an optimum number of MALDI events were accumulated in the ICR cell, and the position of the MALDI target was then adjusted so that the adjacent internal standard spot was next irradiated by the MALDI laser. Internal calibrant ions (usually, from one or two MALDI laser pulses) were accumulated in-cell along with the previously trapped analyte ions, and the combined population of ions was simultaneously mass-analyzed. This procedure was dubbed "internal calibration on adjacent samples" (InCAS) by O'Connor and Costello.[31]

A quadrupole operated in RF-only mode was used as a broadband ion guide to direct ions produced in the external source to the ICR cell. At 10 ms prior to each MALDI laser shot, a pulse of buffer gas (argon) was leaked into the ion guide region of the instrument to vibrationally cool the ions. Positively charged ions were trapped in the cylindrical ICR cell by a 20 V potential applied to the front and rear trapping plates. The potential on the inner trapping rings was held constant at 0.5 V for the duration of the experiment. To allow ions to enter the cell, the rear trapping potential (i.e., the source side trapping plate) was dropped to 4 V for approximately 3 ms, beginning 1 ms prior to each MALDI pulse. Trapped ions in the m/z range 108−2500 were excited by means of an arbitrary waveform pulse (32 k waveform points applied at a DAC rate of 2 MHz, 150 V base to peak amplitude). Following ion acceleration, the front and rear trapping plate potentials were linearly ramped to zero over a 1 s duration. Acquisition of the time domain signal was commenced in broadband mode at an ADC rate of 2 MHz. The time domain signal comprised of 1024 k transient data points was zero-filled once, Blackman-apodized, and fast Fourier-transformed to yield the frequency domain spectrum.

Data Processing and Peptide Mass Fingerprint Analysis. Mass calibration was performed on the $P_{14}R$ internal calibrant masses in the IonSpec Omega software according to standard FTICR-MS calibration relationships.[32] Internally calibrated spectra were default-thresholded and isotope-filtered in the IonSpec PeakHunter software, and the monoisotopic $[M + H]^+$ peak list was exported to Microsoft Excel for further processing. All exact mass calculations were performed with the aid of the IonSpec Exact Mass Calculator.

A macro for refining the monoisotopic peak lists was written in Visual Basic for implementation in Excel. This macro, known in-house as Mass Sieve, was used to perform two operations on each peak list. First, peaks occurring within a specified tolerance of masses in a user-defined list (in this case, a list of internal standard masses) were removed from the list (the Standard Screener function). Any screened masses were reported along with the mass error between the screened mass and the user-specified mass. The screening tolerance for internal standard masses was set at 2 ppm. Second, signals with fractional masses not attributable to peptides were removed from the list (the Peptide Filter function). This was done according to eqs 2 and 3, with slight modifications. Any excluded masses were reported, along with the refined peak list. For comparison, the same original monoisotopic peak lists were processed with the Standard Screener function but not the Peptide Filter function. These raw and refined peak lists were converted to text file format. The text files were submitted to Mascot for PMF database searching via the Mascot Wizard (freely available for download at www.matrixscience.com/wizard.html). The Mass Spectrometry Protein Sequence Database (MSDB) was searched with a taxonomy specified as appropriate for each protein ("*Homo sapiens*" for HAT and HPF; "other mammalia" for BSA). Proteolysis with trypsin was specified, and carbamidomethylation of cysteine residues was included as a fixed modification. No variable modifications were considered. The mass tolerance (3−10 ppm) and the allowed number of missed tryptic cleavages (either 0 or 1) were

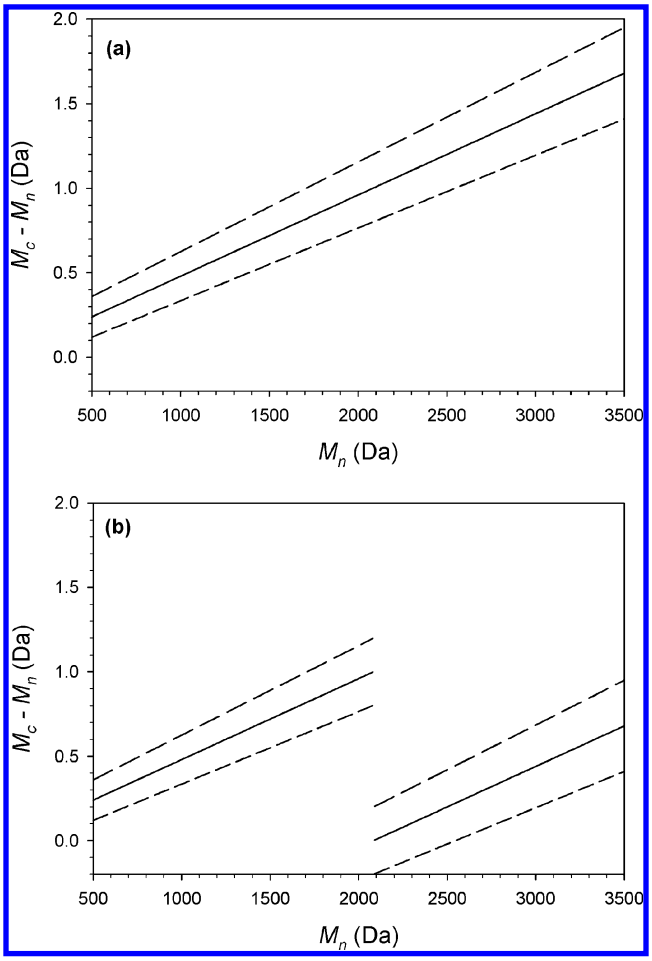

Figure 1. Predicted peptide residual mass range plotted as a function of nominal mass. The values of $M_c - M_n$ are calculated according to eq 2 (a) and eq 4 (b). The solid line is $M_c - M_n$ versus M_n, while the upper and lower dashed lines correspond to $M_c + W_c/2$ and $M_c - W_c/2$ versus M_n, respectively.

set on a case-by-case basis. Signals with m/z < 800 were ignored in all PMF searches.

Results and Discussion

Mass Sieve Peptide Filter Algorithm. While eqs 2 and 3 are well-established and have been deemed appropriately descriptive of the informative mass range for tryptic peptides, two adaptations were necessary for this application. The first of these adjustments was applied to eq 2. In calculating M_c for masses above a critical value of M_n, the fractional mass term (that is, 0.00048 M_n) added to the nominal mass exceeds unity. Specifically, it can be seen that this occurs when the nominal mass exceeds 2083 Da. To ensure addition of the appropriate residual mass to a given nominal mass, eq 2 must be modified to include an additional condition:

$$M_c = \begin{cases} M_n + 0.00048M_n & (M_n \leq 2083) \\ M_n + 0.00048M_n - 1 & (M_n > 2083) \end{cases} \quad (4)$$

Hence, the correct relationship for calculation of M_c over the entire relevant mass range is actually a conditional function, rather than a single equality (Figure 1).

A second complication arises as a consequence of the near-integer predicted values of M_c in the region of 2000 Da. In this area of the mass scale, the permissible fractional mass range

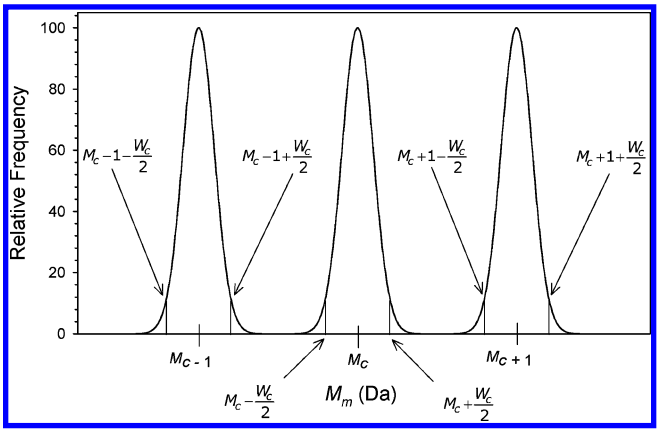

Figure 2. Theoretical peptide mass distributions with annotation of the critical cutoffs for exclusion of non-peptides, as described in eqs 5−8.

encompasses two nominal figures. Consequently, the peptide mass could mistakenly be compared to inappropriate values for M_n and W_c, and the peptide would be wrongly rejected. For the sake of example, it is useful to consider a peptide with an actual monoisotopic mass M_m of 2001.002 Da. For a peptide with a nominal mass M_n of 2001 Da, the predicted permissible mass range $M_c \pm (W_c/2)$ is 2001.960 ± 0.196 Da. Interpreted literally, this calculation would imply that the observed mass is not a peptide; however, the peptide does fall within the predicted mass range for the preceding nominal figure (2000.960 ± 0.195 Da). Unless this incrementation is accounted for in some way, the verbatim application of eqs 3 and 4 in an algorithm for excluding non-peptide masses can result in inadvertent exclusion of legitimate peptide masses. This nominal mass fault can similarly arise when the predicted value of M_c is slightly above its corresponding M_n, but the measured mass is decremented to the preceding nominal figure. To avoid inappropriate rejection of peptides due to single unit nominal mass offset, the Peptide Filter in Mass Sieve includes additional criteria for the rejection of observed masses. One of following four conditions must be met in order to reject a mass as non-peptide:

$$M_m > (M_c + 1) + \frac{W_c}{2} \quad (5)$$

$$M_m < (M_c - 1) - \frac{W_c}{2} \quad (6)$$

$$M_c + \frac{W_c}{2} < M_m < (M_c + 1) - \frac{W_c}{2} \quad (7)$$

$$(M_c - 1) + \frac{W_c}{2} < M_m < M_c - \frac{W_c}{2} \quad (8)$$

These conditions preserve all masses falling within the $M_c \pm (W_c/2)$ range for the submitted nominal mass M_n, as well as the corresponding ranges for $M_n \pm 1$ (Figure 2). For simplicity, W_c is not recalculated for $M_n \pm 1$, as the difference in magnitude of W_c is negligible for consecutive values of M_n. Now, using eqs 5−8, the example peptide of M_m 2001.002 Da is retained by virtue of being within the predicted residual mass range for M_n 2000 (i.e., 2000.960 ± 0.196).

Consideration of Modified Peptides. It is important to ascertain whether this scheme for eliminating extraneous signals based on residual mass also eliminates peptides bearing

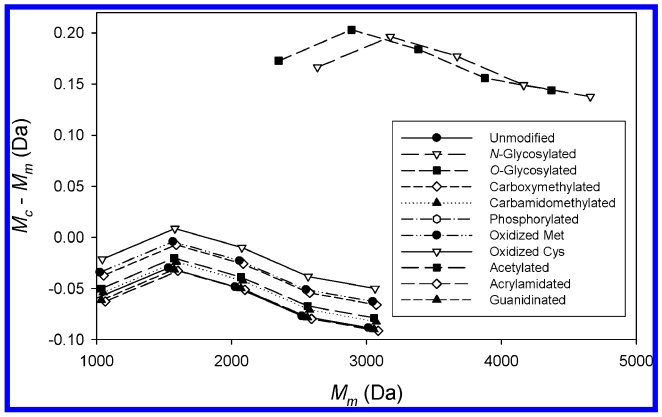

Figure 3. Distance from predicted peptide centroid as a function of monoisotopic mass for averagine peptides ranging in M_m from approximately 1000 to 3000 Da, and several modified counterparts of these peptides.

various post-translational and artificial modifications. In the derivation of eqs 2 and 3, only the compositions of unmodified peptides were considered. For this reason, some have assumed that modified peptides would not fall within the predicted fractional mass range for peptides.[10,16] Conversely, Gay et al. observed fractional mass tendencies very similar to those derived by Mann after assembling a large compilation of tryptic peptide masses from the Swiss-Prot database with consideration of all available modifications.[33]

To directly test whether the Mass Sieve Peptide Filter algorithm implemented here would reject modified peptide masses, theoretical monoisotopic masses of modified peptides were calculated and subjected to the peptide filtering scheme. Specifically, peptide masses approximating 1000, 1500, 2000, 2500, and 3000 Da were calculated using averagine residues ($C_{4.9384}H_{7.7583}N_{1.3577}O_{1.4773}S_{0.0417}$).[34] Each of these peptides was then shifted in mass as appropriate for each modification of interest. The following modifications were considered: carboxymethylation, carbamidomethylation, oxidation (as occurring on cysteine and methionine residues), phosphorylation, acetylation, acrylamidation, guanidination, N-linked glycosylation (with a glycan moiety composed of five hexose residues and four *N*-acetyl hexosamine residues), and O-linked glycosylation (with a glycan moiety composed of two hexose residues, two *N*-acetylhexosamine residues, and two sialic acid residues). All masses were calculated as monoisotopic $[M + H]^+$ masses, except for the glycan-modified peptides. Glycopeptide masses were calculated as the corresponding sodium adducts (i.e., $[M + Na]^+$ for the *N*-glycopeptide; $[M - H + 2Na]^+$ for the *O*-glycopeptide). When the calculated masses for these various modified peptides were processed through the Mass Sieve Peptide Filter, none were rejected as non-peptides. This is a significant finding, as the implementation of this data-processing step does not preclude the use of the refined peak lists for searches including many common natural and artificial modifications encountered in proteomics.

The difference between the modified peptide masses and the corresponding predicted centroid mass ($M_c - M_m$) is plotted as a function of the monoisotopic mass M_m in Figure 3. As expected, there were only small differences between the averagine peptide and the corresponding predicted centroids. Introducing single occurrences of the modifications listed above has rather little impact on the distance from the predicted centroid of a modified averagine peptide. The only exceptions

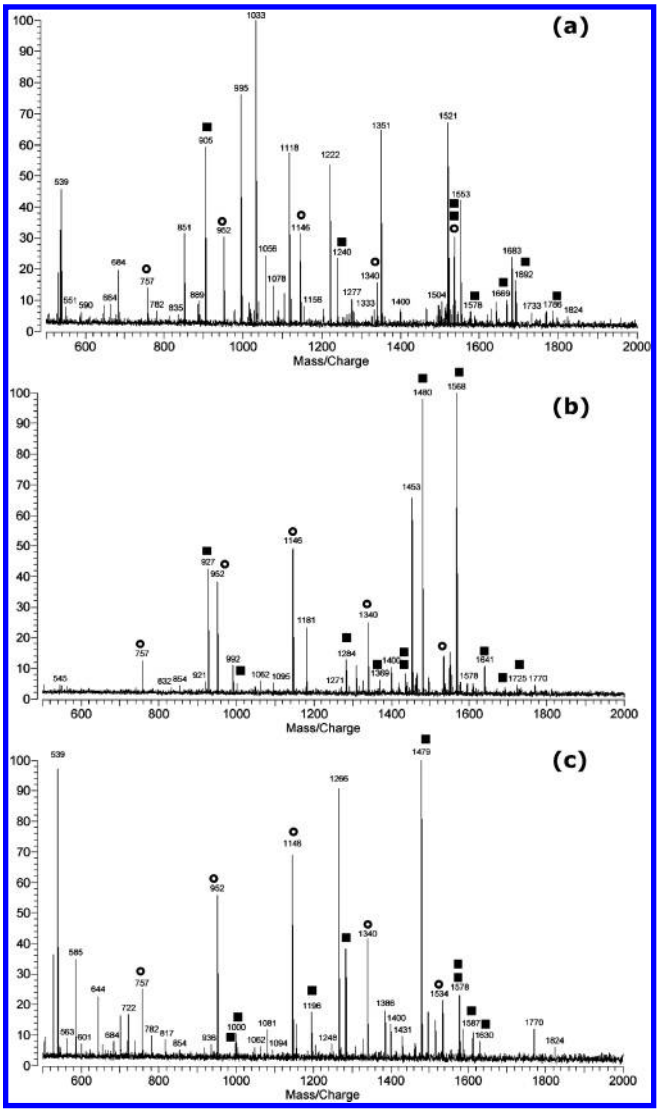

Figure 4. MALDI−FTICR mass spectra for PMF analysis of HPF (a), BSA (b), and HAT (c). Internal calibrant masses from $P_{14}R$ introduced by InCAS are labeled with open circles, while masses matched to the correct protein are labeled with closed squares.

to this observation are the examples involving glycosylated peptides. The glycopeptides are subject to a rather large shift away from the predicted centroid, even if the peptide portion is average in mass. The influence of glycan modifications is large relative to the other modifications, undoubtedly because the glycans are much greater in mass than the other modifications considered. Nevertheless, the glycopeptides still fall within the predicted peptide mass range. These calculations suggest the possibility that a peptide already near a tail of the predicted mass distribution may in fact be rejected upon further offset from the predicted centroid due to glycosylation. A similar situation may arise in the case of multiply-modified peptides. While these possibilities should be taken into consideration, such events would be expected to occur with relatively low frequency.

PMF Analysis Using Raw and Refined Peak Lists. The MALDI−FTICR mass spectra for PMF analysis of HPF, BSA, and HAT are shown in Figure 4. These particular spectra were chosen because they approximate the quality of spectra for realistic unknowns. The monoisotopic peak lists derived from these spectra were used in both their raw (processed using only

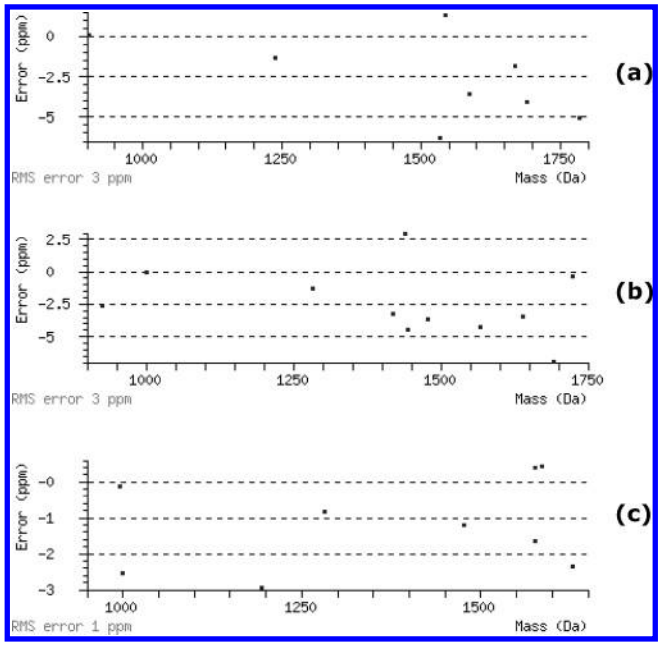

Figure 5. Mass error distributions for peptides matching HPF (a), BSA (b), and HAT (c). These were reported by Mascot following the PMF searches.

the Standard Screener of Mass Sieve) and refined (processed using both the Standard Screener and Peptide Filter functions of Mass Sieve) forms to query MSDB using the Mascot Wizard PMF tool with all other parameters held constant. For each protein, the same number of correctly matching peptides was obtained regardless of whether the raw or refined peak list was submitted. Thus, no useful masses were excluded by the processing. This was as expected, as the spectra were of high mass accuracy. The root-mean-square mass error for the peptides correctly matching a given protein did not exceed 3 ppm (Figure 5). In all cases, the probability-based Mowse scores for the correct protein matches were significantly improved by removal of non-peptide masses as allowed by the algorithm described above. Panels a and b of Figure 6 show the Mowse score distribution for HPF raw and refined peak lists, respectively. While the E and B chains of HPF were both implicated with statistical significance by the raw peak list (due to duplicate cataloging of some HPF sequence information in MSDB), the Mowse score was significantly improved in the case of the refined peak list. Figure 6c,d illustrates the results for PMF using BSA raw and refined peak lists. BSA was the top scoring protein in both searches; however, the score obtained using the raw BSA peak list falls well below the threshold for significance at $p < 0.05$. In the case of the refined peak list, the score for BSA is significantly improved, exceeding the required significance threshold by a considerable margin. In both cases, bovine albumin is clearly distinguished from the closely related sheep albumin. In the case of HAT, shown in Figure 6e,f, the raw peak list yields the correct protein match as the top hit. While the match is well-distinguished from the highest scoring random match, the correct match does not exceed the threshold for statistical significance. However, when the refined peak list was submitted, the score for HAT was drastically improved with concomitant expansion in the difference between the correct and highest incorrect match. Several significant attributes of the PMF outcomes for the raw and refined peak lists are compared in Table 1. Among these attributes are the Mowse score for the correct protein match (S_{Mc}), and the difference ΔS_M in Mowse score between S_{Mc} and

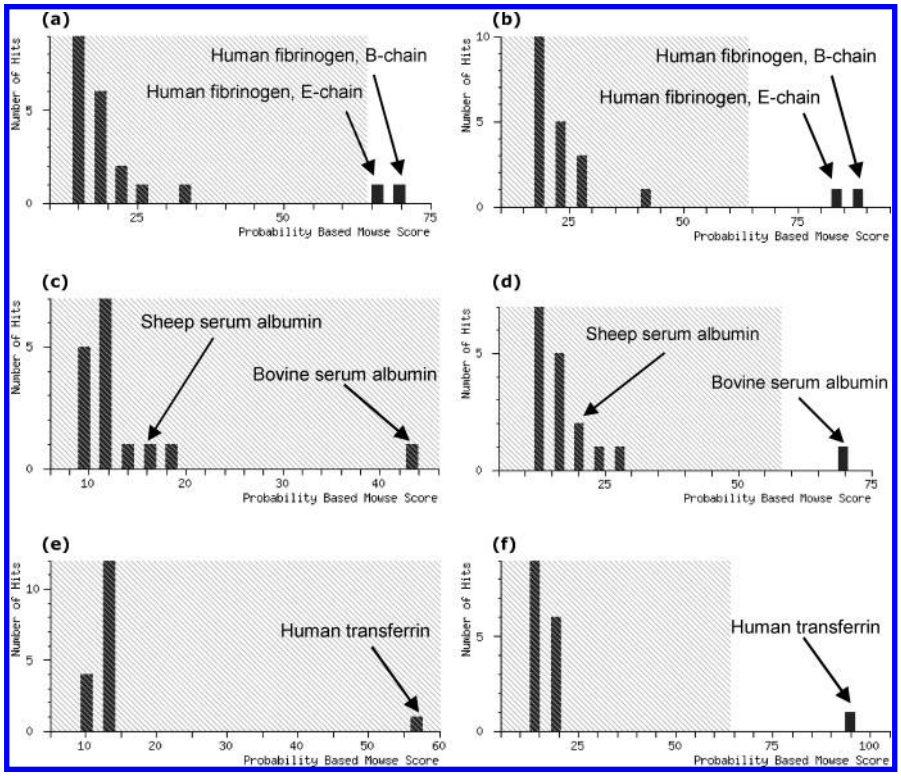

Figure 6. Mowse score distributions reported by Mascot for PMF analysis using raw and refined peak lists. Protein matches falling beyond the shaded region are significant at $p < 0.05$. The distributions are shown for HPF raw (a) and refined (b); BSA raw (c) and refined (d); and HAT raw (e) and refined (f). Note that the *x* axes are scaled differently in each plot.

Table 1. Comparison of Significant Attributes for Mascot PMF Queries Based on Refined and Unrefined Peak Lists

Protein	Attribute	Raw Peak List	Refined Peak List
HPF	S_{Mc}	70	88
	ΔS_M	35	46
	P_c	1.0×10^{-7}	1.6×10^{-9}
	P_c/P_r	3.2×10^{-4}	2.5×10^{-5}
	Matched/total masses	8/126	8/70
BSA	S_{Mc}	43	70
	ΔS_M	25	43
	P_c	5.0×10^{-5}	1.0×10^{-7}
	P_c/P_r	3.2×10^{-3}	5.0×10^{-5}
	Matched/total masses	11/112	11/50
HAT	S_{Mc}	57	95
	ΔS_M	42	75
	P_c	2.0×10^{-6}	3.2×10^{-10}
	P_c/P_r	6.3×10^{-5}	3.2×10^{-8}
	Matched/total masses	9/110	9/37

the Mowse score for the highest ranking random match (S_{Mr}), where

$$\Delta S_M = S_{Mc} - S_{Mr} \tag{9}$$

Notably, S_{Mc} and ΔS_M are markedly improved as the number of unmatched masses decreases. These improvements are especially striking when they are viewed in terms of the corresponding absolute probabilities, that is, the absolute probability of a random match for the correct protein (P_c) and the ratio of the random match probabilities P_c and P_r for the correct and nearest random hits, respectively, where

$$\Delta S_M = S_{Mc} - S_{Mr} = 10 \log\left(\frac{P_r}{P_c}\right) \tag{10}$$

so that

$$\frac{P_c}{P_r} = 10^{-(\Delta S_M/10)} \tag{11}$$

The improvement in P_c ranged from 2 to 4 orders of magnitude, while the improvement in P_c/P_r ranged from 10- to 1000-fold. Examination of the table shows that reducing the number of unmatched masses by some factor x improves P_c and P_c/P_r by a factor on the order of 10^x. Clearly, PMF peak list processing in the manner described here has the ability to resolve ambiguous protein hits into conclusive assignments while offering greater discrimination between significant and random matches. These improvements are accomplished by removing significant numbers of non-peptide masses, which translates to exponential improvements in the statistical significance and discrimination of PMF protein match results.

Mass Accuracy Requirement. There is no clear consensus on what mass accuracy is required in order to make analytical use of peptide residual mass predictions. Gras et al.[13] and Creasey and Cottrell[20] proposed that a mass measurement accuracy of ± 0.5 Da is sufficient to make use of the "maximum likelihood" method of internal calibration correction for MALDI-TOF-MS. Karty et al. suggested that a mass accuracy of about 0.3 Da was sufficient to reliably distinguish peptides from alkali-matrix clusters,[10] while Schmidt et al. implied that a mass error of no more than 0.05 Da was required to confidently eliminate matrix or Coomassie blue-derived masses from peptide masses.[16]

In the application of fractional mass criteria to eliminate non-peptide signals, selectivity in rejection of extraneous

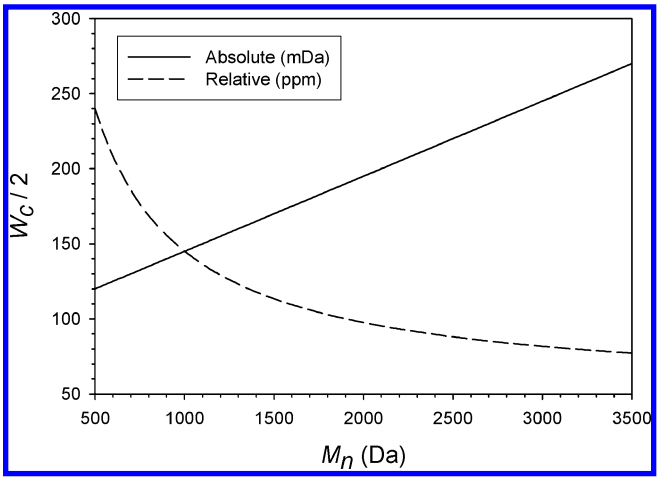

Figure 7. Absolute and relative magnitudes of $W_c/2$ plotted as a function of M_n.

masses is of utmost concern. Importantly, no single mass accuracy requirement can be adopted for the application proposed here due to two crucial factors. First, while the value of W_c increases linearly with increasing M_n, the relative value of W_c with respect to M_n decreases significantly with increasing M_n (Figure 7). Thus, it is important to acknowledge that the permissible range of peptide masses changes significantly across the mass scale, and that the relative mass accuracy requirement is actually greater for larger peptide masses. Second, the minimum error that can result in inappropriate rejection of a true peptide mass depends on the difference between M_m and M_c. Clearly, greater errors are tolerated when M_m is close to M_c, while smaller errors are of increasing concern as the difference between M_c and M_m becomes greater.

For instance, consider a peptide with a nominal mass of 2001 Da. From eqs 3 and 4, the centroid position and width about the centroid are calculated as 2001.960 and 0.390 Da, respectively. Therefore, the range encompassing 95% of peptide masses can be represented as $M_c \pm (W_c/2)$. Since the distribution of peptides about the centroid is essentially Gaussian, the standard deviation s can easily be calculated for the distribution. Recall that in a normally distributed population, approximately 95% of the sampled population occurs within the range $\bar{x} \pm 2s$. Thus, as 95% of peptides are included in the range $M_c \pm (W_c/2)$, an equality relating W_c and s may be derived

$$s \approx \frac{W_c}{4} \tag{12}$$

This provides the means for determining the requirement for mass accuracy to avoid rejecting a peptide that has an actual monoisotopic mass given by, for example, $M_c + 1.7s$ (in this case, equal to 2002.126 Da). Effectively, this will provide an indication of how accurately the mass of a peptide must be determined if the mass falls at the edge of the range encompassing approximately 90% of the population ($\bar{x} \pm 1.7s$). The maximum error not resulting in inappropriate rejection of such a peptide can be easily calculated as the difference between $M_c + 1.7s$ and $M_c + 2s$. This calculation yields a difference of 0.029 Da, or approximately 15 ppm. This implies that a significantly smaller error is required for this application as compared to the acceptable mass errors suggested previously. When similar calculations are used, it can be shown that measurement with an error of 0.05 Da spares only the nearest

83.7% of peptides about M_c. At 2000 Da, a 0.05 Da mass error corresponds to 25 ppm. An error on the order of 0.05 Da is therefore sufficient to narrow the effective permissible peptide range to exclude the outlying 16% of true peptides. This is over 3 times the intended outlier rejection of 5%.

It would appear that the mass accuracy necessary for effective use of this scheme is perhaps greater than previously appreciated, and certainly beyond the capabilities of some instruments used in routine proteomics. This provides at least a partial explanation for the fact that peptide residual mass predictions have not to date been used as the basis of a broadly applied, rigorous filter for non-peptides. Importantly, one should be cautious of accepting any single, generalized figure for maximum allowable mass error appropriate for this data-processing scheme; such a figure would be a significant oversimplification.

Conclusion

This study has demonstrated the usefulness of a PMF peak list processing approach based on the requirement that observed masses must fall within the fractional mass range consistent with peptides. Elimination of masses which cannot possibly be attributed to peptides was demonstrated to significantly improve probability-based Mowse scores for the correct protein, as well as to increase the difference between the correct hit and the nearest random hit by virtue of increasing the ratio of matched to total submitted masses. Furthermore, with the modifications to the original equations as discussed here, no instance of an actual peptide mass being rejected by the filter as compared to the unfiltered peak list was observed. On the basis of calculated masses of modified averagine peptides, it was also demonstrated that the processing of peak lists in this manner does not exclude peptides containing any of 10 natural and artificial modifications often encountered in proteomic analysis. This is noteworthy, as the processing in this manner does not rule out PMF searches involving modified peptides. While the approach described here is a rapid, efficient, and relatively uncomplicated means of removing irrelevant masses from PMF queries, there are two requirements for successful application: sufficient resolution to allow clear selection of the monoisotopic mass, and sufficient mass accuracy to ensure that legitimate peptide masses are not unintentionally excluded from the list. The requirement for mass accuracy to selectively exclude non-peptides based on fractional mass has previously been somewhat underestimated. It appears that this means of data reduction has not previously been applied to PMF because many mass spectrometry platforms routinely used in proteomics are not capable of providing the necessary combination of mass resolution and mass accuracy for the selective exclusion of non-peptide masses solely on the basis of residual mass. Indeed, a similar data-processing step has been incorporated into an approach based on FTICR-MS for shotgun sequencing of peptide mixtures.[35] Unfortunately, the technique has yet to find broader application, most likely for the same reasons cited above. Given the increasing use of high resolution, high accuracy mass spectrometry to proteomics using FTICR-MS[28,36] and the recently introduced Orbitrap MS,[37,38] elimination of non-peptide signals based on accurate mass seems likely to become a valuable and more widely used tool for proteomic data analysis and represents an efficient alternative to more complicated methods of PMF peak list processing.

Acknowledgment. This work was supported by National Institute of Aging Grant AG 24488 (P.J.H.) and by National Institutes of Health General Medicine Grant GM 49077 (C.B.L.).

References

(1) Yates, J. R., III; Speicher, S.; Griffin, P. R.; Hunkapiller, T. Peptide mass maps: a highly informative approach to protein identification. *Anal. Biochem.* **1993**, *214*, 397−408.
(2) Pappin, D. J.; Hojrup, P.; Bleasby, A. J. Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* **1993**, *3*, 327−332.
(3) James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. Protein identification by mass profile fingerprinting. *Biochem. Biophys. Res. Commun.* **1993**, *195*, 58−64.
(4) Henzel, W. J.; Watanabe, C.; Stults, J. T. Protein identification: the origins of peptide mass fingerprinting. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 931−942.
(5) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551−3567.
(6) Olsen, J. V.; Ong, S. E.; Mann, M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell. Proteomics* **2004**, *3*, 608−614.
(7) Hagman, C.; Ramstrom, M.; Jansson, M.; James, P.; Hakansson, P.; Bergquist, J. Reproducibility of tryptic digestion investigated by quantitative Fourier transform ion cyclotron resonance mass spectrometry. *J. Proteome Res.* **2005**, *4*, 394−399.
(8) Stapels, M. D.; Barofsky, D. F. Complementary use of MALDI and ESI for the HPLC-MS/MS analysis of DNA-binding proteins. *Anal. Chem.* **2004**, *76*, 5423−5430.
(9) Nielsen, M. L.; Savitski, M. M.; Kjeldsen, F.; Zubarev, R. A. Physicochemical properties determining the detection probability of tryptic peptides in Fourier transform mass spectrometry. A correlation study. *Anal. Chem.* **2004**, *76*, 5872−5877.
(10) Karty, J. A.; Ireland, M. M. E.; Brun, Y. V.; Reilly, J. P. Artifacts and unassigned masses encountered in peptide mass mapping. *J. Chromatogr., B* **2002**, *782*, 363−383.
(11) Lambert, J. P.; Ethier, M.; Smith, J. C.; Figeys, D. Proteomics: from gel based to gel free. *Anal. Chem.* **2005**, *77*, 3771−3787.
(12) Berndt, P.; Hobohm, U.; Langen, H. Reliable automatic protein identification from matrix-assisted laser desorption/ionization mass spectrometric peptide fingerprints. *Electrophoresis* **1999**, *20*, 3521−3526.
(13) Gras, R.; Muller, M.; Gasteiger, E.; Gay, S.; Binz, P. A.; Bienvenut, W.; Hoogland, C.; Sanchez, J. C.; Bairoch, A.; Hochstrasser, D. F.; Appel, R. D. Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. *Electrophoresis* **1999**, *20*, 3535−3550.
(14) Samuelsson, J.; Dalevi, D.; Levander, F.; Rognvaldsson, T. Modular, scriptable and automated analysis tools for high-throughput peptide mass fingerprinting. *Bioinformatics* **2004**, *20*, 3628−3635.
(15) Breen, E. J.; Hopwood, F. G.; Williams, K. L.; Wilkins, M. R. Automatic Poisson peak harvesting for high throughput protein identification. *Electrophoresis* **2000**, *21*, 2243−2251.
(16) Schmidt, F.; Schmid, M.; Jungblut, P. R.; Mattow, J.; Facius, A.; Pleissner, K. P. Iterative data analysis is the key for exhaustive analysis of peptide mass fingerprints from proteins separated by two-dimensional electrophoresis. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 943−956.
(17) Rognvaldsson, T.; Hakkinen, J.; Lindberg, C.; Marko-Varga, G.; Potthast, F.; Samuelsson, J. Improving automatic peptide mass fingerprint protein identification by combining many peak sets. *J. Chromatogr., B* **2004**, *807*, 209−215.
(18) Mann, M. In Useful Tables of Possible and Probable Peptide Masses. Presented at the 43rd Annual Conference on Mass Spectrometry and Allied Topics, Atlanta, GA, May 21−26, 1995; American Society for Mass Spectrometry, 1995.
(19) Zubarev, R. A.; Hakansson, P.; Sundqvist, B. Accuracy requirements for peptide characterization by monoisotopic molecular mass measurements. *Anal. Chem.* **1996**, *68*, 4060−4063.
(20) Creasy, D. M.; Cottrell, J. S. Error tolerant searching of uninterpreted tandem mass spectrometry data. *Proteomics* **2002**, *2*, 1426−1434.
(21) Andreev, V. P.; Rejtar, T.; Chen, H. S.; Moskovets, E. V.; Ivanov, A. R.; Karger, B. L. A universal denoising and peak picking

algorithm for LC-MS based on matched filtration in the chromatographic time domain. *Anal. Chem.* **2003**, *75*, 6314−6326.
(22) Zhang, X.; Hines, W.; Adamec, J.; Asara, J. M.; Naylor, S.; Regnier, F. E. An automated method for the analysis of stable isotope labeling data in proteomics. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 1181−1191.
(23) Matthiesen, R.; Bauw, G.; Welinder, K. G. Use of performic acid oxidation to expand the mass distribution of tryptic peptides. *Anal. Chem.* **2004**, *76*, 6848−6852.
(24) Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S. Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrom. Rev.* **1998**, *17*, 1−35.
(25) Witt, M.; Fuchser, J.; Baykut, G. Fourier transform ion cyclotron resonance mass spectrometry with NanoLC/microelectrospray ionization and matrix-assisted laser desorption/ionization: analytical performance in peptide mass fingerprint analysis. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 553−561.
(26) Horn, D. M.; Peters, E. C.; Klock, H.; Meyers, A.; Brock, A. Improved protein identification using automated high mass measurement accuracy MALDI FT-ICR MS peptide mass fingerprinting. *Int. J. Mass Spectrom.* **2004**, *238*, 189−196.
(27) He, F.; Emmett, M. R.; Hakansson, K.; Hendrickson, C. L.; Marshall, A. G. Theoretical and experimental prospects for protein identification based solely on accurate mass measurement. *J. Proteome Res.* **2004**, *3*, 61−67.
(28) Bogdanov, B.; Smith, R. D. Proteomics by FTICR mass spectrometry: top down and bottom up. *Mass Spectrom. Rev.* **2005**, *24*, 168−200.
(29) Paizs, B.; Suhai, S. Fragmentation pathways of protonated peptides. *Mass Spectrom. Rev.* **2005**, *24*, 508−548.
(30) Mize, T. H.; Amster, I. J. Broad-band ion accumulation with an internal source MALDI-FTICR-MS. *Anal. Chem.* **2000**, *72*, 5886−5891.
(31) O'Connor, P. B.; Costello, C. E. Internal calibration on adjacent samples (InCAS) with Fourier transform mass spectrometry. *Anal. Chem.* **2000**, *72*, 5881−5885.
(32) Zhang, L. K.; Rempel, D.; Pramanik, B. N.; Gross, M. L. Accurate mass measurements by Fourier transform mass spectrometry. *Mass Spectrom. Rev.* **2005**, *24*, 286−309.
(33) Gay, S.; Binz, P. A.; Hochstrasser, D. F.; Appel, R. D. Modeling peptide mass fingerprinting data using the atomic composition of peptides. *Electrophoresis* **1999**, *20*, 3527−3534.
(34) Senko, M. W.; Beu, S. C.; Mclafferty, F. W. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 229−233.
(35) Nielsen, M. L.; Savitski, M. M.; Zubarev, R. A. Improving protein identification using complementary fragmentation techniques in Fourier transform mass spectrometry. *Mol. Cell. Proteomics* **2005**, *4*, 835−845.
(36) Bergquist, J.; Palmblad, M.; Wetterhall, M.; Hakansson, P.; Markides, K. E. Peptide mapping of proteins in human body fluids using electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Mass Spectrom. Rev.* **2002**, *21*, 2−15.
(37) Olsen, J. V.; de Godoy, L. M.; Li, G.; Macek, B.; Mortensen, P.; Pesch, R.; Makarov, A.; Lange, O.; Horning, S.; Mann, M. Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics* **2005**, *4*, 2010−2021.
(38) Yates, J. R.; Cociorva, D.; Liao, L.; Zabrouskov, V. Performance of a linear ion trap-Orbitrap hybrid for peptide analysis. *Anal. Chem.* **2006**, *78*, 493−500.

PR050486O