Available online at www.sciencedirect.com
ScienceDirect
Analytical Biochemistry 372 (2008) 156–166
ANALYTICAL BIOCHEMISTRY
www.elsevier.com/locate/yabio

Systematic characterization of high mass accuracy influence on false discovery and probability scoring in peptide mass fingerprinting

Eric D. Dodds [a], Brian H. Clowers [a], Paul J. Hagerman [b], Carlito B. Lebrilla [a,b,*]

[a] Department of Chemistry, University of California, Davis, Davis, CA 95616, USA
[b] School of Medicine, Department of Biochemistry and Molecular Medicine, University of California, Davis, Davis, CA 95616, USA

Received 25 June 2007
Available online 11 October 2007

Abstract

Whereas the bearing of mass measurement error on protein identification is sometimes underestimated, uncertainty in observed peptide masses unavoidably translates to ambiguity in subsequent protein identifications. Although ongoing instrumental advances continue to make high accuracy mass spectrometry (MS) increasingly accessible, many proteomics experiments are still conducted with rather large mass error tolerances. In addition, the ranking schemes of most protein identification algorithms do not include a meaningful incorporation of mass measurement error. This article provides a critical evaluation of mass error tolerance as it pertains to false positive peptide and protein associations resulting from peptide mass fingerprint (PMF) database searching. High accuracy, high resolution PMFs of several model proteins were obtained using matrix-assisted laser desorption/ionization Fourier transform ion cyclotron resonance mass spectrometry (MALDI–FTICR–MS). Varying levels of mass accuracy were simulated by systematically modulating the mass error tolerance of the PMF query and monitoring the effect on figures of merit indicating the PMF quality. Importantly, the benefits of decreased mass error tolerance are not manifest in Mowse scores when operating at tolerances in the low parts-per-million range but become apparent with the consideration of additional metrics that are often overlooked. Furthermore, the outcomes of these experiments support the concept that false discovery is closely tied to mass measurement error in PMF analysis. Clear establishment of this relation demonstrates the need for mass error-aware protein identification routines and argues for a more prominent contribution of high accuracy mass measurement to proteomic science.

Keywords: Peptide mass fingerprinting; False positive; Mass measurement accuracy; Matrix-assisted laser desorption/ionization; Fourier transform ion cyclotron resonance mass spectrometry

Over the past decade, peptide mass fingerprinting (PMF)[1] has matured into a rapid and sensitive technique for protein identification and has become an indispensable tool for proteome research [1–5]. Despite significant methodological advances in mass spectrometry (MS)-based proteomic science, PMF has remained a method of first choice for identification of proteins separated by gel electrophoresis. As with any protein identification technology, PMF is subject to the possibility of false positive protein associations. It is generally acknowledged that the risk of false positives is greater with PMF than with other methods that are based on sequence information obtained by tandem mass spectrometry (MS/MS). As aptly put by Perkins et al., "The bane of peptide mass fingerprint searching has always been false positives" [6]. Nevertheless, PMF still

* Corresponding author. Fax: +530 754 5609.
E-mail address: cblebrilla@ucdavis.edu (C.B. Lebrilla).

[1] *Abbreviations used:* PMF, peptide mass fingerprinting; MS, mass spectrometry; MS/MS, tandem mass spectrometry; CEA, chicken egg albumin; GO, glucose oxidase; BSA, bovine serum albumin; HHb, human hemoglobin; HPF, human plasma fibrinogen; HAT, human apo-transferrin; ACN, acetonitrile; TFA, trifluoroacetic acid; DHB, 2,5-dihydroxybenzoic acid; MALDI, matrix-assisted laser desorption/ionization; FTICR, Fourier transform ion cyclotron resonance; InCAS, internal calibration on adjacent samples; ACTH, adrenocorticotropic hormone; BI, bovine insulin; LC, liquid chromatography; RMS, root mean square.

0003-2697/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ab.2007.10.009

occupies an important role in proteome research as a fast, reliable, and inexpensive method for identifying proteins following some degree of fractionation.

The issue of false positive protein identification has recently gained due recognition as a key consideration in interpretation of proteomic results. The development of methods for characterizing and minimizing false discovery rates represents a vigorous area of research and development in all aspects of proteomics, both analytical and informatic [7–16]. One significant contributor to false proteomic discovery is error in mass measurement. As is true of any analytical process, uncertainty in a measurement translates to uncertainty in any results derived from that measurement. When using PMF for protein identification, inaccuracies in measurement of peptide mass/charge ratios (m/z) have the unavoidable consequence of leading to some errors in peptide and protein associations. One simple and effective route to minimizing the possibility of false positive identifications is the use of accurate mass measurement (i.e., mass measurement with <10 ppm error) and the imposition of commensurately restrictive mass tolerances in the database query.

Despite the potential of accurate mass measurement for mitigation of false protein discovery, the significance of high mass accuracy to the field of proteomics has only recently begun to garner broad acknowledgment. Therefore, the application of high accuracy mass measurement to proteomic endeavors stands at the leading edge of modern proteomic science [17–20]. This is true not only of protein identification but also of techniques for extracting further details from proteomic datasets [21–23]. In light of these developments, and with the increasing accessibility of high accuracy mass spectrometers, accurate mass measurement is poised to become the standard for routine proteomic determinations.

Interestingly, the application of high mass accuracy in proteomic approaches involving peptide sequence information has been the object of significantly more study than in proteomic approaches involving PMF. This is an ironic reversal given that approaches yielding peptide sequence information are somewhat less sensitive to mass error than is PMF. For example, the use of a ±3-Da precursor ion tolerance in MS/MS-based proteomics is common even when an instrument capable of high mass accuracy is used [24–26]. Despite this general practice, there are clear and well-documented advantages of high mass accuracy measurement in proteomic approaches based on MS/MS sequencing of peptides [27–37]. Although determination of accurate peptide masses should in principle deliver an even greater margin of improvement in PMF, published developments involving high mass accuracy with regard to PMF have been surprisingly sparse, particularly in comparison with the tremendous volume of proteomic literature [38–42]. Some reports have pointed out that performing PMF database searches with high mass error tolerance returns large numbers of matching proteins and that the likelihood of associating experimental data with a database entry at random decreases with reduced error tolerance [39–42]. These reports notwithstanding, there has been very little in-depth examination of how mass measurement error is linked to false positive peptide and protein associations in PMF, nor has there been a detailed critique of how these issues affect database searching and protein match scoring. Zubarev and Mann recently noted that the benefits afforded by accurate mass measurement remain generally unrealized in the field of proteomics and that most protein identification algorithms and scoring schemes do not take mass measurement error into account when assigning ranks to putative matches [43].

The current article presents a systematic appraisal of the advantages and limitations of high mass accuracy in PMF, with particular emphasis on the issues of false peptide and protein discovery and on the details of database search reporting. The outcomes of this work provide further evidence that mass error and false discovery are closely related. The current findings also present a thorough characterization of this relation at both the peptide and protein levels and, thus, contribute important steps toward establishing a more pivotal role for high accuracy mass measurement in proteome research. Considering the increasing availability of high accuracy MS instrumentation, these findings are timely and assume considerable significance as accurate mass measurement becomes the benchmark for state-of-the-art proteomic determinations.

Materials and methods

Sample preparation

Chicken egg albumin (CEA), *Aspergillus niger* glucose oxidase (GO), bovine serum albumin (BSA), human hemoglobin (HHb), human plasma fibrinogen (HPF), and human apo-transferrin (HAT) served as model proteins (Sigma, St. Louis, MO, USA). Each protein was dissolved in 8 M urea/200 mM total Tris (pH 7.8) at a concentration of 1 μg/μl. Tryptic peptide stocks were then prepared from each protein. A 1-μl aliquot of each 1-μg/μl protein solution was further diluted in 40 μl of 8 M urea/200 mM Tris buffer (pH 7.8). Prior to digestion, proteins were reduced (by the addition of 10 μl of 450 mM dithiothreitol in 50 mM NH_4HCO_3 with incubation at 55 °C for 1 h) and alkylated (by the addition of 10 μl of 500 mM iodoacetamide in 50 mM NH_4HCO_3 with incubation in the dark at ambient temperature for 30 min). Each preparation was then diluted to less than 2 M in urea by the addition of 150 μl of deionized water and treated with 1 μl of a 0.05-μg/μl solution of sequencing-grade modified trypsin (Promega, Madison, WI, USA). Digestion was allowed to proceed for approximately 8 to 10 h with incubation at 37 °C. The reactions were terminated by storing the samples at −20 °C. Aliquots of each tryptic digest (10 μl) were purified by solid phase extraction with C18 ZipTips (Millipore, Billerica, MA, USA). Desalted tryptic peptides were eluted

in 10 μl of 50% acetonitrile (ACN) with 0.1% trifluoroacetic acid (TFA).

Mass spectrometry

A matrix solution of 50 μg/μl of 2,5-dihydroxybenzoic acid (DHB) was prepared in 50% ACN. Samples were prepared for matrix-assisted laser desorption/ionization (MALDI) by combining 1 μl of the purified tryptic digest and 1 μl of DHB on a stainless-steel target and allowing the mixtures to air dry. Each spot contained a quantity of digest corresponding to approximately 100 fmol of protein to approximate a realistic quantity of protein digest. A HiResMALDI Fourier transform ion cyclotron resonance (FTICR) mass spectrometer (IonSpec, Lake Forest, CA, USA) was the platform for all PMF analyses. This instrument featured an actively shielded 7.0 T superconducting magnet and an external MALDI source based on a third harmonic Nd:YAG laser (5 ns pulse width at 355 nm).

All spectra used for PMF were internally calibrated by gas phase combination of analyte and standard ions produced in separate MALDI events [44,45]. This mass calibration technique, known as internal calibration on adjacent samples (InCAS) [45], takes advantage of the pulsed nature of MALDI and the ion trapping capabilities of FTICR–MS. Multiple MALDI pulses (optimized for each individual sample spot) were used to produce analyte ions from the sample spot, and these ions were trapped and stored in the ICR cell. A calibration spot was next irradiated for MALDI, and the standard ions were combined with the analyte ions in the ICR cell. The combined population of trapped analyte and calibrant ions was then mass analyzed. The calibrant spots were prepared by spotting 1 μl of calibrant solution and 1 μl of DHB matrix solution. Two calibrant mixtures were deposited separately on the MALDI target: 1 μM $P_{14}R$ (a labile synthetic peptide yielding *y*-series fragments through metastable decay associated with the "proline effect" [46]) and a mixture of $P_{14}R$, human adrenocorticotropic hormone (ACTH) fragment peptide 18–39, and bovine insulin (BI) oxidized B chain (each at a concentration of 1 μM). Both calibrant solutions were prepared in 50% ACN/0.1% TFA. All standard peptides were obtained from Sigma. A preliminary screening over the m/z range of 500 to 3500 was done to determine which calibration spot was appropriate for the mass range of peptides observed in each digest. $P_{14}R$ produced calibrant ions spanning approximately m/z 750 to 1530, whereas the $P_{14}R$, ACTH, and BI mixture produced calibrant ions spanning approximately m/z 750 to 3500.

An RF-only quadrupole served as a broadband ion guide for injecting externally produced ions into the ICR cell. Ions were vibrationally cooled by a pulse of argon gas into the ion guide and ICR regions of the vacuum chamber. Positively charged ions were trapped in the cylindrical ICR cell by 20 V trapping plate potentials. Ions from multiple MALDI events were accumulated by gating the source side trapping potential to 4 V for a duration dependent on the required m/z range. Immediately preceding ion excitation, the front and rear trapping plate potentials were linearly ramped to zero over a 1-s duration, whereas a potential of 0.5 V was maintained on the inner trapping rings for the entire experiment. An arbitrary waveform pulse (32 k waveform points applied at a DAC rate of 2 MHz, 150 V base-to-peak amplitude) was used to accelerate ions in the m/z range of 500 to 3500. Spectra for PMF analysis were acquired in the mass range of m/z 500 to 2500 or m/z 500 to 3500, depending on the mass range of peptides present in the sample. Each spectrum was acquired by sampling 1024 k time domain data points at an ADC rate of 500 kHz, yielding a 2.097-s transient observation time. Prior to fast Fourier transformation, a Blackman window was applied for apodization and a one order zero fill was appended.

Data processing

Mass calibration was performed based on InCAS masses using the IonSpec Omega software according to standard FTICR–MS calibration relationships [47,48]. Specifically, observed cyclotron frequencies (f_c with units of Hz) were calibrated to m/z according to

$$\frac{m}{z} = \frac{A}{f_c - B}, \tag{1}$$

where A and B are empirically determined calibration constants. The calibrated spectra were further processed using the IonSpec PeakHunter software, and the resulting thresholded, monoisotopic $[M + H]^+$ peak lists were exported as text format files. These peak lists were next relieved of masses that did not have fractional components consistent with peptide elemental composition. This was done based on the previous observation that the permissible residual mass range for all peptide elemental compositions can easily be calculated for a given nominal mass [49,50]. This step is advantageous because interfering signals are common in PMF [51]. The relation of peptide elemental composition to fractional mass has recently been used as a means of ensuring that only ions consistent with peptide mass were interrogated in online liquid chromatography (LC)–MS/MS experiments and as a quality control metric for large proteomic datasets (the so-called "mass deviance" parameter) [29,52]. However, in this context, peptide residual mass predictions were used to allow peptide-inconsistent masses to be excluded from the PMF peak lists. The internal calibrant masses were also removed from the peak lists at this stage. These operations were carried out using the previously described Mass Sieve algorithm [53,54]. Although the original program was written in Visual Basic for implementation in Excel, a new version of Mass Sieve has been developed based on the IGOR Pro 5 software package (Wave Metrics, Lake Oswego, OR, USA). This updated version of Mass Sieve remained essentially equivalent to the original program; however, Mass Sieve in IGOR Pro was improved by the addition of a more convenient graph-

ical user interface, a user-defined low mass cutoff for the refined peak list, and batch processing capabilities. The Mass Sieve processing of all peak lists was performed in both the Visual Basic and IGOR versions, producing identical refined peak lists. Standard masses introduced by InCAS were screened from the peak lists with a 2-ppm error tolerance, and all signals of m/z less than 700 were removed from the peak lists. The final processed peak lists were exported by Mass Sieve in text file format for PMF query submission. It should be noted that the software referred to in the current study was developed in this laboratory and should not be mistaken for an unrelated proteomic program that has more recently assumed the name "MassSieve" [55].

Peptide mass fingerprinting analysis

Processed peak lists were submitted for PMF using the Mascot search engine (www.matrixscience.com) [6]. Although some details of the Mascot search engine and Mowse scoring algorithm are not published, this tool is among the most widely used protein identification packages. Thus, the current work was focused on Mowse scoring with Mascot, allowing the forthcoming results to be of immediate and direct use to the large community of proteomic practitioners using this platform. All queries were submitted via the Mascot Wizard utility (freeware available for download at www.matrixscience.com/wizard.html). The Mass Spectrometry Protein Sequence Database (MSDB) was searched for tryptic peptides with fixed carbamidomethylation of cysteine residues. No variable modifications were considered. The allowed number of missed tryptic cleavages (either zero or one) was set on a case-by-case basis. Taxonomy was set as specifically as possible for each protein of interest. For each PMF, the mass error tolerance of the PMF query was systematically adjusted to simulate varying levels of mass accuracy. All other parameters were held constant unless noted specifically. Although the modulation of mass error tolerance was of interest in this work, in general mass error tolerance should be set based on an empirical assessment of the mass accuracy for a given instrument. This setting should also be made such that the majority of true positive matches are captured. More detailed recommendations on setting mass error tolerance were recently provided by Zubarev and Mann [43].

Results and discussion

Three representative MALDI–FTICR–MS PMFs are shown in Fig. 1. These spectra were chosen such that the quality of realistic samples was approximated. To illustrate the practical accuracy of the mass measurements, the mass errors for all matched tryptic peptides are plotted in Fig. 2. The root mean square (RMS) error of the 50 measurements

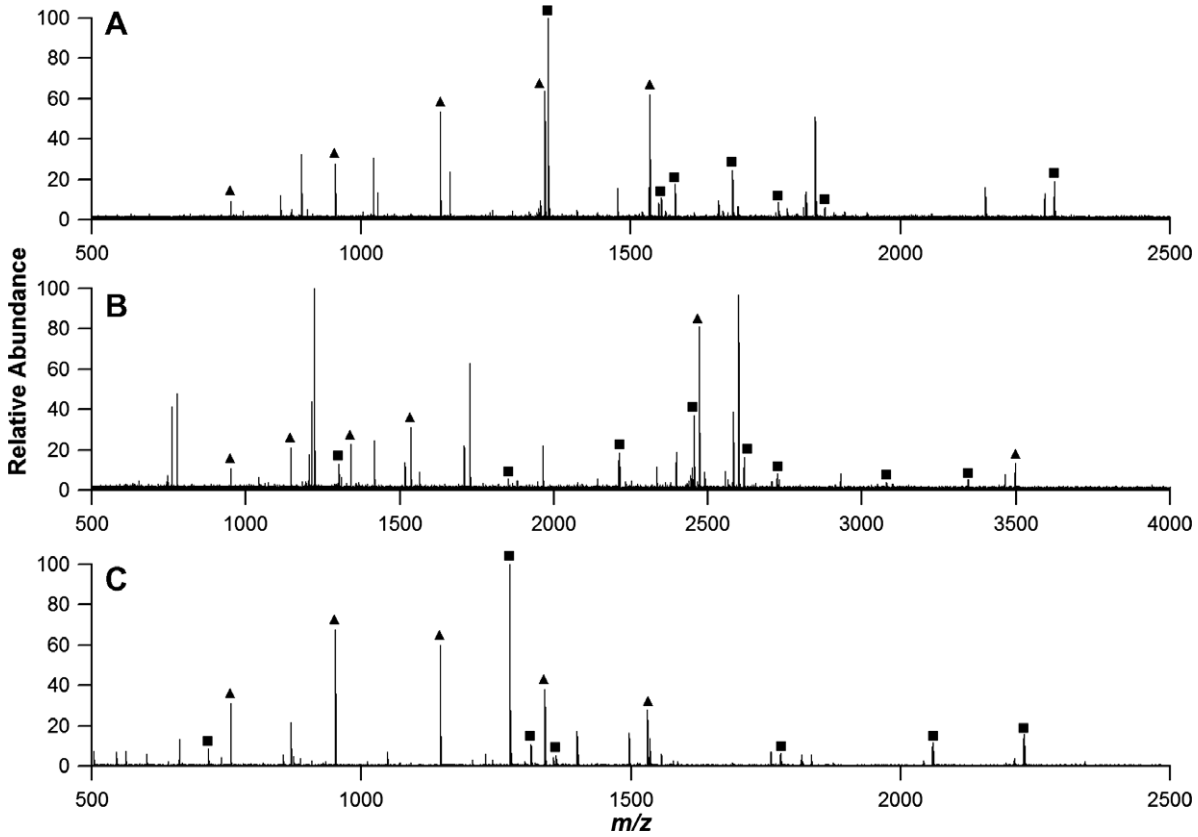

Fig. 1. MALDI–FTICR mass spectra for PMF analysis of CEA (A), GO (B), and HHb (C). Internal calibrant masses introduced by InCAS are labeled with triangles, whereas masses matched to the correct protein to within 10 ppm error are labeled with squares.

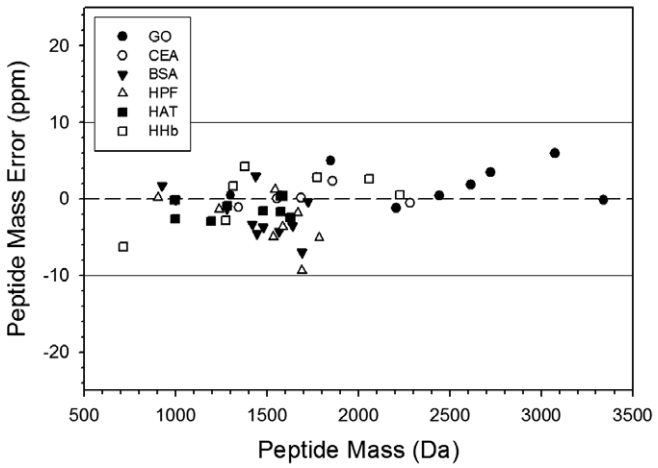

Fig. 2. Mass errors for 50 tryptic peptides correctly matched to their respective target proteins. The RMS mass error was 3 ppm.

was 3 ppm, with no mass errors exceeding 10 ppm. Based on these matching peptides, the sequence coverage for each protein ranged from 24 to 58%. Although these spectra and the associated mass accuracies rendered these PMFs of good quality, close examination of the mass error tolerance effect on PMF results revealed noteworthy trends.

When larger mass errors were tolerated during PMF database searching, the likelihood of false positive peptides inappropriately contributing the Mowse score becomes significantly elevated. This situation is compounded by the fact that, for a given nominal mass, the exact masses of all possible peptides are closely distributed [49,50]. For example, essentially all possible peptide compositions with a nominal mass of 1000 Da occur within a 0.3-Da window. Thus, applying a mass tolerance of as little as ±0.15 Da can effectively reduce the search to unit accuracy (i.e., no more selective than nominal mass).

An excellent example of this situation was observed in the PMF analysis of CEA (Fig. 3). At a mass error tolerance of 500 ppm, the protein was correctly identified with a statistically significant Mowse score of 84 (Fig. 3A); the same PMF submitted with a mass error tolerance of 3 ppm also resulted in a statistically significant and correct protein identification with a Mowse score of 88 (Fig. 3B). These Mowse scores, S_M, are related to the absolute probability, P, that the protein match is a random event according to

$$S_M = -10\log(P). \tag{2}$$

A protein match is generally considered significant when the value of the Mowse score corresponds to a random match probability that is expected to occur with a frequency of less than 5%. Based on these figures alone, the benefits of low mass error tolerance in the analysis of this PMF were not readily apparent considering the close similarity of the Mowse scores obtained at vastly different mass error tolerances. However, a survey of the mass errors of each peptide assignment yielded an important observation. Clearly, in the case of 500 ppm mass error tolerance, the Mowse score was being inflated by four false positive peptides matched with errors as high as 440 ppm (Fig. 3C). These four false positive peptides, identified as such based on their outlying mass errors among the other matched peptides, contributed to the inflation of the Mowse score for CEA at the 500-ppm level of mass error tolerance. These false positive peptides were eliminated at a tolerance of 3 ppm; however, even with a reduced number of peptides matched, the Mowse score was increased and the RMS mass error for the correctly matching peptides was 1.4 ppm (Fig. 3D).

In addition to eliminating false positive peptides, the discrimination of PMF searching is improved greatly at low mass error tolerance. Conducting the PMF query at lower mass error tolerance had the added, and perhaps more noteworthy, benefit of expanding the difference in Mowse scores (ΔS_M) between the score for the highest ranking protein match (S_{M1}) and the score for the second highest ranking match (S_{M2}):

$$\Delta S_M = S_{M1} - S_{M2}. \tag{3}$$

The value of ΔS_M was increased from 32 in the case of 500 ppm tolerance to 62 at 3 ppm tolerance. This improvement in discrimination became even more striking when viewed in terms of the corresponding random match probabilities. Because ΔS_M is defined by

$$\Delta S_M = S_{M1} - S_{M2} = 10\log\left(\frac{P_2}{P_1}\right), \tag{4}$$

the ratio of the corresponding probabilities P_1 and P_2 are given by

$$\frac{P_1}{P_2} = 10^{-(\Delta S_M/10)}. \tag{5}$$

Thus, the ratio of the probabilities was improved by three orders of magnitude solely as a result of more stringent mass error tolerance. This illustrates the elimination of false positive peptide matches to incorrect proteins, much as the false positive peptides matching the correct protein were eliminated.

Based on the CEA example, the values of S_M and ΔS_M appeared to have very different dependencies on mass error tolerance. This trend also held true in many additional PMF experiments (Fig. 4). For BSA, CEA, GO, HAT, and HPF, the value of S_M exhibited remarkably little dependence on mass error tolerances at 100 ppm or less (Fig. 4A). This observation initially might seem counterintuitive, and on the surface it appears to disagree with other published observations that S_M increases with decreasing mass error tolerance [6]. However, those observations were based on varying the mass error tolerance from 10,000 ppm to a minimum of 200 ppm, whereas the current findings are focused on the mass error tolerance range of 500 ppm down to 10 ppm or less. Thus, when mass error tolerances are set in the percentage range, increases in S_M are correlated to decreased mass error tolerance. In comparison,

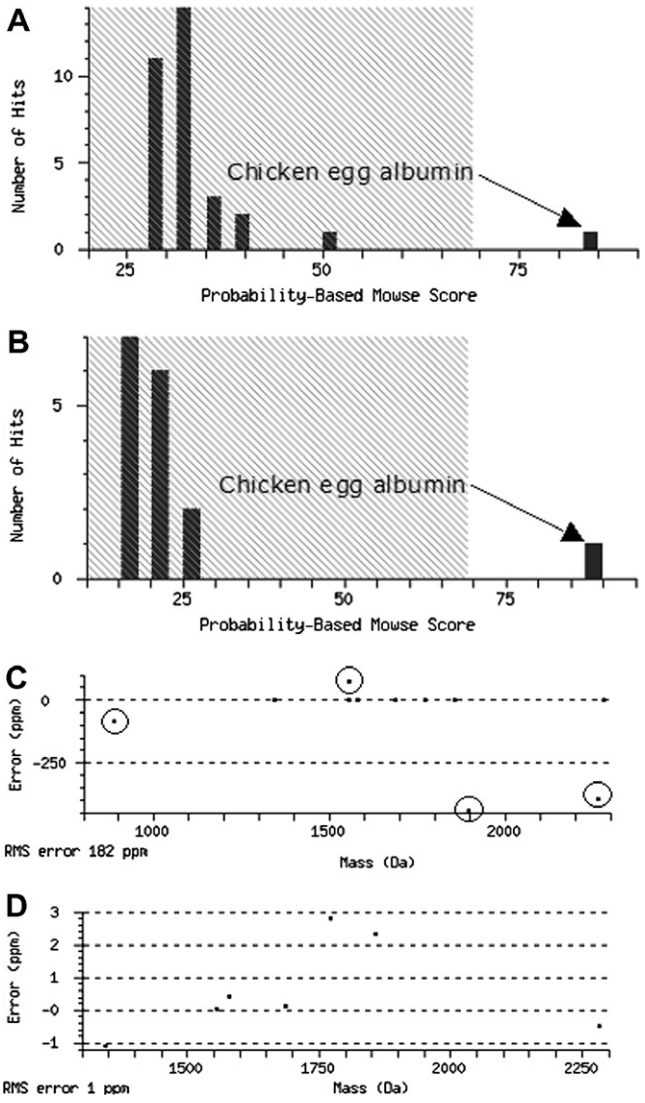

Fig. 3. Mowse score distributions and mass error distributions reported by Mascot for PMF analysis of CEA. PMF queries were conducted with a mass error tolerance of 500 ppm in panels A and C and with a mass error tolerance of 3 ppm in panels B and D. False positive peptides are indicated by circles in panel C. Mowse scores in panels A and B falling beyond the shaded region are significant at $P < 0.05$. Note that the x axes are scaled differently in panels A and B and that the y axes are scaled differently in panels C and D.

the values of S_M are shown here to have little dependence on mass error tolerance in the parts-per-million regime. This apparently is a consequence of the nonuniform distribution of peptide masses that, in the Mowse scoring model, prevents the population of random matches from scaling proportionally to the magnitude of the mass error tolerance [6]. In sharp contrast to the behavior of S_M, the magnitude of ΔS_M showed a strong inverse correlation with mass error tolerance (Fig. 4B). Interestingly, the values of S_M for CEA in Fig. 4A might suggest that the match at 100 ppm tolerance is just as certain as the result at 3 ppm mass error tolerance; however, as shown in Fig. 4B, the benefits of low mass error tolerance not realized in the Mowse statistic are reflected in the value of ΔS_M, a figure of merit seldom reported with PMF results. These trends also hold true for the other proteins considered. It should be noted that these results were derived from searching the most taxonomically specific databases available for each protein of interest. In some cases, querying all taxonomies might be necessary. For those searches involving much larger databases, low mass error tolerance becomes even more valuable in the prevention of false positives.

A closer examination of Fig. 4B revealed a few cases in which the values of S_M and ΔS_M decreased unexpectedly with decreasing mass error tolerance. For example, the Mowse score for BSA dropped significantly at tolerances of less than 100 ppm. A similar situation occurred for HPF, with a mass error tolerance of less than 25 ppm; in this case, the values of both S_M and ΔS_M were depreciated. Although initially it was unclear as to why these metrics were being adversely affected by reduced mass error tolerances, an explanation was found on examination of the mass error distributions of matched peptides. As illustrated in Fig. 5, the higher mass error tolerances allowed false positive peptides to be matched to the protein in question, inappropriately inflating the Mowse scores for the correct matches.

Although the inclusion of false positive peptides in assignment of Mowse scores is troublesome, the assignment of a significant score to an incorrect protein match is of even greater concern. Such an example is depicted in Fig. 6. The PMF analysis of GO at 500 ppm mass error tolerance produced no significant protein matches. When the error tolerance was reduced to 100 ppm, a significant but false hit was obtained for a hypothetical protein from *Yarrowia lipolytica*. At this mass tolerance, a hit for GO was also reported but was scored below the threshold of statistical significance ($P < 0.05$). Although the PMF result at 500 ppm mass error tolerance was inconclusive, the outcome at 100 ppm tolerance exemplifies an even more consequential outcome of relaxed mass error tolerance, resulting in an incorrect protein association with a score exceeding the acceptance threshold. Reducing the mass error tolerance to 25 ppm established GO as the highest ranking and only statistically significant hit, and additional reduction of the mass error tolerance to 6 ppm further improved the S_M and ΔS_M metrics for GO.

Two characteristics of the false positive hit cast significant doubt on this putative protein identification and could have served to eliminate the false positive hit even in the absence of the correct protein identity. First, as seen in Fig. 6B, the highest ranking hit at a mass error tolerance of 100 ppm fell beyond the significance threshold, with $S_M = 68$. However, this hit had a rather low value of $\Delta S_M = 8$ given that the second highest ranking protein (the correct hit for GO) was scored at $S_M = 60$. This small magnitude of ΔS_M is sufficient to warrant a more critical view of the PMF result. Second, the mass error distribution associated with each protein identification provided further reason for skepticism regarding the top hit. As shown in

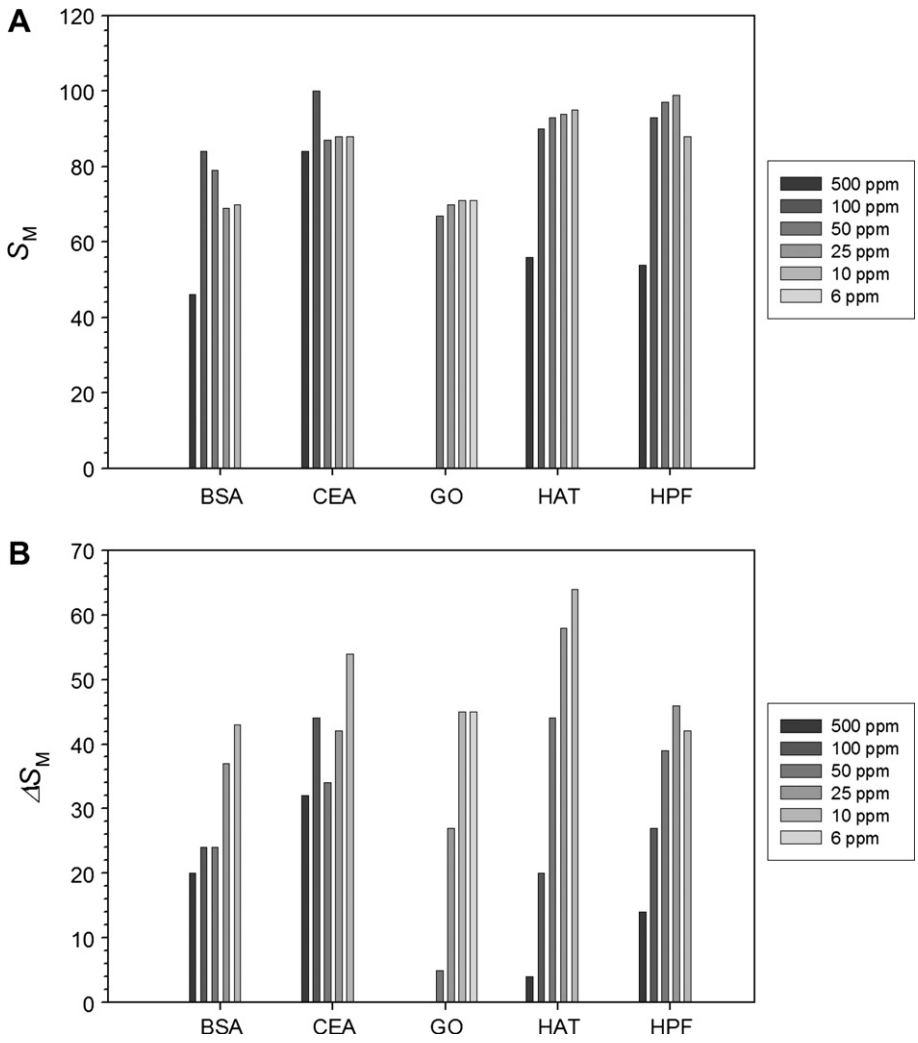

Fig. 4. Effect of mass error tolerance on the Mowse scores (S_M) for correct protein matches (A) and on the differences in Mowse scores (ΔS_M) between the correct matches and nearest random matches (B).

Fig. 7, the RMS peptide mass error for the *Y. lipolytica* hypothetical protein was 44 ppm, with individual peptide mass error approaching the 100-ppm tolerance limits. In the case of 6 ppm tolerance (which established the correct protein identity), the RMS peptide mass error was 3 ppm, a figure much more consistent with expectations for internally calibrated FTICR–MS data. Without the ability to set this stringent mass tolerance, the protein could not have been identified correctly based on the observed peptide signals. Of more critical note is that this PMF would have resulted in a statistically significant false positive at 100 ppm tolerance.

Although mass error tolerance is a critical parameter in PMF, the importance of other search parameters should not be trivialized. This is true even of PMFs measured with high mass accuracy. For example, as shown in Fig. 8, the PMF of HHb did not result in a statistically significant hit at 500 ppm tolerance. Narrowing the mass error tolerance to 100 ppm produced a statistically significant, but not very distinct, match to HHb ($S_M = 69$, $\Delta S_M = 12$). The situation was further improved by reducing the mass error tolerance to 10 ppm ($S_M = 74$, $\Delta S_M = 12$). However, the distinction of the top hit from the nearest random hit (a mutant, recombinant human hemoglobin, designed for use in blood substitutes) remained rather marginal. In all cases, the PMF searches were conducted with a tolerance of one missed tryptic cleavage. Although some allowance for incomplete tryptic digestion is generally applied to proteomic database searches, permitting partially tryptic peptides increases the effective size of the database being queried and can severely diminish specificity. The influence of incomplete tryptic cleavage on PMF searches was recently highlighted by Siepen and coworkers [56]. These authors described a database "masking" approach that removes peptides with likely missed cleavage sites from consideration, resulting in improved databases search results. When searching an unmasked database, similar benefits can be approximated by restricting the search to ideal tryptic peptides (i.e., peptides containing no missed tryptic cleavage sites). Maintaining the mass error tolerance of 10 ppm

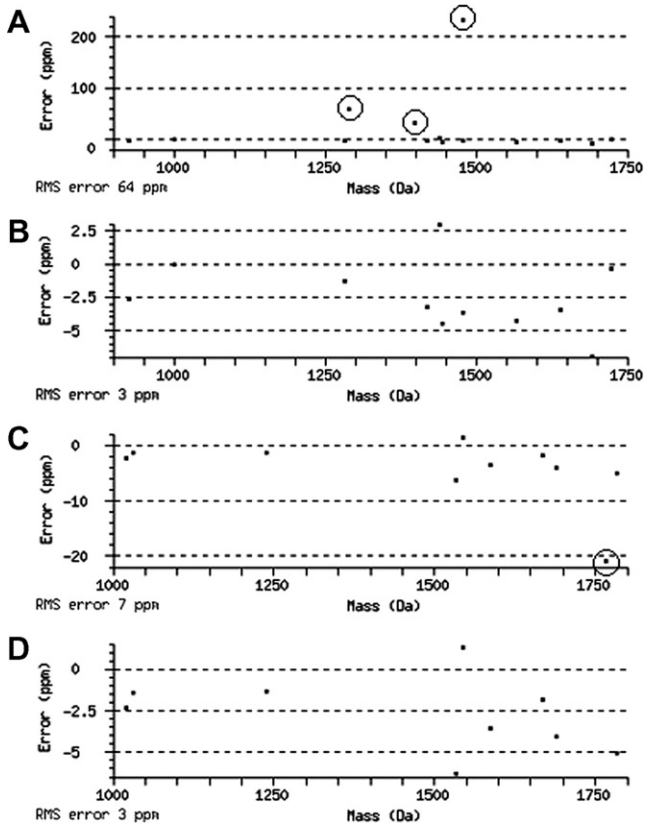

Fig. 5. Mass error distributions reported by Mascot for correct protein matches at varying mass error tolerances: BSA at 500 ppm tolerance (A) and 10 ppm tolerance (B) and HPF at 500 ppm tolerance (C) and 10 ppm tolerance (D). False positive peptides are indicated by circles in panels A and C. Note that the *y* axes are scaled differently in the various plots.

and considering only fully tryptic peptides, the value of S_M was increased slightly to 79 and the value of ΔS_M was expanded dramatically to 40, providing significant confidence in the protein assignment. The RMS mass error for the fully tryptic HHb peptides was 3 ppm. In concert, these figures of merit serve to provide high confidence in the protein identification, demonstrating a high degree of both accuracy and discrimination.

Conclusions

Uncertainty in observed peptide masses inescapably translates to uncertainty in proteomic determinations at both the peptide and protein levels. Here we have presented a thorough evaluation of mass error tolerance effects in PMF, with particular emphasis on the issue of randomly matching peptides and proteins and on how these are represented in the probability-based Mowse scoring scheme used by Mascot. With mass tolerances in the low parts-per-million range, the Mowse score was not necessarily maximized with the most stringent mass tolerances; however, the difference between the correct protein score and the score for the nearest random match was expanded considerably with more rigorous mass error restrictions regardless of the score magnitude.

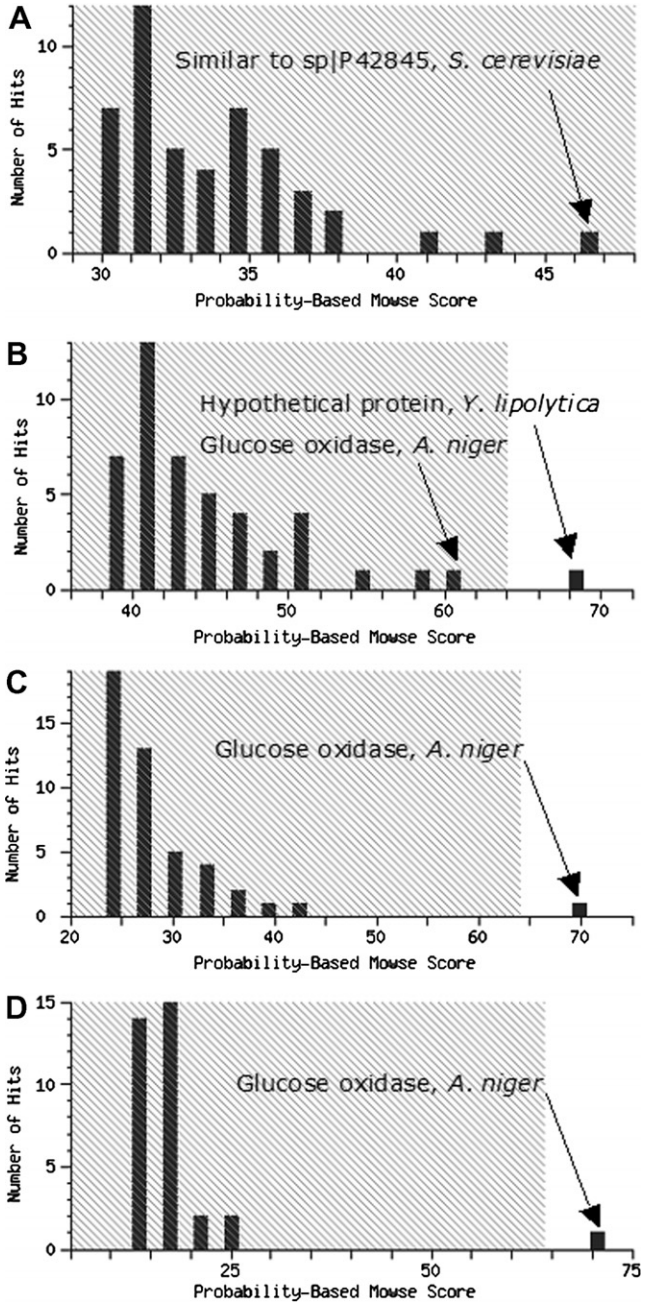

Fig. 6. Mowse score distributions reported by Mascot for PMF analysis of GO. PMF queries were conducted with a mass error tolerance of 500 ppm (A), 100 ppm (B), 25 ppm (C), and 6 ppm (D). Protein matches falling beyond the shaded region are significant at $P < 0.05$. Note that the *x* axes are scaled differently in the various plots. *S. cerevisiae*, *Saccharomyces cerevisiae; Y. lipolytica, Yarrowia lipolytica; A. niger, Aspergillus niger*.

Of particular note is that the benefits of decreased mass error tolerance are not manifest in the value of S_M when operating at tolerances in the low parts-per-million range. The advantages of operating at mass error tolerances on the order of a few parts per million are, however, apparent in the value of ΔS_M. In addition, several examples of inappropriately inflated Mowse scores were observed as a result of false positive peptides matched due to relaxed mass error

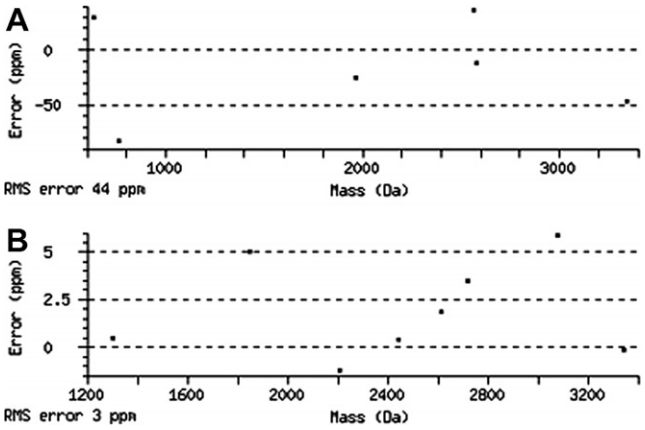

Fig. 7. Mass error distributions reported by Mascot for peptides matching the top-ranking protein hits shown in Figs. 6B and D. (A) Mass errors for peptides matching a hypothetical *Y. lipolytica* protein at 100 ppm tolerance. (B) Mass errors for peptides matching GO at 6 ppm tolerance. Note that the *y* axes are scaled differently in the two plots.

tolerance (i.e., ~100 ppm). A false positive protein identification was also observed with 100 ppm mass error tolerance, underscoring the danger of relaxed mass error tolerance in PMF and again highlighting the importance of the ΔS_M metric. Based on these findings, the reporting of both ΔS_M and S_M is proposed as a standard practice when describing the quality of a potential protein match. This would be analogous to the reporting of C_n (correlation score) and ΔC_n (difference between the two highest correlation scores) in conjunction with Sequest searches using peptide fragment ion spectra [24]. That is, one value indicates the quality of the match (e.g., S_M in Mascot PMF, C_n in Sequest MS/MS correlation) and another value indicates how well the best match is distinguished from other possibilities as a guard against false positives (e.g., ΔS_M, ΔC_n). Although the reporting of ΔC_n is standard practice for Sequest results, the equivalent value ΔS_M is rarely reported in PMF. Implementing acceptance criteria based on both S_M and ΔS_M should be a fairly simple matter and would take advantage of high mass accuracy to provide a helpful additional screen against false positive matches. However, as with Sequest ΔC_n assignments, meaningful cutoff values for ΔS_M will require empirical determination for a given database and suite of search parameters.

Overall, these results illustrate in detail the interplay of mass error and uncertainty in protein identification and, thus, provide impetus for expanding the role of high accuracy mass measurement in proteome research. These results also emphasize the need for protein identification routines that give mass measurement errors due statistical consideration as the increasing availability of high mass accuracy instrumentation calls for bioinformatic tools that take full advantage of accurate mass data. Such database searching algorithms would represent an important contribution toward taking full advantage of modern MS instrumentation and alleviating false positives in proteomics.

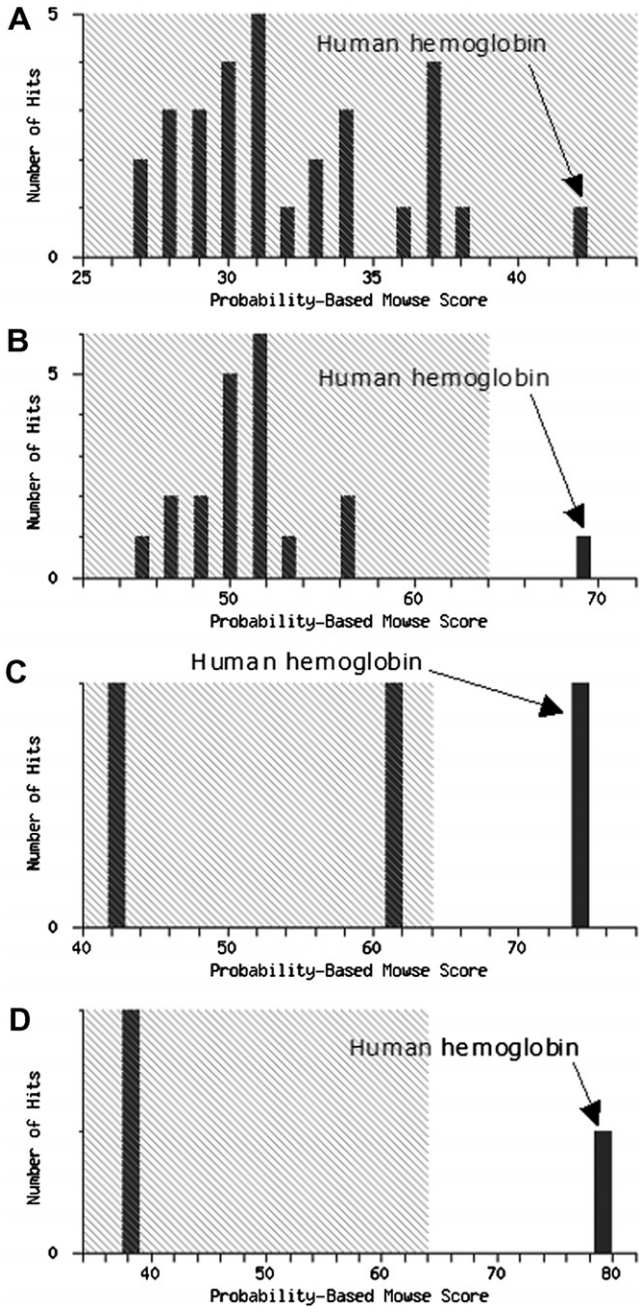

Fig. 8. Mowse score distributions reported by Mascot for PMF analysis of HHb. PMF queries were conducted with the following parameters: 500 ppm mass error tolerance and one missed tryptic cleavage allowed (A), 100 ppm tolerance and one missed cleavage allowed (B), 10 ppm tolerance and one missed cleavage allowed (C), and 10 ppm tolerance with no missed cleavages allowed (D). Protein matches falling beyond the shaded region are significant at $P < 0.05$. Note that the *x* axes are scaled differently in the various plots.

Acknowledgments

Funds provided by the following sources are gratefully acknowledged: National Institutes of Health grants AG 24488 (to P.J.H.) and GM 49077 (to C.B.L.), California Dairy Research Foundation grant 06 LEC-01-NH (to C.B.L.), and a grant from Dairy Management Incorporated.

References

[1] P. James, M. Quadroni, E. Carafoli, G. Gonnet, Protein identification by mass profile fingerprinting, Biochem. Biophys. Res. Commun. 195 (1993) 58–64.

[2] M. Mann, P. Hojrup, P. Roepstorff, Use of mass spectrometric molecular weight information to identify proteins in sequence databases, Biol. Mass Spectrom. 22 (1993) 338–345.

[3] D.J. Pappin, P. Hojrup, A.J. Bleasby, Rapid identification of proteins by peptide mass fingerprinting, Curr. Biol. 3 (1993) 327–332.

[4] J.R. Yates III, S. Speicher, P.R. Griffin, T. Hunkapiller, Peptide mass maps: A highly informative approach to protein identification, Anal. Biochem. 214 (1993) 397–408.

[5] W.J. Henzel, C. Watanabe, J.T. Stults, Protein identification: The origins of peptide mass fingerprinting, J. Am. Soc. Mass Spectrom. 14 (2003) 931–942.

[6] D.N. Perkins, D.J. Pappin, D.M. Creasy, J.S. Cottrell, Probability-based protein identification by searching sequence databases using mass spectrometry data, Electrophoresis 20 (1999) 3551–3567.

[7] J.P. Lambert, M. Ethier, J.C. Smith, D. Figeys, Proteomics: From gel based to gel free, Anal. Chem. 77 (2005) 3771–3787.

[8] P. Berndt, U. Hobohm, H. Langen, Reliable automatic protein identification from matrix-assisted laser desorption/ionization mass spectrometric peptide fingerprints, Electrophoresis 20 (1999) 3521–3526.

[9] J. Eriksson, D. Fenyo, Probity: A protein identification algorithm with accurate assignment of the statistical significance of the results, J. Proteome Res. 3 (2004) 32–36.

[10] W.R. Cannon, K.H. Jarman, B.J.M. Webb-Robertson, D.J. Baxter, C.S. Oehmen, K.D. Jarman, A. Heredia-Langner, K.J. Auberry, G.A. Anderson, Comparison of probability and likelihood models for peptide identification from tandem mass spectrometry data, J. Proteome Res. 4 (2005) 1687–1698.

[11] C. Narasimhan, D.L. Tabb, N.C. Verberkmoes, M.R. Thompson, R.L. Hettich, E.C. Uberbacher, MASPIC: Intensity-based tandem mass spectrometry scoring scheme that improves peptide identification at high confidence, Anal. Chem. 77 (2005) 7581–7593.

[12] W.J. Qian, T. Liu, M.E. Monroe, E.F. Strittmatter, J.M. Jacobs, L.J. Kangas, K. Petritis, D.G. Camp, R.D. Smith, Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and SEQUEST analysis: The human proteome, J. Proteome Res. 4 (2005) 53–62.

[13] P.A. Rudnick, Y.J. Wang, E. Evans, C.S. Lee, B.M. Balgley, Large scale analysis of MASCOT results using a mass accuracy-based threshold (MATH) effectively improves data interpretation, J. Proteome Res. 4 (2005) 1353–1360.

[14] D.B. Weatherly, J.A. Astwood III, T.A. Minning, C. Cavola, R.L. Tarleton, R. Orlando, A heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results, Mol. Cell. Proteomics 4 (2005) 762–772.

[15] H. Xie, T.J. Griffin, Trade-off between high sensitivity and increased potential for false positive peptide sequence matches using a two-dimensional linear ion trap for tandem mass spectrometry-based proteomics, J. Proteome Res. 5 (2006) 1003–1009.

[16] E.L. Huttlin, A.D. Hegeman, A.C. Harms, M.R. Sussman, Prediction of error associated with false-positive rate determination for peptide identification in large-scale proteomics experiments using a combined reverse and forward peptide sequence database strategy, J. Proteome Res. 6 (2007) 392–398.

[17] J. Bergquist, M. Palmblad, M. Wetterhall, P. Hakansson, K.E. Markides, Peptide mapping of proteins in human body fluids using electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry, Mass Spectrom. Rev. 21 (2002) 2–15.

[18] B. Bogdanov, R.D. Smith, Proteomics by FTICR mass spectrometry: Top down and bottom up, Mass Spectrom. Rev. 24 (2005) 168–200.

[19] J.S. Zimmer, M.E. Monroe, W.J. Qian, R.D. Smith, Advances in proteomics data analysis and display using an accurate mass and time tag approach, Mass Spectrom. Rev. 25 (2006) 450–482.

[20] F. He, M.R. Emmett, K. Hakansson, C.L. Hendrickson, A.G. Marshall, Theoretical and experimental prospects for protein identification based solely on accurate mass measurement, J. Proteome Res. 3 (2004) 61–67.

[21] C. Bruce, M.A. Shifman, P. Miller, E.E. Gulcicek, Probabilistic enrichment of phosphopeptides by their mass defect, Anal. Chem. 78 (2006) 4374–4382.

[22] H. Hernandez, S. Niehauser, S.A. Boltz, V. Gawandi, R.S. Phillips, I.J. Amster, Mass defect labeling of cysteine for improving peptide assignment in shotgun proteomic analyses, Anal. Chem. 78 (2006) 3417–3423.

[23] K. Tanaka, S. Takenaka, S. Tsuyama, Y. Wada, Determination of unique amino acid substitutions in protein variants by peptide mass mapping with FT–ICR MS, J. Am. Soc. Mass Spectrom. 17 (2006) 508–513.

[24] J.K. Eng, A.L. Mccormack, J.R. Yates, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database, J. Am. Soc. Mass Spectrom. 5 (1994) 976–989.

[25] J.D. Venable, T. Xu, D. Cociorva, J.R. Yates III, Cross-correlation algorithm for calculation of peptide molecular weight from tandem mass spectra, Anal. Chem. 78 (2006) 1921–1929.

[26] J.R. Yates, D. Cociorva, L. Liao, V. Zabrouskov, Performance of a linear ion trap–Orbitrap hybrid for peptide analysis, Anal. Chem. 78 (2006) 493–500.

[27] A. Brock, D.M. Horn, E.C. Peters, C.M. Shaw, C. Ericson, Q.T. Phung, A.R. Salomon, An automated matrix-assisted laser desorption/ionization quadrupole Fourier transform ion cyclotron resonance mass spectrometer for "bottom-up" proteomics, Anal. Chem. 75 (2003) 3419–3428.

[28] B. Spengler, De novo sequencing, peptide composition analysis, and composition-based sequencing: A new strategy employing accurate mass determination by Fourier transform ion cyclotron resonance mass spectrometry, J. Am. Soc. Mass Spectrom. 15 (2004) 703–714.

[29] M.L. Nielsen, M.M. Savitski, R.A. Zubarev, Improving protein identification using complementary fragmentation techniques in Fourier transform mass spectrometry, Mol. Cell. Proteomics 4 (2005) 835–845.

[30] J.V. Olsen, L.M. de Godoy, G. Li, B. Macek, P. Mortensen, R. Pesch, A. Makarov, O. Lange, S. Horning, M. Mann, Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap, Mol. Cell. Proteomics 4 (2005) 2010–2021.

[31] M.M. Savitski, M.L. Nielsen, F. Kjeldsen, R.A. Zubarev, Proteomics-grade de novo sequencing approach, J. Proteome Res. 4 (2005) 2348–2354.

[32] M.M. Savitski, M.L. Nielsen, R.A. Zubarev, New database-independent, sequence tag-based scoring of peptide MS/MS data validates Mowse scores, recovers below threshold data, singles out modified peptides, and assesses the quality of MS/MS techniques, Mol. Cell. Proteomics 4 (2005) 1180–1188.

[33] S. Wu, N.K. Kaiser, D. Meng, G.A. Anderson, K. Zhang, J.E. Bruce, Increased protein identification capabilities through novel tandem MS calibration strategies, J. Proteome Res. 4 (2005) 1434–1441.

[34] T. Kocher, M.M. Savitski, M.L. Nielsen, R.A. Zubarev, PhosTShunter: A fast and reliable tool to detect phosphorylated peptides in liquid chromatography Fourier transform tandem mass spectrometry data sets, J. Proteome Res. 5 (2006) 659–668.

[35] R.L. Wong, I.J. Amster, Combining low and high mass ion accumulation for enhancing shotgun proteome analysis by accurate mass measurement, J. Am. Soc. Mass Spectrom. 17 (2006) 205–212.

[36] A.M. Frank, M.M. Savitski, M.L. Nielsen, R.A. Zubarev, P.A. Pevzner, De novo peptide sequencing and identification with precision mass spectrometry, J. Proteome Res. 6 (2007) 114–123.

[37] D.K. Williams, A.M. Hawkridge, D.C. Muddiman, Sub parts-per-million mass measurement accuracy of intact proteins and product

ions achieved using a dual electrospray ionization quadrupole Fourier transform ion cyclotron resonance mass spectrometer, J. Am. Soc. Mass Spectrom. 18 (2007) 1–7.
[38] O.N. Jensen, A.V. Podtelejnikov, M. Mann, Identification of the components of simple protein mixtures by high-accuracy peptide mass mapping and database searching, Anal. Chem. 69 (1997) 4741–4750.
[39] K.R. Clauser, P. Baker, A.L. Burlingame, Role of accurate mass measurement ((10 ppm) in protein identification strategies employing MS or MS/MS and database searching, Anal. Chem. 71 (1999) 2871–2882.
[40] M.K. Green, M.V. Johnston, B.S. Larsen, Mass accuracy and sequence requirements for protein database searching, Anal. Biochem. 275 (1999) 39–46.
[41] M. Witt, J. Fuchser, G. Baykut, 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. 14 (2003) 553–561.
[42] D.M. Horn, E.C. Peters, H. Klock, A. Meyers, A. Brock, Improved protein identification using automated high mass measurement accuracy MALDI FT–ICR MS peptide mass fingerprinting, Intl. J. Mass Spectrom. 238 (2004) 189–196.
[43] R. Zubarev, M. Mann, On the proper use of mass accuracy in proteomics, Mol. Cell. Proteomics 6 (2007) 377–381.
[44] T.H. Mize, I.J. Amster, Broad-band ion accumulation with an internal source MALDI–FTICR–MS, Anal. Chem. 72 (2000) 5886–5891.
[45] P.B. O'Connor, C.E. Costello, Internal calibration on adjacent samples (InCAS) with Fourier transform mass spectrometry, Anal. Chem. 72 (2000) 5881–5885.
[46] B. Paizs, S. Suhai, Fragmentation pathways of protonated peptides, Mass Spectrom. Rev. 24 (2005) 508–548.
[47] A.G. Marshall, C.L. Hendrickson, G.S. Jackson, Fourier transform ion cyclotron resonance mass spectrometry: A primer, Mass Spectrom. Rev. 17 (1998) 1–35.
[48] L.K. Zhang, D. Rempel, B.N. Pramanik, M.L. Gross, Accurate mass measurements by Fourier transform mass spectrometry, Mass Spectrom. Rev. 24 (2005) 286–309.
[49] M. Mann, Useful tables of possible and probable peptide masses, 43rd Annual Conference on Mass Spectrometry and Allied Topics, Atlanta, GA, 1995. (American Society for Mass Spectrometry).
[50] R.A. Zubarev, P. Hakansson, B. Sundqvist, Accuracy requirements for peptide characterization by monoisotopic molecular mass measurements, Anal. Chem. 68 (1996) 4060–4063.
[51] J.A. Karty, M.M.E. Ireland, Y.V. Brun, J.P. Reilly, Artifacts and unassigned masses encountered in peptide mass mapping, J. Chromatogr. B 782 (2002) 363–383.
[52] B.D. Piening, P. Wang, C.S. Bangur, J. Whiteaker, H. Zhang, L.C. Feng, J.F. Keane, J.K. Eng, H. Tang, A. Prakash, M.W. McIntosh, A. Paulovich, Quality control metrics for LC–MS feature detection tools demonstrated on *Saccharomyces cerevisiae* proteomic profiles, J. Proteome Res. 5 (2006) 1527–1534.
[53] E.D. Dodds, H.J. An, P.J. Hagerman, C.B. Lebrilla, Enhanced peptide mass fingerprinting through high mass accuracy: Exclusion of non-peptide signals based on residual mass, J. Proteome Res. 5 (2006) 1195–1203.
[54] E. D. Dodds, B. H. Clowers, H. J. An, P. J. Hagerman, C. B. Lebrilla, Low mass error tolerance and the Mass Sieve approach to peptide mass fingerprinting with MALDI–FTICR–MS, 54th Annual Conference on Mass Spectrometry and Allied Topics, Seattle, WA, 2006. (American Society for Mass Spectrometry).
[55] D. J. Slotta, M. A. McFarland, A. J. Makusky, S. P. Markey, MassSieve: A new visualization tool for mass spectrometry-based proteomics, 55th Annual Conference on Mass Spectrometry and Allied Topics, Indianapolis, IN, 2007. (American Society for Mass Spectrometry).
[56] J.A. Siepen, E.J. Keevil, D. Knight, S.J. Hubbard, Prediction of missed cleavage sites in tryptic peptides aids protein identification in proteomics, J. Proteome Res. 6 (2007) 399–408.