Published in final edited form as:

Anal Biochem. 2008 January 15; 372(2): 156–166. doi:10.1016/j.ab.2007.10.009.

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, and Carlito B. Lebrilla^{a,b,*}

^a Department of Chemistry, University of California Davis, Davis, California, USA

^b School of Medicine, Department of Biochemistry and Molecular Medicine, University of California Davis, Davis, California, USA

Abstract

While the bearing of mass measurement error upon protein identification is sometimes underestimated, uncertainty in observed peptide masses unavoidably translates to ambiguity in subsequent protein identifications. While 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. Additionally, the ranking schemes of most protein identification algorithms do not include a meaningful incorporation of mass measurement error. This report 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

*Corresponding Author. Prof. Carlito B. Lebrilla, Department of Chemistry, University of California Davis, One Shields Ave, Davis, CA 95616 USA. E-mail: E-mail: cblebrilla@ucdavis.edu. Telephone: 1-530-752-6364. Fax: 1-530-754-5609.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Introduction

Over the past decade, peptide mass fingerprinting (PMF) 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 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 to 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 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 is only recently garnering broad acknowledgement. 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 data sets [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 those involving PMF. This is an ironic reversal, as approaches yielding peptide sequence information are somewhat less sensitive to mass error than 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]. While determination of accurate peptide masses should in principle deliver an even greater margin of improvement in PMF, published developments involving high mass accuracy with regards to PMF have been surprisingly sparse, particularly in comparison to 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 most protein identification algorithms and scoring schemes do not take mass measurement error into account when assigning ranks to putative matches [43].

This report 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 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 towards establishing a more pivotal role for high accuracy mass measurement in proteome research. Considering the increasing availability of high accuracy mass spectrometry 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). 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 8 M urea/200 mM tris buffer (pH 7.8). Prior to digestion, proteins were reduced (by addition of 10 µL 450 mM dithiothreitol in 50 mM NH₄HCO₃ with incubation at 55°C for one hour) and alkylated (by addition of 10 µL 500 mM iodoacetamide in 50 mM NH₄HCO₃ with incubation in the dark at ambient temperature for 30 min). Each preparation was then diluted to < 2 M in urea by addition of 150 µL deionized water and treated with 1 µL of a 0.05 µg/µL solution of sequencing grade modified trypsin (Promega, Madison, WI). Digestion was allowed to proceed for approximately 8 – 10 hours 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). Desalted tryptic peptides were eluted in 10 µL 50% acetonitrile (ACN) with 0.1% trifluoroacetic acid (TFA).

Mass Spectrometry

A matrix solution of 50 µg/µL 2,5-dihydroxybenzoic acid (DHB) was prepared in 50% ACN. Samples were prepared for MALDI by combining 1 µL of the purified tryptic digest and 1 µL 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 in order to approximate a realistic quantity of protein digest. An IonSpec Corporation HiResMALDI FTICR-MS (Lake Forest, CA) 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 μ L of calibrant solution and 1 μ L of DHB matrix solution. Two calibrant mixtures were separately deposited on the MALDI target: 1 μ M P₁₄R (a labile synthetic peptide yielding y-series fragments through metastable decay associated with the “proline effect” [46]), and a mixture of P₁₄R, human adrenocorticotrophic hormone (ACTH) fragment peptide 18–39, and bovine insulin (BI) oxidized B chain (each at a concentration of 1 μ M). 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 500 – 3500 was done in order to determine which calibration spot was appropriate for the mass range of peptides observed in each digest. P₁₄R produced calibrant ions spanning approximately m/z 750 – 1530, while the P₁₄R, 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 one second duration, while 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 – 3500. Spectra for PMF analysis were acquired in the mass range of m/z 500 – 2500 or m/z 500 – 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:

$$\frac{m}{z} = \frac{A}{f_c - B} \quad (1)$$

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, since 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 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].

While 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). This updated version of Mass Sieve remained essentially equivalent to the original program; however, Mass Sieve in IGOR Pro was improved by addition of a more convenient graphical 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 VB 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 $m/z < 700$ 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 present study was developed in this laboratory, and should not be mistaken for an unrelated proteomics program that has more recently assumed the name “MassSieve” [55].

Peptide Mass Fingerprint 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. The present work was thus focused on Mowse scoring with Mascot, allowing the forthcoming results to be of immediate and direct use to the large community of proteomics 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 specifically noted. While 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 have recently been provided by Zubarev and Mann [43].

Results and Discussion

Three representative MALDI-FTICR-MS PMFs are shown in Figure 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 were plotted in Figure 2. The root mean square (RMS) error of the fifty measurements 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%. While 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 (Figure 3). At a mass error tolerance of 500 ppm, the protein was correctly identified with a statistically significant Mowse score of 84 (Figure 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 (Figure 3b). These Mowse scores, S_M , are related to the absolute probability P that the protein match is a random event according to:

$$S_M = -10 \log(P) \quad (2)$$

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 at random with a frequency of $< 5\%$. 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 (Figure 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, although even with a reduced number of peptides matched the Mowse score was increased and the root mean square mass error for the correctly matching peptides was 1.4 ppm (Figure 3d).

In addition to eliminating false positive peptides, the discrimination of PMF searching is greatly improved 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}):

$$\Delta S_M = S_{M1} - S_{M2} \quad (3)$$

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

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

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

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

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 (Figure 4). For BSA, CEA, GO, HAT and HPF, the value of S_M exhibited remarkably little dependence upon mass error tolerances at 100 ppm and below (Figure 4a). This observation may initially seem counterintuitive, and on the surface 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, while the present 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 percent range, increases in S_M are correlated to decreased mass error tolerance. In comparison, the values of S_M are shown here to have little dependence on mass error tolerance in the parts per million regime. This is evidently a consequence of the non-uniform distribution of peptide masses, which, 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 (Figure 4b). Interestingly, the values of S_M for CEA in Figure 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 Figure 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 may be necessary. For these searches involving much larger databases, low mass error tolerance becomes even more valuable in the prevention of false positives.

A closer examination of Figure 4(b) revealed a few cases in which the values of S_M and ΔS_M unexpectedly decreased with decreasing mass error tolerance. For example, the Mowse score for BSA dropped significantly at tolerances below 100 ppm. A similar situation occurred for HPF with a mass error tolerance below 25 ppm; in this case, the values of both S_M and ΔS_M were depreciated. While it was initially unclear as to why these metrics were being adversely affected by reduced mass error tolerances, an explanation was found upon examination of the mass error distributions of matched peptides. As illustrated in Figure 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.

While the inclusion of false positive peptides in assignment of Mowse scores is troublesome, the assignment of significant score to an incorrect protein match is of even greater concern. Such an example is depicted in Figure 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 *Y. lipolytica*. At this mass tolerance a hit for GO was also reported, but was scored below the threshold of statistical significance ($p < 0.05$). While 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. Firstly, as seen in Figure 6(b), 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 $\Delta S_M = 8$, as 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. Secondly, the mass error distribution associated with each protein identification provided further reason for skepticism regarding the top hit. As shown in Figure 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 expectation for internally calibrated FTICR-MS data). Without the ability to set this stringent mass tolerance, the protein could not have been correctly identified 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.

While 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 Figure 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 although 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. While some allowance for incomplete tryptic digestion is generally applied to proteomics 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 has recently been highlighted by Siepen *et al.* [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 and considering only fully tryptic peptides, the value of S_M was increased slightly to 79 and the value of ΔS_M was dramatically expanded 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 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 considerably expanded with more rigorous mass error restrictions regardless of the score magnitude.

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 ppm range. The advantages of operating at mass error tolerances on the order of a few ppm 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 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 is reported that indicates the quality of the match (such as S_M in Mascot PMF or C_n in Sequest tandem MS correlation) and another value indicates how well the best match is distinguished from other possibilities as a guard against false positives (ΔS_M or ΔC_n). Although the reporting of ΔC_n is standard practice for Sequest results, the equivalent value ΔS_M is rarely reported in PMF. To implement 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 towards taking full advantage of modern MS instrumentation and alleviating false positives in proteomics.

Acknowledgments

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

References

1. James P, Quadroni M, Carafoli E, Gonnet G. Protein identification by mass profile fingerprinting. *Biochem Biophys Res Commun* 1993;195:58–64. [PubMed: 8363627]
2. Mann M, Hojrup P, Roepstorff P. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol Mass Spectrom* 1993;22:338–345. [PubMed: 8329463]
3. Pappin DJ, Hojrup P, Bleasby AJ. Rapid identification of proteins by peptide-mass fingerprinting. *Curr Biol* 1993;3:327–332. [PubMed: 15335725]
4. Yates JR 3rd, Speicher S, Griffin PR, Hunkapiller T. Peptide mass maps: a highly informative approach to protein identification. *Anal Biochem* 1993;214:397–408. [PubMed: 8109726]
5. Henzel WJ, Watanabe C, Stults JT. Protein identification: the origins of peptide mass fingerprinting. *J Am Soc Mass Spectrom* 2003;14:931–942. [PubMed: 12954162]
6. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999;20:3551–3567. [PubMed: 10612281]
7. Lambert JP, Ethier M, Smith JC, Figeys D. Proteomics: from gel based to gel free. *Anal Chem* 2005;77:3771–3787. [PubMed: 15952756]
8. 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. [PubMed: 10612278]
9. Eriksson J, Fenyo D. Probity: a protein identification algorithm with accurate assignment of the statistical significance of the results. *J Proteome Res* 2004;3:32–36. [PubMed: 14998160]
10. Cannon WR, Jarman KH, Webb-Robertson BJM, Baxter DJ, Oehmen CS, Jarman KD, Heredia-Langner A, Auberry KJ, Anderson GA. Comparison of probability and likelihood models for peptide

- identification from tandem mass spectrometry data. *J Proteome Res* 2005;4:1687–1698. [PubMed: 16212422]
11. Narasimhan C, Tabb DL, VerBerkmoes NC, Thompson MR, Hettich RL, Uberbacher EC. MASPIC: intensity-based tandem mass spectrometry scoring scheme that improves peptide identification at high confidence. *Anal Chem* 2005;77:7581–7593. [PubMed: 16316165]
 12. Qian WJ, Liu T, Monroe ME, Strittmatter EF, Jacobs JM, Kangas LJ, Petritis K, Camp DG, Smith RD. Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and SEQUEST analysis: the human proteome. *J Proteome Res* 2005;4:53–62. [PubMed: 15707357]
 13. Rudnick PA, Wang YJ, Evans E, Lee CS, Balgley BM. Large scale analysis of MASCOT results using a mass accuracy-based THreshold (MATH) effectively improves data interpretation. *J Proteome Res* 2005;4:1353–1360. [PubMed: 16083287]
 14. Weatherly DB, Astwood JA 3rd, Minning TA, Cavola C, Tarleton RL, Orlando R. A heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results. *Mol Cell Proteomics* 2005;4:762–772. [PubMed: 15703444]
 15. Xie H, Griffin TJ. 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* 2006;5:1003–1009. [PubMed: 16602709]
 16. Huttlin EL, Hegeman AD, Harms AC, Sussman MR. 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* 2007;6:392–398. [PubMed: 17203984]
 17. Bergquist J, Palmblad M, Wetterhall M, Hakansson P, Markides KE. 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. [PubMed: 12210611]
 18. Bogdanov B, Smith RD. Proteomics by FTICR mass spectrometry: top down and bottom up. *Mass Spectrom Rev* 2005;24:168–200. [PubMed: 15389855]
 19. Zimmer JS, Monroe ME, Qian WJ, Smith RD. Advances in proteomics data analysis and display using an accurate mass and time tag approach. *Mass Spectrom Rev* 2006;25:450–482. [PubMed: 16429408]
 20. He F, Emmett MR, Hakansson K, Hendrickson CL, Marshall AG. Theoretical and experimental prospects for protein identification based solely on accurate mass measurement. *J Proteome Res* 2004;3:61–67. [PubMed: 14998164]
 21. Bruce C, Shifman MA, Miller P, Gulcicek EE. Probabilistic enrichment of phosphopeptides by their mass defect. *Anal Chem* 2006;78:4374–4382. [PubMed: 16808444]
 22. Hernandez H, Niehauser S, Boltz SA, Gawandi V, Phillips RS, Amster IJ. Mass defect labeling of cysteine for improving peptide assignment in shotgun proteomic analyses. *Anal Chem* 2006;78:3417–3423. [PubMed: 16689545]
 23. Tanaka K, Takenaka S, Tsuyama S, Wada Y. Determination of unique amino acid substitutions in protein variants by peptide mass mapping with FT-ICR MS. *J Am Soc Mass Spectrom* 2006;17:508–513. [PubMed: 16490358]
 24. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 1994;5:976–989.
 25. Venable JD, Xu T, Cociorva D, Yates JR 3rd. Cross-correlation algorithm for calculation of peptide molecular weight from tandem mass spectra. *Anal Chem* 2006;78:1921–1929. [PubMed: 16536429]
 26. Yates JR, Cociorva D, Liao L, Zabrouskov V. Performance of a linear ion trap-Orbitrap hybrid for peptide analysis. *Anal Chem* 2006;78:493–500. [PubMed: 16408932]
 27. Brock A, Horn DM, Peters EC, Shaw CM, Ericson C, Phung QT, Salomon AR. An automated matrix-assisted laser desorption/ionization quadrupole Fourier transform ion cyclotron resonance mass spectrometer for “bottom-up” proteomics. *Anal Chem* 2003;75:3419–3428. [PubMed: 14570192]
 28. Spengler B. 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* 2004;15:703–714. [PubMed: 15121200]

29. Nielsen ML, Savitski MM, Zubarev RA. Improving protein identification using complementary fragmentation techniques in Fourier transform mass spectrometry. *Mol Cell Proteomics* 2005;4:835–845. [PubMed: 15772112]
30. Olsen JV, de Godoy LM, 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. [PubMed: 16249172]
31. Savitski MM, Nielsen ML, Kjeldsen F, Zubarev RA. Proteomics-grade de novo sequencing approach. *J Proteome Res* 2005;4:2348–2354. [PubMed: 16335984]
32. Savitski MM, Nielsen ML, Zubarev RA. New data base-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* 2005;4:1180–1188. [PubMed: 15911534]
33. Wu S, Kaiser NK, Meng D, Anderson GA, Zhang K, Bruce JE. Increased protein identification capabilities through novel tandem MS calibration strategies. *J Proteome Res* 2005;4:1434–1441. [PubMed: 16083297]
34. Kocher T, Savitski MM, Nielsen ML, Zubarev RA. PhosTShunter: a fast and reliable tool to detect phosphorylated peptides in liquid chromatography Fourier transform tandem mass spectrometry data sets. *J Proteome Res* 2006;5:659–668. [PubMed: 16512682]
35. Wong RL, Amster IJ. Combining low and high mass ion accumulation for enhancing shotgun proteome analysis by accurate mass measurement. *J Am Soc Mass Spectrom* 2006;17:205–212. [PubMed: 16413206]
36. Frank AM, Savitski MM, Nielsen ML, Zubarev RA, Pevzner PA. De novo peptide sequencing and identification with precision mass spectrometry. *J Proteome Res* 2007;6:114–123. [PubMed: 17203955]
37. Williams DK, Hawkrigde AM, Muddiman DC. 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* 2007;18:1–7. [PubMed: 16979902]
38. Jensen ON, Podtelejnikov AV, Mann M. Identification of the components of simple protein mixtures by high-accuracy peptide mass mapping and database searching. *Anal Chem* 1997;69:4741–4750. [PubMed: 9406524]
39. Clauser KR, Baker P, Burlingame AL. Role of accurate mass measurement (± 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal Chem* 1999;71:2871–2882. [PubMed: 10424174]
40. Green MK, Johnston MV, Larsen BS. Mass accuracy and sequence requirements for protein database searching. *Anal Biochem* 1999;275:39–46. [PubMed: 10542107]
41. 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. [PubMed: 12781456]
42. Horn DM, Peters EC, 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.
43. Zubarev R, Mann M. On the proper use of mass accuracy in proteomics. *Mol Cell Proteomics* 2007;6:377–381. [PubMed: 17164402]
44. Mize TH, Amster IJ. Broad-band ion accumulation with an internal source MALDI-FTICR-MS. *Anal Chem* 2000;72:5886–5891. [PubMed: 11140753]
45. O'Connor PB, Costello CE. Internal calibration on adjacent samples (InCAS) with Fourier transform mass spectrometry. *Anal Chem* 2000;72:5881–5885. [PubMed: 11140752]
46. Paizs B, Suhai S. Fragmentation pathways of protonated peptides. *Mass Spectrom Rev* 2005;24:508–548. [PubMed: 15389847]
47. Marshall AG, Hendrickson CL, Jackson GS. Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrom Rev* 1998;17:1–35. [PubMed: 9768511]

48. Zhang LK, Rempel D, Pramanik BN, Gross ML. Accurate mass measurements by Fourier transform mass spectrometry. *Mass Spectrom Rev* 2005;24:286–309.
49. Mann, M. Useful tables of possible and probable peptide masses. American Society for Mass Spectrometry - 43rd Annual Conference on Mass Spectrometry and Allied Topics; Atlanta, GA, USA. 1995.
50. Zubarev RA, Hakansson P, Sundqvist B. Accuracy requirements for peptide characterization by monoisotopic molecular mass measurements. *Anal Chem* 1996;68:4060–4063.
51. Karty JA, Ireland MME, Brun YV, Reilly JP. Artifacts and unassigned masses encountered in peptide mass mapping. *J Chromatogr B* 2002;782:363–383.
52. Piening BD, Wang P, Bangur CS, Whiteaker J, Zhang H, Feng LC, Keane JF, Eng JK, Tang H, Prakash A, McIntosh MW, Paulovich A. Quality control metrics for LC-MS feature detection tools demonstrated on *Saccharomyces cerevisiae* proteomic profiles. *J Proteome Res* 2006;5:1527–1534. [PubMed: 16823959]
53. Dodds ED, An HJ, Hagerman PJ, Lebrilla CB. Enhanced peptide mass fingerprinting through high mass accuracy: exclusion of non-peptide signals based on residual mass. *J Proteome Res* 2006;5:1195–1203. [PubMed: 16674109]
54. Dodds, ED.; Clowers, BH.; An, HJ.; Hagerman, PJ.; Lebrilla, CB. Low mass error tolerance and the Mass Sieve approach to peptide mass fingerprinting with MALDI-FTICR-MS. American Society for Mass Spectrometry - 54th Annual Conference on Mass Spectrometry and Allied Topics; Seattle, WA, USA. 2006.
55. Slotta, DJ.; McFarland, MA.; Makusky, AJ.; Markey, SP. MassSieve: a new visualization tool for mass spectrometry-based proteomics. American Society for Mass Spectrometry - 55th Annual Conference on Mass Spectrometry and Allied Topics; Indianapolis, IN, USA. 2007.
56. Siepen JA, Keevil EJ, Knight D, Hubbard SJ. Prediction of missed cleavage sites in tryptic peptides aids protein identification in proteomics. *J Proteome Res* 2007;6:399–408. [PubMed: 17203985]

Figure 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, while masses matched to the correct protein to within 10 ppm error are labeled with squares.

Figure 2. Mass errors for fifty tryptic peptides correctly matched to their respective target proteins. The root mean square mass error was 3 ppm.

Figure 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 (a) and (c), and a mass error tolerance of 3 ppm in (b) and (d). False positive peptides are indicated by circling in (c). Mowse scores in (a) and (b) falling beyond the shaded region are significant at $p < 0.05$. Note that the x axes are scaled differently in (a) and (b), and the y axes are scaled differently in (c) and (d).

Figure 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).

Figure 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); HPF 500 ppm tolerance (c) and 10 ppm tolerance (d). False positive peptides are indicated by circling in (a) and (c). Note that the y axes are scaled differently in each plot.

Figure 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 each plot.

Figure 7.

Mass error distributions reported by Mascot for peptides matching the top-ranking protein hits shown in Figures 6(b) and 6(d). Mass errors for peptides matching a hypothetical *Yarrowia lipolytica* protein at 100 ppm tolerance (a); and mass errors for peptides matching GO at 6 ppm tolerance (b). Note that the y axes are scaled differently in each plot.

Figure 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 each plot.