Published in final edited form as:

Anal Chem. 2012 March 20; 84(6): 3026–3032. doi:10.1021/ac203255e.

Improved Mass Defect Model for Theoretical Tryptic Peptides

Indranil Mitra^{1,2}, Alexey V. Nefedov^{1,2}, Allan R. Brasier^{1,2,3}, and Rovshan G. Sadygov^{1,2,*}

¹Department of Biochemistry and Molecular Biology, The University of Texas Medical Branch, Galveston, TX 77555

²Sealy Center for Molecular Medicine, The University of Texas Medical Branch, Galveston, TX 77555

³Institute for Translational Sciences, The University of Texas Medical Branch, Galveston, TX 77555

Abstract

Improvements in the mass accuracy and resolution of mass spectrometers have greatly aided mass spectrometry-based proteomics in profiling complex biological mixtures. With the use of innovative bioinformatics approaches, high mass accuracy and resolution information can be used for filtering chemical noise in mass spectral data. Using our recent algorithmic developments, we have generated the mass distributions of all theoretical tryptic peptides composed of twenty natural amino acids and with masses limited to 3.5 kDa. Peptide masses are distributed discretely, with well defined peak clusters separated by empty or sparsely populated trough regions. Accurate models for peak centers and widths can be used to filter peptide signals from chemical noise. We modeled mass defects, the difference between monoisotopic and nominal masses, peak centers and widths in the peptide mass distributions. We found that peak widths encompassing 95% of all peptide sequences are substantially smaller than previously thought. The result has implications for filtering out larger stretches of the mass axis. Mass defects of peptides exhibit an oscillatory behavior which is damped at high mass values. The periodicity of the oscillations is about 14 Da which is the most common difference between the masses of the twenty natural amino acids.

To determine the effects of amino acid modifications on our findings, we examined the mass distributions of peptides composed of the twenty natural amino acids, oxidized Met and phosphorylated Ser, Thr and Tyr. We found that extension of the amino acid set by modifications increases the 95% peak width. Mass defects decrease, reflecting the fact that the average mass defect of natural amino acids is larger than that of oxidized Met. We propose that a new model for mass defects and peak widths of peptides may improve peptide identifications by filtering chemical noise in mass spectral data.

Keywords

peptide mass distributions; forbidden zones; mass defect; noise filtering; power spectrum; high mass accuracy and resolution

*To whom correspondence should be addressed. Phone: +1 409 772 3287. Fax: +1 409 772 9679. rovshan.sadygov@utmb.edu.

Supporting Information Available: Supplementary materials include Figures S1-S4. These are mass defect, Figure S1, and peak width, Figure S2, of the mass distributions of theoretical peptides composed of twenty natural amino acids and oxidized Met. Figures S3 and S4 are the relative differences in peak center locations between the peptides composed of only natural amino acids and peptides composed of natural amino acids and oxidized Met, Figure S3; and phosphorylated Ser, Thr and Tyr, Figure S4. This material is available free of charge via the Internet at <http://pubs.acs.org>.

Introduction

In recent years, mass spectrometry (MS)-based proteomics has developed into a powerful technology to study complex biological samples.^{1,2} MS coupled to liquid chromatography can identify and quantify thousands of proteins from complex samples in high-throughput experiments. This has been made possible by major advances in instrumentation^{3,4}, where the latest generations of mass spectrometers are capable of achieving 100,000 or better resolution and mass accuracy of less than 5 parts per million (ppm) at the femtomole level of detection.

These technological developments have created interest in profiling mass distributions of all theoretically possible peptides.⁵⁻⁷ Mann was the first to recognize that at high mass accuracy, peptide masses are clustered and non-continuous with disallowed regions (forbidden zones) between the clusters.⁸ This observation can be used, for example, to filter mass spectral data of peptides from chemical/electronic noise.

As an illustration, we show in Figure 1 the distribution of peptide masses for a range of different masses. For smaller masses there are clearly defined forbidden zones. They become smaller for large masses and effectively disappear (still sparsely populated compared to the peak clusters) at 2 kDa. The theoretical mass accuracy used for generating this figure and throughout this work is 0.001 Da. Another observation relates to the positions of peak clusters on the mass axis. They are evolving with peptide mass relative to the integer mass values. As seen from the figure, at smaller masses, e.g., 1528 Da, the peak centers are somewhat in the middle of two consecutive integers. The peak centers “shift” to the next integer value as the peptide masses increase, and, at 2128 Da, the peak center is located on the integer value. After this, the peak centers continue to evolve and, at 2627 Da, they are again located in between two consecutive integers. This peak dynamics represents peptides. It is expected that different chemical species will have their own peak dynamics (due to the different building blocks of each class of species). The concept of the mass defect (defined below) filter is to capture this effect and use it to separate other species from peptides.

Mann⁸ has modeled the distribution of the peak centers and peak widths (PW) encompassing 95% of all peptide sequences with a linear function of nominal masses. The model was developed based on all theoretical peptides up to 2 kDa in mass. It has been used to filter chemical/electronic noise in peptide mass finger printing⁹ and tandem mass spectrometry,¹⁰ to predict masses of glycosylated peptides¹¹ and in mass defect labeling.¹² Others have used a similar approach to model the peak capacity in a sample space of mass spectrometry coupled to liquid chromatography.¹³ It is important to note that Mann’s model was formulated for peptides with masses up to 2 kDa. However, the above mentioned applications employed this model in higher mass regions.

Recently, computational methods for generating mass space of theoretical peptides have been used to study valence parity, in order to distinguish between c' and z' fragment ions in electron capture/electron transfer dissociations^{7,14} and periodicity in monoisotopic mass distributions of peptides.⁵

In this study, we modeled the position and width of peaks in the mass distribution of all theoretical peptides. We used our algorithm^{15,16} for generating amino acid compositions of theoretical peptides to obtain peptide sequences. We found that in general, the observed distributions of peptide peak widths are non-linear. This observation is used in modeling 95% PW. Our model employs a combination of linear and non-linear mass dependencies and provides PW values that are substantially smaller than those derived from Mann’s linear model. Smaller PW values will allow for larger regions of mass axis to be filtered as non-peptidic, and will increase the power of filtering. We also analyzed the effects of amino acid

modifications on the mass distributions of theoretical peptides. The software for generating the mass defects of theoretical peptides under a variety of enzymatic specificities and digest conditions (number of missed cleavages) is available for download at <https://ispace.utmb.edu/users/rgsadygo/Proteomics/MassDefect>.

Methods

Generating the mass distribution of all theoretically possible peptide sequences is a problem of exponential complexity. For a peptide of length L composed of 20 common amino acids, there are 20^L possible sequences. The total number of sequences of length not greater than L is $20*(20^L-1)/19$. Recently, we described a recursive algorithm for computing the masses of amino acid compositions of all theoretically possible peptides.^{15,16} Our algorithm is applicable to different enzymatic as well as cleavage conditions, post-translational modifications and restrictions on the number of amino acids. In the present work, we use this algorithm to generate mass distributions of peptide sequences and to compute total mass defects (whose definition follows below). We use the distributions to model mass defects of all theoretical peptides, their peak centers and 95% PW value. In this work, unless stated otherwise, when we refer to the theoretical tryptic peptides we mean all theoretical tryptic peptides comprised of twenty natural amino acids without modifications and peptide masses not greater than 3.5 kDa.

Determining Peak Center and 95% Peak Widths

The number of peptide sequences is generated from the number of amino acid compositions. For a composition of length L , the total number of sequences is given by a multinomial coefficient:

$$\frac{L!}{n_1!n_2!\dots n_N!}$$

where, n_i is equal to the number of times the i -th amino acid occurs in the peptide. The width of the mass bins was 0.001 Da. The algorithm computes protonated peptide masses up to the 8th digit after the decimal point (using monoisotopic masses of amino acids, proton and H_2O), and then rounds them to the 3rd digit after the decimal point. A higher mass accuracy does not incur additional computational expense. However, considering the current mass accuracy available in many modern mass spectrometers (~ 5 ppm) we have chosen to limit the theoretical mass accuracy to 0.001 Da.

Using generated mass distributions we computed the position of peak center and peak's width. For mass values up to 1.8 kDa, peaks are well isolated. In this mass range there were well-defined forbidden zones between peak clusters, Figure 1. For these peaks, the centers were determined as the center of mass of all amino acid sequences (see below) whose mass fell into the peak interval. 95% PW values for these peaks were determined by incrementally moving from the peak center to the peak tails and determining the percentage of total sequences encompassed in each interval.

For larger peptides (mass values > 1.8 kDa), we used a two-step procedure to determine peak positions. In the first step, we determined preliminary peak positions by assuming that the number of sequences in the peak center must be the maximum among the number of sequences in 200 consecutive mass values. The number 200 was determined empirically and worked well for non-specific and tryptic peptide distributions. In the second step, we computed the center of mass of a peak by starting from the preliminary peak center positions and including all peaks in the 1 Da interval around this peak:

$$M_c = \frac{\sum_{i=1}^n M_i n_i}{\sum_{i=1}^n n_i}$$

where, n_i is the number of peptide sequences with mass M_i . The newly computed center of masses served as peak center positions for subsequent computation of 95% PW values.

Mass Defects of Peak Centers and All Peptide Sequences

Following earlier work,¹³ we distinguished two mass defects for peptide species, mass defect (MD) and total mass defect (TMD). The MD of a peptide is defined as the difference between its monoisotopic mass (MI) and nominal mass (NM), where NM is equal to the floor of the monoisotopic mass:

$$MD = MI - NM = MI - [MI]$$

The total mass defect of a peptide is defined as the sum of mass defects of its constituent amino acid residues:

$$TMD = \sum_{i=1}^L MD(aa_i)$$

where L is the length of a peptide, aa_i is its i^{th} amino acid. Note that the term mass excess has alternatively been used in the literature.¹³ Here we use the more widely used term, “mass defect”.

All peptides with masses up to 3.5 kDa can be divided into three groups according to their total mass defect: group 1 (Gr I) contains all peptides with TMD less than 1, group 2 (Gr II) contains all peptides with TMD between 1 (inclusive) and 2 (exclusive), and group 3 (Gr III) contains all peptides with TMD between 2 (inclusive) and 3 (exclusive). There are no peptide sequences with TMD greater than 3 Da in the mass range of 3.5 kDa or less.

Periodic Patterns in Mass Defects

We observed periodicity in the distributions of MDs. While the periodicity of peptide mass distributions has been shown previously,^{5,17} studies were limited to peptide mass spaces below 1.1 kDa. Here we examined this effect for MDs of peptides up to 3.5 kDa. To identify periodic patterns in the MD distributions we computed the spectral power of the MDs, similar to the computations of power density for mass spectral profiles.¹⁸

Power spectrum is computed via a periodogram¹⁹ estimator. At a frequency f_k the corresponding value of the power spectrum is:

$$P(f_k) = \frac{1}{N^2} \left(|C_k|^2 + |C_{N-k}|^2 \right),$$

where C_k is a discrete Fourier transform of the MD at the frequency f_k , $k = 1, 2, \dots, (N/2 - 1)$, and N is the number of data points.

We examined power spectrum to determine the maximum power frequency. The maximum frequency corresponds to the most frequent peak spacing in the mass defect distribution. The

mass interval that corresponds to a specific frequency, f_k , is obtained by back-transformation into the mass domain:

$$dMass = N / (2 * f_c * k) = N * \Delta / k,$$

where f_c is the Nyquist critical frequency. Δ is the mass scan rate and N is the number of data points.

Results and Discussions

Mann⁸ modeled mass distributions of theoretical peptides by a linear dependence between the peak centers and their nominal masses for peptides up to 2 kDa. It was found that the masses of peak centers, M_C , and 95% PW values, W_C , have the following approximate relationships with respect to the nominal masses M_N of the peak centers:

$$\begin{aligned} M_C &= 1.00048 M_N \\ W_C &= 0.19 + 0.0001 M_N \end{aligned} \quad (1)$$

Equation (Eq.) 1 is well suited for use in filtering chemical and electronic noise from mass spectral data^{9,10}. Using Eq. 1, for every mass value, one can determine if the mass falls within 95% of a corresponding PW value. This equation has also been used for estimating accessible proteomic space for multidimensional separations coupled to mass spectrometry.¹³ The difference between the peak center mass and its nominal mass is its MD. Based on Eq. 1, the MD of peak centers is a linear function of their nominal mass with an intercept of 0, and a slope of 1.00048.

To extend Eq. 1 to higher mass values, we generated the mass distribution of all theoretical tryptic peptide sequences with masses up to 3.5 kDa. For each sequence, we first computed its monoisotopic mass and then its nominal mass and MD. The heat map in Figure 2 shows the base two logarithm of the number of peptide sequences for each pair of monoisotopic mass and MD. Violet and blue colors correspond to sparsely populated areas of monoisotopic masses and MDs. As Figure 2 illustrates, when the mass range of peptides is expanded beyond 2 kDa, MDs go through an inflection point - increasing up to 1 Da then returning to 0 and starting to increase again.

The effect described above is also seen in MD distributions of peak centers, depicted in Figure 3. Figure 3 shows the mass defects of peak centers as a function of nominal mass. Mass defect fluctuates for smaller peptides and passes through an inflection point at 2128 Da. This mass value differs significantly from the prediction⁹ of 2083 Da obtained from Eq. 1.

The original model, Eq. 1, seems to have assumed a zero intercept. We found that if the linear model is restricted to an intercept of 0, then we obtain a result which is very close to the coefficients in Eq. 1. Thus, for peptides with mass up to 2128 Da, we obtained a slope coefficient of 1.000479. However, we believe that the non-zero intercept model is more appropriate for peak center masses. Peptide masses are obtained by adding monoisotopic masses of amino acid residues and the monoisotopic masses of H₂O and proton. The intercept allows us to account for the nonzero MD of H₂O and proton. If we do not force the linear model to have a zero intercept, it provides intercept and slope values of 0.04299 and 1.00045, respectively. While both models (with zero and nonzero intercepts) have R^2 values of 1.0, the residual sum of squares is smaller for the model with the non-zero intercept. The results for peak centers are summarized in Eq. 2 below:

$$\begin{aligned} M_c &= 0.04298599 + 1.00044946 M_n, M_n \leq 2128; (a) \\ M_c &= -0.9554922 + 1.0004483 M_n, M_n > 2128; (b) \end{aligned} \quad (2)$$

The extended precision of the coefficients in Eq. 2 is dictated by the necessity of continuity between the two parts of the equation, obtained by separate fittings. This is shown in Figure 4, which depicts the residuals. Note that Figure 1 shows mass distributions in a mass interval encompassing 2128 Da. As seen from the figure, peak centers are in fact located on integer mass values in this interval.

The apparent discontinuity of MDs in Figures 2 and 3 is explained by changes in the relative contribution of components of total mass defect, and by the assumed maximum MD of 1 Da. To illustrate this, Figure 5 depicts normalized plots of different TMD groups defined in the Methods section. The switch between Eq.'s 2a and 2b is related to change in relative abundance levels of TMD's as peptide mass increases. For the mass range of up to 3.5 kDa considered here, peptides in Gr I constitute the majority of peptides in the range of small and medium masses, Eq. 2a; the TMD for these peptides is smaller than 1 Da. Peptides in Gr II constitute the majority of peptides with high mass values; they have TMD between 1 and 2 Da., and their MD is better described by Eq. 2b. Peptides in Gr III do not contribute significantly to peptide abundance for the mass range below 3.5 kDa. Therefore, we have not represented this group by an independent equation.

In Table 1 we list the amino acid compositions and masses of the lightest Gr II and Gr III peptides. This information is useful when setting the lower mass limits for computing Gr II and Gr III peptide sequences. Note that for a given peptide mass value it is not possible to predict its TMD value *a priori* (except for peptides smaller than 1.3 kDa). One has to consider all possible TMD's that a given mass would allow. Therefore, for an unknown species, both Eq. 2a and 2b should be considered if the mass range would allow more than one TMD. These equations were obtained from all peptides, without filtering peptides in any one the groups.

We next computed the 95% PW values for all theoretical tryptic peptides with masses up to 3.5 kDa. Figure 6 illustrates these results, as well as peak width computed using Eq. 1. As Figure 6 shows, empirical width distributions are, in general, not linear with nominal mass. For small peptides (< 1.4 kDa), PWs are highly oscillatory. This is explained by the relatively small numbers of peptide species and occupied mass bins in these mass ranges. Because of these factors, reaching the 95% PW value often encompasses the whole peak width. As the number of peptide sequences in the peaks increases, 95% PW value becomes linear with peptide mass. For large stretches of the mass axis, Eq. 1 (red line in Figure 6) overestimates the true PW. PW models are used for filtering non-peptides and their overestimation may reduce the efficiency of filtering. To more accurately describe 95% PW values, we have used two separate fitting functions, a quadratic function for nominal peaks up to 1.4 kDa and a linear function for peaks larger than 1.4 kDa. Our model is shown below:

$$\begin{aligned} W_c &= -0.2733 + 0.001123 M_n - 0.0000005369 M_n^2, M_n \leq 1400 \\ W_c &= 0.15 + 0.00007172 M_n, M_n > 1400 \end{aligned} \quad (3)$$

Profiles of PW values computed using Eq. 3 are shown with blue lines in Figure 6. Note that a quadratic model was chosen as the simplest non-linear model to describe non-linear behavior. This model performed well, where we observed that 95.1 % of all peptides up to 3.5 kDa in mass were encompassed by the PW values modeled by Eq. 3, when peak centers

were determined using Eq. 2. Also shown in Figure 6, as a red line, is the PW values determined using Eq. 1. This model seems to linearly estimate the boundaries of PW distributions at low masses; the region of non-linear distributions. However, as seen in Figure 6, 95% PW values are largely overestimated for a long stretch of mass values. At 2 kDa, the width of 95% PW from our model is 25% less than that computed using Eq. 1. This result compares well with the width reduction (31.6 %) of a previous model that only considered proteins of human serum and seminal fluids¹⁰ from a protein sequence database. The results from the present model are more consistent, as we observe a 25 % reduction of 95% PW value at 3 kDa mass. For the model accounting only for human proteins¹⁰ the relative 95% PW reduction decreases to 25.6%. Since our model was developed for all theoretical peptides, we expect that it should generalize better to proteomes of different biological species.

The residual of the MD in Figure 4 shows a strong oscillatory behavior. We note that the amplitude of the oscillations are damped for larger mass values. A power spectral analysis reveals the oscillation frequency, Figure 7. There is a clear maximum in the power spectrum at the frequency, f_k , where $k = 220$. In the mass domain, this frequency corresponds to $N*\Delta/k \approx 3096/(1.0*220) \approx 14$ Da. 3096 is the number of the peak centers in the distribution of theoretical tryptic peptides with masses not greater than 3.5 kDa. Δ , distance between peak centers, is about 1 Da, Figure 1. The mass interval (14 Da) that corresponds to the maximum power frequency, is the most frequent mass difference (this value corresponds to the nominal mass of the CH_2 group) between the masses of the twenty natural amino acids used to generate the theoretical peptides.

The peptide sequences used in modeling Eqs 2 and 3 are those obtained from the twenty naturally occurring amino acids. In general, the models can be affected by amino acid modifications. We have explored several common modifications - oxidation of Met, phosphorylations of Ser, Thr and Tyr. These modifications change peptide mass distributions slightly. Figures S1 and S2 of the Supplementary Materials compare the distributions of mass defects and peak widths with and without the inclusion of oxidized Met. The figures illustrate that MDs are slightly reduced and 95% PW values are increased when oxidized Met is included in generating theoretical peptides. Figure S3 shows the quantification of changes in peak center locations using the corresponding PW values. For peptides larger than 650 Da, changes in the peak center positions are less than 10% of the corresponding PW values. For smaller peptides (<650 Da), these changes can be larger compared to the corresponding PWs. This behavior is due to smaller absolute values of PWs in this region, Figure 6. These observations are explained by the nature of the building blocks of theoretical peptides - twenty natural amino acids and modified amino acids. From a computational point of view, each modification is an addition of a new element into the set of building blocks. The MD value of oxidized Met, 0.0354 Da, is smaller than the average MD of the twenty natural amino acids, 0.05 Da. Therefore, the addition of oxidized Met into the set generally reduces the MDs of theoretical peptides. The PW values of theoretical peak clusters will increase as modified amino acids are added to the set. This behavior simply reflects an increase in variance due to increased uncertainty in system.

Quantification of relative changes in the peak center locations of distributions including phosphorylated Ser, Thr and Tyr and shown in Figure S4. The relative changes are again dependent on nominal mass and are somewhat higher for this set ~ 15 % compared to the corresponding PW values for peptide distributions composed of the twenty natural amino acids. This observation is due to the larger number of modified amino acids added into the original set of building blocks (twenty naturally occurring amino acids).

In our next work we will apply MD and PW information in combination with information about the forbidden zones to assess the effect of MD filtering on peptide identifications in shotgun proteomics. In most of the modern mass spectrometers employed for high-throughput peptide identifications, full mass scans of precursors are acquired with high mass resolution and accuracy. Selected precursor ions are fragmented and precursor mass and masses of fragment ions are used to identify the peptide. It has been estimated from experimental data (by the comparison of labeled and unlabeled species) that 30% of all eluting species are chemical noise.²⁰ High mass resolution and accuracy potentially allows us to separate chemical noise from peptides based on accurate mass information. Note that highly charged species can be reduced to +1 charge, and only monoisotopic peaks are retained. Possible neutral losses of water and/or ammonia do not change the MDs of parent ions significantly. Therefore, peak centers and widths obtained for intact peptides are valid for peptides with neutral losses.

Conclusions

We have modeled MDs, TMDs, locations of peak centers and 95% PW values in the mass distributions of all theoretical peptides composed of twenty natural amino acids with masses less than 3.5 kDa. Our algorithm can take into account different enzymatic and digestion conditions, and restrictions of amino acids to generate mass distributions. We have shown how nonlinear effects dominate the distributions of PW values for lower masses and have modeled this relationship, giving rise to new formulas for obtaining 95% PW values using a combination of linear and quadratic models. We found that mass intervals enclosed by 95% PW values are smaller than previously thought. The smaller 95% PW values will improve noise filtering in high mass accuracy spectral data. In modeling the peak centers we have fine-tuned MDs accurately and shown that they are explained by three groups of TMDs that are observed for peptides under 3.5 kDa. Mass defects exhibit oscillatory behavior. Spectral power analysis of the MDs revealed that the frequency of the oscillations is 14 Da. This mass value corresponds to the most frequent mass difference between the masses of the twenty natural amino acids. Addition of modified amino acids, oxidized Met, phosphorylated Ser, Thr and Tyr, slightly changes peak center positions and widths.

We believe that this methodology has a wider scope and can be a powerful supplement to the recent upsurge of studies of mass defect labeling. Of particular interest is looking into unique amino acid signatures of MDs which is an important marker of Bromine/Iodine based mass defect labeling²¹ of Cys-rich residues²² and phosphopeptides.²³

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported, in part, by UL1RR029876 UTMB CTSA (ARB), HHSN272200800048C NIAID Clinical Proteomics Center (ARB) and NIH-NLBIHHSN268201000037C NHLBI Proteomics Center for Airway Inflammation (Alex Kurosky, UTMB).

Abbreviations

Da	Dalton
MD	Mass Defect
TMD	Total Mass Defect

PW	peak width
MI	Monoisotopic Mass
NM	Nominal Mass
MS	Mass Spectrometry
ppm	parts per million
Eq	equation
Gr	group

References

1. Mann M, Kelleher NL. Special Feature: Precision proteomics: The case for high resolution and high mass accuracy. *Proc Natl Acad Sci USA*. 2008
2. Domon B, Aebersold R. Mass spectrometry and protein analysis. *Science*. 2006; 312:212–17. [PubMed: 16614208]
3. Andrews GL, Simons BL, Young JB, Hawkrige AM, Muddiman DC. Performance characteristics of a new hybrid quadrupole time-of-flight tandem mass spectrometer (TripleTOF 5600). *Anal Chem*. 2011; 83:5442–46. [PubMed: 21619048]
4. Makarov A, Denisov E, Lange O, Horning S. Dynamic range of mass accuracy in LTQ Orbitrap hybrid mass spectrometer. *J Am Soc Mass Spectrom*. 2006; 17:977–82. [PubMed: 16750636]
5. Yu L, Xiong YM, Polfer NC. Periodicity of monoisotopic mass isomers and isobars in proteomics. *Anal Chem*. 2011; 83:8019–23. [PubMed: 21932815]
6. 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]
7. Mao Y, Tipton JD, Blakney GT, Hendrickson CL, Marshall AG. Valence Parity to Distinguish c' and z(*) Ions from Electron Capture Dissociation/Electron Transfer Dissociation of Peptides: Effects of Isomers, Isobars, and Proteolysis Specificity. *Anal Chem*. 2011; 83:8024–28. [PubMed: 21932844]
8. Mann, M. Useful Tables of Possible and Probable Peptide Masses. Annual Conference on Mass Spectrometry and Allied Topics; 5-1-1995; Atlanta, GA: American Society of Mass Spectrometry; Ref Type: Conference Proceeding
9. 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; 5:1195–203. [PubMed: 16674109]
10. Toumi ML, Desaire H. Improving mass defect filters for human proteins. *J Proteome Res*. 2010; 9:5492–95. [PubMed: 20731397]
11. Lehmann WD, Bohne A, von Der Lieth CW. The information encrypted in accurate peptide masses-improved protein identification and assistance in glycopeptide identification and characterization. *J Mass Spectrom*. 2000; 35:1335–41. [PubMed: 11114093]
12. Yao X, Diego P, Ramos AA, Shi Y. Averagine-scaling analysis and fragment ion mass defect labeling in peptide mass spectrometry. *Anal Chem*. 2008; 80:7383–91. [PubMed: 18778085]
13. Frahm JL, Howard BE, Heber S, Muddiman DC. Accessible proteomics space and its implications for peak capacity for zero-, one- and two-dimensional separations coupled with FT-ICR and TOF mass spectrometry. *J Mass Spectrom*. 2006; 41:281–88. [PubMed: 16538648]
14. Hubler SL, Jue A, Keith J, McAlister GC, Craciun G, Coon JJ. Valence parity renders z(*)-type ions chemically distinct. *J Am Chem Soc*. 2008; 130:6388–94. [PubMed: 18444621]
15. Nefedov AV, Mitra I, Brasier AR, Sadygov RG. Examining troughs in the mass distribution of all theoretically possible tryptic peptides. *J Proteome Res*. 2011; 10:4150–57. [PubMed: 21780838]
16. Nefedov AV, Sadygov RG. A parallel method for enumerating amino acid compositions and masses of all theoretical peptides. *BMC Bioinformatics*. 2011; 12:432. [PubMed: 22059886]

17. Demirev PA, Zubarev RA. Probing combinatorial library diversity by mass spectrometry. *Analytical Chemistry*. 1997; 69:2893–900. [PubMed: 21639310]
18. Sadygov RG, Zhao Y, Haidacher SJ, Starkey JM, Tilton RG, Denner L. Using power spectrum analysis to evaluate (18)O-water labeling data acquired from low resolution mass spectrometers. *J Proteome Res*. 2010; 9:4306–12. [PubMed: 20568695]
19. Press, WH.; Teukolsky, SA.; Vetterling, WT.; Flannery, BP. The art of scientific computing. 3. Cambridge University Press; Cambridge, UK: 2007. Numerical recipes.
20. Michalski A, Cox J, Mann M. More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority is Inaccessible to Data-Dependent LC-MS/MS. *J Proteome Res*. 2011; 10:1785–93. [PubMed: 21309581]
21. Hall MP, Ashrafi S, Obegi I, Petesch R, Peterson JN, Schneider LV. “Mass defect” tags for biomolecular mass spectrometry. *J Mass Spectrom*. 2003; 38:809–16. [PubMed: 12938101]
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; 78:3417–23. [PubMed: 16689545]
23. Bruce C, Shifman MA, Miller P, Gulcicek EE. Probabilistic enrichment of phosphopeptides by their mass defect. *Anal Chem*. 2006; 78:4374–82. [PubMed: 16808444]

FIGURE 1.

Peak dynamics of theoretical tryptic peptides for selected mass intervals. As peptides become larger, the peak centers “move” with respect to the integer masses. Peak center of theoretical tryptic peptides “crosses” the integer mass at 2128 Da (see the main text). Forbidden zones are seen in the distributions of peptides up to 1.5 kDa mass.

FIGURE 2.

MD distributions for all theoretical tryptic peptides with masses up to 3.5 kDa. The color bar is in base 2 log scale for the number of peptide sequences for each pair of monoisotopic mass and MD, with violet and blue indicating low-populated areas and red indicating the most populated areas.

FIGURE 3.

Mass defect as a function of nominal mass for peak centers of theoretical tryptic peptides. The MD grows (with oscillations) for peptides in the mass range up to 2128 Da. At 2128 Da, MD passes through an inflection point, i.e., its value drops to 0. After this abrupt change MD increases again. The empirical distribution of MD necessitates two separate fits: one for mass values up to 2128 Da and one for mass values between 2128 Da and 3.5 kDa.

FIGURE 4.

Distribution of residuals of the MDs of the peak centers obtained by using Eq. 2. This distribution exhibits an oscillatory behavior and for large mass values the amplitude of the oscillations is damped.

FIGURE 5.

Normalized plots of the total mass defects for Gr I (black solid line), Gr II (black broken line) and Gr III (red broken line). MD modeling and the necessity of two separate fittings stems from the relative abundances of Gr I and Gr II peptides. For Gr I peptides TMD is 1 Da and Eq. 2 (a) is an adequate description. In the mass range where Gr II peptides constitute the majority, Eq. 2 (b) is a better description. Note that even for higher masses there still will be peptides for which Eq. 2 (a) describes the correct MD.

FIGURE 6.

95% PWs as a function of peptide nominal masses for peak centers (black lines). Also shown are predictions from the model distributions of Eq. 3 (blue line) and Eq. 1 (red line). The model using Eq. 1 is more conservative in estimating PWs. PWs can be substantially reduced by using a combination of nonlinear (small mass values) and linear (larger mass values) models.

FIGURE 7.

Spectral power of MD residual values from Figure 4. A prominent frequency, f_k $k = 220$, is observed. When converted back to mass domain, this frequency corresponds to a mass value of 14 Da (see the main text).

TABLE 1

Masses and compositions of the lightest Gr II and Gr III peptides. Note that degeneracy with respect to the I/L substitutions is not accounted for in the table. However, our algorithm itself generates different configurations for I- and L- containing sequences. Degenerate compositions are removed from the table to save space.

Mass	Compositions
1363.006 (Gr II)	V ₂ L ₉ K
2720.015(Gr III)	C ₈ L ₄ M ₄ H ₃ F ₁ W ₁ R; T ₁ C ₃ L ₁ M ₉ H ₄ F ₂ R V ₄ C ₆ M ₆ H ₃ F ₁ W ₁ R; S ₁ P ₁ C ₃ M ₇ F ₃ Y ₁ W ₃ K P ₄ V ₁ C ₄ M ₈ H ₃ W ₁ R