Published in final edited form as:

J Proteome Res. 2011 September 2; 10(9): 4150–4157. doi:10.1021/pr2003177.

Examining Troughs in the Mass Distribution of All Theoretically Possible Tryptic Peptides

Alexey V. Nefedov^{1,2}, Indranil Mitra^{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, Texas 77555

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

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

Abstract

This work describes the mass distribution of all theoretically possibly tryptic peptides made of 20 amino acids, up to the mass of 3 kDa, with resolution of 0.001 Da. We characterize regions between the peaks of the distribution, including gaps (forbidden zones) and low-populated areas (quiet zones). We show how the gaps shrink over the mass range, and when they completely disappear. We demonstrate that peptide compositions in quiet zones are less diverse than those in the peaks of the distribution, and that by eliminating certain types of unrealistic compositions the gaps in the distribution may be increased. The mass distribution is generated using a parallel implementation of a recursive procedure that enumerates all amino acid compositions. It allows us to enumerate all compositions of tryptic peptides below 3 kDa in 48 minutes using a computer cluster with 12 Intel Xeon X5650 CPUs (72 cores). The results of this work can be used to facilitate protein identification and mass defect labeling in mass spectrometry-based proteomics experiments.

Keywords

distribution of peptide masses; forbidden zones; quiet zones; amino acid compositions of all theoretically possible peptides; accurate peptide masses; mass accuracy

Introduction

Peptides are made of twenty amino acids with specific masses, which leads to an inhomogeneous, clustered distribution of their masses. This can be observed both for all theoretically possible peptides, and for real protein sequence databases. In the distribution, peptides form a series of peaks separated by approximately 1 Da, which grow taller and wider as the mass increases. Consecutive peaks are separated by low populated areas (quiet zones) and gaps (forbidden zones) – that is, the mass ranges for which there exist no possible sequences of amino acids.

*To whom correspondence should be addressed: Tel. 409-772-3287; Fax 409-772-9670; rovshan.sadygov@utmb.edu.

Supporting Information Available: Supplementary materials include CSV files with the mass distributions of tryptic peptides and all peptides (no enzymatic specificity), files listing gaps in these distributions, Microsoft Excel file with statistics on AA occurrences in tryptic peptides from IPI human proteome database, a table with masses of amino acids used in this work, and Figures S1–S4. This material is available free of charge via the Internet at <http://pubs.acs.org>.

These features of the distribution of peptide masses play important role in the mass spectrometry-based proteomics. Forbidden zones are used to filter out non-peptide masses in a variety of experiments and workflows, including peptide mass fingerprinting¹ and mass defect filtering². They also find application in recently proposed mass defect labeling^{3–5}. A very promising research is being done in de novo peptide sequencing based on highly accurate mass measurements and the discrete nature of the distribution of peptide masses^{6–9}.

A recent work by Mickalski¹⁰ and colleagues has suggested a practical way of separating peptide species from co-eluting contaminants. It was observed that in an equal mixture of completely labeled and label-free SILAC¹¹ samples all peptide species would have isotopically labeled counterparts. Absence of a labeled counterpart indicates that the species is non-peptidic. It was observed that up to 30% of all eluting species were of non-peptide origin. Ability to filter out these species, or even better not to choose them for fragmentation in the first place, will significantly improve the efficiency of experiments.

Thus, an increasing number of proteomics workflows face the need for efficient generation and characterization of mass distributions of different classes of peptides. Often, it is also desirable to permanently store these distributions (along with the atomic, amino acid, or other compositions of the molecules) and use them as look up tables. Since the number of all sequences comprised of N letters and having length L is equal to N^L , the main challenge here is a high computational cost associated with generation of these distributions and huge amounts of data to be stored.

One of the pioneering works on generation and description of the mass distribution of all theoretically possible peptides was done by Mann in 1995¹². He considered the mass range up to 2 kDa and suggested linear equations linking nominal peptide masses with the position and width of the peaks formed by monoisotopic peptide masses. He described a range of potential applications of his results, including the possibility of the use of suggested equations to identify non-peptide masses. It was noted that the computational time needed to generate a 50 Da wide range of the distribution was less than one second for masses below 500 Da and 18 hours for masses around 2 kDa.

Since then, a number of papers were published in which theoretical or observed distributions of peptide masses were generated^{1,8,13–17} for different proteomics applications. These distributions usually covered the mass range up to 2 kDa, which did not allow full characterization of the forbidden zones. Also, their generation often required long computational times or/and extensive computational capabilities. In many cases, the distributions were not made easily available for the research community for independent validation, study and use.

In this paper, we describe the mass distribution of all theoretically possibly tryptic peptides made of 20 amino acids, up to the mass of 3 kDa, with resolution of 0.001 Da. We characterize regions between the peaks of the distribution, including gaps (forbidden zones) and low-populated areas (quiet zones). In particular, we describe how the gaps shrink over the mass range and when they completely disappear. We also show that peptide compositions in quiet zones are less diverse than those in the peaks of the distribution, and that by eliminating certain types of unrealistic (i.e., not likely to be observed) compositions the gaps in the distribution may be increased.

The mass distribution is generated using a parallel implementation of a recursive procedure that enumerates all amino acid (AA) compositions. Suggested parallel implementation of the procedure yields significant reduction in the computation time. For instance, the distribution described in this paper was generated in 48 minutes on a computer cluster with 12 Intel Xeon X5650 CPUs (72 cores). We believe that, given the fact that parallel computing

architectures become increasingly available, this result opens new perspectives on the use of brute-force enumeration of amino acids compositions in proteomics studies. In particular, if peptide mass distributions can be generated quickly, then they can be generated “to order” to facilitate protein identification in particular experiments or experimental setups, taking into account specific proteases used in sample preparation, possible post-translational or chemical modifications, and other factors.

Methods

A peptide may be considered as a sequence of letters from a 20-letter alphabet $A = \{a_1, a_2, \dots, a_{20}\}$, whose letters correspond to 20 amino acids. This sequence can be coded by a numerical vector $(n_1, n_2, \dots, n_{20})$, whose i -th component is the number of occurrences of the i -th letter (amino acid) in the sequence, $i = 1, 2, \dots, 20$. This vector is called a composition (also amino acid composition or peptide composition). For example, for sequence $a_1 a_{20} a_1 a_1$ we get composition $(3, 0, \dots, 0, 1)$. We generate all theoretically possible peptide masses by using amino acid compositions and a recursive procedure for enumerating these compositions, which is executed in parallel.

It is worthwhile to note some basic properties of compositions. For the sake of generality, consider alphabet A of N letters $a_1, a_2, \dots, a_N$, and composition $(n_1, n_2, \dots, n_N)$. The length of the composition is defined as $L = n_1 + n_2 + \dots + n_N$. For a single sequence of letters we have a uniquely defined composition, while for a single composition of length L we have

$$\frac{L!}{n_1! n_2! \dots n_N!} \quad (1)$$

corresponding sequences, since the composition codes only the number of letters used in the sequence but not their order. The number of compositions of length L that use N letters is equal to the number of ways to choose L elements from a set of N elements if repetitions are allowed:

$$\binom{N+L-1}{L} = \frac{(N+L-1)!}{L!(N-1)!}, \quad (2)$$

while the number of corresponding sequences is equal to N^L . Table 1 shows the number of compositions and sequences for all peptides and tryptic peptides up to the length 10. Note how the number of sequences quickly overgrows the number of compositions as L increases.

We use compositions as a convenient means to build and study the mass distribution of all theoretically possible peptides. Specifically, we consider tryptic peptides with no missed cleavages, up to the mass of 3 kDa, but any other class of peptides or mass limit can be considered by following the same methodology. A useful property of compositions is that one composition codes multiple sequences (Eq. (1)) with equal masses. Thus, in order to enumerate all possible peptide masses, it is sufficient to enumerate all amino acid compositions instead of all amino acid sequences.

Still, enumeration of all compositions presents a significant computational challenge given almost exponential growth of their number with length L (Table 1). To address this challenge, we have developed a parallel program utilizing a recursive procedure for composition enumeration. Figure 1 shows pseudocode of the procedure Generate which generates all compositions of length not greater than L for an alphabet of N letters and prints their masses. The procedure should be called with parameters $(L, 1)$. Array c holds current

composition $(n_1, n_2, \dots, n_N)$ and is indexed from one. Procedure Mass returns the mass of the input composition.

Procedure Generate starts with composition $(0, 0, \dots, 0)$ and enumerates all compositions with n_N ranging from 0 to L . It then assigns n_{N-1} to 1, and enumerates all compositions with n_N ranging from 0 to $L-1$, and so on. The last composition in this enumeration process is $(L, 0, \dots, 0)$. For example, if $N = 3$ and $L = 2$ then the procedure will enumerate all compositions of length between 0 and 2 in the following way: $(0, 0, 0)$, $(0, 0, 1)$, $(0, 0, 2)$, $(0, 1, 0)$, $(0, 1, 1)$, $(0, 2, 0)$, $(1, 0, 0)$, $(1, 0, 1)$, $(1, 1, 0)$, $(2, 0, 0)$.

Procedure Generate can be made significantly faster by using an upper limit on the mass of compositions. If at some point the mass of the defined components of the current composition exceeds the limit, the further construction of this composition can be canceled. For example, enumeration of all tryptic compositions up to the length 30 takes about 37 hours without an upper mass limit, and about 7 hours with the upper mass limit of 3 kDa (one core, Intel Xeon X5650 CPU).

Execution of procedure Generate in non-parallel fashion is feasible for relatively short lengths L . For example, it takes about 9 minutes to generate the mass distribution of all tryptic peptides up to the length 20 and mass 3 kDa, but this time rises to about 8 hours for all tryptic peptides up to the length 40 (one core, Intel Xeon X5650 CPU). Parallel execution of procedure Generate is based on the fact that a single call to this procedure with parameters $(L, 1)$ can be expanded into $L+1$ independent calls with $start = 2$: $(L, 2)$, $(L-1, 2)$, $\dots$, $(0, 2)$, while n_1 is set to 0, 1, $\dots$, L , correspondingly. Each of these calls, which we call jobs, can be expanded into a set of jobs with $start = 3$, and so on. Note that jobs $(L, start + 1)$ and $(L-1, start)$ require less computations than job $(L, start)$.

To generate the mass distribution of all tryptic peptides up to the length 51, we split the primary problem into 229,426 jobs. Table 2 shows the maximum values of parameter L for each value of $start$ that were used in this process. First, the primary job $(51, 1)$ was split into 52 jobs $(51, 2)$, $(50, 2)$, $\dots$, $(0, 2)$. Then, every job with $start = 2$ and $L > 20$ was split into a set of jobs with $start = 3$. Next, every job with $start = 3$ and $L > 24$ was split into a set of jobs with $start = 4$, and so on; jobs with $start = 6$ were not split into smaller jobs (with $start \geq 7$). Parameters listed in Table 2 directly affect computational time and their choice is dependent upon the available computational resources (the number and speed of the CPU cores).

The general scheme of the computations can be described in the following way. A master process creates a list of jobs for a given primary problem described by parameters $(L, 1)$. It then iterates through the list of jobs and available worker processes and assigns jobs to free workers. A worker generates the mass histogram for the given parameters $(L, start)$ and sends it back to the master. The master adds histograms obtained from the workers to the final mass histogram corresponding to the primary problem. The master and workers can exchange data by using the Message Passing Interface (MPI)¹⁸.

Our program for generation of the mass distribution of all tryptic peptides up to the given length (with the maximum mass of 3 kDa) is freely available for download at <https://ispace.utmb.edu/users/rgsadygo/Proteomics/AACompositions/>. It was implemented using Microsoft Visual C++ and MPI library from Microsoft HPC SDK Pack.

For the mass distribution described in this paper, we enumerate compositions that correspond to all theoretically possible tryptic peptides without missed cleavages (that is, peptides ending with Lys or Arg, without other Lys and Arg), of length between 2 and 51

amino acids ($2 \leq L \leq 51$), and having the monoisotopic mass no greater than 3 kDa. Note that no tryptic peptides longer than 51 amino acids have the mass of 3 kDa or smaller.

The monoisotopic mass of a (protonated) peptide is calculated as the sum of the monoisotopic masses of its constituent amino acid residues, plus monoisotopic masses of H_2O and a proton. We use amino acid masses defined up to the 8th digit after the decimal point, and then round the peptide mass to three digits after the decimal points (thus, the mass distribution has accuracy of 0.001 Da). Monoisotopic masses of amino acids that have been used in this work are listed in the supplemental materials.

Results

Mass Distribution of Peptide Compositions

Figure 2 shows a logarithmic plot of the number of tryptic peptide compositions versus their mass (this distribution, for $2 \leq L \leq 51$, is available in the supplemental materials as CSV file tryptic_compositions.txt). Note the almost exponential growth of the number of compositions as the mass increases. Overall, the distribution has about 1.3×10^{12} compositions; the lightest peptide, GK, has the mass of 232.140 Da, and the heaviest peptides of 3,000.000 Da are described by 333,571 compositions. The maximum number of compositions, 27,381,800, corresponds to the mass of 2999.399 Da.

Though the mass distribution of compositions for $2 \leq L \leq 51$ looks solid on the scale of Figure 2, it actually consists of separate peaks, low populated zones, and gaps (Figure 3). As the mass increases, these peaks become wider and taller, and their centers move towards the next integer mass value. Mann¹² noticed that for all (not only tryptic) peptides below 2 kDa, the centers of the peaks M_c and their widths W_c encompassing about 95% of all peptides have the following approximate relationship with the nominal masses M_n of peptides:

$$M_c = M_n + 0.00048 M_n \quad (3)$$

$$W_c = 0.19 + 0.0001 M_n \quad (4)$$

We have determined that for the mass range below 2 kDa, 95% of the tryptic peptides lie within the regions described by equations (3) and (4), while for the mass range between 2 and 3 kDa this figure decreases to 90%. For the entire mass range up to 3 kDa this figure is 90% (note that the number of compositions between 2 and 3 kDa is approximately 255 times higher than the number of compositions below 2 kDa).

To characterize the symmetry of the mass peaks, we calculated the skewness coefficient¹⁹ according to the following equation:

$$g_1 = \frac{\frac{1}{n} \sum_{i=1}^n (m_i - \bar{m})^3}{\left(\frac{1}{n} \sum_{i=1}^n (m_i - \bar{m})^2 \right)^{3/2}},$$

where n is the total number of compositions in a given peak, m_i is the mass of the i -th composition, and $\bar{m}$ is the average mass of all compositions in the peak. Coefficient g_1 is equal to zero for distributions with symmetrical shape, is positive for distributions with longer right tail, and negative for distributions with longer left tail.

For the peaks in the mass distribution of tryptic peptide compositions, coefficient g_1 oscillates around 0.08 with damping amplitude (Figure S1). It starts to oscillate between -1.25 to 1.25 for peaks located in the beginning of the mass range, becomes positive for all peaks located above approximately 1500 Da, and oscillates with small constant amplitude after 2000 Da. This means that with increasing mass the shape of the peaks stabilizes and becomes more symmetrical, with slightly longer right tail.

Forbidden Zones

From the practical perspective, an important feature of the distribution of peptide masses is the presence of gaps, which means that for some mass values there exist no amino acid sequences that have these masses. Extended continuous gaps in the distribution of masses of amino acid sequences are called forbidden zones. They can be used to filter out non-peptide masses in a variety of experiments and workflows, including peptide mass fingerprinting¹ and mass defect filtering².

Consider a series of consecutive masses M_1, M_2, M_3, M_4, M_5 and assume that we have $K_1, 0, 0, 0, K_5$ compositions corresponding to these masses, where $K_1 > 0$ and $K_5 > 0$. Then the gap between M_1 and M_5 can be described by a pair $(M_2, 3)$ – that is, the mass, at which the gap starts, and the width of the gap, measured in the units of the mass scale ($M_i - M_{i-1}$). If the mass scale unit is 0.001 Da, then the gap's width can be converted into Da by multiplying 3 by 0.001 Da.

Figure S2 in the supplemental materials shows gaps' widths for three different mass ranges of our distribution (the full table of gaps is available in the supplemental materials as CSV file `tryptic_compositions_gaps.txt`). In the range 230–400 Da we have gaps as wide as 10 Da, due to the boundary effect of the distribution. In the range 400–2,500 Da gap's width varies between 0.001 and 1 Da, and the amplitude of its variation slowly decreases as the mass increases, so that after 1,160 Da the maximum gap is about 0.5 Da; after 1,530 Da it is about 0.25 Da; and after 1752 Da it is about 0.12 Da. The last gap in the distribution is located at mass 2501.517 Da and has width 0.001 Da; after that, the distribution of masses of all theoretically possible tryptic peptides with no missed cleavages (measured at 0.001 Da resolution) has no gaps. Note that the total width of all gaps in the distribution between 232 and 2,502 Da is about 1,203 Da, or 53% of this mass range.

Figure 4 gives another illustration of how the gaps are closing towards the mass of 2502 Da. Points of dark purple show the total width of all gaps between consecutive peaks, while points of light purple – the width of the maximum continuous gaps between consecutive peaks. Note that the slope of the curves is fairly constant for the most part of the mass range.

Quiet Zones and Corresponding Peptide Compositions

As Figure 3 shows, two consecutive peaks in the mass distribution are separated by gaps and low populated areas, also called quiet zones. We found that many compositions falling into the quiet zones correspond to sequences with long repetitions of a single amino acid. For example, all compositions in the range 2000.436–2000.445 Da have between 10 and 14 Cys residues, as shown in Table 3. This particular observation can be explained by the fact that Cys residue has the smallest ratio of the mass defect (the difference between the exact mass and the nominal mass, also referred to as the mass excess¹⁷) to the monoisotopic mass among all 20 amino acids². Therefore, the right tail of the mass peak located at 1999.96 Da is expected to be populated by the compositions enriched with Cys residues.

To make the mass distribution more realistic, we introduced the following restriction on peptide compositions: every composition $(n_1, n_2, \dots, n_N)$ must have $n_i \leq 8$, $i = 1, 2, \dots, 20$, and $N_L + n_I \leq 8$, where N_L is the number of leucines and n_I is the number of isoleucines.

These peptides are rare, which has been confirmed by our examination of the IPI human proteome database²⁰, where only about 1.5% of all tryptic peptides below 3 kDa have single amino acid occurring more than 8 times. Prohibited compositions (about 18% of all compositions) were removed from the original distribution to form the restricted distribution.

The top graph in Figure 5 shows a fragment of the non-restricted mass distribution where red peaks correspond to compositions that violate the introduced constraints. We can see that the proportion of prohibited compositions is larger in the quiet zones than in the peaks. On the bottom graph in Figure 5, red peaks correspond to uniformly chosen 18% of compositions. We can see that the proportion of uniformly chosen compositions is higher in peaks and lower in quiet zones than the proportion of prohibited compositions.

The distribution of gaps' widths for restricted compositions is very similar to that for non-restricted composition, with one anticipated difference: for restricted compositions, the last gap (of 0.001 Da) in the distribution is located at mass 2,842.723 Da, which is about 341 Da further than the last gap in the unrestricted distribution. The total width of all gaps in the restricted distribution between 232 and 2,843 Da is about 1,272 Da, or 49% of this mass range.

Figure 4 illustrates how the width of the gaps in the restricted distribution changes as the mass increases. Points of dark orange show the total width, and points of light orange – the maximum width of all gaps between consecutive peaks. Note that for the end of the distribution the orange curves have more gentle slopes than purple curves, and they approach zero at larger mass values.

In another experiment, we produced the third mass distribution with restrictions on peptide compositions derived from IPI human protein database (HPD). We made *in silico* trypsin digestion of all proteins in HPD and calculated length-specific average μ_{ij} and standard deviation σ_{ij} of the number of times each AA occurs in tryptic peptides, $i = 2, 3, \dots, 51$ (length index), $j = 1, 2, \dots, 18$ (amino acid index, K and R are excluded). We then calculated the maximum allowed number of AA j in a composition of length i as $\mu_{ij} + 2\sigma_{ij}$. For example, the average number of glycines in tryptic peptides of length 4 was 0.22, and the standard deviation of this number was 0.46 (see file "AA_statistics.xls" in the supplemental materials). Therefore, the maximum allowed number of glycines for compositions of length 4 was set to $\text{ceiling}(0.22 + 2 \cdot 0.46) = 2$. Prohibited compositions (about 45% of all compositions) were removed from the original distribution to form the restricted distribution. The effects of restricting number of an amino acid per a composition based on amino acid occurrences in HPD and absolute maximum are similar, Figures S3 and S4. The forbidden zones are extended up to 2200 Da. The all gaps distribution extends somewhat farther with the AA proportions from HPD up to 3 kDa. The details of the distributions are described in the supplemental materials (Figures S3, S4).

Quiet zones and entropy

To further characterize the diversity of peptide compositions making up peaks and quiet zones of the mass distribution, we calculated the average entropy of peptide compositions having the same mass. For a composition $(n_1, n_2, \dots, n_N)$ of length L its entropy is defined as

$$H = - \sum_{n_i > 0} \frac{n_i}{L} \ln\left(\frac{n_i}{L}\right).$$

As it follows from this equation, for sequences comprised of only one letter, entropy of the corresponding compositions is zero (lowest diversity). On the other hand, for sequences comprised of L different letters, entropy of the corresponding compositions is $\ln(L)$ (highest diversity).

Figure 6 shows the average entropy of compositions and the corresponding margin of three standard deviations for a short range of mass distribution. The figure demonstrates that the average entropy of compositions is higher in peaks and lower in quiet zones, which means that peaks are made of more diverse compositions than quiet zones. The latter agrees with the observation made above about the distribution of prohibited compositions.

Conclusions and Future Work

We described the mass distribution of all theoretically possible tryptic peptides up to the mass of 3 kDa. Specific focus of this study was on characterization of forbidden and quiet zones of the distribution. Mass defect filtering and mass defect labeling require the exact knowledge of the location and width of forbidden zones in the mass distribution of peptides. We gave detailed description of these zones until they disappear at about 2.5 kDa (for the non-restricted distribution). An interesting observation is that the total width of all gaps in the distribution constitutes 53% of the mass range between 232 and 2,502 Da.

Analyzing peptide compositions from the quiet zones we found that many of them correspond to sequences with long repetitions of a single amino acid. This raised the question about realistic (i.e., those that may occur in nature) and nonrealistic peptide sequences, and the probability of specific amino acid patterns occurring. As a first attempt to exclude nonrealistic peptide compositions and generate a more plausible distribution of peptide masses, we excluded compositions where each amino acid occurred more than 8 times, or where the cumulative number of leucines and isoleucines was more than 8. This rule was justified by examining tryptic peptides obtained from the IPI human protein database. It may be refined by making it dependent on the amino acid, as well as on the length and mass of the peptide.

We showed that the proportion of prohibited compositions (i.e., containing long repetitions of a single amino acid) is larger in the quiet zones than in the peaks. This was confirmed by the analysis based on entropy of compositions: we found that entropy (diversity) of the compositions making up peaks of the distribution is higher than entropy (diversity) of compositions making up quiet zones.

We also described the method used to generate this distribution. Our method gives substantial reduction in the computation time, allowing us to enumerate all peptide compositions below 3 kDa in 48 minutes on a computer cluster with 12 Intel Xeon X5650 CPUs (72 cores). Fast computation times give an opportunity to routinely generate and use look up tables of all theoretically possible peptide masses in various proteomics experiments. In our next work, we are going to use this approach to enhance the accuracy of protein identification in real mass spectrometry data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Abbreviations

Da Dalton

AA	amino acid
MPI	message passing interface
IPI	international protein index

Acknowledgments

This work was supported in part by "Clinical Proteomics Centers in Biodefense and Emerging Infectious Diseases" (NIAID contract HHSN272200800048C)

References

1. 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]
2. Toumi ML, Desaire H. Improving mass defect filters for human proteins. *J. Proteome Res.* 2010; 9:5492–5495. [PubMed: 20731397]
3. Yao X, Diego P, Ramos AA, Shi Y. Averagine-scaling analysis and fragment ion mass defect labeling in peptide mass spectrometry. *Anal. Chem.* 2008; 80:7383–7391. [PubMed: 18778085]
4. 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]
5. Hall MP, Ashrafi S, Obegi I, Petesch R, Peterson JN, Schneider LV. 'Mass defect' tags for biomolecular mass spectrometry. *Journal of Mass Spectrometry.* 2003; 38:809–816. [PubMed: 12938101]
6. Spengler B. Accurate mass as a bioinformatic parameter in data-to-knowledge conversion: Fourier transform ion cyclotron resonance mass spectrometry for peptide de novo sequencing. *Eur. J. Mass Spectrom.* (Chichester, Eng). 2007; 13:83–87.
7. 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. *Journal of the American Society for Mass Spectrometry.* 2004; 15:703–714. [PubMed: 15121200]
8. Olson MT, Epstein JA, Yergey AL. De novo peptide sequencing using exhaustive enumeration of peptide composition. *J. Am. Soc. Mass Spectrom.* 2006; 17:1041–1049. [PubMed: 16735127]
9. 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]
10. Mann M, Michalski A, Cox J. 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
11. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell Proteomics.* 2002; 1:376–386. [PubMed: 12118079]
12. 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; 5-11-1995. Ref Type: Conference Proceeding
13. Gay S, Binz PA, Hochstrasser DF, Appel RD. Modeling peptide mass fingerprinting data using the atomic composition of peptides. *Electrophoresis.* 1999; 20:3527–3534. [PubMed: 10612279]
14. Zubarev RA, Hakansson P, Sundqvist B. Accuracy requirements for peptide characterization by monoisotopic molecular mass measurements. *Analytical Chemistry.* 1996; 68:4060–4063.
15. Demirev PA, Zubarev RA. Probing combinatorial library diversity by mass spectrometry. *Analytical Chemistry.* 1997; 69:2893–2900. [PubMed: 21639310]

16. Fenyo D, Qin J, Chait BT. Protein identification using mass spectrometric information. Electrophoresis. 1998; 19:998–1005. [PubMed: 9638946]
17. 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; 41:281–288. [PubMed: 16538648]
18. Pacheco, P. Parallel Programming with MPI. 1st ed.. San Francisco: Morgan Kaufman; 1996.
19. Press, WH. Numerical recipes in C++ the art of scientific computing. 2nd ed.. Cambridge, UK: Cambridge University Press; 2002.
20. Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R. The International Protein Index: an integrated database for proteomics experiments. Proteomics. 2004; 4:1985–1988. [PubMed: 15221759]

```
procedure Generate( $L, start$ )  
  if  $start < N$  then  
    for  $i \leftarrow 0$  to  $L$   
       $c[start] \leftarrow i$   
      Generate( $L - i, start + 1$ )  
  else  
    for  $i \leftarrow 0$  to  $L$   
       $c[N] \leftarrow i$   
       $m \leftarrow \text{Mass}(c)$   
      print  $m$ 
```

Figure 1.

Pseudocode of the recursive procedure Generate that prints masses of all compositions of length not greater than L for an alphabet of N letters. The procedure should be called with parameters $(L, 1)$.

Figure 2.

Logarithmic plot of the number of tryptic peptide compositions with masses up to 3 kDa (separate plots for all lengths between 2 and 51, and five lengths from 5 to 25). Note the almost exponential growth of the number of compositions of all lengths as the mass increases.

Figure 3. Number of peptide compositions for three mass ranges. Red points mark peak centers, green points mark peak widths calculated according to equations (3), (4).

Figure 4.
Width of gaps between peaks of the mass distribution.

Figure 5.

Top graph: a fragment of the non-restricted mass distribution with red peaks corresponding to prohibited compositions, making up 18% of all compositions. Bottom graph: red peaks correspond to uniformly chosen compositions making up 18% of all compositions.

Figure 6.

The average entropy (red) and the margin of three standard deviations (aquamarine) plotted against the number of compositions (black) for the mass range 1900.25–1901.50 Da. The average entropy is higher for compositions making up the peak and lower for compositions making up the quiet zones.

Table 1

Number of compositions and sequences corresponding to all peptides and tryptic peptides of length 3–10.

Peptides' length (<i>L</i>)	All peptides (20 AA)		Tryptic peptides (18 AA, K or R in the end)	
	Number of compositions	Number of sequences	Number of compositions	Number of sequences
3	1,540	8,000	342	648
4	8,855	160,000	2,280	11,664
5	42,504	3,200,000	11,970	209,952
6	177,100	64,000,000	52,668	3,779,136
7	657,800	$1,280 \times 10^6$	201,894	68,024,448
8	2,220,075	$25,600 \times 10^6$	692,208	$1,224 \times 10^6$
9	6,906,900	$512,000 \times 10^6$	2,163,150	$22,040 \times 10^6$
10	20,030,010	$10,240,000 \times 10^6$	6,249,100	$396,719 \times 10^6$

Table 2

Maximum values of parameter L used in the job generation.

<i>start</i>	L_{max}
2	20
3	24
4	28
5	34

Table 3

Compositions from quiet zone 2000.436–2000.445 Da. Note long repetitions of amino acid Cys.

Mass (Da)	Number of compositions	Compositions
2000.436	1	P ₁ C ₁₃ D ₁ H ₂ R
2000.437	0	–
2000.438	1	P ₁ C ₁₁ D ₄ Y ₁ K
2000.439	1	S ₁ T ₁ C ₁₃ Y ₂ K
2000.440	3	C ₁₃ H ₁ Y ₁ W ₁ R, C ₁₃ N ₁ W ₂ R, G ₂ C ₁₃ W ₂ R
2000.441	5	T ₁ C ₁₀ D ₄ M ₂ K, T ₁ C ₁₁ D ₂ E ₂ M ₁ K, T ₁ C ₁₂ E ₄ K, S ₁ C ₁₀ D ₃ E ₁ M ₂ K, S ₁ C ₁₁ D ₁ E ₃ M ₁ K
2000.442	5	P ₁ C ₁₁ D ₂ M ₃ K, P ₁ C ₁₂ E ₂ M ₂ K, S ₂ C ₁₂ D ₁ H ₁ Y ₁ R S ₂ C ₁₂ N ₁ D ₁ W ₁ R, G ₂ S ₂ C ₁₂ D ₁ W ₁ R
2000.443	2	V ₁ C ₁₄ H ₁ F ₁ R, A ₁ C ₁₃ M ₁ H ₁ F ₁ R
2000.444	0	–
2000.445	2	S ₄ C ₁₁ N ₁ D ₂ R, G ₂ S ₄ C ₁₁ D ₂ R