Approximating Isotope Distributions of Biomolecule Fragments

Dennis Goldfarb†‡, Michael J. Lafferty‡, Laura E. Herring, Wei Wang, and Michael B. Major†‡

1Department of Computer Science, 2Lineberger Comprehensive Cancer Center, 3Bioinformatics and Computational Biology, 4UNC Michael Hooker Proteomics Core Facility, 5Department of Pharmacology, and 6Department of Cell Biology and Physiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, United States
7Department of Computer Science, University of California, Los Angeles, California 90095, United States

ABSTRACT: In mass spectrometry (MS)-based proteomics, protein and peptide sequences are determined by the isolation and subsequent fragmentation of precursor ions. When an isolation window captures only part of a precursor’s isotopic distribution, the isotope distributions of the fragments depend on the subset of isolated precursor isotopes. Approximation of the expected isotope distributions of these fragments prior to sequence determination enables MS2 deisotoping, monoisotopic mass calculation, charge assignment of fragment peaks, and deconvolution of chimeric spectra. However, currently such methods do not exist, and precursor isotope distributions are often used as a proxy. Here, we present methods that approximate the isotope distribution of a biomolecule’s fragment given its monoisotopic mass, the monoisotopic mass of its precursor, the set of isolated precursor isotopes, and optionally sulfur atom content. Our methods use either the Averagine model or splines, the latter of which have similar accuracy to the Averagine approach, but are 20 times faster to compute. Theoretical and approximated isotope distributions are consistent for fragments of in silico digested peptides. Furthermore, mass spectrometry experiments with the angiotensin I peptide and HeLa cell lysate demonstrate that the fragment methods match isotope peaks in MS2 spectra more accurately than the precursor Averagine approach. The algorithms for the approximation of fragment isotope distributions have been added to the OpenMS software library. By providing the means for analyzing fragment isotope distributions, these methods will improve MS2 spectra interpretation.

1. INTRODUCTION

In mass spectrometry (MS)-based proteomics, peptide sequences are determined by tandem MS scans (MS2), which isolate and subsequently fragment precursor ions. Frequently, only a part of a precursor’s isotopic distribution is captured by an isolation window due to isolation windows that are too narrow or are offset relative to the precursor. For example, data-dependent acquisition experiments typically use isolation windows that are 1.4–4 m/z wide. With a 1.4 m/z wide isolation window, only one to three isotopic peaks of a charge +2 peptide can fit within its boundaries. For example, if the window is centered >0.2 m/z below the monoisotopic peak, then only the monoisotopic peak would be isolated. This can occur for co-eluting peptides that were not the intended target of an MS2 scan because their m/z position relative to the isolation window is random. Since co-fragmentation is encountered in as many as 50% of MS2 spectra of complex samples, partial isolation of unexpected isotopes is common. This partial isolation leads to fragments with complex isotope distributions; these distributions depend on the subset of isolated precursor isotopes and the elemental compositions of both the precursor and the fragment of interest. Although a general approach to calculate the theoretical isotope distribution of a fragment from an individual precursor isotopic peak has been developed, this approach requires exact knowledge of elemental compositions. Typically, peptide sequences and elemental compositions are unknown a priori. Therefore, computational tasks that occur prior to sequence determination, including MS2 deisotoping, monoisotopic mass calculation, charge assignment of fragment peaks, and chimeric spectra deconvolution, do not take full advantage of fragment isotope distributions. To improve these preprocessing endeavors and to increase protein and peptide identifications, an efficient method is needed to approximate theoretical fragment isotope distributions based on observed mass to charge ratios (m/z) and isolation window parameters.

The isotope distribution of a molecule arises from the varying number of neutrons in its individual elements. Given
the elemental composition of a molecule, its theoretical isotope distribution is computed using methods such as polynomial expansion, multinomial expansion or the fast Fourier transform (FFT).1–7 If a molecule’s elemental composition is not known, but is comprised of similar structural units, such as amino acids or nucleotides, then its theoretical isotope distribution can be approximated. The most common approach is to first approximate the elemental composition using the Averagine model and then to compute the corresponding theoretical isotope distribution.9 A fractional Averagine approach was later developed that allowed continuous values for element counts and therefore avoided discontinuities due to element rounding.2 An alternative method utilizes the relationship between mass and isotope ratios. In the case of peptides, approximate isotope distributions are reconstructed by evaluating polynomial functions that are fit to the isotope ratios and masses of peptides generated in silico.10,11 Because of its unique isotope distribution, the number of sulfur atoms within a peptide creates a divergence in these patterns, particularly for shorter peptide sequences. If the number of sulfurs can be determined, then a more accurate prediction of isotope ratios can be achieved by utilizing models that are fit specifically to peptides with the same sulfur count.

Distributions of fragment isotopes are more complex than precursor isotope distributions. During the isolation and fragmentation of an individual precursor isotopic peak, each precursor in the population has the same number of neutrons, though the locations of the extra neutrons vary. Consequently, the isotope distribution of a fragment arises from the stochastic arrangement of neutrons within the precursor. The isotope distribution of a specific fragment is governed by the probabilities of extra neutrons residing in the given fragment versus its complementary fragment. Isolating multiple precursor isotopes adds further complexity, as the resultant fragment isotope distributions are linear combinations of the fragment isotope distributions from individual precursor isotopic peaks. Conveniently, isolation of a precursor’s entire isotopic distribution creates fragments whose distributions are equivalent to their natural isotope distributions as if they themselves were precursors. Software to calculate isotope probabilities after the dissociation of an individual precursor isotopic peak has been developed.12 Unfortunately, utilization of this method has been minimal.13 Extending the framework to subsets of precursor isotopes and providing a method to approximate isotope distributions will increase its utility and range of applications. Such opportunities exist in the preprocessing of MS2 spectra of unknown elemental compositions. Current algorithms for deisotoping and charge reduction in MS2 spectra use approximate precursor isotope distributions, and despite not being the ideal model, still increase peptide identifications.14–22

Here, we developed methods that approximate fragment isotope distributions when exact elemental compositions are not known. The existing general framework for fragment isotope distributions of individual precursor isotopes was rederived and then extended for subsets of isotopes. Next, the Averagine model was incorporated within this framework to support biomolecules of unknown elemental compositions. Given that sulfurs have a large effect on the isotope distributions of small peptides, which are frequent in MS2 spectra, a sulfur-specific Averagine approach was developed and evaluated for both precursors and fragments. Furthermore, individual precursor isotope probabilities followed a smooth nonlinear pattern that were summarized with splines and used in place of the Averagine model. The accuracy and speed of these methods were evaluated in silico digested peptides, mass spectrometry experiments utilizing the angiotensin I peptide, and in complex peptide mixtures from HeLa cells lysate.

2. RESULTS AND DISCUSSION

2.1. Derivation of a Model for Fragment Isotope Distributions. A common scenario in mass spectrometry proteomics is the isolation and dissociation of a partial isotopic distribution, which results in fragments with complex isotope distributions. To predict these distributions, we modeled the probabilities of a fragment’s isotopic state given its elemental composition, the elemental composition of its precursor, and the boundaries of the employed isolation window. This model requires explicitly stating the five assumptions employed, of which the first three are self-evident: (1) a single molecule cannot simultaneously be in multiple isotopic states; (2) the molecule cannot have fewer neutrons than its monoisotopic state; and (3) when no other information is available, the number of neutrons in a fragment and its complementary fragment are independent of each other. However, the other two assumptions are not entirely accurate: (4) there is uniform isolation within the isolation window, and (5) there is zero isolation outside of the isolation window’s boundaries. Current mass spectrometry instrumentation does not achieve the perfect box shape for an isolation window; isolation efficiency often decreases near the edges of the window and is nonzero just outside of its boundaries.2,23 Therefore, our model reflects an idealized scenario. For the fragmentation of a single precursor isotopic peak, the model is equivalent to the framework by Rockwood. The method described here for determining the exact theoretical fragment isotope distributions was added to the OpenMS library along with unit tests to ensure correctness.

2.2. Averagine Model Incorporation. In discovery experiments, the identities and elemental compositions of the molecules in each MS2 scan are unknown. For such cases, the model for calculating fragment isotope distributions cannot be used directly, and an approximation method must be used instead. The application of the Averagine model to approximate isotope distributions of a precursor is referred to as the precursor Averagine method. However, the method described here uses the Averagine model to approximate the isotope distribution of a fragment and is therefore referred to as the fragment Averagine method. In the fragment Averagine method, the Averagine model is used to approximate the elemental composition of a fragment and its complementary fragment from their average masses. An alternative method is to first approximate the compositions of the precursor and fragment and then subtract the composition of the fragment from the precursor to determine the composition of the complementary fragment. This is the more computationally efficient method when approximating multiple fragment isotope distributions for the same precursor. However, this approach will often lead to negative hydrogen counts for the complementary fragment due to rounding and hydrogen compensation performed by the Averagine approach. For example, the Averagine approach approximates an elemental composition of C_{60}H_{102}O_{46}N_{95}S_{6} for a mass of 1340 Da and C_{60}H_{102}O_{46}N_{95}S_{6} for a fragment mass of 1220 Da. Subtracting the two compositions leads to an approximate elemental
composition of $C_{6}H_{30}O_{2}N_{1}S_{1}$ for the complementary fragment of mass 120 Da, which is not compatible with software that calculates isotope distributions.

The standard Averagine approach uses the average mass calculated from observed peaks, however, when only part of an isotopic distribution is isolated, a fragment’s isotope distribution is no longer representative of its average mass. Furthermore, difficulty arises for low abundant ions where the monoisotopic peak may not be observed due to low abundance. For the evaluations performed in this work, elemental compositions were known, and the correct average masses were used. When average masses are not known, an approach based on observed peaks will be necessary and will result in some mass errors, however, the effect on approximate isotope distributions due to inaccuracy of a few Daltons is negligible. The approximation methods for fragment isotope distributions using the Averagine and sulfur-specific Averagine models have also been added to the OpenMS library.

2.3. Spline Construction. Although the Averagine model combined with the FFT has successfully been used to approximate isotope distributions, they have some undesirable properties. The Averagine model has discontinuities due to the rounding of element counts, with the largest effect due to sulfurs as demonstrated by the vertical jumps within the blue lines of Figure 1. Additionally, the FFT is often replaced with a precomputed lookup table at several Dalton intervals when extremely fast computation is necessary. The fractional Averagine method avoids discontinuities, but requires five additional convolutions and is therefore slower to compute. As an alternative, we used splines to model isotope probabilities in a compact and efficient data structure (Figures 1 and S1). Although the probabilities follow a consistent pattern, divergence results from the distinct isotope distribution of sulfur-containing peptides (Figure 1). To address this, sulfur-specific splines were fitted separately to peptides containing the matching number of sulfurs (Figures 1 and S2). Both the sulfur-specific and average splines showed excellent goodness of fit with $>0.99$ $R^{2}$ values (Table S1), and the best fits were exhibited by the sulfur-specific models. Computing approximate isotope distributions with splines is nearly 20 times faster than the Averagine and FFT approach (Figure S3). The disadvantage of splines is that the requested mass must be within the mass range to which that spline was fitted. This can be mitigated by training the model to the anticipated range of queries or by defaulting to the Averagine and FFT method when the requested mass is out of range. A sample Java program to parse and compute approximate isotope distributions using the spline models is available at our Github repository.

2.4. In silico Evaluation. To determine how well approximate distributions matched to theoretical fragment isotope distributions, we calculated chi-squared statistics between approximate and theoretical distributions for each $b$ and $y$ fragment from all tryptic peptides in the human proteome (Figure 2). The precursor Averagine approach was included as a baseline and to demonstrate that it is inappropriate for fragment isotopes except when most of the precursor isotopic distribution is isolated. As shown in the first row of Figure 2, the precursor Averagine approach improves as more isotopes are isolated. For the fragment methods, the sulfur-specific Averagine approach and sulfur-specific splines were the best matches. The sulfur-specific splines were slightly better, having a 10% smaller median $\chi^2$ score and 7% smaller mean (Table S2). The fragment Averagine approach and splines were nearly identical to the sulfur-specific methods when isotopes less than $M + 2$ were isolated. Interestingly, the
fragment Averagine approach has a 37% smaller median $\chi^2$ score than the splines, but it has a 23% larger mean. The fragment Averagine approach has a better best case because it can sometimes approximate a peptide’s exact or near exact elemental composition, but in rare situations the compositions are very inaccurate and negatively skew the mean. Overall, the sulfur-specific methods are the best matches to theoretical distributions, but the sulfur-specific methods require that the number of sulfur atoms be known. Furthermore, the fragment Averagine approach is a better match than the splines in most cases.

2.5. Angiotensin I Evaluation. To experimentally validate the theoretical calculations and approximation methods, we directly infused Angiotensin I peptide into the mass spectrometer and isolated and fragmented different subsets of precursor isotopes (Figure 3). The two most abundant fragment ions, B$_5^+$ and B$_9^{++}$, displayed minor deviation from the theoretical distributions at least partially due to sample sizes and nonuniform isolation efficiency within and beyond the isolation window boundaries. Evidence for isolation outside of the isolation window is the small M+1 peaks observed for both fragments when only the monoisotopic precursor should have been isolated. Once again, the precursor Averagine approach was only appropriate when most of the precursor isotopic distribution was isolated. Conversely, all of the fragment methods recapitulated the observed isotopic distributions. It is notable that many of the isotopic distributions are visibly distinguishable from each other except when the only difference is the isolation of a low abundant precursor isotope. This implies that the set of isolated precursor isotopes that created a fragment could be inferred from the fragment’s isotopic distribution and can potentially be used to deconvolve chimeric spectra generated by the co-isolation of multiple precursors with different sets of isotopes.

2.6. Whole-Cell Lysate Evaluation. To test the accuracy of these methods on complex samples utilizing typical instrument settings, we performed a shotgun proteomics experiment with whole-cell lysate from HeLa cells using data-dependent acquisition. After a database search to identify peptide-spectrum matches (PSMs), fragment isotopic distributions were compared to theoretical and approximate isotope distributions (Figure 4 and Table 1). Observed fragment isotopic distributions were compared with theoretical and approximate isotope distributions if the first two or three isotope peaks were detected. The multimodal nature of the chi-squared scores is due to separate, but overlapping, distributions that correspond to the number of missing fragment isotopes. The leftmost distributions have no missing isotopes, whereas more undetected isotopes result in greater chi-squared scores. The precursor Averagine approach had 34 and 74% higher chi-squared scores on average compared to all of the other approximation methods. The fragment methods’ average chi-squared scores were all within 2% of each other (including the sulfur-specific methods), suggesting that in a high-throughput and complex setting, experimental sources of variance, bias, and interference far outweigh the theoretical impact of sulfurs.

3. CONCLUSIONS

Experimental fragment isotopic distributions can be accurately computed and approximated despite the inability of current mass spectrometers to employ perfect box-shaped isolation windows. Taking a probabilistic approach, the equations for theoretical fragment isotope distributions were rederived and expanded to handle the isolation of multiple precursor isotopes. Two approximation approaches were developed: one using the Averagine model and the other using splines. Although the spline models can be slightly less accurate than the Averagine model when compared to theoretical distribu-
Figure 3. MS2 scans were performed on directly infused angiotensin I peptide using various isolation windows. Different sets of precursor isotopes were captured in each scan (right axis labels and diagrams). Profile data are displayed of the two most abundant fragments of angiotensin I after collision-induced dissociation (CID) fragmentation: B5+ and B9++. All peaks within 1 m/z of a fragment’s isotopic distribution were extracted from the profile data, and computed isotope distributions were scaled to the extracted base peak. The circles and squares represent the predicted abundances.
tions, in a high-throughput shotgun experiment the splines were equally accurate. Therefore, the spline models are a viable alternative, especially when speed is a top priority. Furthermore, sulfur-specific variants of both approaches were introduced. Their usage requires knowledge of the number of sulfur atoms in a peptide, or iteratively testing each possibility and choosing the best fit. However, the sulfur-specific variants did not produce better matches to experimental fragment isotopic distributions. The worst performing method was the precursor Averagine approach, which is only appropriate for calculating fragment isotopic distributions when a precursor’s entire isotopic distribution is isolated. The methods to calculate theoretical and approximate fragment isotopic distributions using the Averagine and sulfur-specific Averagine models have been added to the OpenMS library so that they can be utilized by future approaches to process MS2 spectra.

4. EXPERIMENTAL SECTION

4.1. Probabilistic Model for Fragment Isotope Distributions. The nominal isotope probabilities of a fragment after the isolation and fragmentation of a single precursor ion were modeled. A molecule’s nominal isotopic state is its difference in neutrons relative to its monoisotopic form. In the remaining text, nominal isotopes are referred to simply as isotopes. For notation: random variables are represented with capital letters; specific values are represented with lowercase letters; a collection of specific values is denoted by bold lowercase letters; and unions represent logical “or”. Model variables are defined in Table 2. The term isotope distribution refers to a probability distribution, whereas the term isotopic distribution refers to the isotopic peaks observed in a mass spectrum. The five assumptions underlying our model are as follows:

(1) Mutual exclusivity of isotopic states: A single molecule cannot simultaneously exist in multiple isotopic states.
(2) Independence of isotopic states between a fragment and its complementary fragment: The isotopic state of a fragment does not influence the isotopic state of its complementary fragment when not conditioned upon another event.
(3) Non-negativity of isotopic states: A molecule cannot have fewer neutrons than its monoisotopic form.

Table 1. Summary of Chi-Squared Statistics from HeLa Cell Lysate Experiment

| method                  | median  | mean   | sample size | isotope count |
|-------------------------|---------|--------|-------------|---------------|
| theoretical fragment    | 0.0931  | 0.1463 | 69 027      | 2             |
| precursor Averagine      | 0.1711  | 0.2586 | 69 027      | 2             |
| fragment Averagine       | 0.0957  | 0.1486 | 69 027      | 2             |
| splines                 | 0.0911  | 0.1459 | 69 027      | 2             |
| sulfur-specific Averagine| 0.0956  | 0.1488 | 69 027      | 2             |
| sulfur-specific splines  | 0.0909  | 0.1459 | 69 027      | 2             |
| theoretical fragment    | 0.1679  | 0.3008 | 20 131      | 3             |
| precursor Averagine      | 0.2527  | 0.4121 | 20 131      | 3             |
| fragment Averagine       | 0.1710  | 0.3064 | 20 131      | 3             |
| splines                 | 0.1685  | 0.3017 | 20 131      | 3             |
| sulfur-specific Averagine| 0.1695  | 0.3064 | 20 131      | 3             |
| sulfur-specific splines  | 0.1671  | 0.3021 | 20 131      | 3             |

Table 2. Variable Descriptions for Isotope Probability Model

| symbol | description                                                                 |
|--------|----------------------------------------------------------------------------|
| p      | random variable for the nominal isotopic state of precursor with known elemental composition |
| f      | random variable for the nominal isotopic state of fragment with known elemental composition |
| c      | random variable for the nominal isotopic state of a complementary fragment, whose elemental composition is that of the precursor minus the fragment |
| p      | specific value for the precursor’s nominal isotopic state                  |
| f      | specific value for the fragment’s nominal isotopic state                   |
| p      | subset of precursor isotopes that can be isolated by the isolation window  |
4.4. Approximation Using Splines. Splines were fitted to theoretical isotope probabilities of in silico generated amino acid sequences. Each isotope had a training data set consisting of average masses and corresponding isotope probabilities for simulated sequences. The simulated sequences were varied in length from one to 1000 amino acids with a maximum mass of 100 kDa. For each sequence length, 1000 sequences were generated by choosing a random amino acid for each position. To mimic the distribution of amino acid combinations observed in nature, the amino acids were sampled from a probability distribution corresponding to the observed frequencies for the 20 most common amino acids found in the human canonical SwissProt database (downloaded 11/28/16). After sequence generation, theoretical isotope distributions were computed up to the first 100 isotopes. Sulfur-specific training sets containing zero to five sulfurs were generated separately for each case. The construction of the sulfur-containing sequences was identical to the procedure described above, except that once the sequence contained the desired number of sulfurs (from methionine or cysteine residues), the rest of the sequence was derived from the remaining 18 amino acids. Random sequences were chosen over in silico proteome digests to minimize gaps and biases in mass coverage.

Individual cubic splines were fit for each isotope (\(M - M + 100\)) and sulfur count (0–5, all) combinations using MATLAB’s Curve Fitting Toolbox (version R2016a). Initially, knots were uniformly spaced along the mass axis at 2 kDa intervals with the first and last knots repeated four times to force the splines to have two continuous derivatives. Next, cubic B-splines were fit using a least-squares approximant and the initial knot sequence. Knot selections were then adjusted to uniformly distribute the errors of the least-squares approximant, and the B-splines were refit. The final B-splines were converted to piece-wise polynomial format and written to an XML file.

4.5. Chemicals and Standards. Angiotensin I was purchased from Sigma (St. Louis, MO; catalog number A9650) and reconstituted to a final concentration of 1 pmol/μL in a solution of 50:50 (methanol:water) containing 0.1% acetic acid. Pierce HeLa Protein Digest Standard was purchased from Thermo Fisher Scientific (Waltham, MA; catalog number 88328) and diluted to a final concentration of
200 ng/μL in a solution of 98:2 (water/acetonitrile) containing 0.1% formic acid.

4.6. Mass Spectrometry. Angiotensin I peptide was analyzed by direct infusion into an Orbitrap Fusion Lumos mass spectrometer (Thermo Scientific). The syringe pump was operated at a flow rate of 3 μL/min. The heated electrospray ionization ion source voltage was +3.5 kV; the radio frequency (RF) lens was set to 30%; and the ion-transfer tube was maintained at 300 °C. MS2 scans were acquired by the Orbitrap analyzer at 15 k resolution using a 5e4 AGC target, 30 ms max injection time, and collision-induction dissociated (CID) at 30% collision energy. The MS2 scans were performed in a targeted manner using an inclusion list to isolate and fragment varying isotopes of the precursor in the +3 charge state. The inclusion list consisted of isolation windows with widths ranging from 0.4 to 2.4 m/z at 0.1 m/z intervals and isolation window offsets ranging from −1.2 to 1.2 m/z at 0.05 m/z intervals relative to the +3 precursor monoisotope (m/z = 432.9).

Trypsinized peptides (200 ng) from HeLa cell lysate were separated via reverse-phase chromatography using a nano-ACQUITY UPLC system (Waters Corporation; Milford, MA) and analyzed by an Orbitrap Fusion Lumos. Peptides were trapped on a 2 cm column (Pepmap 100, 3 μm particle size, 100 Å pore size) and separated in a 25 cm EASY-spray analytical column (75 μm ID, 2.0 μm C18 particle size, 100 Å pore size) at 300 nL/min and 35 °C using a 180 min gradient from 2 to 25% buffer B (0.1% formic acid in acetonitrile). The EASY-spray ion source voltage was set to +1.95 kV; the RF lens was set to 30%; and the transfer tube was maintained at 275 °C. The mass spectrometer was operated in data-dependent acquisition mode with a 3 s cycle time (TopSpeed). Full MS scans were obtained at 60 k resolution by the Orbitrap mass analyzer, with a 400–1550 m/z scan range, 4e5 AGC target, and 50 ms maximum injection time. For MS2 selection, peptide monoisotopic peak determination was enabled, and dynamic exclusion was set to 60 s with a 10 ppm mass tolerance. Further MS2 selection criteria included a 5e4 intensity threshold and inclusion of charges 2−7. Isolation was performed by a quadrupole using isolation windows of 1.6 m/z width and centered on the most abundant isotopic peak. MS2 scans were obtained by the Orbitrap mass analyzer at 15 k resolution using a 5e4 AGC target, 50 ms maximum injection time, and 25% CID collision energy.

4.7. Data Analysis. Angiotensin I data were processed via custom programs utilizing the OpenMS library. Prior to analysis, raw data were converted twice into mzML format using ProteoWizard’s MSConvert. In one conversion, the profile data were centroided; in the second conversion, the profile data were preserved for plotting purposes. Scans that included contiguous subsets of the first four precursor isotopes were identified based on isolation window parameters, and isotopic peaks from the two most abundant fragment ions (B9+, DRVYI and B9++ DRVYIHPFH) were extracted. The extraction process consisted of searching the centroided data for the monoisotopic fragment peak up to the largest isolated isotope using a 10 ppm mass tolerance. Observed isotope intensities for each fragment within each scan were normalized to a sum of one. Theoretical and approximate isotope distributions were computed using the OpenMS implementations of the previously described methods. When calculating precursor isotope distributions, the first seven isotope probabilities were computed, and isotopes were removed if both of the following were true: (1) their abundance was less than 10% of the most abundant isotope and (2) the isotope was greater than the maximum isolated isotope. After filtering, the isotope probabilities were renormalized such that they sum to one. To evaluate goodness of fit between observed and computed distributions, chi-squared (χ²) statistics were calculated using the computed distributions as the expected values.

HeLa cell lysate data were analyzed by database search within an OpenMS workflow, followed by a custom program to evaluate the fits of approximated isotope distributions. After conversion to mzML, a database search was performed using MSGF+ against the human canonical SwissProt database (downloaded 11/28/16) appended with reversed decoy sequences. Search parameters included a static Carbamidemethyl (C) modification, variable oxidation (M) modification, maximum of two modifications, 10 ppm precursor mass tolerance, fully tryptic digest, 6–40 amino acid length, charge states of 2−4, no isotope error, and the Q-Exactive instrument parameter. Peptide-spectrum matches (PSMs) were scored using Percolator (version 3.0) and filtered for a 1% false discovery rate. The custom program then extracted MS2 spectra for each PSM and calculated the m/z for each b and y ion of charge +1 up to one less than the precursor charge. Using the same procedure as described above for angiotensin peptide, the fragment isotopes were found in the spectrum; their theoretical and approximate isotope distributions were computed, and chi-squared statistics were calculated. The source code used to generate all figures in this manuscript is available at www.github.com/MajorLab/Fragment-Isotope-Distribution-Paper/.
REFERENCES

(1) 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, 1785–1793.

(2) Scheltema, R. A.; Hauschild, J.-P.; Lange, O.; Hornburg, D.; Denisov, E.; Damoc, E.; Kuehn, A.; Makarov, A.; Mann, M. The Q Exactive HF, a Benchtop Mass Spectrometer with a Pre-filter, High-performance Quadrupole and an Ultra-high-field Orbitrap Analyzer. Mol. Cell. Proteomics 2014, 13, 3698–3708.

(3) Houel, S.; Abernathy, R.; Renganathan, K.; Meyer-Arendt, K.; Ahn, N. G.; Old, W. M. Quantifying the Impact of Chimera MS/MS Spectra on Peptide Identification in Large-Scale Proteomics Studies. J. Proteome Res. 2010, 9, 4152–4160.

(4) Rockwood, A. L.; Kushnir, M. M.; Nelson, G. J. Dissociation of individual isotopic peaks: predicting isotopic distributions of product ions in Msn. J. Am. Soc. Mass Spectrom. 2003, 14, 311–322.

(5) Brownawell, M. L.; Filippo, J. S. Chemical instrumentation. II: A program for the synthesis of mass spectral isotopic abundances. J. Chem. Educ. 1982, 59, No. 663.

(6) Yergey, J. A. A general approach to calculating isotopic distributions for mass spectrometry. Int. J. Mass Spectrom. Ion Phys. 1983, 52, 337–349.

(7) Rockwood, A. L.;Orden, S. L. V.;Smith, R. D. Rapid Calculation of Isotope Distributions. Anal. Chem. 1995, 67, 2699–2704.

(8) Senko, M. W.; Beu, S. C.; Mclafferty, F. W. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. J. Am. Soc. Mass Spectrom. 1995, 6, 229–233.

(9) Renard, B. Y.; Kirchner, M.; Steen, H.; Steen, J. A.; Hamprecht, F. A. NITPICK: peak identification for mass spectrometry data. BMC Bioinformatics 2008, 9, No. 355.

(10) Valkenborg, D.; Jansen, I.; Burzykowski, T. A model-based method for the prediction of the isotopic distribution of peptides. J. Am. Soc. Mass Spectrom. 2008, 19, 703–712.

(11) Glavidlo, F. Z.; Mertens, I.; Baggerman, G.; Laukens, K.; Burzykowski, T.; Valkenborg, D. The use of the isotopic distribution as a complementary quality metric to assess tandem mass spectra results. J. Proteomics 2014, 98, 150–158.

(12) Ramaley, L.; Herrera, L. C. Software for the calculation of isotope patterns in tandem mass spectrometry. Rapid Commun. Mass Spectrom. 2008, 22, 2707–2714.

(13) Rockwood, A. L.; Palmblad, M. Mass Spectrometry Data Analysis in Proteomics; Humana Press, 2013.

(14) Carvalho, P. C.; Xu, T.; Han, X.; Cociorva, D.; Barbosa, V. C.; Yates, J. R. YADA: a tool for taking the most out of high-resolution spectra. Bioinformatics 2009, 25, 2734–2736.

(15) Xiao, K.; Yu, F.; Fang, H.; Xue, B.; Liu, Y.; Tian, Z. Accurate and Efficient Resolution of Overlapping Isotopic Envelopes in Protein Tandem Mass Spectra. Sci. Rep. 2015, 5, No. 14755.

(16) Chen, L.; Sze, S. K.; Yang, H. Automated Intensity Descent Algorithm for Interpretation of Complex High-Resolution Mass Spectra. Anal. Chem. 2006, 78, 5006–5018.

(17) Horn, D. M.; Zubarev, R. A.; Mclafferty, F. W. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. J. Am. Soc. Mass Spectrom. 2000, 11, 320–332.

(18) Zabrouskov, V.; Senko, M. W.; Du, Y.; Leduc, R. D.; Kelleher, N. L. New and automated MSn approaches for top-down identification of modified proteins. J. Am. Soc. Mass Spectrom. 2005, 16, 2027–2038.

(19) Liu, X.; et al. Deconvolution and Database Search of Complex Tandem Mass Spectra of Intact Proteins: A Combinatorial Approach. Mol. Cell. Proteomics 2010, 9, 2772–2782.