Published in final edited form as:

J Proteome Res. 2018 January 05; 17(1): 751–758. doi:10.1021/acs.jproteome.7b00807.

Poisson Model To Generate Isotope Distribution for Biomolecules

Rovshan G. Sadygov[*]

Department of Biochemistry and Molecular Biology, Sealy Center for Molecular Medicine, The University of Texas Medical Branch, Galveston, Texas 77555, United States

Abstract

We introduce a simplified computational algorithm for computing isotope distributions (relative abundances and masses) of biomolecules. The algorithm is based on Poisson approximation to binomial and multinomial distributions. It leads to a small number of arithmetic operations to compute isotope distributions of molecules. The approach uses three embedded loops to compute the isotope distributions, as compared with the eight embedded loops in exact calculations. The speed improvement is about 3-fold compared to the fast Fourier transformation-based isotope calculations, often termed as ultrafast isotope calculation. The approach naturally incorporates the determination of the masses of each molecular isotopomer. It is applicable to high mass accuracy and resolution mass spectrometry data. The application to tryptic peptides in a UniProt protein database revealed that the mass accuracy of the computed isotopomers is better than 1 ppm. Even better mass accuracy (below 1 ppm) is achievable when the method is paired with the exact calculations, which we term a hybrid approach. The algorithms have been implemented in a freely available C/C++ code.

Graphical Abstract

[*] Corresponding Author: Phone: 409-772-3287. Fax: 409-772-9670. rovshan.sadygov@utmb.edu.

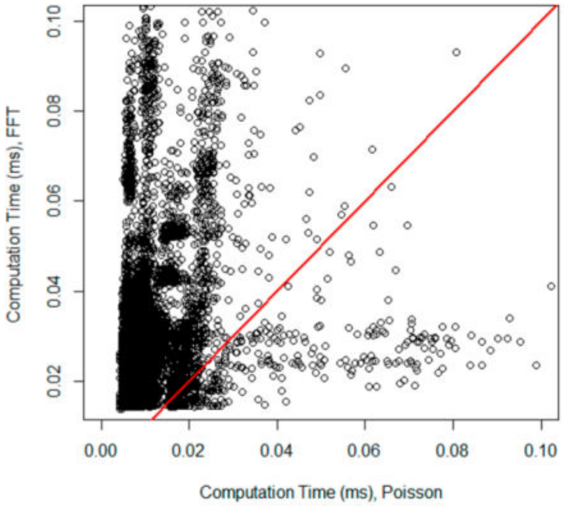
Notes

The author declares no competing financial interest.

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.7b00807.

Table S1 and Figures S1 and S2 (PDF)



Keywords

Poisson distribution; isotope envelope; mass spectrometry; stable isotope labeling

INTRODUCTION

Calculations of isotope distributions of biomolecules are important in diverse areas where biomedical mass spectrometry is used.[1,2] Mass spectrometry-based metabolomics and proteomics employ isotope distributions of biomolecules routinely. Recently, stable isotope labeling has been a major element for static and as well as dynamic proteome or metabolome studies. Atom-based labeling approaches, such as ^{2}H,[3–6] ^{13}C,[7] or ^{15}N,[8–10] for example, or residue-based labeling, such as SILAC, are popular techniques in quantitative proteomics.[11–15] The studies use isotopomer distributions of natural peptides and isotopically labeled peptides[16] to determine the relative expression of labeled proteins. Therefore, generating isotope distributions of biomolecules from their elemental composition has been a focus of research for some time.[16–20] Two main approaches may be classified. The first approach uses binomial and multinomial distributions to compute the isotope distribution of atoms and then uses their convolution to determine the isotope distribution of a biomolecule.[21] The approach can be computationally costly when a molecule is large. Its computational complexity scales with the number of atoms polynomially, and techniques have been developed to speed up the calculations.[17,22] The second approach uses fast Fourier transform (FFT)-based convolution to compute the isotope distributions of atoms and then of a molecule.[19,20] This approach can be faster, but in general, there are advantages to each approach depending on the molecular weight, number of isotopomers, and desired mass resolution.[16] The FFT-based approach has large memory requirements for large molecules and high mass resolution data. Due to the computational speed requirements, some bioinformatics applications needing isotope distributions preprocessed the sequences first and store their distributions.[21]

In this work, we describe a computationally efficient and conceptually simple approach to calculate isotope distributions of biomolecules using a Poisson approximation of binomial and multinomial distributions. We show that, in most of the cases, a single Poisson formula is enough to compute the isotope distribution. The mass distribution of isotopes can be computed from the same Poisson distribution as an expectation value of a distribution. The paper is organized as follows. First, we introduce the notion of isotope distributions of atoms. Then we show how Poisson distribution approximates the combined isotope distributions of hydrogen (H), carbon (C), and nitrogen (N) atoms. Next, we approximate the multinomial isotope distributions of oxygen (O) and sulfur (S) with multiple Poisson distributions. Single isotope distributions from the O and S atoms are naturally incorporated into the combined Poisson distribution of H, C, and N. The final isotope distribution of a molecule is a convolution of relevant isotope distributions of its atoms. The computations use a number of arithmetic operations which are linear in the sum of the O and S atoms. The mass distributions of isotopes of a biomolecule are obtained as expectation values from Poisson distributions.

METHODS

Isotope Distribution of a Peptide

We will consider isotope distributions of biomolecules made of five types of atoms, H, C, N, O, and S. The first three atoms have two naturally occurring isotopes. Isotope distributions of any of these atoms can be computed using a binomial distribution formula:[23]

$$p(k_{\mathrm{A}}, N_{\mathrm{A}}) = \binom{N_{\mathrm{A}}}{k_{\mathrm{A}}} p_{\mathrm{A}}^{k_{\mathrm{A}}} (1-p_{\mathrm{A}})^{N_{\mathrm{A}}-k_{\mathrm{A}}} \qquad (1)$$

where N_A is the number of atoms of type A (any of H, C, and N) and p_A is the naturally occurring frequency of the isotope of A. Formula 1 is the probability that there are k_A isotopes of N_A atoms of A.

In the case of O and S atoms, there are more than two (three for O and five for S) naturally occurring isotopes. The corresponding exact probabilities of isotope distributions for these atoms are calculated using a multinomial distribution. For example, for a sulfur atom with four isotopes the distribution is

$$p(k_0, k_1, k_2, k_3) = \frac{N_{\mathrm{S}}!}{k_0! k_1! k_2! k_3} p_0^{k_0} p_1^{k_1} p_2^{k_2} p_3^{k_3} \qquad (2)$$

where N_S is the number of S atoms and k_0, k_1, k_2, and k_3 are the number of monoisotopic, first, second, and third isotope atoms of S, with the corresponding probabilities of p_0, p_1, p_2, and p_3. The general relationships of multinomial distributions hold:

$$N_{\mathrm{S}}=k_0+k_1+k_2+k_3 \text{ and } p_0+p_1+p_2+p_3=1$$

Computation of the isotope distributions using eq 2 is somewhat computationally expensive, and an approach has been developed to do the computations “recursively”.[17] We note that another often occurring in a biomolecular atom, phosphorus, has only one naturally occurring isotope and does not contribute to the isotopic complexity of biomolecules and will not be considered here.

The isotope distribution of a biomolecule comprising H, C, N, O, and S atoms can be thought of as the distribution of a random variable, X, which is the sum of the random variables corresponding to each atomic isotope:

$$X=X_{\mathrm{H}}+X_{\mathrm{C}}+X_{\mathrm{N}}+X_{\mathrm{O}}+X_{\mathrm{S}} \quad (3)$$

where X_H, X_C, and X_N are the binomial random variables distributed according to eq 1 and X_O and X_S are distributed according to the multinomial distribution (eq 2). While it is possible to compute the isotope distributions of each atom using eqs 1 and 2, there is no exact (or approximate) closed formula for the isotope distribution of a molecule composed of several different atoms (each with a different isotope distribution). The probability distribution of a sum of random variables is a convolution (in the case of the isotope distributions, a discrete convolution) of probability distributions of each random variable and is formally expressed as

$$P(X)=P(X_{\mathrm{H}}) \otimes P(X_{\mathrm{C}}) \otimes P(X_{\mathrm{N}}) \otimes P(X_{\mathrm{O}}) \otimes P(X_{\mathrm{S}}) \quad (4)$$

Due to the absence of a closed form formula, it is computationally costly to compute the isotope distributions using eq 4, as the number of atoms in biomolecules is large and the number of possible configurations grows polynomially (different combinations of atomic isotopes may contribute to the same molecular isotope). In addition, in a typical proteomics application, tens of thousands of peptides are analyzed in a single workflow. Each of the peptides may need isotope distributions computed. Equation 4 provides a way to compute the relative abundances of the peaks of the isotope patterns. The sum of all isotope abundances is normalized to be equal to 1. The peaks of the isotope pattern are denoted with integers (0, 1, …, n), counting the mass offset from the light (i = 0) isotope (for the atoms that we consider here, it is termed the monoisotope). The monoisotopic peak is made of the atomic isotopes with highest natural abundance, and it corresponds to the lowest mass-to-charge ratio (m/z) in an isotope pattern. To denote the isotopes of a molecule, we use the following nomenclature: (m_i, M_i), where the index i is the mass shift from the monoisotope (i = 0). m_i and M_i denote the mass of the ith isotope (also referred here to as isotopomer) and its relative abundance, respectively.

In mass spectrometric applications, in addition to the isotope distribution, it is also important to accurately compute the (averaged) mass of each isotope. High mass accuracy and resolution mass spectrometers allow mass accuracy in few parts per million (ppm) or better. Therefore, it becomes necessary to compute the isotope masses with high mass accuracy to take advantage of the capabilities of modern mass spectrometers and compare the computed isotope masses to the experimentally measured values. The mass of a certain isotopomer, m_i, is determined as an expectation value of masses of atomic isotopes whose configuration leads to this mass:

$$E[m_i]=m_0+\sum_{\substack{k_{\mathrm{H}}:k_{\mathrm{C}}:k_{\mathrm{N}}:k_{\mathrm{O}}:k_{\mathrm{S}}\\ m_i=k_{\mathrm{H}}*m_{\mathrm{H}}+k_{\mathrm{N}}*m_{\mathrm{N}}+k_{\mathrm{C}}*m_{\mathrm{C}}+k_{\mathrm{O}}*m_{\mathrm{O}}+k_{\mathrm{S}}*m_{\mathrm{S}}}}\left(k_{\mathrm{H}}*m_{\mathrm{H}}+k_{\mathrm{N}}*m_{\mathrm{N}}+k_{\mathrm{C}}*m_{\mathrm{C}}+\vec{k}_{\mathrm{O}}*\vec{m}_{\mathrm{O}}+\vec{k}_{\mathrm{S}}*\vec{m}\right.$$
$$\left.*P\left(k_{\mathrm{H}},k_{\mathrm{C}},k_{\mathrm{N}},\vec{k}_{\mathrm{O}},\vec{k}_{\mathrm{S}}\right)\right.$$

(5)

where m_0 is the mass of the monoisotope, $P(k_H,k_C,k_N,k_O,k_S)$ is the probability of the numbers of H, C, N, O, and S atoms that results in the mass m_i, m_H is the mass difference between the deuterium and hydrogen atoms, and $\vec{k}_S*\vec{m}_S$ denotes the fact that there multiple isotopes of a sulfur atom (the vector $\vec{m}_S$ consists of mass differences for each of its heavy isotopes). In eq 5, the sum is taken over all configurations of number of atoms that lead to the ith isotope of a biomolecule. In Figure 1, we show implementation of eq 5 to compute the relative isotope abundances and the corresponding isotope masses. We termed this approach an exact algorithm.

Poisson Approximation for Isotope Distributions of Multiatomic Molecules

Poisson distribution has a well-known approximation to the binomial distribution[24] when the probability at each trial is small (as is the case with the heavy atom isotopes). For example, the probability to observe k_A heavy isotopes in an isotope distribution of the H, C, and N atoms is

$$P_{\mathrm{A}}(k_{\mathrm{A}})=\frac{\lambda_{\mathrm{A}}^{k_{\mathrm{A}}}}{k_{\mathrm{A}}!}e^{-\lambda_{\mathrm{A}}} \qquad (6)$$

where λ_A is the average number of successes, $\lambda_A = N_A \times p_A$. Note the difference between the capital, P_A (probability of a Poisson distribution), and lower case, p_A (the probability of success in each trial – Bernoulli distribution). The subscript “A” can be any of H, C, and N. For example, λ_H for 200 H atoms willbe $\lambda_H = 200 \times 0.000115 = 0.023$. To compute probability that among 200 randomly chosen hydrogen atoms there will be (for example) 15 deuteriums, we will set $k_A = 15$ and $\lambda_A = 0.023$ in the eq 6. The probability will be the relative intensity of the 16th isotope in the isotopic distribution of 200 H atoms.

Next, we use the simplifications that are provided in the sum of Poisson random variables. The sum of independent Poisson random variables is also a Poisson variable, with the mean equal to the sum of means of original Poisson variables. This is shown in many different ways, and a simple way to show this is to use the moment generating function of each Poisson random variable and the fact that the moment generating function of a sum of independent random variables is the product of their moment generating functions. The random variables corresponding to H, C, and N atoms are independent. We replace the first three terms in the eq 3 with a single Poisson random variable, X_{HCN}, with the mean, $\lambda_{\mathrm{HCN}} = \lambda_{\mathrm{H}} + \lambda_{\mathrm{C}} + \lambda_{\mathrm{N}}$.

Note that the Poisson distribution, unlike the binomial distribution, does not place a limit on the number successes (mini-isotopes, e.g., ^{2}H, ^{13}C, ^{15}N occurrences). In the latter, there cannot be more success than the number of trials. There is no such limit in Poisson distribution. So potentially an artificial outcome where the number of isotopes is larger than the number of atoms seems to be allowed. However, in practice, when the number of successes is more than the actual number of atoms of a given type, the probability is practically zero (long before reaching the limit). In addition, it is possible to limit the contribution to any isotope from any of the H, C, and N atoms by the corresponding number of atoms.

Multinomial distributions also allow a Poisson approximation.[25,26] For example, the distribution of minor (heavy) isotopes of oxygen (^{17}O and ^{18}O isotopes) is approximated with

$$P_{\mathrm{O}}(X_{^{17}\mathrm{O}}=k_1;X_{^{18}\mathrm{O}}=k_2)=\frac{\lambda_{^{17}\mathrm{O}}^{k_1}}{k_1!}e^{-\lambda_{^{17}\mathrm{O}}}\frac{\lambda_{^{18}\mathrm{O}}^{k_2}}{k_2!}e^{-\lambda_{^{18}\mathrm{O}}} \qquad (7)$$

The above approximation uses the fact the major isotope of oxygen has large abundance (probability) compared to the other two, minor isotope abundances. In eq 7, $\lambda_{^{17}\mathrm{O}} = p_{^{17}\mathrm{O}} N_{\mathrm{O}}$ is the mean of the Poisson distribution for the ^{17}O isotope, $p_{^{17}\mathrm{O}}$ is the frequency of this isotope in the nature, and N_{O} is the number of oxygen atoms in a molecule. Probability properties of the ^{18}O isotope are determined similarly.

Similar approach is extended to the isotope distribution resulting from sulfur atoms. The error of approximations, Δ_n, between the original multinomial distribution and its approximate Poisson distribution is proportional to the sum of the probabilities of minor isotopes:[26,27]

$$\Delta_n=\frac{\sum_{i=1}^{m}p_i}{\sqrt{2\pi e}}\left(1+O\left(\frac{1}{\sqrt{n\sum_{i=1}^{m}p_i}}\right)\right), \quad n\to\infty$$

where n is the number of atoms, O is a “big O”, e is the Euler’s constant, and the summation is over the elemental isotope probabilities that do not include the main isotope. Thus, for this approximation to work well, the sum of minor isotope relative abundances has to be small. In the case of the sulfur atom, the abundance of the second heavy isotope is relatively large, 0.0421, and the sum becomes relatively large and the accuracy of the approximation reduces. Note that the difference with the exact multinomial distribution is less than 1% (as computed using the probabilities of the heavy isotopes of S, and setting $n = 1$). For the sake of accuracy, it may be preferable to compute the sulfur isotope distribution separately using the multinomial probabilities. In the freely available C/C++ code (https://ispace.utmb.edu/users/rgsadygo/Proteomics/IsotopeDistributions), we have implemented both options. In this work, we have termed this approach as a hybrid (between exact for S atoms and Poisson approximation for H, C, N, and O atoms) approach.

In practical applications, the masses and relative abundances of the monoisotope ($i = 0$) and of the first heavy isotope ($i = 1$) require few numerical operations, and they can be computed exactly. The complexity of the isotope calculations arise for the isotopes with indices $i \geq 2$.

Once the Poisson approximations for H, C, N, O, and S atoms are used, we can further simplify the computations by recognizing that the first heavy isotopes from both O (^{17}O) and S (^{33}S) atoms have a common property with the H, C, and N isotopes. They all shift the mass value approximately by 1 Da. Therefore, both of these distributions are integrated into the common distribution with H, C, and N atoms. In addition, the second heavy isotopes of O (^{18}O) and S (^{34}S) can be combined to create a single Poisson distribution. At this point, eq 3 is transformed into the following equation:

$$X = X_{\mathrm{A}} + X_{^{18}\mathrm{O}^{34}\mathrm{S}} + X_{^{36}\mathrm{S}}$$

where X_{A} is a Poisson random variable which integrates the distribution of ^{2}H, ^{13}C, ^{15}N, ^{17}O, and ^{33}S isotopes, $X_{^{18}\mathrm{O}^{34}\mathrm{S}}$ is a Poisson random variable of joint O (^{18}O) and S (^{34}S) isotopes, and $X_{^{36}\mathrm{S}}$ is a Poisson random variable describing the distributions of ^{36}S isotope. No further simplifications are applicable as each term in eq 7 carries a unique mass shift.

The expected mass of an isotopomer, m_i, is calculated using features of the Poisson distribution:

$$\begin{aligned} m_i &= E[m_i] \\ &= m_0 + (\lambda_{\mathrm{A}} m_{\mathrm{A}} + \lambda_{^{18}\mathrm{O}^{34}\mathrm{S}} m_{^{18}\mathrm{O}^{34}\mathrm{S}} + \lambda_{^{36}\mathrm{S}} m_{^{36}\mathrm{S}}) * P(k_{\mathrm{A}}) * P(k_{^{18}\mathrm{O}^{34}\mathrm{S}}) * P(k_{^{36}\mathrm{S}}) \end{aligned}$$

The indices have the same meaning as in eq 7. In Figure 2, we provide an algorithm that implements the relative abundance and mass calculations of isotopes using Poisson approximation.

FFT-Based Computation of Isotope Distributions

In one application of an FFT-based isotope distribution computation, the isotope envelope of a basic unit (such as an atom’s isotopes) is coded by an array, X. If isotope fine structure is not a purpose, the arrays of all elements are one-dimensional. Each of the distributions is fast Fourier transformed to obtain the FFT of the elemental isotopes. Each one of the FFTs of elemental isotope distribution is raised to the corresponding power (the number of atoms of the particular type in the molecule). The product of all FFTs is obtained. The product is inverse Fourier transformed for the isotope distribution of a molecule. The equivalence between the convolution of two vectors and the inverse FFT of a product of forward FFTs of each vector is used. The FFT-based algorithms scale as O(K log(K)), where K is the length of a vector (in this case, the number of isotopes to be computed). The FFT approach to the isotope calculations has been described previously.[20] Recent publications have provided codes of applications of this algorithm in *R* (ecipex algorithm[28]) and in Matlab (a Matlab version[29] of the Taverna workflow in *R*). These applications are particularly suitable for calculations of isotope fine structure. The mass accuracy and resolution of the modern mass spectrometers continuous to improve. At high mass resolution, it is possible to discern isotope fine structure–isotope distributions of particular isotopologues (molecules that have the same chemical and isotope compositions and differ only on the position of the isotopes). It should be noted the mass resolution requirement to resolve isotope fine structure is higher for resolution of higher isotopologues (isotopologues having more than two or more nonabundant atomic isotopes). In the Supporting Information's Figure S1, we provide an algorithm and its C/C++ and *R* implementations.

Recursive Computation of Isotope Distributions

Another approach for generating isotope fine structures computes isotopologue distributions recursively.[30,31] This approach uses the fact that the binomial and multinomial probability mass functions can be computed recursively. For example, if two isotopologues (x and y) of an atom differ in exchange of one atomic isotope (assumed natural abundance, p_1) to another (assumed natural abundance, p_2), then the probability of new isotopologue, P_y, resulting from this transition is the same (in terms of P_x) for both binomial and multinomial distributions and is equal to

$$P_y = P_x * \frac{n_x^1}{n_y^2}\frac{p_2}{p_1} \quad (8)$$

where n_x^1 is the number of first isotope types in the isotopologue x, n_y^2 is the number of the second isotope types in isotopologue y. enviPat uses eq 8 to first estimate the highest probability atomic isotopes, then applies a user-adjustable filter to filter out low probability isotopologues.[30] An indexing scheme is used to avoid redundant transitions. The isotopologue probabilities of each atom type computed using eq 8 are combined via the Cartesian product to compute the isotopologue probabilities of molecules.

Note that the recursive calculations based on eq 8 use five arithmetic operations to determine probability of each isotopologues resulting from a single exchange of two different atomic isotopes. The analogous formula for Poisson distribution will have only three arithmetic operations in the case of atoms with only two isotopes (C, H, N):

$$P_y = P_x * \frac{\lambda}{n_y}$$

In the case of the three or more atomic isotopes (O and S), the transition probability formula is similar to the one in eq 8, with p_2 and p_1 replaced with λ_y and λ_x, respectively. Thus, in applications to recursive algorithms, there are computational gains in computing probabilities of isotopologues of atoms with two isotopes. For each atom the total number of isotopologues is computed using a binomial coefficient, $\binom{n+i-1}{n}$, where n is the number of atoms of a given type and i is the number of its isotopes.[32] The formula is obtained as a sampling from a set (isotopes of an atom) with replacement and without ordering. The number of isotopologues can become large, and pruning techniques are used filter out low probability isotopologues.[30,33]

RESULTS AND DISCUSSION

For our calculations, we used atomic isotope distributions recently published in the report of the Institute of Pure and Applied Chemistry.[34] In the report, there are two possible isotope distribution choices for each atomic element. One presents the range of values for each isotope, the second is the most accurately determined isotope ratios. We used the most accurately determined isotope ratio values. The natural isotope distributions used in this study are provided in the Supporting Information, Table S1.

To compare the computational speed, we applied Poisson-based and FFT-based isotope distribution to compute isotope distributions of all tryptic peptides in the UniProt database[35] downloaded in May of 2015. Three missed cleavages were allowed, and peptides ranging in length between 7 and 41 amino acids were considered. Only 16 isotopes for each tryptic peptide were calculated to exclude the possible effects of memory requirements in the FFT-based method. Theoretical isotope distributions were generated for more than 46 million peptide sequences from 548 454 protein entries. It took the Poisson-based method about 4.8 min to compute isotope distributions of all peptides. The FFT-based method took 14.1 min for the same task (on the same computer). In Figure 3, we show the density of the ratios of relative abundances of the first heavy isotopes as computed by FFT- and Poisson-based methods. As it is seen from this Figure, Poisson-based method accurately predicts the isotope distributions (median and mean of the deviations were the same and equal to 2%; the range of the accuracy was between 0.05 and 7%, and the mean and standard deviation were 0.15 and 0.09%, respectively).

In a separate comparison, we compared the time it took to compute isotope distributions of each peptide sequence by the FFT- and Poisson-based algorithms. Figure 4 shows the detailed comparison of isotope generation times by the FFT-and Poisson-based algorithms. It is a scatter plot of the times for all peptides with each methods. The red line is the line of unity. For less than 0.1% of peptides, the FFT-based algorithm was faster. For the rest of the peptides Poisson-based algorithm performed faster. The time comparison was done sequentially—for each peptide sequence, isotope distribution was first generated by one method and then the other one. We changed the ordering of methods to make sure that there was no dependence on the order in which a method was called. The general features of time distribution for the methods agreed with the overall time comparisons mentioned about (4.8 min versus 14.1 min). Thus, the median of isotope calculation time with the Poisson method was 0.0053 ms, and the FFT was 0.018 ms.

We have also compared the running times of our code with those of enviPat. enviPat is available as an *R* code. *R* is an interpretable environment, and normally codes take longer to run in this environment. For the comparison, we only recorded the time it takes to run the "isopattern" function in enviPat for atomic composition of each peptide. It took considerably longer to compute the isotope distributions in enviPat (Supporting Information, Figure S2). We believe that the fact the enviPat runs in *R* environment has contributed to the execution time difference. In addition, enviPat provide fine isotope structure, while our code provided only isotopomer distributions.

Breen et al.[36] have used Poisson distribution to predict the isotope distributions of peptides based on the peptide mass. For this purpose, they used the concept of averagine[37] amino acid, generated 15 peptides composed of one through 15 averagine amino acids, used a fitting to these data (15 data points) to determine the mean of the Poisson distribution. The mean was a linear function of peptide mass and this information can be used to predict isotope distribution of a peptide based only on peptide mass. Valkenborg et al.[38] extended this approach to include sulfur atom, as well. These Poisson model-based approaches predict the isotope distribution of a peptide species based on its mass only. They are important in different applications, for example, for peak picking or deconvolving overlapping isotope profiles[36] where the amino acid sequence of a peptide is unknown. Our approach will facilitate the computations of theoretical isotopes in high-throughput data processing of stable isotope labeling proteomics experiments such as experiments with metabolic labeling. As described above our approach has no restrictions in terms of the sulfur atoms.

Patterns of relative isotope abundances are regularly used in interrogation of mass spectrometric data, e.g., for relative quantification or proteome turnover studies.[39,40] Some of the application tools use information on both relative isotope abundances and isotope masses.[41] The calculations of the masses of isotopes are computationally the most challenging element of isotope generation. The algorithms for relative isotope abundance calculations (such as FFT-based algorithm) implicitly uses the fact that occupational numbers of isotopes are integer values. On the other hand, the masses of isotopes are not integers and therefore may need to be explicitly calculated. In general, there are eight, (3*1 + 2 + 3), embedded loops for computing isotope masses. This makes the computational problem of polynomial complexity. The fully Poisson approach described above avoids most of the embedded loops, and has only three embedded loops. However, the technique did not produce highly accurate masses for all isotopes of peptides. While for the peptides that did not contain sulfur atoms, the largest error in an isotope distribution (of all isotopes with relative abundance large than 0.001 of the tallest isotope peak in the isotope envelope) was less than 1 ppm for peptides that contained sulfur atoms, the mass accuracy was low as 10 ppm (for high mass isotopes). To address this problem, we tested a hybrid approach to simplify the isotope mass calculation and increase the mass accuracy of the calculations by including exact calculations for S atoms. In this approach, we, at first, compute the isotope masses of joint distribution of hydrogen, carbon, and nitrogen atoms using Poisson approximation for these atoms. At the second stage, we compute the isotope masses of joint distribution from the first stage and that of oxygen atoms also computed using Poisson approximation. At the third, final stage, we incorporate the mass shifts due to the isotopes of the sulfur atoms using an exact approach. In this approach, we have three embedded loops (one for each: combined isotope distributions of hydrogen, carbon, and nitrogen atoms; and the isotopes of sulfur atoms). As a result of this “decomposition” of isotope mass calculations into simplified calculations with smaller number of embedded loops, we obtain improved speed of computations. The mass accuracy from this approach was very high—better than 1 ppm for sulfur-containing peptides (the mean of difference was 0.06 ppm and standard deviation 0.012 ppm) and better than 1 part per billion (ppb) for peptides that do not contain a sulfur atom (the mean of difference was 0.4 ppb and standard deviation 0.02 ppb). The hybrid algorithm is shown in Figure 5. The results of its application and the scatter plot of mass accuracy (the worst accuracy for all isotopes that have relative abundance larger than 0.1% of the maximum abundance isotope) against the relative isotope abundance is illustrated in Figure 6.

In the case of Poisson approximation, as with many computational approaches, various combinations of approximations may be beneficial dependent on specific needs of an application. Here, we have tested a fully Poisson approximated and hybrid approach to generate isotope distributions of peptides from their amino acid compositions. If only relative abundances are needed, then the fully Poisson approximation is very fast and accurate. If masses of isotopes are also required, then (in particular, for high mass isotopes) a combination of Poisson approximations with exact convolution is a preferable approach.

CONCLUSIONS

We have developed an approximation to accurately generate isotope distributions of biomolecules. Our approach is based on the Poisson approximations to multinomial and binomial distributions. It uses a closed formula and has no requirements for memory storage. We show that the relative isotope abundances are calculated with a high accuracy (Figure 3). The relative isotope calculations are 3-fold faster than the ultrafast isotope calculations using FFT, which are also provided in the freely available code.

The mass accuracy of the masses of isotopomers is in low parts per million. To improve this accuracy, we have proposed and validated a hybrid approach which uses the Poisson approximation for H, C, N, and O atoms and exact calculations for the S atom. In the hybrid approach, we obtain mass accuracy better that 1 ppm.

We conclude that for peptides that do not contain a sulfur atom, the Poisson approximation provides very accurate computation of masses and relative abundances of isotopomers. For peptides containing sulfur atoms, a hybrid approach is suggested, where we use Poisson approximations for all atoms but the sulfur atom. The hybrid approach results in mass accuracy better than 1 ppm for all isotopes with abundance more than 1000th of the maximum abundance isotope (tallest isotope peak in an isotope envelope).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Research reported in this publication was supported in part by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM112044.

ABBREVIATIONS

Binom	binomial distribution
FFT	fast Fourier transform
invFFT	inverse Fourier transform
LC-MS	liquid chromatography–mass spectrometry
Multinom	multinomial distribution
MS	mass spectrometry
ppm	parts per million
ppb	parts per billion

References

1. Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. Nature. 2016; 537:347–355. [PubMed: 27629641]
2. Zhang Y, Fonslow BR, Shan B, Baek MC, Yates JR III. Protein analysis by shotgun/bottom-up proteomics. Chem Rev. 2013; 113:2343–2394. [PubMed: 23438204]
3. Kasumov T, Ilchenko S, Li L, Rachdaoui N, Sadygov RG, Willard B, McCullough AJ, Previs S. Measuring protein synthesis using metabolic (2)H labeling, high-resolution mass spectrometry, and an algorithm. Anal Biochem. 2011; 412:47–55. [PubMed: 21256107]
4. Kim TY, Wang D, Kim AK, Lau E, Lin AJ, Liem DA, Zhang J, Zong NC, Lam MP, Ping P. Metabolic labeling reveals proteome dynamics of mouse mitochondria. Mol Cell Proteomics. 2012; 11:1586–1594. [PubMed: 22915825]
5. Salisbury JP, Liu Q, Agar JN. QUDeX-MS: hydrogen/deuterium exchange calculation for mass spectra with resolved isotopic fine structure. BMC Bioinf. 2014; 15:403.
6. Naylor BC, Porter MT, Wilson E, Herring A, Lofthouse S, Hannemann A, Piccolo SR, Rockwood AL, Price JC. DeuteRater: a tool for quantifying peptide isotope precision and kinetic proteomics. Bioinformatics. 2017; 33:1514–1520. [PubMed: 28093409]
7. Jehmlich N, Fetzer I, Seifert J, Mattow J, Vogt C, Harms H, Thiede B, Richnow HH, von Bergen M, Schmidt F. Decimal place slope, a fast and precise method for quantifying 13C incorporation levels for detecting the metabolic activity of microbial species. Mol Cell Proteomics. 2010; 9:1221–1227. [PubMed: 20064840]
8. Guan S, Price JC, Prusiner SB, Ghaemmaghami S, Burlingame AL. A data processing pipeline for mammalian proteome dynamics studies using stable isotope metabolic labeling. Mol Cell Proteomics. 2011; 10:M111.010728.

9. Vogt JA, Schroer K, Holzer K, Hunzinger C, Klemm M, Biefang-Arndt K, Schillo S, Cahill MA, Schrattenholz A, Matthies H, Stegmann W. Protein abundance quantification in embryonic stem cells using incomplete metabolic labelling with 15N amino acids, matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry, and analysis of relative isotopologue abundances of peptides. Rapid Commun Mass Spectrom. 2003; 17:1273–1282. [PubMed: 12811750]

10. Rauniyar N, McClatchy DB, Yates JR 3rd. Stable isotope labeling of mammals (SILAM) for in vivo quantitative proteomic analysis. Methods. 2013; 61:260–268. [PubMed: 23523555]

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. Li L, Willard B, Rachdaoui N, Kirwan JP, Sadygov RG, Stanley WC, Previs S, McCullough AJ, Kasumov T. Plasma proteome dynamics: analysis of lipoproteins and acute phase response proteins with 2H2O metabolic labeling. Mol Cell Proteomics. 2012; 11:M111.014209.

13. Wu CC, MacCoss MJ, Howell KE, Matthews DE, Yates JR III. Metabolic labeling of mammalian organisms with stable isotopes for quantitative proteomic analysis. Anal Chem. 2004; 76:4951–4959. [PubMed: 15373428]

14. Zhang T, Price JC, Nouri-Nigjeh E, Li J, Hellerstein MK, Qu J, Ghaemmaghami S. Kinetics of precursor labeling in stable isotope labeling in cell cultures (SILAC) experiments. Anal Chem. 2014; 86:11334–11341. [PubMed: 25301408]

15. Ahmed Z, Zeeshan S, Huber C, Hensel M, Schomburg D, Munch R, Eisenreich W, Dandekar T. Software LS-MIDA for efficient mass isotopomer distribution analysis in metabolic modelling. BMC Bioinf. 2013; 14:218.

16. Sperling E, Bunner AE, Sykes MT, Williamson JR. Quantitative analysis of isotope distributions in proteomic mass spectrometry using least-squares Fourier transform convolution. Anal Chem. 2008; 80:4906–4917. [PubMed: 18522437]

17. Yergey J, Heller D, Hansen G, Cotter RJ, Fenselau C. Isotopic Distributions in Mass-Spectra of Large Molecules. Anal Chem. 1983; 55:353–356.

18. Yergey JAA. General-Approach to Calculating Isotopic Distributions for Mass-Spectrometry. Int J Mass Spectrom Ion Phys. 1983; 52:337–349.

19. Rockwood AL, Van Orden SL, Smith RD. Rapid Calculation of Isotope Dsitributions. Anal Chem. 1995; 67:2699–2704.

20. Rockwood AL, VanOrden SL. Ultrahigh-speed calculation of isotope distributions. Anal Chem. 1996; 68:2027–2030. [PubMed: 21619291]

21. Bocker S, Letzel MC, Liptak Z, Pervukhin A. SIRIUS: decomposing isotope patterns for metabolite identification. Bio-informatics. 2009; 25:218–224.

22. Claesen J, Dittwald P, Burzykowski T, Valkenborg D. An efficient method to calculate the aggregated isotopic distribution and exact center-masses. J Am Soc Mass Spectrom. 2012; 23:753–763. [PubMed: 22351289]

23. Lee WN, Byerley LO, Bergner EA, Edmond J. Mass isotopomer analysis: theoretical and practical considerations. Biol Mass Spectrom. 1991; 20:451–458. [PubMed: 1768701]

24. Feller, W. An Introduction to Probability Theory and Its Applications. 3. Vol. I. Wiley & Sons; 1968.

25. Arenbaev NK. Asymptotic-Behavior of Multinomial Distribution. Theory Probab Its Appl. 1977; 21:805–810.

26. Deheuvels P, Pfeifer D. Poisson Approximations of Multinomial Distributions and Point-Processes. J Multivariate Anal. 1988; 25:65–89.

27. McDonald DR. On the Poisson Approximation to the multinomial distribution. Canadian Journal of Statistics. 1980; 8:115–118.

28. Ipsen A. Efficient calculation of exact fine structure isotope patterns via the multidimensional Fourier transform. Anal Chem. 2014; 86:5316–5322. [PubMed: 24841326]

29. Rockwood AL, Palmblad M. Isotopic distributions. Methods Mol Biol. 2013; 1007:65–99. [PubMed: 23666722]

30. Loos M, Gerber C, Corona F, Hollender J, Singer H. Accelerated isotope fine structure calculation using pruned transition trees. Anal Chem. 2015; 87:5738–5744. [PubMed: 25929282]

31. Li L, Karabacak NM, Cobb JS, Wang Q, Hong P, Agar JN. Memory-efficient calculation of the isotopic mass states of a molecule. Rapid Commun Mass Spectrom. 2010; 24:2689–2696. [PubMed: 20814974]

32. Sadygov RG. Using SEQUEST with Theoretically Complete Sequence Databases. J Am Soc Mass Spectrom. 2015; 26:1858–1864. [PubMed: 26238326]

33. Lacki MK, Startek M, Valkenborg D, Gambin A. IsoSpec: Hyperfast Fine Structure Calculator. Anal Chem. 2017; 89:3272–3277. [PubMed: 28234451]

34. Meija J, Coplen TB, Berglund M, Brand WA, De Bievre P, Groning M, Holden NE, Irrgeher J, Loss RD, Walczyk T, Prohaska T. Isotopic compositions of the elements 2013 (IUPAC Technical Report). Pure Appl Chem. 2016; 88:293–306.

35. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. The Universal Protein Resource (UniProt). Nucleic Acids Res. 2004; 33:D154–D159.

36. Breen EJ, Hopwood FG, Williams KL, Wilkins MR. Automatic poisson peak harvesting for high throughput protein identification. Electrophoresis. 2000; 21:2243–2251. [PubMed: 10892735]

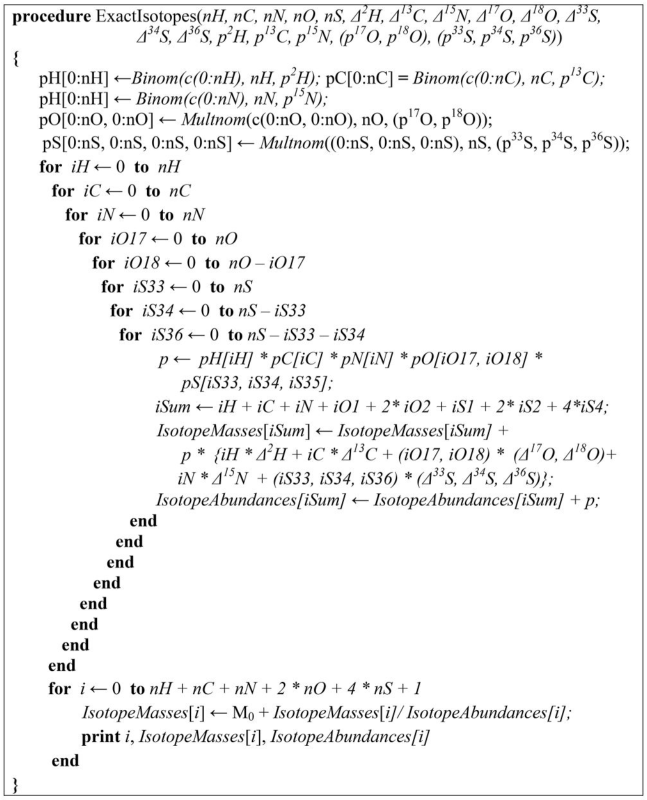
37. Senko MW, Beu SC, McLaffertycor FW. Determination of Monoisotopic Massed and Ion Populations for Large Biomolecules from Resolved Isotopic Distributions. J Am Soc Mass Spectrom. 1995; 6:229–233. [PubMed: 24214167]

38. Valkenborg D, Assam P, Thomas G, Krols L, Kas K, Burzykowski T. Using a Poisson approximation to predict the isotopic distribution of sulphur-containing peptides in a peptide-centric proteomic approach. Rapid Commun Mass Spectrom. 2007; 21:3387–3391. [PubMed: 17891751]

39. Rahman M, Previs SF, Kasumov T, Sadygov RG. Gaussian Process Modeling of Protein Turnover. J Proteome Res. 2016; 15:2115–2122. [PubMed: 27229456]

40. Lam MP, Wang D, Lau E, Liem DA, Kim AK, Ng DC, Liang X, Bleakley BJ, Liu C, Tabaraki JD, Cadeiras M, Wang Y, Deng MC, Ping P. Protein kinetic signatures of the remodeling heart following isoproterenol stimulation. J Clin Invest. 2014; 124:1734–1744. [PubMed: 24614109]

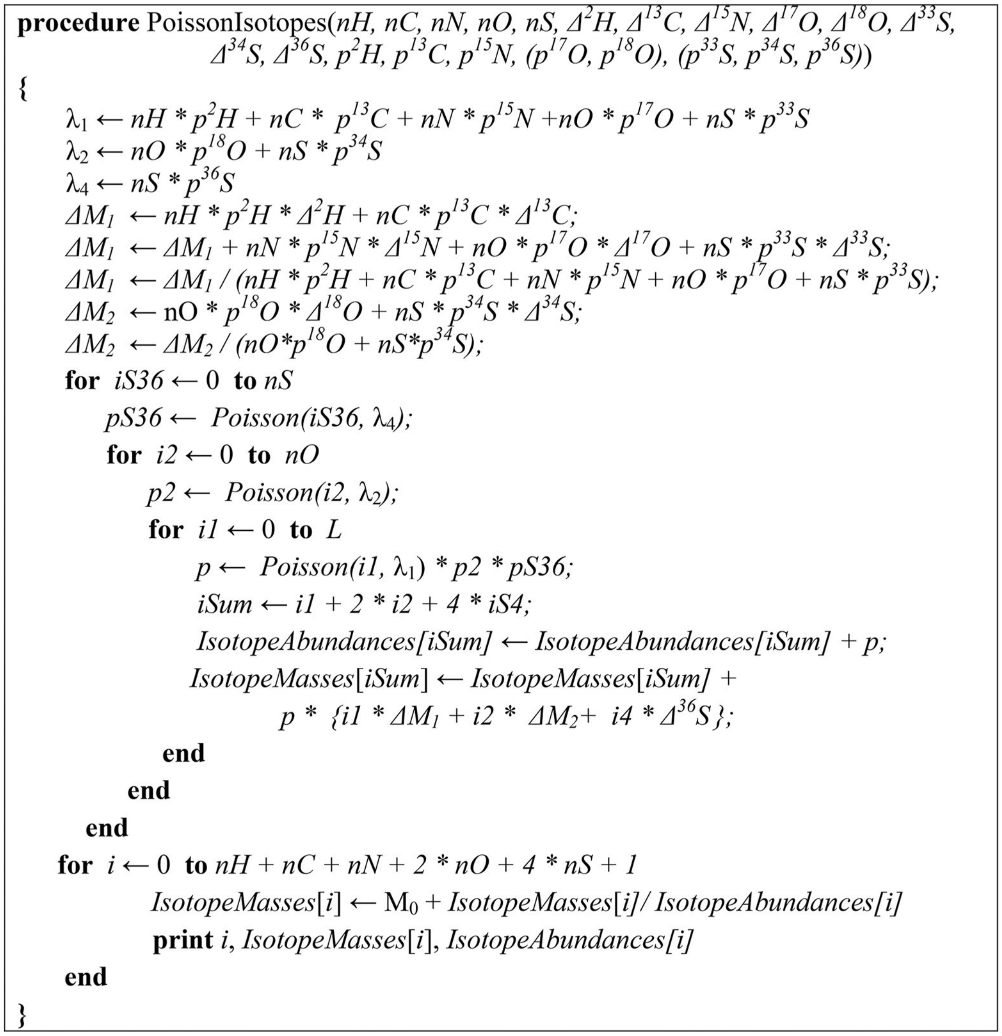
41. Naylor BC, Porter MT, Wilson E, Herring A, Lofthouse S, Hannemann A, Piccolo SR, Rockwood AL, Price JC. DeuteRater: a Tool for Quantifying Peptide Isotope Precision and Kinetic Proteomics. Bioinformatics. 2017:btx009.



```
procedure ExactIsotopes($nH, nC, nN, nO, nS, \Delta^{2}H, \Delta^{13}C, \Delta^{15}N, \Delta^{17}O, \Delta^{18}O, \Delta^{33}S,$
                 $\Delta^{34}S, \Delta^{36}S, p^{2}H, p^{13}C, p^{15}N, (p^{17}O, p^{18}O), (p^{33}S, p^{34}S, p^{36}S))$
{
    pH[0:nH] ←Binom(c(0:nH), nH, $p^{2}H$); pC[0:nC] = Binom(c(0:nC), nC, $p^{13}C$);
    pH[0:nH] ← Binom(c(0:nN), nN, $p^{15}N$);
    pO[0:nO, 0:nO] ← Multnom(c(0:nO, 0:nO), nO, ($p^{17}O$, $p^{18}O$));
    pS[0:nS, 0:nS, 0:nS, 0:nS] ← Multnom((0:nS, 0:nS, 0:nS), nS, ($p^{33}S$, $p^{34}S$, $p^{36}S$));
    for iH ← 0 to nH
      for iC ← 0 to nC
        for iN ← 0 to nN
          for iO17 ← 0 to nO
            for iO18 ← 0 to nO – iO17
              for iS33 ← 0 to nS
                for iS34 ← 0 to nS – iS33
                  for iS36 ← 0 to nS – iS33 – iS34
                        p ← pH[iH] * pC[iC] * pN[iN] * pO[iO17, iO18] *
                            pS[iS33, iS34, iS35];
                        iSum ← iH + iC + iN + iO1 + 2* iO2 + iS1 + 2* iS2 + 4*iS4;
                        IsotopeMasses[iSum] ← IsotopeMasses[iSum] +
                            p * {iH * $\Delta^{2}H$ + iC * $\Delta^{13}C$ + (iO17, iO18) * ($\Delta^{17}O$, $\Delta^{18}O$)+
                           iN * $\Delta^{15}N$ + (iS33, iS34, iS36) * ($\Delta^{33}S$, $\Delta^{34}S$, $\Delta^{36}S$)};
                        IsotopeAbundances[iSum] ← IsotopeAbundances[iSum] + p;
                    end
                  end
                end
              end
            end
          end
        end
      end
      for i ← 0 to nH + nC + nN + 2 * nO + 4 * nS + 1
          IsotopeMasses[i] ← $M_0$ + IsotopeMasses[i]/ IsotopeAbundances[i];
          print i, IsotopeMasses[i], IsotopeAbundances[i]
      end
}
```

Figure 1.

Pseudocode of algorithm to compute exact isotope distributions. The inputs are the relevant characteristics of each atom. For example, nO is the number of O atoms, $\Delta^{17}O$ and $\Delta^{18}O$ are the mass shifts of ^{17}O and ^{18}O isotopes (relative to ^{16}O isotope), respectively, and $p^{17}O$ and $p^{18}O$ are the corresponding (normalized) abundances. M_0 is the monoisotopic mass. Binom and Multinom stand for binomial and multinomial distributions, respectively.



```
procedure PoissonIsotopes($nH, nC, nN, nO, nS, \Delta^2H, \Delta^{13}C, \Delta^{15}N, \Delta^{17}O, \Delta^{18}O, \Delta^{33}S,$
                          $\Delta^{34}S, \Delta^{36}S, p^2H, p^{13}C, p^{15}N, (p^{17}O, p^{18}O), (p^{33}S, p^{34}S, p^{36}S)$)
{
    $\lambda_1 \leftarrow nH * p^2H + nC * p^{13}C + nN * p^{15}N + nO * p^{17}O + nS * p^{33}S$
    $\lambda_2 \leftarrow nO * p^{18}O + nS * p^{34}S$
    $\lambda_4 \leftarrow nS * p^{36}S$
    $\Delta M_1 \leftarrow nH * p^2H * \Delta^2H + nC * p^{13}C * \Delta^{13}C;$
    $\Delta M_1 \leftarrow \Delta M_1 + nN * p^{15}N * \Delta^{15}N + nO * p^{17}O * \Delta^{17}O + nS * p^{33}S * \Delta^{33}S;$
    $\Delta M_1 \leftarrow \Delta M_1 / (nH * p^2H + nC * p^{13}C + nN * p^{15}N + nO * p^{17}O + nS * p^{33}S);$
    $\Delta M_2 \leftarrow \mathrm{nO} * p^{18}O * \Delta^{18}O + nS * p^{34}S * \Delta^{34}S;$
    $\Delta M_2 \leftarrow \Delta M_2 / (nO*p^{18}O + nS*p^{34}S);$
    for $iS36 \leftarrow 0$ to $nS$
        $pS36 \leftarrow Poisson(iS36, \lambda_4);$
        for $i2 \leftarrow 0$ to $nO$
            $p2 \leftarrow Poisson(i2, \lambda_2);$
            for $i1 \leftarrow 0$ to $L$
                $p \leftarrow Poisson(i1, \lambda_1) * p2 * pS36;$
                $iSum \leftarrow i1 + 2 * i2 + 4 * iS4;$
                $IsotopeAbundances[iSum] \leftarrow IsotopeAbundances[iSum] + p;$
                $IsotopeMasses[iSum] \leftarrow IsotopeMasses[iSum] +$
                        $p * \{i1 * \Delta M_1 + i2 * \Delta M_2 + i4 * \Delta^{36}S\};$
            end
        end
    end
    for $i \leftarrow 0$ to $nH + nC + nN + 2 * nO + 4 * nS + 1$
        $IsotopeMasses[i] \leftarrow M_0 + IsotopeMasses[i] / IsotopeAbundances[i]$
        print $i, IsotopeMasses[i], IsotopeAbundances[i]$
    end
}
```

Figure 2.

Pseudocode of algorithm to compute isotope distributions using Poisson approximations. Poisson denotes a Poisson probability with the specified parameters. The rest of the nomenclature is the same as that in Figure 1.

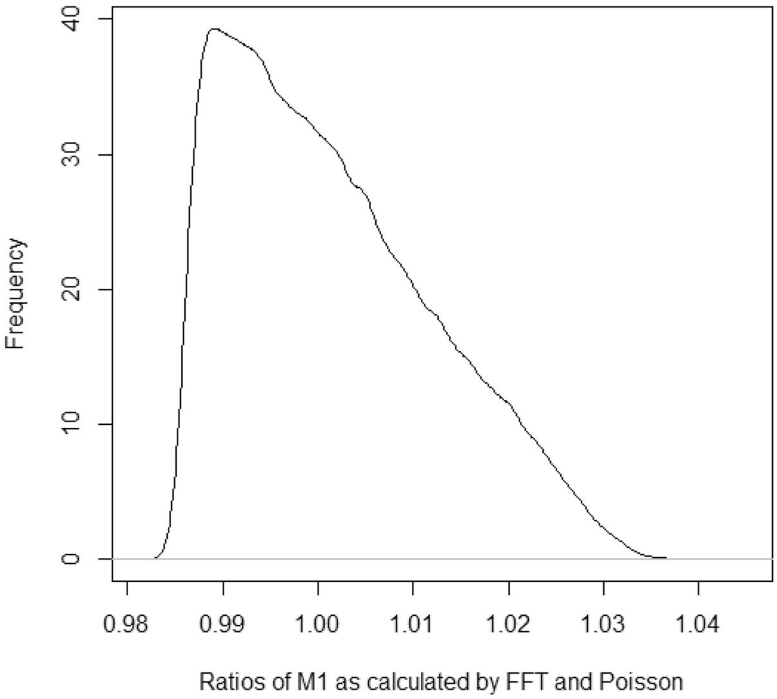


Figure 3.

Ratios of M_1 intensities as calculated by Poisson to FFT.

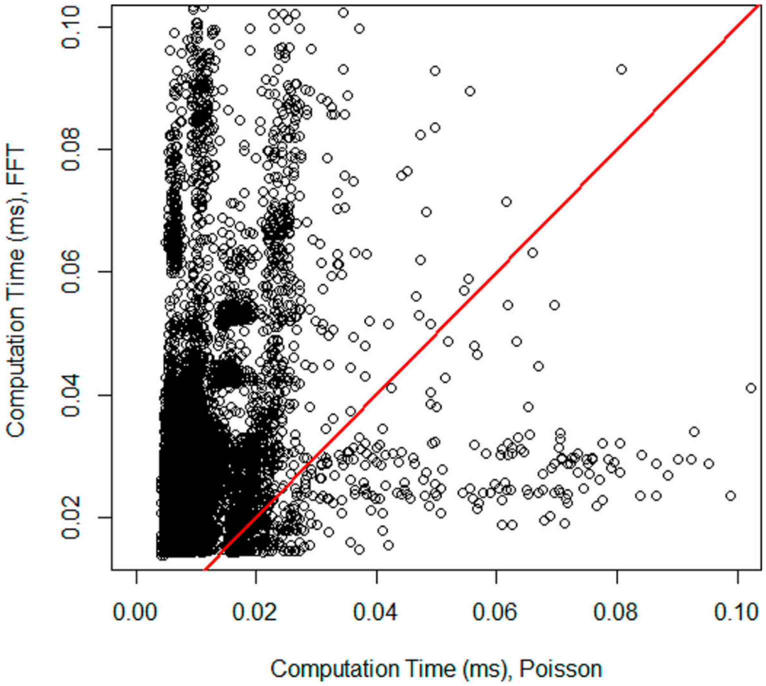
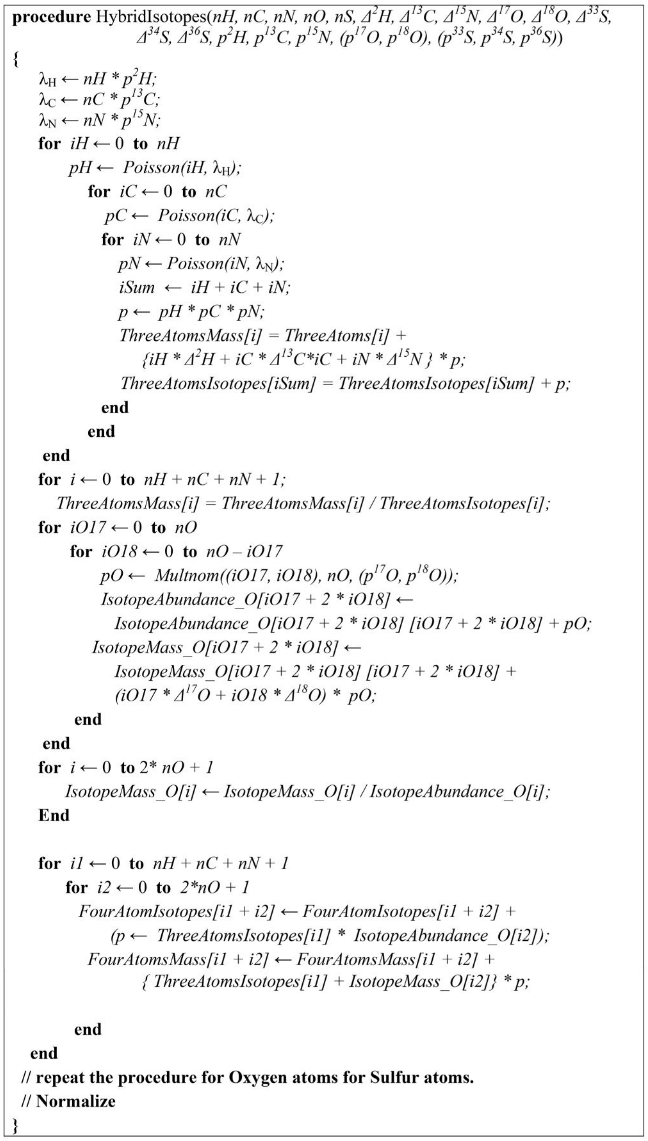


Figure 4.

Scatter plot of computation times of isotope generation using FFT-based (*y*-axis) and Poisson-based (*x*-axis) methods. The red line is the line of unity.



```
procedure HybridIsotopes($nH, nC, nN, nO, nS, \Delta^2H, \Delta^{13}C, \Delta^{15}N, \Delta^{17}O, \Delta^{18}O, \Delta^{33}S,$
                    $\Delta^{34}S, \Delta^{36}S, p^2H, p^{13}C, p^{15}N, (p^{17}O, p^{18}O), (p^{33}S, p^{34}S, p^{36}S)$)
{
    $\lambda_H \leftarrow nH * p^2H;$
    $\lambda_C \leftarrow nC * p^{13}C;$
    $\lambda_N \leftarrow nN * p^{15}N;$
    for $iH \leftarrow 0$ to $nH$
        $pH \leftarrow Poisson(iH, \lambda_H);$
            for $iC \leftarrow 0$ to $nC$
              $pC \leftarrow Poisson(iC, \lambda_C);$
              for $iN \leftarrow 0$ to $nN$
                $pN \leftarrow Poisson(iN, \lambda_N);$
                $iSum \leftarrow iH + iC + iN;$
                $p \leftarrow pH * pC * pN;$
                $ThreeAtomsMass[i] = ThreeAtoms[i] +$
                    $\{iH * \Delta^2H + iC * \Delta^{13}C*iC + iN * \Delta^{15}N\} * p;$
                $ThreeAtomsIsotopes[iSum] = ThreeAtomsIsotopes[iSum] + p;$
              end
            end
     end
    for $i \leftarrow 0$ to $nH + nC + nN + 1;$
      $ThreeAtomsMass[i] = ThreeAtomsMass[i] / ThreeAtomsIsotopes[i];$
    for $iO17 \leftarrow 0$ to $nO$
        for $iO18 \leftarrow 0$ to $nO - iO17$
            $pO \leftarrow Multnom((iO17, iO18), nO, (p^{17}O, p^{18}O));$
            $IsotopeAbundance\_O[iO17 + 2 * iO18] \leftarrow$
              $IsotopeAbundance\_O[iO17 + 2 * iO18] [iO17 + 2 * iO18] + pO;$
           $IsotopeMass\_O[iO17 + 2 * iO18] \leftarrow$
              $IsotopeMass\_O[iO17 + 2 * iO18] [iO17 + 2 * iO18] +$
              $(iO17 * \Delta^{17}O + iO18 * \Delta^{18}O) * pO;$
         end
     end
    for $i \leftarrow 0$ to $2* nO + 1$
       $IsotopeMass\_O[i] \leftarrow IsotopeMass\_O[i] / IsotopeAbundance\_O[i];$
    End

    for $i1 \leftarrow 0$ to $nH + nC + nN + 1$
       for $i2 \leftarrow 0$ to $2*nO + 1$
         $FourAtomIsotopes[i1 + i2] \leftarrow FourAtomIsotopes[i1 + i2] +$
             $(p \leftarrow ThreeAtomsIsotopes[i1] * IsotopeAbundance\_O[i2]);$
          $FourAtomsMass[i1 + i2] \leftarrow FourAtomsMass[i1 + i2] +$
                $\{ThreeAtomsIsotopes[i1] + IsotopeMass\_O[i2]\} * p;$

        end
   end
 // repeat the procedure for Oxygen atoms for Sulfur atoms.
 // Normalize
}
```

Figure 5.

Pseudocode of a “hybrid” algorithm to compute isotope distributions using Poisson approximation and exact distributions. The nomenclature is the same as that in Figures 1 and 2. For brevity, we did not show the computations for S atoms (they are similar to the O atom calculations).

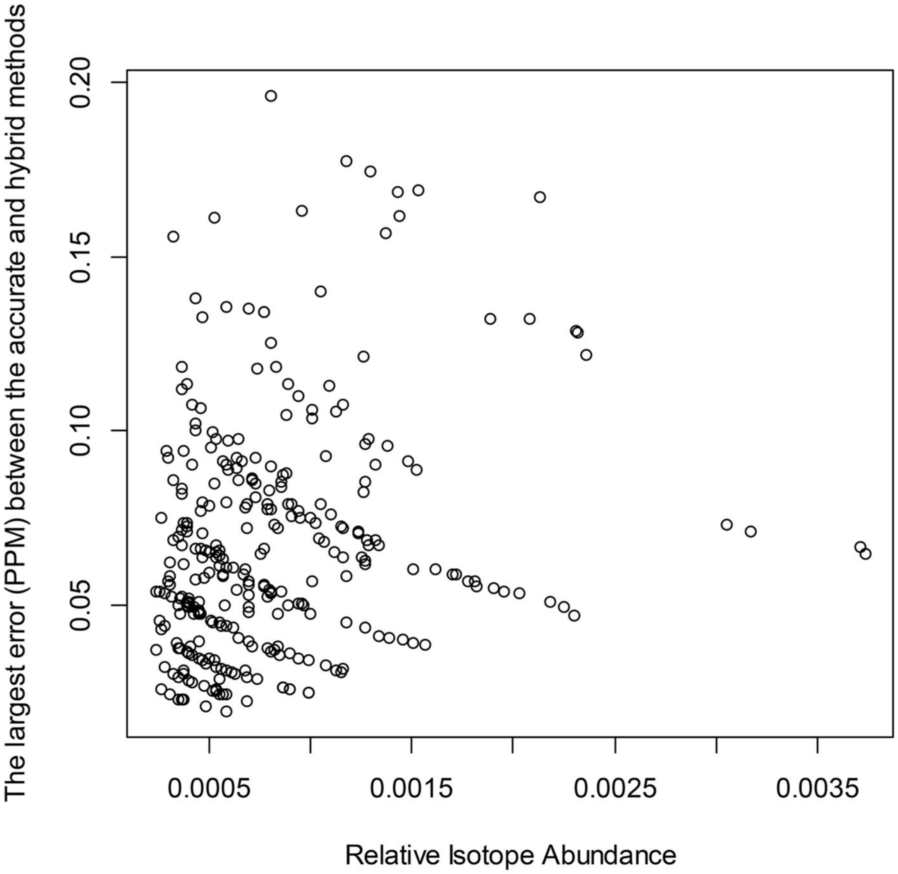


Figure 6.

Scatter plot of mass accuracy. Shown are the worst accuracy in each peptide's isotope cluster. The only filtering used was the requirement that an isotopomer abundance needed to be larger than 0.1% of the maximum abundance isotope for the particular peptide under consideration. Data for 1000 randomly chosen peptides are shown. The *x* axis shows the absolute values of the relative abundances. The sum of all isotope abundances is normalized to equal 1.