An efficient method to calculate the aggregated isotopic distribution and exact center-masses

Jürgen Claesen[1,*], Piotr Dittwald[2,3,*], Tomasz Burzykowski[1], Dirk Valkenborg[1,4,5,†]

[1] I-BioStat, Hasselt University, Belgium
[2] Institute of Informatics, University of Warsaw, Poland
[3] College of Inter-Faculty Individual Studies in Mathematics and Natural Sciences, University of Warsaw, Poland
[4] Flemish Institute for Technological Research (VITO), Belgium
[5] CfP-CeProMa, University of Antwerp, Belgium

Address reprint requests to Dirk Valkenborg, Boeretang 200, 2400 Mol, Belgium, +32 14 33 50 43, dirk.valkenborg@vito.be

*Both authors contributed equally to this manuscript.
†Corresponding author: dirk.valkenborg@vito.be.

Abstract

In this manuscript, we present a computation- and memory-efficient method to calculate the probabilities of occurrence and exact center-masses of the aggregated isotopic distribution of a molecule. The method uses fundamental mathematical properties of polynomials given by the Newton-Girard theorem and Viete's formulae. The calculation is based on the atomic composition of the molecule and the natural abundances of the elemental isotopes in normal terrestrial matter. To evaluate the performance of the proposed method, which we named *BRAIN*, we compare it to the results obtained from five existing packages (*IsoPro*, *Mercury*, *Emass*, *NeutronCluster*, and *IsoDalton*) for ten biomolecules. Additionally, we compare the computed mass centers with the results obtained by calculating, and subsequently aggregating, the fine isotopic distribution for two of the exemplary biomolecules. The algorithm will be made available as a Bioconductor package [6], and is also available upon request.

1 Introduction

The *isotopic distribution* is an important, but often forgotten, concept in the field of mass spectrometry (MS). Yet, it is particularly useful for the interpretation of the complex patterns observed in mass spectral data. For example, a peptide molecule visualized by MS should exhibit a characteristic signal in the form of series of regularly spaced peaks of a specific profile. The profile is related to the isotopic distribution of the peptide. Prior knowledge about the distribution can thus be used to develop strategies for searching for the profile in the spectra and, hence, for efficient processing of the spectral information [21, 22, 3, 5, 19]. Another application can be found in the field of metabolomics. For example, a comparison of the observed pattern of peaks in a mass spectrum with a set of hypothesized isotopic distributions from moieties with a similar mass as the observed molecule can be used to construct a confidence score for the identification.

The isotopic distribution reflects the number and probabilities of occurrence of different *isotopic variants* of a molecule. The occurrence probabilities are reflected in the mass spectrum by the relative heights of the series of peaks related to the molecule; whilst the different variants result from the fact that there are different isotopes of chemical elements.

Every isotopic variant of a molecule has, in principle, a different mass[‡]. If we ignore the small deviations of the masses from integer values, we can define *aggregated isotopic variants* of a molecule, with masses differing approximately by 1 Da. The *aggregated isotopic distribution* provides the number and occurrence probabilities for the aggregated isotopic variants. In fact, given the finite resolution of mass spectrometers, the profile of peak heights observed in a spectrum for a molecule is directly related to the aggregated isotopic distribution.

The calculation of the (aggregated) isotopic distribution for a molecule of a known atomic composition is thus a relevant and important problem. Several methods have already been proposed to this aim. In the early sixties of the 20th century, Biemann suggested a step-wise procedure [2]. In the late seventies, Yamamoto and McCloskey [24] and Brownawell and Fillippo [4] argued that, for large molecules, the isotopic distribution could be easily obtained by symbolically expanding a polynomial function. Later in the eighties, Yergey and colleagues [25, 26] generalized the concept of polynomial expansion to a multinomial expansion. In the nineties, Rockwood and co-workers propagated the use of the convolution [12]. An overview of the different procedures to calculate isotopic distributions has been recently provided by Valkenborg *et al.* [23].

A vital element in the calculation of aggregated isotopic distributions is the assignment of the center-masses to the aggregated isotope variants. To this aim, the center-mass is calculated as a probability-weighted sum of the masses of the isotopic variants that contribute to an aggregated variant, as defined by Roussis and Proulx [17]. The accuracy of this mass calculation depends on the number of isotopic variants accounted for. Rockwood *et al.* [14] solved this problem by a linear transformation based on the average mass and standard deviation of the isotopic distribution to acquire semi-accurate masses. In a later paper, Rockwood and colleagues focussed

[‡]In this terminology, we ignore location isomers, e.g., ${}^{12}C{}^{12}C{}^{13}C$ and ${}^{13}C{}^{12}C{}^{12}C$, which obviously do have the same mass.

on the accurate mass calculation of a pre-selected aggregated isotope variant [15, 13]. Another solution to aforementioned problem was proposed by Olson and Yergey [10], who developed the idea of using equatransneutronic isotopes. The method, however, induces some error in the mass assignments of the aggregated isotope variants. To overcome this inaccuracy, Olson and Yergey proposed to estimate the error and to account for it in the calculation of the center-masses.

In this manuscript, we present an alternate, computation- and memory-efficient method to calculate the probabilities of occurrence and exact center-masses of the aggregated isotopic distribution of a molecule. The calculation is based on the atomic composition of the molecule and the natural abundances of stable elemental isotopes in normal terrestrial matter [16]. Note that this excludes unstable radio-isotopes, and that our use of the term "exact center-masses" is conditional on this assumption. Our method, which we name *BRAIN* (**B**affling **R**ecursive **A**lgorithm for **I**sotopic distributio**N** calculations), allows computing the exact center-masses, because it accumulates the mass information along a recursive calculation of the aggregated isotopic distribution. The algorithm will be made available as a Bioconductor package [6], and is also available upon request.

To evaluate the performance of the proposed method, we compare it to the results obtained from five existing packages (*IsoPro* [18], *Mercury* [14], *Emass* [13], *NeutronCluster* [10], and *IsoDalton* [20]) for ten biomolecules. Additionally, we compare the computed exact mass-centers with the results obtained by the calculation of the fine isotopic distribution and subsequent aggregation of this distribution by the method of Roussis and Proulx [17] for two of the ten exemplary biomolecules.

For the purposes of the current manuscript, we restrict the calculation of the isotopic distribution to molecules containing only carbon (C), nitrogen (N), hydrogen (H), oxygen (O), and sulphur (S), unless specified otherwise. The most-abundant (and lightest) isotopes for the latter elements are ^{12}C, ^{1}H, ^{14}N, ^{16}O, and ^{32}S. A molecule composed out of only these elemental isotopes is called the *monoisotopic variant*. In addition, we only consider stable isotopes, that is, the isotopes just mentioned, together with ^{13}C, ^{2}H, ^{15}N, ^{17}O, ^{18}O, ^{33}S, ^{34}S, and ^{36}S. Extending the presented algorithm to molecules containing other poly-isotopic elements is straightforward.

2 Methods

Yamamoto *et al.* [24] and Brownawell *et al.* [4] argued that, for large molecules, the isotopic forms could be easily obtained by symbolically expanding a polynomial function. In the case of proteins or peptides with a composition $C_vH_wN_xO_yS_z$, this polynomial takes the following form:

$$\begin{aligned}&(^{12}C + {}^{13}C)^v \times (^{1}H + {}^{2}H)^w \times (^{14}N + {}^{15}N)^x \times \\ &(^{16}O + {}^{17}O + {}^{18}O)^y \times (^{32}S + {}^{33}S + {}^{34}S + {}^{36}S)^z \,. \end{aligned} \tag{1}$$

Symbolic expansion of (1) results in many product terms, which correspond to different isotopic variants of a molecule. By substituting the probabilities of occurrence for $^{12}C, {}^{13}C, \ldots, {}^{36}S$ from Table 1 in each term, the prevalence of the variants of the peptide could be obtained.

Table 1: List of stable isotopes for carbon, hydrogen, nitrogen, oxygen, and sulphur. Source: IUPAC 1997 standard [16].

Isotope	Mass (ma/u)	Abundance (%)	Isotope	Mass (ma/u)	Abundance (%)
^{12}C	12.0000000000	98.93	^{16}O	15.9949146	99.757
^{13}C	13.0033548378	1.07	^{17}O	16.9991312	0.038
^{1}H	1.0078250321	99.9885	^{18}O	17.9991603	0.205
^{2}H	2.0141017780	0.0115	^{32}S	31.97207070	94.93
^{14}N	14.0030740052	99.632	^{33}S	32.97145843	0.76
^{15}N	15.0001088984	0.368	^{34}S	33.96786665	4.29
			^{36}S	35.96708062	0.02

Given that the deviations of the masses of the isotopes of C, N, H, O, and S from integer values are different (see Table 1), every isotopic variant of a molecule has, in principle, a different mass. By ignoring the small deviations, we obtain the aggregated isotopic variants, with masses differing by approximately 1 Da. The aggregated variants are represented in the expansion of (1) by multiple product terms. To identify these components more explicitly, we introduce in (1) an indicator variable I. The introduction explicitly expresses the calculation of the isotopic distribution in terms of the additional neutron content, i.e., as an aggregated isotopic distribution. The modified form of (1) is given as follows:

$$\begin{aligned} Q(I;v,w,x,y,z) = \left(P_{C_{12}}I^0 + P_{C_{13}}I^1\right)^v \quad &\times \quad \left(P_{H_1}I^0 + P_{H_2}I^1\right)^w \times \left(P_{N_{14}}I^0 + P_{N_{15}}I^1\right)^x \times \\ \left(P_{O_{16}}I^0 + P_{O_{17}}I^1 + P_{O_{18}}I^2\right)^y \quad &\times \quad \left(P_{S_{32}}I^0 + P_{S_{33}}I^1 + P_{S_{34}}I^2 + P_{S_{36}}I^4\right)^z \ , \end{aligned} \tag{2}$$

where $P_{C_{12}}$, $P_{C_{13}}$, ..., $P_{S_{36}}$ represent the natural abundances (probabilities of occurrence) of the isotopes of carbon, hydrogen, nitrogen, oxygen, and sulphur in normal terrestrial matter, as displayed in Table 1. Note that the power of the symbolic indicator I represents the additional neutron content (or discrete mass shift) with respect to the monoisotopic variant. This indicator serves a book keeping-device to keep track of the different aggregated isotopic variants.

It should be stressed that equation (2) makes abstraction of the mass information, as the aggregated isotopic variants are presented by their additional neutron count. Later in the manuscript we discuss how the exact center-masses can be calculated.

In what follows, we will also be referring to the following, abbreviated form of (2):

$$Q(I;v,w,x,y,z) = \{Q_C(I)\}^v \times \{Q_H(I)\}^w \times \{Q_N(I)\}^x \times \{Q_O(I)\}^y \times \{Q_S(I)\}^z \ , \tag{3}$$

with $Q_C(I) = \left(P_{C_{12}}I^0 + P_{C_{13}}I^1\right)$, etc.

Generally, the expansion of the polynomial (2) can be written as

$$Q(I;v,w,x,y,z) \equiv \sum_{j=0}^{n} q_j I^j \ , \tag{4}$$

where $n = v + w + x + 2y + 4z$ is a function of the atomic composition of the molecule. The coefficient q_j represents the occurrence probability of the j-th aggregated isotopic variant of

the molecule. Hence, the problem of calculating the aggregated isotopic distribution may be reformulated as the problem of finding values of the coefficients $q_0, q_1, \ldots, q_n$ of the expanded polynomial (4).

To clarify the role of the polynomial (2), consider a very simple example of ozone (O_3). For this molecule, the polynomial takes the following form:

$$Q(I;0,0,0,3,0) = \left(P_{O_{16}} I^0 + P_{O_{17}} I^1 + P_{O_{18}} I^2\right)^3 = \{Q_O(I)\}^3 = \sum_{j=0}^{6} q_j I^j \ , \tag{5}$$

where the coefficients $q_0, \ldots, q_6$ are the result of expanding 5:

$$\begin{aligned} &q_0 = P_{O_{16}}^3, \quad q_1 = 3P_{O_{16}}^2 P_{O_{17}}, \quad q_2 = 3P_{O_{16}}^2 P_{O_{18}} + 3P_{O_{16}} P_{O_{17}}^2, \qquad q_3 = P_{O_{17}}^3 + 6P_{O_{16}} P_{O_{17}} P_{O_{18}}, \\ &q_4 = 3P_{O_{17}}^2 P_{O_{18}} + 3P_{O_{16}} P_{O_{18}}^2, \quad q_5 = 3P_{O_{17}} P_{O_{18}}^2, \quad q_6 = P_{O_{18}}^3 \ . \end{aligned} \tag{6}$$

Thus, the coefficients indeed provide the probabilities of occurrence of aggregated isotopic variants with masses differing from the monoisotopic one by a specified integer number of mass units. In particular, q_0 gives the occurrence probability of the monoisotopic variant of O_3.

Note that, even for this seemingly simple example, the form of the coefficients is already quite complex. They are obtained by summing the occurrence probabilities of all isotope variants with exactly j additional neutrons, as compared to the monoisotopic variant. In a general case, however, such a naive approach to the calculation of the values of the coefficients is numerically not feasible.

Rockwood [12] proposed to approach the problem of the calculation of the coefficients by using the Fast Fourier Transform. The approach is numerically efficient and has been widely used. In what follows, we outline an alternate method by using the properties of the elementary symmetric polynomials and power sums of the roots of the polynomial in equation (2).

2.1 The new method for calculating the aggregated isotopic distribution

By applying the Newton-Girard theorem and Viete's formulae [9], we can express the coefficients q_j in the following recursive form:

$$q_j = -\frac{1}{j} \sum_{l=1}^{j} q_{j-l} \psi_l, \tag{7}$$

where ψ_l is a linear combination of the $(-l)$-th power of the roots of $Q_C(I)$, $Q_H(I)$, $Q_N(I)$, $Q_O(I)$, and $Q_S(I)$, defined in (3). More specifically, for C, H, and N, the roots become equal to

$$r_C = -\frac{P_{C_{12}}}{P_{C_{13}}}, \quad r_H = -\frac{P_{H_1}}{P_{H_2}}, \quad \text{and} \quad r_N = -\frac{P_{N_{14}}}{P_{N_{15}}} \ . \tag{8}$$

The roots of $Q_O(I)$ are conjugate complex numbers r_O and $\bar{r}_O$, defined as follows:

$$r_O = \frac{-P_{O_{17}} + \sqrt{P_{O_{17}}^2 - 4P_{O_{16}} P_{O_{18}}}}{2P_{O_{18}}}, \quad \bar{r}_O = \frac{-P_{O_{17}} - \sqrt{P_{O_{17}}^2 - 4P_{O_{16}} P_{O_{18}}}}{2P_{O_{18}}} \ . \tag{9}$$

The roots of $Q_S(I)$, a fourth-order polynomial, are less trivial, but can be expressed in a closed form. The expression is not very transparent, though; it can also be calculated using numerical root finding methods. There are two pairs of complex and conjugate roots of $Q_S(I)$, which we will denote by $(r_{S,1}, \bar{r}_{S,1})$ and $(r_{S,2}, \bar{r}_{S,2})$.

Using the roots defined above, the coefficients ψ_l can be expressed in general as follows:

$$\psi_l = v(r_C)^{-l} + w(r_H)^{-l} + x(r_N)^{-l} + y(r_O)^{-l} + y(\bar{r}_O)^{-l} + z(r_{S,1})^{-l} + z(\bar{r}_{S,1})^{-l} + z(r_{S,2})^{-l} + z(\bar{r}_{S,2})^{-l} \; . \tag{10}$$

Note that the sum of powers for conjugate complex numbers r and $\bar{r}$ can be written as

$$r^{-l} + \bar{r}^{-l} = |r|^{-l} \cos\{-l\varphi(r)\} \; , \tag{11}$$

where $|r|$ and $\varphi(r)$ indicate the modulus and argument of r and $\bar{r}$, respectively. From (11) it follows that the sum on the right-hand side of (10) can be simplified by replacing the sum of the powers of the conjugate roots of oxygen and sulphur by their reduced forms.

As it was already noted, equation (7) is recursive. To start the recursion, we need to compute the value of the coefficient q_0. In this case, the computation is trivial, as q_0 corresponds to the probability of occurrence of the monoisotopic variant. As pointed out by Beynon [1], the probability that no heavy isotopes would occur in a peptide of composition $C_v H_w N_x O_y S_z$ is

$$q_0 \;\; = \;\; P_{C_{12}}^{v} \times P_{H_1}^{w} \times P_{N_{14}}^{x} \times P_{O_{16}}^{y} \times P_{S_{32}}^{z} \; . \tag{12}$$

After having computed q_0, we can use (7) to compute q_1, q_2, etc.

Let us consider an example. For propane C_3H_8, the polynomial (2) assumes the following form:

$$Q(I;3,8,0,0,0) = \left(P_{C_{12}} I^0 + P_{C_{13}} I^1\right)^3 \times \left(P_{H_1} I^0 + P_{H_2} I^1\right)^8 = \{Q_C(I)\}^3 \{Q_H(I)\}^8 = \sum_{j=0}^{11} q_j I^j .$$

Following (12), the probability of occurrence of the monoisotopic variant (see Table 1) is given by

$$q_0 = P_{C_{12}}^3 \times P_{H_1}^8 = 0.9893^3 \times 0.999885^8 = 0.967352.$$

From (7), the probability of occurrence of the first aggregated isotopic variant is obtained as $q_1 = q_0 \times \psi_1$, where, according to (10),

$$\psi_1 = 3 \times r_C^{-1} + 8 \times r_H^{-1} = 3 \times \left(-\frac{0.9893}{0.0107}\right)^{-1} + 8 \times \left(-\frac{0.999885}{0.000115}\right)^{-1} = -0.033367.$$

Hence, $q_1 = -q_0 \times \psi_1 = -0.967352 \times (-0.033367) = 0.032278$. Thus, the probability of occurrence of an isotopic variant heavier by approximately 1 mass unit than the monoisotopic one is equal to 0.032278.

Next, we have $q_2 = -(q_0 \times \psi_2 + q_1 \times \psi_1)/2$, where

$$\psi_2 = 3 \times r_C^{-2} + 8 \times r_H^{-2} = 3 \times \left(-\frac{0.9893}{0.0107}\right)^{-2} + 8 \times \left(-\frac{0.999885}{0.000115}\right)^{-2} = 0.000351.$$

It follows that $q_2 = -(0.967352 \times 0.000351 + 0.032278 \times (-0.033367))/2 = 0.000369$. And so on up to q_{11}. The resulting aggregated isotopic distribution of propane is a follows:

$$\begin{array}{llll} q_0 = 0.967352, & q_1 = 0.032278, & q_2 = 0.000369, & q_3 = 1.55 \times 10^{-6}, \\ q_4 = 1.25 \times 10^{-9}, & q_5 = 4.83 \times 10^{-13}, & q_6 = 1.09 \times 10^{-16}, & q_7 = 1.54 \times 10^{-20}, \\ q_8 = 1.40 \times 10^{-24}, & q_9 = 8.01 \times 10^{-29}, & q_{10} = 2.62 \times 10^{-33}, & q_{11} = 2.26 \times 10^{-38}. \end{array}$$

A few comments are worth giving here.

- It can be observed that the complexity of calculations depends primarily on the number of different chemical elements present in the molecule (for peptides: C, H, N, O, S). It does not depend on the numbers of atoms for each element present in the molecule, but on the number of the aggregated isotopic variants, for which computations are required. In practice, one would stop the computations when the value of q_j falls below a particular (very small) threshold or when a preset percentage of the isotopic distribution is covered. Alternately, the computation of a fixed number of the q_j coefficients might be of interest.

- The method is very memory-efficient. In particular, it requires the storage of the mono-isotopic variant and only two variables, namely, q_j and ψ_l, for each desired aggregated isotopic variant. Hence, calculating the first, e.g., 100 aggregated isotopic variants requires only 201 numbers to be stored.

- It is possible to reduce the number of computations by computing in advance the roots and their powers (by using the logarithmic transformation for improved numerical stability) needed to compute the coefficients ψ_l and storing them for consecutive calculation steps.

- For chemical elements with more than four isotopic variants, a closed form solution of the roots is in general infeasible (the Abel-Ruffini theorem). The roots can be calculated by using numerical root-finding methods, such as the Newton-Raphson or Dandelin-Graeffe method. Again, the computed roots and their powers can be stored for further calculations.

- The value of ψ_l may be easily calculated by using vectorization and recursive formulae. For instance, because $b^{-l} = b^{-1}b^{-(l-1)}$, if we have already calculated $\psi_1, \psi_2, \ldots, \psi_{(l-1)}$, we can use the values to calculate ψ_l.

In the next section we show how the method can be used to compute the center-masses of the aggregated isotopic variants.

2.2 The new method for calculating the center-masses of the aggregated isotopic variants

As discussed by Roussis and Proulx [17], the center-mass $\bar{m}_j$ of the j-th aggregated variant is calculated as a probability-weighted sum of masses of the contributing isotopic variants:

$$\bar{m}_j = \frac{\sum_k m_{jk} p_{jk}}{\sum_k p_{jk}}, \tag{13}$$

where p_{jk} and m_{jk} denote, respectively, the probability of occurrence and the mass of the k-th isotopic variant contributing to the j-th aggregated variant. Note that the sum in the denominator of the fraction at the right-hand side of (13) is the occurrence probability of the j-th aggregated isotopic variant. Thus, $\sum_k p_{jk}$ is equal q_j and can be computed as outlined in Section 2.1.

It is obvious that accurate computations of the center-masses can only be achieved if all the isotopic variants contributing to the particular aggregated one are considered. Again, computations for all individual isotopic variants are in general infeasible due to the combinatorial explosion of the number of the variants for large molecules. However, we can circumvent this exhaustive method of calculation by resorting to the use of the Newton-Girard theorem and Viete's formulae.

To this aim, we first consider the following polynomial:

$$U(I; v, w, x, y, z) = \sum_j \left(\sum_k m_{jk} p_{jk} \right) I^j \equiv \sum_j q_j^\star I^j. \tag{14}$$

Note that we are interested in the coefficients $q_j^\star \equiv \sum_k m_{jk} p_{jk}$, which correspond to the numerator of the fraction at the right-hand side of the equation (13).

In order to obtain information about $q_j^\star$, we define a new polynomial by adding an additional indicator variable K in the polynomial (2):

$$\begin{aligned} Q^*(I, K; v, w, x, y, z) = \\ \left(P_{C_{12}} K^{M_{C_{12}}} I^0 + P_{C_{13}} K^{M_{C_{13}}} I^1\right)^v \times \left(P_{H_1} K^{M_{H_1}} I^0 + P_{H_2} K^{M_{H_2}} I^1\right)^w \times \\ \left(P_{N_{14}} K^{M_{N_{14}}} I^0 + P_{N_{15}} K^{M_{N_{15}}} I^1\right)^x \times \left(P_{O_{16}} K^{M_{O_{16}}} I^0 + P_{O_{17}} K^{M_{O_{17}}} I^1 + P_{O_{18}} K^{M_{O_{18}}} I^2\right)^y \times \\ \left(P_{S_{32}} K^{M_{S_{32}}} I^0 + P_{S_{33}} K^{M_{S_{33}}} I^1 + P_{S_{34}} K^{M_{S_{34}}} I^2 + P_{S_{36}} K^{M_{S_{36}}} I^4\right)^z \ , \end{aligned} \tag{15}$$

where $M_{C_{12}}$, $M_{C_{13}}$, ..., $M_{S_{36}}$ represent the masses of the isotopes of carbon, hydrogen, nitrogen, oxygen, and sulphur in normal terrestrial matter, as displayed in Table 1. The indicator variable K acts as a tracking device for the masses.

By using argumentation similar to the one used in Section 2.1, we can express the polynomial (15) as follows:

$$Q^*(I, K; v, w, x, y, z) \equiv \sum_j \left(\sum_k p_{jk} K^{m_{jk}} \right) I^j \ . \tag{16}$$

We will use $Q^*(I, K; v, w, x, y, z)$ to obtain the polynomial $U(I; v, w, x, y, z)$ from the equation (14). To this aim, we differentiate $Q^*(I, K; v, w, x, y, z)$ with respect to K:

$$\frac{\partial}{\partial K} Q^*(I, K; v, w, x, y, z) = \sum_j \left(\sum_k m_{jk} p_{jk} K^{m_{jk}-1} \right) I^j \ . \tag{17}$$

Then, by setting $K = 1$ in (17), we obtain:

$$
\begin{aligned}
U(I; v, w, x, y, z) \;=\; & \\
& vQ^*(I; v-1, w, x, y, z)\left(P_{C_{12}} M_{C_{12}} + P_{C_{13}} M_{C_{13}} I^1\right) && (18)\\
+\; & wQ^*(I; v, w-1, x, y, z)\left(P_{H_1} M_{H_1} + P_{H_2} M_{H_2} I^1\right) && (19)\\
+\; & xQ^*(I; v, w, x-1, y, z)\left(P_{N_{14}} M_{N_{14}} + P_{N_{15}} M_{N_{15}} I^1\right) && (20)\\
+\; & yQ^*(I; v, w, x, y-1, z)\left(P_{O_{16}} M_{O_{16}} + P_{O_{17}} M_{O_{17}} I^1 + P_{O_{18}} M_{O_{18}} I^2\right) && (21)\\
+\; & zQ^*(I; v, w, x, y, z-1) \times && (22)\\
& \left(P_{S_{32}} M_{S_{32}} + P_{S_{33}} M_{S_{33}} I^1 + P_{S_{34}} M_{S_{34}} I^2 + P_{S_{36}} M_{S_{36}} I^4\right) \;. && (23)
\end{aligned}
$$

By using the method outlined in Section 2.1, we can compute the coefficient q_j, i.e., the occurrence probability of the j-th aggregated isotopic variant, separately for each of the $Q^*(\cdot)$ polynomials, present in (18)–(23). Consequently, we can compute the coefficients of the five polynomials included in the sum on the right-hand side of (18)–(23). By adding the coefficients corresponding to I^j for the five polynomials, we obtain $q_j^\star$ for (14). Finally, the centered mass for the j-th aggregated isotopic variant is obtained from the equation (13) as $q_j^\star / q_j$.

3 Results and Discussion

We compared our method, named *BRAIN*, with five other algorithms. All methods were used to compare the aggregated isotopic distribution of 10 biomolecules shown in Table 2. The 10 biomolecules are the same as those used in the paper of Olson and Yergey [10]. The size of the molecules ranges from considerably small to very large.

3.1 Compared algorithms

The five packages considered in our comparison with *BRAIN* are *IsoPro*, *Mercury*, *Emass*, *NeutronCluster*, and *IsoDalton*.

IsoPro [18] is an implementation of the multinomial expansion method proposed by Yergey [25]. *Mercury* contains an implementation of the convolution method of Rockwood and Van Orden [14]. *Emass* calculates the masses and intensities of isotopic peaks by the linear transformation of Rockwood and Haimi [13]. *NeutronCluster* uses the equatransneutronic isotopes proposed by Olson and Yergey [10]. *IsoDalton* [20] efficiently calculates the fine isotopic distribution by means of dynamic programming. The outcome can be used as an intermediate step to retrieve the aggregated isotopic structure of a biomolecule.

For large molecules, the implementations of *IsoPro* and *IsoDalton* become computationally inefficient in terms of the memory usage and computation time. Because of this limitation, we report for these two methods the masses for the first seven and first five biomolecules, respectively.

All the algorithms were used with their default parameter settings, except of *IsoPro* and *NeutronCluster*. For the former the permutation threshold was set to 10^{-6}, while for the latter the

Table 2: List of selected biomolecules.

No.	Common Name	Molecular Formula	Mass (Da)	
			Monoisotopic	Average
(1)	Angiotensin II	$C_{50}H_{71}N_{13}O_{12}$	1045.534515	1046.181107
(2)	Bovine insulin	$C_{254}H_{377}N_{65}O_{75}S_6$	5729.600867	5733.510759
(3)	Human insulin	$C_{520}H_{817}N_{139}O_{147}S_8$	11616.849350	11624.448751
(4)	Human myoglobin	$C_{744}H_{1224}N_{210}O_{222}S_5$	16812.954775	16823.321352
(5)	Human intrinsic factor	$C_{2023}H_{3208}N_{524}O_{619}S_{20}$	45387.007033	45415.679370
(6)	Bovine serum albumin	$C_{2934}H_{4615}N_{781}O_{897}S_{39}$	66389.862474	66432.455561
(7)	Human Na/K ATPase Renal isoform, subunit	$C_{5047}H_{8014}N_{1338}O_{1495}S_8$	112823.879546	112895.125932
(8)	Human ATP binding cassette protein	$C_{8574}H_{13378}N_{2092}O_{2392}S_{77}$	186386.799265	186506.052594
(9)	Human intrinsic factor -hydroxocobalamin receptor	$C_{17600}H_{26474}N_{4752}O_{5486}S_{197}$	398470.366994	398722.972484
(10)	Human dynein heavy chain	$C_{23832}H_{37816}N_{6528}O_{7031}S_{170}$	533403.475090	533735.214651

required ion current coverage was changed to 0.999. As four out of the five algorithms use different values for the atomic masses and abundances of the isotopes (see Table S1 in the appendix), we have changed the abundances and masses to the values used by *IsoDalton*, which correspond to the IUPAC 1997 standard [16] as displayed in Table 1.

All the algorithms, except of *NeutronCluster*, have been run on a Dell Latitude E6500 with an Intel dual core P8400 2.26 GHz and 4 GB RAM. *NeutronCluster* has been run on a Apple MacBook with 4GB RAM, due to technical incompatibilities with the BigFloat and BigInt packages of Perl on a Windows operating system (personal communication with Olson and Yergey).

3.2 Results of the comparison

Table 3 presents the comparison of the mass of the first peak returned by *BRAIN* and by the other selected algorithms with the theoretical monoisotopic mass of the molecules presented in Table 2. The theoretical monoisotopic mass was simply computed as follows:

$$Monoisotopic\ mass = vM_{C_{12}} + wM_{H_1} + xM_{N_{14}} + yM_{O_{16}} + zM_{S_{32}}, \tag{24}$$

assuming that the atomic composition of the molecule is of the form $C_vH_wN_xO_yS_z$. In the remainder of this manuscript, we use the term "peak" to indicate the aggregated isotopic variant and not a variant of the isotopic fine structure as calculated via *IsoDalton*.

Negative values in Table 3 indicate that the mass of the first returned peak is higher than the monoisotopic mass. As *BRAIN*'s algorithm uses the same formula for the calculation of

the monoisotopic mass as (24), there is no difference between the reported and theoretical monoisotopic mass for *BRAIN*. For *NeutronCluster*, awkwardly, there is a considerably large difference for molecule no. 4. The returned monoisotopic mass of the molecule is identical to the mass reported in the paper of Olson and Yergey[10]. All other monoisotopic masses returned by *NeutronCluster* are identical to the ones calculated by (24).

For the first four molecules, the mass of the returned first peak is the same or close to the monoisotopic-variant mass in the case of *Emass* and *IsoPro*. Starting from molecule no. 5, the difference between the monoisotopic and the returned mass increases. To put these results in perspective, it is worth noting that for large molecules, e.g., (5)-(10) in Table 2, the probability of occurrence of the monoisotopic variants is very small, $< 10^{-10}$. In practice, such small values fall below the detection limits of a mass spectrometer and will go unnoticed. Therefore, one can argue that returning isotopic variants with such low probabilities is not meaningful. It can happen that some of the non-reported peaks were actually calculated inside the compared programs, but were not reported due to a reporting threshold built into the method. In our method, we have chosen to return all aggregated isotope variants regardless of their probability of occurrence.

For *Mercury*, the first returned peaks have masses which are lower than the monoisotopic mass in the case of small biomolecules. A possible explanation for this behavior could be the numerical imprecisions of the (discrete) Fast Fourier Transform, e.g., distortion of the signal due to aliasing. This could be fixed by a slight modification of the computer code to widen the calculation window. Generally, peaks with lower masses than the monoisotopic one should just be ignored.

For larger molecules, the masses returned for the first peaks by *Emass* and *IsoPro* are higher than the theoretical monoisotopic mass. This is probably the result of pruning techniques applied by these methods. Note that this is also the case for *Mercury*.

Table 3: Differences between the monoisotopic mass according to (24) (see Table 2) and the mass of the first returned peak by the algorithm. Negative values correspond to higher reported masses.

Molecule	*BRAIN*	*Emass*	*Mercury*	*NeutronCluster*	*IsoPro*	*IsoDalton*
(1)	0	0	7.019535	0	-0.000005	0
(2)	0	0	12.019012	0	-0.000023	-0.000001
(3)	0	0	8.013385	0	0.000030	-0.000003
(4)	0	0	22.053225	-360	-0.000055	-0.000005
(5)	0	-2.005731	2.996274	0	-8.024597	-0.000014
(6)	0	-8.022072	22.028846	0	-19.055126	
(7)	0	-24.065517	-7.029971	0	-51.133714	
(8)	0	-55.149869	-55.161957	0		
(9)	0	-155.399787	-124.322610	0		
(10)	0	-220.583942	-203.597009	0		

Although *IsoDalton* calculates the fine structure of the isotopic distribution, there are slight differences between the returned and theoretical monoisotopic mass. These differences can possibly

be explained by the fact that, in the calculations presented in Table 3, the "*exact_probability*" module has been used, as was advised by Snider (personal communication), instead of the "*exact_mass*" module. The latter results in more accurate estimates for the masses, but is less accurate for the average aggregated mass and its corresponding expected peak abundance.

To check the overall accuracy of the computation of an aggregated isotopic distribution, we considered the theoretical average mass of the molecules presented in Table 2. The average mass is computed according to the following definition:

$$\begin{aligned} \textit{Average mass} &= vM_{C_{12}} \times P_{C_{12}} + vM_{C_{13}} \times P_{C_{13}} \\ &+ wM_{H_1} \times P_{H_1} + wM_{H_2} \times P_{H_2} \\ &+ xM_{N_{14}} \times P_{N_{14}} + xM_{N_{15}} \times P_{N_{15}} \\ &+ yM_{O_{16}} \times P_{O_{16}} + yM_{O_{17}} \times P_{O_{17}} + yM_{O_{18}} \times P_{O_{18}} \\ &+ zM_{S_{32}} \times P_{S_{32}} + zM_{S_{33}} \times P_{S_{33}} + zM_{S_{34}} \times P_{S_{34}} + zM_{S_{36}} \times P_{S_{36}} \ . \end{aligned} \quad (25)$$

Table 4 presents the results of the comparison of the theoretical average mass, as computed in (25), and the weighted average based upon the predicted masses and occurrence probabilities for all peaks returned by a particular algorithm. There is virtually no difference between the two average values for *BRAIN* and for *Emass*. Somewhat larger, but small deviations are obtained for *IsoDalton*. The differences are larger for *Mercury* and *NeutronCluster* and they increase as the molecules become larger. *IsoPro* is unexpectedly the least accurate method in our comparison; the large differences are most likely a side effect of the pruning step, which removes low probability variants during the calculation. Pruning is a necessity to calculate the isotope fine structure for large molecules in order to maintain the computational complexity memory usage within limits.

It is worth mentioning that, when the results in Table 4 are viewed in relative terms, all of the reported numbers are quite satisfactory. The reported differences between *Emass*, *Mercury*, *NeutronCluster*, and *IsoDalton* are in fact not measurable with the accuracy available in the current generation of mass spectrometers.

Table 4: Difference between the theoretical (see Table 2) and calculated (using all returned peaks) average mass. Negative values correspond to higher calculated masses. The values in parentheses are the relative differences in ppb.

Molecule	*BRAIN*	*Emass*	*Mercury*	*NeutronCluster*	*IsoPro*	*IsoDalton*
(1)	0 (0)	-0.000001 (0.956)	0.000090 (86.027)	0.000238 (227.494)	0.001297 (1.240e+3)	-0.000001 (0.956)
(2)	0 (0)	0 (0)	0.000323 (56.336)	0.002474 (431.498)	0.011478 (2.002e+3)	-0.000001 (0.174)
(3)	0 (0)	0 (0)	0.000225 (19.356)	0.006620 (569.489)	0.093513 (8.045e+3)	0.000245 (21.076)
(4)	0 (0)	0 (0)	0.002916 (173.331)	-360.315145 (-2.142e+7)	0.155448 (9.240e+3)	-0.000005 (0.297)
(5)	0 (0)	0 (0)	-0.003078 (67.774)	-0.008751 (-192.687)	0.947604 (20.865e+3)	-0.000013 (0.286)
(6)	0 (0)	0 (0)	-0.004153 (62.515)	-0.003685 (-55.470)	2.094637 (31.530e+3)	
(7)	0 (0)	0 (0)	0.003699 (32.765)	-0.021463 (-190.114)	1.944364 (17.223e+3)	
(8)	0 (0)	0 (0)	-0.005138 (27.549)	-0.078241(-419.509)		
(9)	0 (0)	0 (0)	0.017207 (43.155)	-0.057899 (-145.211)		
(10)	0 (0)	0 (0)	-0.047547 (89.084)	0.092907 (174.069)		

The differences observed in Table 4 are mainly due to the fact that every algorithm returns a different number of peaks (i.e., aggregated isotopic variants), with a different first and last

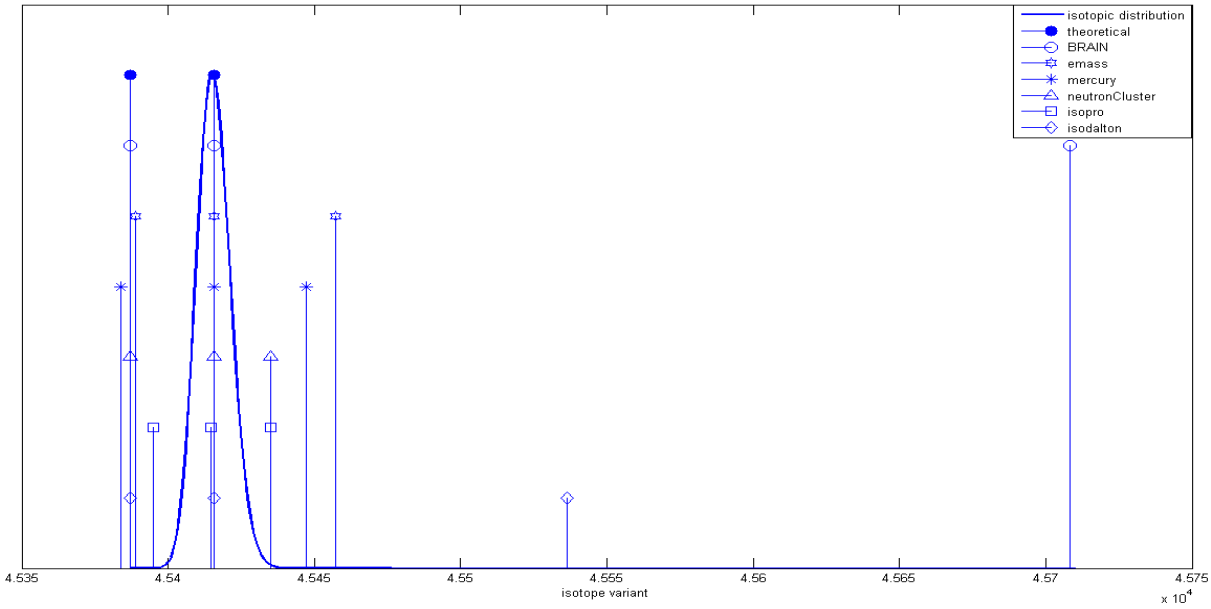

Figure 1: The calculated aggregated isotopic distribution of human intrinsic factor. (The height of the lines have no meaning, they are only chosen to facilitate the interpretation of the graph.)

reported peak. Figure 1 illustrates how the methods perform in the tail of the distribution for the molecule no. 5 from Table 2. The figure shows three vertical lines for each method, indicating the mass of the first reported peak, the average mass, and the last reported peak. In other words, the representation in Figure 1 can be seen as the coverage of the isotopic distribution by a particular method. Note that the lines indicating the mass of the first reported peak for *BRAIN*, *NeutronCluster*, and *IsoDalton* overlap with the line corresponding to the theoretical monoisotopic mass, in agreement with the results presented in Table 3. Similarly, the lines indicating the average mass practically overlap with the line corresponding to the theoretical expected mass for *BRAIN*, *Emass*, *IsoDalton*, *Mercury*, and *NeutronCluster*, in agreement with the results presented in Table 4. A clear difference can be observed in the mass of the last reported peak. *IsoPro* and *NeutronCluster* report a peak with the smallest mass, followed by *Mercury*, *Emass*, *IsoDalton*, and *BRAIN*. Depending on the shape of the true aggregated isotopic distribution, the different number and location of the reported peaks may lead to a difference between the value of the average mass computed for a particular algorithm and the theoretical value, obtained from (25). For the case presented in Figure 1, the difference is small, though visible for *IsoPro*.

As it was mentioned in Section 2.1, for *BRAIN*, the calculations can be stopped when the computed occurrence probabilities become too small or when the required number of aggregated isotopic variants has been reached. The latter number can be heuristically obtained. For this purpose, we propose the following rule of thumb: compute the difference between the theoretical

monoisotopic mass and the theoretical average mass, multiply this number by two, and subsequently round it to the nearest integer greater than or equal to the multiplied difference. For instance, for the molecule no. 10 in Table 2, the heavy chain of the human dynein protein, this method gives 664 as the number of the aggregated isotopic variants to be included in the calculations. Note, however, that the method may return a too small number for smaller molecules. For instance, in the case of the molecule no. 1 in Table 2, angiotensin II, the obtained number is equal to 2. For such small molecules, the minimal number of peaks should be four or five. As already mentioned before, schemes based on the percentage coverage of the isotopic distribution can also be used.

The number of isotopic variants used in the computation of the average masses in Table 4 and the corresponding computation time for our method are listed in Table 5. Increasing the number of the requested variants influences the computation time, but the effect is minor. Comparison of the computation time of *BRAIN* with the other algorithms is difficult, as the methods are implemented using different software and platforms. In general terms, *Emass* and *Mercury* are faster than our method, but the differences are negligible small. It is worth noting, however, that *BRAIN* is now operated by an interpreted language. We believe that a compiled version of *BRAIN* will be as fast as *Emass* and *Mercury*.

Table 5: Requested number of aggregated isotopic variants and the associated computation time for *BRAIN*. The calculations were performed in Matlab.

Molecule	Requested no. of variants	Time (s)
(1)	50	0.037523
(2)	50	0.037040
(3)	50	0.037627
(4)	100	0.037019
(5)	322	0.072257
(6)	400	0.075427
(7)	643	0.155975
(8)	807	0.216821
(9)	1163	0.355737
(10)	1325	0.408562

The results presented in Table 4 already indicate the proper functioning of *BRAIN*. If the calculation of the occurrence probabilities and/or center-masses were wrong, then the average masses of the molecules would deviate from their theoretical values. In order to further investigate the accuracy of the calculations for our method in more detail, we computed the isotopic distribution for the molecules no. 1 (angiotensin II) and 2 (bovine insulin) from Table 2 by considering all possible isotopic variants, while using an implementation of the multinomial expansion [25]. From the obtained result we derived the aggregated isotopic distribution. Tables 6 and 7 present the center-masses and occurrence probabilities for the first 50 aggregated isotopic variants for angiotensin II and bovine insulin, respectively. We can confirm that *BRAIN* provides exactly the same masses and occurrence probabilities for these aggregated isotopic variants.

As mentioned previously, for large molecules, the occurrence probabilities for the monoisotopic

Table 6: The first 50 aggregated isotopic variants for angiotensin II.

Mass	Abundance	Mass	Abundance	Mass	Abundance
1045.534515	0.536241	1062.576779	0	1079.617682	0
1046.537411	0.322570	1063.579164	0	1080.620130	0
1047.540111	0.108627	1064.581550	0	1081.622584	0
1048.542719	0.026442	1065.583936	0	1082.625044	0
1049.545270	0.005141	1066.586324	0	1083.627509	0
1050.547780	0.000842	1067.588713	0	1084.629979	0
1051.550262	0.000120	1068.591105	0	1085.632454	0
1052.552722	0.000015	1069.593499	0	1086.634932	0
1053.555164	0.000002	1070.595897	0	1087.637413	0
1054.557593	0	1071.598298	0	1088.639897	0
1055.560011	0	1072.600703	0	1089.642381	0
1056.562421	0	1073.603113	0	1090.644866	0
1057.564824	0	1074.605527	0	1091.647350	0
1058.567221	0	1075.607947	0	1092.649831	0
1059.569614	0	1076.610372	0	1093.652310	0
1060.572004	0	1077.612803	0	1094.654784	0
1061.574392	0	1078.615239	0		

Table 7: The first 50 aggregated isotopic variants for bovine insuline.

Mass	Abundance	Mass	Abundance	Mass	Abundance
5729.6008666	0.0298940	5746.6269490	0.0000057	5763.6514943	0
5730.6037205	0.0928879	5747.6282361	0.0000017	5764.6531171	0
5731.6060166	0.1565624	5748.6295395	0.0000005	5765.6547575	0
5732.6079855	0.1874710	5749.6308606	0.0000001	5766.6564152	0
5733.6097364	0.1774096	5750.6322007	0	5767.6580896	0
5734.6113345	0.1404106	5751.6335606	0	5768.6597801	0
5735.6128224	0.0962370	5752.6349409	0	5769.6614863	0
5736.6142300	0.0584802	5753.6363420	0	5770.6632076	0
5737.6155792	0.0320421	5754.6377643	0	5771.6649435	0
5738.6168866	0.0160312	5755.6392077	0	5772.6666936	0
5739.6181650	0.0073961	5756.6406722	0	5773.6684573	0
5740.6194246	0.0031713	5757.6421577	0	5774.6702342	0
5741.6206735	0.0012719	5758.6436640	0	5775.6720238	0
5742.6219182	0.0004797	5759.6451908	0	5776.6738256	0
5743.6231641	0.0001709	5760.6467376	0	5777.6756394	0
5744.6244157	0.0000577	5761.6483041	0	5778.6774645	0
5745.6256763	0.0000185	5762.6498899	0		

and (several) consecutive aggregated isotopic variants can be very small. In that case, given that

the recursive relationship (7) implies starting the calculations from the monoisotopic variant, the computations of these initial probabilities can be affected by the level of the available numerical precision. As seen from the results presented, e.g., in Table 4, these numerical precision issues do not influence the calculations for the meaningful region of the aggregated isotopic distribution, i.e., for the aggregated isotopic variants with non-negligible occurrence probabilities. However, for extremely large molecules this does not hold. These molecules have abundances for the monoisotopic and consecutive peaks that are extremely small, i.e., $\ll 10^{-100}$. As these molecules are exceptional and difficult to measure with the accuracy avaible in the current generation of mass spectrometers, we can ignore this numerical issue.

The method we propose is predominantly conceived for calculating the aggregated isotopic distribution. From a practical point of view, ignoring the isotopic fine structure is not a serious limitation. This is because for large molecules like, e.g., intact proteins, the resolution in MS does not allow for observing the fine structure of aggregated isotopic variants. For large molecules, the calculation of exact center-masses of aggregated variants becomes fundamental, and the calculation is taken care of by our method. When information about the isotopic fine structure is required, other methods proposed in, e.g., [7], [8], [11], or [20] can be used. If the molecule is not too large, the multinomial expansion [25] can be applied to infer the isotopic fine structure.

4 Conclusions

The proposed *BRAIN* method allows a fast computation of the aggregated isotopic distribution. It provides the correct values of the occurrence probabilities of various aggregated isotopic variants and the center-masses. In terms of speed and accuracy, *BRAIN* yields results comparable to those obtained by existing algorithms like *Emass*, but is more memory-efficient and simpler to implement. The *BRAIN* method will be made available within the Bioconductor package in R.

5 Acknowledgment

JC gratefully acknowledges financial support from Bijzonder Onderzoeksfonds Universiteit Hasselt (grant BOF09NI006). PD gratefully acknowledges the LLP Erasmus Placement Programme for supporting his visit at Hasselt University. The authors are grateful to Ross Snider, Alan Rockwood, and Matthew Olson and Alfred Yergey for providing the software, and to Peter Boyen for the help with implementing *NeutronCluster*.

The authors are grateful to the editor and the reviewers for their insightful comments. All of these comments were most helpful and have resulted in an improved text.

References

[1] J.H. Beynon. *Mass Spectrometry and its Applications to Organic Chemistry*. New York: Elsevier, 1960.

[2] K. Biemann. *Mass Spectrometry, Organic Chemical Applications*. McGraw-Hill, New York, 1962.

[3] E.J. Breen, F.G. Hopwood, K.L. Williams, and M.R. Wilkins. Automatic Poisson peak harvesting for high throughput protein identification. *Electrophoresis*, 21:2243–2251, 2000.

[4] M. Brownawell and J.S. Fillippo. A program for the synthesis of mass spectral isotopic abundances. *Journal of Chemical Education*, 59(8):663–665, 1982.

[5] S. Gay, P. Binz, D Hochstrasser, and R. Appel. Modeling peptide mass fingerprinting data using the atomic composition of peptides. *Electrophoresis*, 20:3527–3534, 1999.

[6] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A.J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J.YH. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

[7] L. Li, M. Karabacak, J. Cobb, Q. Wang, P. Hong, and J. Agar. Memory-efficient calculation of the isotopic mass states of a molecule. *Rapid Communications in Mass Spectrometry*, 24:2689–2696, 2010.

[8] L. Li, J. Kresh, M. Karabacak, J. Cobb, J. Agar, and P. Hong. A hierarchical algorithm for calculating the isotopic fine structures of molecules. *Journal of American Society for Mass Spectrometry*, 19:1867–1874, 2008.

[9] I. G. Macdonald. *Symmetric Functions and Hall Polynomials*. Clarendon Press ; Oxford University Press, Oxford : New York, 1979.

[10] M. Olson and A. Yergey. Calculation of the isotope cluster for polypeptides by probability grouping. *Journal of American Society for Mass Spectrometry*, 20:295–302, 2009.

[11] A. L. Rockwood, S. L. Van Orden, and R. D. Smith. Ultrahigh resolution isotope distribution calculations. *Rapid Communication in Mass Spectrometry*, 10:54–59, 1996.

[12] A.L. Rockwood. Relationship of fourier transforms to isotope distribution calculations. *Rapid Communications in Mass Spectrometry*, 9:103–105, 1995.

[13] A.L. Rockwood and P. Haimi. Efficient calculation of accurate masses of isotopic peaks. *Journal of the American Society for Mass Spectrometry*, 17:415–419, 2006.

[14] A.L. Rockwood and S.L. Van Orden. Ultrahigh-speed calculation of isotope distributions. *Analytical Chemistry*, 68:2027–2030, 1996.

[15] A.L. Rockwood, J.R. Van Orman, and D.V. Dearden. Isotopic compositions and accurate masses of single isotopic peaks. *Journal of the American Society for Mass Spectrometry*, 15:12–21, 2004.

[16] K.J.R. Rosman and P.D.P. Taylor. Isotopic compositions of the elements 1997. *Pure and Applied Chemistry*, 70(1):217–235, 1998.

[17] S.G. Roussis and R. Proulx. Reduction of chemical formulas from the isotopic peak distributions of high-resolution mass spectra. *Analytical Chemistry*, 75(6):1470–1482, 2003.

[18] M.W. Senko. Isopro computer program 3.0.

[19] M.W. Senko, S.C. Beu, and F.W. McLafferty. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distribution. *Journal of the American Society for Mass Spectrometry*, 6:229–233, 1995.

[20] R.K. Snider. Efficient calculation of exact mass isotopic distributions. *Journal of the American Society for Mass Spectrometry*, 18:1511–1515, 2007.

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

[22] D. Valkenborg, I. Jansen, and T. Burzykowski. A model-based method for the prediction of the isotopic distribution of peptides. *Journal of the American Society for Mass Spectrometry*, 19(5):703–712, 2008.

[23] D. Valkenborg, I. Mertens, F. Lemière, E. Witters, and T. Burzykowski. The isotopic distribution conundrum. *Mass Spectrometry Review*, DOI: 10.1002/mas.20339, 2011.

[24] H. Yamamoto and J. A. McCloskey. Calculations of isotopic distribution in molecules extensively labeled with heavy isotopes. *Analytical Chemistry*, 49:281–283, 1977.

[25] J. A. Yergey. A general approach to calculating isotopic distributions for mass spectrometry. *International Journal of Mass Spectrometry and Ion Physics*, 52:337–349, 1983.

[26] J. A. Yergey, D. Heller, G. Hansen, R.J. Cotter, and C. Fenselau. Isotopic distributions in mass spectra of large molecules. *Analytical Chemistry*, 55:353–356, 1983.

Appendix

Table S1: Atomic masses and abundance of C,H,N,O and S according to the tested algorithms.

Element	Mass IsoPro	Abundance (%) IsoPro	Mass Emass	Abundance(%) and Mercury	Mass IsoDalton	Abundance(%) IsoDalton	Mass NeutronCluster	Abundance(%) NeutronCluster
^{1}H	1.007825017	99.9850	1.0078246	99.985	1.0078250321	99.9885	1.007825	99.9886
^{2}H	2.013999939	0.0105	2.0141021	0.015	2.0141017780	0.0115	2.014102	0.011570
^{12}C	12.0	98.90	12	98.893	12	98.93	12	98.938
^{13}C	13.003350258	1.10	13.0033554	1.1070	13.0033548378	1.07	13.003354	1.078
^{14}N	14.003069878	99.640	14.003072	99.6337	14.0030740052	99.632	14.00307	99.6327
^{15}N	15.000109673	0.360	15.0001088	0.3663	15.0001088984	0.368	15.000108	0.3687
^{16}O	15.994910240	99.760	15.9949141	99.7590	15.9949146	99.757	15.99491	99.75716
^{17}O	16.999130249	0.040	16.9991322	0.0374	16.9991315	0.038	16.99913	0.0381
^{18}O	17.999160767	0.20	17.9991616	0.2036	17.9991604	0.205	17.99916	0.20514
^{32}S	31.972070694	95.0	31.972070	95.02	31.97207069	94.93	31.97207069	94.9331
^{33}S	32.971458435	0.760	32.971456	0.75	32.97145850	0.76	32.97146	0.762
^{34}S	33.967861176	4.220	33.967866	4.21	33.96786683	4.29	33.96787	4.2928
^{36}S	35.967090607	0.020	35.967080	0.021	35.96708088	0.02	35.96708088	0.021