The isotopic distribution conundrum

Dirk Valkenborg[1, 2, 3], Inge Mertens[1], Filip Lemière[4], Erwin Witters[1, 4], Tomasz Burzykowski[2, 3]

[1] Flemish Institute for Technological Research, VITO, Mol, Belgium
[2] I-BioStat, Hasselt University, Diepenbeek, Belgium
[3] I-BioStat, Catholic University of Leuven, Leuven, Belgium
[4] Center for Proteome Analysis and Mass Spectrometry, University of Antwerp, Antwerp, Belgium

Corresponding author:
Dirk Valkenborg
Boeretang 200
B-2400 Mol
Belgium
dirk.valkenborg@vito.be

Abstract:

Alhought access to high-resolution mass spectrometry, especially in the field of biomolecular mass spectrometry, is becoming readily available due to recent advances in MS technology, the accompanied information on isotopic distribution in high-resolution spectra is not used at its full potential, mainly because of lack of knowledge and or awareness.

In this review, we give an insight on the practical problems when calculating the isotopic distribution for large biomolecules, and present an overview of methods for the calculation of the isotopic distribution. We will discuss the key events that triggered the development of various algorithms and explain to the reader the rationale of how and why the various isotopic-distribution calculations were performed. The review is focused around the developmental stages as briefly outlined below, starting with the first observation of an isotopic distribution.

The observations of Beynon in the field of organic mass spectrometry that chlorine appeared in a mass spectrum as two variants as odds 3:1 lie at the basis of the first wave of algorithms for the calculation of the isotopic distribution, based on the atomic composition of a molecule. From here on, we will explain why more complex biomolecules such as peptides exhibit a highly complex isotope pattern when assayed by mass spectrometry, and explain how combinatorial difficulties complicate the calculation of the isotopic distribution on computers. For this purpose, we highlight three methods, which were introduced in the eighties. These are the stepwise procedure introduced by Kubinyi, the polynomial expansion from Brownawell and Fillippo, and the multinomial expansion from Yergey.

The next development was instigated by Rockwood, who suggested to decompose the isotopic distribution in terms of their nucleon count instead of the exact mass. In this respect, we could claim that the term 'aggregated' isotopic distribution is more appropriate. Due to the simplification of the isotopic distribution to its aggregated counterpart, Rockwood was able to use the convolution for the calculation of the 'aggregated' isotopic distribution. Convolution methods are computationally efficient and economic in their memory usage. We will spend a section on the work introduced by Rockwood during the nineties.

Due to recent breakthroughs in mass spectrometric technology and the widespread high-resolution instruments (e.g., FTICR-MS, FTOrbitrap-MS, TOF-MS) that provide high-resolution, isotope-resolved, accurate mass data, there is an emerging need for algorithms that can calculate isotopic distributions for large biomolecules. The number of recent publications on this topic does witness this trend. The new methods are mostly based on complex mathematical developments such as, e.g., cellular automata [25][25][25][25][25], dynamic programming [43][43][43][43][43], and hierarchical models[19][19][19][19] [19].

We also comment on the ideas to use Punnet squares and Pascal's triangle to introduce the concept of the isotopic distribution for educational and didactic purposes.

Introduction

In 1913, the mental father of mass spectrometry, J. J. Thomson, made a controversial discovery during a canal ray experiment. He observed that neon in a positive ray tube existed as atoms with different atomic masses[44][44][44][44] [44]. Moreover, he observed parabolas of Neon that corresponded to two different isotopes, ^{20}Ne and ^{22}Ne. These were the first non-radioactive isotopes to be discovered and measured with mass spectrometry. More-accurate mass spectrometers were built by Aston and by Dempster to study stable- and radio-isotopes. Doing so, they instigated the field of modern mass spectrometry. These events have led to a change in the meaning of molecular and atomic masses, and introduced mathematical combinatorics in this field. An interesting review on the history of isotope research by Budzikiewicz and Grigsby [6][6][6][6][6] is available for more background information. In this review, we provide an overview of different methods that have been proposed to calculate the *isotopic distribution* of a molecule given its elemental composition. We focus on how the isotopic distribution can be applied in the context of interpretation of mass spectrometry data.

On Earth, particular chemical elements appear as different stable or radioactive variants. These variants have a different number of neutrons in their atomic nucleus and are known as *isotopes*. For example, there are two stable isotopes of carbon (*C*) that appear in nature: $^{12}_{6}C$ and $^{13}_{6}C$. In the notation, the upper-left index indicates the *mass* number or nucleon number and represents the total number of protons and neutrons, i.e., nucleons, in the atomic nucleus. The lower-left index indicates the atomic number, i.e., the number of protons in the nucleus. It should be noted that each elemental isotope has a different mass number. Hence, the number of neutrons in the nucleus can be easily determined by subtracting the atomic number from the nucleon number. In the remainder of the paper, we will suppress the atomic number from the notation. The natural abundances of the elemental isotopes in normal terrestrial matter are known, and are described by a Commission on Atomic Weights and Isotopic Abundances of the International Union of Pure and Applied Chemistry (IUPAC) in [1][1][1][1][1].

Terrestrial molecules incorporate elemental isotopes according to their natural abundances. Thus, a molecule can have different isotopic variants with different masses that depend on the number of different isotopes of the atomic composition of the molecule. The probability of occurrence of these isotopic variants can be calculated given the atomic composition and the known elemental isotope abundances. The result of this calculation is known as the isotopic distribution. Caution should be applied when computing the isotopic distribution for molecules that deviate from the IUPAC description. For example, some groups of plants can have a preference to incorporate ^{13}C in their molecules. More information on this topic can be found in the review of Farquhar *et al.* [12][12][12][12][12]. Moreover, isotope distribution, predominantly those of oxygen, in seawater have varied over geologic timescales in response to climate alterations.

The knowledge of the isotopic distribution of a molecule can be useful for several purposes. We will indicate its usefulness by describing two popular applications. As a first application, it can be used to filter out the relevant information about peptide molecules from high-resolution mass spectra obtained from a biological sample. A mass spectrum usually contains a lot of noise and

redundant information. However, a peptide molecule should be represented in a spectrum by a characteristic signal in the form of (short) series of regularly spaced peaks; the isotope envelope, which should exhibit a specific profile related to the isotopic distribution of the peptide. This typical signal originates from the fact that a population of ions for a particular peptide is measured, instead of a single ion. In this population of ions, a percentage of molecules will carry, for instance, one ^{13}C isotope which leads to an increase in mass of approximately one dalton (Da) due to the extra neutron in ^{13}C, etc. From prior information about the expected form of the isotopic distribution, the information about peptide abundance can be extracted from the spectrum. Such prior information can be obtained from, e.g., a prediction based on an average peptide [41, 15, 46, 47][41, 15, 46, 47][41, 15, 46, 47][41, 15, 46, 47][41, 15, 46, 47]. For example, McIlwain *et al.* [21][21][21][21][21] proposed a scoring algorithm to annotate isotopic peaks by means of Bayes nets in combination with dynamic programming. As a second application, in high-resolution mass spectrometry (MS), isotopes of sulfur can be resolved separately in a spectrum due to the relatively large difference between the mass of the atomic nucleus and the sum of the masses of its fundamental particles, i.e., the mass defect. Therefore, high-resolution isotopic profiles of large molecules, observed in a mass spectrum, can be compared to a hypothesized isotopic distribution so that the information on the elemental composition is revealed. It should be noted that, in such a case, the elemental composition of the molecule is identified without the requirement of a tandem MS scan [42, 38][42, 38][42, 38][42, 38][42, 38].

Thus, the calculation of the isotopic distribution is an important practical tool. The need for an efficient calculation method is proved by the substantial number of publications on this topic. In this paper, we provide a chronological overview of the publications, which have led to the current status of algorithms for the calculation of the isotopic distribution. In our presentation, we ignore the methods used for the graphical representation of theoretical mass spectra. Instead, we concentrate on information about the isotopic distribution in the form of a discrete probability distribution for different mass isotopic variants of a molecule. The information can be stored as a list of masses (m) of the isotopic variants and their corresponding occurrence probabilities (p). The graphical representation of theoretical mass spectra is trivial and is usually done by rendering a Gaussian or Laplacian approximation which is convoluted with the obtained list of mass and probabilities, as discussed by Werlen [48][48][48][48][48].

The chronological overview is divided into three main parts. The first part mainly focuses on the methods based on the polynomial and multinomial expansion and stepwise addition. The second part is devoted entirely to the methodology proposed by A. Rockwood, in which the polynomial expansion is linked to the functional convolution. This link has resulted in a straightforward and computational efficient algorithm. The third part presents approaches, that use more advanced computational methods, such as dynamic programming and hierarchical algorithms, to elaborate the isotopic fine structure of a molecule.

In the literature, these methods are often compared and adjusted as if they were totally different procedures. However, they are different formulations of the same concept. The reason for this confusion can tentatively be explained by emerging technologies for scientific computing in the early eighties. Computing power was limited, and the usage of internal memory and of the numerical machine precision were important issues those days. The different techniques focus on how to

efficiently use these resources in order to obtain the isotopic distribution within a reasonable time interval.

Unless specified differently, in our presentation, we focus on peptides which are mainly composed of carbon (*C*), hydrogen (*H*), nitrogen (*N*), oxygen (*O*), and sulfur (*S*). However, the methods mentioned in this manuscript can also be applied to other polyisotopic elements. An advantage of focusing on peptides is that the monoisotopic mass of a molecule corresponds to the lightest isotopic variant. In the remainder of this review, we will term this variant as "the monoisotopic variant". The monoisotopic variant is composed of the most-abundant isotopes. For carbon, hydrogen, nitrogen, oxygen, and sulfur, the most-abundant (and lightest) isotopes are ^{12}C, ^{1}H, ^{14}N, ^{16}O, and ^{32}S. Further, we only consider stable isotopes, i.e., the isotopes just mentioned, together with ^{13}C, ^{2}H, ^{15}N, ^{17}O, ^{18}O, ^{33}S, ^{34}S, and ^{36}S. The considered isotopes are listed in Table 1 together with their atomic mass and natural abundance. It is not our intention to subject the algorithms and software packages to a benchmark set to compare the performance, speed, and memory use. Instead, we will intuitively explain the concepts behind each algorithm. An isotopic-abundance calculation involves probability theory and combinatorial analysis. Although mathematics and statistics cannot be avoided completely, we will emphasize practical matters, and aiming to answer the advantages and disadvantages related to the method.

Table 1: Standard atomic weights (IUPAC 1997) for the isotopes of elements as they exist naturally in normal terrestrial material.

Isotopes	**Atomic mass (Da)**	**Natural abundance (%)**
^{12}C	12.0000000000	98.93
^{13}C	13.0033548378	1.07
^{1}H	1.0078250321	99.9885
^{2}H	2.0141017780	0.0115
^{14}N	14.0030740052	99.632
^{15}N	15.0001088984	0.368
^{16}O	15.9949146	99.757
^{17}O	16.9991312	0.038
^{18}O	17.9991603	0.205
^{32}S	31.97207070	94.93
^{33}S	32.97145843	0.76
^{34}S	33.96786665	4.29

^{36}S	35.96708062	0.02

PART I: The early methods

The stepwise procedure and the polynomial expansion

In 1960, Beynon [3][3][3][3][3][3][3][3][3] indicated that the probability that no heavy isotopes would occur in a peptide with elemental composition $C_vH_wN_xO_yS_z$ was equal to

$$P = Pr(^{12}C)^v \times Pr(^{1}H)^w \times Pr(^{14}N)^x \times Pr(^{16}O)^y \times Pr(^{32}S)^z \quad (1)$$

where the terms Pr(^{12}C), Pr(^{1}H),Pr(^{14}N), Pr(^{16}O), and Pr(^{32}S) represent the natural abundances or, more specifically, the probability of the occurrence, of the particular isotopes of carbon, hydrogen, nitrogen, oxygen, and sulfur, respectively, as displayed in Table 1. In other words, equation (1) gives the probability of occurrence for the monoisotopic variant of the peptide molecule.

The calculation of the remaining isotopic variants is less trivial, and becomes increasingly complex in the case of molecules that contain many polyisotopic elements; typical for biomolecules. Thus, as the molecule mass increases, the calculation complexity also increases. In theory, a peptide with a molecular formula of $C_vH_wN_xO_yS_z$ has $2^v2^w2^x3^y4^z$ possible isotopic variants, when taking into account the location of the isotope in the molecule. For example, if one ^{13}C atom is incorporated in the molecule, then it can be present as any of the *v* carbon atoms.

It is worth noting that a mass spectrometer cannot discern among the different locations of an elemental isotope in an intact molecule, because the isotope location variants are isobaric. Thus, what is important is the number of different elemental isotopes, and not their actual position in a molecule. Therefore, the number of isotopic variants, from a point of view of their mass, of a peptide with a molecular formula of $C_vH_wN_xO_yS_z$ is actually smaller than $2^v2^w2^x3^y4^z$. For completeness, we should discern among three types of isotopic variants. First, mathematically, we can define a set of all possible isotope location variants and neglect molecule symmetry. Second, chemically, previous set can be reduced taking into consideration the symmetry in the molecule. Third, analytically, we can further reduce the set towards only isobar isotopic variants.

Initially, to calculate the isotopic variants, the stepwise procedure, introduced by Biemann [4][4][4][4] [4] in 1962, was used. The procedure is based on the isotope ratio of the elements that compose the molecule. Th isotopic distribution is calculated by the stepwise addition of the isotope ratio information of a particular atom to accumulate to the full molecule. For example, a chlorine atom has two isotopes: ^{35}Cl and ^{37}Cl, with a natural occurrence probability of 0.7577 and 0.2423, respectively. Note that the probability of prevalence for ^{35}Cl is approximately three times larger than the probability of occurrence of ^{37}Cl, what we can represent by odds 3:1. Now, the isotopic distribution of a molecule composed out of exactly two chlorine atoms can be calculated by

multiplying the odds corresponding to the first atom (3:1) by the odds corresponding to the second atom (3:1). This calculation leads to the odds of 9: 3: 3: 1 for the nominal masses (35+35), (35+37), (37+35), and (37+37), respectively. Summing the terms with equal mass gives rise to the odds of 9: 6: 1 for the isotopic variants with nominal masses 70, 72, and 74, respectively. This procedure is repeated by atom-wise adding and multiplying of the odds for each chemical element, until all the polyisotopic elements have been considered. By doing so, "stick spectra" are generated with one stick per dalton. The approach is very laborious, and it is impossible to execute it for large molecules with many polyisotopic atoms. For this reason, the stepwise procedure was considered only for small molecules and not for biomolecules. Another drawback of the method is that the probabilities (real numbers) of occurrences of different isotopic variants of a molecule were rounded to obtain odds expressed by ratios of integers (rational numbers), that can lead to bias in the calculation of the isotopic distribution for large molecules. Although the method is straightforward to use, it consumes a tremendous amount of computer memory and processor time. Many variants of the step-wise algorithm were introduced to aim at the most efficient use of these resources [30, 17][30, 17][30, 17][30, 17][30, 17]. A case in point is the method proposed by Kubinyini[18][18][18][18] [18], where the calculation is subdivided into a binary series of steps. Instead of the atom-wise addition, Kubinyini introduced the concept of hypothetical clusters of atoms, called hyperatoms, e.g., C_4 . Splitting up the calculation of the isotopic distribution of a molecule into the multiplication of the odds of larger hyperatoms will improve the computational efficiency. How this calculation is achieved will be explained in more detail at the of the section about complexity reduction.

In the seventies, Yamamoto and McCloskey [49][49][49][49][49] and Brownawell and Fillippo [5][5][5][5][5] argued that, for large molecules, the isotopic distribution could be easily obtained by symbolically expanding a polynomial function. In the case of peptides, this becomes:

$$\left({}^{12}C + {}^{13}C\right)^{v} \times \left({}^{1}H + {}^{2}H\right)^{w} \times \left({}^{14}N + {}^{15}N\right)^{x} \times \left({}^{16}O + {}^{17}O + {}^{18}O\right)^{y} \times \left({}^{32}S + {}^{33}S + {}^{34}S + {}^{36}S\right)^{z} \quad (2)$$

The terms ^{12}C, ..., ^{36}S are symbolic indicators that refer to the corresponding atoms. Symbolic expansion of equation (2) results in many equivalent isobaric product terms, which correspond to molecules with the same isotope combination, but at a different location. As we have already mentioned, the isobaric peptide variants, i.e., all variants with identical isotope composition and equal mass, are indistinguishable. Therefore, these isotopic variants, often referred to as like-terms, can be collected. Collecting these equivalent terms and substituting the probabilities and masses from Table 1 into the symbolic indicator gives the probability of occurrence and its exact-mass value for the corresponding isotopic variant of the peptide.

In the nineteen-eighties and nineties, the symbolic polynomial expansion was explicitly performed on computers [27, 14, 29, 8][27, 14, 29, 8][27, 14, 29, 8][27, 14, 29, 8][27, 14, 29, 8]. The advantage, as compared to the stepwise procedure, was that, in the case of the polynomial expansion, the isotopic variants for each element could be calculated separately before further multiplying across the elements. The separate calculation of elemental-specific isotopic variants opened the possibility to perform adjustments to the algorithm. These adjustments could accelerate the calculations. It should also be noted that, in the polynomial expansion, the exact probabilities can be used instead of rounded ratios.

To illustrate the concepts, we will present an example of the symbolic expansion for propane C_3H_8. In this case, we can expand the terms for carbon and hydrogen separately. For carbon, this calculation results in 2^3 possible terms:

$$\begin{aligned}&\left({}^{12}C + {}^{13}C\right)^3 \\ &= {}^{12}C\,{}^{12}C\,{}^{12}C \\ &+ {}^{12}C\,{}^{12}C\,{}^{13}C + {}^{12}C\,{}^{13}C\,{}^{12}C + {}^{13}C\,{}^{12}C\,{}^{12}C + {}^{12}C\,{}^{13}C\,{}^{13}C + {}^{13}C\,{}^{13}C\,{}^{12}C + {}^{13}C\,{}^{12}C\,{}^{13}C + {}^{13}C\,{}^{13}C\,{}^{13}C\end{aligned}$$

Note that the eight hydrogen atoms mathematically give rise to 2^8=256 possible isotope location variants. From previous expansion, it can be observed that even a simple molecule like propane can already have as many as 2048 isotopic variants. However, collecting the isobaric variants (like-terms) results in only four terms for the C_3 fragment of propane:

$$\left({}^{12}C + {}^{13}C\right)^3 = \left({}^{12}C\right)^3 + 3\left({}^{12}C\right)^2{}^{13}C + 3\,{}^{12}C\left({}^{13}C\right)^2 + \left({}^{13}C\right)^3$$

In a similar way, collecting the like-terms from the 256 isotopic variants of H_8 results in only nine different isobaric variants. The obtained polynomials for the symbolic indicator C and H should now be multiplied to calculate the complete isotopic distribution for the propane molecule. This multiplication results in 36 terms which represent the isotopic variants with unique masses that range from the monoisotopic mass of 44.062 Da to 55.1228 Da. In order to calculate the isotopic probabilities and accurate mass, each of the 36 product terms should be separately evaluated by replacing the symbolic indicator variables ${}^{12}C$, ${}^{13}C$, ${}^{1}H$, and ${}^{2}H$ by the corresponding occurrence probabilities and masses from Table 1.

The Multinomial expansion

As mentioned earlier, symbolically expanding the polynomial is an exhaustive and computer-intensive method. However, there exists a more elegant way to calculate the frequency, with which a particular product term will appear in the expansion. Already in 1963, Margrave and Polansky[20][20][20][20] [20] described the binomial nature of the probability of occurrence of isotopic variants. This observation was also made by Carrick and Glocklin in 1967 [7][7][7][7][7]. Later, in 1973, Genty[13][13][13][13] [13] used the multinomial distribution to calculate the isotopic ratios for small molecule. In 1983, Yergey et al**.** [50, 51][50, 51][50, 51][50, 51][50, 51] generalized the latter findings and stated that the isotopic distribution for a peptide could be calculated by the expansion of the multinomial distribution, because the following relationship holds:

$$\left(x_1 + x_2 + \cdots + x_k\right)^n = \sum \frac{n!}{n_1! \times n_2! \times \cdots \times n_k!} x_1^{n_1} \times x_2^{n_2} \times \cdots \times x_k^{n_k} \quad (3)$$

n_i is the number of isotopes x_i of element x present in the molecule. Note that the sum in equation (3) is taken over all combination of $n_1, n_2, \ldots, n_k$ such that they sum to n, where n denotes the number of atoms of element x. The search for all possible isotope combinations $n_1, n_2, \ldots, n_k$ is known as the money-exchange problem, and can be seen as the solution of a Diophantine equation, which will be discussed later in this manuscript. The left-hand side of equation (3) represents the symbolic expansion of the polynomial, whereas the right-hand side is its direct representation in terms of a sum of multinomial probability-like terms. The cumbersome process of collecting like-terms for each

possible isotopic variant from the symbolic expansion is now replaced by calculating the multinomial coefficients.

It should be noted that the previous equation is only used to calculate the probability of occurrence of an isotopic variant of a molecule composed out of only one type of element, e.g., C_{112}. The probability of occurrence of an isotopic variant of the entire biomolecule is calculated as the product of the probabilities for the corresponding element-specific isotopic variants. For example, the probability of occurrence of the isotopic variant $^{12}\boldsymbol{C_{110}}\ ^{13}\boldsymbol{C_2}\ ^{1}H_{165}\ ^{14}N_{27}\ ^{16}\boldsymbol{O_{36}}$ of a peptide with atomic composition $C_{112}H_{165}N_{27}O_{36}$ (nucleon number: 2463) is described by the product of the following terms:

$$\Pr\left(^{12}C_{110}\ ^{13}C_2\right) = \frac{112!}{110!\times 2!}\Pr\left(^{12}C\right)^{110}\times\Pr\left(^{13}C\right)^{2} = 0.2180$$

$$\Pr\left(^{1}H_{165}\ ^{2}H_0\right) = \frac{165!}{165!\times 0!}\Pr\left(^{1}H\right)^{165}\times\Pr\left(^{2}H\right)^{0} = 0.9812$$

$$\Pr\left(^{14}N_{27}\ ^{15}N_0\right) = \frac{27!}{27!\times 0!}\Pr\left(^{14}N\right)^{27}\times\Pr\left(^{15}N\right)^{0} = 0.9057$$

$$\Pr\left(^{16}O_{36}\ ^{17}O_0\ ^{18}O_0\right) = \frac{36!}{36!\times 0!\times 0!}\Pr\left(^{16}O\right)^{36}\times\Pr\left(^{17}O\right)^{0}\times\Pr\left(^{18}O\right)^{0} = 0.9161$$

The exact mass of this isotopic variant is 2466.1978 Da, with the probability of occurrence equal to 0.2180 × 0.9821 × 0.9057 × 0.9161 = 0.1775. Now, consider the isotopic variant $^{12}\boldsymbol{C_{112}}\ ^{1}H_{165}\ ^{14}N_{27}\ ^{16}\boldsymbol{O_{35}}\ ^{18}\boldsymbol{O_1}$ with the mass of 2466.1953 Da. Its probability of occurrence is described by the product of the terms with the following *C* and *O* elements:

$$\Pr\left(^{12}C_{112}\ ^{13}C_0\right) = \frac{112!}{112!\times 0!}\Pr\left(^{12}C\right)^{112}\times\Pr\left(^{13}C\right)^{0} = 0.2997$$

$$\Pr\left(^{16}O_{35}\ ^{17}O_0\ ^{18}O_1\right) = \frac{36!}{35!\times 0!\times 1!}\Pr\left(^{16}O\right)^{35}\times\Pr\left(^{17}O\right)^{0}\times\Pr\left(^{18}O\right)^{1} = 0.0678$$

The resulting probability is 0.2997 × 0.9821 × 0.9057 × 0.0678 = 0.0180. It should be noted that, although both isotopic variants $^{12}\boldsymbol{C_{110}}\ ^{13}\boldsymbol{C_2}\ ^{1}H_{165}\ ^{14}N_{27}\ ^{16}\boldsymbol{O_{36}}$ and $^{12}\boldsymbol{C_{112}}\ ^{1}H_{165}\ ^{14}N_{27}\ ^{16}\boldsymbol{O_{35}}\ ^{18}\boldsymbol{O_1}$ have the same nucleon number, i.e., contain the same additional amount of neutrons; their masses slightly differ by **0.0025** Da. This example also indicates that isotopic elements with negligible individual occurrence probabilities have substantial contributions to the isotopic distribution in the middle-molecule mass range and should not be ignored. Previous statement is especially true in the case for carbon, where the probability of occurrence of two different carbon isotopes in the molecule is quite large. Figure 1(a) presents the result of the multinomial expansion of $C_{112}H_{165}N_{27}O_{36}$ for the 166 components in the mass interval 2464-2470 Da. Figure 1(b) presents a zoom-in of the isotopic cluster with nominal mass 2468 Da. The zoom-in indicates that there are many isotopic variants of the molecule with a similar mass and a very small probability of occurrence.

To calculate the isotopic distribution of a molecule, the multinomial expansion should be evaluated separately for all possible isotope combinations of atom. In 1984, Hsu [16][16][16][16][16]

suggested to this aim the use of Diophantine equations. A Diophantine equation is an equation in which the coefficients and solutions are required to be integers, and can be expressed as

$$\sum_{i=1}^{k} n_i = n \quad (4)$$

where *n is* the number of atoms for a particular element and n_i denotes the number of occurrences of the *i*th isotope in a polyisotopic element with *k* isotopes. The solution of the Diophantine equation in (4) returns all possible sets of n_i's. For the calculation of the probability of occurrences for the isotopic variant, Hsu also propagated the use of the multinomial distribution.

It appears that the implementation of the multinomial expansion works well for low-mass peptides. However, for larger molecules, the calculation of the multinomial coefficients becomes problematic due to overflow in the computation of factorials. An overflow takes place when the result of a calculation is larger than the maximum representable number on a computer. For example, a 64-bit floating-point representation in the IEEE Standard 754 will generate an overflow when calculating the multinomial coefficients for a molecule, like bovine insulin $C_{254}H_{337}N_{65}O_{75}S_6$. To circumvent this problem, Yergey and Hsu both proposed to use a logarithm of the multinomial coefficient to facilitate the numerical computations. This representation of the multinomial increases the speed of their algorithms. Further, they propose to compute the multinomial coefficient in a recursive way. This method reduces the number of computer manipulations because the probability of occurrence a new isotope combination is based on a previously calculated isotope combination, which is updated by a ratio computed from the differences between the two isotope combinations. It should be noted that the concept of using ratios to update previous probabilities was also implicitly suggested by Biemann when considering the step-wise addition of atoms as adding abundance ratios**.** To conclude this section, many implementations of the multinomial expansion are available as software packages; e.g., 'IsoPro' from Senko [11][11][11][11][11] and 'Isotopica' from Fernandez-de-Cossio et al. [40][40][40][40][40].

Combinatorial complexity and memory usage

To illustrate the combinatorial complexity of the isotopic distribution, we provide a rule to calculate the memory space required to store the isotopic distribution of a peptide $C_vH_wN_xO_yS_z$. The calculation of the number of possible isotopic variants with different masses for each element separately is straightforward. For example, the number of possible different-mass isotopic combinations that correspond to *v* carbon atoms is equal to (v+1), because the peptide can carry 0 to *v* ^{13}C isotopes. Similarly, hydrogen and nitrogen will contribute (*w*+1) and (*x*+1) isotopic combinations with different masses, respectively. Oxygen has three stable isotopes; therefore, *y* oxygen atoms will result in $\frac{(y+1)(y+2)}{2}$ possible combinations. Finally, sulfur has four stable isotopes, which results in $\frac{(z+1)(z+2)(z+3)}{6}$ isotopic combinations with different masses. The calculation of the isotopic distribution for each element separately does not require much computing time, and is not demanding from a point of view of memory resources. However, to calculate the isotopic distribution for the whole molecule, the numbers of possible isotopic combinations for each element have to be multiplied by each other. This multiplication dramatically increases the total number of combinations to be computed and stored. For example, the multinomial expansion of peptide $C_{112}H_{165}N_{27}O_{36}S_0$ results in 113 x 166 x 28 x 703 = 369,232,472 possible isotopic variants of the peptide. In a standard

software package, storing this amount of data for the masses and probabilities of occurrence in double precision (8 bytes) requires approximately 2 x 2.9539 gigabytes (GB) of memory. Adding one sulfur atom to the example described above would quadruple the complexity, and would require 2 x 11.8154 GB of memory. Note that this amount would be the memory resources required to fully store the isotopic distribution of this peptide, irrespectively of which algorithms are used for the calculation. From previous, it can be seen that the polynomial and multinomial expansion methods have unfavorable scaling properties due to the combinatorial increase of the required storage space and are only suited for molecules in the lower-mass range.

Complexity reduction

To deal with the memory issues mentioned in the previous section, many methods for the calculation of the isotopic distribution employ a pragmatic technique called 'pruning' to control the number of isotopic variants. For example, element-specific isotopic variants can be removed from consideration if their probabilities of occurrence fall below a user-defined threshold. As a consequence, the total number of variants, which need to be stored, is controlled. However, the deletion of an element-specific variant with a low probability of occurrence can result in the removal of an abundant variant for a whole molecule, because the probability will be multiplied by probabilities of isotopic variants for the other elements. This numerical artifact can be observed when comparing Figure 1(a) and Figure 2. Figure 2 represents a pooled version of the isotopic distribution in Figure 1(a), such that isotopic variants with similar nominal mass are summed. The dash-dotted lines in Figure 1(a) and 2 indicate that low-abundance isotopic variants observed in Figure 1(b) contribute substantially to the isotopic distribution shown in Figure 2.

A more-efficient and commonly employed method to reduce the complexity of the isotopic distribution is to restrict the distribution to a mass region in which the isotopic variants experimentally can be observed. For this purpose, a threshold can be determined to stop the procedure after a given number of calculations for an increasing number of isotopes in the molecule. Isotopic variants outside the defined mass range are neglected. For instance, the computations can be stopped when, for each element, only those isotopic variants are considered that contain up to 9 additional neutrons. Truncating the isotopic variants per element within this interval is justified, because, in a mass spectrum due to the dynamic range limitations of the detector, e.g., 37dB, peptide peaks are rarely visible beyond this mass interval of M+9 Da for the particular middle-mass peptide. Further, it is of importance that the isotopic variant with the maximum probability of occurrence in the elemental-specific isotopic distribution is maintained in the mass interval specified by the user.

Applying previous strategy to $C_{112}H_{165}N_{27}O_{36}$ results in a Diophantine solution that considers 10, 10, 10, and 33 isotopic variants for carbon, hydrogen, nitrogen, and oxygen, respectively. A multinomial expansion of the element-specific isotopic distributions yields 33,000 possible variants that contribute to the isotopic distribution of the molecule. Note, that the corresponding 33,000 probabilities add up to 0.999999574429216. This high value indicates that the complete isotopic distribution is approximately covered by the 33,000 components, and that the remaining 369,199,472 components contribute only a miniscule fraction to the molecule's distributional

information. This phenomenon is exploited in the various algorithms for the calculation of the isotopic distribution when pruning the low probable isotopic variants.

As mentioned earlier, the calculation of the factorial in the multinomial coefficient can be problematic for large molecules. Kubinyi [18][18][18][18][18] solved this issue in an original way by subdividing the calculation into steps of a binary series. For example, to calculate the possible terms for C_{256}, one could reformulate the problem as $\left(\left({}^{12}C + {}^{13}C\right)^{128} \times \left({}^{12}C + {}^{13}C\right)^{128}\right)$,and apply the stepwise procedure. The factorial of 128 can be easily represented as a 64-bit floating-point number. This calculation can be further subdivided by the series C_{64}, C_{32}, C_{16}, C_8, C_4 and C_2, if necessary. These hyper-atoms [18][18][18][18][18] or super-atoms [38][38][38][38][38] can be pre-calculated and stored in a look-up table for further calculations; that process leads to an accelerated computation.

However, the most-efficient method to reduce the complexity is achieved by taking into consideration only nominal masses of isotopic variants; i.e., isotopic variants with the same nucleon number. This premise is justified, because the resolution of a MS instrument is finite; e.g., 10000. Therefore, in a mass spectrum, isotopic variants with the same nominal mass are pooled and accumulate to an aggregated probability. It should be stressed that, by pooling the isotopic variants in mass bins of approximately 1 Da length, we no longer deal with the isotopic distribution, but with an *aggregated isotopic distribution*. For example, applying such a strategy to the 166 isotopic variants in Figure 1(a) would lead to an aggregated isotopic distribution with only 6 aggregated isotopic variants, which would combine the information about the 166 variants. This topic will be fully discussed in the next section.

PART II: the methodology of Alan Rockwood

Nominal mass and the nucleon number

The masses of the isotopic variants of a molecule can be easily computed by summing the masses of the isotopes that compose the molecule. For example, the mass of isotopic variant $^{12}C_{110}\ ^{13}C_2\ ^{1}H_{165}\ ^{14}N_{27}\ ^{16}O_{36}$ of a peptide with atomic composition $C_{112}H_{165}N_{27}O_{36}$ is calculated as *110 mass(^{12}C) + 2 mass(^{13}C) + 165 mass(^{1}H) + 27 mass(^{14}N) +36 mass(^{16}O)*. As mentioned earlier, isotopic variants with the same nucleon number can have different masses. However, one can easily show that, for isotopic combinations with the same nucleon numbers, the exact mass is spread less than 50 ppm [7][7][7][7][7]. Therefore, it is reasonable to consider only nominal masses, which aggregate isotopic variants with the same nucleon number. For example, consider the isotopic variants $^{12}C_{110}\ ^{13}C_2\ ^{1}H_{165}\ ^{14}N_{27}\ ^{16}O_{36}$ and $^{12}C_{112}\ ^{1}H_{165}\ ^{14}N_{27}\ ^{16}O_{35}\ ^{18}O_1$ of a peptide with atomic composition $C_{112}H_{165}N_{27}O_{36}$. They have slightly different masses, 2466.1978 and 2466.1953, respectively. In a mass spectrum, the isotopic variants would not appear as separate contributions, but as a single peak or equivalently, as aggregated isotopic variants , because they have the same nucleon number of 2465.

The center mass m_{center} of such an aggregated isotopic variant can be calculated as a probability-weighted sum of the masses of the *k* isotope combinations that contribute to this variant, as defined by Roussis and Proulx [38][38][38][38][38]:

$$m_{center} = \frac{\sum_{j=1}^{k} p_j \times m_j}{\sum_i p_i} \quad (5),$$

where p_j denotes the probability of occurrence of isotopic variant with mass m_j.

It is worth noting that all isotopic variants with equal nucleon number contribute to the aggregated isotopic variant and therefore are required to accurately compute the mass. Employing a 'pruning' to reduce the complexity of the calculation would result in missing isotopic variants and would compromise the calculation of the accurate mass of a aggregated isotopic variant[37, 32][37, 32][37, 32][37, 32] [37, 32].

An interesting characteristic of aggregated isotopic variants is that they are separated in mass by approximately 1 Da. Instead of using the masses to denote the aggregated variant, we can make abstraction of the mass and regard only the additional neutron content. In this respect, the symbolic expansion of the polynomial in equation (2) can be reformulated in function of the additional neutron content:

$$\left(\Pr(^{12}C)\,I^0 + \Pr(^{13}C)\,I^1\right)^v \times \left(\Pr(^{1}H)\,I^0 + \Pr(^{2}H)\,I^1\right)^w \times \left(\Pr(^{14}N)\,I^0 + \Pr(^{15}N)\,I^1\right)^x \times \left(\Pr(^{16}O)\,I^0 + \Pr(^{17}O)\,I^1 + \Pr(^{18}O)\,I^2\right)^y \times \left(\Pr(^{32}S)\,I^0 + \Pr(^{33}S)\,I^1 + \Pr(^{34}S)I^2 + \Pr(^{36}S)\,I^3\right)^z \quad (6)$$

where the terms Pr(^{12}C), Pr(^{13}C), ..., Pr(^{36}S) denote the probability of occurrence of the element-specific isotopes, as displayed in Table 1. Note that the power of the symbolic indicator I^P represents the additional neutron content with respect to the monoisotopic variant. With this extra indicator variable, it is easier to symbolically expand and multiply the polynomials, what implies a serious reduction in the number of possible isotopic variants. For instance, a multinomial expansion for O_{75}, which yields 703 isotopic variants, could be replaced by 73 aggregated isotopic variants with distinct masses separated by approximately 1 Da or, equivalently, with the same additional neutron content.

In summary, molecules which have the same additional neutron content, can be pooled together to represent an aggregated isotopic variant. The mass deficit which possibly occurs, can be ignored for instruments with a limited resolution. Interestingly, by considering the neutron count, we can easily relate the polynomial expansion to the convolution which is a mathematical operator. Algebraically, convolution of the coefficients of the polynomials corresponds to multiplication of the polynomials, as described by Rockwood in 1995 [31][31][31][31][31]. [26][26][26][26][33][33][33][33][25][25][25][25]

The convolution method by the Fourier transform

Between 1995 and 1996, Rockwood et al. published a series of papers that discussed the relationship between the multiplication of polynomials and the convolution by means of a Fourier transform method[31, 35][31, 35][31, 35][31, 35] [31, 35, 36, 34]. For convenience, we start the discussion with the method proposed in 1996 [34], which uses the convolution theorem to produce a 'stick' mass spectrum. Rockwood and colleagues obtained the convolution via the Fourier transform by replacing the symbolic indicator I, used in equation (6), by the Fourier term $e^{i2\pi m\mu}$, where i denotes the imaginary number, m indicates the nominal mass, and μ is a real number between 0 and 1. The parameter μ can be interpreted as a grid over which the harmonic function $e^{i2\pi m\mu}$ is evaluated, or

equivalently, sampled. The number of points in the grid defines the resolution of the result. By making these substitutions to equation (6), we obtain the following expression

$$\left(\Pr\left({}^{12}C\right)e^{i2\pi 12\mu} + \Pr\left({}^{13}C\right)e^{i2\pi 13\mu}\right)^{v} \times \left(\Pr\left({}^{1}H\right)e^{i2\pi 1\mu} + \Pr\left({}^{2}H\right)e^{i2\pi 2\mu}\right)^{w} \times \left(\Pr\left({}^{14}N\right)e^{i2\pi 14\mu} + \Pr\left({}^{15}N\right)e^{i2\pi 15\mu}\right)^{x} \times \left(\Pr\left({}^{16}O\right)e^{i2\pi 16\mu} + \Pr\left({}^{17}O\right)e^{i2\pi 17\mu} + \Pr\left({}^{18}O\right)e^{i2\pi 18\mu}\right)^{y} \times \left(\Pr\left({}^{32}S\right)e^{i2\pi 32\mu} + \Pr\left({}^{33}S\right)e^{i2\pi 33\mu} + \Pr\left({}^{34}S\right)e^{i2\pi 34\mu} + \Pr\left({}^{36}S\right)e^{i2\pi 36\mu}\right)^{z} \quad (7)$$

It should be noted that the exponential $e^{i2\pi m\mu}$ represents a harmonic function, for which the frequency of oscillation is determined by the discrete mass value *m*. By casting the calculation of the polynomial expansion in function of harmonics, i.e., sines and cosines, we enter the field of digital signal processing. Consequently, the jargon, like 'Shannon-Nyquist-Kotelnikov criterium', 'subsampling', 'heterodyning ', etc. appears repeatedly in the manuscripts by Rockwood *et al.*

The important advantage of the use of the Fourier transform is that a multiplication is equivalent to a convolution after the Fourier transformation (and vice-versa). This property can now be exploited by evaluating expression (7) and applying the Fourier transform. Afterwards, the terms are multiplied with each other and transformed back to the original space with the inversed Fourier transform. This computational process can be seen as an analogy of the logarithmic transformation, which allows one to replace multiplication with addition.

Practically, the calculation of the aggregated isotopic distribution is obtained with the Fast Fourier Transform (FFT), and yields a distribution for which the aggregated isotopic variants are equally spaced with integer values, as can be observed in Figure 2. Note that the aggregated isotopic variant near 2468 Da in Figure 2 is twice as likely as the isotopic variant with the maximum probability of occurrence in the isotopic distribution near 2468, observed in Figure 1(b). As stated previously, this example clearly demonstrates that the accumulation of low-probability variants, that are measured by mass spectrometry, have a substantial contribution to the aggregated isotopic distribution of a molecule.

After the extremely fast calculation of the aggregated isotopic distribution based on integer point spaces, a correction is applied such that masses are placed at non-integer values. The non-integer masses are "semi-accurate", i.e. typically a few millidalton, and are based on a linear transformation of the mass scale. This transformation forces the new distribution to have the correct mean and standard deviation based on information from the monoisotopic and average masses. It is worth mentioning that, despite the suboptimal mass accuracy for the aggregated isotopic variant, the correct isotopic probabilities are obtained with the convolution method. This benefit in probability precision is achieved because no "explicit" pruning technique is used. As a consequence, no information about low-probability isotopic variants is lost. Instead, the variants are aggregated in distinct, equally spaced bins with semi-accurate mass. Obviously, numerical errors due to digital filtering, FFT and rounding will occur, but these disturbances are minor compared to the loss of precision due to pruning. Later in this manuscript, we dedicate a paragraph to explain what exactly is meant by "none-explicit" pruning.To relate the polynomial expansion to the convolution, a trade-off has to be made between the isotopic fine structure and computational speed. Whether this trade-off is acceptable depends on the application in which information about the isotopic distribution is used.

For example, in a peptide-centric proteomic setting, the isotopic fine structure is usually not of interest. However, when the focus is on elemental speciation analysis, the fine structure becomes much more important.

In the manuscript published in 1995 [35], Rockwood *et al.* approached the isotopic distribution calculation differently by combining the computation of the isotopic variants with a peak shape convolution. The method described in this manuscript produces profile-mode distributions instead of the 'stick' spectrum discussed earlier. Profile-mode means that on top of the 'stick spectrum' a particular shape is convoluted as indicated by the dashed lines in Figure 3. Contributions of the convoluted isotopic variants (dashed lines) are aggregated to form a continuous function (line), which is sampled at equidistant intervals (stars). It should be noted that the aggregated isotopic variants are placed at accurate masses and no longer at integer values as described previously. This degree of accuracy can be achieved by substituting the masses from Table 1 in the corresponding Fourier terms, $e^{i2\pi m\mu}$, from equation (7). [36, 34][36, 34][36, 34][36, 34]

In summary, the method is able to calculate the aggregated isotopic distribution in profile-mode with accurate masses. It should be pointed out that in this scheme; the aggregated isotopic distribution is represented as a sampled version of a continuous function with finite peak width, instead of a list of isotopic variants. In theory, it is even possible to resolve the isotopic fine structure by adjusting the width of the peak shape and the sampling (a "stick spectrum" can be seen as having infinite small peak width). Factors that determine the sampling or, equivalently, the resolution of the convolution method, are the array size of μ and the frequency of the harmonics controlled by the variables *m;* the masses of the chemical elements. Hence, by increasing the array size and decreasing the peak width, one can do a calculation that resolves the isotopic fine structure, but the manuscript does not present this ultrahigh-resolution because array size was restricted by the computer compiler and limited memory resources [2].

Rockwood *et al.*, presented a variant of the profile-mode distribution in 1996 [36] that does not require a large array size for the calculation of the isotopic fine structure. This method also uses Fourier transforms. By zooming into a limited mass range (typically 0.2 Dalton) and focusing the sampling in that region, an isotopic distribution with a higher resolution is obtained. Technically, this zooming involves heterodyning and digital filtering in order to shift the isotopic distribution towards 0 Da. Via subsampling, the available datapoints, i.e., the array size of μ, are confined towards a particular region such that a higher resolution (millidalton) is obtained. This is also a sampled version of a profile-mode distribution with a peak shape convoluted into the results. This calculation is not particularly fast, primarily because of the need to do digital filtering, but it does allow one to do an ultrahigh-resolution calculation using a small array size. It should be noted that filtering and subsampling to a specific mass window can attenuate the signal.

[34][34][34][34]To reiterate, the FFT-based convolution method does not seem to require an "explicit" user-defined threshold in order to improve computational speed, though the choice of the array size of μ, used to sample or bin the isotopic variants, can be seen as a global threshold. Nevertheless, we should stress that, in the approach of Rockwood *et al.*, isotopic variants are assembled when the resolution is limited, instead of being pruned. In this respect, an analogy with MS instrumentation can be made. In a low-resolution mass spectrometer that does not resolve the isotopic variants of a molecule such that only one molecule peak is visible in a spectrum, no ions are

omitted from this peak, but they are pooled over time on the mass detector. However caution should be applied. If the mass range defined in μ is inadequate then there will be a kind of aliasing in the calculation. For example, high mass peaks that are outside of the computational window will wrap around and appear at the low mass end of the mass window. This distortion will be dependent on the implementation of the algorithm, and in particular whether the implementation uses a strategy that selects a mass window that is sufficiently wide or is externally supplied. To conclude, previous versions of the algorithm are implemented in the Mercury software. Mercury contains two methods for calculating isotope distributions, one giving the full isotopic distribution and the other giving an ultrahigh resolution picture of a single nominal isotopic variant.

In 2004 Rockwood and colleagues changed focus towards the calculation of isotopic combinations and the exact-center masses of the aggregated isotopic variants, i.e., no resolved isotopic fine structure. This algorithm does not require the use of Fourier transforms[37, 32][37, 32][37, 32][37, 32] [37]. Thus far, we presented the aggregated isotopic variants as labeled with an integer number; i.e., the nucleon number or with a semi-accurate mass obtained via a linear transformation. Little attention was given to the computation of the exact-center masses for the aggregated isotopic variants. Exact-center masses are especially important for large molecules because accumulation of the mass defect causes considerable deviation from the nominal mass. Moreover, Thurman and Ferrer proposed to use this phenomenon as a tool to filter molecular formulas[45][45][45][45] [45].

One approach to compute the accurate mass is to determine first all isotopic combinations that contribute to the aggregated variant. Based on the obtained isotopic combinations list, the exact masses and the probability of occurrence can be calculated. As mentioned earlier, the center mass of the aggregated isotopic variant can now be computed based on the previous isotopic probabilities and the exact isotopic masses, as stated by Roussis and Proulx [38][38][38][38] [38]. Therefore, Rockwood *et al*. suggest a two-step procedure that is based on the idea that one can determine the isotopic composition of an aggregated isotopic variant by removing atoms from the molecule. First, an aggregated isotopic variant of the molecule for which exact mass is required, is selected. Keep in mind that this calculation will produce the correct probabilities if the FFT-method in [34] is used. Second, for each element in the molecule, the isotope contribution to the selected aggregated isotopic variant is calculated. The isotope contribution is based on the intermediate distribution of the molecule for which an atom is omitted. Together with the probability of occurrence associated to the isotope of each element, the isotope contribution can be calculated. [18][18][18][18][38][38][38][38]In general this algorithm is intermediate in speed between the method presented in references [35] and [36].

A variant on previous theme was published in 2006 [32]. This method also calculates the exact-mass center and probability of a selected aggregated isotopic variant, as does the method in the previous paragraph, but it does not generate composition information. Instead the method breaks the calculation down to a series of direct convolution steps arranged in a certain order to minimize the time required for the calculation. Internally it throws away the information required to resolve isotopic fine structure, but it keeps the accurate mass information. The procedure is strongly related to the step-wise procedure of Kubinyini][18], because it uses the concept of super-atoms to facilitate the calculations. Finally, to calculate the exact center mass, it uses the probability-weighted

sum of masses in equation (5) proposed by Rousis and Proulx][38], where the probabilities represent the contribution of the elemental isotopes.

In 2003, Rockwood and co-authors extended the idea of the convolution towards the problem defined by O'Connor *et al.* [26] for the isotopic assignment of product-ion spectra from a tandem MS experiment, when only a single non-monoisotopic isotopic peak of the precursor ion was selected for fragmentation [33]. The approach can be seen as the reverse of the stepwise-addition of Kubinyini, because in each step the precursor ion is deciphered into smaller "molecules" to calculate the isotopic distribution of the neutral fragment that leaves the precursor ion. Meija and Caruso also apply a separation strategy to the aforementioned problem and use the method of Rockwood *et al.* for the isotope pattern deconvolution of isobaric interferences in tandem MS [25].

PART III: Current progress

Modern methods

As discussed by Roussis and Proulx in 2003[38][38][38][38] [38], the stepwise addition of hypothetical atom clusters in Kubinyi's algorithm is computationally efficient, produces the correct probabilities, and does not have the unfavorable scaling properties of the polynomial and multinomial method. However, a limitation is that it does not yield the correct masses, because it assumes that the isotopic variants have mass differences of exactly one Dalton; i.e., they also use aggregated isotopic variants. To resolve the mass-accuracy issue, Rousis and Proulx proposed to calculate the accurate mass of the aggregated isotopic variant by a probability-weighted sum of the masses of the isotopic variants that contribute to that aggregated variant. Whenever a new 'superatom 'or 'hyperatom' is added, the mass is updated accordingly. Kubinyini defined a hyperatom as a cluster with 2, 4, 8,…, 2^n atoms of the same type; e.g., C_2 , $C_{4,}$ … , Cn. Rousis and Proulx extended this idea to superatoms which are hypothetical atom clusters across the different atoms; e.g., CO. Latter approaches can be seen as a built-up method, where the isotopic probabilities and masses are calculate by stepwise combination of hyper- or superatoms. In contrary, the approach of Rockwood et al. [37, 32] can be seen as a break-down method where the aggregated isotopic probabilities are decomposed into clusters of isotopic variants.

In 2007, Snider [43][43][43][43][43] presented a variant of Biemann's stepwise addition of elements algorithm based on dynamic programming. Instead of working with aggregated isotopic variants, Snider proposed to keep track of all obtained isotopic variants when adding a single element to the molecule. The probability related to each isotopic variant after addition of an element depends completely on the probabilities of the isotopic variants before the addition and on the isotope probabilities of the added element. To efficiently calculate the isotopic distribution, Snider frames the problem in terms of a Markov process that operates on a discrete state space.

The different states can be interpreted as the isotopic variants, whereas the discrete steps correspond to the addition of elements. Obviously, at each discrete step, the number of possible states or, equivalently, the isotopic variants increases. Figure 4 shows how the Markov process looks in the case of carbon monoxide. At step 1, only two states are available. In this case, the probabilities of these two states correspond to the probability of occurrence of the carbon isotopes. In step 2, oxygen is added and the number of possible states becomes equal to six. According to the Markov property, the probabilities of future states depend on the probabilities of the current state. Future

probabilities are calculated by summation over all possible state sequences, while taking into consideration the isotopic distribution of the newly added element. The probabilities that correspond to the states can now be efficiently calculated via the forward algorithm, which is a form of dynamic programming. Snider also recognizes that a brute-force calculation of all possible isotopic variants is not feasible given the vast amount of data. To deal with the intrinsic numerical complexity, two reduction modes were proposed:

1) Most-Probable Exact Masses: after adding an element, only the N_{max} most-probable states with exact masses are maintained. Note that eliminating states will prune potential path combinations in the Markov chain, and therefore will distort the probability values, as described by Rockwood [31, 35, 36, 34][31, 35, 36, 34][31, 35, 36, 34][31, 35, 36, 34][31, 35, 36, 34].
2) Exact Probability Distribution: after adding an element, combine the probabilities of isotopic variants that are closest together in terms of mass values, until N_{max} states are maintained. This binning approach is similar in spirit to the one proposed by Rockwood . However, the method of Rockwood, i.e., with fixed array size, is a global binning of isotopic variants, whereas the method proposed by Snider groups the isotopic variants locally. Local binning uses memory resources in a more optimal manner than global binning. The mass for binned isotopic variant is calculated as a probability-weighted sum of the masses [38][38][38][38][38].

In 2008, Li *et al.* [19][19][19][19][19] rediscovered the properties of the Markov process to calculate the isotopic distribution, but called it a hierarchical algorithm, by making an analogy with atomic physics, which relates the possible isotope configurations to the excitation states of atoms. They described the isotopic variants with equal nominal mass as the main electron shells of atoms which are symbolized by the principal quantum number. This representation corresponds to the gross structure of line spectra. The fine structures are caused by spin-orbit coupling, and describe the splitting of the spectral lines of atoms, which represent the isotopic variants within an aggregated isotopic variant. In other words, isotopic variants with an equal nucleon number are grouped at the same level. The top level (level 0) only contains the lightest isotope (ground state) of that molecule; the bottom level only contains the heaviest isotope (highest excited-state) of that molecule. So keep in mind that each level corresponds to a particular nucleon number. For each level, the split across the possible isotope combinations is calculated. To keep track of the isotope combinations, a bookkeeping device, similar in spirit to the ones proposed by Kubinyini [18][18][18][18][18] and Rockwood [32][32][32][32][32], is used to record the isotope combination as a binary series. If the levels are arranged by nucleon number, then they form a hierarchical structure, as can be seen in Figure 5. The difference in nucleon numbers between adjacent levels is equal to one. As a consequence, the Boolean difference between the binary representations of isotope combinations between adjacent levels is also equal to one.

The hierarchical aspect comes from the fact that the probabilities of the isotopic variants in a higher level, i.e., for a higher nominal mass, only depend on the probabilities of the isotopic variants in the previous level. In fact, the hierarchical algorithm of Li et al. is a different representation of the Markov process proposed by Snider. The probabilities of transition from one excitation level, i.e., isotope combination, to a neighboring level are calculated via the recursive formula already

proposed by Yergey [50][50][50][50][50] to avoid repetitive calculation of the factorial. In this respect, Yergey implicitly assumed a form of hierarchy for the calculation of the isotopic distribution.

The fine structure of the isotopic distribution can be calculated as an iterative process that starts with the ground state. To calculate the next level, or equivalently, isotopes with a higher nucleon number, only the information about the two previous levels should be kept in memory. To save memory resources, the information for each level is output separately to the hard-drive.

For large molecules, the number of isotopic variants in a level, or equivalently, aggregated isotopic variant, can be large. Therefore, Li et al. limit the possible number of isotopic variants in a level to 100,001 and propose a truncation scheme to calculate only the major mass states of molecules. For this purpose, they use the monotonic decreasing property of isotopic probabilities once it has reached a maximum value. After this maximum, an isotopic variant with a low probability in a particular level will have a lower probability in the next level; i.e., 1 additional neutron. This truncation scheme is similar in spirit to the pruning approach discussed in the section on complexity reduction, where isotopic variants for each element can be restricted to a certain mass window.

As mentioned earlier, Snider and Li *et al.* both present the calculation of the isotopic distribution as a Markov process, though there are some fundamental differences. They both interpret the possible isotope combinations as states, but the transition is differently formulated. A transition in the Markov model of Snider can be seen as the addition of an atom, whereas the transition in the context of Li *et al.* can be seen as the addition of a neutron. Li *et al.* start with a molecule with no isotope present, and gradually go through all the possible combinations per level. Snider starts with a single atom and considers all possible isotope combinations for each newly formed molecule. Only the last step, i.e., the entire molecule, returns the final isotopic distribution. Calculations for the previous "molecules" can be seen as intermediate steps. In contrast, the method of Li *et al.*, generates the complete isotopic distribution from the beginning; each step adds all isotopic variant within an additional neutron. For example, if only the variants up to M+4 Da are needed, then the calculation can be stopped at level 3. Such an approach is not possible in the case of the method proposed by Snider.

More recently, Olson and Yergey [28][28][28][28][28] proposed an alternative approach to calculate the isotopic distribution. Instead of pruning or aggregating low-probability isotopic variants at the molecule level, Olson and Yergey introduce a virtual element which enables them to calculate the isotopic distribution in terms of equatransneutronic isotopes (ETN). The ETN isotope is a group of isotopes that differ from their element's most abundant isotope by the same number of neutrons. In other words, elemental isotopes with the same number of additional neutrons are clustered. In the case of peptides, which are composed out of C, H, N, O, and S, three sets of ETN isotopes can be defined. Namely, isotopes with 1 additional neutron [^{13}C, ^{2}H, ^{15}N, ^{17}O, ^{33}S], isotopes with 2 additional neutrons [^{18}O, ^{34}S], and isotopes with 4 additional neutrons [^{36}S]. Next, the probability of occurrence for each ETN is calculated by taking into consideration the number of elements in the molecule and their elemental isotopic distribution. The calculation of the molecule's isotopic distribution is defined in terms of the ETN's instead of the individual elements. As a result, the solution of the Diophantine equation becomes feasible, and the computation of the probabilities is more efficient. By grouping elemental isotopes according to their additional neutron content, Olson and Yergey actually present the masses of the elemental isotopes with their nominal mass, similarly to the stepwise addition

proposed by Biemann[4][4][4][4] [4] and Kubinyini[18][18][18][18] [18]. In fact, the method is the probabilistic explanation of the convolution and digital filtering approach of Rockwood *et al.* It is worth keeping in mind that Rockwood *et al.* implicitly assumed discrete elemental masses. This assumption can be observed in equation (6), where the indicator variable *I* is phrased in terms of the additional neutron count. Rockwood *et al.* did not directly compute the combinatorial complexity via, e.g., the Diophantine or multinomial coefficient, but rather counted it via the FFT-variant of the polynomial method.

It is straightforward to see that working with discrete element masses facilitates the calculation and circumvents the need for pruning, as argued earlier. Nevertheless, mass accuracy remains an important issue. Although the method of Olson and Yergey adds in a correction for the mass of the most-abundant isotope, it suffers from an inaccuracy, especially for high-mass molecules.
This issue can be solved by the methods which enable high-accuracy mass calculations for the aggregated isotopic variants, as proposed by Rockwood *et al.,* [37, 32][37, 32][37, 32][37, 32][37, 32].

In 2010, Fernandez-de-Cossio [10, 9] extended the idea of Rockwood et al. [31, 35, 36] to use the FFT for the calculation of the isotopic distribution. They propagate the use of the two-dimensional FFT, in order to split the calculation of the isotopic distribution into separate dimensions. The extreme levels of detail, i.e., the coarse structure (Da) and details about the fine structure, i.e., millidaltons, allow to concentrate the sampling in the informative surroundings of the isotopic variants. By doing so, a more optimal and efficient algorithm is obtained which is implemented in the DEUTERIUM software.

Educational aspects

Isotopic information is crucial to interpret mass spectra, and to gain insight in the structure of molecules. However, the concept of isotopic variants and their distribution can be difficult, especially for students with little training in mathematics. For this purpose, Sein [39][39][39][39][39] propagates the use of Punnett squares in order to make an analogy between isotopic variants and genetic trait variants. Sein argues that most university students are familiar with the concept of Punnett square from basic courses in Medelian law, and that this multiplication table can conveniently represent the probabilistic formulation for the calculation of isotopic distributions. For example, Figure 6 depicts the normalized Punnett square for a molecule composed of two chlorine atoms. The areas of the boxes are proportional to the probability of the isotopic variant. Isotopic variants with equal nominal mass are displayed in the same color. Using Punnet squares as a visual aid, would help understanding the formation of isotopic variants

Molecular Isotope patterns are combinatorial convolution products of atomic isotopes. To facilitate the understanding about binomial combinatorics, the result of a binomial expansion can be arranged in a triangle; i.e. the Pascal's triangle. To this aim, Meija organized a challenge where the fractal patterns of a Sierpinski triangle were given, whereas the competitors had to infer the element responsible for this isotopic pattern [22, 23][22, 23][22, 23][22, 23][22, 23]. Furthermore, related to the use of Pascal's and Sierpinski's triangle, Meija suggested that cellular automata could be used to graphically represent the construction of the isotopic distribution [24][24][24][24][24].

As a didactic example in numerical mathematics, the calculation of the multinomial expansion for large numbers could be issued. From previous discussion, we know that the calculation of isotopic distributions for large molecules via the multinomial distribution is problematic. The first problem is that the multinomial coefficient in equation (3), e.g.,

$$\Pr(^{12}C_{700}\,^{13}C_{300}) = \frac{1000!}{700!\times 300!}\,\Pr(^{12}C)^{700}\times \Pr(^{13}C)^{300}$$

cannot be calculated due to the overflow when evaluating the factorial. A second problem is the multiplication of very small values, because of the high powers the probabilities are raised to. For example, calculating the probability $\Pr(^{13}C)^{300} = 0{,}0107^{300}$ which can lead to underflow (cfr. overflow). Naively, the first problem can be solved by vectorizing the calculation in the following algorithm:

- Vector A = [701, 702, …, 1000]
- Vector B = [1, 2, …, 300]
- C = elemental-wise division of vector A and vector B
- D = product of the elements in C

D is equal to the multinomial coefficient described above. The calculation of D circumvents the direct evaluation of the factorials.

The calculation can be further simplified by taking the logarithm of the multinomial distribution. This transformation leads to the calculation of

$$\log\Pr(^{12}C_{700}\,^{13}C_{300}) = \sum(\log A - \log B) + 700\log\Pr(^{12}C) + 300\log\Pr(^{13}C)$$

for which the exponential is taken afterwards. Simple rewriting the formulae improves numerical stability and avoids the use of approximations for the multinomial distribution. Note that the aforementioned formulation can easily be extended for elements with more than two isotopes.

The method of Monte Carlo sampling, mentioned by Senko *et al.* [41] and Rockwood *et al.* [37], is worth further exploring to generate the isotopic distribution.. It is well known that Monte Carlo methods can be used to break the computational efficiency logjam for some problems. However, the main problems with Monte Carlo methods are that:

1) the results are non-deterministic (assuming a different seeds for the starting of the random number generator)
2) convergence to the true answer slows dramatically as the calculation proceeds because it converges as the square root of the number of sampling points.

Nevertheless, convergence to a usefully accurate result may be faster than polynomial-based methods which suffer from a combinatorial explosion of terms when applied to large problems. As an additional benefit, it is possible to get some statistical estimates of the variability one might expect from an experimental result obtained with a finite number of ions. Further, Monte Carlo simulations can be used to simulate ultrahigh-resolution spectra.

Conclusion

A number of methods have been employed to elucidate the isotopic distribution and the aggregated isotopic distribution of biomolecules. Every method tries to describe the multinomial character of the isotopic distribution, and deals with computation efficiency and memory resources in different ways.

A broadly used method to calculate isotopic distributions is the polynomial method and the multinomial expansion. Straightforward implementations of the polynomial method on computers are computationally involved and memory intensive, particularly for large biomolecules like DNA and proteins. Alternate approaches which focus on the calculation of the aggregated isotopic distribution of a molecule have been proposed. Some approaches trade accuracy for speed. This trade-off, is often justified in a mass spectrometry-based peptide-centric proteomic approach where the infinite resolution of the polynomial method is often not required. All these approaches, however, have two things in common. First, they all try to avoid the combinatorial explosion with a pruning or binning strategy. Second, they all contain a user-specified threshold to control the accuracy and computational speed.

The main reason for the development of the different methods is that, in the early eighties, due to improved mass spectrometry instruments, the calculations of isotopic distributions became of importance. But back then, computer power and memory were limited. To solve this problem, creative algorithms which focused on numerical stability, memory-usage, and computation time were needed.

As a first bottleneck, the multinomial expansion becomes unfeasible for large molecules on computers and underflow occurs because of the multiplication of low probabilities. Modern methods like, e.g., dynamic programming, are used to circumvent the calculation of the multinomial coefficient [43][43][43][43][43]. However, inventive numerical representations to calculate the multinomial expansion [50][50][50][50][50] or approximations of the multinomial coefficients via Stirling's approximation can also solve the problem.

To circumvent the second bottleneck related to the large number of possible isotopic variants, a user-defined threshold can be used to limit the number of isotopic variants. Alternatively, information about isotopic variants can be stored on hard-drives. If accurate masses are required in a mass region, i.e., for a specific nominal mass with low-probability isotopic variants, i.e., region not covered by the user-defined threshold, then it is still possible to calculate the isotopic terms by restricting the variants to the desired mass interval. This idea is in the same spirit as described by Rockwood *et al.* [36, 37][36, 37][36, 37][36, 37][36] and Li *et al.* [19][19][19][19] [19].

To conclude, the multinomial expansion of the separate elements, followed by a polynomial expansion across the elements, is the mathematical method that describes most closely the underlying physical mechanism, by which isotopic distributions are generated. However, the method often computes more details than required. If the fine isotopic resolution of the isotopic variants is not required, then the convolution method, which uses the Fourier transformation and produces the correct probabilities[31][31][31][31] [31], is an elegant and efficient alternative to calculate the aggregated isotopic distribution.

Acknowledgements

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. We especially thank Dr. Alan Rockwood for his contribution to this manuscript.

References

[1] Naturally occurring isotope abundances: Commission on atomic weights and isotopic abundances report for the international union of pure and applied chemistry in isotopic compositions of the elements 1989. *Pure and Applied Chemistry*, 70:217, 1989.

[2] Personal communication with dr. alan rockwood. 2011.

[3] J.H. Beynon. *Mass spectrometry and its applications to organic chemistry*. New York: Elsevier, 1960.

[4] K. Biemann. *Mass spectrometry, organic chemical applications*. McGraw-Hill, New York, 1962.

[5] 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.

[6] H. Budzikiewicz and R.D. Grigsby. Mass spectrometry and isotopes: a century of research and discussion. *Mass Spectrometry Reviews*, 25:146– 157, 2006.

[7] A. Carrick and Glocklin. F. Mass and abundance data for polyisotopic elements. *Journal of the Chemical Society A*, pages 40–42, 1967.

[8] B. P. Datta. Polynomial method of molecular isotopic abundance calculations: A computational note. *Rapid Communications in Mass Spectrometry*, 11:1767–1774, 1997.

[9] Fernandez de Cossio J. Computation of the isotopic distribution in two dimensions. *Analytical Chemistry*, 82(15):6726–6729, 2010.

[10] Fernandez de Cossio J. Efficient packing fourier-transform approach for ultrahigh resolution isotopic distribution calculations. *Analytical Chemistry*, 82(5):1759–1765, 2010.

[11] Fernandez de Cossio J., Gonzalez L.J., Satomi Y., L. Betancourt, Y. Ramos, V. Huerta, A. Amaro, V. Besada, G. Padron, Minamino N., and T. Takao. Isotopica: a tool for the calculation and viewing of complex isotopic envelopes. *Nucleic Acids Research*, 32:W674–W678, 2004.

[12] Farquhar G.D., Ehleringer J.R., and Hubick K.T. Carbon isotope discrimination and photosynthesis. *Review of Plant Physiology and Plant Molecular Biology*, 40:503–537, 1989.

[13] C. Genty. Application of a statistical method to isotopic analysis. *Analytical Chemistry*, 45(3):505–511, 1973.

[14] D. B. Hibbert. A prolog program for the calculation of isotope distributions in mass-spectrometry. *Chem. Intelligent Lab. Syst*, 6:203–212, 1989.

[15] D.M. Horn, R.A. Zubarev, and F.W. McLafferty. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *(J Am Soc Mass Spectrom*, 11:320–332, 2000.

[16] C. S. Hsu. Diophantine approach to isotopic abundance calculations. *Analytical Chemistry*, 56:1356–1361, 1984.

[17] E.; Loliger J. Hugentob. General approach to calculating isotope abundance ratios in mass spectroscopy. *Journal of Chemical Education*, 49:610–612, 1972.

[18] H. Kubinyi. Calculation of isotope distributions in mass spectrometry. a trivial solution for a non-trivial problem. *Analytica Chimica Acta*, 247:107–119, 1991.

[19] 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 Spectrome*, 19:1867–1874, 2008.

[20] J. L. Margrave and R. B. Polansky. Relative abundance calculations for isotopic molecular species. *Journal of chemical education*, 39:335–337, 1962.

[21] S. McIlwain, D. Page, E. Huttlin, and M. Sussman. Using dynamic programming to create isotopic distribution maps from mass spectra. *Bioinformatics*, 23:i328–i336, 2008.

[22] J. Meija. Isotope pattern geometry challenge. *Analytical and Bioanalytical Chemistry*, 380:3–4, 2004.

[23] J. Meija. Solution to isotope pattern geometry challenge. *Analytical and Bioanalytical Chemistry*, 381:13, 2005.

[24] J. Meija. Understanding isotopic distributions in mass spectrometry. *Journal of Chemical Education*, 83(12):1761, 2006.

[25] J. Meija and JA. Caruso. Deconvolution of isobaric interferences in mass spectra. *Journal of the American Society for Mass Spectrometry*, 15(5):654–658, 2004.

[26] P. O'Connor, D. Little, and F. McLafferty. Isotopic assignment in large-molecule mass spectra by fragmentation of a selected isotopic peak. *Analytical Chemistry*, 68:542–545, 1996.

[27] C. E. Olsen. A pascal program for micro-computers for calculations of compositions and isotope clusters from accurate mass measurements. *International Journal of Mass Spectrometry Ion Processes*, 47:337–340, 1983.

[28] 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.

[29] J.D. Pulfer and Derrick P.J. Simulation of isotopic peak patterns for high-mass oligomers and polynuclidic transition-metal salts. *Australian Journal of Chemistry*, 44(6):799–807, 1991.

[30] R. J. Robinson, C. G. Warner, and R. S. Gohlke. Calculation of relative abundance of isotope clusters in mass spectrometry. *Journal of Chemical Education*, 47:467–468, 1970.

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

[32] 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.

[33] A.L. Rockwood, M.M. Kushnir, and G.J. Nelson. Dissociation of individual isotopic peaks: predicting isotopic distributions of product ions in ms^n. *Journal of the American Society for Mass Spectrometry*, 14:311–32, 2003.

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

[35] A.L. Rockwood, S.L. Van Orden, and R.D. Smith. Rapid calculation of isotope distributions. *Analytical Chemistry*, 67:2699–2704, 1995.

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

[37] 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.

[38] 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.

[39] L.T. Sein. Using punnett squares to facilitate students' understanding of isotopic distributions in mass spectrometry. *Journal of Chemical Education*, 83:228–232, 2006.

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

[41] 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.

[42] S.D. Shi, C.L. Hendrickson, and A.G. Marshall. Counting individual sulfur atoms in a protein by ultrahigh-resolution fourier transform ion cyclotron resonance mass spectrometry: experimental resolution of isotopic fine structure in proteins. *Proceeding of the Natational Academy of Science*, 95:11532–11537, 1998.

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

[44] J.J. Thomson. Bakerian lecture: Rays of positive electricity. *Proceedings of the Royal Society A*, 89:11–20, 1913.

[45] E. Thurman and I. Ferrer. The isotopic mass defect: a tool for limiting molecular formulas by accurate mass. *Analytical and Bioanalytical Chemistry*, 397(7):2807–2816, 2010.

[46] 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.

[47] 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.

[48] R. C. Werlen. Effect of resolution on the shape of mass spectra of proteins: Some theoretical considerations. *Rapid Communication in Mass Spectrometry*, 8:976–980, 1994.

[49] H. Yamahato and J. A. McCloskey. *Analytical Chemistry*, 49:281, 1977.

[50] 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.

[51] 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.