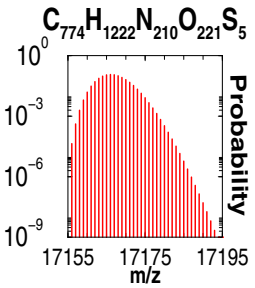
J. Am. Soc. Mass Spectrom. (2014) 25:57–70
DOI: 10.1007/s13361-013-0733-7

RESEARCH ARTICLE

Molecular Isotopic Distribution Analysis (MIDAs) with Adjustable Mass Accuracy

Gelio Alves, Aleksey Y. Ogurtsov, Yi-Kuo Yu

National Center for Biotechnology Information, National Library of Medicine, NIH, Bethesda, MD 20894, USA



Abstract. In this paper, we present Molecular Isotopic Distribution Analysis (MIDAs), a new software tool designed to compute molecular isotopic distributions with adjustable accuracies. MIDAs offers two algorithms, one polynomial-based and one Fourier-transform-based, both of which compute molecular isotopic distributions accurately and efficiently. The polynomial-based algorithm contains few novel aspects, whereas the Fourier-transform-based algorithm consists mainly of improvements to other existing Fourier-transform-based algorithms. We have benchmarked the performance of the two algorithms implemented in MIDAs with that of eight software packages (BRAIN, Emass, Mercury, Mercury5, NeutronCluster, Qmass, JFC, IC) using a consensus set of benchmark molecules. Under the proposed evaluation criteria, MIDAs's algorithms, JFC, and Emass compute with comparable accuracy the coarse-grained (low-resolution) isotopic distributions and are more accurate than the other software packages. For fine-grained isotopic distributions, we compared IC, MIDAs's polynomial algorithm, and MIDAs's Fourier transform algorithm. Among the three, IC and MIDAs's polynomial algorithm compute isotopic distributions that better resemble their corresponding exact fine-grained (high-resolution) isotopic distributions. MIDAs can be accessed freely through a user-friendly web-interface at http://www.ncbi.nlm.nih.gov/CBBresearch/Yu/midas/index.html

Key words: Isotopic Distribution, Accurate mass, Mass spectrometry, Proteomics

Received: 26 April 2013/Revised: 31 July 2013/Accepted: 2 August 2013/Published online: 20 November 2013

Introduction

Most biomolecules are composed of hydrogen, carbon, nitrogen, oxygen, and sulphur. It is known that the natural isotopes of these elements occur with different probabilities [1, 2], and in some experiments the relative abundances of an element's isotopes can be manipulated by using a technique known as stable isotopic labeling [3, 4]. The relative abundances of isotopes determine a molecule's isotopic distribution (ID), which can be measured experimentally using a mass spectrometer. The measured ID constrains the elemental composition when compared with the in-silico computed ID and, hence, helps in identifying the underlying molecule. The realization of this goal, however, demands accurate in-silico ID prediction [5–11].

The information content in an experimentally measured ID depends on the resolution of the mass spectrometer. An ID generated by a low resolution instrument contains less information than that by an ultra-high resolution instrument [12–15]. Based on the instrument resolution, three different types of IDs are commonly mentioned in the literature: the aggregated, the fine structure, and the hyper-fine structure IDs [16]. The aggregated ID is computed by merging isotopic variants that have the same nucleon number into one aggregated isotopic variant [17, 18] whose corresponding molecular mass (MM) and occurrence probability are computed respectively from the probability-weighted sum of masses and from the sum of the probabilities of the isotopic variants merged. The fine and hyper-fine structure IDs are computed similarly to the aggregated ID, except that one merges only isotopic variants whose molecular mass differences are within some pre-specified mass accuracy.

To make practical use of experimentally measured IDs, it is imperative to have methods that can compute in-silico IDs when given molecular formulas. Rockwood et al. [19] mentioned several criteria for a sound ID-computing method (IDCM): an IDCM must accurately compute in a very short time the masses and intensities without consuming much computational resource. We propose a few additional criteria by which to assess an IDCM's application value: to handle experimentally generated IDs from both low-resolution and high-resolution instruments, an IDCM should allow adjustable mass accuracy; given that customized isotopic labeling has become a common experimental technique for quantitative

Correspondence to: Yi-Kuo Yu; *e-mail:* yyu@ncbi.nlm.nih.gov

analyses, an IDCM should be able to handle customized (or user-specified) isotopic abundances (or occurrence probabilities) of all chemical elements considered; finally, an IDCM should be able to compute IDs for a wide mass range and be user-friendly. Although there are several available methods [16] that can compute an aggregated ID [17, 18, 20–23], fine structure ID [19], and hyper-fine structure ID [23–26], there are not many methods that can satisfy all the requirements mentioned above.

In this manuscript, we present MIDAs, a software tool satisfying all the requirements above. MIDAs provides users with two accurate and efficient algorithms to compute IDs: the first algorithms belongs to the class of polynomial methods [27, 28], whereas the other algorithm belongs to the class of Fourier transform methods [29, 30]. The latter consists mainly of changes made to the existing Fourier transform method [19], and the changes made are shown to improve significantly the accuracy of the computed ID. Both algorithms can compute low and high resolution IDs, referred to as the coarse-grained isotopic distribution (CGID) and the fine-grained isotopic distribution (FGID), respectively, for the remainder of this manuscript. Also both algorithms implemented in MIDAs are capable of computing CGID and FGID with adjustable mass accuracy.

To evaluate the performance of MIDAs, we have benchmarked it against eight methods: four of these methods—Mercury [19], NeutronCluster (NC) [17], Emass [21], and BRAIN [18, 31]—are the four best performing methods taking from a recent publication by Claesen et al. [18]; four other methods included are Mercury5 (a new version of Mercury2) [32], Qmass [20], Isotope Calculator (IC) [33], and a Fourier-transform-based method recently published [34], which we refer to as JFC. JFC is an improved version of Isotopica [35], which incorporates BRAIN's generating function. The program of JFC was downloaded from http://bioinformatica.cigb.edu.cu/isotopica/centermass.html. The BRAIN code was downloaded from http://www.bioconductor.org/packages/release/bioc/html/BRAIN.html. The program IC was downloaded from http://agarlabs.com/. The rest of the programs were provided by the code authors, whom we acknowledge in the Acknowledgment section.

The performance evaluation was conducted using 25 molecules. Ten of these molecules are benchmark proteins previously used to evaluate the accuracy of computed CGIDs [17, 18]. Another 10 are hydrocarbon molecules whose CGIDs and FGIDs can be exactly computed, making them ideal for evaluating the accuracy of computed IDs. The remaining five molecules, made of a combination of sulfur, mercury, carbon and hydrogen, are used together with some of the other 20 molecules to evaluate the computational time of MIDAs's algorithms. Results from our investigation show that MIDAs [both the polynomial-based algorithm (MIDAs^a) and the Fourier-transform-based algorithm (MIDAs^b)], Emass, and JFC compute CGIDs with equivalent accuracy and are more accurate than the other methods evaluated. When computing the FGIDs, IC and MIDAs^a yield FGIDs that are closest to the exact FGIDs. The results also show that MIDAs^a and MIDAs^b satisfy all aforementioned requirements to be considered a valuable tool, providing the community with two new options for computing accurate IDs.

Methods

In the subsections below we explain in detail the two algorithms implemented in MIDAs. The first subsection explains MIDAs^a, a polynomial-based algorithm. The second subsection describes MIDAs^b, a fast Fourier transform (FFT) based algorithm. Both algorithms can be used to compute CGIDs and FGIDs.

MIDAs Polynomial Multiplication Algorithm (MIDAs^a)

It is well known that the ID of a molecule can be obtained by expanding the corresponding product of polynomials: each expanded term corresponds to an isotopic composition of the molecule's elements. For example, the ID of a molecule having molecular formula (MF) $x_N\, y_M$ is given by expanding

$$\left[p(x_1)I^{m(x_1)} + \cdots + p(x_p)I^{m(x_p)}\right]^N \left[p(y_1)I^{m(y_1)} + \cdots + p\left(y_q\right)I^{m(y_q)}\right]^M, \quad (1)$$

where I is an indicator variable, x_i and y_i are the isotopes of elements x and y, respectively, $p(x_i)$ and $p(y_i)$ are normalized probabilities of occurrence, and $m(x_i)$ and $m(y_i)$ are the exact atomic masses.

There are several polynomial-based methods designed to compute an ID from the MF. Methods such as the stepwise procedure and its improvement [36, 37], symbolic expansion [4], and multinomial expansion [28, 38] have been proposed to compute the expansion of the above polynomial. Although these methods have been shown to perform well for small molecules, they fail to handle large molecules, yielding inaccurate IDs, requiring a significant amount of computer memory, and taking a considerable amount of computational time [16].

Here we present MIDAs^a, a polynomial-based algorithm that is simple and easy to understand. Our algorithm computes the molecule's CGID by directly performing polynomial multiplication. To simplify the explanation, define the polynomial between the brackets in Equation (1) containing the probabilities and atomic masses of an element's isotopes as the element fundamental polynomial (EFP). Let us represent the EFP of element x by $\mathbf{P}_x$, and also define the following recursion operation that multiplies together polynomials $\mathbf{Q}_x$ and $\mathbf{P}_x$ and assigns the resulting polynomial back to $\mathbf{Q}_x$ as $\mathbf{Q}_x \leftarrow (\mathbf{Q}_x \times \mathbf{P}_x)$.

Substituting these definitions in Equation (1) with $\mathbf{Q}_x$ initialized to one gives

$$\begin{aligned}[\mathbf{P}_x]^N[\mathbf{P}_y]^M &= \left\{[\mathbf{P}_x]^{\lfloor\frac{N}{10}\rfloor}\right\}^{10}\times[\mathbf{P}_x]^{N-10\lfloor\frac{N}{10}\rfloor}\left\{[\mathbf{P}_y]^{\lfloor\frac{M}{10}\rfloor}\right\}^{10}\times[\mathbf{P}_y]^{M-10\lfloor\frac{M}{10}\rfloor}\\ &= \{[(\mathbf{Q}_x\times\mathbf{P}_x)\times\cdots\times\mathbf{P}_x]\}^{10}[\mathbf{P}_x]^{N-10\lfloor\frac{N}{10}\rfloor}\\ &\times\left\{\left[(\mathbf{Q}_y\times\mathbf{P}_y)\times\cdots\times\mathbf{P}_y\right]\right\}^{10}[\mathbf{P}_y]^{M-10\lfloor\frac{M}{10}\rfloor},\end{aligned} \tag{2}$$

where $\lfloor z\rfloor$ represents the integer part of z for any positive number z. Using the recursion operation mentioned earlier, all the x-element related polynomials finally merge into $\mathbf{Q}_x$ and all the y-element related polynomials finally merge into $\mathbf{Q}_y$ as shown in algorithms 1 and 2. By first computing $[\mathbf{P}_x]^{\lfloor\frac{N_X}{10}\rfloor}$ in Equation (2), one considerably reduces the computational time needed to obtain the polynomial expansion of an EFP. The logic in computing $\left([\mathbf{P}_x]^{\lfloor\frac{N}{10}\rfloor}\right)^{10}\times[\mathbf{P}_x]^{N-10\lfloor\frac{N}{10}\rfloor}$ (or $\left([\mathbf{P}_y]^{\lfloor\frac{M}{10}\rfloor}\right)^{10}\times[\mathbf{P}_y]^{M-10\lfloor\frac{M}{10}\rfloor}$) and not $[\mathbf{P}_x]^N$ (or $[\mathbf{P}_y]^M$) is that the former requires a smaller number of arithmetic operations. This is due to two heuristic procedures of MIDAS^a, prune and merge, which reduce the number of retained terms in the expanded polynomial $[\mathbf{P}_x]^{\lfloor\frac{N_X}{10}\rfloor}$. These heuristics are similar to the hyperatom concept [37] and the superatom concept [10], except that the number of atoms $\lfloor N/10\rfloor$ in a superstructure is not fixed in our case. The choice of using 10 in Equation (2) was somewhat arbitrary but seemed to generate an accurate ID for each molecule used in our investigation in a short amount of computational time. Evidently, one may use a number other than 10. Choosing a smaller number, however, means that we need a larger memory to hold $\mathbf{Q}$. Choosing a larger number, on the other hand, results in a longer computation time. We find using the number 10 seems to provide a good balance between the two.

The first heuristic employed by the MIDAS^a algorithm prunes terms from the polynomial $\mathbf{Q}$ that have probability smaller than a pre-set probability value (η). The second heuristic procedure merges polynomial terms from $\mathbf{Q}$ that are within some user specified mass accuracy (ϵ) of each other into a new polynomial term. The new polynomial term is assigned a new mass ($\overline{m}$) that is equal to the probability-weighted sum of the merged terms

$$\overline{m} = \frac{\sum_i p_i m_i}{\sum_i p_i} \tag{3}$$

where m_i and p_i stand for the mass and probability of the merged terms, respectively. This new term associated with $\overline{m}$ is then assigned a probability equal to the sum of the probabilities of the merged terms. The pseudo-code for computing a CGID is given by algorithm 1, which is used by MIDAS^a.

```
initialization;
η = 10^−150;
Q = 1 ;
for i ← 1 to (number of unique elements in MF) do
    Q_i = 1 ;
    for j ← 1 to (number of times element i appears in MF) do
        Q_i ← P_i × Q_i;
        if (number of term in Q_i) ≥ 1000 then
            Prune(Q_i, η) ;
            Sort(Q_i);
            Merge(Q_i, ϵ);
        end
    end
    Q ← Q × Q_i;
    Prune(Q, η) ;
    Sort(Q);
    Merge(Q, ϵ);
end
```

Algorithm 1. Computes Coarse-Grained Isotopic Distribution

To compute the FGID for an MF, for every element x, MIDAS^a first computes the expected number of occurrences $\mu[x_i]$ for each isotope x_i of x. MIDAS^a then computes $\sigma^2[x_i]$, the variance of the number of occurrences. As an example, for the molecular formula $x_N y_M$, the expectation and variance in the number of atoms for a given isotope of element x is given by

$$\mu[x_i] = Np(x_i) \tag{4}$$

and

$$\sigma^2[x_i] = Np(x_i)(1-p(x_i)). \tag{5}$$

Using the computed expectation and variance values, we denote the range $[\mathcal{B}(x_i),\mathcal{U}(x_i)]$ as allowable for $\mathcal{N}(x_i)$, the number of atoms of isotope x_i. The upper bound $\mathcal{U}(x_i)$ and the lower bound $\mathcal{B}(x_i)$ are given by

$$\mathcal{U}(x_i) = \mu[x_i] + 10\sqrt{(1+\sigma^2[x_i])} \tag{6}$$

$$\mathcal{B}(x_i) = \begin{cases}\mu[x_i]-10\sqrt{(1+\sigma^2[x_i])}, & \mu[x_i]-10\sqrt{(1+\sigma^2[x_i])}>0\\ 0, & \mu[x_i]-10\sqrt{(1+\sigma^2[x_i])}\le 0\end{cases}$$

For isotope x_i, we choose $10\sqrt{(1+\sigma^2[x_i])}$ to be the span of sum as this quantity is guaranteed to be simultaneously larger then 10σ $[x_i]$ and 10 daltons. For each element x, the $\mathcal{U}(x_i)$s and $\mathcal{B}(x_i)$s are used to construct a polynomial, $\tilde{\mathbf{P}}_x$, by means of the multinomial expansion formula

$$\begin{aligned}\tilde{\mathbf{P}}_x = \underbrace{\sum_{k_1=\mathcal{B}(x_1)}^{\mathcal{U}(x_1)}\sum_{k_2=\mathcal{B}(x_2)}^{\mathcal{U}(x_2)}\sum_{k_3=\mathcal{B}(x_3)}^{\mathcal{U}(x_3)}\cdots\sum_{k_p=\mathcal{B}(x_p)}^{\mathcal{U}(x_p)}}_{k_1+k_2+k_3+\cdots+k_p=N}\frac{N!}{k_1!k_2!k_3!\cdots k_p!}\left[p(x_1)I^{m(x_1)}\right]^{k_1}\\ \times\left[p(x_2)I^{m(x_2)}\right]^{k_2}\times\left[p(x_3)I^{m(x_3)}\right]^{k_3}\times\cdots\times\left[p(x_p)I^{m(x_p)}\right]^{k_p}.\end{aligned} \tag{7}$$

By summing only the contributions bounded by $\mathcal{B}$ and $\mathcal{U}$, we direct the calculations to the relevant part of the ID. It has counter-part in FT based method, namely the heterodyning of in [24].

For numerical accuracy and efficiency we employ the following simple identity

$$\frac{N!}{k_1!\cdots k_p!}p(x_1)^{k_1}\cdots p(x_p)^{k_p} = \exp\big(\ln(N!)-\ln(k_1!)-\cdots-\ln(k_p!) + k_1\log(p(x_1)) + \cdots + k_p\log(p(x_p))\big), \quad (8)$$

and $\ln(n!) = \sum_{k=1}^{n}\ln k$. This representation reduces computational time of Equation (7) since by tabulation one only enumerates all the logarithmic terms in Equation (8) once. Once all $\widetilde{\mathbf{P}}_x$s have been computed, they are used together with a user specified ϵ to compute a FGID using algorithm 2.

```
initialization;
η = 10^−16;
Q = 1 ;
for i ← 1 to (number of unique elements in MF) do
    Q ← P̃_i × Q;
    Prune(Q, η) ;
    Sort(Q);
    Merge(Q, ϵ);
end
```

Algorithm 2. Computes Fine-Grained Isotopic Distribution 2

MIDAs Fast Fourier Transform Algorithm (MIDAs^b)

The MIDAs^b algorithm is similar to an early FFT algorithm by Rockwood et al. [19], which was implemented in a computer program called Mercury. These two algorithms differ, however, in a few aspects. First, using the exact isotopic masses in discrete FFT (DFFT) [39, 40], Mercury produces IDs with leakages (assigning nonzero probabilities to masses where exactly zero probability is expected) and employs an apodization function to minimize leakage [41]. On the other hand, by assigning each isotope mass to a point on a fixed grid, MIDAs^b avoids the leakage problem. Using discrete masses to avoid leakage is not new: Rockwood and Van Orden [32] have written a computer program, whose latest version is called Mercury5, to compute IDs based on the nucleon numbers (or roughly using one dalton mass grid). The improvement we made was to allow the users to specify the mass accuracy other than 1 Da. Second, Mercury uses a fixed number of sample points with the DFFT, whereas in MIDAs^b the number of sample points used depends on the mass accuracy, which is a parameter adjustable by the user.

Every FFT based method relies on the convolution theorem, which states that a convolution can be performed as multiplication in the Fourier domain.

As we shall discuss in the Appendix, there are two key conditions in order for the convolution theorem to be used in the discrete case while computing IDs. The first one is that the masses of each isotope must lie on grid points. Using a mass that is not on the grid causes the "leakage" phenomenon [41]. If the masses considered all reside on grid points, the leakage problem no longer exists. The second important condition is that the mass domain must be large enough so that the "folded-back" phenomenon (which is also known as "aliasing", "fold over", or "wrap around" in the signal processing community) near the tail of the distribution is negligible (see Appendix).

Prior to delving into detail constructs of MIDAs^b, let us first describe how one may compute the theoretical molecular mass variance σ^2_{MM}. Using our example molecule $x_N\, y_M$, one note that the molecular mass variance of this molecule can be rigorously written as $\sigma^2_{MM} = N\sigma^2[m_x] + M\sigma^2[m_y]$, where $\sigma^2[m_x]$ is the molecular mass variance associated with element x. Explicitly, one may calculate $\sigma^2[m_x]$ as follows

$$\sigma^2[m_x] = \left[\sum_i p(x_i)m^2(x_i)\right] - \left[\sum_i p(x_i)m(x_i)\right]^2,$$

where the index i runs over all isotopes of element x and $p(x_i)$ again represents the occurrence probability of isotope x_i.

A key constraint of DFFT based ID method is that the total number of sample points, denoted by S, must be an integral power of two [42]. For a given molecule and specified mass accuracy ϵ, the total number of sample points S used in MIDAs^b's DFFT is given by

$$S = 2^{\alpha}, \text{where } \alpha = \left\lceil \ln\left(\frac{15\sqrt{1+\sigma^2_{MM}}}{\epsilon}\right)/\ln 2 \right\rceil, \quad (9)$$

and $\sigma_{MM}{}^2$ is the theoretical variance in MM due to the elements' isotopes [32]. The quantity $\lceil z \rceil$ represents the smallest integer that is larger than z for any positive number z. Again the quantity $15\sqrt{1+\sigma^2_{MM}} > \max(15, 15\sigma_{MM})$ is chosen so that S covers on both ends more than 7.5 standard deviations from the mean molecular mass, which prevents *folded-back* mass regions from having significant probabilities.

In order to avoid the problem of *leakage*, instead of using exact masses of isotopes and then applying filtering windows, we pin all isotopic masses to grid points. For each isotope mass $m(x_i)$, we first find a corresponding grid index $n(x_i)$ by the following formula

$$n(x_i) = \left\lfloor \frac{m(x_i)}{\epsilon} + 0.5 \right\rfloor. \quad (10)$$

Using this discrete approach, the probability function of the mass of element x becomes

$$Prob\ (m_x = n\epsilon) = \sum_i p(x_i)\delta_{n,n(x_i)}$$

where $n(x_i)\epsilon$ is the approximate expression for the exact mass $m(x_i)$, and the Kronecker delta function takes value one if its two indices coincide and zero otherwise.

Consider the mass distribution of our example molecule $x_N y_M$. By the convolution theorem, the Fourier transform of the mass distribution, denoted by $\Psi(v)$, can be written as

$$\Psi(v) = \left[p(x_1)e^{2\pi i n(x_1)v/S} \cdots + p(x_p)e^{2\pi i n(x_p)v/S}\right]^N \times \left[p(y_1)e^{2\pi i n(y_1)v/S} + \cdots + p\left(y_q\right)e^{2\pi i n(y_q)v/S}\right]^M,$$

where v takes S discrete values: 0,1,…, $S-1$. The sample function $\Psi(v)$ is heterodyned to have zero average mass by multiplying it by $e^{-2\pi i n_o v/S}$, where n_o is equal to $\overline{n}$ (the molecule's probability-weighted average grid index computed using $n(x_i)$ and $p(x_i)$) rounded to the nearest integer.

Once the function $\Psi(v)$ has been calculated, three other operations are performed in order to generate the final FGID and CGID. The first operation performed is the inverse discrete fast Fourier transform (IDFFT), which transforms the sample function $\Psi(v)$ to $\Phi(n)$ on the mass grid. Second, we apply a denoising procedure to remove small amplitudes due to rounding errors that occur during IDFFT. The rounding errors are expected to create small positive and negative amplitudes of equal amounts in the mass domain. MIDAs^b thus removes all amplitudes whose absolute magnitude are smaller than that of the most negative amplitude. As a matter of fact, to be more conservative, MIDAs^b uses an amplitude cutoff value that is twice the absolute value of the most negative amplitude. This means that only terms having amplitude greater than the cutoff value are reported in a computed FGID and CGID, with the amplitude values renormalized to sum to one. Figure 1 shows an example of the overlap between the positive amplitude histogram and the negative amplitude histogram. Right below the cutoff absolute amplitude, we see that the two histograms resemble each other, reflecting the fact that rounding errors have equal probability to be positive and negative. Following Rockwood and Van Orden [32], in the third step, MIDAs^b applies a linear transformation to rescale the masses associated with the IDs to ensure a good agreement between the theoretically calculated and the numerically computed mean molecular mass as well as standard deviation of the molecular mass. The procedure described above is summarized in the pseudo code give by algorithm 3.

initialization;
$\Delta\nu = \frac{1}{S}$;
$\mathbf{D}[2(S)]$;
for $k \leftarrow 1$ **to** *(S/2)* **do**
  $j = (S/2) + k$;
  $\nu_k = (k-1)\Delta\nu$;
  $\nu_j = (j - N - 1)\Delta\nu$;
  $\mathbf{D}[2k-1] = \mathbf{Re}[e^{-2\pi i n_o \nu_k}\Psi(\nu_k)]$;
  $\mathbf{D}[2k] = \mathbf{Im}[e^{-2\pi i n_o \nu_k}\Psi(\nu_k)]$;
  $\mathbf{D}[2j-1] = \mathbf{Re}[e^{-2\pi i n_o \nu_j}\Psi(\nu_j)]$;
  $\mathbf{D}[2j] = \mathbf{Im}[e^{-2\pi i n_o \nu_j}\Psi(\nu_j)]$;
end
IDFFT(D) ;
Remove_Rounding_Errors(**D**) ;
Mass_Scale_Transformation(**D**) ;

Algorithm 3. Computes Fine-Grained and Coarse-Grained Isotopic Distribution

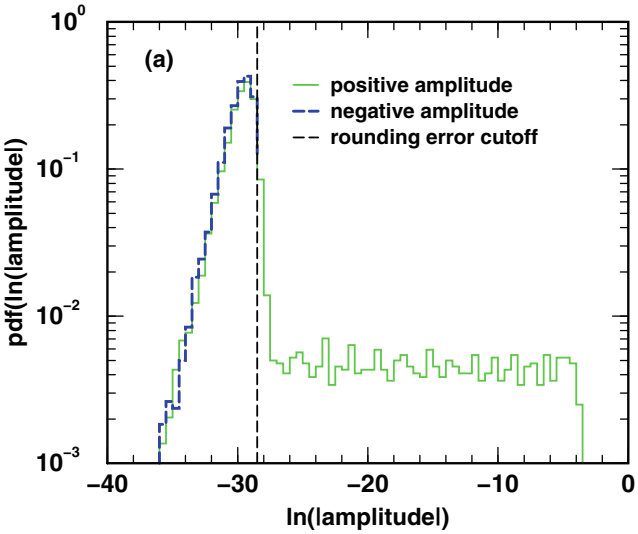


Figure 1. Example of rounding errors. The curves plotted above are the histogram for the logarithm of absolute value of the positive (green solid line histogram) and negative (blue long-dashed line histogram) amplitudes obtained after applying the discrete Fourier transform to compute an isotopic distribution for molecule $C_{2023}H_{3208}N_{524}O_{619}S_{20}$ using a mass accuracy of 0.01 Da (≈0.22 ppm). Absent the leakage, the negative amplitude can only come from rounding errors, among which equal amounts of small positive amplitudes and negative amplitudes are expected. The above histograms overlap for terms that have magnitude in amplitude less than 4.2e–10, displayed above by a dash black line, and at this point is the rounding error cutoff value used by MIDAs^b

Results and Discussion

All methods used in our investigation were evaluated using their default parameter settings, except for a few parameter changes made to ensure that the atomic masses and abundances of elements' isotopes were the same for all methods (see Table 1). To conduct the evaluation, we used the 22 biomolecules and three inorganic compounds, all of which are listed in Table 2. Ten of these biomolecules, (1)–(10), are benchmark proteins previously used to evaluate the computed CGID [17, 18]; 10 biomolecules, (11)–(20), are hydrocarbon molecules whose FGID can be exactly computed and were employed to evaluate the FGID; the remaining five molecules, (21)–(25), made of a combination of sulfur, mercury, carbon, and hydrogen, are used together with the some of the other 20 biomolecules to evaluate the computational time of MIDAs's algorithms.

Overview of Methods Benchmarked

MIDAs's performance was evaluated against eight published methods: Mercury [19], Mercury5 [32], JFC [34], Isotope Calculator (IC) [33], Qmass [20], BRAIN [18, 31, 43], NeutronCluster (NC) [17], and Emass [21]. The first three published methods are Fourier-transform-based, IC utilizes a

Table 1. Atomic Masses and Abundances used for Benchmark Test in this Paper

Isotope	Atomic mass Da	Abundance (%)
Atomic masses and naturally occurring isotopic abundances [1]		
^{12}C	12.0000000000	98.9300
^{13}C	13.0033548378	1.0700
^{1}H	1.0078250321	99.9885
^{2}H	2.0141017780	0.0115
^{14}N	14.0030740052	99.6320
^{15}N	15.0001088984	0.3680
$^{16}0$	15.9949146	99.7570
$^{17}0$	16.9991312	0.0380
$^{18}0$	17.9991603	0.2050
^{32}S	31.97207070	94.9300
^{33}S	32.97145843	0.7600
^{34}S	33.96786665	4.2900
^{36}S	35.96708062	0.0200
^{196}Hg	195.965833	0.0015
^{198}Hg	197.966769	0.0997
^{199}Hg	198.968279	0.1687
^{200}Hg	199.968326	0.2310
^{201}Hg	200.970302	0.1318
^{202}Hg	201.970643	0.2986
^{204}Hg	203.973493	0.0687
Atomic Masses and Enriched Carbon's Isotopic Abundances		
^{12}C	12.0000000000	1.0000
^{13}C	13.0033548378	99.0000

divide-and-recursively-combine algorithm, Qmass has its core based on FFT, BRAIN and NeutronCluster are polynomial-based, whereas Emass is based on a direct convolution approach related to the stepwise procedure and its improvement [36, 37]. BRAIN, Qmass, NC, Emass, JFC, and Mercury5 all use nucleon numbers to classify molecule's isotopic variants, while all but the last assign to a given nucleon number the average isotopic mass of all variants of that nucleon number.

IC is suitable for computing FGIDs, not CGIDs. Qmass, BRAIN, NeutronCluster, and Emass are suitable for computing CGIDs, not FGIDs. The remaining three Fourier-transform-based methods are also suitable for computing CGIDs, although Mercury is the only one that has FGID computing capacity. To benchmark the FGIDs computed by MIDAs against those of Mercury, however, would require post-processing of Mercury data files such as removing noise from leakage and rounding errors, as well as compiling output from different specified molecular masses. All of these steps may be done differently and make the benchmark test less meaningful. For these reasons, we only evaluated MIDAs's FGIDs against that of IC, not that of Mercury.

Benchmarking of Computed CGIDs

Following previous publications [18, 19, 24], the accuracy of a method is gauged by how accurately it yields ID mean, ID standard deviation, lightest and heaviest molecular masses, while computing a CGID. In our evaluation, the lightest mass and heaviest molecular mass are defined as a molecule's molecular mass computed using the masses of the lightest and heaviest isotopes, respectively.

Lightest masses comparisons for biomolecules, numbered (1)–(10) in Table 2, with elements having their naturally occurring isotopic abundances taken from Table 1, are displayed in Table 3. Unexpectedly, the lightest masses for the first six

Table 2. Molecules for which the Isotopic Distribution was Computed by Various Methods

No. [1]	Molecular formula	Lightest Mass (Da)[2]	Average Mass (Da)
(1)	$C_{50}H_{71}N_{13}O_{12}$	1045.5345145467	1046.1811074558
(2)	$C_{254}H_{377}N_{65}O_{75}S_{6}$	5729.6008666397	5733.5107592120
(3)	$C_{520}H_{817}N_{139}O_{147}S_{8}$	11616.8493497485	11624.4487510271
(4)	$C_{744}H_{1224}N_{210}O_{222}S_{5}$	16812.9547750824	16823.3213522608
(5)	$C_{2023}H_{3208}N_{524}O_{619}S_{20}$	45387.0070331016	45415.6793695079
(6)	$C_{2934}H_{4615}N_{781}O_{897}S_{39}$	66389.8624747027	66432.4555603617
(7)	$C_{5047}H_{8014}N_{1338}O_{1495}S_{48}$	112823.8795468070	112895.1259319964
(8)	$C_{8574}H_{13378}N_{2092}O_{2392}S_{77}$	186386.7992654122	186506.0525933526
(9)	$C_{17600}H_{2674}N_{4752}O_{5486}S_{197}$	398470.3669960258	398722.9724824960
(10)	$C_{23832}H_{37816}N_{6528}O_{7031}S_{170}$	533403.4750914392	533735.2146493989
(11)	$C_{5}H_{5}$	65.0391251605	65.0933832534
(12)	$C_{10}H_{10}$	130.0782503209	130.1867665069
(13)	$C_{50}H_{50}$	650.3912516049	650.9338325345
(14)	$C_{100}H_{100}$	1300.7825032099	1301.8676650690
(15)	$C_{1000}H_{1000}$	13007.8250320999	13018.6766506902
(16)	$C_{10000}H_{10000}$	130078.2503209999	130186.7665069023
(17)	$C_{20000}H_{20000}$	260156.5006419999	260373.5330138047
(18)	$C_{30000}H_{30000}$	390234.7509629999	390560.2995207072
(19)	$C_{40000}H_{40000}$	520313.0012839999	520747.0660276095
(20)	$C_{50000}H_{50000}$	650391.2516049999	650933.8325345119
(21)	S_{20000}	639441.4139999999	641321.6938997399
(22)	Hg_{5000}	159860.3534999999	160330.4234749349
(23)	$Hg_{1000}S_{1000}$	227937.9037000000	232665.2510595869
(24)	$S_{1000}C_{1000}H_{1000}$	44979.8957320999	45084.7613456772
(25)	$Hg_{1000}C_{1000}H_{1000}$	208973.6580321000	213617.8430152902

[1]Reference number associated with a molecular formula (biomolecule or inorganic compound).
[2]The unified atomic mass unit dalton (Da).

Table 3. Coarse - Grained Isotopic Distribution Results using Naturally Occurring Isotopes

Difference in lightest mass

No.[1]	MIDAs[a]	MIDAs[b]	BRAIN	Emass	Mercury	Mercury5	NC	Qmass	JFC
(1)	0	-3.4e - 05	-2.6e - 10	2.2e - 13	7.0	7.3	0	12.2	0
(2)	0	-1.7e - 03	-1.3e - 09	0	12.0	12.1	0	15.9	0
(3)	0	-2.6e - 03	-2.8e - 09	0	8.0	8.4	0	18.2	-1.0e - 10
(4)	0	-2.1e - 03	-4.2e - 09	0	22.0	21.6	-360	39.1	8.0e - 10
(5)	7.2e - 12	-7.4e - 03	1.0e - 08	1.4e - 11	2.9	3.3	0	0.045	-1.6e - 01
(6)	-1.4e - 11	-5.0	-1.6e - 08	0	22.0	21.3	0	65.2	-4.6
(7)	1.5e - 11	-19.1	-2.7e - 08	-8.0	-8.0	-7.2	0	-69.7	-18.1
(8)	0	-49.1	-4.1e - 08	-31.1	-55.2	-55.3	0	-90.7	-48.8
(9)	-5.8e - 11	-147.4	-1.1e - 07	-114.3	-124.3	-124.7	0	-188.3	-118.9
(10)	0	-210.6	-1.2e - 07	-172.5	-203.7	-203.7	0	-355.6	-146.4

Difference in heaviest isotopic mass

No.	MIDAs[a]	MIDAs[b]	BRAIN	Emass	Mercury	Mercury5	NC	Qmass	JFC
(1)	7.4e + 01	1.4e + 02	1.5e + 02	1.4e + 02	1.5e + 02	1.5e + 02	1.5e + 02	1.4e + 02	1.4e + 02
(2)	7.2e + 02	8.4e + 02	8.7e + 02	8.4e + 02	8.5e + 02	8.5e + 02	8.6e + 02	8.4e + 02	8.4e + 02
(3)	1.6e + 03	1.8e + 03	1.8e + 03	1.7e + 02	1.8e + 03	1.8e + 03	1.8e + 03	1.8e + 03	1.8e + 03
(4)	2.5e + 03	2.6e + 03	2.6e + 03	2.6e + 03	2.6e + 03	2.6e + 03	2.3e + 03	2.6e + 03	2.6e + 03
(5)	6.8e + 03	7.0e + 03	7.0e + 03	7.0e + 03	7.0e + 03	7.0e + 04	7.0e + 03	6.9e + 03	7.0e + 03
(6)	9.9e + 03	1.0e + 04	1.0e + 04	1.0e + 04	1.0e + 04	1.0e + 04	1.0e + 04	1.0e + 04	1.0e + 04
(7)	1.7e + 04	1.7e + 04	1.7e + 04	1.7e + 04	1.7e + 04	1.7e + 04	1.7e + 04	1.8e + 04	1.7e + 04
(8)	2.9e + 04	2.9e + 04	2.9e + 04	2.9e + 04	2.9e + 04	2.9e + 04	2.9e + 04	2.9e + 04	2.9e + 04
(9)	6.0e + 04	6.0e + 04	6.0e + 04	6.0e + 04	6.0e + 04	6.0e + 04	6.0e + 04	6.0e + 04	6.0e + 04
(10)	8.2e + 04	8.3e + 04	8.3e + 04	8.3e + 04	8.3e + 04	8.3e + 04	8.3e + 04	8.2e + 04	8.2e + 04

[1]Reference number associated with a molecular formula (biomolecule).

molecules, reported by Mercury, Mercury5, and Qmass, are even lighter than their exact lightest masses, which should be the lightest masses possible in these six IDs. This observation was also described by [18]. For Mercury5 (and Mercury), this is caused by the rounding errors (and the leakage) when applying the DFFT. (In principle, both methods can avoid this problem by not reporting any terms in the computed ID that have masses lighter than the *exact* lightest mass.) For Qmass, this seems to arise from computing ID terms that are outside of the allowed mass range imposed by the biomolecule's MF. This is because in the Qmass output file the reported masses lighter than the *exact* lightest mass are associated with elemental compositions that differ from the biomolecule's MF used in the evaluation.

The software NC reports correct lightest masses for nine out of the 10 molecules. For biomolecule number four, NC reports a mass that is 360 Da heavier. This same result has also been observed independently by others [17, 18].

For MIDAs[a], BRAIN, and Emass, the differences between exact and computed lightest masses, for small and medium size biomolecules [numbered (1)–(6)], are smaller than 1.0e–08 Da. As for JFC and MIDAs[b], although they do not perform as well as the polynomial-based methods above, they are not inferior to other Fourier-transform-based methods such as Mercury and Mercury5. When the biomolecules become heavier [say molecules numbered (7)–(10)], the chance of experimentally observing the *exact* lightest masses rapidly decreases, and the computed difference between exact and computed lightest masses becomes less important.

The evaluation of getting the correct heaviest mass is not as important under natural conditions. This is because heavy isotopes typically carry very low natural occurrence probabilities so that it is impossible to observe the *exact* heaviest isotopic variant of the molecule. Of course, when artificial isotopic abundances are enforced, obtaining the correct heaviest masses can become important, while obtaining the correct lightest masses can become unimportant. Since the current evaluation is using the natural isotopic abundances, we do not expect any method to provide *correct* heaviest masses. Indeed, because most methods are computing terms of an ID that are concentrated around a molecule's average molecular mass, which is closer to the *exact* lightest mass under natural isotopic abundances, the mass range used for computing IDs usually will not include the heaviest masses. For biomolecules numbered (1)–(10), the differences between the exact heaviest masses and the heaviest masses computed by all methods considered are all of the same order of magnitude.

Displayed in the upper (lower) half of Table 4 are the relative differences of computed CGIDs derived molecular mass averages (standard deviations) to their theoretical values. Molecules numbered (1)–(10) in Table 2 are used with elements assuming isotopic abundances shown in Table 1. In terms of the average masses, MIDAs[a,b], JFC, and Emass have comparable errors and have slightly smaller errors than the other methods. In terms of mass standard deviations, MIDAs[a,b], JFC, and Emass have slightly smaller errors than the other methods. In principle, the accuracy of BRAIN might be improved by increasing the number of aggregated isotopic variants computed for each computed CGID. However, to accomplish this would require changing its default option. As mentioned earlier, to keep the benchmarking test simple, we only use the default option for each method considered. From Table 4, one can also infer that Qmass yields small errors for

Table 4. Coarse - Grained Isotopic Distribution Results using Naturally Occurring Isotopes

Difference in average mass

No.[1]	MIDAs[a]	MIDAs[b]	BRAIN	Emass	Mercury	Mercury5	NC	Qmass	JFC
(1)	2.3e - 13	6.8e - 13	4.4e - 03	-4.5e - 13	9.1e - 05	-2.9e - 05	6.51e - 05	-4.6e - 13	2.3e - 12
(2)	1.8e - 12	3.5e - 11	3.2e - 01	-1.8e - 12	8.1e - 04	7.6e - 05	3.7e - 03	-7.3e - 12	3.6e - 12
(3)	-5.4e - 12	-5.4e - 12	8.2e - 02	-3.6e - 12	2.2e - 04	-5.1e - 05	5.9e - 03	5.4e - 12	0
(4)	-7.3e - 12	2.0e - 10	4.6e - 02	0	2.9e - 03	-7.3e - 04	-360	5.8e - 11	7.3e - 12
(5)	4.3e - 11	2.6e - 10	1.4e - 04	7.3e - 12	-3.1e - 03	1.8e - 04	3.7e - 03	-7.3e - 12	-2.9e - 11
(6)	0	1.3e - 10	1.7e - 06	-5.8e - 11	-4.1e - 03	3.1e - 03	-8.5e - 04	4.2e - 09	1.2e - 10
(7)	4.3e - 11	-2.4e - 09	-2.7e - 08	-1.4e - 11	-4.1e - 03	2.1e - 03	-3.9e - 03	-1.3e - 10	5.8e - 11
(8)	-2.9e - 11	1.6e - 09	-4.1e - 08	0	-5.1e - 03	-7.9e - 03	-1.0e - 02	-7.5e - 01	2.6e - 10
(9)	-3.5e - 10	7.2e - 09	-1.1e - 07	-3.5e - 10	1.7e - 02	7.7e - 03	-4.0e - 02	-9.7e - 03	7.6e - 10
(10)	-1.2e - 10	7.6e - 09	-1.2e - 07	-1.2e - 10	-1.1e - 01	3.3e - 02	-3.8e - 02	-4.4e+02	-4.6e - 10

Difference in standard deviation

No.[1]	MIDAs[a]	MIDAs[b]	BRAIN	Emass	Mercury	Mercury5	NC	Qmass	JFC
(1)	1.1e - 06	-4.2e - 10	1.2e - 02	1.1e - 06	1.6e - 06	-1.2e - 04	-3.6e - 04	1.1e - 06	1.1e - 06
(2)	6.5e - 06	-5.0e - 09	3.6e - 01	6.5e - 06	1.2e - 06	-1.8e - 04	-1.2e - 03	6.5e - 06	6.5e - 06
(3)	8.0e - 06	1.5e - 08	1.2e - 01	8.0e - 06	9.8e - 05	-3.3e - 05	2.2e - 03	7.9e - 06	8.0e - 06
(4)	7.2e - 06	2.5e - 08	7.3e - 02	7.2e - 06	-3.0e - 07	-4.6e - 04	-4.5e - 02	7.0e - 06	7.1e - 06
(5)	1.3e - 05	1.8e - 07	3.7e - 04	1.3e - 05	9.7e - 06	-2.7e - 04	-1.8e - 03	1.3e - 05	1.3e - 05
(6)	1.8e - 05	-1.9e - 07	2.3e - 05	1.8e - 05	-3.9e - 07	-9.0e - 04	-8.4e - 03	-2.7e - 06	1.7e - 05
(7)	2.0e - 05	-8.0e - 07	2.1e - 05	2.0e - 05	-2.7e - 07	-7.1e - 04	-7.5e - 03	2.2e - 05	2.0e - 05
(8)	2.5e - 05	2.1e - 06	2.5e - 05	2.5e - 05	4.4e - 06	-5.4e - 04	-8.7e - 03	-4.8e+00	2.6e - 05
(9)	4.2e - 05	-7.8e - 06	4.1e - 05	4.5e - 05	-5.9e - 07	-1.5e - 03	-5.2e - 03	-9.9e - 02	3.9e - 05
(10)	4.8e - 04	-1.0e - 05	5.0e - 05	5.4e - 05	-1.2e - 05	-1.3e - 03	9.6e - 03	-1.4e+02	3.8e - 05

[1]Reference number associated with a molecular formula (biomolecule).

small and medium size molecules, but the error increases as the molecular mass increases.

We have also considered the possibility of deviations from the natural frequencies of occurrence of an element's isotopes. Such customized modifications can be accomplished experimentally by a technique known as isotopic labeling [3], which is frequently employed in quantitative proteomics [44]. To mimic such a situation, we have computed CGIDs for various molecules assuming different carbon isotopic abundances: 99% ^{13}C and 1% ^{12}C as listed in Table 1. We then derive

Table 5. Coarse - Grained Isotopic Distribution Evaluation using Abundances for Carbon's Isotopes of 99% ^{13}C and 1% ^{12}C

Difference in average mass

No.[1]	MIDAs[a]	MIDAs[b]	BRAIN	Emass	Mercury	Mercury5	NC	Qmass	JFC
(1)	-6.8e - 13	3.9e - 12	-1.7e + 01	2.3e - 13	5.0e - 05	-4.0e - 05	3.3e - 02	-2.3e - 13	1.6e - 12
(2)	0	4.5e - 11	NR[2]	3.6e - 12	3.3e - 04	7.3e - 05	1.7e - 01	-1.8e - 12	4.5e - 12
(3)	1.8e - 12	2.2e - 10	NR	-1.8e - 12	-6.3e - 04	-3.1e - 04	3.1e - 01	-6.4e - 11	-2.4e - 11
(4)	-7.3e - 12	4.4e - 11	NR	0	3.9e - 03	-2.5e - 04	NR	-1.1e - 11	-1.1e - 11
(5)	-2.9e - 11	-5.8e - 11	NR	7.3e - 12	-5.2e - 03	6.4e - 04	1.8e + 03	-4.1e - 07	-7.3e - 12
(6)	5.8e - 11	-1.0e - 10	NR	2.9e - 11	-5.4e - 03	2.1e - 03	2.7e + 03	-2.9e - 11	2.9e - 11
(7)	1.4e - 11	6.8e - 10	NR	-7.3e - 11	-8.7e - 04	6.6e - 04	4.8e + 03	-4.9e - 07	1.4e - 10
(8)	3.2e - 11	-3.5e - 10	NR	8.7e - 11	-8.1e - 03	6.4e - 03	8.4e + 03	-1.1e + 02	1.2e - 10
(9)	0	4.2e - 09	NR	0	-3.7e - 03	8.7e - 03	1.7e + 04	-1.5e + 02	-4.1e - 10
(10)	2.3e - 10	7.7e - 09	NR	8.15e - 10	-1.2e - 01	5.4e - 03	2.4e + 04	-5.1e + 02	-1.2e - 10

Difference in standard deviation

No.[1]	MIDAs[a]	MIDAs[b]	BRAIN	Emass	Mercury	Mercury5	NC	Qmass	JFC
(1)	7.9e - 07	-3.3e - 10	-1.4e + 01	7.9e - 07	-2.6e - 08	-1.2e - 04	-4.9e + 00	7.4e - 07	7.9e - 07
(2)	5.9e - 06	5.9e - 09	NR[2]	5.9e - 06	2.2e - 07	-1.9e - 04	-1.1e + 01	5.9e - 06	5.9e - 06
(3)	7.6e - 06	1.4e - 08	NR	7.6e - 06	8.2e - 06	-1.3e - 04	-1.6e + 01	7.6e - 06	7.6e - 06
(4)	7.1e - 06	1.4e - 08	NR	7.1e - 06	-1.7e - 07	-4.7e - 04	NR	7.1e - 06	7.1e - 06
(5)	1.3e - 05	-2.0e - 07	NR	1.3e - 05	3.6e - 07	-2.8e - 04	5.2e + 00	1.2e - 05	1.3e - 05
(6)	1.7e - 05	-5.7e - 07	NR	1.7e - 05	-1.2e - 07	-9.2e - 04	6.6e + 00	1.8e - 05	1.8e - 05
(7)	2.1e - 05	-1.1e - 06	NR	2.1e - 05	-6.9e - 09	-7.2e - 04	8.6e + 00	1.9e - 05	2.0e - 05
(8)	2.2e - 05	6.3e - 07	NR	2.4e - 05	-1.7e - 06	-5.5e - 04	1.1e + 01	-1.1e + 02	2.5e - 05
(9)	4.0e - 05	1.0e - 06	NR	4.4e - 05	-3.5e - 06	-1.5e - 03	1.6e + 01	-2.0e + 02	5.1e - 05
(10)	3.3e - 05	1.0e - 05	NR	3.5e - 05	-1.5e - 05	-1.4e - 03	1.9e + 01	-2.3e - 04	6.8e - 05

[1]Reference number associated with a molecular formula (biomolecule).
[2]Nothing reasonable reported (NR).

from the computed CGIDs the average molecular masses and standard deviations, and compare them to the corresponding theoretical values that can be analytically calculated.

The results of using the above mentioned customized carbon abundances are shown in Table 5. The differences displayed in Table 5 show that MIDAs[a,b], JFC, and Emass have the smallest errors. However, in terms of ID's standard deviations, Mercury and Mercury5 yield comparable errors to MIDAs[a,b], JFC, and Emass. The results for Qmass are similar to the ones obtained in Table 4: in terms of average masses and standard deviations, it yields small errors for small to medium sizes biomolecules. Table 5 also shows that the current versions of BRAIN and NC are not able to compute IDs using the modified isotopic abundances for carbon. However, the developers of NC have mentioned how NC could be modified to handle stable isotope enrichment by partition of the elements of enriched isotopes away from the equatransneutronic isotopes groups [17]. This option is not currently available in NC. Also, the proposed

Table 6. Coarse-Grained Isotopic Distribution (CGID) Fidelity Assessment Results τ is the Number of Terms in the Exact CGID Having Probability Greater than 5e - 12. $\Delta\tau$ is the Difference Between τ and the Number of Terms of a Computed CGID. $\Delta\chi$ is the Difference Between the Sum of Probability Terms from the Exact CGID and the Sum of Probability terms from the Computed CGID; σ_m is the Root-Mean-Square Differences of Masses Between Exact and Computed CGID, see Equation (11); U is the Number of Terms from the Computed CGID that are not with $\pm 2\epsilon$ (ϵ = 1 Da) from any Terms in the Exact CGID; E is the Number of Terms in the Exact CGID that Have at Least One Corresponding Term in Computed CGID that are with $\pm 2\epsilon$; ρ is the Weighted Correlation Between Computed and Exact CGID

No.[1]	τ	$\Delta\tau$	$\Delta\chi$	σ_m	U	E	ρ	Method
(11)	6	0	-5.6e - 16	5.1e - 13	0	6	1.0	MIDAs[a]
		0	-7.4e - 15	2.2e - 04	0	6	1.0	MIDAs[b]
		0	-4.7e - 04	1.2e - 14	0	6	0.99999988	Emass
		0	-4.4e - 16	2.1e - 05	0	6	1.0	JFC
(12)	7	0	-8.9e - 16	7.7e - 12	0	7	1.0	MIDAs[a]
		0	7.8e - 16	7.5e - 05	0	7	1.0	MIDAs[b]
		0	-8.3e - 04	2.6e - 14	0	7	0.99999963	Emass
		0	-5.6e - 16	1.3e - 05	0	7	1.0	JFC
(13)	12	0	-5.0e - 15	1.5e - 11	0	12	1.0	MIDAs[a]
		0	-2.2e - 14	3.3e - 05	0	12	1.0	MIDAs[b]
		0	-1.6e - 03	7.3e - 14	0	12	0.99999885	Emass
		0	-5.0e - 15	8.3e - 04	0	12	1.0	JFC
(14)	15	0	-1.0e - 14	1.2e - 11	0	15	1.0	MIDAs[a]
		0	1.5e - 14	2.3e - 05	0	15	1.0	MIDAs[b]
		0	-1.3e - 03	1.9e - 13	0	15	0.99999957	Emass
		0	-1.2e - 14	3.6e - 03	0	15	1.0	JFC
(15)	40	0	-9.8e - 14	1.0e - 11	0	40	1.0	MIDAs[a]
		0	1.3e - 12	6.3e - 06	0	40	1.0	MIDAs[b]
		0	-5.6e - 04	3.2e - 12	0	40	0.99999996	Emass
		0	-9.7e - 14	2.5e - 03	0	40	1.0	JFC
(16)	139	0	-9.6e - 13	4.4e - 10	0	139	1.0	MIDAs[a]
		0	3.8e - 12	6.3e - 06	0	139	1.0	MIDAs[b]
		0	-1.9e - 04	5.2e - 11	0	139	1.0	Emass
		0	-6.6e - 13	4.1e - 02	0	139	1.0	JFC
(17)	195	0	-2.0e - 12	5.5e - 10	0	195	1.0	MIDAs[a]
		0	1.3e - 12	6.3e - 06	0	195	1.0	MIDAs[b]
		0	-1.3e - 04	1.3e - 10	0	195	1.0	Emass
		0	-1.9e - 12	6.1e - 02	0	195	1.0	JFC
(18)	238	0	-3.0e - 12	9.4e - 10	0	238	1.0	MIDAs[a]
		0	2.5e - 11	6.4e - 06	0	238	1.0	MIDAs[b]
		0	-1.1e - 04	2.3e - 10	0	238	1.0	Emass
		0	-2.6e - 12	6.0e - 02	0	238	1.0	JFC
(19)	274	0	-4.1e - 12	1.2e - 09	0	274	1.0	MIDAs[a]
		1	2.5e - 11	6.1e - 02	0	274	1.0	MIDAs[b]
		0	-9.5e - 05	3.0e - 10	0	274	1.0	Emass
		0	-5.4e - 12	5.9e - 02	0	274	1.0	JFC
(20)	306	0	-4.8e - 12	1.6e - 09	0	306	1.0	MIDAs[a]
		0	2.6e - 11	6.8e - 06	0	306	1.0	MIDAs[b]
		0	-8.5e - 05	4.2e - 10	0	306	1.0	Emass
		0	-4.6e - 12	5.6e - 02	0	306	1.0	JFC

[1]Reference number associated with a molecular formula (biomolecule).

solution reduces NC back to a polynomial method algorithm, which, if not efficiently implemented, can significantly influence the overall computation time. In BRAIN's case, there are no reasonable IDs reported and it is difficult to speculate what might have happened.

Assessing Fidelity of Computed CGIDs and FGIDs

To evaluate the fidelity of CGIDs and FGIDs reported, we used 10 hydrocarbon molecules [numbered (11)–(20) in Table 2] because the "exact" CGIDs and FGIDs can be calculated for these molecules. *Exact* CGID is defined as follows. First, one merges isotopic variants that have the same nucleon number into one aggregated isotopic variant, whose corresponding molecular mass (MM) and occurrence probability are computed respectively from the probability-weighted sum of masses and from the sum of the probabilities of the isotopic variants merged. However, only aggregated isotopic variants having probability greater than 5e–12 were retained for accuracy evaluation. The *exact* FGIDs were obtained/defined similarly to the exact CGIDs, except that one merges only isotopic variants whose molecular mass differences are within some pre-specified mass accuracy, here set to 0.01 Da. The probability cutoff of 5e–12, for typical sample loads, probably already surpasses the detection capability of current mass spectrometer. Furthermore, it is also a small enough cutoff that ignoring terms below the cutoff has negligible effect in the ID profile.

Four quantities were then utilized to evaluate the fidelity of computed IDs. The first quantity is the difference in the numbers of terms ($\Delta\tau$) kept by a computed ID and by its corresponding exact ID, be it the exact CGID or the exact FGID. The second quantity was the difference in the

Table 7. Fine - Grained Isotopic Distribution (FGID) Fidelity Assessment Results τ is the number of terms in the exact FGID having probability greater than 5e - 12; $\Delta\tau$ is the difference between τ and the number of terms of a computed FGID; $\Delta\chi$ is the difference between the sum of probability terms from the exact FGID and the sum of probability terms from the computed FGID; σ_m is the root-mean-square differences of masses between exact and computed FGID, see Equation (11); U is the number of terms from the computed FGID that are not with $\pm 2\epsilon$ (ϵ=0.01 Da) from any terms in the exact FGID; E is the number of terms in the exact FGID that have at least one corresponding term in computed FGID that are with $\pm 2\epsilon$; ρ is the weighted correlation between computed and exact FGID

No.[1]	(ppm)[2]	τ	$\Delta\tau$	$\Delta\chi$	σ_m	U	E	ρ	Method
(11)	307.25	6	0	-5.6e - 16	1.1e - 09	0	6	1.0	MIDAs[a]
			0	8.9e - 14	7.3e - 05	0	6	1.0	MIDAs[b]
			0	1.5e - 08	1.8e - 09	0	6	1.0	IC
(12)	153.63	7	0	-2.7e - 15	4.2e - 10	0	7	1.0	MIDAs[a]
			0	2.3e - 12	1.6e - 05	0	7	1.0	MIDAs[b]
			0	3.7e - 07	2.3e - 09	0	7	1.0	IC
(13)	30.72	13	1	-5.7e - 13	3.1e - 03	0	13	0.99999591	MIDAs[a]
			0	3.0e - 13	4.4e - 03	0	12	0.99999592	MIDAs[b]
			1	-2.9e - 07	3.1e - 03	0	13	0.99999591	IC
(14)	15.36	16	3	2.7e - 12	5.3e - 03	0	15	0.99937104	MIDAs[a]
			3	3.6e - 12	7.1e - 03	0	15	0.99937227	MIDAs[b]
			4	-3.5e - 07	5.1e - 03	0	16	0.99937103	IC
(15)	1.53	65	-6	-4.6e - 11	1.7e - 03	0	58	0.99927870	MIDAs[a]
			3	1.2e - 11	4.0e - 03	0	64	0.98806083	MIDAs[b]
			5	2.4e - 05	3.6e - 03	0	65	0.98803755	IC
(16)	0.15	291	-5	2.6e - 11	5.2e - 03	0	257	0.99999001	MIDAs[a]
			53	6.9e - 11	5.6e - 03	1	282	0.99958599	MIDAs[b]
			82	1.4e - 07	5.2e - 03	0	280	0.99998237	IC
(17)	0.077	500	-18	1.8e - 10	4.0e - 03	0	453	0.99785805	MIDAs[a]
			126	1.3e - 10	7.3e - 03	13	488	0.95950051	MIDAs[b]
			124	-1.6e - 08	4.5e - 03	0	496	0.99951891	IC
(18)	0.051	715	-16	-5.4e - 10	4.5e - 03	0	636	0.99466182	MIDAs[a]
			242	1.5e - 10	7.0e - 03	10	681	0.71069880	MIDAs[b]
			19	-5.0e - 08	4.3e - 03	0	690	0.99735244	IC
(19)	0.038	881	57	7.6e - 11	4.8e - 03	0	824	0.95007936	MIDAs[a]
			437	1.9e - 10	7.8e - 03	33	866	0.59270671	MIDAs[b]
			-26	-1.7e - 06	5.9e - 03	0	713	0.97935224	IC
(20)	0.031	1143	93	-1.4e - 10	5.7e - 03	0	1011	0.85638390	MIDAs[a]
			498	2.2e - 10	8.4e - 03	47	1092	0.63960000	MIDAs[b]
			-173	-1.7e - 05	4.6e - 03	0	838	0.98564325	IC

[1]Reference number associated with a molecular formula (biomolecule).
[2]Using a mass accuracy of 2ϵ the equivalent resolution in parts per million.

probability sums ($\Delta\chi$), one from the computed ID and the other from the exact FGID (or the exact CGID). The third quantity was the root-mean-square differences of masses (σ_m) between computed and the exact CGIDs (or the exact FGIDs).

$$\sigma_m^2 \equiv \frac{1}{N}\sum_{i=1}^{N} \min_{j\epsilon \text{exact}} \left|m_i - m_j\right|^2. \tag{11}$$

In the equation above, m_i represents a computed mass term while each m_j represents a mass term in the exact FGID (or the exact CGID). N is the number of terms retained in the computed ID. That is, for every mass term in a computed ID, the closest mass term within the exact FGID (or the exact CGID) is found and their difference square is summed. The average of such sum of squares constitutes ${\sigma_m}^2$. The fourth quantity computed was the weighted correlation (ρ) between computed and exact IDs. The weighted correlation (ρ) is defined as follows. Let $p(m_i)$ and $p(m_j)$ be the terms of a computed ID and the corresponding exact FGID (or exact CGID), respectively. We first introduce the weight (w_{ij}) between a computed ID term (index i) and exact FGID (or exact CGID) term (index j) as

$$w_{ij} = \begin{cases} e^{-\zeta}, \zeta = \min_j \left|m_i - m_j\right| \le 2\epsilon \\ 0\ , \ \min_j \left|m_i - m_j\right| \ > \ 2\,\epsilon, \end{cases} \tag{12}$$

where, in the above equation, $min_j\ |m_i - m_j|$, is the minimum mass difference between a term (m_i) from the computed ID and terms (m_j) from the exact ID. The computed weights (w_{ij}) are then normalized by the normalization factor, $W_j = \sum_i w_{ij}$, by summing over all i terms from the computed ID that are close to a common term j in the exact FGID (or the exact CGID). The weighted correlation using the above definitions is given by

$$\rho = \frac{\sum_{j=1}^{M}\sum_{i=1}^{N} p(m_i)p\left(m_j\right)w_{ij}/W_j}{\sqrt{\sum_{i=1}^{N} p(m_i)^2}\sqrt{\sum_{j=1}^{M} p\left(m_j\right)^2}}. \tag{13}$$

For CGIDs, ϵ is set to one Da, while for FGIDs, ϵ is set to 0.01 Da.

Using molecules numbered (11)–(20) in Table 2, we document the analysis results of the four quantities mentioned above in Table 6 (for CGIDs) and Table 7 (for FGIDs). For CGIDs, we include in Table 6 only four methods that largely satisfy the

Table 8. Computation Time in Seconds (s) and Number of Terms Reported with MIDAs's Computed Coarse-Grained (CG) and Fine-Grained (FG) Isotopic Distributions (ID) Using 1.0 Da and 0.01 Da Mass Accuracy, Respectively

MIDAs[a]

No.[1]	Number of terms CGID	CGID time(s)	Number of terms FGID	FGID time(s)
(1)	85	0.006	42	0.001
(2)	154	0.01	151	0.002
(3)	186	0.02	246	0.001
(4)	203	0.02	301	0.001
(5)	290	0.04	809	0.006
(6)	341	0.05	1269	0.01
(7)	423	0.1	1945	0.05
(8)	540	0.13	3145	0.06
(9)	820	0.14	6579	0.3
(10)	956	0.2	7850	0.4
(21)	3022	0.35	13834	0.8
(22)	5908	1.0	74994	1.0
(23)	2706	0.23	28508	2.1
(24)	617	0.05	2805	0.01
(25)	2623	0.21	18261	0.5

MIDAs[b]

No.	Number of terms CGID	CGID time(s)	Number of terms FGID	FGID time(s)
(1)	15	0.0006	29	0.025
(2)	29	0.001	114	0.041
(3)	38	0.001	193	0.043
(4)	42	0.001	241	0.08
(5)	78	0.002	740	0.08
(6)	95	0.002	1166	0.14
(7)	123	0.004	1784	0.14
(8)	157	0.004	2953	0.14
(9)	230	0.004	6527	0.3
(10)	257	0.01	7818	0.3
(21)	794	0.005	12405	0.4
(22)	1500	0.01	74994	1.0
(23)	706	0.01	23367	0.7
(24)	188	0.003	2209	0.2
(25)	698	0.01	16384	0.8

[1]Reference number associated with a molecular formula (biomolecule or inorganic compound).

criteria for being a sound IDCM of application value. For FGIDs, only IC, MIDAs[a] and MIDAs[b] are included in Table 7 since they are the only methods that can do FGID-computing reasonably fast and without additional post processing.

For fidelity assessment of CGIDs, all four methods shown in Table 6 yield small $\Delta\tau$ and ρ values close to one. In terms of σ_m and $\Delta\chi$, more differences are revealed. Emass always yields small σ_m, reflecting good fidelity in terms of mass locations, but seems to give a larger $|\Delta\chi|$, reflecting less accuracy in amplitudes. JFC and MIDAs[b] seem to yield less precise mass locations, evidenced by a larger σ_m, but seem to provide more accurate amplitudes, evidenced by a smaller $|\Delta\chi|$. MIDAs[a] yields both accurate mass locations and accurate amplitudes.

The values of $\Delta\chi$ and σ_m in Table 7 indicate that IC, MIDAs[a], and MIDAs[b] report FGID terms with similar mass accuracy and with probability sums that are close to the expected value. For small to medium molecules, numbered (11)–(15), IC, MIDAs[a], and MIDAs[b] have equivalently accurate results. For molecules numbered (16)–(20), IC and MIDAs[a] have comparable performances, both slightly better than MIDAs[b]. The values for $\Delta\tau$ indicates that MIDAs[b] reports many more terms than expected in its computed FGID. Not expecting any leakage, MIDAs[b] gains these extra terms mainly due to rounding errors associated with the DFFT numerical procedure.

The difference observed in $\Delta\tau$ for MIDAs[a] is caused by the pruning and merging procedures employed by the algorithm. All the FGID terms computed by IC and MIDAs[a] are within 2ϵ from the exact FGID terms, which is shown by the number of unexplained term (U) being zero in Table 7. It is also true that most of the terms computed from MIDAs[b] are within 2ϵ from the exact FGID terms with the exception of molecules (17)–(20) where the number U ranges from 1 to 47. The computed weighted correlation also shows that for heavier molecules, (18)–(20), both IC and MIDAs[a] produce FGIDs that are more similar to the exact FGIDs than MIDAs[b].

What causes MIDAs[b] to perform worse here might be related to the fact that pinning the elemental masses to grid points may introduce appreciable mass errors while computing IDs for larger molecules. In the worst case scenario, the mass error introduced is apparently proportional to the number of atoms contained in the molecule. Even though MIDAs[b] employs a mass rescaling [32] to bring the computed average masses and standard deviations close to their theoretical values, the linear mass rescaling is not sufficient to guarantee the full profile resemblance between the computed ID and the exact ID. The non-negligible discrepancy (indicated by the weighted correlation ρ not very close to one) between the computed FGID and the exact FGID for molecules (18)–(20) is reflecting this problem.

MIDAs Web Interface

MIDAs web interface http://www.ncbi.nlm.nih.gov/CBBresearch/Yu/midas/index.html is user-friendly, but at the same time offers considerable flexibility. For example, in terms of the input molecule, the user may type in the box an elemental composition, a molecular formula, a peptide, or even a protein sequence. The program recognizes the input molecule in all formats above and extracts the corresponding elemental compositions for computing CGIDs and FGIDs. The isotopic abundances and elements' masses can also be customized within the web interface. The user simply clicks on the "change" button to edit the abundance table of all elements. Other fields that can be easily customized and specified by the user are the charge of the input molecule and the cutoff probability. MIDAs displays both CGID and FGID together using user-specified accuracies, one for each. The "algorithm" drop down box allows the user to select either the FFT or the polynomial algorithms. The output, including the lightest mass, theoretical average mass, theoretical mass standard deviation, computed average mass, computed mass standard deviation, FGID peak list, and CGID peak list can be exported to a flat file by clicking on the "download output" button on the result page. There is also a contextual help for every functional button.

Conclusion and Outlook

The two algorithms introduced here, MIDAs[a] and MIDAs[b], for the 25 molecules tested, seem to be able to compute IDs quickly and accurately. Between the two, MIDAs[a] seems slightly more accurate. For CGIDs MIDAs[b] appears to be faster (see Table 8), whereas for FGIDs they are of comparable speed. Both algorithms benchmark well with existing methods and stand out because of their ability to compute CGIDs and FGIDs using a user-specified accuracy. These two algorithms were also shown to accurately compute IDs for molecules labeled with stable isotopes, which was not the case for some of the methods evaluated. In summary, in terms of CGIDs derived average masses, MIDAs[a], MIDAs[b], JFC, and Emass yield smaller errors than other methods. In terms of CGIDs derived standard deviation, our investigation shows that MIDAs[a], MIDAs[b], JFC, and Emass yield smaller errors than other methods. When computing the FGID, MIDAs[a] computes a FGID that better resembles the exact FGID than MIDAs[b] using our evaluation gauges. Both algorithms described here were coded using the C++ programming language in a computer program called MIDAs that is available for download at http://www.ncbi.nlm.nih.gov/CBBresearch/Yu/downloads/MIDAs.html. To make these algorithms widely accessible, we have made them available through a user-friendly web-interface at http://www.ncbi.nlm.nih.gov/CBBresearch/Yu/midas/index.html.

Acknowledgments

The authors thank Alfred Yergey for sending them the NeutronCluster code, and Alan Rockwood for providing them with codes of Mercury, Emass, Qmass, and Mercury5. The authors thank the administrative group of the National Institutes of Health Biowulf Clusters, where all the computational tasks were carried out. They also thank the National Institutes of Health Fellows Editorial Board for editorial assistance. This work was supported by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health.

Funding for Open Access publication charges for this article was provided by the National Institutes of Health.

Open Access

Appendix

Using Convolution theorem in Discrete Fourier Transform

For completeness, we first review a few important properties of the DFT. Consider a function H sampled at L equally-spaced values. We shall denote $H(x = n\epsilon)$ by H_n with $n = 0, 1, 2, \ldots, L-1$. One may then consider the DFT of this function by constructing

$$h_k = \sum_{n=0}^{L-1} e^{-2\pi i k(n\epsilon)/(L\epsilon)} H_n,$$

with $k = 0, 1, 2, \ldots, L-1$. With h_k given, one can also invert the expression above to yield

$$H_n = \frac{1}{L} \sum_{k=0}^{L-1} e^{2\pi i k(n\epsilon)/(L\epsilon)} h_k,$$

where the identity (when d is an integer)

$$\sum_{n=0}^{L-1} e^{2\pi i d n/L} = L\delta_{d,0}$$

is used. Evidently, ϵ is the spacing between each pair of sampled values along the variable x. Although we only specify H on L points, using the Fourier expression of H_n, we can easily see that $H_{n+L} = H_n$. That is, the DFT effectively makes the function considered, say H, periodic with period L.

Given two periodic functions H and G of period L, we may consider the following convolution

$$W_n = \sum_{l=0}^{L-1} H_{n-l}\, G_l \ .$$

We can now compute

$$w_k = \sum\nolimits_{n=0}^{L-1} e^{-2\pi i k n/L} \sum\nolimits_{l=0}^{L-1} H_{n-l} G_l$$
$$= \sum\nolimits_{l=0}^{L-1} e^{-2\pi i k l/L} G_l \left[\sum\nolimits_{n=0}^{L=1} e^{-2\pi i k(n-l)/L} H_{n-l} \right] = g_k h_k,$$

where we have employed the periodic property of H and

$$g_k = \sum_{l=0}^{L-1} e^{-2\pi i k l/L} G_l$$

is the Fourier transform of G. The inverse transform of w_k of course leads to W_n without any *leakage* issue involved.

When applying the DFT to compute an ID, however, one also need to pay attention to the issue of *folded-back*. To illustrate this problem, let us consider a toy example where an element E has two isotopes with masses ϵ and 2ϵ. For simplicity, let us also assume that both isotopes occur with equal probability 1/2. If one chooses to use grid size L=4 and compute the ID of the molecule E_2 using DFT, one starts with the following EFP

$$\frac{1}{2}\left(e^{-2\pi\pi i k/4} + e^{-4\pi\pi i k/4}\right)$$

and raise it to the second power to yield

$$w_k(\mathrm{E}_2) = \frac{1}{2^2}\left(e^{-4\pi i k/4} + 2e^{-6\pi i k/4} + e^{-8\pi i k/4}\right)$$
$$= \frac{1}{2^2}\left(e^{-4\pi i k/4} + 2e^{-6\pi i k/4} + 1\right)$$

which yields, upon inverting back to the mass domain, probability 1/4 for masses 2ϵ and zero and probability 1/2 for mass 3ϵ. The reason for a zero mass is due to the periodicity. By identifying zero with 4ϵ, one gains the results expected: a probability of 1/4 for both masses 2ϵ and 4ϵ. Absent the molecular masses of 1ϵ and 5ϵ and so on, we see that the mass distribution of the molecule E_2 is resolved and appears correctly within the mass range from ϵ to 4ϵ.

However, if one continues to keep the grid size $L = 4$ while considering the molecule E_4, the Fourier transform of the molecule's mass distribution becomes

$$\frac{e^{-8\pi i k/4}}{2^4}\left(1 + 4e^{-2\pi i k/4} + 6e^{-4\pi i k/4} + 4e^{-6\pi i k/4} + e^{-8\pi i k/4}\right)$$
$$= \frac{1}{2^4}\left(1 + 4e^{-2\pi i k/4} + 6e^{-4\pi i k/4} + 4e^{-6\pi i k/4} + 1\right),$$

which upon inversion yields masses zero, ϵ, 2ϵ, and 3ϵ respectively with probabilities $2/2^4$, $4/2^4$, $6/2^4$, and $4/2^4$. Remembering the periodicity, one may recognize that it is the set of masses 4ϵ, 5ϵ, 6ϵ, and 7ϵ (instead of zero, ϵ, 2ϵ, and 3ϵ) that acquires the set of probabilities $2/2^4$, $4/2^4$, $6/2^4$, and $4/2^4$. However, a simple calculation yields the possible masses to be 4ϵ, 5ϵ, 6ϵ, 7ϵ, and 8ϵ with respective probabilities $1/2^4$, $4/2^4$, $6/2^4$, $4/2^4$ and $1/2^4$. What has happened is that the mass 8ϵ is now *folded-back* to 4ϵ due to the inherent periodicity caused by DFT with L=4. With this illustrative example, one can see that in order to avoid the *folded-back* artifact, one needs to have enough sample points so that the mass range used for the DFT is larger than the mass span of the molecule considered. However, if the tails of the mass distribution have very small probabilities, one might be able to use a smaller number of sample points with only

a weak *folded-back* effect that only causes negligible distortion on the ID profile.

In general, when the number L is fixed, the folded-back problem should be less severe for the CGID when compare to its FGID counter-part. This is because if one keeps L fixed but decreases the mass difference between adjacent points, the effective mass range shrinks and there exists the possibility when regions with significant probabilities are now folded back to a particular mass window, where much smaller probabilities are assumed if no folded-back occurs. It is for this reason that MIDAs does not fix the number of sampled points, but rather increases it in proportion to $1/\epsilon$.

References

1. Rosman, K., Taylor, P.: Isotopic compositions of elements 1997. Pure App. Chem. **70**, 217–235 (1998)
2. Michener, R., Lajtha, K.: Stable Isotopes in Ecology and Environmental Science. Wiley-Blackwell, Massachusetts (2007)
3. Becker, G.W.: Stable isotopic labeling of proteins for quantitative proteomic applications. Brief. Funct. Genom. Proteom. **7**(5), 371–382 (2008)
4. Yamamoto, H., McCloskey, J.A.: Calculations of isotopic distribution in molecules extensively labeled with heavy isotopes. Anal. Chem. **49**(2), 281–283 (1977)
5. Rockwood, A.L.: Deconvoluting isotopic distributions to evaluate parent/fragment ion relationships. Rapid Commun. Mass Spectrom. **11**(3), 241–248 (1997)
6. Gay, S., Binz, P.A., Hochstrasser, D.F., Appel, R.D.: Modeling peptide mass fingerprinting data using the atomic composition of peptides. Electrophoresis **20**(18), 3527–3534 (1999)
7. Blank, P., Sjomeling, C., Backlund, P., Yergey, A.: Use of cumulative distribution functions to characterize mass spectra of intact proteins. J. Am. Soc. Mass Spectrom. **13**, 40–46 (2002)
8. Goodlett, D.R., Bruce, J.E., Anderson, G.A., Rist, B., Pasa-Tolic, L., Fiehn, O., Smith, R.D., Aebersold, R.: Protein identification with a single accurate mass of a cysteine-containing peptide and constrained database searching. Anal. Chem. **72**(6), 1112–1118 (2000)
9. Polacco, B.J., Purvine, S.O., Zink, E.M., Lavoie, S.P., Lipton, M.S., Summers, A.O., Miller, S.M.: Discovering mercury protein modifications in whole proteomes using natural isotope distributions observed in liquid chromatography-tandem mass spectrometry. Mol. Cell. Proteom. (2011). doi:10.1074/mcp.M110.004853
10. Roussis, S.G., Proulx, R.: Reduction of chemical formulas from the isotopic peak distributions of high-resolution mass spectra. Anal. Chem. **75**(6), 1470–1482 (2003)
11. Nikolaev, E.N., Jertz, R., Grigoryev, A., Baykut, G.: Fine structure in isotopic peak distributions measured using a dynamically harmonized Fourier transform ion cyclotron resonance cell at 7 t. Anal. Chem. **84**(5), 2275–2283 (2012)
12. Russell, D.H., Edmondson, R.D.: High-resolution mass spectrometry and accurate mass measurements with emphasis on the characterization of peptides and proteins by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. J. Mass Spectrom. **32**, 263–276 (1997)
13. Werlen, R.C.: Effect of resolution on the shape of mass spectra of proteins: some theoretical considerations. Rapid Commun. Mass Spectrom. **8**(12), 976–980 (1994)
14. Marshall, A.G., Hendrickson, C.L.: High-resolution mass spectrometers. Annu. Rev. Anal. Chem. **1**(1), 579–599 (2008)
15. Michalski, A., Damoc, E., Lange, O., Denisov, E., Nolting, D., Muller, M., Viner, R., Schwartz, J., Remes, P., Belford, M., Dunyach, J.J., Cox, J., Horning, S., Mann, M., Makarov, A.: Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. Mol. Cell. Proteom. (2012). doi:10.1074/mcp.O111.013698
16. Valkenborg, D., Mertens, I., Lemiere, F., Witters, E., Burzykowski, T.: The isotopic distribution conundrum. Mass Spectrom. Rev. **31**(1), 96–109 (2012)
17. Olson, M., Yergey, A.: Calculation of the isotope cluster for polypeptides by probability grouping. J. Am. Soc. Mass Spectrom. **20**, 295–302 (2009)
18. Claesen, J., Dittwald, P., Burzykowski, T., Valkenborg, D.: An efficient method to calculate the aggregated isotopic distribution and exact center-masses. J. Am. Soc. Mass Spectrom. **23**, 753–763 (2012)
19. Rockwood, A.L., Van Orden, S.L., Smith, R.D.: Rapid calculation of isotope distributions. Anal. Chem. **67**(15), 2699–2704 (1995)
20. Rockwood, A.L., Van Orman, J.R., Dearden, D.V.: Isotopic compositions and accurate masses of single isotopic peaks. J. Am. Soc. Mass Spectrom. **15**(1), 12–21 (2004)
21. Rockwood, A., Haimi, P.: Efficient calculation of accurate masses of isotopic peaks. J. Am. Soc. Mass Spectrom. **17**, 415–419 (2006)
22. Senko, M.W.: IsoPro computer program 3.0.
23. Snider, R.: Efficient calculation of exact mass isotopic distributions. J. Am. Soc. Mass Spectrom. **18**, 1511–1515 (2007)
24. Rockwood, A.L., Van Orden, S.L., Smith, R.D.: Ultrahigh resolution isotope distribution calculations. Rapid Commun. Mass Spectrom. **10**(1), 54–59 (1996)
25. Li, L., Kresh, J.A., Karabacak, N.M., Cobb, J.S., Agar, J.N., Hong, P.: A hierarchical algorithm for calculating the isotopic fine structures of molecules. J. Am. Soc. Mass Spectrom. **19**(12), 1867–1874 (2008)
26. Fernandez-de Cossio, J.: Efficient packing Fourier-transform approach for ultrahigh resolution isotopic distribution calculations. Anal. Chem. **82**(5), 1759–1765 (2010)
27. Brownawell, M.L., San Filippo, J.: Simulation of chemical instrumentation. ii: A program for the synthesis of mass spectral isotopic abundances. J. Chem. Educ. **59**(8), 663 (1982)
28. Yergey, J.A.: A general approach to calculating isotopic distributions for mass spectrometry. Int. J. Mass Spectrom. Ion Phys. **52**, 337–349 (1983)
29. Rockwood, A.L.: Relationship of Fourier transforms to isotope distribution calculations. Rapid Commun. Mass Spectrom. **9**(1), 103–105 (1995)
30. Cooley, J.W.: The rediscovery of the fast Fourier transform algorithm. Mikrochim. Acta **3**, 33–45 (1987)
31. Dittwald, P., Claesen, J., Burzykowski, T., Valkenborg, D., Gambin, A.: BRAIN: a universal tool for high-throughput calculations of the isotopic distribution for mass spectrometry. Anal. Chem. **85**(4), 1991–1994 (2013)
32. Rockwood, A.L., Van Orden, S.L.: Ultrahigh-speed calculation of isotope distributions. Anal. Chem. **68**(13), 2027–2030 (1996)
33. Li, L., Karabacak, N.M., Cobb, J.S., Wang, Q., Hong, P., Agar, J.N.: Memory-efficient calculation of the isotopic mass states of a molecule. Rapid Commun. Mass Spectrom. **24**(18), 2689–2696 (2010)
34. Fernandez-de Cossio Diaz, J., Fernandez-de Cossio, J.: Computation of isotopic peak center-mass distribution by fourier transform. Anal. Chem. **84**(16), 7052–7056 (2012)
35. Fernandez-de Cossio, J., Gonzalez, L.J., Satomi, Y., Betancourt, L., Ramos, Y., Huerta, V., Amaro, A., Besada, V., Padron, G., Minamino, N., Takao, T.: Isotopica: a tool for the calculation and viewing of complex isotopic envelopes. Nucleic Acids Res. **32**(Web Server issue), W674–W678 (2004)
36. Beynon, J.H.: Mass Spectrometry and Its Applications to Organic Chemistry. Elsevier, New York (1960)
37. Kubinyi, H.: Calculation of isotope distributions in mass spectrometry. A trivial solution for a non-trivial problem. Anal. Chim. Acta **247**(1), 107–119 (1991)
38. Margrave, J.L., Polansky, R.B.: Relative abundance calculations for isotopic molecular species. J. Chem. Educ. **39**(7), 335 (1962)
39. Bohman, H.: Approximate Fourier analysis of distribution functions. Arkiv Matematik **4**, 99–157 (1961)
40. Harris, F.J.: High-resolution spectral analysis with arbitrary spectral centers and arbitrary spectral resolutions. Comput. Electr. Eng. **3**(2), 171–191 (1976)
41. Harris, F.: On the use of windows for harmonic analysis with the discrete Fourier transform. Proc. IEEE **66**(1), 5–83 (1978)
42. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes in C. The Art of Scientific Computing, 2nd edn. Cambridge University Press, New York, NY (1992)
43. Gould, H.: Fibonacci Q. **37**(2), 135–140 (1999)
44. Elliott, M.H., Smith, D.S., Parker, C.E., Borchers, C.: Current trends in quantitative proteomics. J. Mass Spectrom. **44**(12), 1637–1660 (2009)