analytical chemistry

Letter

pubs.acs.org/ac

BRAIN: A Universal Tool for High-Throughput Calculations of the Isotopic Distribution for Mass Spectrometry

Piotr Dittwald,*[,†,‡] Jürgen Claesen,[¶] Tomasz Burzykowski,[¶] Dirk Valkenborg,[¶,§,∥] and Anna Gambin[†,⊥]

[†]Institute of Informatics, University of Warsaw, Poland

[‡]College of Inter-faculty Individual Studies in Mathematics and Natural Sciences, University of Warsaw, Poland

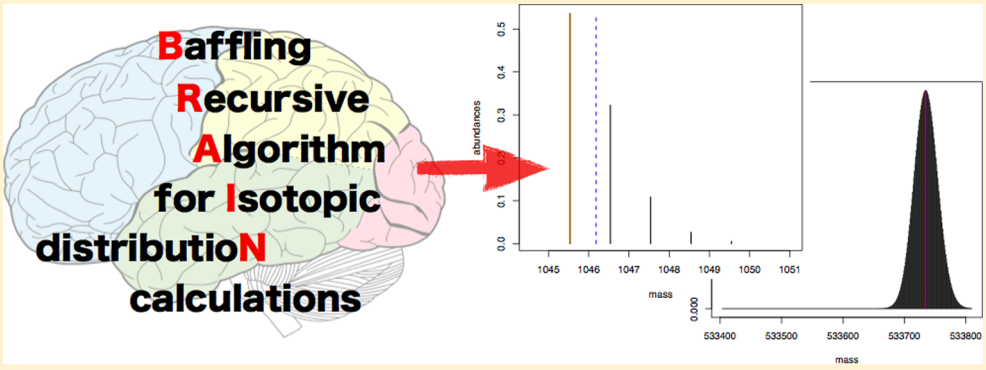
[¶]I-BioStat, Hasselt University, Belgium

[§]Flemish Institute for Technological Research (VITO), Belgium

[∥]CFP-CeProMa, University of Antwerp, Belgium

[⊥]Mossakowski Medical Research Centre, Polish Academy of Sciences, Warsaw, Poland

Supporting Information



ABSTRACT: This Letter presents the R-package implementation of the recently introduced polynomial method for calculating the aggregated isotopic distribution called BRAIN (Baffling Recursive Algorithm for Isotopic distributioN calculations). The algorithm is simple, easy to understand, highly accurate, fast, and memory-efficient. The method is based on the application of the Newton-Girard theorem and Viète's formulae to the polynomial coding of different aggregated isotopic variants. As a result, an elegant recursive equation is obtained for computing the occurrence probabilities of consecutive aggregated isotopic peaks. Additionally, the algorithm also allows calculating the center-masses of the aggregated isotopic variants. We propose an implementation which is suitable for high-throughput processing and easily customizable for application in different areas of mass spectral data analyses. A case study demonstrates how the R-package can be applied in the context of protein research, but the software can be also used for calculating the isotopic distribution in the context of lipidomics, metabolomics, glycoscience, or even space exploration. More materials, i.e., reference manual, vignette, and the package itself are available at Bioconductor online (http://www.bioconductor.org/packages/release/bioc/html/BRAIN.html).

Mass spectrometry has become a fundamental tool in proteomics, metabolomics, lipidomics, and other high-throughput studies of complex biochemical samples. The successful interpretation of mass spectrometry data often depends on the comparison of the detected signals with theoretical features of a putative molecule.[1−3] One such feature, referred to as isotopic distribution, originates from the fact that most biomolecules are composed out of polyisotopic elements. For small molecules, the calculation of the theoretical isotopic distribution is relatively easy. However, the calculation becomes very complex for large molecules such as proteins or polymers, because of a combinatorial explosion of the number of terms that must be computed. Therefore, there is a need for efficient algorithms that overcome the combinatorial problem.[4] Most methods that have been proposed are subject to limitations like, e.g., loss of accuracy or high time/memory complexity.

In this short note, we present an implementation of the method for the efficient calculation of aggregated isotopic distributions called BRAIN (Baffling Recursive Algorithm for Isotopic distributioN calculations). It should be noted that BRAIN does not calculate the isotope fine structure of the molecule, as observed by high-resolution mass spectrometry, such as FTICR MS. Instead, the BRAIN method lumps together the isotopic variants with equal nucleon count, hence the term aggregated isotope distribution. From this perspective,

Received: November 27, 2012
Accepted: January 25, 2013
Published: January 25, 2013

   dx.doi.org/10.1021/ac303439m | *Anal. Chem.* 2013, 85, 1991−1994

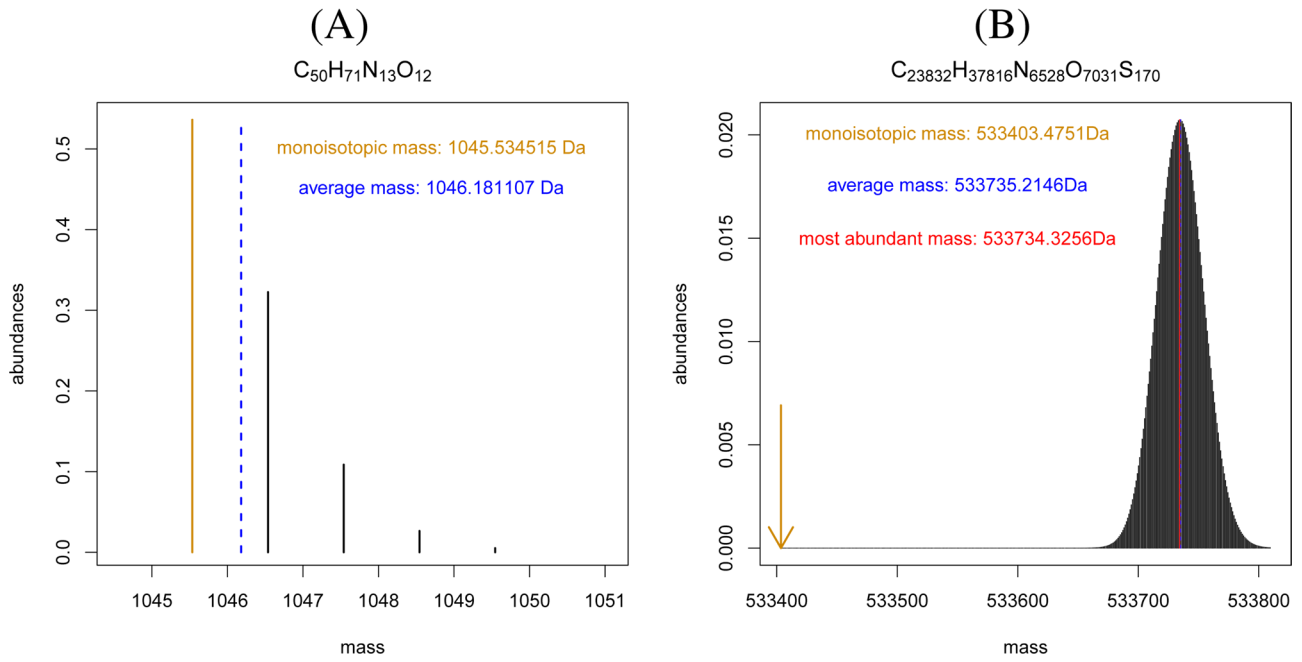


Figure 1. Stick representation of the aggregated isotopic distribution of $C_{50}H_{71}N_{13}O_{12}$ (panel A) and $C_{23832}H_{37816}N_{6528}O_{7031}S_{170}$ (panel B). In panel B, the monoisotopic mass is indicated by an arrow. Note that the BRAIN package returns the average mass of the molecule computed by a closed formula taking into consideration the average mass of the composing elements (function *calculateAverageMass*).

the result of BRAIN is similar to the isotope distribution observed in mass instruments with a resolution of 10 000−50 000. Small mass differences between isotope variants with an equal nucleon count cannot be distinguished in such mass spectrometers, because coalescing ion bundles in the mass detector. For the sake of brevity of presentation, we do not compare here our approach with competing software packages. It should be emphasized that a comprehensive comparison with other programs has been already conducted for a prototype implementation in MATLAB.[5−7]

The BRAIN algorithm requires the elemental composition and the elemental isotope distribution as main input parameters. The output of BRAIN is the aggregated isotope distribution in stick representation. Figure 1 displays the aggregated isotope distribution for a light and heavy molecule. This aggregated approach does not limit the BRAIN application for the identification of molecular species in a complex mass spectrum. Only in the case of extremely high-mass resolution on FTICR, the assumption of coalescing ion bundles is not valid anymore, and the isotopic distribution does not appear as an aggregated distribution. However, identification based upon MS1 spectra is complex. The monoisotopic and average masses (separately or combined) can be used to reduce the list of potential candidates and, in some cases, will return one single candidate. Therefore, combining MS1-based identification with MS2-based analysis is a more promising approach and can also use BRAIN to generate isotope distributions for spectral comparison.

Our approach uses the concept of the polynomial expansion[8,9] and applies the Newton-Girard theorem and Viète's formulae to obtain a simple recursive equation to calculate the occurrence probabilities of consecutive aggregated isotopic variants. The method also provides the exact center-masses of the aggregated isotopic variants.[5]

IMPLEMENTATION

In the formulation of the BRAIN algorithm, we use the representation proposed by Rockwood and Van Orden.[10] Consider a molecule with a composition $C_vH_wN_xO_yS_z$, i.e., with v carbon (C) atoms, w hydrogen (H) atoms, x nitrogen (N) atoms, y oxygen (O) atoms, and z sulfur (S) atoms. For such a molecule, let us formulate the following polynomial:

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

where I is a variable representing the additional neutron content of a molecule relative to the monoisotopic variant. The coefficients $P_{C_{12}}$, $P_{C_{13}}$, ..., $P_{S_{36}}$ correspond to the natural abundances of isotopes, e.g., $P_{C_{12}}$ = 98.93% and $P_{C_{13}}$ = 1.07%. The following expansion of the polynomial Q is of interest:

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

with $n = v + w + x + 2y + 4z$ indicating the order of the polynomial or, equivalently, the number of aggregated isotopic variants (without the monoisotopic variant). The coefficient q_j is equal to the occurrence probability of the j-th isotopic variant, i.e., the molecule with j additional neutrons as compared to the monoisotopic one. Using the Newton-Girard theorem and Viète's formulae, the following system of recursive equations for coefficients q_j can be obtained:

$$q_j = -\frac{1}{j}\sum_{l=1}^{j} q_{j-l}\psi_l \qquad (1)$$

dx.doi.org/10.1021/ac303439m | *Anal. Chem.* 2013, 85, 1991−1994

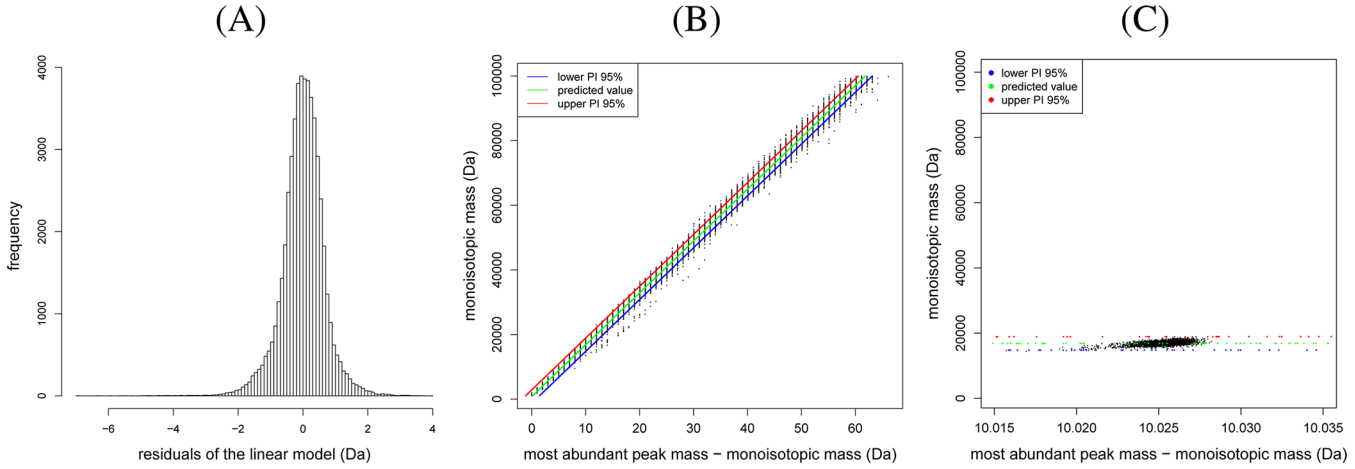


Figure 2. (A) Residuals of the linear model predicting the monoisotopic mass from the most abundant peak mass. The global shape of the residual distribution is a bell-shape curve with a standard deviation of 0.656. (B) The difference between the masses of the most abundant and monoisotopic variants is shown as a function of the monoisotopic mass. It is worth mentioning that the mass difference can exceed 60 Da or more for heavy molecules above 100 000 Da. The green line shows this difference modeled by linear regression. Blue and red colors (B, C) indicate lower and upper bounds of the 95% prediction interval (PI). Note that prediction intervals refer to a probability interval in which future responses will fall, based on previous observations, i.e., the Uniprot data mentioned in the manuscript. Panel C illustrates a close-up of the region near 10 Da in panel B. It should be noted that the vertical lines in panel B are in fact slightly distributed over the mass axis. The regression line in panel B provides a better insight into the residuals for a given mass interval (C).

where ψ_l is a power sum of roots of the polynomials $Q_C(I)$, $Q_H(I)$, ..., $Q_S(I)$. This power sum can be represented as follows:

$$\psi_l = \frac{v}{r_C^l} + \frac{w}{r_H^l} + \frac{x}{r_N^l} + \frac{y}{r_O^l} + \frac{y}{\overline{r}_O^l} + \frac{z}{r_{S,1}^l} + \frac{z}{\overline{r}_{S,1}^l} + \frac{z}{r_{S,2}^l} + \frac{z}{\overline{r}_{S,2}^l}$$

with r_C, r_H, and r_N denoting the unique roots of $Q_C(I)$, $Q_H(I)$, and $Q_N(I)$, respectively. The pair of the conjugate roots of the second-order polynomial $Q_O(I)$ are denoted by r_O and $\overline{r}_O$, with a similar notation used for the two pairs of the roots of the fourth-order polynomial $Q_S(I)$: $r_{S,1}$, $\overline{r}_{S,1}$, $r_{S,2}$, and $\overline{r}_{S,2}$. Note that in this case all roots can be expressed in a closed form. The recursion in Eq 1 starts by the calculation of the occurrence probability of the monoisotopic variant: $q_0 = P_{C_{12}}^v P_{H_1}^w P_{N_{14}}^x P_{O_{16}}^y P_{S_{32}}^z$. The occurrence probabilities corresponding to the consecutive aggregated isotopic variants are calculated recursively using Eq 1.

The recursive algorithm of BRAIN has been implemented as an R (http://www.r-project.org) package distributed via Bioconductor[11] under GNU General Public License. The computational complexity of the algorithm depends on the required number of the aggregated isotopic variants to be computed. If only the first 10 variants are required, only 10 iterations are needed. Moreover, the algorithm is memory-efficient: for each j-th aggregated isotopic variant, only two values (q_j and ψ_j) need to be stored. The function use*BRAIN* in the R package implements several stopping criteria, such as an upper bound on the coverage (expressed as a percentage) of the isotopic distribution or on the number of aggregated isotopic variants to be computed. The R-implementation of BRAIN is perfectly adapted to high-throughput computation and batch processing. As in some specific environments (e.g., in meteorites), the abundances of stable isotopes of particular atoms may differ from those reported in terrestrial matter; the user may change values of these abundances in the configuration file (see file input.R in the source code).

PROTEOMIC CASE STUDY

To illustrate the computational performance of the BRAIN implementation in a high-throughput setting, we have selected over 50 000 human proteins from the Uniprot database release 2011_11[12,13] with the monoisotopic masses lower than 100 000 Da. The goal of this simple case was to explore and model the relationship between the masses of the monoisotopic and the most abundant aggregated isotopic variants. The motivation comes from the fact that, in a mass spectrum, the peak representing the monoisotopic mass of a molecule does not necessarily correspond to the most abundant isotopic variant. In fact, for large molecules, the monoisotopic variant is often not even visible by mass spectrometry, as can be observed in Figure 1B. On the other hand, the most abundant variant can be estimated on the basis of the observed isotope pattern. Thus, an interesting question is whether the monoisotopic mass of a molecule can be predicted from the observed mass of the most abundant variant.

To this aim, the aggregated isotopic distributions of the selected proteins were calculated. More specifically, the part of the distribution between the monoisotopic and most abundant variants was obtained. Afterwards, a regression model was used to estimate the relationship between the monoisotopic mass and the mass of the most abundant aggregated variant. As a result, the following linear formula was obtained: $monoMass = 0.482 + 0.9994 \times mostAbundantPeakMass$.

From the distribution of residuals (c.f. Figure 2A), it can be seen that an error tolerance of approximately 2 Da should be allowed when predicting the monoisotopic mass based on the observed most abundant mass. For this reason, the mass spectrometry community prefers to characterize the mass of large molecules by means of the average mass of an observed isotope cluster. Figure 2B,C shows another representation of the regression model, with the mass of the monoisotopic variant regressed against the difference between the masses of the monoisotopic and the most abundant variants. Analogously, the linear model predicting average mass from the most

dx.doi.org/10.1021/ac303439m | *Anal. Chem.* 2013, 85, 1991−1994

abundant peak mass can also be built using BRAIN (c.f. Suppl. Figure S1, Supporting Information, for residuals of the model).

Finally, this application study illustrates the power of BRAIN in high-throughput processing. All computations have been performed using BRAIN (version 1.4.0.) as an open-source Bioconductor package on PC with two Intel(R) Core(TM)2 2.40 GHz CPUs and took around 80 min in total to process the 52 589 proteins. To facilitate a comparison with the MATLAB version of the BRAIN algorithm, Suppl. Table S1 is added to the Supporting Information.

■ ASSOCIATED CONTENT

Supporting Information

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: p.dittwald@mimuw.edu.pl.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research is supported in part by the Polish National Science Center grant 2011/01/B/NZ2/00864 and by the EU through the European Social Fund, contract number UDA-POKL.04.01.01-00-072/09-00. J.C. gratefully acknowledges the support by the BOF09NI006 grant. Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is also gratefully acknowledged (J.C., T.B.). D.V., P.D., and A.G. gratefully acknowledge the support of the bilateral FWO grant VS.005.13N.

■ REFERENCES

(1) Gambin, A.; Dutkowski, J.; Karczmarski, J.; Kluge, B.; Kowalczyk, K.; Ostrowski, J.; Poznański, J.; Tiuryn, J.; Bakun, M.; Dadlez, M. *Int. J. Mass Spectrom.* **2007**, *260*, 20−30.
(2) Valkenborg, D.; Jansen, I.; Burzykowski, T. *J. Am. Soc. Mass Spectrom.* **2008**, *19*, 703−712.
(3) Valkenborg, D.; Thomas, G.; Krols, L.; Kas, K.; Burzykowski, T. *J. Mass Spectrom.* **2009**, *44*, 516−529.
(4) Valkenborg, D.; Mertens, I.; Lemière, F.; Witters, E.; Burzykowski, T. *Mass Spectrom. Rev.* **2012**, *31*, 96−106.
(5) Claesen, J.; Dittwald, P.; Burzykowski, T.; Valkenborg, D. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 753−763.
(6) Böcker, S. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 1826−1827.
(7) Claesen, J.; Dittwald, P.; Burzykowski, T.; Valkenborg, D. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 1828−1829.
(8) Yamamoto, H.; McCloskey, J. A. *Anal. Chem.* **1996**, *49*, 281−290.
(9) Brownawelld, M.; Fillippo, J. S. *J. Chem. Educ.* **1996**, *59*, 663−665.
(10) Rockwood, A. L.; Van Orden, S. L. *Anal. Chem.* **1996**, *68*, 2027−2030.
(11) Gentleman, R. C.; et al. *Genome Biol.* **2004**, *5*, R80.
(12) Yamamoto, H.; McCloskey, J. A. *Nucleic Acids Res.* **2012**, *40*, D71−D75.
(13) Jain, E.; Bairoch, A.; Duvaud, S.; Phan, I.; Redaschi, N.; Suzek, B. E.; Martin, M. J.; McGarvey, P.; Gasteiger, E. *BMC Bioinf.* **2009**, *10*, 136.