Fast and accurate computation of the isotopic distribution in two dimensions (2D)

Jorge Fernandez-de-Cossio
Bioinformatics Department. Center for Genetic Engineering and Biotechnology, P.O. Box 6162, CP 10600, C. Habana, CUBA.
Jorge.cossio@cigb.edu.cu

Supporting Information

Probe compounds

Various compounds are used in the proof of concept and performance evaluation of the algorithm described in this manuscript.

Tribromo-trichloro-ethane ($C_2Br_3Cl_3$) and the related hypothetical compound $(C_2Br_3Cl_3)_5$ are peculiar in that Br and Cl provide each two isotopes separated in mass by two nucleons, while the isotopes of C are separated in mass by one nucleon. The relative abundances of the isotopes of Br and Cl are comparable; while the abundance of ^{13}C, the heaviest isotope of C, is considerably lower than its lighter partner ^{12}C. The mass defect of the heaviest isotopes of Br and Cl are in opposite direction to those of C, introducing variability in the inter-peak distances, an issue addressed by the change of metric (see Result and Discussion).

Natural bio-molecules with a varying composition of Sulfur, both in absolute amounts and in relative proportions: Bovine ubiquitin ($C_{378}H_{629}N_{105}O_{118}S$), tumor suppression protein P16 ($C_{681}H_{1100}N_{216}O_{208}S_5$), Bovine insulin ($C_{254}H_{378}N_{65}O_{75}S_6$), Epidermal Grow Factor ($C_{270}H_{401}N_{73}O_{83}S_7$), Human interferon α-2 ($C_{964}H_{1531}N_{251}O_{283}S_{12}$) , Human Substance-Preceptor ($C_{2135}H_{3243}N_{517}O_{579}S_{26}$), Human Plasminogen ($C_{3948}H_{6073}N_{1123}O_{1213}S_{59}$). The content of Sulfur is relevant since it contributes with four stable isotopes, one of them (^{36}S) increases by two the number of nucleons of its closest partner (^{34}S), the latter being more abundant than the heavy isotopes of the other elements (C, H, O, and N).

The metallic clusters Sn_{10} and Sn_{100} of the element Tin (Stannum) are challenging compounds for isotopic distribution calculations[13]. Stannum has 10 natural stable isotopes; seven of them are more abundant than its lightest isotope.

Accuracy in the isotopic distribution calculation of small and large compounds

Table S- 1: Enumeration of the isotopic species of $C_2Br_3Cl_3$. Masses and abundances were accurately calculated from the isotope composition.

mass	abundance	isotopic species	mass	abundance	isotopic species
365.66157	0.054883169	12C2 35Cl3 79Br3	372.65788	0.0033904	12C 13C 35Cl2 37Cl 79Br 81Br2
366.66492	0.001228715	12C 13C 35Cl3 79Br3	372.65878	0.0011311	12C 13C 35Cl3 81Br3
367.65862	0.053345577	12C2 35Cl2 37Cl 79Br3	373.65067	0.0054474	12C2 37Cl3 79Br2 81Br
367.65952	0.16016704	12C2 35Cl3 79Br2 81Br	373.65157	0.0490663	12C2 35Cl 37Cl2 79Br 81Br2
367.66828	6.88E-06	13C2 35Cl3 79Br3	373.65248	0.0491062	12C2 35Cl2 37Cl 81Br3
368.66197	0.001194292	12C 13C 35Cl2 37Cl 79Br3	373.65943	2.34E-07	13C2 37Cl3 79Br3
368.66288	0.003585793	12C 13C 35Cl3 79Br2 81Br	373.66033	6.32E-06	13C2 35Cl 37Cl2 79Br2 81Br
369.65567	0.017283687	12C2 35Cl 37Cl2 79Br3	373.66123	1.898E-05	13C2 35Cl2 37Cl 79Br 81Br2
369.65657	0.15567984	12C2 35Cl2 37Cl 79Br2 81Br	373.66214	6.33E-06	13C2 35Cl3 81Br3
369.65747	0.1558066	12C2 35Cl3 79Br 81Br2	374.65403	0.000122	12C 13C 37Cl3 79Br2 81Br
369.66533	6.68E-06	13C2 35Cl2 37Cl 79Br3	374.65493	0.0010985	12C 13C 35Cl 37Cl2 79Br 81Br2
369.66623	2.00695E-05	13C2 35Cl3 79Br2 81Br	374.65583	0.0010994	12C 13C 35Cl2 37Cl 81Br3
370.65902	0.000386944	12C 13C 35Cl 37Cl2 79Br3	375.64862	0.0052991	12C2 37Cl3 79Br 81Br2
370.65993	0.003485334	12C 13C 35Cl2 37Cl 79Br2 81Br	375.64953	0.0159102	12C2 35Cl 37Cl2 81Br3
370.66083	0.003488172	12C 13C 35Cl3 79Br 81Br2	375.65738	6.83E-07	13C2 37Cl3 79Br2 81Br
371.65272	0.001866608	12C2 37Cl3 79Br3	375.65828	6.15E-06	13C2 35Cl 37Cl2 79Br 81Br2
371.65362	0.050439453	12C2 35Cl 37Cl2 79Br2 81Br	375.65919	6.15E-06	13C2 35Cl2 37Cl 81Br3
371.65452	0.15144157	12C2 35Cl2 37Cl 79Br 81Br2	376.65198	0.0001186	12C 13C 37Cl3 79Br 81Br2
371.65543	0.050521625	12C2 35Cl3 81Br3	376.65288	0.0003562	12C 13C 35Cl 37Cl2 81Br3
371.66238	2.17E-06	13C2 35Cl 37Cl2 79Br3	377.64658	0.0017183	12C2 37Cl3 81Br3
371.66328	1.95073E-05	13C2 35Cl2 37Cl 79Br2 81Br	377.65534	6.64E-07	13C2 37Cl3 79Br 81Br2
371.66418	1.95232E-05	13C2 35Cl3 79Br 81Br2	377.65624	1.99E-06	13C2 35Cl 37Cl2 81Br3
372.65607	4.17893E-05	12C 13C 37Cl3 79Br3	378.64993	3.847E-05	12C 13C 37Cl3 81Br3
372.65698	0.00112923	12C 13C 35Cl 37Cl2 79Br2 81Br	379.65329	2.15E-07	13C2 37Cl3 81Br3

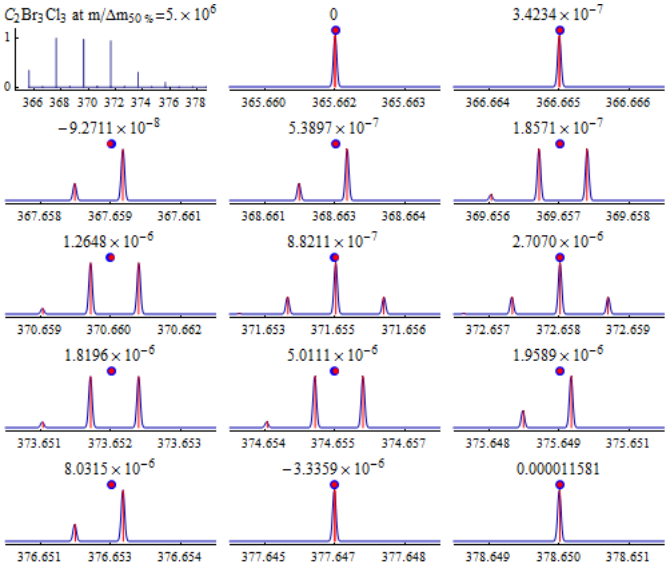

Figure S- 1: Accuracy evaluation of the isotopic distribution of $C_2Br_3Cl_3$, calculated by DEUTERIUM at resolving power 5x10^6. Continue blue profiles correspond to the calculated isotopic distribution and the blue dots locate their individual peak centroids. The red vertical lines locate the accurate position of each isotopic species and the red dots locate their individual peak centroids. The relative abundances are normalized to maximum 1. A) The whole profile. B) Individual peaks details.

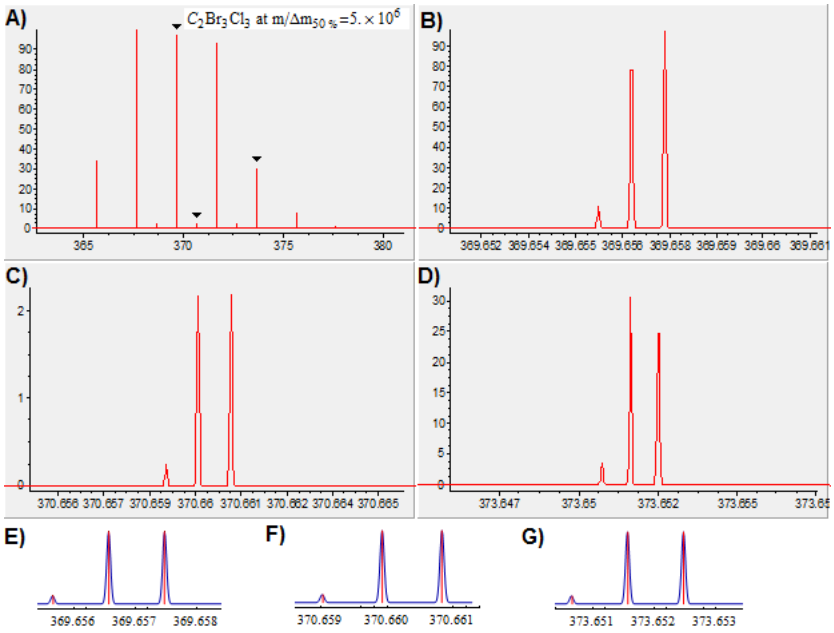

Figure S- 2: Isotopic distribution of $C_2Br_3Cl_3$ at resolving power 5x10^6. A) The whole profile as calculated by IsoPro. B-D) Zoom-in of some individual peaks as calculated by IsoPro. E-G) Zoom-in of the same chosen peaks as calculated by DEUTERIUM (blue profiles). The exact mass and abundances of the isotopic species (Table S1) are represented with red bars.

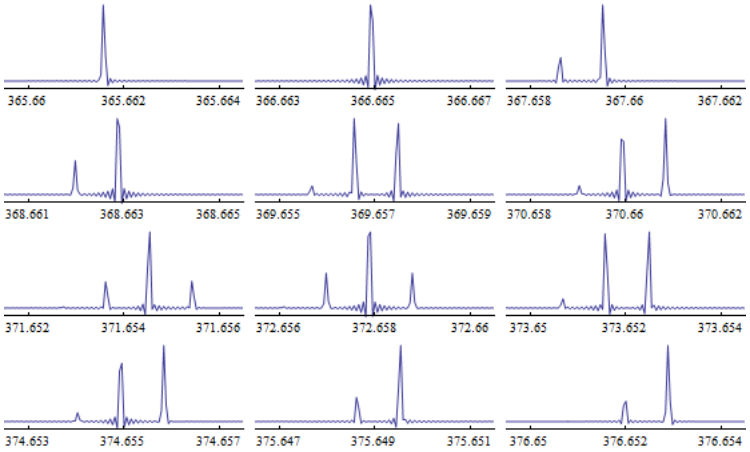

Figure S- 3: Isotopic distribution of $C_2Br_3Cl_3$ calculated with the ultrahigh option of Mercury. Apodization subscript parameter was set to 2 and high resolution range was limited to 0.1 Da about each peak centroid.

Enumeration of the species is obviously impractical for larger compounds like Human plasminogen (90.5 *k*Da). Though highly overlapped, the fine structure of the isotopic species remains exposed in peaks of lower masses at resolving power $3x10^8$ (Figure S-4 B-K). The detailed structures of the peak at 90 567 *Da* accurately match with those obtained with the ultrahigh option of Mercury (not shown) and with the packing procedure[7] using a spacer length of 0.85 (Figure S-4K).

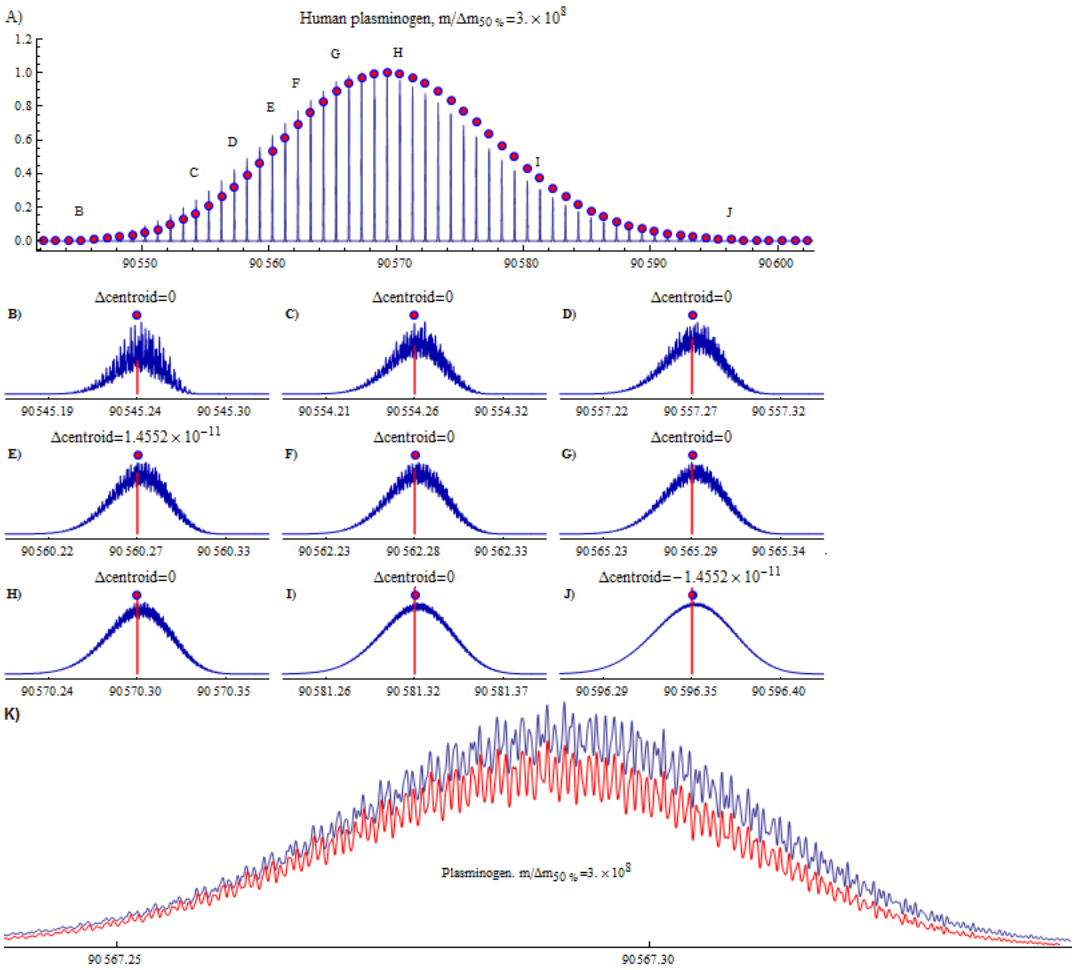

Figure S- 4: Isotopic distribution of Human plasminogen at resolving power 3x10^8. Continuous blue profiles correspond to the isotopic distribution calculated by DEUTERIUM and the blue dots locate their individual peak centroids. Red vertical lines indicate the position and relative abundance of the centroids calculated by accurate polynomial method. Red dots also locate their individual peak centroids. The relative abundances are normalized to maximum 1. A) The whole profile. B-J) Details of some chosen individual peaks. K) Zoom-in of the isotope peak at 90 567 Da. The profile calculated by the packing procedure is in red. The profile calculated by DEUTERIUM is in blue. The dynamic range axis of the packed profile was slightly reduced so as to clearly visualize both profiles. They accurately match.

Truncation arising with polynomial methods

At resolving power 1x10^7 and default parameters, the isotopic distribution of Substance P calculated by IsoPro appears truncated ~10 *Da* at the higher tail (Figure S5). A more drastic truncation takes place with the cluster compound Sn_{100} (Figure S7A-B), confining the whole isotopic distribution to a mass range of length < 40 *Da*. Since the theoretical standard deviation is > 20 *Da*, more than 30% of the masses have been swept away by IsoPro (and IsoDalton). DEUTERIUM produces a smooth bell shaped distribution with standard deviation differing from the theoretical one by ~0.3 Da (Figure S7C).

Truncation of the mass range is a sensible issue of current implementations of polynomial methods. Pruning techniques introduced to take over the growing number of terms produced at each polynomial expansion are in need for improvements. Any advance in this direction shall be desirable in a parameter-less manner, so as to be amenable for automation. While the nature of the polynomial-truncation issue

is absent by construction in Fourier transform methods[13], folding-over can arise by aliasing if the sampling range falls too short, but this is avoided with the Rockwood simple rule $10(1+\sigma^2)^{1/2}$.

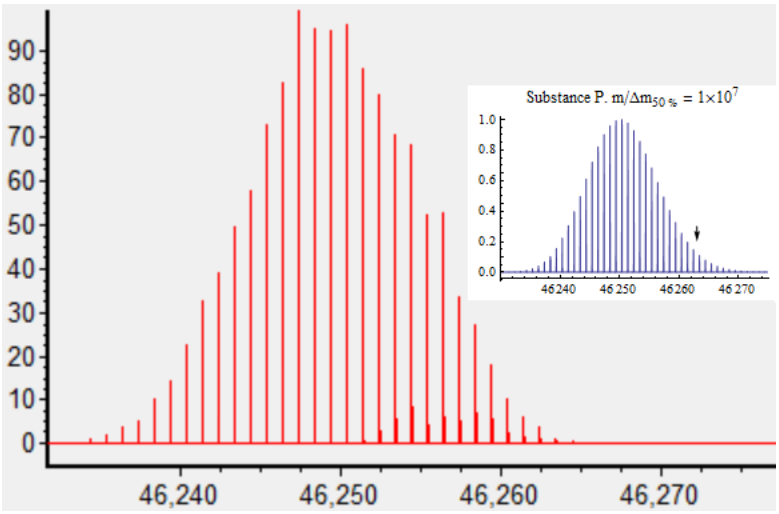

Figure S- 5: Isotopic distribution of Substance P at resolving power 1x10^7 as calculated by IsoPro with default parameters. In the inset is the isotopic distribution as calculated by DEUTERIUM. IsoPro distribution terminates ~10 Da before DEUTERIUM distribution.

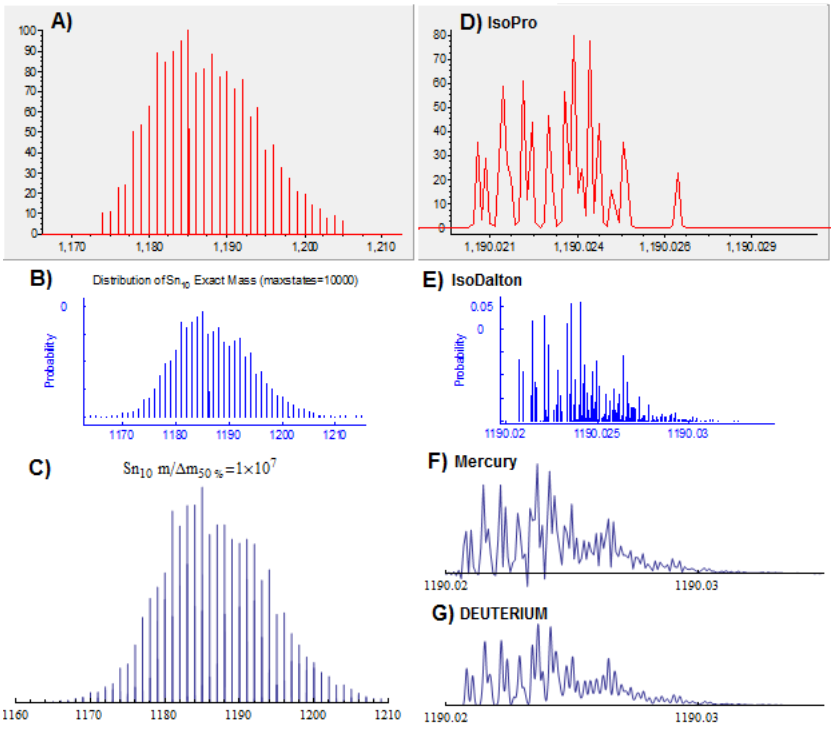

Figure S- 6: Isotopic distribution of the metallic cluster Sn_{10} at resolving power 1x10^7. A) Calculated by IsoPro with default parameters. B) Calculated by IsoDalton with a maximum of 10000 states. C) Calculated by DEUTERIUM. D-G) Zoom-in of peak 1190 as calculated respectively by IsoPro, IsoDalton, Mercury, and DEUTERIUM. Details of peak 1190 are consistent, except for the high mass tail, which appears less populated in the IsoPro Zoom-in.

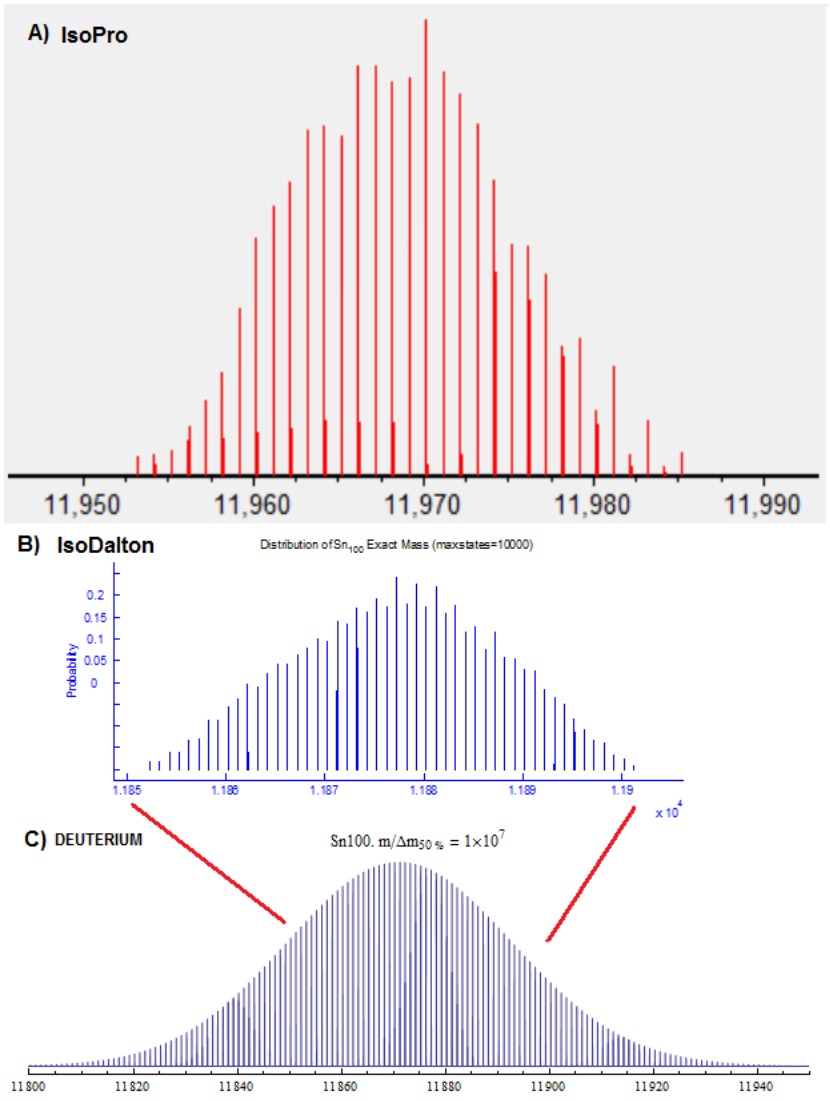

Figure S- 7: Isotopic distribution of the metallic cluster Sn_{100} at resolving power $1x10^7$. A) Calculated by IsoPro with default parameters. B) Calculated by IsoDalton with 10 000 maximum states. C) Calculated by DEUTERIUM. Both IsoPro and IsoDalton calculated distributions appear drastically truncated (confined to a range of length ~40 Da)[1].

[1] A large mass error shift (~150 Da) is observed in the IsoPro calculated distribution. It is certainly not attributable to polynomial methods performance, but a software bug.

Nomenclature for resolution ranges

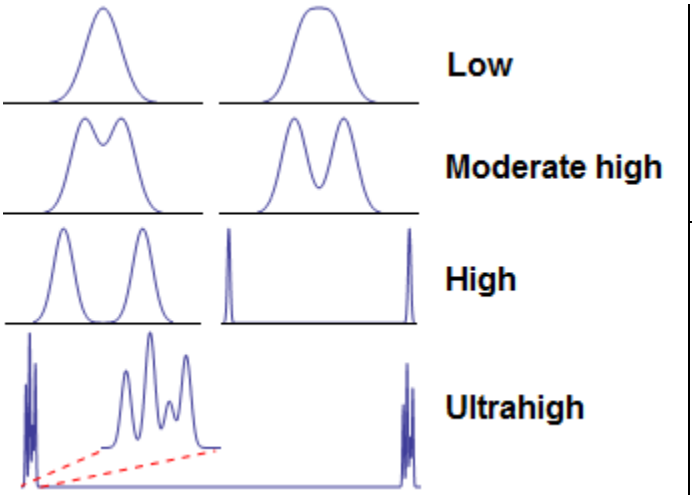

Figure S- 8: Sketchy representation of the nomenclature adopted within this manuscript for the different level of resolutions. The two (mixed) peaks are ~1*Da* apart.

The 2D approach introduced in this manuscript exhibit superior performance for high and ultrahigh resolutions levels. The original formulation of the method for the calculation of the isotopic distribution[13] is more efficient for low and moderate high resolutions levels.

Inter-peak distances variability and optimization

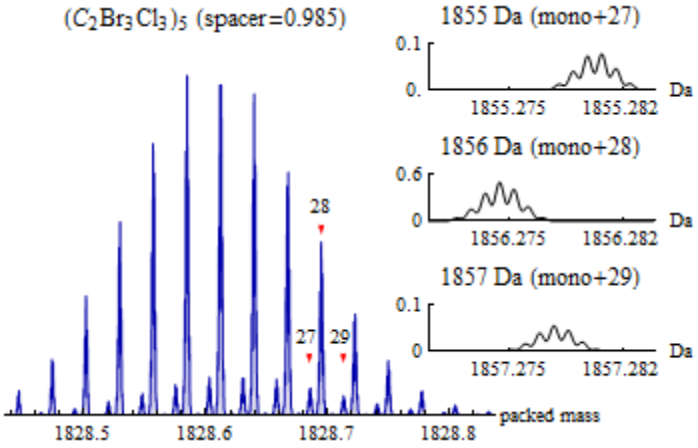

Figure S- 9: Variability of consecutive inter-peak distances of the compound $(C_2Br_3Cl_3)_5$ is exposed in the packed isotopic distribution at resolving power $3x10^6$. Consecutive distances between peaks 27, 28 and 29 are ~0.99 Da and ~1.005 Da respectively, as is apparent from the inset (natural scale).

The change of metrics appreciably reduces the range spanned by the mass defect. However, a "zigzag" arrangement of the consecutive peaks persists in the metrics-adjusted green profile of Figure 1, widening the required range. This is entirely due to the variability of the inter-peak distance, as is appreciated by the relative offset between odd and even peaks, visualized in Figure S- 9. The peaks in the compound $C_2Br_3Cl_3$ with odd number of nucleons (odd peaks) necessarily contains an odd number of ^{13}C, whereas the contribution of multiple ^{13}C to peaks with an even number of nucleons (even peaks) is depreciable (^{13}C is more than 24 times less abundant than the isotopes of Br and Cl). The mass defect

per nuclei of ^{81}Br (= -0.0010235) and ^{37}Cl (= -0.00147485) are in opposite direction to the mass defect of ^{13}C (=0.0033554), and the latter is more than 2 times larger in magnitude. Hence, the distance from odd peaks to the next (one Da heavier) is smaller than the distance from even peaks to the next.

The mass defect range could be further reduced (optimally) by straightening the “zigzag” in a recoverable manner, spanning only a length covering the maximum peak-spread. This can be achieved by “heterodyning” the Fourier transform along the w dimension, so that the mass defect averages to zero at each isotopic peak centroid. The change of metric then becomes superfluous by this more optimal procedure, but require knowing a priori the centroids, obtainable by accurate calculations[14].