S-1. Derivation of Lemma 1

For a polypeptide $C_{n_C}H_{n_H}N_{n_N}O_{n_O}S_{n_S}$, we can compute I_k by the coefficient of x^k in the expansion of the following polynomial[1, 2, 3].

$$\begin{aligned} P(x) = &(P_{C_0} + P_{C_1}x)^{n_C}(P_{H_0} + P_{H_1}x)^{n_H}(P_{N_0} + P_{N_1}x)^{n_N} \\ &(P_{O_0} + P_{O_1}x + P_{O_2}x^2)^{n_O}(P_{S_0} + P_{S_1}x + P_{S_2}x^2 + P_{S_4}x^4)^{n_S} \end{aligned} \quad (1)$$

That is, intensity I_k in an isotopic distribution of a polypeptide is regarded as the sum of existential probabilities of all polypeptide instances with mass difference k. Intensity I_0 is the probability of there being no isotopes in $C_{n_C}H_{n_H}N_{n_N}O_{n_O}S_{n_S}$, which is the constant term of polynomial $P(x)$, defined as follows.

$$I_0 = {P_{C_0}}^{n_C}{P_{H_0}}^{n_H}{P_{N_0}}^{n_N}{P_{O_0}}^{n_O}{P_{S_0}}^{n_S} \quad (2)$$

Intensity I_1 is the probability of there being only one +1 isotope, which is the coefficient of x in $P(x)$, and I_2 is the probability of there being two +1 isotopes or one +2 isotope, which is the coefficient of x^2, and they are defined as follows.

$$\begin{aligned} I_1 &= n_C{P_{C_0}}^{n_C-1}P_{C_1}{P_{H_0}}^{n_H}{P_{N_0}}^{n_N}{P_{O_0}}^{n_O}{P_{S_0}}^{n_S} + n_H{P_{C_0}}^{n_C}{P_{H_0}}^{n_H-1}P_{H_1}{P_{N_0}}^{n_N}{P_{O_0}}^{n_O}{P_{S_0}}^{n_S} + \ldots \\ &= I_0\sum_{X\in A} n_X\frac{P_{X_1}}{P_{X_0}} \end{aligned} \quad (3)$$

$$
\begin{aligned}
I_2 = & \binom{n_{\mathrm{C}}}{2} {P_{\mathrm{C}_0}}^{n_{\mathrm{C}}-2} {P_{\mathrm{C}_1}}^2 {P_{\mathrm{H}_0}}^{n_{\mathrm{H}}} {P_{\mathrm{N}_0}}^{n_{\mathrm{N}}} {P_{\mathrm{O}_0}}^{n_{\mathrm{O}}} {P_{\mathrm{S}_0}}^{n_{\mathrm{S}}} + \binom{n_{\mathrm{H}}}{2} {P_{\mathrm{C}_0}}^{n_{\mathrm{C}}} {P_{\mathrm{H}_0}}^{n_{\mathrm{H}}-2} {P_{\mathrm{H}_1}}^2 {P_{\mathrm{N}_0}}^{n_{\mathrm{N}}} {P_{\mathrm{O}_0}}^{n_{\mathrm{O}}} {P_{\mathrm{S}_0}}^{n_{\mathrm{S}}} \\
& + \ldots + \binom{n_{\mathrm{S}}}{2} {P_{\mathrm{C}_0}}^{n_{\mathrm{C}}} {P_{\mathrm{H}_0}}^{n_{\mathrm{H}}} {P_{\mathrm{N}_0}}^{n_{\mathrm{N}}} {P_{\mathrm{O}_0}}^{n_{\mathrm{O}}} {P_{\mathrm{S}_0}}^{n_{\mathrm{S}}-2} {P_{\mathrm{S}_1}}^2 \\
& + n_{\mathrm{C}} n_{\mathrm{H}} {P_{\mathrm{C}_0}}^{n_{\mathrm{C}}-1} P_{\mathrm{C}_1} {P_{\mathrm{H}_0}}^{n_{\mathrm{H}}-1} P_{\mathrm{H}_1} {P_{\mathrm{N}_0}}^{n_{\mathrm{N}}} {P_{\mathrm{O}_0}}^{n_{\mathrm{O}}} {P_{\mathrm{S}_0}}^{n_{\mathrm{S}}} \\
& + n_{\mathrm{C}} n_{\mathrm{N}} {P_{\mathrm{C}_0}}^{n_{\mathrm{C}}-1} P_{\mathrm{C}_1} {P_{\mathrm{H}_0}}^{n_{\mathrm{H}}} {P_{\mathrm{N}_0}}^{n_{\mathrm{N}}-1} P_{\mathrm{N}_1} {P_{\mathrm{O}_0}}^{n_{\mathrm{O}}} {P_{\mathrm{S}_0}}^{n_{\mathrm{S}}} \\
& + \ldots \\
& + n_{\mathrm{O}} {P_{\mathrm{C}_0}}^{n_{\mathrm{C}}} {P_{\mathrm{H}_0}}^{n_{\mathrm{H}}} {P_{\mathrm{N}_0}}^{n_{\mathrm{N}}} {P_{\mathrm{O}_0}}^{n_{\mathrm{O}}-1} P_{\mathrm{O}_2} {P_{\mathrm{S}_0}}^{n_{\mathrm{S}}} + n_{\mathrm{S}} {P_{\mathrm{C}_0}}^{n_{\mathrm{C}}} {P_{\mathrm{H}_0}}^{n_{\mathrm{H}}} {P_{\mathrm{N}_0}}^{n_{\mathrm{N}}} {P_{\mathrm{O}_0}}^{n_{\mathrm{O}}} {P_{\mathrm{S}_0}}^{n_{\mathrm{S}}-1} P_{\mathrm{S}_2} \\
= & I_0 \left(\sum_{X \in A} \binom{n_X}{2} \frac{{P_{X_1}}^2}{{P_{X_0}}^2} + \sum_{\substack{X,Y \in A \\ X \neq Y}} n_X n_Y \frac{P_{X_1} P_{Y_1}}{P_{X_0} P_{Y_0}} + \sum_{X \in A} n_X \frac{P_{X_2}}{P_{X_0}} \right)
\end{aligned}
\tag{4}
$$

Now consider intensity I_k for an arbitrary $k \geq 1$. The instances with a mass difference $k = k_1 + 2k_2 + 4k_4$ consist of all the instances with k_1 isotopes of +1Da, k_2 isotopes of +2Da, and k_4 isotopes of +4Da. For a polypeptide instance, let t_{X_1}, t_{X_2}, and t_{X_4} be the number of +1, +2, and +4 isotopes of atom X, respectively. Then, the probability of all instances with given k_1, k_2, and k_4 is the sum of the probabilities of there being t_{X_1}, t_{X_2}, and t_{X_4} isotopes for each atom X such that the sum of t_{X_1} for all atoms X is k_1, that of t_{X_2} is k_2, and that of t_{X_4} is k_4. I_k is the probability sum of all combinations of k_1, k_2, and k_4 such that $k = k_1 + 2k_2 + 4k_4$ as follows.

$$
\begin{aligned}
I_k = & \sum_{k_1+2k_2+4k_4=k} \; \sum_{t_{\mathrm{C}_1}+t_{\mathrm{H}_1}\ldots+t_{\mathrm{S}_1}=k_1} \; \sum_{t_{\mathrm{C}_2}+t_{\mathrm{H}_2}\ldots+t_{\mathrm{S}_2}=k_2} \; \sum_{t_{\mathrm{C}_4}+t_{\mathrm{H}_4}\ldots+t_{\mathrm{S}_4}=k_4} \\
& \prod_X \binom{n_X}{t_{X_1} \quad t_{X_2} \quad t_{X_4}} P_{X_1}^{t_{X_1}} P_{X_2}^{t_{X_2}} P_{X_4}^{t_{X_4}} P_{X_0}^{n_X - t_{X1} - t_{X_2} - t_{X_4}}
\end{aligned}
\tag{5}
$$

Since n_X (number of atom X) is much larger than t_{X_1}, t_{X_2}, and t_{X_4} in practice, we employ the following approximation.

$$\binom{n_X}{t_{X_1}\quad t_{X_2}\quad t_{X_4}} = \frac{n_X!}{(n_X - t_{X1} - t_{X_2} - t_{X_4})! t_{X_1}! t_{X_2}! t_{X_4}!} \approx \frac{{n_X}^{t_{X1}+t_{X_2}+t_{X_4}}}{t_{X_1}! t_{X_2}! t_{X_4}!} \tag{6}$$

Then, we obtain the approximation of I_k in Lemma 1 by algebraic manipulations.

$$I_k \approx \sum_{k_1+2k_2+4k_4=k} \sum_{t_{X_1}} \sum_{t_{X_2}} \sum_{t_{X_4}} \prod_X \frac{{n_X}^{t_{X1}+t_{X_2}+t_{X_4}}}{t_{X_1}! t_{X_2}! t_{X_4}!} {P_{X_1}}^{t_{X_1}} {P_{X_2}}^{t_{X_2}} {P_{X_4}}^{t_{X_4}} {P_{X_0}}^{n_X - t_{X1} - t_{X_2} - t_{X_4}}$$

$$= \sum_{k_1+2k_2+4k_4=k} \sum_{t_{X_1}} \sum_{t_{X_2}} \sum_{t_{X_4}} \prod_X {P_{X_0}}^{n_X} \frac{1}{t_{X_1}!} \left(\frac{n_X P_{X_1}}{P_{X_0}} \right)^{t_{X_1}} \frac{1}{t_{X_2}!} \left(\frac{n_X P_{X_2}}{P_{X_0}} \right)^{t_{X_2}} \frac{1}{t_{X_4}!} \left(\frac{n_X P_{X_4}}{P_{X_0}} \right)^{t_{X_4}}$$

$$= \sum_{k_1+2k_2+4k_4=k} \prod_X {P_{X_0}}^{n_X} \left(\sum_{t_{X_1}} \prod_X \frac{1}{t_{X_1}!} \left(\frac{n_X P_{X_1}}{P_{X_0}} \right)^{t_{X_1}} \right) \left(\sum_{t_{X_2}} \prod_X \frac{1}{t_{X_2}!} \left(\frac{n_X P_{X_2}}{P_{X_0}} \right)^{t_{X_2}} \right) \left(\sum_{t_{X_4}} \prod_X \frac{1}{t_{X_4}!} \left(\frac{n_X P_{X_4}}{P_{X_0}} \right)^{t_{X_4}} \right)$$

$$= \sum_{k_1+2k_2+4k_4=k} \frac{\prod_X {P_{X_0}}^{n_X}}{k_1! k_2! k_4!} \left(\sum_{t_{X_1}} k_1! \prod_X \frac{1}{t_{X_1}!} \left(\frac{n_X P_{X_1}}{P_{X_0}} \right)^{t_{X_1}} \right) \left(\sum_{t_{X_2}} k_2! \prod_X \frac{1}{t_{X_2}!} \left(\frac{n_X P_{X_2}}{P_{X_0}} \right)^{t_{X_2}} \right) \left(\sum_{t_{X_4}} k_4! \prod_X \frac{1}{t_{X_4}!} \left(\frac{n_X P_{X_4}}{P_{X_0}} \right)^{t_{X_4}} \right)$$

$$= \sum_{k_1+2k_2+4k_4=k} \frac{I_0}{k_1! k_2! k_4!} \left(\sum_X \frac{n_X P_{X_1}}{P_{X_0}} \right)^{k_1} \left(\sum_X \frac{n_X P_{X_2}}{P_{X_0}} \right)^{k_2} \left(\sum_X \frac{n_X P_{X_4}}{P_{X_0}} \right)^{k_4} \tag{7}$$

S-2. Methods for Improving Accuracy

Some more techniques were implemented to improve the accuracy of our method. We apply different techniques according to the mass, because the number of peaks of a pseudo cluster becomes larger as the mass increases.

When a peptide mass is less than 4000 Da, all the peaks of a pseudo cluster are considered important because there are only a few peaks in the cluster. In such cases, the number of terms summed in score calculation is also small and the effect of missing peaks (often excluded during the peak picking step) on the score can be significant. To overcome such difficulties, we additionally select some of the missing peaks from the raw spectrum. Assume that p peaks ($p \geq 1$) are missing in a pseudo cluster (i.e., the first peak of the pseudo cluster is the (p+1)-st peak of the isotopic distribution). First, we select I'_{-1} which is the intensity of a (noise) peak at the position of the p-th peak of the isotopic distribution. I'_{-1} is expected to be larger than $I'_0/R_{max}(p\text{-}1, m)$ because $R_{max}(p\text{-}1, m)$ represents the upper bound of $I_p/I_{p\text{-}1}$ which corresponds to I'_0/I'_{-1}. Therefore, if I'_{-1} is smaller than $I'_0/R_{max}(p\text{-}1, m)$, the program calculate $scoreR(-1, p, m)$ and add it to the score of the cluster ($scoreR(-1, p, m)$ is always negative). Second, we select I'_n which is the intensity of a (noise) peak at the position of the (n+p+1)-st peak of the isotopic

distribution. The program calculates *scoreR*(n-1, p, m) and add it to the score of the cluster if I'_n is smaller than I'_{n-1}*R_{min}(n+p-1, m).

If a peptide mass is greater than 4000 Da, we give more weight to the ratio and ratio product terms computed from high intensity peaks on the ground that high intensity peaks are more reliable than low intensity peaks. Note that a pseudo cluster contains few peaks in the case of m < 4000, so the use of weight may cause loss of information. Specifically, for each ratio and ratio product term, we use the smallest intensity in each ratio and ratio product term as its weight. In addition, the total score is normalized by $I_{\max}$, which is the intensity of the highest peak in the cluster. Hence, the weighted score function is as follows.

$$\frac{1}{I_{\max}}\left(\sum_{k=0}^{n-2}\min(I'_k, I'_{k+1})\,scoreR(k,p,m)+\sum_{k=0}^{n-3}\min(I'_k, I'_{k+1}, I'_{k+2})\,scoreRP(k,p,m)\right)$$

Another technique for $m \geq 4000$ is an introduction of so called "*bias*". Intuitively speaking, the bias is a trend in the ratio values that indicates whether a majority of them is larger than or smaller than the average values R_{avg}. The bias is defined as follows.

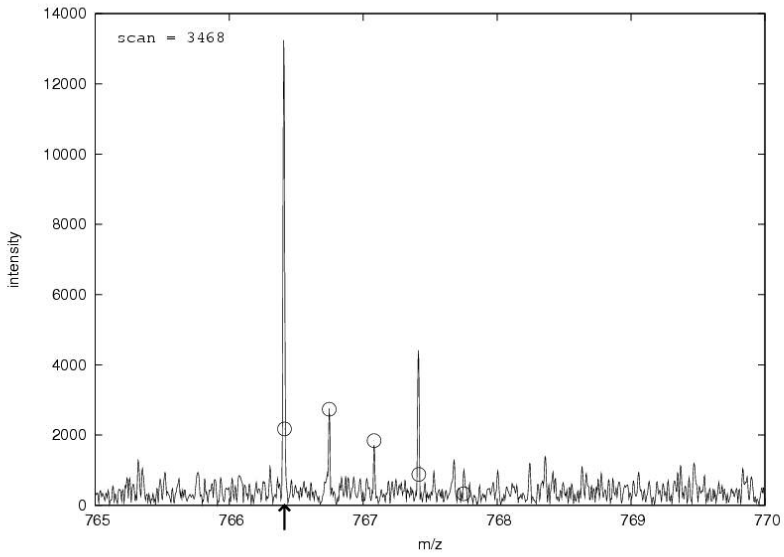
$$bias_{-1} = 0$$

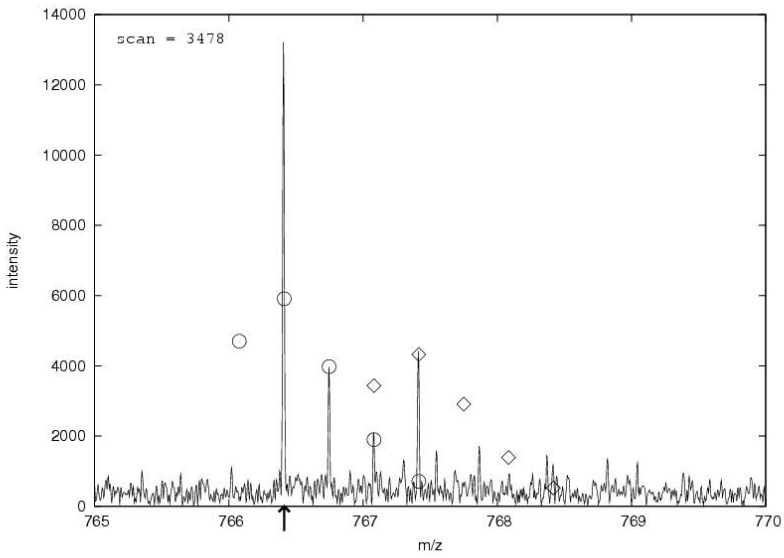
$$bias_k = \begin{cases} bias_{k-1} + \dfrac{I'_{k+1}/I'_k - R_{avg}(k+p,m)}{R_{\max}(k+p,m) - R_{avg}(k+p,m)} & \text{if} \quad I'_{k+1}/I'_k > R_{avg}(k+p,m) \\ bias_{k-1} - \dfrac{R_{avg}(k+p,m) - I'_{k+1}/I'_k}{R_{avg}(k+p,m) - R_{\min}(k+p,m)} & \text{otherwise} \end{cases}$$

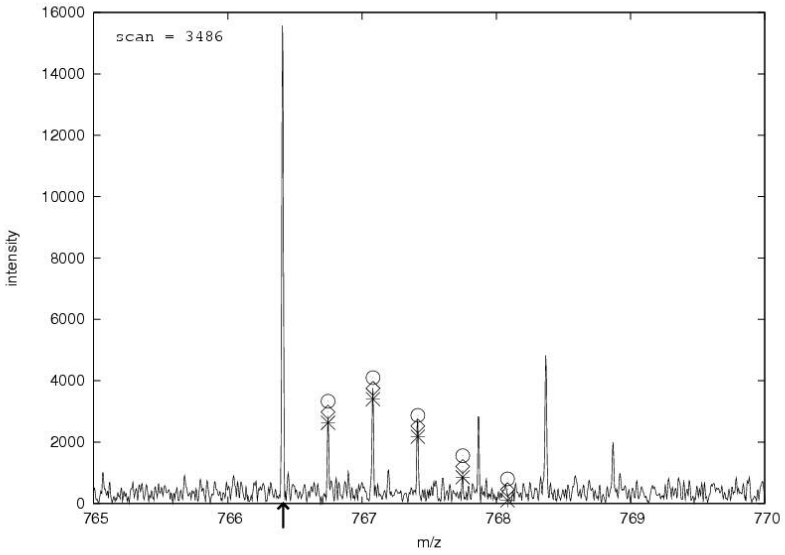
The value $bias_k$ represents an accumulation of biases of I'_1/I'_0 to I'_{k+1}/I'_k. A large bias value means that there are many ratios which are larger than $R_{avg}(k+p, m)$. When the mass of a pseudo cluster is determined to be 1 Da larger than the correct mass, in general each ratio I'_{k+1}/I'_k is smaller than $R_{avg}(k+p, m)$ because $I_{k+1}/I_k \geq I_{k+2}/I_{k+1}$. On the contrary, most of ratios are larger than average when the mass of a cluster is determined to be 1 Da smaller than the correct mass. Therefore the absolute values of biases get large when a monoisotopic mass is determined incorrectly. And if the bias values are close to 0, this means that the shape of this cluster is similar to that of the isotopic distribution. The weighted score function with the bias included for $m \geq 4000$ is like the following:

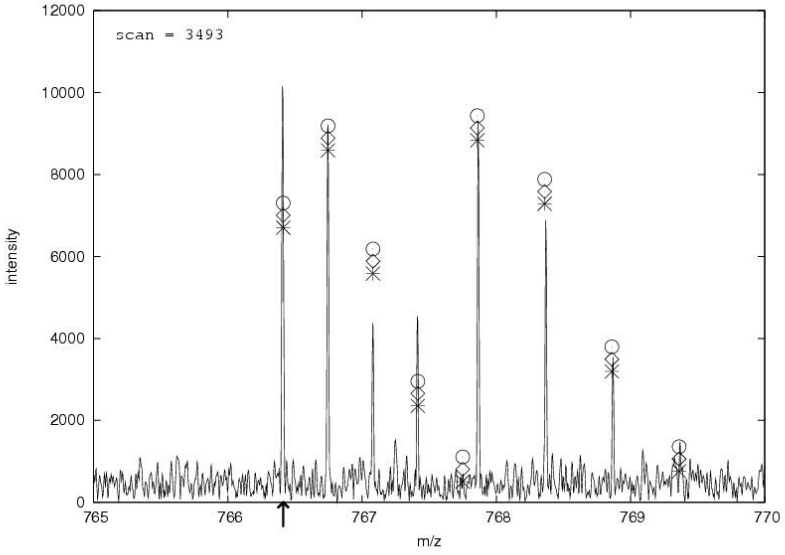
$$\frac{1}{I_{\max}}\left(\sum_{k=0}^{n-2}\min(I'_k, I'_{k+1})(1-bias_k)scoreR(k,p,m) + \sum_{k=0}^{n-3}\min(I'_k, I'_{k+1}, I'_{k+2})(1-bias_k)scoreRP(k,p,m)\right)$$

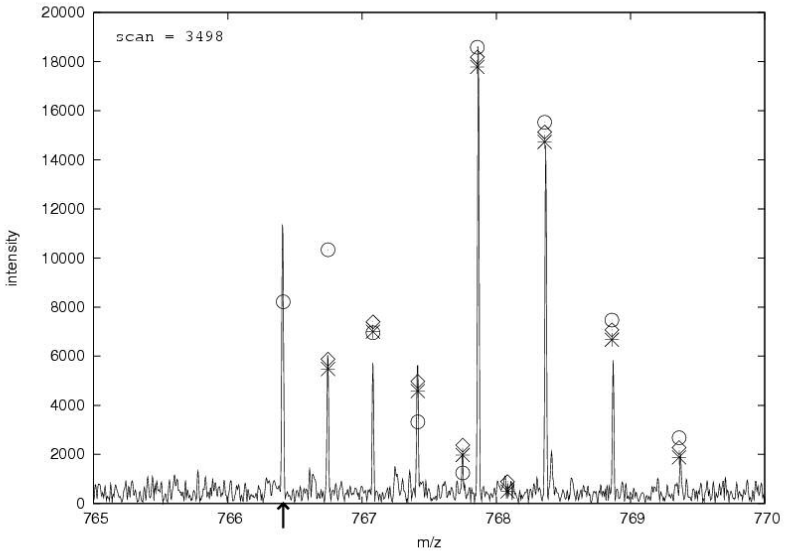
Figure S-1. Ten different scans of a peptide summarized in Table 4 ($C_{101}H_{165}N_{29}O_{32}$, 2296.22 Da). An arrow represents the monoisotopic peak of this peptide and circles, diamonds and stars represent the theoretical isotopic distributions of this peptide calculated by each of our method, Decon2LS and ICR2LS, respectively.

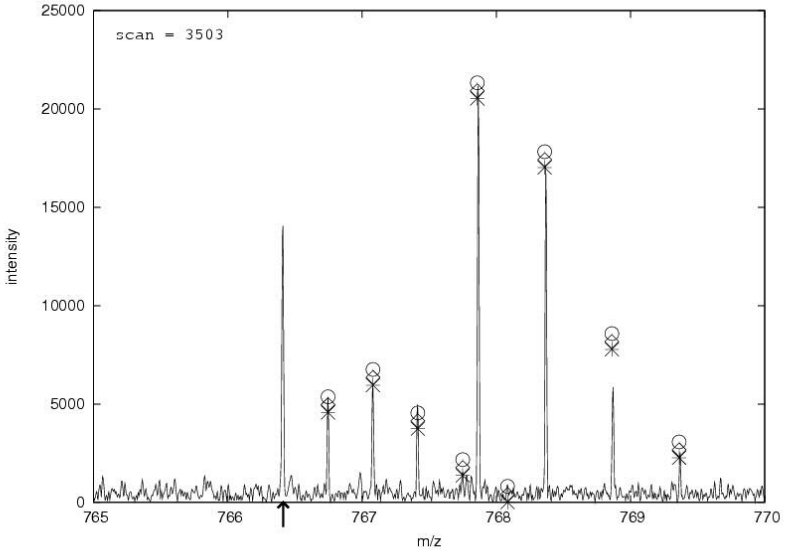


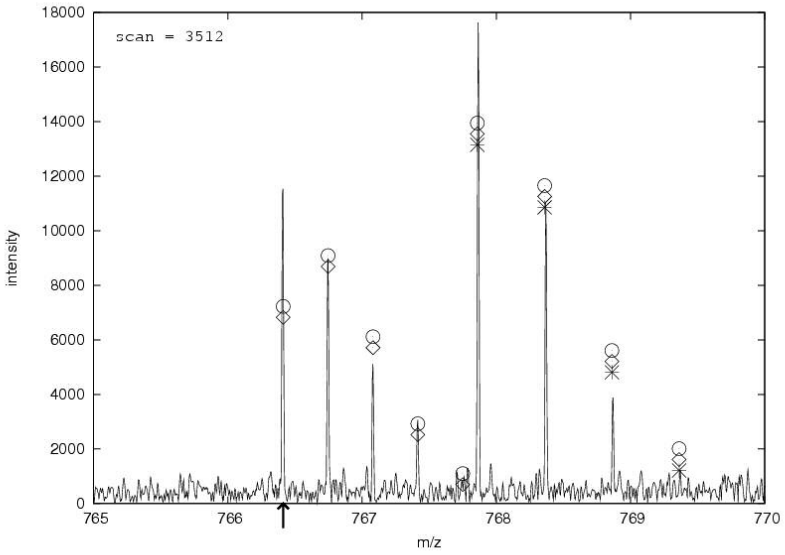


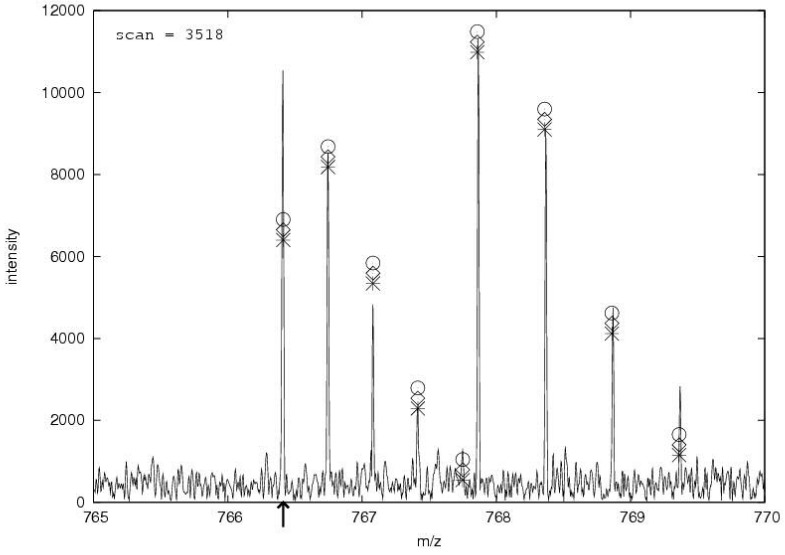

scan = 3486
intensity
m/z
16000
14000
12000
10000
8000
6000
4000
2000
0
765
766
767
768
769
770

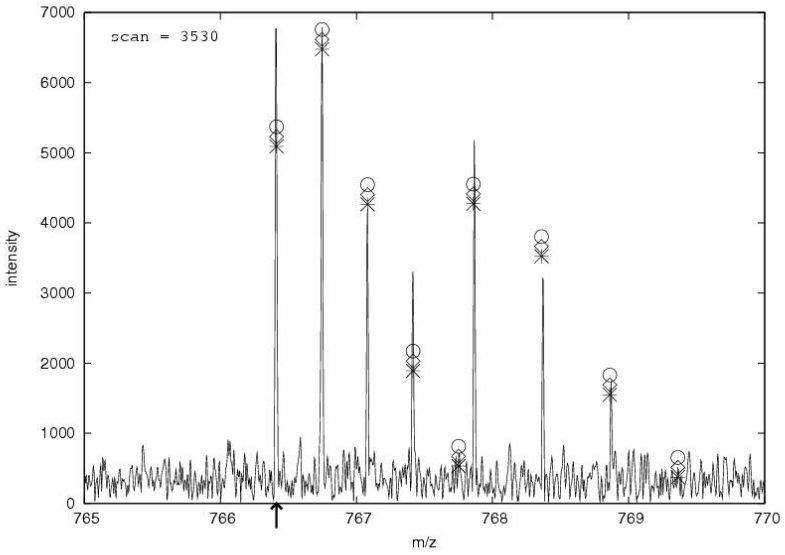

scan = 3493
intensity
m/z
12000
10000
8000
6000
4000
2000
0
765
766
767
768
769
770


scan = 3498
intensity
m/z
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
765
766
767
768
769
770


scan = 3503
intensity
m/z
0
5000
10000
15000
20000
25000
765
766
767
768
769
770


scan = 3512
intensity
m/z
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
765
766
767
768
769
770


scan = 3518
intensity
m/z
0
2000
4000
6000
8000
10000
12000
765
766
767
768
769
770


scan = 3530
intensity
m/z
0
1000
2000
3000
4000
5000
6000
7000
765
766
767
768
769
770

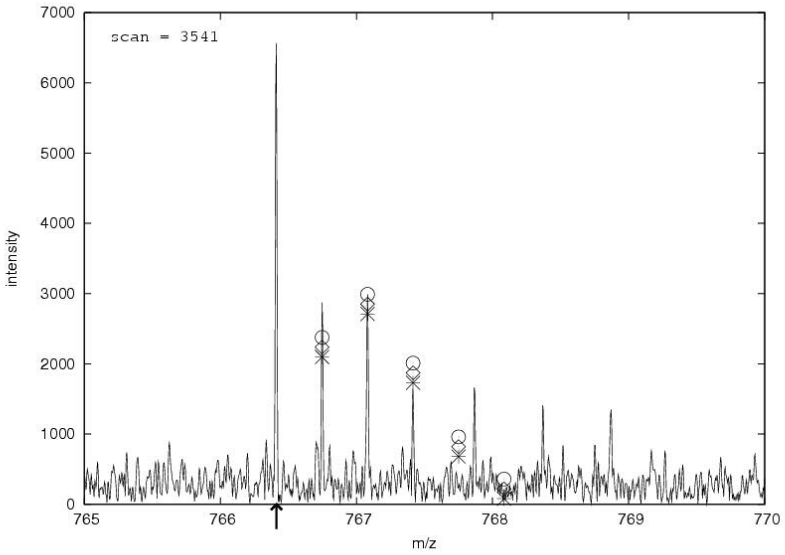

scan = 3541
intensity
m/z
0
1000
2000
3000
4000
5000
6000
7000
765
766
767
768
769
770

Table S-1. Computations of average I_{k+1} / I_k by the *averagine* model and by our fitting.

	Averagine	Our fitting result	
		$m < 1800$	$m \geq 1800$
I_1 / I_0	$5.43\times10^{-4} m$	$5.42\times10^{-4} m + 1.00\times10^{-30}$	$5.44\times10^{-4} m + 1.00\times10^{-30}$
I_2 / I_1	$2.71\times10^{-4} m + 8.17\times10^{-2}$	$2.79\times10^{-4} m + 5.94\times10^{-2}$	$2.74\times10^{-4} m + 6.75\times10^{-2}$
I_3 / I_2	$1.81\times10^{-4} m + 1.63\times10^{-1}$	$\frac{1.93\times10^{-4} m^2 + 1.28\times10^{-1} m}{m + 3.01\times10^{2}}$	$1.86\times10^{-4} m + 7.41\times10^{-2}$

Table S-2. Computations of average $I_k I_{k+2} / {I_{k+1}}^2$ by the *averagine* model and by our fitting.

	Averagine	Our fitting result	
		$m < 1800$	$m \geq 1800$
$I_0 I_2 / {I_1}^2$	$5.00\times10^{-1} + \frac{1.51\times10^{2}}{m}$	$5.12\times10^{-1} + \frac{1.14\times10^{2}}{m}$	$5.03\times10^{-1} + \frac{1.29\times10^{2}}{m}$
$I_1 I_3 / {I_2}^2$	$6.67\times10^{-1} + \frac{2.01\times10^{2}}{m + 6.02\times10^{2}}$	$7.30\times10^{-1} + \frac{-4.51\times10^{-1}}{m - 3.95\times10^{2}}$	$6.66\times10^{-1} + \frac{2.31\times10^{2}}{m + 2.06\times10^{3}}$
$I_2 I_4 / {I_3}^2$	$7.50\times10^{-1} + \frac{2.64\times10^{2}}{m + 1.81\times10^{3}}$	$7.56\times10^{-1} + \frac{1.50\times10^{2}}{m + 1.81\times10^{3}}$	$7.59\times10^{-1} + \frac{1.51\times10^{2}}{m + 1.81\times10^{3}}$

Table S-3. Average number of peaks and execution time of three programs. It shows that the number of peaks is a major factor in execution time.

Segment[1]	Number of scans	Average number of peaks			Time (s)		
		Our method	Decon2LS	ICR2LS	Our method	Decon2LS	ICR2LS
1	737	123	123	135	625	757	4877
2	921	164	164	178	760	1027	7391
3	1085	231	231	260	861	1650	12269
4	1205	586	586	683	993	4834	43993
5	1048	360	360	413	816	2759	21531

[1] A total of five segment LC/MS/MS data set was obtained from a whole LC gradient experiment (see experimental section for details).

Supplementary References

(1) Snider, R. K. *J. Am. Soc. Mass Spectrom.* **2007**, 18, 1511-1515.

(2) Yergey, J. A. *Int. J. Mass Spectrom. Ion Phys.* **1983**, 52, 337-349.

(3) Rockwood, A. L.; Van Orden, S. L.; Smith, R. D. *Anal. Chem.* **1995**, 67, 2699-2704.