Received: 31 December 2020 | Revised: 6 July 2021 | Accepted: 6 July 2021

DOI: 10.1002/rcm.9162

RESEARCH ARTICLE

Rapid Communications in Mass Spectrometry WILEY

Predicting the number of sulfur atoms in peptides and small proteins based on the observed aggregated isotope distribution

Jürgen Claesen[1,2,3] | Dirk Valkenborg[3] | Tomasz Burzykowski[3,4]

[1]Department of Epidemiology and Data Science, Amsterdam UMC, VU University Amsterdam, Amsterdam, The Netherlands

[2]Microbiology Unit, SCK-CEN, Mol, Belgium

[3]I-Biostat, Data Science Institute, Hasselt University, Hasselt, Belgium

[4]Department of Statistics and Medical Informatics, Medical University of Bialystok, Bialystok, Poland

Correspondence
J. Claesen, Department of Epidemiology and Data Science, Amsterdam UMC, VU University Amsterdam, Amsterdam, The Netherlands.
Email: j.claesen@amsterdamumc.nl

Rationale: Identification of peptides and proteins is a challenging task in mass spectrometry–based proteomics. Knowledge of the number of sulfur atoms can improve the identification of peptides and proteins.

Methods: In this article, we propose a method for the prediction of S-atoms based on the aggregated isotope distribution. The Mahalanobis distance is used as dissimilarity measure to compare mass- and intensity-based features from the observed and theoretical isotope distributions.

Results: The relative abundance of the second and the third aggregated isotopic variants (as compared to the monoisotopic one) and the mass difference between the second and third aggregated isotopic variants are the most important features to predict the number of S-atoms.

Conclusions: The mass and intensity accuracies of the observed aggregated isotopic variants are insufficient to accurately predict the number of atoms. However, using a limited set of predictions for a peptide, rather than predicting a single number of S-atoms, has a reasonably high prediction accuracy.

1 | INTRODUCTION

In a mass spectrum, peptides and proteins appear as a series of correlated peaks corresponding to the fine or aggregated isotope distribution (Figure 1). The fine isotope distribution reflects the probabilities of occurrence of every isotopic variant of a molecule. If we ignore small deviations of the masses from integer values, the isotopic variants can be grouped into the aggregated isotopic variants. The aggregated isotope distribution provides the number and occurrence probabilities of these aggregated isotopic variants.[1] The fine or aggregated isotope distribution can be used, for instance, to interpret the mass spectral data[2] or to predict the elemental composition of biomolecules.[3,4]

The mass and the probabilities of occurrence of the isotopic variants of a molecule are a function of the elemental composition of the molecule and the elemental isotope definition.[5] Consequently, the presence of atoms with a distinctive elemental isotope definition has a profound effect on the isotope distribution of the biomolecule. For example, the presence of a monoisotopic element such as a phosphorus atom shifts the (aggregated) isotope distribution to a higher mass (by $\approx$ 31 Da) without changing the probabilities of occurrence of the isotopic variants. Another example is sulfur. Sulfur has four stable isotopes, ^{32}S, ^{33}S, ^{34}S, and ^{36}S, of which the first and third isotopes are the most abundant, with the probability of occurrence equal to about 94.85% and 4.365%, respectively (Table 1). Therefore, the probability of occurrence of the third (aggregated)

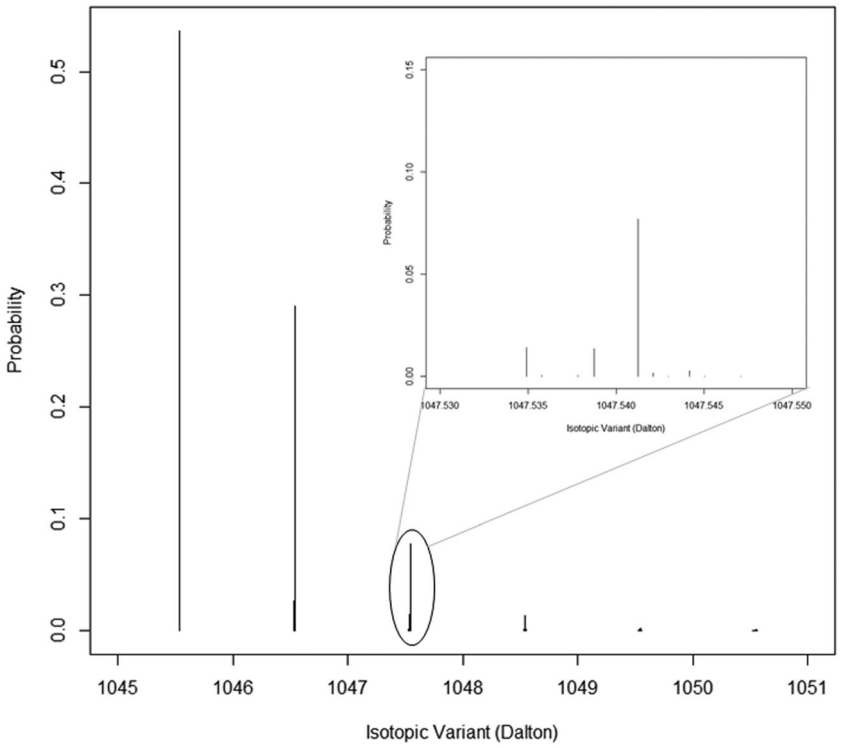

FIGURE 1 The aggregated isotope distribution of angiotensin II. The third aggregated peak consists of 11 isotopic variants

TABLE 1 Sulfur isotopes according to IUPAC2018 (Holden et al. 2018)

Isotope	Mass	Probability of occurrence
^{32}S	31.972071174	[0.944100, 0.952900]
^{33}S	32.971458910	[0.007290, 0.007970]
^{34}S	33.967867000	[0.039600, 0.047700]
^{36}S	35.967081000	[0.000129, 0.000187]

isotopic variant of a molecule with one or more sulfur (S-)atoms is larger than that for a molecule without S-atoms. In addition, the masses of the S-isotopes (31.972, 32.971, 33.968, and 35.967) influence the mass differences between the isotopic variants of a molecule. A molecule without S-atoms has a larger difference between the masses of the second and third (aggregated) isotopic variants as compared to a molecule with one or more S-atoms.[6]

Successful prediction of the number of S-atoms is beneficial for the identification of peptides and proteins in mass spectrometry experiments, because contradicting identifications can be flagged as false-positive findings. The prediction can also guide *de novo* identification. A sulfur prediction method can be useful to screen for disulfide-rich peptides.[7,8]

In the past two decades, several methods[9–14] have been introduced to determine the number of S-atoms of peptides and metabolites based on the fine isotopic distribution extracted from MS1 spectra. These methods derive the information about the mass and the isotope abundance from the fine isotopic variant containing ^{34}S ions and compare it to the monoisotopic variant.

In this paper, we introduce an approach to predict the number of S-atoms of peptides and small proteins based on the observed aggregated isotope distribution. In contrast to existing methods, our approach does not use information from the fine isotopic variant containing ^{34}S ions, as this variant is not resolved in aggregated isotope distributions observed in MS1 spectra. Therefore, we predict the number of S-atoms based on the isotope abundance and masses of the monoisotopic, second, and third aggregated isotopic variants that are found in MS1 spectra.

2 | METHODS

The probabilities of occurrence of the aggregated isotope distribution can be used to calculate the relative isotopic ratios, that is, the ratio between the probability of occurrence of the $(i + 1)$th isotopic variant and the ith isotopic variant.[15] Plotting the first $(r = 1)$ theoretical relative ratio (RR) against the second $(r = 2)$ RR shows distinctive groups, as indicated in Figure 2. Each group corresponds to a specific number of S-atoms. The differences between these distinctive groups are mainly due to the second RR, that is, the ratio of the probabilities of occurrence of the third and the second aggregated isotopic variants. Note that, on the atomic level, the isotopic abundance of the third sulfur isotope is five to six times larger than that of the second isotope (Table 1). Therefore, at the molecular level, the probability of occurrence of the third (aggregated) isotopic variant of a molecule increases much faster than that of the second (aggregated) isotopic variant when the number of S-atoms increases.

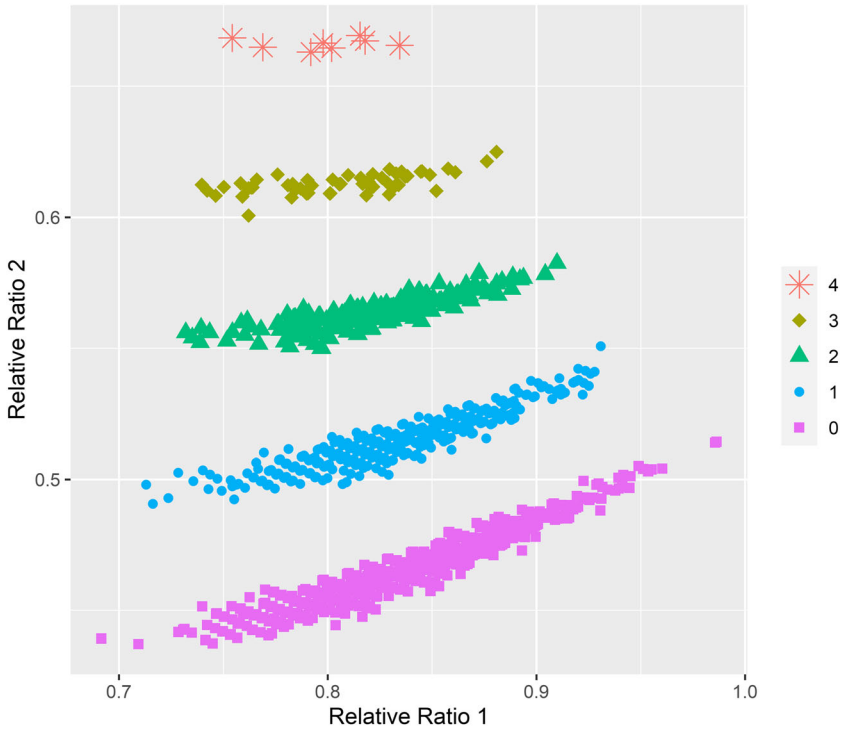

FIGURE 2 Theoretical relative isotope ratios of peptides of the human proteome (UniProtKB 9606, keyword 181, release 2011-11) with a monoisotopic mass between 1500 and 1505 Da. Each peptide is colored according to its number of sulfur atoms

The mass differences between the (aggregated) isotopic variants of biomolecules present a pattern similar to that presented in Figure 3, that is, distinct groups of peptides differing by the number of sulfur atoms. The mass difference between the second and third sulfur isotopes (Table 1), that is, 0.99640809 Da, is the main cause of the occurrence of the distinct groups. In particular, the mass difference is smaller than the mass differences between ^{12}C and ^{13}C (1.00354835 Da), ^{14}N and ^{15}N (0.99734895 Da), ^{16}O and ^{17}O (1.004317138), ^{17}O and ^{18}O (1.00002785), and ^{32}S and ^{33}S (0.999387736). In combination with the high occurrence probability of the third sulfur isotope, this mass defect has a substantial effect on the mass of the third aggregated isotopic variant: the mass difference between the second and third aggregated isotopic variants decreases when the number of S-atoms increases (Figure 3).

In a mass spectrum, the probabilities of occurrence of the (aggregated) isotopic variants of a molecule are reflected by the intensity or height of the peaks of the (aggregated) isotope distribution. Therefore, the first ($r = 1$) and second ($r = 2$) RRs can be estimated from MS1 spectra by computing the ratios of the intensities. By comparing these observed RRs with their theoretical values, one could determine the number of S-atoms in a peptide or protein by using a high-quality MS1 spectrum. Similarly, comparing the theoretical mass differences with the mass differences of the observed (aggregated) isotopic variants could also be used to predict the number of S-atoms.

It is worthwhile to mention that, due to the limited accuracy of the spectral intensities and the masses of the (aggregated) isotopic variants measured in MS1 spectra obtained by the currently available equipment, the observed RRs and mass differences of a biomolecule may significantly deviate from their theoretical values. In addition, within each distinctive cluster, there is a substantial correlation between the RRs (Figure 2), between the mass differences (Figure 3), and between the RRs and the mass differences (Figure FIGURE S1, supporting information). These deviations and correlations should be considered when constructing a prediction algorithm.

The Mahalanobis distance[16] is a dissimilarity measure that captures, in a multidimensional space, the distance between two points (row vectors) $\boldsymbol{x}$ and $\boldsymbol{y}$ that come from the same distribution with the variance–covariance matrix $\boldsymbol{\Sigma}$. It is defined as follows:

$$D(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\Sigma}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{y})^T}. \quad (1)$$

Thus, the distance measure accounts for the variation and correlation present in a distribution.

To construct a prediction algorithm, we considered characterizing each peptide by the first and second RRs (which will be denoted by *RR1* and *RR2*, respectively) and/or by the mass differences between the first and second (aggregated) isotopic variants and between the second isotopic and third (aggregated) isotopic variants, which will be denoted by $\Delta m21$ and $\Delta m32$, respectively. More concretely, we investigated the predictive value of following combinations of metrics represented by the following four vectors: (*RR1*, *RR2*), ($\Delta m21$, $\Delta m32$), (*RR1*, *RR2*, $\Delta m32$), and ($\Delta m12$, $\Delta m32$, *RR2*).

The proposed algorithm is summarized in Figure 4. Assume that a peptide with monoisotopic mass m is observed in an MS1 spectrum. We will term it an "observed" peptide. For this observed peptide, the value $\boldsymbol{x}$ of a particular form of the vector of metrics (e.g., (*RR1*, *RR2*)) is derived from the observed (aggregated) isotope distribution of the peptide.

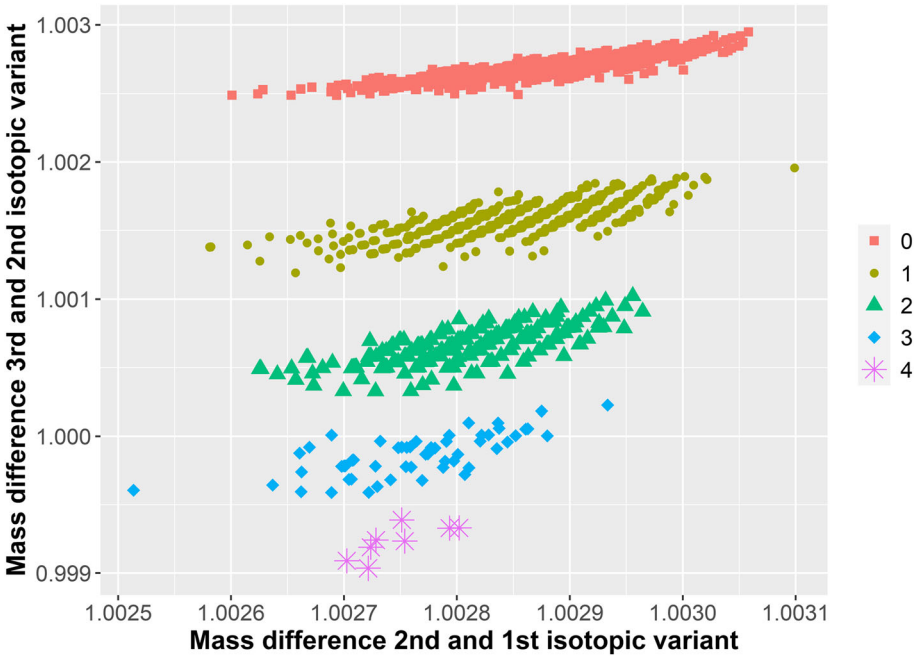

FIGURE 3 Theoretical mass differences of the isotopic variants of peptides of the human proteome (UniProtKB 9606, keyword 181, release 2011-11) with a monoisotopic mass between 1500 and 1505 Da. Each peptide is colored according to its number of sulfur atoms

Step 1: calculate the ***x***-vector from the observed isotope distribution

Step 2: select from the human proteome all peptides with monoisotopic masses within a 20 Da wide interval around the observed monoisotopic mass

Step 3: group the peptides according to the number of S-atoms

Step 4: for each group of theoretical peptides:

a. calculate the ***y***-vector for each theoretical peptide in the selected group
b. calculate the variance-covariance matrix $\boldsymbol{\Sigma}_s$ of vectors ***y***
c. calculate the Mahalanobis distance between the ***x***-vector and the ***y***-vectors
d. average the Mahalanobis distances

Step 5: determine the smallest Mahalanobis distance

Step 6: number of S-atoms of the observed peptide is equal to the number of S-atoms of the group of theoretical peptides with the smallest Mahalanobis distance

FIGURE 4 Proposed algorithm for the prediction of S-atoms based on the observed aggregated isotope distribution

Next, peptides from an *in silico* tryptic digest (without missing cleavages) of the human proteome (UniProtKB 9606, keyword 181, release 2011-11), with monoisotopic masses within a 20 Da wide interval around *m*, are selected. We will term them "theoretical" peptides. For each of those theoretical peptides, the value ***y*** of the particular form of the vector of metrics is computed by using the theoretical (aggregated) isotope distribution based on the peptide's elemental composition.

Subsequently, the theoretical peptides are split into separate groups containing the same number of S-atoms. We will index these groups by index *s* equal to 0, 1, 2, and so on. For group *s*, the variance–covariance matrix $\boldsymbol{\Sigma}_s$ of vectors ***y*** is computed. Then, using the obtained matrix $\boldsymbol{\Sigma}_s$, the Mahalanobis distance is computed for vector ***x*** for the observed peptide and vector ***y*** for every theoretical peptide from group *s*. Finally, the obtained values of the Mahalanobis distance are averaged.

The calculation is repeated for each group *s*. In the final step, the number of S-atoms of the observed peptide, which was characterized by a vector ***x***, is predicted as being equal to the number of S-atoms of the group *s* that was, according to the average Mahalanobis distance, the closest to ***x***.

Inaccuracies of mass and intensity of the isotopic variants of a peptide may lead to incorrect predictions. To compensate for these inaccuracies, one can extract, if available, multiple aggregated isotope distributions of the same peptide (e.g., with different charges and/or with different retention times) and average these isotope distributions. (Alternative approaches to combine multiple aggregated isotope distributions of one peptide can also be considered.) The averaged aggregated isotope distribution can be used to calculate the ***x***-vector and subsequently used as an input to the proposed prediction algorithm. We will refer to this approach as the "average-***x***" prediction rule. The approach in which the extracted isotope distributions of a peptide are not averaged, but for each individual isotope distribution the ***x***-vector is computed, will be referred to as the "individual-***x***" prediction rule.

3 | DATA

To illustrate the proposed method, we selected two data sets from two different mass spectrometers. The first data set is a tryptic digest of the *Candida albicans* plasmid pHis3 (PXD011194[17]) measured using an Orbitrap Q Exactive mass spectrometer (ThermoFisher Scientific, Waltham, Massachusetts, US). The second data set is a HeLa cell tryptic digest (PXD001592[18]) recorded using an Impact II ESI-Q-TOF (Bruker, Billerica, Massachusetts, US). For both data sets, lists of peptides and proteins identified with MaxQuant[19] were available. According to MaxQuant, the average mass resolution of the pHis3 data set is equal to 54 805.74 and the average uncalibrated mass error is 0.682 ppm for pHis3 and 8.516 ppm for the HeLa data set. For the latter, no information on the mass resolution was available.

4 | RESULTS AND DISCUSSION

We randomly selected 333 identified peptides containing 0–4 S-atoms observed in the pHis3 data set and 560 peptides containing 0–7 S-atoms observed in the HeLa data set (Table 2). For each selected peptide, we attempted to extract multiple aggregated isotope distributions with different charges from the MS1 spectrum and from 10 adjacent spectra (5 before and 5 after the corresponding MS1 spectrum) such that the first three isotopic variants were present. A 50 ppm wide mass-tolerance window was used to select the aggregated isotopic variants based on the expected masses of the aggregated isotopic variants.

Therefore, we found 3654 aggregated isotope distributions corresponding to 294 unique peptides and 8005 aggregated isotope distributions corresponding to 453 unique peptides in the pHis3 and the HeLa data sets, respectively.

For each data set and each peptide with an observed aggregated isotope distribution, we applied the "individual-***x***" and "average-***x***" prediction rules described earlier (Figure S2, supporting information). For both prediction rules, we evaluated their performance when considering prediction based on the *s* group of the theoretical peptides with the smallest, the second-smallest, and the third-smallest averaged Mahalanobis distance. We also considered the performance when using the list of predicted numbers of S-atoms suggested by the three smallest Mahalanobis distances. Note that we assess the performance of the proposed prediction rules under the assumption that the randomly selected peptides are correctly identified. Consequently, the "true" performance of the prediction rules might differ from the performance reported here.

4.1 | pHis3 data set

Table 3 summarizes the results of the "individual-***x***" prediction rule applied to the 3654 individual aggregated isotope distributions. In particular, the table presents the accuracy of the prediction by

TABLE 2 Number of selected peptides and aggregated isotope distributions for the pHis3 and HeLa data sets

	pHis3			HeLa		
Number of sulfur atoms	Number of selected peptides	Number of found peptides	Number of found aggregated isotope distributions	Number of selected peptides	Number of found peptides	Number of found aggregated isotope distributions
0	100	92	1066	100	91	1587
1	100	94	1114	100	81	1509
2	100	84	1149	100	90	1594
3	29	20	287	100	78	1257
4	4	4	38	100	73	1344
5	0	0	0	54	37	658
5	0	0	0	5	3	56
7	0	0	0	1	0	0
Total	333	294	3654	560	453	8005

TABLE 3 Number of correctly predicted S-atoms with the "individual-***x***" prediction rule for the pHis3 data set

Number of S-atoms in the molecule	Smallest distance	Second-smallest distance	Third-smallest distance	Three smallest distances
0	800	84	58	942
1	149	735	93	977
2	118	128	677	923
3	26	52	48	126
4	8	6	3	17
Total	1101	1005	879	2985

Note. For each individual aggregated isotope distribution of a peptide, the number of S-atoms has been predicted based on the (Δ*m32*, *RR1*, *RR2*)-vector.

TABLE 4 Number of correctly predicted S-atoms with the "average-***x***" prediction rule for the pHis3 data set

Number of S-atoms in the molecule	Smallest distance	Second-smallest distance	Third-smallest distance	Three smallest distances
0	75	5	7	87
1	15	69	3	87
2	11	10	51	72
3	2	2	4	8
4	0	1	1	2
Total	103	87	66	256

Notes. The observed masses and intensities of the isotope distributions of the same peptide across multiple spectra were averaged. For each peptide, the number of S-atoms has been predicted based on the (Δ*m32*, *RR1*, *RR2*)-vector.

comparing the number of S-atoms derived from the MaxQuant peptide identification with the sulfur prediction from the smallest, second-smallest, or third-smallest averaged Mahalanobis distance or by the inclusion of the correct number of S-atoms in the list of predictions obtained by considering simultaneously the three distances.

When the smallest averaged Mahalanobis distance was considered, the number of S-atoms was correctly predicted for 30.1% of the peptides (1101/3654) characterized by vector (*RR1*, *RR2*, Δ*m32*). The majority of these peptides (800) did not contain any S-atom. When the second-smallest averaged Mahalanobis distance was used, the number of S-atoms was predicted correctly for 27.5% of the peptides (1005/3654), including 735 cases with one S-atom. When the number of S-atoms was predicted by using the peptide group corresponding to the third-smallest averaged Mahalanobis distance, the number of S-atoms was correctly predicted for 24.1% of the peptides (879/3654), the majority (677) of which included two S-atoms. The accuracy of the prediction based on the use of the set of the three smallest averaged Mahalanobis distances was equal to 81.7% (2985/3654).

When peptides were characterized by vectors (*RR1*, *RR2*), (Δ*m21*, Δ*m32*), and (Δ*m12*, Δ*m32*, *RR2*), the results of the "individual-***x***" prediction rule were less accurate, though a pattern similar to that observed in Table 3 was present (see Supplementary File 1).

Table 4 summarizes the results of the "average-***x***" prediction rule combined with the use of the vector (*RR1*, *RR2*, Δ*m32*) to characterize the peptides. As compared to the "individual-***x***" prediction rule, the prediction accuracy increased to 35.0% (103 correct predictions of 294) when the smallest Mahalanobis distance was used and to 87.1% (256/294) when the set of the predictions for the three smallest Mahalanobis distances was used. The prediction accuracy when the smallest Mahalanobis distance was used was the highest for peptides without any S-atoms. For peptides with one S-atom, the accuracy was the highest when the second-smallest Mahalanobis distance was used, whereas for peptides with two S-atoms, it was the highest when the third-smallest Mahalanobis distance was considered. Similar to the "individual-***x***" prediction rule, the accuracy of the "average-***x***" prediction rule was lower when peptides were characterized by vectors (*RR1*, *RR2*), (Δ*m21*, Δ*m32*), and (Δ*m12*, Δ*m32*, *RR2*) (see Supplementary File 1).

4.2 | HeLa data set

Table 5 presents the results of the "individual-***x***" prediction rule for the aggregated isotope distributions of 8005 nonunique peptides. The peptides were characterized by vector (*RR1*, *RR2*, Δ*m32*). When the peptide group corresponding to the smallest averaged Mahalanobis distance was used to predict the number of S-atoms, prediction accuracy of 26.0% (2084/8005) was obtained. When the set of the predictions for the three smallest Mahalanobis distances was used, the prediction accuracy was equal to 65.7% (5256/8005). Peptides with no S-atoms were most often correctly predicted (1163/2084, i.e., 55.8%) when the smallest Mahalanobis distance was used, peptides with one S-atom were most often correctly predicted (885/1509, i.e., 58.6%) when the second-smallest Mahalanobis

TABLE 5 Number of correctly predicted S-atoms with the "individual-***x***" prediction rule for the HeLa data set

Predicted number of S atoms	Smallest distance	Second-smallest distance	Third-smallest distance	Three smallest distances
0	1163	167	102	1432
1	249	885	173	1307
2	194	261	898	1353
3	183	180	145	508
4	234	180	100	514
5	61	36	43	140
6	0	1	1	2
Total	2084	1710	1462	5256

Note. For each individual aggregated isotope distribution of a peptide, the number of S-atoms has been predicted based on the ($\Delta m32$, *RR1*, *RR2*)-vector.

TABLE 6 Number of correctly predicted S-atoms with the "average-***x***" prediction rule for the HeLa data set

Predicted number of S-atoms	Smallest distance	Second-smallest distance	Third-smallest distance	Three smallest distances
0	45	9	14	68
1	19	21	14	54
2	15	16	33	64
3	19	22	15	56
4	20	14	12	46
5	11	10	3	24
6	1	0	0	1
Total	130	82	88	288

Notes. The observed masses and intensities of the isotope distributions of the same peptide across multiple spectra were averaged. For each peptide, the number of S-atoms has been predicted based on the (*RR1*, *RR2*)-vector.

distance was used, and peptides with two S-atoms were mainly correctly predicted when the third-smallest Mahalanobis distance (898/1594, i.e., 56.3%) was used.

Using the "average-***x***" prediction rule did not lead to any substantial improvement in the prediction accuracy when characterizing peptides by vector (*RR1*, *RR2*, $\Delta m32$) (see Supplementary File 1). Using vector (*RR1*, *RR2*) led to an improvement in the prediction accuracy of about 3% (from 26.0% to 28.7%) for the smallest Mahalanobis distance and to no improvement (from 63.9% to 63.6%) for the set of the three distances, as compared to the "individual-***x***" prediction rule (Table 6). As observed in the pHis3 data set, the prediction accuracy for peptides without any S-atoms was the highest (34.6%) when the smallest Mahalanobis distance was used, and for peptides with two sulfur atoms, it was the highest when the third-smallest Mahalanobis distance (37.5%) was considered. For the second-smallest Mahalanobis distance, the prediction accuracies for peptides with one or three sulfur atoms were the highest and equal to 25.6% and 26.8%, respectively.

The differences in the prediction accuracy of the HeLa and pHis3 data sets may be explained by the difference in the observed mass accuracies. The average mass accuracy of the HeLa data set was equal to 10.79 ppm, as compared to 1.6 ppm for the pHis3 data set (Table FIGURE S1, supporting information). The mass accuracy of both data sets improved when the extracted isotope distributions were averaged. However, the improvement was much more substantial for the pHis3 data set (±1.6 ppm) than for the HeLa data set (±0.5 ppm). This might explain why no or little improvement in the prediction accuracy could be observed for the HeLa data set when comparing the "average-***x***" prediction rule with the "individual-***x***" prediction rule.

For the majority of peptides, multiple aggregated isotopic distributions were extracted. We checked if the accuracy of predicting the number of S-atoms was influenced by peptide charge, intensity of the monoisotopic peak, mass accuracy, differences between the theoretical and observed RRs, and differences between the theoretical and observed differences of the masses of the second and third aggregated isotopic variants.

In particular, for each peptide with multiple extracted isotopic distributions, we compared the number of cases with correctly predicted number of S-atoms across the charges. Differences between the number of correctly predicted number of S-atoms could be observed (Figures S3 and S4, supporting information). However, these differences were limited and centered around zero, indicating that charge did not have any systematic effect on the precision of the "individual-***x***" prediction rule.

Similarly, we studied the effect of the intensity of the monoisotopic peak. First, we categorized the intensity into five distinct classes, ranging from a very low to a very high intensity

(Table S2, supporting information). Subsequently, we compared the number of correctly predicted S-atoms for each peptide with multiple aggregated isotopic distributions for which the monoisotopic intensities were categorized in at least two different classes. For the majority of peptides, no or limited differences in the number of correctly predicted S-atoms were found (Figures S5 and S6, supporting information). This indicates that the intensity of the monoisotopic peak did not influence the outcome of the "individual-***x***" prediction rule.

The effect of the mass accuracy and deviations from the first ($r = 1$) and second ($r = 2$) theoretical RRs, categorized into five classes (Tables S3 and S4, supporting information), on the prediction accuracy of the "individual-***x***" prediction rule was also limited (Figures S7–S10, supporting information).

Deviations from the theoretical difference between the mass of the second and third isotopic variants (Table S5, supporting information) influenced the performance of the prediction rule (Figures S11 and S12, supporting information). The differences in the number of correct predictions increased when the deviations from the theoretical $\Delta m32$ values increased.

When the multiple isotope distributions of one peptide were compared with the distributions of the other peptides, the prediction accuracy of the "individual-***x***" prediction rule decreased when the mass accuracy deteriorated for the pHis3 data set, whereas this was not the case for the HeLa data set (Table S6, supporting information). For the deviations from the first ($r = 1$) and second ($r = 2$) theoretical RRs, the prediction accuracy decreased when the deviations increased for the pHis3 data set, whereas for the HeLa data set, the accuracy remained the same or even increased when the deviations from the expected RRs increased (Table S7, supporting information). A potential explanation for the latter effect might be that, whereas the deviations from the theoretical isotope ratios increase, the mass accuracy increases and/or the deviations from the theoretical $\Delta m32$ values decrease. Deviations from the theoretical $\Delta m32$ values have a negative effect on the accuracy of the prediction rule. When these deviations increase, that is, above 0.06 Da, the accuracy of the prediction rule decreases (Table S8, supporting information).

We also evaluated the effect of posttranslational modifications (acetylation and oxidation) and the effect of cysteine carbamidomethylation on the performance of the "individual-***x***" prediction rule (Table S9, supporting information).

In the case of acetylation, we found 11 different isotope distributions for 1 acetylated peptide (pHis3) and 104 isotope distributions for 7 acetylated peptides (HeLa). With the "individual-***x***" prediction rule, when the three smallest distances were combined, a prediction accuracy of 91% (10/11) and 90% (104/115) was found for the pHis3 and HeLa data sets, respectively. The prediction accuracy increased to 100% (1/1) for the pHis3 data set but decreased to 71.4% (5/7) for the HeLa data set when the "average-***x***" prediction rule was used.

In the pHis3 data set we found 28 oxidized peptides; there was one such peptide in the HeLa data set. For the oxidized peptides, 435 and 23 aggregated isotopic distributions were extracted. The number of S-atoms was predicted correctly for 82.1% (357/435) and 91.3% (21/23) of the distributions with the "individual-***x***" prediction rule, respectively, and for all peptides with the "average-***x***" prediction rule in both data sets.

In the pHis3 data set, 689 aggregated isotope distributions for 65 peptides with one or more carbamidomethylated cysteines were extracted. In the HeLa data set, we found 3769 aggregated isotope distributions for 220 peptides with one or more cysteines that were carbamidomethylated. The prediction accuracy of the "individual-***x***" prediction rule for these peptides was equal to, respectively, 78.5% (541/689) and 56.2% (2117/3769), whereas the prediction accuracy of the "average-***x***" rule was lower, that is, equal to 72.4% and 55.3% for the pHis3 and HeLa data sets, respectively.

Based on the aforementioned results we can conclude that the occurrence of the evaluated (posttranslational) modifications had little or no effect on the accuracy of the proposed prediction rule(s).

5 | SUMMARY AND CONCLUSIONS

In this paper, we investigated the prediction of the number of S-atoms of a peptide or a protein based on the observed isotope distribution. Our analysis indicates that, although the theoretical isotope ratios and theoretical mass differences clearly show distinct groups of peptides and proteins with differing number of S-atoms, the mass and intensity accuracies of the observed aggregated isotopic variants are insufficient to accurately predict the number of the atoms. Averaging the observed intensities and masses of the isotopic variants moderately improves the prediction accuracy. Using the extracted ion chromatograms to determine which aggregated isotope distributions of a peptide should be averaged may lead to higher prediction accuracies. A reasonably high accuracy can be obtained if, instead of predicting the correct number of S-atoms for an observed peptide, one focuses on including the correct number in a limited set of predictions.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available at the PRIDE database with reference numbers PXD011194 and PXD001592.

ORCID

Jürgen Claesen https://orcid.org/0000-0001-7615-5322

REFERENCES

1. Valkenborg D, Mertens I, Lemière F, Witters F, Burzykowski T. The isotopic distribution conundrum. *Mass Spectrom Rev*. 2011;31(1):96-109.
2. Burzykowski T, Claesen J, Valkenborg D. The analysis of peptide-centric mass spectrometry data utilizing information about the expected isotope distribution. In: Datta S, Mertens BJA, eds. *Statistical Analysis of Proteomics, Metabolomics, and Lipidomics Data Using Mass Spectrometry*. Switzerland: Frontiers in Probability and the Statistical Sciences Springer International Publishing; 2017:45-64.

3. Kind T, Fiehn O. Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*. 2007;8(1):105-124.
4. Claesen J, Valkenborg D, Burzykowski T. De novo prediction of the elemental composition of peptides and proteins based on a single mass. *J Mass Spectrom*. 2019;55(8):e4367. https://doi.org/10.1002/jms.4367
5. Holden NE, Coplen TB, Böhlke JK, et al. IUPAC periodic table of the elements and isotopes (IPTEI) for the education community (IUPAC technical report). *Pure Appl Chem*. 2018;90(12):1833-2092.
6. Sleno L. The use of mass defect in modern mass spectrometry. *J Mass Spectrom*. 2012;47(2):226-236.
7. Neitz S, Jürgens M, Kellmann M, Schulz-Knappe P, Schrader M. Screening for disulfide-rich peptides in biological sources by carboxyamidomethylation in combination with differential matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom*. 2001;15(17):1586-1592.
8. Samgina TY, Vorontsov EA, Gorshkov VA, et al. Novel cysteine tags for the sequencing of non-tryptic disulfide peptides of anurans: ESI-MS study of fragmentation efficiency. *J Am Soc Mass Spectrom*. 2011;22(12):2246-2255.
9. Shi SDH, Hendrickson CL, Marshall AG. Counting individual sulfur atoms in a protein by ultrahigh-resolution Fourier transform ion cyclotron resonance mass spectrometry: Experimental resolution of isotopic fine structure in proteins. *Proc Natl Acad Sci U S A*. 1998;95(20):11532-11537.
10. Miura D, Tsuji Y, Takahashi K, Wariishi H, Saito K. A strategy for the determination of the elemental composition by Fourier transform ion cyclotron resonance mass spectrometry based on isotopic peak ratios. *Anal Chem*. 2010;82(13):5887-5891.
11. Moseley HNB. Correcting for the effects of natural abundance in stable isotope resolved metabolomics experiments involving ultra-high resolution mass spectrometry. *BMC Bioinform*. 2010;11(1):139.
12. Nagaoa T, Yukihiraa D, Fujimurab Y, et al. Power of isotopic fine structure for unambiguous determination of metabolite elemental compositions: In silico evaluation and metabolomic application. *Anal Chim Acta*. 2014;813:70-76.
13. Yang M, Zhou Z, Guo D. A strategy for fast screening and identification of sulfur derivatives in medicinal *Pueraria* species based on the fine isotopic pattern filtering method using ultra-high-resolution mass spectrometry. *Anal Chim Acta*. 2015;894:44-53.
14. Nakaybashi R, Sato K. Ultrahigh resolution metabolomics for S-containing metabolites. *Curr Opin Biotechnol*. 2017;43:8-16.
15. Valkenborg D, Jansen I, Burzykowski T. A model-based method for the prediction of the isotopic distribution of peptides. *J Am Soc Mass Spectrom*. 2008;19(5):703-712.
16. Mahalanobis P. On the generalised distance in statistics. *Proc Natl Acad Sci India*. 1936;2(1):49-55.
17. Perez-Riverol Y, Csordas A, Bai J, et al. The PRIDE database and related tools and resources in 2019: Improving support for quantification data. *Nucleic Acids Res*. 2019;47(D1):D442-D450.
18. Beck S, Michalski A, Raether O, et al. The impact II, a very high-resolution quadrupole time-of-flight instrument (QTOF) for deep shotgun proteomics. *Mol Cell Proteomics*. 2015;14(7):2014-2029.
19. Cox J, Mann M. Maxquant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*. 2008;26(12):1367-1372.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Claesen J, Valkenborg D, Burzykowski T. Predicting the number of sulfur atoms in peptides and small proteins based on the observed aggregated isotope distribution. *Rapid Commun Mass Spectrom*. 2021;35(19):e9162. https://doi.org/10.1002/rcm.9162