A Model-Based Method for the Prediction of the Isotopic Distribution of Peptides

Dirk Valkenborg, Ivy Jansen, and Tomasz Burzykowski
Center for Statistics, Hasselt University, Diepenbeek, Belgium

The process of monoisotopic mass determination, i.e., nomination of the correct peak of an isotopically resolved group of peptide peaks as a monoisotopic peak, requires prior information about the isotopic distribution of the peptide. This points immediately to the difficulty of monoisotopic mass determination, whereas a single mass spectrum does not contain information about the atomic composition of a peptide and therefore the isotopic distribution of the peptide remains unknown. To solve this problem a technique is required, which is able to estimate the isotopic distribution given the information of a single mass spectrum. Senko et al. calculated the average isotopic distribution for any mass peptide via the multinomial expansion (Yergey 1983) [1], using a scaled version of the average amino acid Averagine (Senko et al. 1995) [2]. Another method, introduced by Breen et al., approximates the result of the multinomial expansion by a Poisson model (Breen et al. 2000) [3]. Although both methods perform well, they have their specific limitations. In this manuscript, we propose an alternative method for the prediction of the isotopic distribution based on a model for consecutive ratios of peaks from the isotopic distribution, similar in spirit to the approach introduced by Gay et al. (1999) [5]. The presented method is computationally simple and accurate in predicting the expected isotopic distribution. Further, we extend our method to estimate the isotopic distribution of sulphur-containing peptides. This is important because the naturally occurring isotopes of sulphur have an impact on the isotopic distribution of a peptide. (J Am Soc Mass Spectrom 2008, 19, 703–712) © 2008 American Society for Mass Spectrometry

When analyzing a mass spectrometry (MS) experiment, measurements made for a mass spectrum usually need to be turned into a simple peak list that contains the monoisotopic masses and abundances of the corresponding peptides. This process is called monoisotopic mass determination, also referred to as deisotoping or deconvoluting of a mass spectrum. To nominate the correct peak of an isotopically resolved group of peptide peaks as a monoisotopic peak and to distinguish it from peaks accidentally generated by error, one can consider using a method that predicts the isotopic distribution related to a peptide.

Existing methods for the estimation of the isotopic distribution, as described by, e.g., Senko et al., calculate the average isotopic distribution for any mass peptide via the multinomial expansion [1], using a scaled version of Averagine [2]. Another method, introduced by Breen et al., approximates the result of the multinomial expansion by a Poisson model [3]. Although both methods perform well, they have their specific limitations. The method of Breen et al. is fast but not accurate for sulphur-containing peptides [4], while the method of Senko et al. is computationally involved and, as we explain later in the manuscript, not accurate enough for low mass peptides.

For these reasons, we have developed an alternative method for the prediction of the isotopic distribution based on a model for consecutive ratios of peaks from the isotopic distribution, similar in spirit to the approach introduced by Gay et al. [5]. The presented method is computationally simple and accurate in predicting the expected isotopic distribution. However, it should be noted that all the aforementioned methods are only applicable when the isotopic peaks are individually resolved. Thus, for this purpose, a high-resolution mass spectrometer is mandatory.

Experimental

Materials

A dataset of 2154 tandem MS sequenced peptides (MASCOT score above 40 at 95% significance level) found in human serum was used for the validation of the proposed method. The isotopic distribution of the peptides was calculated by the multinomial expansion implemented in the Isotopic Pattern Calculator (IPC) [6].

Although sulphur is a scarce atom in amino acids (as only cysteine and methionine contain a sulphur atom), this is not equivalent to a low abundance of sulphur

Address reprint requests to Dr. Dirk Valkenborg, Hasselt University, Center for Statistics, Agoralaan 1 - building D, 3590 Diepenbeek, Belgium. E-mail: dirk.valkenborg@uhasselt.be

1044-0305/08/$32.00
doi:10.1016/j.jasms.2008.01.009

Published online January 31, 2008
Received August 22, 2007
Revised January 8, 2008
Accepted January 22, 2008

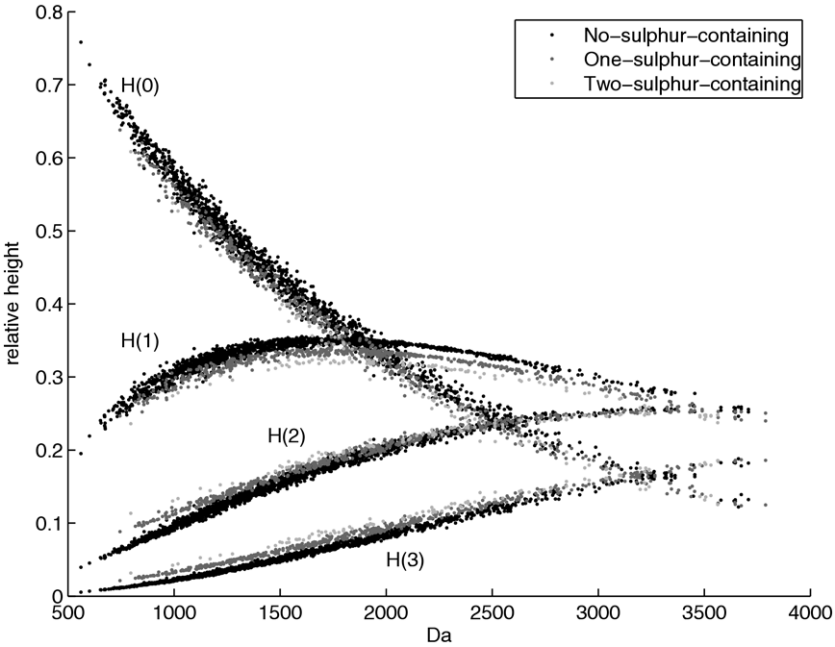

Figure 1. The relative height $H(i)$ of the first four isotopic peaks ($i = 0,\ldots,3$) of the 2137 peptides.

containing peptides. Out of the 2154 peptides, 1542 (71.59%) did not contain any sulphur atom, 502 (23.31%) contained one sulphur atom, and 93 (4.32%) contained two sulphur atoms. Only 17 (0.78%) peptides out of 2154 contained more than two sulphur atoms. Because the methods developed in this manuscript only consider the prevalence of one and two sulphur atoms in peptides, the set of 17 peptides was removed from the data. Thus, finally, in our study a dataset of 2137 peptides was used.

It should be noted that the aforementioned estimate of the number of sulfur atoms in a peptide was determined using a limited dataset. Possibly, the prevalence of peptides with more than two sulphur atoms is more pronounced. To retrieve a more precise determination about the sulphur prevalence in peptides, one could consider using a database such as the PIR protein database, and performing, e.g., an in silico tryptic digest for all human proteins. However, this is beyond the scope of this manuscript, where we introduce a method to estimate the isotopic distribution for sulphur-containing peptides.

Figure 1 shows the relative height $H(i)$ of the first four isotopes ($i = 0,\ldots,3$) of the calculated isotopic distribution of the 2137 peptides. The effect of a sulphur atom on the isotopic distribution of a peptide is clearly visible (i.e., the isotopes have a different relative height). The isotopic distributions of sulphur-containing peptides are different from those of no-sulphur peptides. This indicates that, in contrast to the assumption made by Breen et al. [3], the effect of sulphur should not be ignored in the approximation of the distributions.

Methods

In this section, two methods for approximating the isotopic distribution of an average peptide are reviewed, together with a short discussion of their properties. Then, a new method, which avoids the limitations of the two reviewed approaches, is proposed.

Approximation by Averagine

To calculate the isotopic distribution of a peptide, Senko et al. [2] proposed the concept of a virtual amino acid Averagine, constructed using the statistical occurrences of amino acids in the PIR protein database. The resulting average amino acid was

$$\text{Averagine} = C_{4.9384}H_{7.7583}O_{1.4773}N_{1.3577}S_{0.0417}, \quad (1)$$

with a mass of 1111.254 Da. Note the low abundance of sulphur in Averagine.

Assume that a peptide of 1111.254 Da is observed in the mass spectrum. This peptide is 10 times heavier than Averagine; thus if we multiply the elements of Averagine by 10, we obtain the average elemental composition for a peptide with mass 1111.254. After rounding the elements to the nearest integer, the peptide's isotopic distribution can be computed from the known relative abundances of the isotopic variants of carbon (C), hydrogen (H), nitrogen (N), oxygen (O), and sulphur (S) using the multinomial expansion [1]. With this straightforward method the isotopic distribution for a peptide is easily estimated. However, this method has some limitations:

- The multinomial expansion is computer intensive. Although algorithms exist that can speed up computations, it still requires considerable computing power.
- Scaling Averagine is not accurate enough for peptides. Consider a peptide with a mass below 1332.4388 Da. Scaling and rounding of Averagine will always result in an average peptide with no sulphur in its elemental composition. Peptides with a mass between 1332.4388 Da and 3997.3165 Da will always be assumed to contain one sulphur atom, etc. In other words, according to this model, the presence of sulphur is purely dependent on the mass of the observed peptide. Of course, it is more likely for a heavy peptide to contain a sulphur atom than it is for a low mass peptide, but this distinction is more subtle then the yes–no case implied by the method of Senko et al. This can also be observed from Figure 1, where sulphur-containing peptides are present over the entire mass range.

Poisson Approximation

Breen et al. [3] suggested to approximate the result of the multinomial expansion (i.e., the expected proportional heights of different isotopic peaks) by a Poisson distribution. To compute the distribution, its mean, say M, needs to be known. The mean depends on the number of atoms n of a particular type (C, H, N, O, and S), as well as on the proportional abundances p of the isotopic variants. In practice, only a peptide's molecular monoisotopic mass, say m, will be available. Breen et al. have developed a mapping from m to $M = np$. To this aim, they constructed an average amino acid (AA)

$$u = C_{10}H_{16}O_3N_3 \tag{2}$$

by averaging all AAs from all proteins in the SWISS-PROT protein database. Note that this approximately corresponds to Averagine after multiplication of the elements by 2. Next, a set of theoretical peptides

$$H - (u \ldots u) - OH - H^+, \tag{3}$$

were used to span a mass range from $m_1 = 245.1376$ Da to $m_{15} = 3410.8059$ Da. For each so-constructed theoretical peptide of mass m_i ($i = 1,\ldots,15$) the isotopic distribution E' was calculated using Protein Prospector [7]. The mean M_i of a Poisson distribution giving the best approximation to the isotopic distribution at m_i was then found by minimizing the sum of absolute deviations between the components of both distributions:

$$M^*(m) = \underset{M}{\operatorname{argmin}} \sum_{x=1}^{\infty} \left| E'(0,m)\frac{P(x;M)}{P(0;M)} - E'(x,m) \right|, \tag{4}$$

with $E'(x,m)$, the relative height of the xth isotopic peak for a theoretical peptide with mass m; $P(x;M) = e^{-M}M^x/x!$, the Poisson distribution.

As a result, a linear relationship between Poisson mean M and monoisotopic mass m was found:

$$M = 0.000594m - 0.03091. \tag{5}$$

The relationship allows to compute the mean M of a Poisson approximation to the isotopic distribution of a peptide with monoisotopic mass m. The approximation can then be used to compute expected proportional heights of peaks observed in a spectrum and, in turn, to decide whether the observed peaks can correspond to a series generated by a peptide. This results in a fast alternative for the multinomial expansion. However, the method also has limitations. In particular, sulphur is completely ignored in the construction of the theoretical peptides. This results in a biased estimate of the expected isotopic distribution for sulphur-containing peptides as reported by Valkenborg et al. [4].

Modeling the Ratios of the Peaks of Isotopic Distributions

To account for the effect of sulphur on the isotopic distribution, we propose an alternative approach. It is similar in spirit to the method developed by Gay et al. [5], but we use the ratios between peak heights rather then the relative height of the isotopic peaks. Also note that Gay et al. do not use Averagine to build theoretical peptides, but instead use a local average of all peptides within a 1 Da mass range after an in silico tryptic digest of an entire protein sequence database.

In the proposed approach, we use an average amino acid u, which is equal to Averagine, but does not include information on sulphur:

$$u = C_{4.9384}H_{7.7583}O_{1.4773}N_{1.3577}. \tag{6}$$

This average amino acid is used for the construction of three sets of theoretical peptides, namely:

- no-sulphur-containing set:

$$H - \text{round}(u \ldots u) - OH - H^+, \tag{7}$$

- one-sulphur-containing set:

$$H - \text{round}(u \ldots u) - S_1 - OH - H^+, \tag{8}$$

- two-sulphurs-containing set:

$$H - \text{round}(u \ldots u) - S_2 - OH - H^+. \tag{9}$$

Note that the three sets of theoretical peptides correspond to the three possible sulphur abundances observed in the dataset. Further, instead of working with a rounded average amino acid, as in Breen et al., we

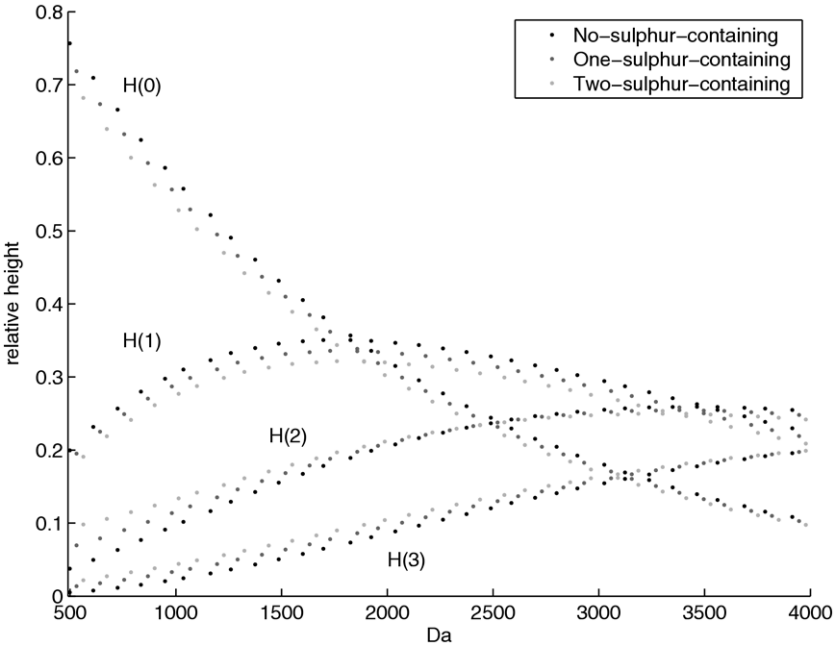

Figure 2. The relative height $H(i)$ of the peaks of the first four isotopic variants ($i = 0,\ldots,3$) of the theoretical peptides.

round after taking a multiple of average amino acids. This is done to prevent accumulation of rounding errors. In this context, rounding means that the elements of the average amino acid are rounded to the nearest integer after multiplication of the elements of u.

Each set consists of 32 theoretical peptides. The peptide with the lowest mass is formed with four average amino acids u and has a monoisotopic mass m equal to 498.257(S_0) Da, 530.228(S_1) Da, 562.200(S_2) Da, respectively, while the peptide with the highest mass is formed with 35 multiples of u resulting in masses 3915.040(S_0) Da, 3947.0125(S_1) Da, and 3978.985(S_2) Da, respectively. For all the theoretical peptides, the isotopic distribution is calculated using the Isotopic Pattern Calculator. The result is displayed in Figure 2 where, for each of the three sets of theoretical peptides, the dots indicate the relative heights of the peaks for the first four isotopes of the isotopic distribution.

Instead of working with relative intensities, however, we build a model for the ratios of subsequent peaks in the isotopic distribution. The rationale behind this decision is 3-fold:

- Ratios are dimensionless and their use allows to avoid scaling of the expected and observed intensity distribution.
- Ratios are not sensitive to multiplicative noise.
- Subsequent ratios produce smaller errors than ratios with a common reference, and will be explained at a later stage.

The $(x + 1)$th ratio R is calculated from the heights of subsequent isotopic peaks H with monoisotopic mass m as

$$R(x+1,m) = \frac{H(x+1,m)}{H(x,m)} \quad \text{with} \quad x = 0,\ldots,n, \qquad (10)$$

where $H(0,m)$ is the height of the monoisotopic peak. The first three ratios computed from the three sets of theoretical peptides are shown in Figure 3. The relation between the three subsequent ratios and the monoisotopic mass of the theoretical peptides can be easily modeled by the following polynomial model

$$R(x+1,m) = \beta_0 + \beta_1\left(\frac{m}{1000}\right) + \beta_2\left(\frac{m}{1000}\right)^2 + \beta_3\left(\frac{m}{1000}\right)^3 + \beta_4\left(\frac{m}{1000}\right)^4, \qquad (11)$$

where the monoisotopic mass m is divided by 1000 to limit the magnitude of the β parameters. The order of the model was chosen empirically by assessing the improvement in the adjusted coefficient of determination with respect to the addition of an extra parameter.

The parameters $\beta_0,\ldots,\beta_4$ of the polynomial model are estimated using the least-squares method, separately for each of the three sets of theoretical peptides. Note that in this way the nonlinear optimization for the Poisson mapping, proposed by Breen et al. and indicated in eq 4, is replaced by a simple and fast least-squares solution.

The estimated parameters for the three first ratios of the peaks in the isotopic distribution are presented in Table 1, separately for each set of the theoretical peptides. The fitted models are displayed as solid lines in Figure 3.

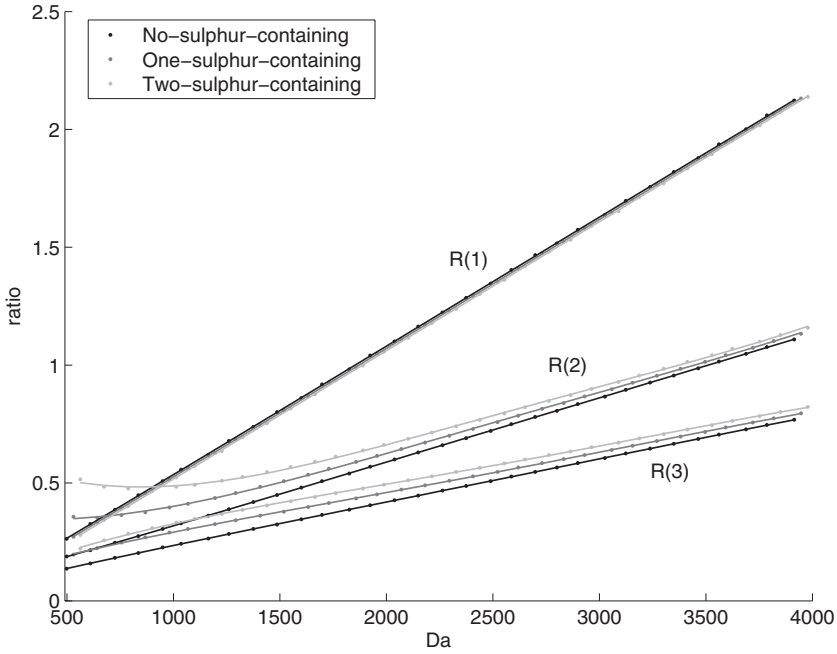

Figure 3. Fit of the model in eq 11 to the subsequent ratios $R(i)$ of the theoretical peptides with i = 1,. . .,3.

Although the first four isotopic peaks always capture at least 79% of the available isotopic information between a range of 500 and 4000 Da, it can be useful, especially for heavier peptides, to have information about the fifth, sixth, and seventh isotopic peak. Therefore, for masses starting at 1000, 1250, and 1500 Da, extra information is included in the model for the fourth, fifth, and sixth isotopic ratio, respectively. The parameters for these extra ratios are displayed in Table 2.

It should be stressed that this model is only valid for the prediction of isotopic ratios for peptides with a monoisotopic mass inside the intervals specified in Tables 1 and 2.

In a practical application, by comparing the ratios between a series of peaks observed in a spectrum with the ratios predicted from the proposed model (eq 11), and using the parameter estimates from Table 1, we can decide whether the series of peaks might be generated by a peptide. The choice of an appropriate statistic for this comparison is not trivial. We propose the Pearson χ^2 statistic because the variability seems to depend on the value of the ratio. The Pearson χ^2 statistic is calculated as:

Table 1. Estimated parameters of the model for ratios $R(1,m)$, $R(2,m)$, and $R(3,m)$, for the indicated mass ranges

β	$R(1)$	$R(2)$	$R(3)$
No-sulphur-containing peptides, mass range: 498–3915 Da			
β_0	−0.00142320578040	0.06258138406507	0.03092092306220
β_1	0.53158267080224	0.24252967352808	0.22353930450345
β_2	0.00572776591574	0.01729736525102	−0.02630395501009
β_3	−0.00040226083326	−0.00427641490976	0.00728183023772
β_4	−0.00007968737684	0.00038011211412	−0.00073155573939
One-sulphur-containing peptides, mass range: 530–3947 Da			
β_0	−0.01040584267474	0.37339166598255	0.06969331604484
β_1	0.53121149663696	−0.15814640001919	0.28154425636993
β_2	0.00576913817747	0.24085046064819	−0.08121643989151
β_3	−0.00039325152252	−0.06068695741919	0.02372741957255
β_4	−0.00007954180489	0.00563606634601	−0.00238998426027
Two-sulphur-containing peptides, mass range: 562–3978 Da			
β_0	−0.01937823810470	0.68496829280011	0.04215807391059
β_1	0.53084210514216	−0.54558176102022	0.40434195078925
β_2	0.00580573751882	0.44926662609767	−0.15884974959493
β_3	−0.00038281138203	−0.11154849560657	0.04319968814535
β_4	−0.00007958217070	0.01023294598884	−0.00413693825139

Table 2. Estimated parameters of the model for ratios $R(4,m)$, $R(5,m)$, and $R(6,m)$, for the indicated mass ranges

β	$R(4)$	$R(5)$	$R(6)$
No-sulphur-containing peptides, mass range:			
	907–3915 Da	1219–3915 Da	1559–3915 Da
β_0	−0.02490747037406	−0.19423148776489	0.04574408690798
β_1	0.26363266501679	0.45952477474223	−0.05092121193598
β_2	−0.07330346656184	−0.18163820209523	0.13874539944789
β_3	0.01876886839392	0.04173579115885	−0.04344815868749
β_4	−0.00176688757979	−0.00355426505742	0.00449747222180
One-sulphur-containing peptides, mass range:			
	939–3947 Da	1251–3947 Da	1591–3947 Da
β_0	0.04462649178239	−0.20727547407753	0.27169670700251
β_1	0.23204790123388	0.53536509500863	−0.37192045082925
β_2	−0.06083969521863	−0.22521649838170	0.31939855191976
β_3	0.01564282892512	0.05180965157326	−0.08668833166842
β_4	−0.00145145206815	−0.00439750995163	0.00822975581940
Two-sulphur-containing peptides, mass range:			
	971–3978 Da	1283–3978 Da	1623–3978 Da
β_0	0.14015578207913	−0.02549241716294	−0.14490868030324
β_1	0.14407679007180	0.32153542852101	0.33629928307361
β_2	−0.01310480312503	−0.11409513283836	−0.08223564735018
β_3	0.00362292256563	0.02617210469576	0.01023410734015
β_4	−0.00034189078786	−0.00221816103608	−0.00027717589598

$$e = \sum_{x=1}^{3} \frac{[R_E(x,m) - R_O(x,m)]^2}{R_E(x,m)}, \tag{12}$$

with R_E denoting the predicted and R_O the observed subsequent ratios. The smaller the error e, the better the agreement between the observed and predicted ratios, and the more likely that the series of peaks are generated by a peptide.

It is difficult to give a general strategy to choose a threshold for the error e to distinguish between a series of noise peaks and peptide peaks. This is because of the diversity of mass spectrometers and the possible settings to conduct an MS experiment. Therefore, before applying this method, one should conduct an experiment to determine an optimal threshold for peptide peak detection for the particular MS technique at hand.

As we argued earlier, subsequent ratios are preferable over the ratios with a common reference. Figure 4 shows the error e calculated with eq 12 for the 1542 no-sulphur containing human peptides using the proposed model (eq 11). The calculated errors in Figure 4a are obtained from ratios with the monoisotopic peak as a reference, while the errors in Figure 4b are obtained from subsequent ratios. Note that the only difference between the two results is the way the ratios were calculated. It can be observed that the errors from ratios with a reference peak are much larger than the errors from subsequent ratios. Normally, one would expect no difference, but a larger variability in the height of the monoisotopic peak, as can be seen from Figure 1, leads to larger errors, because an error in the first peak is always present in the three ratios.

Valkenborg et al. state that the method of Poisson approximation is not suited for sulphur-containing peptides [4]. This can be explained by the fact that, for the Poisson model, a ratio of subsequent peaks can be analytically expressed as

$$R(x+1,m) = \frac{P(x+1;m)}{P(x;m)} = \frac{e^{-M}M^{x+1}x!}{e^{-M}M^{x}(x+1)!} = \frac{M}{x+1}. \tag{13}$$

The previous formula (eq 13) indicates that the mean M of the Poisson model (eq 4) is equivalent to the first subsequent ratio $R(1,m)$. Note that Breen et al. found a linear relationship between M and m, as given in eq 5. From eq 13, it follows that the linearity is also implied for ratios $R(2,m)$ and $R(3,m)$. This actually coincides with the observations for no-sulphur-containing peptides, as can be seen in Figure 3.

On the other hand, from Figure 3 we can observe that the presence of sulphur atoms introduces a curvature in the relationship of the second and the third ratio with m. Consequently, the Poisson approximation used in eq 4 can be expected to perform worse for sulphur-containing peptides because the implied linearity no longer holds.

Results and Discussion

Figure 5a shows the error e between the observed subsequent ratios of the 2137 human peptides and the ratios predicted by the model (eq 11) for no-sulphur-containing peptides (for the coefficients, see Table 1). The effect of ignoring sulphur on the error is clearly seen, with sulphur-containing peptides showing large errors that decrease with the mass. This indicates that sulphur has a major effect on low mass peptides, which is logical, because the heavier the peptide, the more C,

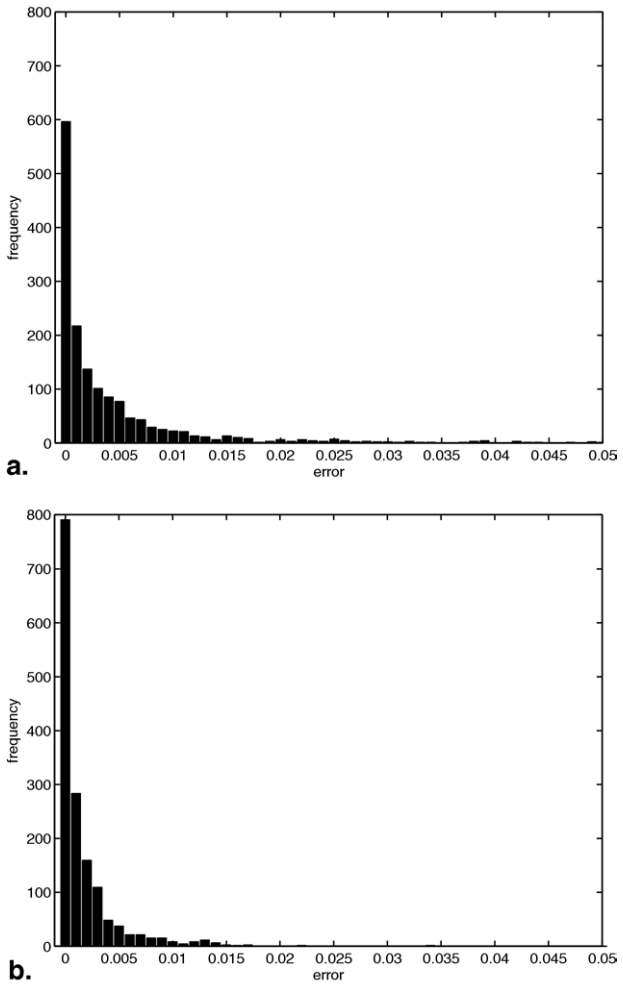

Figure 4. Comparison of the error distribution for different calculation methods for the ratio. (**a**) Distribution of errors e calculated via ratios with a reference. (**b**) Distribution of errors e calculated via subsequent ratios.

H, N, O atoms, and the less influential sulphur is on the isotopic distribution.

However, when the errors are computed using the model (eq 11) that accounts for the occurrence of sulphur atoms in the peptide, then the magnitude of the error is reduced, as shown in Figure 5b.

It should be noted that in this theoretical setting, the prevalence of sulphur in a peptide is known. Of course, in practice we do not know how many sulphur atoms might be present in a peptide that generates the peaks in a single mass spectrum. The following procedure can be considered:

- For a particular mass m, calculate three sets of predicted ratios, assuming different number of sulphur atoms present in the peptide, using the coefficients from Table 1.
- Compare each set of ratios with the observed ones using the Pearson χ^2 error statistic (eq 12).
- Choose the set of predicted ratios with the smallest error. If this error is smaller than a chosen threshold, conclude that the series of peaks is generated by a peptide with the corresponding number of sulphur atoms. For this particular case, a threshold of 0.05 is used.

Applying the classification strategy on the 2137 human peptides resulted in 10 misclassified peptides, which were distributed as follows: one peptide out of 1542 no-sulphur-containing peptides was misclassified as a one-sulphur-containing peptide, 8 peptides out of 502 one-sulphur-containing peptides were misclassified as no-sulphur-containing peptides, and one peptide out of 93 two-sulphur-containing peptides was misclassified as a one-sulphur-containing peptide.

The technique we developed is in the spirit of the methods proposed by Breen et al. and Gay et al. Therefore, it is of interest to check how these methods perform on our dataset. Figure 6a compares the predicted isotopic distribution obtained via the method of Breen et al. (dashed line) and the method of Gay et al. (dotted line) with the (true) isotopic distribution of the no-sulphur containing peptides. It can be observed that, for the method of Breen et al., the predicted relative height of the monoisotopic peaks fits quite well to the (true) height of the monoisotopic peaks of the no-1542 sulphur containing peptides. However, it fits less well the height of the other isotopic variants of the peptides. This can perhaps be explained by the fact that Breen et al. calculated the multinomial expansion by using Protein Prospector, which provides only an approximation of the complete multinomial expansion. Nevertheless, the relative height of a monoisotopic peak is easy to calculate and no bias should be observed for it. However, this bias is present in the isotopic distribution predicted by the method of Gay et al. A possible explanation for this is that Gay et al. incorrectly assumed that 100% of the isotopic ions is distributed over the first six isotopic peaks. Figure 6b displays the ratios of the heights of the subsequent isotopic peaks, predicted by the methods of Breen et al. (dashed line) and Gay et al. (dotted line) and the results of the application of the model (eq 11) (solid line).

Conclusions

For no-sulphur containing peptides, the prediction of the isotopic distribution by using Poisson approximation, as proposed by Breen et al., gives satisfactory results. However, as suggested by our results, the complex Poisson mapping, eq 4, can be replaced by directly modeling the relation between monoisotopic mass m and the first isotopic ratio $R(1,m)$. For sulphur-containing peptides, the prediction of the isotopic ratios with the polynomial model will correct for the presence of sulphur, as indicated in Figure 5b.

The results from the aforementioned classification strategy suggest that it is possible to predict the number

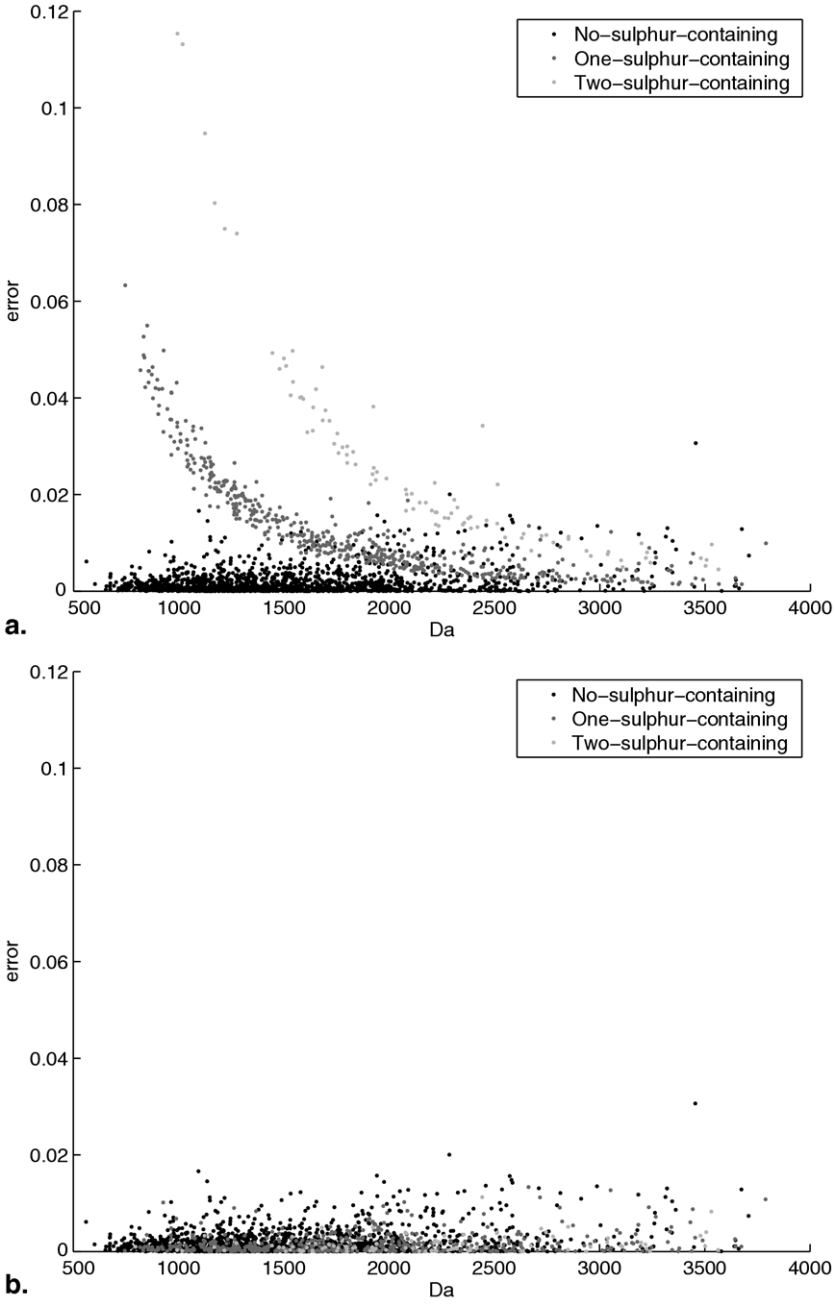

Figure 5. Error e between the observed subsequent ratios of the 2137 peptides and the ratios predicted by the model eq 11. (**a**) Shows the errors not accounting for the presence of sulphur. (**b**) Shows the error when accounting for the presence of sulphur.

of sulphur atoms in a peptide using the proposed model (eq 11). Note that this does not require an extra tandem MS. In this way, we could rapidly screen single mass spectra for sulphur-containing peptides. However, it should be underlined that these results are based on theoretical isotopic distributions computed via a multinomial expansion, and not on series of peaks observed in a spectrum. Thus, it does not reflect the extra influence of error in a mass spectrum.

The method developed in this manuscript accounts for the prevalence of one or two sulphur atoms in a peptide. This restriction was chosen based on the observed prevalence of sulphur atoms in the dataset of 2154 peptides available for the study. In principle, it is possible to extend the method in such a way that it accounts for the presence of three or more sulphur atoms in a peptide. This can be done by constructing an extra set of theoretical peptides with the required

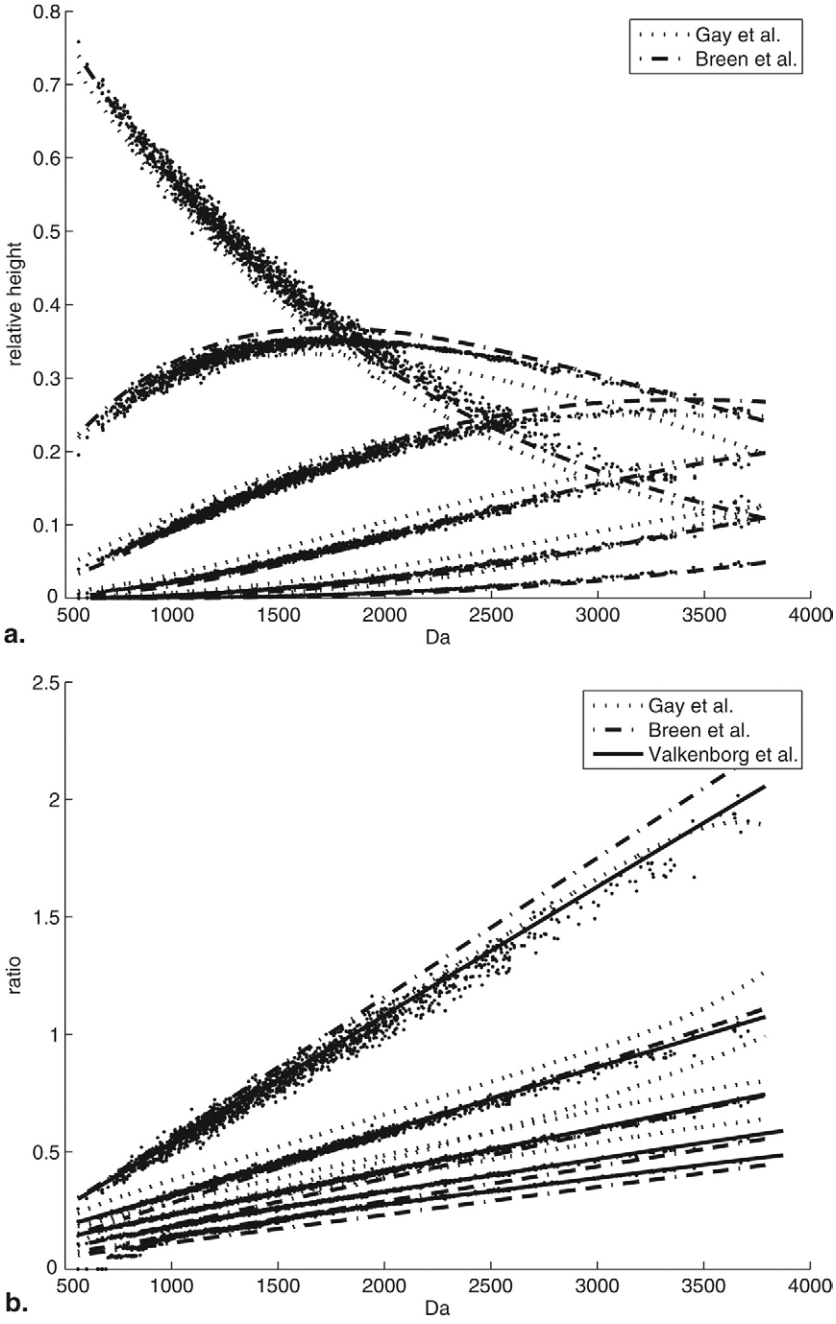

Figure 6. Observed and predicted (using the methods of Breen et al. and Gay et al.) heights (**a**) and ratios (**b**) of the isotopic variants of the 1542 no sulphur-containing peptides.

number of sulphur atoms and applying the model in eq 11 to them.

It should be noted that the developed model was validated with a set of human serum peptides. However, the construction of the theoretical peptides was done with a variant of Senko's Averagine, which was derived from the PIR protein database by using the statistical occurrences of all the amino acids. This implies that Averagine is a valid approximation for an average amino acid of all species. Therefore, the method proposed in our paper should be applicable not only to human serum peptides but also to other peptides. One could ask the question, however, whether the method could be improved by calculating an organism-specific (e.g., human) average amino acid using a subset of proteins from, e.g., the PIR protein database. In this respect, it is worth noting that by using the available set of 2154 peptides we found the following observed average amino acid:

$$C_{4.7810}H_{7.6042}O_{1.5079}N_{1.4067}S_{0.0238}. \quad (14)$$

The only discrepancy between eq 14 and Senko's Averagine (eq 1) lies in the relative average number of sulphur atoms, which is possibly related to the in vivo protein processing. Therefore, it is our conjecture that calculating an organism-specific average amino acid may not necessarily offer much improvement to the developed approach.

In principle, one might consider extending this method so that it can be used to detect other types of modifications of a peptide, e.g., phosphorylation, or a family of glycoproteins. However, this is only possible if the modification has a substantial effect on the isotopic distribution to discern it from an unmodified peptide.

Acknowledgments

The authors gratefully acknowledge financial support from the IAP research network no. P6/03 of the Belgian government (Belgian Science Policy). DV gratefully acknowledges support from Bijzonder Onderzoeksfonds Universiteit Hasselt (grant BOF04G01). The authors are grateful to the editor and reviewers for their insightful comments, which resulted in an improved manuscript. They also thank Pronota (Ghent, Belgium) for providing the data used in the study.

References

1. Yergey, J. A. A General Approach to Calculating Isotopic Distributions for Mass Spectrometry. *Int. J. Mass Spectrom. Ion Phys.* **1983,** *52,* 337–349.
2. Senko, M. W.; Beu, S. C.; McLafferty, F. W. Determination of Monoisotopic Masses and Ion Populations for Large Biomolecules from Resolved Isotopic Distribution. *J. Am. Soc. Mass Spectrom.* **1995,** *6,* 229–233.
3. Breen, E. J.; Hopwood, F. G.; Williams, K. L.; Wilkins, M. R. Automatic Poisson Peak Harvesting for High Throughput Protein Identification. *Electrophoresis* **2000,** *21,* 2243–2251.
4. Valkenborg, D.; Assam, P.; Krols, L.; Thomas, G.; Kas, K.; Burzykowski, T. Using a Poisson Approximation to Predict the Isotopic Distribution of Sulphur-Containing Peptides in a Peptide-Centric Proteomic Approach. *Rapid Commun. Mass Spectrom.* **2007,** *21,* 3387–3391.
5. Gay, S.; Binz, P.; Hochstrasser, D.; Appel, R. Modeling Peptide Mass Finger-Printing Data Using the Atomic Composition of Peptides. *Electrophoresis* **1999,** *20,* 3527–3534.
6. •. http://sourceforge.net/projects/isotopatcalc/
7. •. http://prospector.ucsf.edu/