

**Using a Poisson approximation to predict the isotopic distribution of sulphur-containing peptides in a peptide-centric proteomic approach.**

Journal:	<i>Rapid Communications in Mass Spectrometry</i>
Manuscript ID:	RCM-07-0361.R1
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	10-Aug-2007
Complete List of Authors:	Valkenburg, Dirk; Hasselt University, Center for Statistics Assam, Pryseley; Hasselt University, Center for Statistics Thomas, Grégoire; Pronota Krols, Luc; Pronota Kas, Koen; Pronota Burzykowski, Tomasz; Hasselt University, Center for Statistics
Keywords:	COFRADIC N-terminal peptides, Predicting the isotopic distribution, Sulphur contribution



view

## Abstract

Breen et al. (2000) proposed a method for finding monoisotopic peptide peaks in mass spectra based on an approximation of the distribution of different isotopic variants of a peptide by a Poisson distribution. They developed the method using all protein sequences from the SWISS-PROT database. We investigated the suitability of this method to predict the isotopic distribution in an environment which enriches for peptides carrying sulphur. More specifically, we focussed on mass spectra obtained by a COmbined FRActional DIagonal Chromatography (COFRADIC) approach, developed by Gevaert et al. (2003), targeting a specific subset of peptides, in this case the N-terminal peptides. One can therefore ask whether the original results of Breen et al. apply to spectra generated by the particular COFRADIC method. We investigate whether the proposed approximation holds for N-terminal peptides. We also evaluate whether ignoring sulphur atoms while developing the approximation, as proposed by Breen et al., does not increase the risk of missing monoisotopic peaks corresponding to sulphur-containing peptides. Finally, we check sensitivity of the quality of the approximation to optimization criteria used in the development process. The results are not simply restricted to a COFRADIC setting but are also applicable more generally, for any method which enriches for sulphur-containing peptides.

**Keywords:** COFRADIC N-terminal peptides, Predicting the isotopic distribution, Sulphur contribution.

## Introduction

In the prediction of isotope envelopes, sulphur atoms have been often discarded because of the low abundance of methionine and cysteine that carry them. However, several gel-free techniques enriches for peptides carrying sulphur. Some techniques isolate only cysteinyl peptides (e.g. ICAT [1]) or methionyl peptides, such that the problem of predicting isotopic distributions may even be more pronounced, because the isotopes of sulphur are unique.

In this manuscript we specifically focus of N-terminal COmbined FRActional DIagonal Chromatography (COFRADIC) [2] peptides, which targets only N-terminal enzymatic digested proteomic peptides. Despite the fact that most serum proteins will lose their initiator methionine upon secretion, we believe that the subset of N-terminal COFRADIC peptides is enriched for peptides carrying sulphur.

COFRADIC, introduced by Geveart *et al.*, is a novel suite of technologies for the identification of biomarkers in complex mixtures. The basic strategy of COFRADIC comprises the following steps: (1) cleavage of proteins to peptides: in this case by tryptic digest, (2) chromatographic fractionation of the complex peptide mixture, (3) chemical or enzymatic modification of a target subset of peptides to alter the column retention properties: in this case the hydrophobicity of internal peptides, (4) refractionation of the unmodified N-terminal peptides by a secondary automated chromatography, the modified internal peptides shift out of the original collection interval, (5) mass spectrometrical analysis of the N-terminal peptides for each fraction using matrixassisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOFMS).

The generated mass spectra are compared over different experimental conditions to search for possible biomarkers, with identification of the peptides/proteins obtained via MASCOT [3].

In order to process the massive amount of mass spectral data obtained in (5), we have developed an algorithm [4] using the method of monoisotopic peak validation based on Poisson modelling, proposed by Breen *et al.* [5]. By exploiting characteristic, theoretical features that should be associated with peaks corresponding to an isotopically resolved group, the method allows us to discriminate series of peaks possibly related to peptides from those generated by error.

In this manuscript, we evaluate the applicability of the method developed by Breen *et al.* to the N-terminal COFRADIC peptides. Further, we study whether algorithmic modifications of the method developed by Breen *et al.* lead to an improvement in the estimate of the isotopic distribution. Finally, we investigate whether the modified method could be used to predict the isotopic distribution of sulphur-containing peptides, as the method of Breen *et al.* excluded the effect of sulphur from their model.

## Experimental

### Material

A dataset of 1562 tandem MS sequenced N-terminal peptides (MASCOT score above 50 at 95% significance level) found in human serum was used to evaluate the method proposed by Breen *et al.* The isotopic distribution for these N-termini was calculated via polynomial expansion [6] with the software tool, Isotopic Pattern Calculator (IPC) [7].

The peptide molecular masses ranged between 741.8 and 3790.0. Of 1562 N-terminal tryptic peptides, 1059 (68%) did not contain any sulphur atom, 409 (26%) contained one sulphur atom and 80 (5%) contained two sulphur atoms. Only 14 N-termini out of 1562 contained three or more sulphur atoms; and these 14 peptides were removed from the data. Thus, finally, in our study a dataset of 1548 peptides was used.

Table 1 shows which sulphur-containing amino acids contribute the most to the pool of sulphur-containing peptides. It can be observed that methionine is more abundant than cysteine, as expected for N-terminal peptides.

### Method

The isotopic distribution for a peptide of any mass can be calculated using a polynomial expansion [6]. Breen *et al.* [5] approximate the result of the polynomial expansion (i.e. the expected proportional heights of different isotopic peaks) by a Poisson distribution. In order 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: Carbon (C), Hydrogen (H), Nitrogen (N), Oxygen (O), and Sulphur (S), as well as on the proportional abundance  $p$  of various isotopic variants. In practice, only the molecular monoisotopic weight of a peptide, 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)  $u = C_{10}H_{16}N_3O_3$  by averaging all AAs from all proteins in the SWISS-PROT protein database. Note that the average amino acid  $u$  is equivalent to the duplication of Averagine's elements (with rounding to the nearest integer) with

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

as reported by Senko *et al.* [8]. In  $u$ , however, the contribution of sulphur was ignored.

Multiples of the average AA, given by  $(H - i \times u - OH - H^+)$ , with  $i = 1, \dots, 15$ , were then used to span a mass range from  $m_1 = 245.1376$  to  $m_{15} = 3410.8059$ . For each so-constructed theoretical peptide the isotopic distribution  $E'$  was calculated using the Protein Prospector [9]. The mean  $M_i$  of a Poisson distribution  $E^*$  giving the best approximation to the theoretical isotopic distribution at  $m_i$  was then found by minimizing the sum of absolute deviations between the components of both

distributions:

$$M^*(m) = \operatorname{argmin}_M \sum_{x=1}^{\infty} \left| E'(0, m) \frac{P(x; M)}{P(0; M)} - E'(x, m) \right|, \quad (2)$$

with

$E'(x, m)$  : the  $x$ th isotope from a theoretical peptide with mass  $m$ ,

$$P(x; M) = \frac{e^{-M} M^x}{x!}.$$

As a result, a linear relationship between the Poisson mean  $M_i$  and the monoisotopic mass  $m_i$  was found

$$M_i = F(m_i) = a \times m_i + b, \quad (3)$$

which took the following form:

$$M_i = 0.000594m_i - 0.03091. \quad (4)$$

The relationship allows us 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 to decide whether the observed peaks can correspond to a series generated by a peptide.

Application of the approach developed by Breen *et al.* to mass spectra obtained using the COFRADIC technology raises some issues:

- In the considered variant of the COFRADIC-methodology, only N-terminal peptides are targeted. It is not clear whether the results obtained by Breen *et al.* from all protein sequences would apply to N-terminal peptides.
- Breen *et al.* found the mean  $M_i$  of a Poisson approximation by minimizing the sum of absolute deviations (2) between the components of the Poisson distribution and the distribution obtained via Protein Prospector. One could ask whether other criteria (e.g. squared deviations or Pearson's chi-squared deviations) would yield an improvement in the isotopic estimate.
- Due to the low abundance of sulphur-containing AAs, sulphur has been ignored in the average AA construction reported by Breen *et al.*. However, a low abundance of sulphur-containing AAs does not imply a low abundance of sulphur-containing peptides. The approximation of the isotopic distribution for peptides with sulphur might be inaccurate when ignoring sulphur. Thus, one may risk not detecting them in a mass spectrum.

## Results and discussion

To answer the first question, it is sufficient to check whether the elemental composition in Averagine proposed by Senko *et al.* corresponds to the average AA obtained from the

collection of N-terminal COFRADIC peptides. The average AA for the N-terminal peptides, excluding the extra acetyl group, is equal to

$$C_{4.7519}H_{7.5478}N_{1.3893}O_{1.5183}S_{0.0251} \quad (5)$$

The average AA is in close agreement with Averagine (1), that was constructed using the statistical occurrences of AAs from the PIR protein database. Hence, the method of Breen *et al.* is applicable to N-terminal COFRADIC peptides if we incorporate the extra acetyl group. It can be observed that the average abundance of sulphur in (5) is lower than in Averagine (1). This indicates that, in general, the presence of sulphur-containing peptides might be even higher than the 32% reported in the material section.

The set of theoretical peptides should now be constructed as

$$CH_3CO-[i \times \text{Averagine} \setminus S]-OH-H^+ \quad (6)$$

Note, that Averagine is used without the sulphur component,  $\text{Averagine} \setminus S$ . The square brackets in (6) indicate that the elements of the average amino acid are rounded to the nearest integer after multiplication of the elements of Averagine by  $i$ .

Breen *et al.* developed their method via a non-linear mapping as shown in equation (2) of the Poisson mean on the theoretical isotopes obtained via Protein Prospector. To investigate whether a change in the method of fitting of the Poisson distribution would improve the prediction of the isotopes, we looked at different optimization criteria. In particular, we investigated the use of squared deviations

$$M^*(m) = \operatorname{argmin}_M \sum_{x=1}^{\infty} \left[ E'(0, m) \frac{P(x; M)}{P(0; M)} - E'(x, m) \right]^2 \quad (7)$$

and Pearson chi-squared deviations

$$M^*(m) = \operatorname{argmin}_M \sum_{x=1}^{\infty} \frac{\left[ E'(0, m) \frac{P(x; M)}{P(0; M)} - E'(x, m) \right]^2}{E'(x, m)} \quad (8)$$

Figure 1 shows that the relationship between the Poisson mean and the mass of the theoretical peptides (6) for the different mappings is virtually the same. This means that using a different optimization criteria does not lead to different predictions of the isotopic distribution. As there is no obvious preference for any optimization criteria, we continue using the Poisson mapping with absolute deviations as specified in (2).

As mentioned earlier, Breen *et al.* ignored sulphur in the construction of theoretical peptides because sulphur is a low abundance atom in Averagine. This is correct, as only methionine and cysteine contain a sulphur atom. However, this does not imply that peptides, which contain methionine or cysteine, are of low abundance. This is especially true for the aforementioned gel-free techniques, which focus on sulphur-containing peptides. For instance, 32% of the N-terminal COFRADIC peptides (in our case study) contain a sulphur atom. If sulphur does not affect the isotopic distribution of a peptide, the Poisson approximation developed by Breen *et al.* can be safely used. To assess the effect of sulphur on the isotopic distribution, we calculated the Pearson chi-squared errors  $X^2 = \sum_{i=1}^N (O_i - E_i)^2 / E_i$ , with  $O_i$  the observed and  $E_i$  the expected relative isotope intensity, respectively, for the comparison of the isotopic distribution

predicted using the Poisson approximation and the theoretical distribution returned by IPC, for the set of 1548 N-terminal COFRADIC peptides. The influence of sulphur on the isotopic distribution is presented in Figure 2, where sulphur-containing peptides yield larger  $X^2$  errors. In particular, sulphur-containing peptides with a low molecular mass exhibit much larger errors than their sulphur-free counterparts. This suggests that we should account for the occurrence of sulphur by adjusting the Poisson model. To this aim, different sets of theoretical peptides can be generated for:

- Peptides with one sulphur atom:



- Peptides with two sulphur atoms:



For these sets of theoretical peptides, the isotopic distribution can be calculated using IPC. Next, the principle of the Poisson mapping can be applied to the IPC result. However, the Poisson mapping becomes problematic for sulphur-containing peptides. In Figure 3, it can be observed that the errors from a Poisson mapping obtained using absolute deviations (2) are systematically larger when sulphur is present. This indicates that the Poisson model has difficulty in approximating the theoretical distribution of sulphur-containing peptides. Another illustration of this difficulty can be found in Figure 4, where Poisson means  $M$  (obtained via absolute deviation Poisson mapping) are shown for the set of theoretical peptides containing two sulphur atoms. The linear relationship between monoisotopic mass  $m$  and Poisson mean  $M$ , reported by Breen *et al.*, is replaced by an apparent piecewise linear relationship. This means that the linear relation present in Figure 1 breaks apart in three different line fragments in Figure 4 for two sulphur-containing peptides.

By the nature of the COFRADIC procedure, cysteines are alkylated with iodoacetamide to S-carbamoylmethylcysteine. In this respect, equations (9) and (10) should ideally account for the number of alkyl groups present in the peptide due to the number of cysteines. The effect of alkylation on the isotopic distribution is an interesting research question. However, it is omitted from this study because we cannot discriminate between sulphur atoms coming from a cysteine or from a methionine. Normally, the increase in molecular weight due to the presence of the extra alkyl groups will adjust the prediction of the isotopic distribution. For example, an alkylated cysteine has two extra carbons, hydrogens, and oxygens in its atomic composition, with a total weight of 58.0361 Da. The Scaling Averagine according to the additional mass corresponds to 2.5792 carbon, 4.0519 hydrogen, 0.7091 nitrogen, 0.7715 oxygen and 0.0218 sulphur atoms. Although, this is not the exact atomic composition of the extra alkyl-group, it at least gives a coarse correction.

## Conclusion

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

In this paper we evaluated the suitability of the Poisson approximation proposed by Breen *et al.* for detecting valid peptide peaks in a sample enriched for peptides carrying sulphur, in this case, COFRADIC N-terminal data. The average amino acid, derived from the frequencies of occurrence of amino acids in an N-terminal data set, is in agreement with the Average reported by Senko *et al.*. This means that the atomic composition of an AA is similar for N-terminal peptides and for peptides retrieved from the PIR protein database. Hence, the method of Breen *et al.* can be used for N-terminal COFRADIC peptides when accounting for the extra acetyl group.

Modification of the optimization criteria (2), did not change the quality of the prediction. Therefore, the absolute deviation is a reasonable optimization criterion.

Due to the low occurrence of sulphur in Average, sulphur was ignored in the construction of theoretical peptides proposed by Breen *et al.*. However, a low abundance of sulphur-containing AAs is not equivalent to a low occurrence of sulphur-containing peptides. In our dataset 32% of the peptides contained at least one sulphur atom. The presence of sulphur has an important influence on the isotopic distribution of a peptide. Therefore, the presence of sulphur should not be ignored. One way to deal with the issue is to obtain a separate Poisson approximation for peptides with one, two, or more sulphur atoms. Unfortunately, it seems that the Poisson assumption might be problematic for sulphur-containing peptides (as seen in Figures 3 and 4). Therefore, caution should be applied when using the method proposed by Breen *et al.* to such peptides. An alternative approach is a topic of further research.

Finally, we stress that the described method is only suited for the detection of unlabeled peptides. Often relative peptide quantification is done by labeling peptides with a stable, naturally low abundance isotope. This labeling causes a complex mixing of isotopic distributions. Methods of dealing with such data are a topic of further research.

## References

- [1] Gygi S, Rist B, Gerber S, Turecek F, Gelb M, Aebersold R. *Nature Biotechnology* 1999; **17**: 994.
- [2] Gevaert K, Goethals M, Martens L, Van Damme J, Staes A, Thomas G, Vandekerckhove J. *Nature Biotechnology* 2003; **21**: 566.
- [3] Perkins D, Pappin D, Creasy D, Cottrell J. *Electrophoresis* 1999; **20**: 3551.
- [4] Valkenburg D, Burzykowski T, Krols L, Thomas G, Kas K. *The Tenth Annual International Conference on Research in Computational Biology: Poster Abstract* 2006.
- [5] Breen E, Hopwood F, Williams K, Wilkins M. *Electrophoresis* 2000; **21**: 2243.
- [6] McCloskey J. *Methods in Enzymology* 1990; **193**: 882.

1  
2  
3  
4 [7] URL: <http://sourceforge.net/projects/isotopatcalc/> Accessed on 31 May, 2006.  
5  
6

7 [8] Senko M, Beu S, McLafferty F. *J. Am Soc Mass Spectrom* 1995; **6**: 229.  
8  
9

10 [9] URL: <http://prospector.ucsf.edu/> Accessed on 8 November, 2006.  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For Peer Review

## Acknowledgment

Financial support from the IAP research network nr P6/03 of the Belgian government (Belgian Science Policy) is gratefully acknowledged by the first two and the last author.

The first author gratefully acknowledges support from Bijzonder Onderzoeksfonds Universiteit Hasselt (grant BOF04G01).

We are grateful to the reviewers for their insightful comments which resulted in an improved manuscript.

For Peer Review

## Tables

		Methionine				
	number	0	1	2	3	4
Cysteine	0	1059	288	46	8	1
	1	121	27	3	0	0
	2	7	0	0	0	0
	3	2	0	0	0	0

Table 1: Distribution of methionine/cysteine over the pool of 1562 peptides

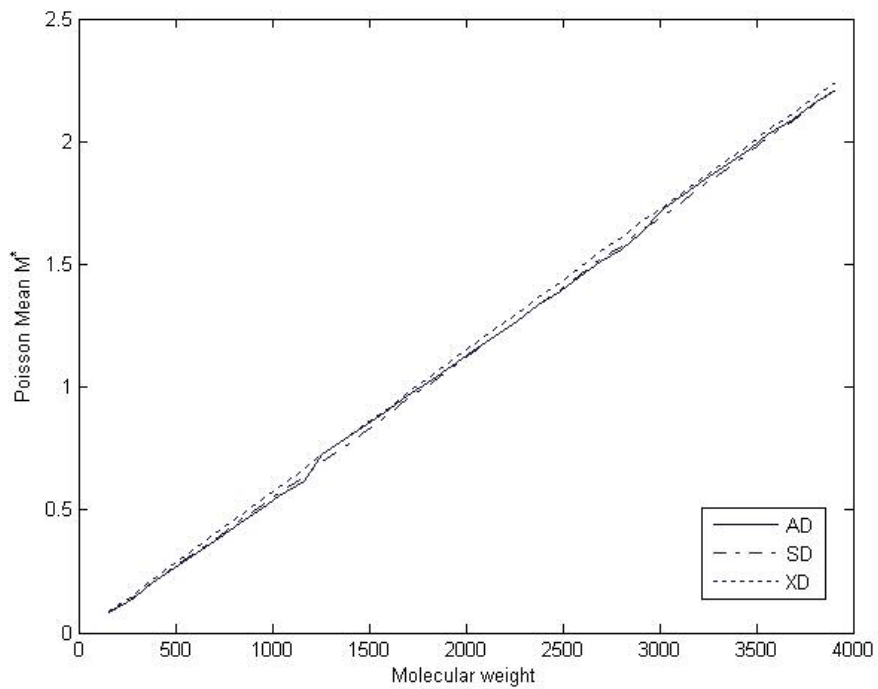
## Legends

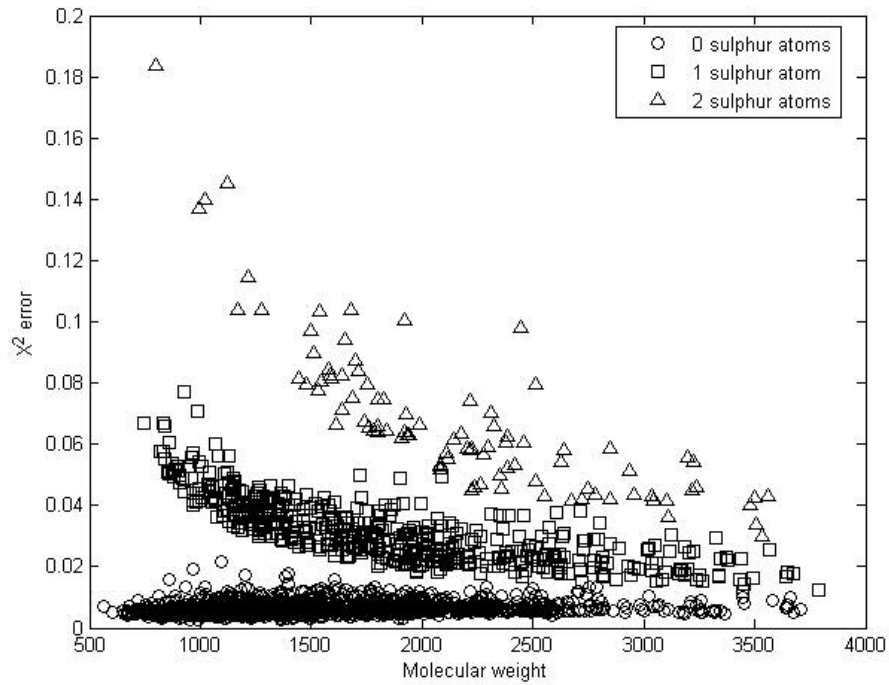
Figure 1: Optimal Poisson mean after mapping the Poisson function to the theoretical isotopic distribution. The solid line indicates the result with absolute deviation (AD) as the optimization criterion, while the dashed and dotted lines represent the squared deviation and Pearson's chi squared deviation, respectively.

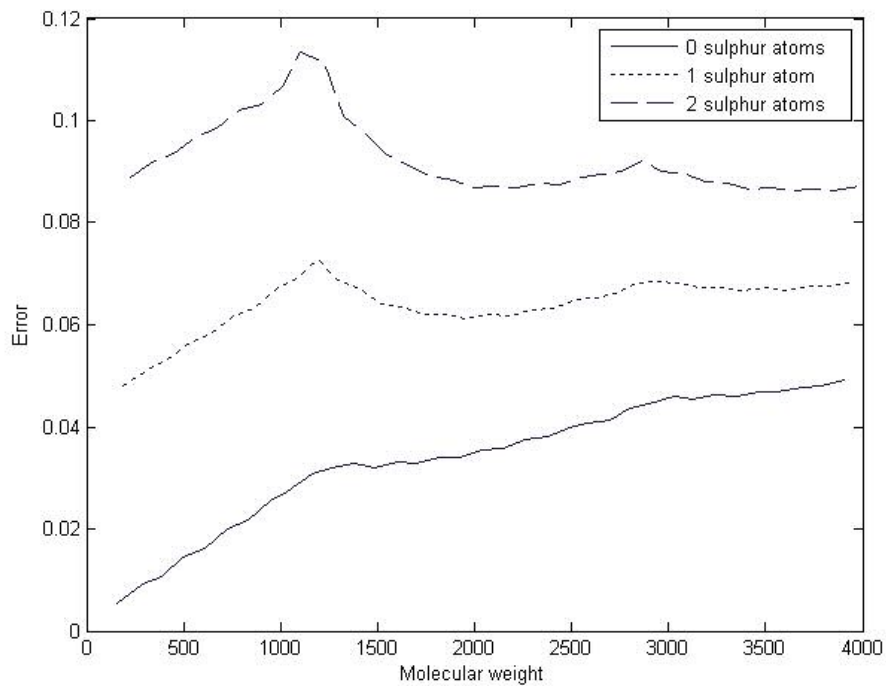
Figure 2: Pearson's chi-squared error  $X^2$  for the theoretical isotopic distribution compared with the approximation obtained via the Poisson distribution of the 1548 N-terminal COFRADIC peptides. The circles represent the errors for sulphur-free peptides. The errors for one sulphur atom and two sulphur atom-containing peptides are indicated by boxes and triangles, respectively.

Figure 3: The minimum error of the absolute deviation cost function for the Poisson mapping.

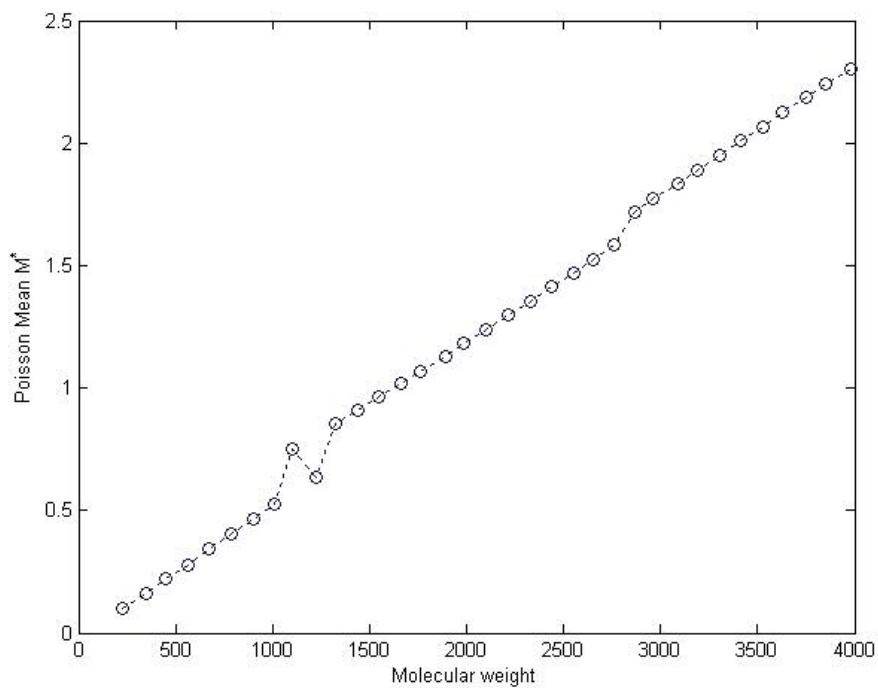
Figure 4: Poisson mean for the approximation of the theoretical isotopic distribution for a two sulphur atoms containing peptide. The results are obtained with the sum of absolute deviations as the optimization criterion.







Review



Review