An Article Submitted to

Statistical Applications in Genetics and Molecular Biology

Manuscript 1538

A Markov-Chain Model for the Analysis of High-Resolution Enzymatically ^{18}O-Labeled Mass Spectra

Dirk Valkenborg* Tomasz Burzykowski†

*VITO - Flemish Institute for Technological Research, dirk.valkenborg@vito.be

†Hasselt University, tomasz.burzykowski@uhasselt.be

Copyright ©2010 Berkeley Electronic Press. All rights reserved.

A Markov-Chain Model for the Analysis of High-Resolution Enzymatically ^{18}O-Labeled Mass Spectra*

Dirk Valkenborg and Tomasz Burzykowski

Abstract

The enzymatic ^{18}O-labeling is a useful quantification technique to account for between-spectrum variability of the results of mass spectrometry experiments. One of the important issues related to the use of the technique is the problem of incomplete labeling of peptide molecules, which may result in biased estimates of the relative peptide abundance. In this manuscript, we propose a Markov-chain model, which takes into account the possibility of incomplete labeling in the estimation of the relative abundance from the observed data. This allows for the use of less precise but faster labeling strategies, which should better fit in the high-throughput proteomic framework. Our method does not require extra experimental steps, as proposed in the approaches developed by Mirgorodskaya et al. (2000), López-Ferrer et al. (2006) and Rao et al. (2005), while ~~including~~ the model proposed by Eckel-Passow et al. (2006) as a special case. The method estimates information about the isotopic distribution directly from the observed data and is able to account for biases induced by the different sulphur content in peptides as reported by Johnson and Muddiman (2004). The method is integrated in a statistically sound framework and allows for the calculation of the errors on the parameter estimates based on model theory. In this manuscript, we describe the methodology in a technical matter and assess the properties of the algorithm via a thorough simulation study. The method is also tested on a limited dataset; more intense validation and investigation of the operational characteristics is being scheduled.

KEYWORDS: quantitative proteomics, high-resolution mass spectrometry, stable isotope labeling, ^{18}O-labeling, Markov-chain model

---

*Dirk Valkenborg, VITO - Flemish Institute for Technological Research. Tomasz Burzykowski, I-BioStat, Hasselt University. Financial support from the IAP research network nr. P6/03 of the Belgian government (Belgian Science Policy) is gratefully acknowledged by both authors. The first author also acknowledges support from Bijzonder Onderzoeksfonds Universiteit Hasselt (grant BOF04G01). The authors are grateful to the editor and the reviewers for their insightful comments, which resulted in an improved manuscript.

1 Introduction

Peptide-centric techniques are gaining a lot of interest for the search of new protein biomarkers, surrogate endpoints, or markers for classification of diseases. Typically, such techniques extensively use liquid chromatography (LC) combined with mass spectrometry (MS) for protein-expression profiling, because they promote the high-throughput quantitative characterization of a proteome. By comparing the protein abundances between different samples, differentially expressed proteins can be found. By analyzing these proteins, important information about, e.g., mechanisms of disease can be obtained. However, the LC-MS measurements are influenced by different sources of variability, which can obstruct the detection of differentially expressed proteins. In order to reduce the effect of the variability on the data, a labeling approach can be considered. There are two generally accepted labeling strategies for the relative quantification of proteins.

The first strategy is based on the principle of isobaric labeling, i.e., the reporter protocol. In this approach, peptides from multiple samples are coded with isobaric mass tags (e.g., iTRAQ from Applied Biosystems, TMT from Proteome Sciences, ExacTag from Perkin Elmer, etc.) and are mixed together. Peptides labeled with different tags are indistinguishable in a precursor scan. To quantify the relative abundance of a peptide in the labeled samples, an additional tandem MS interrogation is required. Quantification of the relative abundance is based on the observed intensities of the reporter molecules.

The second strategy is based on isotopic labeling, i.e., the precursor protocol. In this approach, peptides are coded with stable isotope tags (e.g., ICPL from TopLab, ICAT from Applied Biosystems, etc.) and mixed together with an unlabeled sample. The stable isotope tag will result in an increase of the peptide's mass. Due to this increased mass, a peptide from the labeled sample is discernable from its unlabeled counterpart in a precursor scan. Quantification of the relative abundance is based on the observed intensities.

Isobaric labeling has several advantages over the isotopic labeling strategy. For instance, it allows for multiplexing up to eight samples in one LC-MS run. The quantification results are also complemented by the identification of the putative peptide. However, a disadvantage of the isobaric labeling strategy is that for the quantification of a peptide a tandem MS fragmentation is required. The selection of these peptides is data dependent and limited by the available amount of material and instrument acquisition. Usually only the most abundant peptides are selected for tandem MS and often undersampling occurs (Li *et al.*, 2005). It is also well possible that uninteresting peptides, which are not differentially expressed, are selected.

For these reasons, we regard the isobaric labeling as highly reliable and excellent for screening samples containing only a limited number of peptides. This

makes the technology suited for the hypothesis-driven evaluation of differentially expressed proteins. However, when the scope is hypothesis-generation, i.e., discovery by analyzing complex proteomes, this approach is not optimal. For the screening of whole proteomes for differentially expressed proteins it would be convenient if a thorough data analysis could indicate the peptides, which are differentially expressed prior to a second interrogation on the tandem MS. This targeted approach should improve the dynamic range and sensitivity of the method, and should lead to a more efficient use of the mass spectrometry device.

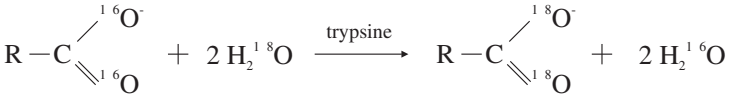


Figure 1: Chemical reaction scheme for the two-step enzymatic ^{18}O-labeling procedure.

In this respect, a relatively low-cost and open-source technique for stable isotope labeling is the enzymatic ^{18}O-labeling, where the two ^{16}O oxygen atoms in the carboxyl-terminus of a peptide are replaced with oxygen isotopes from heavy-oxygen-water. By putting emphasis on the enzymatic aspect of the labeling, we want to distinguish between the method considered in this manuscript and the commonly used proteolytic labeling. Enzymatic labeling is performed in two steps. In the first step, protein digestion is done in normal water. In the second step, the labeling is done in heavy-oxygen-water. The oxygen replacements in the carboxyl-terminus are a continuous process and are enzymatically catalyzed by a proteolytic reagent. This labeling reaction is schematically depicted in Figure 1. In principle, this reaction is the reverse of the protease-catalyzed peptide-bond (Miyagi and Rao, 2007). We assume that both oxygen-atoms on the carboxyl-terminus will react equally favorable with the ^{18}O-atoms. Hence, in ideal circumstances, the labeling should lead to an increase of the mass of the peptide molecule by 4 dalton (Da)[1].

For example, the labeled peptides from, say, Sample II, can now be pooled together with the unlabeled peptides from, say, Sample I, and processed simultaneously by LC and MS. Without the enzymatic ^{18}O-labeling, the isotopic peaks, corresponding to the isotopic distribution[2] of a peptide present in both samples, would appear at the same location in the resulting *joint mass spectrum*. This is

[1]The term dalton (Da) is a mass unit and is defined as $1/12$ of the mass of an unbound atom of ^{12}C at rest and in its ground state.

[2]A peptide molecule has different isotopic variants with a different mass because of the presence of carbon (^{12}C, ^{13}C), hydrogen (^{1}H, ^{2}H), nitrogen (^{14}N, ^{15}N), oxygen (^{16}O, ^{17}O, ^{18}O), and sulphur (^{32}S, ^{33}S, ^{34}S, ^{36}S) isotopes. The probability of occurrence of the variants can be computed and

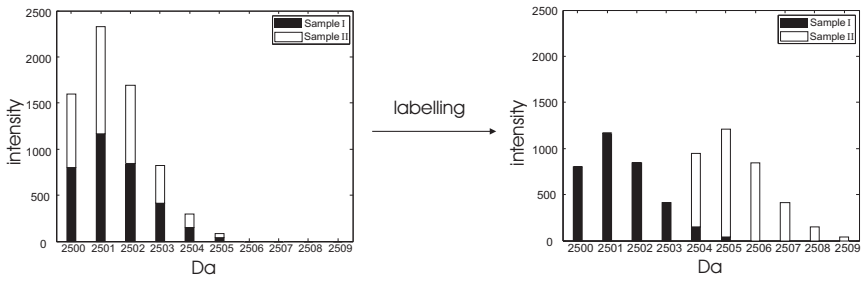


Figure 2: Effect of enzymatic ^{18}O-labeling in a mass spectrum in "stick" representation. Left panel: "sticks" can be seen as a representation of the distribution of the isotopic variants of the peptide. Right panel: labeling causes accumulation of different isotopic variants in a joint spectrum.

graphically illustrated in the left panel of Figure 2. In this situation, no distinction could be made between the contributions of the different biological samples to the peptide peaks observed in the joint spectrum (cfr. isobaric labeling). However, with the enzymatic ^{18}O-labeling, the isotopic peaks which correspond to the labeled peptide will shift 4 Da to the right in the mass spectrum, as shown in the right-hand side panel of Figure 2. This allows for making a distinction between the peaks related to peptides from different samples. Consequently, a direct comparison of the peptide abundance in the two samples is possible because the abundance measurements are affected by the same amount of machine noise. A "naïve" approach to compute the relative abundance of the peptide in the two samples would be to take the ratio of the heights of the first and fifth peak observed for the peptide in the joint mass spectrum (see the right-hand side panel of Figure 2), as these peaks would correspond to the monoisotopic variants of the peptide in the unlabeled and labeled sample, respectively. However, as it can be observed from Figure 2, some isotopic peaks of the unlabeled peptide will still overlap with the monoisotopic peak of the labeled peptide. Thus, even in this ideal setting, where a mass shift of 4 Da is acquired, the ratio would yield a biased estimate of the relative abundance, because it does not take into account the overlap of the isotopic peaks.

In practice, however, there are more problems related to the use of the enzymatic ^{18}O-labeling strategy. First, the heavy-oxygen water does not contain 100% pure ^{18}O-water. It can also contain ^{16}O- and ^{17}O-atoms. We term these *water impurities*. Note that, if the two carboxyl-terminus oxygen atoms are replaced by, e.g., ^{17}O-atoms, the peptide molecule becomes heavier by only 2, and not 4 Da, as it

is called the isotopic distribution. When assayed by high-resolution mass spectrometry a peptide produces a series of peaks, called isotopic peaks. The peaks are separated by approximately one mass-unit and their intensity correspond to the isotopic distribution of the peptide. The lightest isotopic variant of a peptide is called the monoisotopic variant.

ideally would be the case in 100% pure ^{18}O-water. Second, the speed of the enzymatic reaction, i.e. λ the oxygen incorporation rate, depends on multiple unobserved factors and therefore can differ for different peptides. As a result, at the end of the enzymatic reaction, not all peptide molecules from Sample II may have been actually labeled. The isotopic peaks for these molecules will overlap with the peaks from Sample I, which results in a biased estimate of the relative abundance.

These problems imply that the peaks, observed for a peptide in a joint spectrum, will correspond to a complex mixture of shifted and overlapping isotopic peaks that are related to the isotopic distributions of the peptide molecules in the unlabeled and labeled samples. In order to obtain an unbiased estimate of the relative abundance of the peptide in the two samples, the overlap of the isotopic peaks has to be taken into account (Ye *et al.*, 2009).

Several methods have been proposed to deal with the issue. On one hand, efforts aimed at the optimization of the labeling process, have been undertaken. For instance, methods that prohibit the back-exchange have been investigated (Storms *et al.*, 2006; Staes *et al*, 2004). Alternatively, techniques that only allow for the incorporation of a single ^{18}O-atom have been proposed (Rao *et al.*, 2005).

On the other hand, approaches that address the issue at the data analysis stage have been developed. Mirgorodskaya *et al.* (2000) have formulated a regression approach, which uses information about the isotopic distribution and about the labeling efficiency of the labeled peptide. The information is extracted from an additional mass spectrum of the labeled peptides, obtained before mixing the unlabeled and labeled sample. This extra MS step complicates the conduct of the experiment. López-Ferrer *et al.* (2006) and Rao *et al.* (2005) have suggested to identify the amino acid sequence of the peptide via an additional MS identification (tandem MS). Consequently, they can calculate the isotopic distribution of the peptide. The extra MS identification and the calculation of the isotopic distribution are computationally involved and require extra mass spectrometer time. Eckel-Passow *et al.* (2006) have proposed a regression approach similar in spirit to the method of Mirgorodskaya *et al.* (2000). They have used the method of Senko *et al.* (1995) to estimate the average isotopic distribution. This method is fast and does not need extra MS steps. However, it can lead to biased relative abundance estimates, as the actual isotopic distribution of a peptide can substantially deviate from the average isotopic distribution when, e.g., the peptide contains sulphur atoms (Johnson and Muddiman, 2004; Valkenborg *et al.*, 2007). Other methods treat the problem as a normalization issue similar to the one related to microarray data, but by doing so they ignore valuable information regarding the labeling processes.

In this manuscript, we rigorously describe an alternate, model-based approach to estimate the relative abundance of a peptide from enzymatically ^{18}O-labeled MS data. The approach uses the regression framework, considered by Mirgorodskaya

et al. and combines the framework with a probabilistic model, which describes the kinetics of the enzymatic ^{18}O-labeling reaction. An important advantage of the method is that it allows to estimate the peptide's isotopic distribution directly from the observed data, which in turn can be used to validate if the peaks are indeed originating from a bonafide peptide (Valkenborg *et al.*, 2008a). This implies that no additional MS steps are required for quantification, while the information is unbiasedly extracted from the observed spectra. The method is able to accommodate additional joint mass spectra for a given peptide, which can arise from, e.g., neighboring LC-fractions or technical replicates. Further, we extended the method such that it can account for the possible presence of ^{17}O atoms in the heavy-oxygen water. The proposed method is integrated into a sound statistical framework and the properties are thoroughly evaluated by means of a simulation study. A controlled MS experiment, limited to one commercially available purified protein, is conducted in order to demonstrate the correct functioning of the method on mass spectrometry data. More complex experiments are being set up in order to further investigate the operational characteristics of the method.

2 Methods

We assume that, prior to the statistical analysis of a series of peaks observed in a MALDI-TOF spectrum, the spectrum was appropriately pre-processed. To this aim, we use the strategy proposed by Valkenborg *et al.* (2009; 2008b). The pre-processing strategy extracts the information about the mass location and the height (intensity) of peaks, which are most likely due to a peptide. Thus, we represent the peaks in a mass spectrum by "sticks", disregarding their shape.

In what follows, we present the development of our approach. In the first subsection, we describe the basic model for peptide peaks observed in a joint mass spectrum obtained from an enzymatic ^{18}O-labeling experiment. The observed peaks are expressed in terms of the unobserved isotopic peaks in both samples before labeling. Unfortunately, the model is not practical, as it is over-parameterized. To address the issue, in the second subsection, we formulate a parsimonious model for the kinetics of the enzymatic reaction, which drastically reduces the number of parameters in the basic model and makes the latter estimable.

In developing the model, we assume availability of a single joint spectrum. However, in practice, multiple spectra, resulting from analysis of, e.g., several replicated measurements for the same or different biological samples will usually be available. Thus, in the third section, we discuss how the inclusion of multiple spectra can be handled via the construction of the log-likelihood function. Finally, in the last subsection, we discuss the issues related to the numerical algorithms used for the practical implementation of the model.

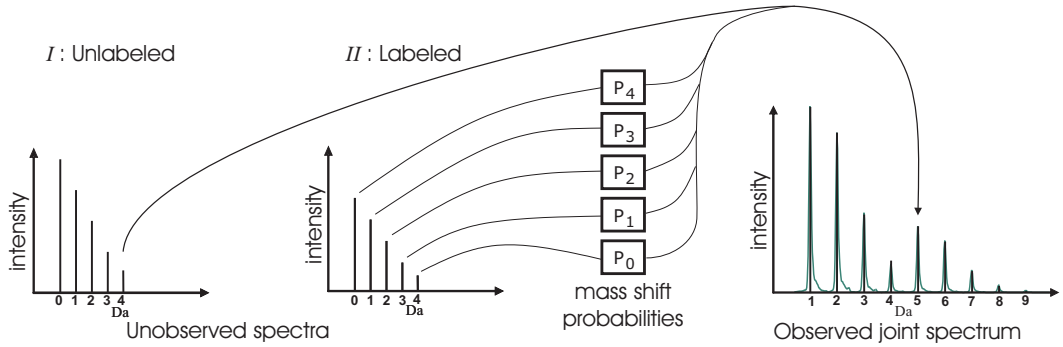


Figure 3: The height of the fifth peak of the observed joint spectrum can be defined in terms of the unobserved peptide peak intensities before labeling and the mass shift probabilities. Due to the imprecise labeling of a peptide, five potential mass shifts can occur, P_0, P_1, P_2, P_3 and P_4. In this way, the set of isotopic peaks from the labeled peptide can contribute to the fifth peak in the joint spectrum via the earlier defined mass shift probabilities.

In this manuscript, various assumptions are made during the construction of the model. In order to improve the understanding of the method, a guideline to evaluate the validity of the assumptions is added to the Appendix.

2.1 A model for the joint spectrum

As it was mentioned in the introduction, the heavy-oxygen water contains water impurities. We denote the proportions of ^{16}O, ^{17}O, and ^{18}O atoms in the heavy-oxygen water by p_{16}, p_{17}, and p_{18}, respectively, with $p_{16} + p_{17} + p_{18} = 1$. Due to water impurities, the carboxyl-terminus of a peptide can contain different isotopes of oxygen. Let us consider the triplet (n_{16}, n_{17}, n_{18}), where n_{16}, n_{17}, and n_{18} denote the number of ^{16}O, ^{17}O, and ^{18}O atoms in a carboxyl-terminus, respectively. Clearly, $n_{16} + n_{17} + n_{18} = 2$. The possible isotope combinations can now be expressed as follows:

$$\begin{aligned} X(1) &= (2,0,0), \quad X(3) = (1,0,1), \quad X(5) = (0,1,1), \\ X(2) &= (1,1,0), \quad X(4) = (0,2,0), \quad X(6) = (0,0,2), \end{aligned} \tag{1}$$

For example, configuration $X(3) = (1,0,1)$ indicates that one of the carboxyl-terminus oxygen atoms was replaced by an ^{16}O-atom, while the other was replaced by an ^{18}O-atom.

For different configurations in (1), peaks corresponding to the isotopic distribution of a labeled peptide will shift with multiples of 1 Da. The mass shift depends

on the configuration. The probability of a particular mass shift follows from the probability distribution of the six possible configurations of the carboxyl-terminus:

$$\begin{aligned} P_0 = P\{X(1)\}, \quad P_2 = P\{X(3)\} + P\{X(4)\}, \\ P_1 = P\{X(2)\}, \quad P_3 = P\{X(5)\}, \quad P_4 = P\{X(6)\}, \end{aligned} \tag{2}$$

where the index of the probability indicates the mass shift which ranges from 0 to 4 Da. It should be noted that we define the mass shifts relative to a carboxyl-terminus, which contains two ^{16}O-atoms. The probabilities $P_0, \ldots, P_4$ can also be interpreted as a neutron count, where the index denotes the number of additional neutrons due to the presence of oxygen isotopes in the carboxyl-terminus.

Figure 3 illustrates a single joint mass spectrum for a certain peptide. It presents how the fifth peak of the observed joint spectrum at the right-hand side is composed out of the unobserved isotopic variants of the peptide in Sample I and Sample II at the left-hand side before the labeling. Note that the first five isotopic variants of the peptide in Sample II (labeled sample) contribute to the fifth peak of the observed joint spectrum via the mass shift probabilities induced by the labeling. Hence, in order to estimate the relative abundance of the peptide in the two samples, we need to retrieve the information about the unobserved abundances of the isotopic variants before the labeling. To this aim, we propose a model, which expresses the m observed peak intensities y_j in the joint spectrum as a function of the abundance of the l unobserved isotopic variants of the peptide in Sample I and Sample II before the labeling. The function is parameterized in terms of the mass shift probabilities, defined in (2). Note that the observed peak intensities y_j in a joint mass spectrum are most likely also affected by instrument noise. Therefore, we need to consider a model that incorporates an error structure. Thus, we assume that

$$y_j = x_j + \varepsilon_j \ , \tag{3}$$

with $\varepsilon_j \sim N(0, \sigma^2)$ and that ε_j's are independent. A correlated, heteroscedastic error structure for the random error terms ε_j might also be plausible, but this leads to a more complex model and is a topic of further research. The index $j = 1, 2, \ldots, m$ denotes the position of the peak in the observed series of peaks in a joint spectrum (see Figure 3), with $j = 1$ referring to the first peak in the joint spectrum.

It should be noted that there is a special relation between the m observed peaks in a joint mass spectrum and the l unobserved isotopic variants. For example, consider a peptide, which has $l \geq 5$ isotopic variants[3] (including the monoisotopic

[3]A sensitive mass spectrometer is able to visualize up to five isotopic variants for a peptide. However, low-abundance and low mass peptides can have isotopic variants which fall under the limit of detection. The structure of the model can be easily adjusted to accommodate for this. Generally, the value of m ranges between 9 and 11.

variant). Based on this information, we can calculate the number of observed peaks expected in a joint mass spectrum. Enzymatic ^{18}O-labeling and mixing of such a peptide with its unlabeled counterpart will result in an observed joint spectrum of $m = l + 4$ peaks, due to the mass shift of 4 Da. Equivalently, in order to determine the structure of the system of equations in (5) for an observed joint spectrum, we need to specify $l = m - 4$ for a series of $m \geq 9$ peaks.

The mean intensity, x_1, of the first peak in the joint mass spectrum can now be expressed as

$$x_1 \;=\; H_0^I + P_0 H_0^{II} \;, \tag{4}$$

where H_0^I is the unobserved abundance of the monoisotopic variant in Sample I (unlabeled) and $P_0 H_0^{II}$ denotes the contribution of the monoisotopic variant from Sample II (labeled) before the labeling. It should be noted that P_0 indicates the probability that a peptide will not receive an isotope label, i.e, does not shift to a higher mass. Now, for the expected intensities of a series of peaks observed in the joint spectrum, we can write down such a decomposition, which describes how the unobserved isotopic variants of both samples before the labeling contribute to them:

$$\begin{aligned}
x_2 &= H_1^I + P_0 H_1^{II} + P_1 H_0^{II}, \\
x_3 &= H_2^I + P_0 H_2^{II} + P_1 H_1^{II} + P_2 H_0^{II}, \\
x_4 &= H_3^I + P_0 H_3^{II} + P_1 H_2^{II} + P_2 H_1^{II} + P_3 H_0^{II}, \\
&\vdots \\
x_{m-4} &= H_{l-1}^I + P_0 H_{l-1}^{II} + P_1 H_{l-2}^{II} + P_2 H_{l-3}^{II} + P_3 H_{l-4}^{II} + \\
&\quad P_4 H_{l-5}^{II}, \\
x_{m-3} &= P_1 H_{l-1}^{II} + P_2 H_{l-2}^{II} + P_3 H_{l-3}^{II} + P_4 H_{l-4}^{II}, \\
x_{m-2} &= P_2 H_{l-1}^{II} + P_3 H_{l-2}^{II} + P_4 H_{l-3}^{II}, \\
x_{m-1} &= P_3 H_{l-1}^{II} + P_4 H_{l-2}^{II}, \\
x_m &= P_4 H_{l-1}^{II}.
\end{aligned} \tag{5}$$

Terms $P_0, \ldots, P_4$ denote the contributions of the unobserved isotopic variants from Sample II to the observed peaks from the joint spectrum. The contributions depend on the mass shift probabilities, which were defined in (2). Note that, for the peaks $(m-3)$ to m, there are no contributions from the unobserved isotopic variant of the peptide in Sample I (unlabeled).

We can reduce the number of parameters, involved in (5), by exploiting the fact that the isotopic distribution of a peptide is the same for the two samples, disregarding the oxygen atoms in the carboxyl-terminus. It follows that the ratio of

abundance of any of the unobserved isotopic variants of the peptide should be the same in both samples. Thus, let us define the set of *isotopic ratios* as

$$R_i \;=\; \frac{H_i^I}{H_0^I} = \frac{H_i^{II}}{H_0^{II}}\;, \tag{6}$$

with $i = 1, \ldots, (l-1) = (m-5)$. The abundances of the isotopic variants can be written as a function of the isotopic ratios R_i and the abundances of the monoisotopic variants H_0^I and H_0^{II}. Obviously, the ratio R_0 is equal to one.

As we are mainly interested in the relative abundance of a peptide present in Samples I and II, we further reparameterize the abundances of the monoisotopic variants H_0^I and H_0^{II} as

$$H_0^I = H \qquad \text{and} \qquad H_0^{II} = HQ\;, \tag{7}$$

where H is called the reference intensity and $Q = H_0^{II}/H_0^I$ is the relative abundance. By combining (6) and (7) with (5), we obtain the system of equations depicted in (8) with $5 + 2 + (m-5) = m + 2$ parameters. Note, however, that this still is more than the number of observed peaks m. Thus, we need to consider some additional simplifying assumptions. These are discussed in the next section.

$$\begin{aligned}
x_1 &= H + P_0HQ,\\
x_2 &= HR_1 + P_0HQR_1 + P_1HQ,\\
x_3 &= HR_2 + P_0HQR_2 + P_1HQR_1 + P_2HQ,\\
x_4 &= HR_3 + P_0HQR_3 + P_1HQR_2 + P_2HQR_1 + P_3HQ,\\
&\vdots\\
x_{m-4} &= HR_{l-1} + P_0HQR_{l-1} + P_1HQR_{l-2} + P_2HQR_{l-3} +\\
&\quad P_3HQR_{l-4} + P_4HQR_{l-5},\\
x_{m-3} &= P_1HQR_{l-1} + P_2HQR_{l-2} + P_3HQR_{l-3} + P_4HQR_{l-4},\\
x_{m-2} &= P_2HQR_{l-1} + P_3HQR_{l-2} + P_4HQR_{l-3},\\
x_{m-1} &= P_3HQR_{l-1} + P_4HQR_{l-2},\\
x_m &= P_4HQR_{l-1}.
\end{aligned} \tag{8}$$

2.2 A model for the enzymatic ^{18}O-labeling

A way to further reduce the number of parameters is to assume a model for the enzymatic ^{18}O-labeling reaction, such that the shift probabilities $P_0, \ldots, P_4$ can be replaced by a smaller number of parameters. To this aim, we consider a Markov-model.

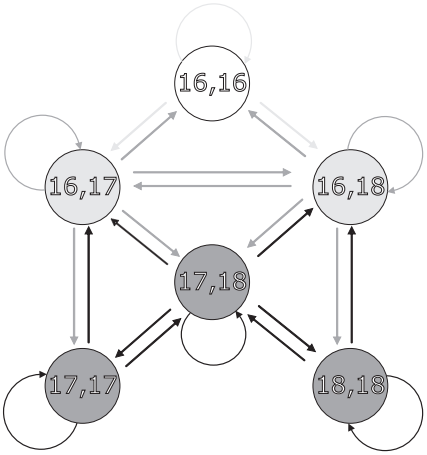


Figure 4: Possible transitions between the carboxyl-terminus states in the Markov-chain.

In equation (1) we introduced different configurations $X(\cdot)$, indicating the combination of oxygen isotopes present at the carboxyl-terminus of a peptide. We will refer to the configurations as states. As argued in the previous section, we assume that the carboxyl-terminus of all isotopic variants of a peptide from Sample II before the labeling contains two ^{16}O-atoms, i.e., it is in state $X(1)$. This is depicted in Figure 4, where the white circle denotes state $X(1)$. After one oxygen-atom replacement ($k = 1$), the peptide's carboxyl-terminus will stay with certain probability in state $X(1)$ or moves to states $X(2)$ or $X(3)$. The probability depends on the proportions of the heavy-oxygen water impurities p_{16} and p_{17} (see previous section). The new states are indicated by the light gray color in Figure 4, where the arrows indicate the possible direction of transitions. After two oxygen replacements ($k = 2$), the probabilities for the carboxyl-terminus to remain in states $X(1)$, $X(2)$, or $X(3)$ will change. Moreover, three additional states can be reached, namely, $X(4)$, $X(5)$, and $X(6)$ (see the dark gray circles in Figure 4). A third oxygen-replacement reaction ($k = 3$) will allow for eight new transitions, indicated by the black arrows in Figure 4, and so on. This process can be seen as a discrete-time Markov-chain, with the discrete time steps interpreted as the oxygen replacements.

The Markov-chain can now be defined more formally. Given the transition probability matrix $\boldsymbol{T}$, the state probabilities can be expressed as follows:

$$\boldsymbol{S}'_k \;=\; \boldsymbol{S}'_0 \boldsymbol{T}^k P(k)\;, \tag{9}$$

with $\boldsymbol{S}_k$ denoting a 6×1 column vector containing the state probabilities after k ($k = 0, 1, \ldots$) oxygen replacements and $P(k)$ denoting the probability that k replacement reactions will take place. Under the assumption that at the beginning of the labeling process the isotopic variants of a peptide in Sample II contain 100% ^{16}O-atoms at the carboxyl-terminus, the 6×1 initial state vector is given by $\boldsymbol{S}_0 = (1, 0, 0, 0, 0, 0)'$.

Recall that we assume that the enzymatic reaction is equally likely on both reaction sites of the carboxyl-terminus. We also assume that previous oxygen replacements do not influence the enzymatic reaction for future oxygen replacements, i.e., that the transition probabilities, specified in matrix $\boldsymbol{T}$, are independent of the number of oxygen replacements k. The transition probability matrix $\boldsymbol{T}$ with transition probabilities P_{ab} can then be constructed in the following way from the known water impurities, p_{16} and p_{17}:

$$\begin{pmatrix} p_{16} & p_{17} & p_{18} & 0 & 0 & 0 \\ \frac{p_{16}}{2} & \frac{p_{16}+p_{17}}{2} & \frac{p_{18}}{2} & \frac{p_{17}}{2} & \frac{p_{18}}{2} & 0 \\ \frac{p_{16}}{2} & \frac{p_{17}}{2} & \frac{p_{16}+p_{18}}{2} & 0 & \frac{p_{17}}{2} & \frac{p_{18}}{2} \\ 0 & p_{16} & 0 & p_{17} & p_{18} & 0 \\ 0 & \frac{p_{16}}{2} & \frac{p_{16}}{2} & \frac{p_{17}}{2} & \frac{p_{17}+p_{18}}{2} & \frac{p_{18}}{2} \\ 0 & 0 & p_{16} & 0 & p_{17} & p_{18} \end{pmatrix}, \tag{10}$$

Row ($a = 1, \ldots, 6$) and column ($b = 1, \ldots, 6$) indices correspond to states $X(1)$ to $X(6)$, respectively. The transition probabilities P_{ab} give the probability to move from state $X(a)$ to state $X(b)$. For example, the probability to move from state $X(3) = (1, 0, 1)$ to state $X(1) = (2, 0, 0)$ equals $P_{31} = p_{16}/2$, because only if the ^{18}O-atom in state $X(3)$ is replaced by an ^{16}O-atom, we reach state $X(1)$. The chance that a carboxyl-oxygen is replaced with an ^{16}O-atom depends on the water impurity p_{16} of the heavy-oxygen water. We assume that the concentration of water impurities is constant over time. This achieved by performing the enzymatic reaction in an abundance of heavy-oxygen-water, such that dilution by exchanged ^{16}O is negligible.

Term $P(k)$ in (9) represents the probability of k oxygen replacements. The number of oxygen replacements k during the labeling reaction is unknown and depends on the reaction speed and duration. The duration of the enzymatic reaction is usually known and kept constant across multiple labeling experiments. We will denote the duration by τ. The reaction speed depends on many factors and is specific for a peptide. We express the speed as the peptide-specific incorporation rate λ, which gives the number of reactions per time unit. We assume that, for a particular peptide, λ is constant over time.

Under these assumptions, the probability for k oxygen replacements can be modeled by a Poisson process with rate λ and time τ. As a result, after summing

over all possible values of k and rearranging terms, equation (9) can be expressed as follows:

$$\boldsymbol{S}'(\lambda, \tau, p_{16}, p_{17}) \quad = \quad \boldsymbol{S}'_0 e^{-\lambda\tau} e^{\boldsymbol{T}\lambda\tau} \ , \tag{11}$$

where $\boldsymbol{S}'(\lambda, \tau, p_{16}, p_{17})$ is the vector containing the state probabilities for the isotope combination on the carboxyl-terminus of a peptide with incorporation rate λ after a reaction time τ in heavy-oxygen water with impurities p_{16} and p_{17}. Note that, to simplify notation, we will suppress the use of τ, p_{16}, and p_{17} in subsequent formulae.

The term $e^{\lambda\tau}e^{\boldsymbol{T}\lambda\tau}$ in equation (11) can be seen as a transition matrix resulting from a solution of the Kolmogorov backward equation for a continuous-time Markov model with generator $\mathcal{Q} = \lambda(\boldsymbol{T} - \boldsymbol{I}_6)$, where $\boldsymbol{I}_6$ is the 6×6 identity matrix and $\boldsymbol{T}$ as defined in (10).

Now, the probabilities of the mass shifts, defined in (2), are computed as follows:

$$\begin{aligned} P_0(\lambda) = S_1(\lambda), \quad P_2(\lambda) = S_3(\lambda) + S_4(\lambda), \\ P_1(\lambda) = S_2(\lambda), \quad P_3(\lambda) = S_5(\lambda), \quad P_4(\lambda) = S_6(\lambda), \end{aligned} \tag{12}$$

where the index denotes the element of the state probability vector $\boldsymbol{S}(\lambda)$.

Figure 5 shows the values of the mass shift probabilities as a function of λ for a labeling reaction of $\tau = 120$ in heavy-oxygen water with impurities $p_{16} = 4\%$ and $p_{17} = 1\%$. Note that, for $\lambda \geq 0.1$, the shift probabilities are basically constant. A similar plot would be obtained for the dependence of the probabilities on the reaction duration. It follows that, for a peptide with $\lambda \geq 0.1$, the enzymatic reaction is basically completed after 120 time units, e.g., minutes; extending the duration does not change the mass shift probabilities, because the reaction has reached a stationary condition. This means that, if we consider a peptide with $\lambda = 0.1$, after $\tau = 120$ minutes, only 89.8% of the molecules will receive two ^{18}O-atoms on their carboxyl group in the current setting. In other words, the isotopic peaks of only 89.8% of the peptide molecules from Sample II will shift by 4 Da to the right in the joint mass spectrum. Further, the peaks of 1.89%, 8.04%, 0.08%, and 0.18% of the labeled molecules will shift by 3, 2, 1, and 0 Da, respectively. The analysis of a labeled mass spectrum should correct for these different overlaps to avoid biased estimates of the relative peptide abundance. Further, it should be stressed that the presence of the ^{17}O-isotope may lead to mass shifts of 1 and 3 Da. Although ^{17}O is a non-abundant isotope of oxygen, its contribution to the bias may not be ignorable. In the case of 1% ^{17}O contamination, this leads to 1.97% of the peptides labeled with ^{17}O.

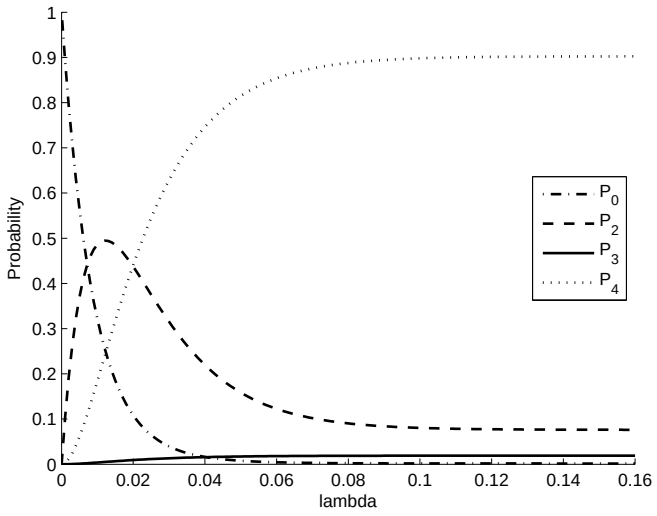


Figure 5: Shift probabilities P_0, P_2, P_3 and P_4 as a function of λ for an enzymatic reaction of 120 minutes with heavy-oxygen water impurities of $p_{16} = 4\%$ and $p_{17} = 1\%$. Shift probabilities P_1 are small and not shown in this figure.

It is important to point out that the sensitivity of the estimation method with respect to assumptions made about the percentage ^{17}O and ^{16}O contamination is not a concern, because the degree of contamination present in the heavy-oxygen water are measured prior to the labeling experiment.

From this perspective, by using (11) and (12), we replace the five shift probabilities by a single parameter, namely, λ. Consequently, we further reduce the number of parameters in (5) to $3 + (m-5) = m-2$, which is less then the number of available observations m. This allows us to fit the model, specified by (8) and (11)–(12), to the observed data.

2.3 Estimation and inference

As described in the previous sections, by using (8) and (11)–(12), we can express the expected values x_j of the peaks observed in the joint spectrum as a function of parameter vector $\boldsymbol{\theta} = [Q, H, R_1, \ldots, R_{m-5}, \lambda]$. The parameter of interest is the relative abundance Q, defined in (7). By using the assumed form of the model, given in (3), the likelihood for the joint spectrum with m observed peaks of intensities y_j can be expressed as follows:

$$L(\boldsymbol{\theta}) \;=\; \prod_{j=1}^{m} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\{y_j - x_j(\boldsymbol{\theta})\}^2} \;. \tag{13}$$

It is difficult to express the function $x_j(\boldsymbol{\theta})$ explicitly in the general case. An example of the model structure in matrix formulation can be found in the Appendices.

The extension of (13) to accommodate additional, say n, spectra resulting from, e.g., technical replicates or peptides, which appear in multiple mass spectra due to a high-dimensional LC-step, is obvious:

$$L(\boldsymbol{\theta}) \;=\; \prod_{s=1}^{n}\prod_{j=1}^{m} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\{y_{sj} - x_{sj}(\boldsymbol{\theta})\}^2} \;. \tag{14}$$

For each additional spectrum s ($s = 1, \ldots, n$), an extra reference intensity parameter H_s may need to be added to vector $\boldsymbol{\theta}$, to allow for the between-spectrum intensity-scale variability. The reference intensities account for the relationship between the abundances in multiple mass spectra due to LC. It should be stressed that this relationship is not additive, but multiplicative. This means that all peak heights of a particular peptide differ by a multiplicative constant H_s across the LC-runs. Also, a separate residual variance parameter σ_s^2 may be used. Note that the number of observations $(n \times m)$ increases more rapidly than the number of parameters $(2 + n + m - 5)$ when additional mass spectra are available for a given peptide. Thus, inclusion of additional spectra improves the efficiency of the estimation, as it increases the number of degrees of freedom.

The estimates $\hat{\boldsymbol{\theta}}$ are found by maximizing likelihood (14). The residual variance(s) are estimated by the usual mean residual sum of squares. Under the normality and homoscedasticity assumptions, the approximate variance-covariance matrix of the estimated parameters can be obtained by

$$\boldsymbol{V}(\hat{\boldsymbol{\theta}}) \;=\; \hat{\sigma}^2(\boldsymbol{J}'\boldsymbol{J})^{-1} \;, \tag{15}$$

where $\boldsymbol{J}$ is the Jacobian of (possibly, modified for multiple spectra) likelihood function (13), evaluated at $\hat{\boldsymbol{\theta}}$. Moreover,

$$\frac{\hat{\theta}_i - \theta_i}{s(\hat{\theta}_i)} \sim t_d, \tag{16}$$

where $\hat{\theta}_i$ is the ith element of $\hat{\boldsymbol{\theta}}$, $s(\hat{\theta}_i)$ is the standard deviation of $\hat{\theta}_i$, and t_d is the t-distribution with $d = n \times m - (2 + n + m - 5)$ degrees of freedom.

In the context of the proposed model, inference is especially important for the parameters Q and λ. It is straightforward to assign p-values to the obtained estimates. For example, if we want to test whether the relative abundance Q, differs ~~significantly~~ from one ($H_0 : \ Q = 1$), we calculate following statistic:

$$t_{\text{score}} = \frac{\hat{Q} - 1}{s(\hat{Q})} \ . \tag{17}$$

The p-value can now be calculated from the t-score using the cumulative t-distribution with d degrees of freedom.

As mentioned earlier, the structure of the model is rigid and is determined by the number of observed peptide peaks m. Parameters which are found non-significant by previously described test will not restrict the model.

2.4 Practical implementation of the estimation procedure

Practically, the maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ are obtained by minimizing the term

$$\sum_{s=1}^{n} \sum_{j=1}^{m} (y_{sj} - x_{sj})^2 \tag{18}$$

in function of parameters $\boldsymbol{\theta} = [Q, H_1, \ldots, H_n, R_1, \ldots, R_{m-5}, \lambda]$. Because function $x_{sj}(\boldsymbol{\theta})$ is non-linear in the parameters and because the parameters are constrained to positive values, minimizing (18) becomes a constrained non-linear optimization problem. In order to transform this to an unbounded optimization problem, the logarithm of the parameters can be estimated. This is sufficient for all parameters except for λ, because the derivatives of the shift probabilities are near zero for large values of λ. This can be observed in the example of Figure 5. To avoid numerical instability during the minimization, parameter λ should be constrained to an upper bound. For instance, we propose an upper bound for λ equal to $20/\tau$. This upper bound was implemented via an extension of Box's idea:

$$\lambda = \frac{20/\tau \exp(\lambda')}{\exp(\lambda') + 1} \ , \tag{19}$$

with the inverse transformation given by

$$\lambda' = log\left(\frac{\lambda}{20/\tau - \lambda}\right) \ . \tag{20}$$

The transformation is depicted in Figure 6 for $\tau = 120$. It can be observed that λ' can take any real value, while λ lies between 0 and 0.166. From Figure 5, it can be

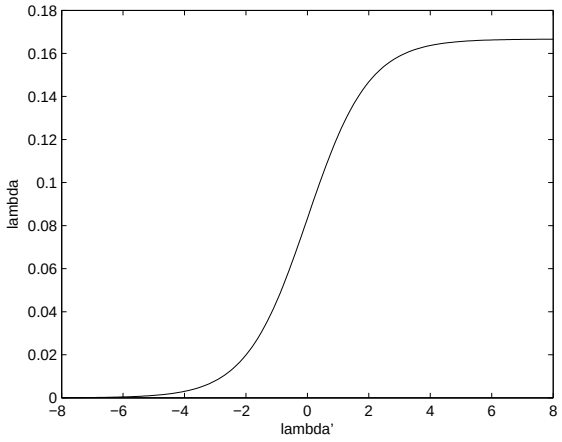


Figure 6: Transformation (19) of λ

seen that for large λ-values ($\lambda > 0.1$), the shift probabilities $P_0, \ldots, P_4$ stabilize. Thus, in a steady-state, larger λ-values do not influence the shift probabilities much. Therefore, it is reasonable to assume an upper bound for λ equal to 0.166.

The minimization of (18) can now be solved as an unconstrained non-linear optimization problem. For this purpose, the Gauss-Newton method (Hartley *et al.*, 1961) can be used. The required Jacobian matrix can be easily calculated analytically, because the derivative of the matrix exponential in (11) has the same form as the derivative of a scalar exponential. The Gauss-Newton method converges fast when the starting values are close to the true values of the parameters. Therefore, a rough estimate of the parameter values is needed. The starting value for the relative abundance Q can be calculated as the ratio between the fifth and the first peak intensity observed in the joint spectrum. The reference intensity H for the joint spectrum is chosen as the peak intensity of the first peak. The isotopic ratios $R_1, \ldots, R_{m-5}$ are calculated by the method of Valkenborg *et al.* (2008a). The starting value for λ is chosen to be constant at the upper bound, defined in (19).

Further, in order to improve the numerical stability of the optimization problem, the matrix exponential of $\boldsymbol{T}$ is computed by using a scaling and squaring algorithm with a Padé approximation (Higham *et al.*, 2005).

3 Results

In this section, we present results of a simulation study, undertaken to check the statistical properties and robustness of the developed model. We also show results

of an application of the model to a controlled experiment of the enzymatic labeling of bovine Cytochrome C peptides.

3.1 Simulation study

We considered five tryptic peptides found in human blood and identified on tandem MS. The peptides were chosen such that their mass falls in the range between 1000 Da to 3000 Da in steps of approximately 500 Da. The isotopic distribution of these peptides was calculated via the multinomial expansion, as described by Yergey (1983). This resulted in, respectively, 5, 6, 6, 7, and 8 isotopic variants for the peptides (in the order of peptides' increasing masses). The joint spectra were generated by using the model, defined by (3), (8), and (11)–(12). Nine different parameter settings were considered. Relative abundance Q was allowed to take the values $1/2$, 1, and 2. Peptide-specific oxygen incorporation rate λ was set at 0.08, 0.02, and 0.008. The choice was motivated by the form of the plot shown in Figure 5. The duration of the enzymatic reaction was kept constant at $\tau = 120$ minutes. For illustration purposes, the proportions of heavy-oxygen water impurities were assumed to equal $p_{16} = 4\%$ and $p_{17} = 1\%$. Finally, a small amount of normal instrument noise with $\sigma^2 = 5$, compatible with values observed in well-controlled, unlabeled MALDI-TOF spectra, was added to the generated expected values of the peaks; negative values were truncated at zero and are regarded as isotopic peptide variants under the limit of detection. In order to assess if inference holds for small sample sizes, we considered only two replicates ($n = 2$) for this simulation study. This means that, for each simulation setting, two joint spectra are generated. Inter-spectra variability due to, e.g., laser fluctuations, LC-variability, ionization efficiency or inefficient crystallization, was simulated by using different reference intensities H_s for the joint spectra. They took the value of $H_1 = 1800$ and $H_2 = 2200$ for $Q = 0.5$ and 1, and $H_1 = 900$ and $H_2 = 1100$ for $Q = 2$. In this way, the joint mass spectra, generated for $Q = 0.5$ and $Q = 2$, should be equally affected by the instrument noise.

The simulated data were analyzed by using the model modified for multiple spectra, as explained in the previous section. The intention was to check the statistical properties of estimation and inference under the model assumptions. We generated 2500 data sets for each setting and calculated the coverage of the confidence intervals (CIs) of the estimated parameters, obtained by using the t-distribution (16). Furthermore, for each of the estimated parameters, the average relative bias $\bar{b}$ and the empirical variance s^2_{emp} were computed from the 2500 estimated parameter estimates. The average model-based variance s^2_{mb} was calculated from the 2500 model-based variances, obtained from the diagonal elements of the variance-covariance matrix, estimated by (15). We mainly discuss the results of estimation

Table 1: Simulation results for the peptide with mass 1000.5 Da: estimation of Q. The relative bias $\bar{b}$, empirical variance s^2_{emp}, and the average model-based variance s^2_{mb}.

Q	λ	$\bar{b} \times 10^{-5}$	$s^2_{\text{emp}} \times 10^{-5}$	$s^2_{\text{mb}} \times 10^{-5}$	CI coverage
0.5	0.008	-590.91	219.41	207.96	93.88
0.5	0.02	-63.28	3.98	4.12	95.04
0.5	0.08	-23.16	0.48	0.52	95.12
1	0.008	-165.43	107.82	103.04	94.48
1	0.02	-20.26	1.90	1.94	95.24
1	0.08	-8.18	0.27	0.28	95.08
2	0.008	-76.00	292.93	273.84	94.40
2	0.02	-16.50	6.19	6.12	95.04
2	0.08	-7.52	1.40	1.41	94.60

of parameters Q and λ for the peptide with mass 1000.5 Da. The results for the other peptides are similar.

Table 1 shows the results for relative abundance Q. In general, the results for $\lambda > 0.008$ are satisfactory: the estimation bias is negligible, the model-based variance is close, but slightly higher than the empirical one, and the CI coverage is close to the desired level of 95%. The standard error for the estimated coverage is equal to $\sqrt{0.05 \times 0.95/2500} = 0.004$. The bias and difference between the variances decrease with λ. For $\lambda = 0.008$, the results show a larger bias in the estimate of Q and an underestimation of the empirical variance. In the case of $Q = 0.5$ and $\lambda = 0.008$, this results in the CI coverage statistically significant smaller than the desired level of 95%. Note that, as shown in Figure 5, when $\tau = 120$, the incorporation rate $\lambda = 0.008$ leads to a very inefficient labeling. As a consequence, a substantial proportion of labeled peptides do not receive a heavy-oxygen isotope and there is a large overlap of isotopic peaks for the peptides from the two samples. In such circumstances, one can expect difficulties in estimating the relative abundance; in the extreme case, with a total failure of labeling, it would be impossible to distinguish between the isotopic peaks from the two samples. In such a case the relative abundance cannot be estimated.

In order to asses the validity of the t-distribution, we constructed Q-Q-plots of the statistic in (16) for the relative abundance parameter Q. Figure 1 in the Appendix displays the Q-Q-plots of the statistic against a t-distribution with $2 \times 9 - (2 + 2 + 4) = 10$ degrees of freedom. The Q-Q-plots re-iterate the suitability of the use of the t-distribution, perhaps with the exception of $\lambda = 0.008$ (panels (c), (f) and (i)).

Table 1 in the Appendix presents the results for the peptide-specific incorporation rate λ. They indicate that the relative bias of estimates for λ decreases with increasing Q. For $Q > 1$ it is below 1%, while for $Q = 0.5$ it is between 0.5% and 7%. This trend may be seen as an expected one: a larger abundance of peptide molecules in Sample II results in a larger amount of ill-labeled molecules and a larger overlap of isotopic peaks with the unlabeled peptide molecules in Sample I. This, in turn, has a positive effect on the estimation of parameter λ. For $\lambda < 0.08$, the model-based variance is close to, but slightly smaller than the empirical one. For $\lambda = 0.08$, the empirical variance is much larger than for $\lambda < 0.08$, and it is overestimated, on average, by the model-based estimates. As a result, for $\lambda = 0.08$ and $Q = 0.5$, the CI coverage is statistically significantly higher than the nominal level. A fast incorporation rate (in this case, $\lambda > 0.08$) will emerge in a reaction, which reaches its stationary condition very rapidly. This means that the labeling reaction is complete and stable after a duration of $\tau = 120$ minutes for λ values larger than 0.08, as can be seen in Figure 5. In this region, the derivatives of the mass shift probabilities with respect to λ are close to zero. As a consequence, there is a degree of uncertainty and the algorithm cannot precisely determine the value of the λ parameter. It should be noted that, in an extreme case, this can lead to unidentifiability issues. On the other hand, a fast incorporation rate will often lead to a complete reaction and this will positively influence the precision of the estimates for the relative abundance Q.

To check whether the peptide mass influences the precision of estimation of Q, we plotted the empirical and mean model-based variances for the five peptides considered in the simulations. Panel (a) from Figure 2 in the Appendix shows the plot for $\lambda = 0.08$. It indicates that the smallest variance is obtained for $Q = 1$. This can be explained by the fact that, in this case, the influence of the instrument noise is relatively small, because more peaks in the joint spectra have a large intensity. For $Q = 0.5$ and $Q = 2$, the influence of noise is larger for the small peaks.

The variances increase for decreasing λ (see panel (b) of Figure 2 in the Appendix). A smaller incorporation rate leads to inefficient labeling and it increases the number of the peptide molecules from the labeled sample that do not receive two ^{18}O-atoms. As a consequence, there is a larger overlap of isotopic peaks related to those molecules with the isotopic peaks of the peptide from the unlabeled sample. This results in a larger uncertainty about the relative abundance and a larger variance of estimated Q. A reverse pattern can be observed when inspecting the variance of the estimates of λ (see panel (a) and panel (b) of Figure 3 in the Appendix). This is because a substantial amount of overlapping isotopic peaks is required to accurately estimate λ. On the other hand, when estimating Q, we want to avoid the overlap.

3.2 Bovine Cytochrome C data

In this section, we describe the application of the proposed method to a data set of six replicated joint mass spectra obtained from the tryptic peptides of bovine Cytochrome C from LC Packings. The peptide mixture was divided into two parts. One part was enzymatically labeled with a stable ^{18}O-isotope, with trypsine as a catalyst, while the other part remained unlabeled. Next, three units from the unlabeled part where mixed with one unit from the labeled part, which should result in the relative abundance ratio of $Q = 0.33$. The composed mixture was automatically spotted six times on one stainless steel plate by a robot. The plate was processed by a 4800 MALDI-TOF/TOF analyzer (Applied Biosystems) mass spectrometer. More details about the exact procedure can be found in the manuscript by Staes *et al.* (2004). We restrict the analysis to three Bovine Cytochrome C peptides, for which joint spectra were obtained. The other Bovine Cytochrome C peptides could not be retrieved from the spectra. The amino acid compositions of these peptides are as follows: peptide CC1 (mass 1167.61 Da) - TGPNLHGLFGR; peptide CC2 (mass 1455.66 Da) - TGQAPGFSYTDANK; peptide CC3 (mass 1583.75 Da) - KTGQAPGFSYTDANK. The data are processed by the method presented by Valkenborg *et al.* (2009) and the resulting joint spectra in stick representation with $m = 10$ peptide peaks are displayed in panel (a) of Figures 4, 6, and 8, respectively, in the Appendix. The quality of the peak selection is manually curated in order to confirm that all the found peaks are members of the corresponding isotopic distribution.

Table 2 shows the parameter estimates of the model modified for multiple spectra, defined by (3), (8), and (11)–(12), obtained by fitting the model to the observed peak heights of the six joint spectra for each of the three considered peptides. The proportions of water impurities of the heavy-oxygen water were reported by the lab experimentalists and equal to $p_{16} = 2\%$ and $p_{17} = 0.9\%$. The true values of isotopic ratios R_i were calculated from the atomic composition of the peptides by using the convolution method developed by Rockwood (1995). As we do not know the values of the peptide-specific incorporation ratios λ, we only estimate products $\lambda\tau$.

Panel (b) of Figures 5, 7, and 9 in the Appendix display the estimated expected values of the peaks of the joint spectra shown in corresponding panels (a). For peptides CC2 and CC3, the observed and estimated peak heights seem to be in agreement, while for CC1 marked differences are observed. The fluctuating reference intensities, i.e., first peak in the joint spectrum, are worth noting, which indicate the between-spectrum variability due to, e.g., laser fluctuations, crystallization effects, etc. For this reason, we consider the reference intensities as a nuisance and

Table 2: Parameter estimates (Est.) and standard errors (SE) based on six technical replicates for the tryptic bovine Cytochrome C peptides at mass 1167.61, 1455.66, and 1583.75 Da.

	Q	R_1	R_2	R_3	R_4	R_5	$\lambda\tau$
Peptide CC1 (1167.61 Da)							
True	0.33	0.6552	0.2394	0.0631	0.0133	0.0065	-
Est.	0.5518	0.8266	0.2949	0.0318	0.0629	0.000001	4.7980
SE	0.0318	0.0119	0.0192	0.0161	0.0191	0.0154	0.3998
Peptide CC2 (1455.66 Da)							
True	0.33	0.7902	0.3414	0.0947	0.0329	0.0077	-
Est.	0.3340	0.7903	0.3367	0.0966	0.0322	0.0063	7.7350
SE	0.0129	0.0035	0.0083	0.0066	0.0078	0.0056	1.1780
Peptide CC3 (1583.75 Da)							
True	0.33	0.8581	0.4119	0.1428	0.0396	0.0098	-
Est.	0.3318	0.8615	0.4056	0.1309	0.0411	0.0099	7.8289
SE	0.0068	0.0020	0.0045	0.0038	0.0042	0.0031	0.6704

we do not include their estimates in Table 2, but we provide them in Table 2 in the Appendix.

The results for peptides CC2 and CC3, shown in Table 2, confirm that the estimated relative abundance Q is in agreement with the targeted value of 0.33. Equally, the estimates of the isotopic ratios are virtually identical to their theoretical values. Thus, the model seems to adequately describe the data. Interestingly, the estimated value of $\lambda\tau$ is similar for the two peptides, suggesting a similar incorporation rate.

For example, in order to test if the relative abundance Q of peptide CC2 is different from ~~one~~, we calculate the t-score:

$$\frac{(0.3340 - 1)}{0.0129} = -51.6279\ . \tag{21}$$

For a t-distribution with $d = 6 \times 10 - (2 + 6 + 10 - 5) = 47$ degrees of freedom, the t-score corresponds to a p-values of 2.3797×10^{-43}. At a significance level of 5%, we can reject the null hypothesis $H_0 : \ Q = 1$. However, for peptide CC1, the estimated relative abundance markedly deviates from 0.33. The estimated isotopic ratios for peptide CC1 are also statistically significantly different from the theoretical values. Moreover, for a peptide within this mass range, we expect the isotopic ratios to decrease monotonically. The estimates in Table 2 show a clear deviation from monotonicity. This non-conformity of the isotopic distribution can be used as an indicator for model misspecification or method failure. Finally, the residual error variance σ^2 and reference intensities (see Table 2 in the Appendix) are markedly different from the corresponding values for peptides CC2 and CC3.

Table 3: Results of the application of the Eckel-Passow *et al.* (2006) model based on Averagine (left-panel) and naive isotope ratios (right-panel) to the six technical replicates of the tryptic bovine Cytochrome C. Mean values, over the six replicates, of the estimates of Q and $\lambda\tau$.

	Eckel-Passow			Naive ratio		
Parameter	CC1	CC2	CC3	CC1	CC2	CC3
Q	0.5671	0.2838	0.2856	0.5336	0.3473	0.3556
$\lambda\tau$	2.6330	.	.	.	.	.

To assess the fit of the model in more detail, Figures 5, 7, and 9 in the Appendix presents the residuals for the analysis of the data for peptide CC1, CC2, and CC3, respectively. The plots suggests that a model with a heteroscedastic error structure, in which the residual variance decreases with the mean intensity of the peaks observed in the joint spectrum, might be more appropriate. An extension of the model to deal with such a residual variance is an important step for further research.

In order to visualize the gain in stability when incorporating multiple spectra, we refitted the model on all possible groups of $1, \ldots, 6$ spectra for peptide CC3. Figure 10 in the Appendix, shows how the estimate (panel (a)) and precision (panel (b)) for relative abundance Q improves when using multiple spectra. Regardless of the number of spectra used, the algorithm converged swiftly. The same results are observed for peptide CC1 and CC2 (data not shown). Figure 12 displays the Q-Q plot for the estimates of the relative abundance Q, reference intensity H, and peptide-specific incorporation rate λ for the six spectra fitted individually. The three outliers at the left of the figure are originating from the same spectrum. Disregarding the outliers, the distributional assumptions seem valid.

We mentioned earlier, that the model structure is rigid and depends on the number of observed peptide peaks in a joint spectrum. To investigate the consequence of missing the last peptide peaks, we refitted the model based on six replicates for peptide CC3 for $m = 6, \ldots, 10$ observed peptide peaks. The result is shown in Figure 11 in the Appendix. It follows that ignoring the available information alters the structure of the model, which violates the observed data. This leads to biased estimates. Therefore, it is important to use all available information.

To compare the obtained results with other methods, we used the model developed by Eckel-Passow *et al.* (2006) based on Averagine. We fitted it to each of the six technical replicates separately, as the model does not accommodate multiple spectra. The mean values of the estimates of $\theta_{c2s}/\theta_{c1s}$ and $K_{cs}t$, which correspond to Q and $\lambda\tau$ in our notation, are displayed in the left part of Table 3. For peptides CC2 and CC3, the method seriously underestimates the relative abundances and

failed to compute a value for $K_{cs}t$. For peptide CC1, the relative abundance is overestimated even more than in the case of the estimate shown in Table 2. We found this method very sensitive regarding the assumed isotopic distribution.

A naive approach to calculate the relative abundance is to take the ratio of the fifth and the first peak in a joint spectrum. By doing so, we obtain the results displayed at the right-hand part of Table 3. For peptides CC2 and CC3, the estimates of the relative abundance are, on average, further away from the true value of 0.33 than the estimates given in Table 2. However, it should be noted that, in this case, the results from the naive approach are acceptable. The efficient ^{18}O-labeling (overnight) and the use of highly purified heavy-oxygen water (expensive) caused a clear separation between the labeled and unlabeled spectra, which justifies the naive assumptions. In a realistic setting, however, labeling might be inefficient and it would be cost-efficient to use less purified oxygen labels, which is taken care of by the presented method.

4 Discussion

As we have mentioned in the introduction, several methods have already been proposed to analyze data from enzymatic ^{18}O-labeling experiments, (Mirgorodskaya *al.*, 2000; Rao *al.*, 2005; López-Ferrer *al.*, 2006; Eckel-Passow *al.*, 2006). Most of them, however, postulate the use of additional experimental steps, which is an important limitation. Our method does not require such steps. It is similar in spirit to the approach developed by Eckel-Passow *et al.* (2006). In fact, we can show that the Markov-model, which we propose, includes the model developed by Eckel-Passow *et al.* for the probabilities of particular mass shifts of the labeled peptide molecules (see equations (1) and (2) in their paper). However, our model extends in several ways the one used by Eckel-Passow *et al.* First, Eckel-Passow *et al.* suggest to estimate the isotopic distribution of a peptide by using the average distribution developed by Senko *et al.* (1995). Although, the actual isotopic distribution of a peptide can markedly deviate from the average one when, e.g., the peptide contains sulphur atoms (Johnson and Muddiman, 2004; Valkenborg *et al.*, 2007). Instead, we propose to estimate the parameters of the isotopic distribution directly from the observed data. The advantage of this solution is that the information about the ratios can be used to automatically annotate whether the observed series of mass spectrum peaks are truly generated by a peptide (Valkenborg *et al.*, 2008a) or originating from noise. Note, however, that it is also possible to use our model with a fixed isotopic distribution. Second, it allows the model to account for the possible presence of ^{17}O-atoms in the heavy-oxygen water, although the bias introduced by ^{17}O-atoms is expected to be minor. Finally, we developed a unified modeling framework, in

which all parameters of interest, like the relative abundance Q, isotopic ratios R, and the peptide-specific incorporation rate λ, are simultaneously estimated from the data. It can easily accommodate different parameterizations, and provide necessary estimates of precision. It can also be scaled up to more complicated experimental designs, with several groups of samples with technical and biological replicates, etc.

We studied the performance of the proposed approach by means of a simulation study and by a controlled MS experiment. The simulation results indicate satisfactory properties of the estimates obtained from the model under its correct specification. They point to the importance of the peptide-specific incorporation rate λ for the performance of the labeling strategy: if λ is too low, the incomplete labeling may cause bias in the estimation of the relative abundance. This underscores the importance of a careful choice of the duration of the labeling experiment. From this point of view, the possibility of using the model to obtain a preliminary estimate of λ from, e.g., a limited pilot-experiment, is an important advantage when optimizing the experimental protocol. The influence of the purity of the heavy-oxygen water and the duration of the reaction on the optimal model performance for, e.g., λ or Q, is a topic for further research.

The results of the application to the controlled MS experiment were consistent with the true parameter values for two out of three analyzed peptides. For one peptide, however, the results were biased both for our model and for the method of Eckel-Passow *et al.* (2006). As we encountered issues with the quality of MS-measurements in the available spectra, it is possible that the bias may be caused by some experimental factors unknown to us. On the other hand, the model, which we have developed, entails several assumptions. It is conceivable that, e.g., the chemical composition of the peptide (arginine/lysine C-terminus) causes a violation of some of these assumptions. For instance, it could be possible that some peptides are not amenable to any further reaction after receiving one oxygen-isotope. This would imply that transition matrix $\boldsymbol{T}$, which assumes that such reactions take place, is misspecified. Also, the assumption regarding the Poisson process, which imply that subsequent oxygen replacements are independent of each other, could be an issue. These topics are subject to further investigation.

Several extensions of the proposed methodology can be considered. For instance, inclusion of heteroscedastic and/or serially correlated errors might be achieved by using appropriate variance- and correlation functions (Pinheiro and Bates, 2000). Also, the possibility of including random effects, which would allow estimating, e.g., the between-sample biological variability, can be thought of. These extensions require the use of more advanced estimation methods and will be addressed in the future.

Finally, mathematical methods to derive accurate quantification from incomplete labeled peptides are very important, but the incomplete labeling described in this manuscript is not restricted to ^{18}O-labeling. Other types of stable isotope labeling such as the more popular SILAC, ICAT and ^{15}N labeling all suffer from essentially the same phenomenon of incomplete labeling, and therefore the described method could be modified to be more generally applied to all isotopic labeling strategies. The necessary modifications will include, e.g., the adaptation of the Markov-model to the particular features of the labeling reaction.

The method presented in this manuscript is implemented as a Matlab-toolbox and is available on request. The method is computational fast and can evaluate approximately 100 peptides based on two technical replicates, i.e., two joint spectra, in one second on a standard laptop (Dell Latitude E6500).

Appendix

Example

Assume we have observed a series of $m = 11$ peaks in the joint mass spectrum from an enzymatic ^{18}O-labeling experiment with water impurities $p_{16} = 4\%$ and $p_{17} = 1\%$ and a duration of $\tau = 120$ minutes. This series can be generated by a peptide which has $l = m - 4 = 7$ seven isotopic variants or, equivalently 6 isotopic ratios. The function $x(Q, H, R_1, \ldots, R_6, \lambda)$ can be expressed as a matrix which takes the following form:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \\ x_{10} \\ x_{11} \end{pmatrix} = \left(\begin{array}{ccccccc|ccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & P_0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & P_1 & P_0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & P_2 & P_1 & P_0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & P_3 & P_2 & P_1 & P_0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & P_4 & P_3 & P_2 & P_1 & P_0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & P_4 & P_3 & P_2 & P_1 & P_0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & P_4 & P_3 & P_2 & P_1 & P_0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & P_4 & P_3 & P_2 & P_1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & P_4 & P_3 & P_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & P_4 & P_3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & P_4 \end{array}\right) \begin{pmatrix} H \\ HR_1 \\ HR_2 \\ HR_3 \\ HR_4 \\ HR_5 \\ HR_6 \\ HQ \\ HQR_1 \\ HQR_2 \\ HQR_3 \\ HQR_4 \\ HQR_5 \\ HQR_6 \end{pmatrix}$$

where $x_j(\boldsymbol{\theta})$ in equation (13) denotes the jth element of the above result vector.

From this set of equations it should be observed that the matrix on the right-hand side has a special structure. The structure of this matrix, say $\boldsymbol{A}$, is completely determined by the 11 observed peaks in the joint spectrum. More generally, when m peaks are observed from the joint spectrum, the matrix $\boldsymbol{A}$ can be represented as follows:

$$\boldsymbol{A} = \left(\begin{array}{c|c} \boldsymbol{I}_{m-4} & \\ \hline \boldsymbol{0}_{4\times(m-4)} & \boldsymbol{L}_{m\times(m-4)} \end{array} \right) ,$$

with $\boldsymbol{I}_{m-4}$ denoting the identity matrix of dimension $(m-4)\times(m-4)$ and $\boldsymbol{0}_{4\times(m-4)}$ denoting a matrix of zeros's of dimension $4 \times (m-4)$. Matrix $\boldsymbol{L}$ has a dimension of $m \times (m-4)$ and has a banded diagonal structure.

The probabilities $P_0, \ldots, P_4$ are a function of λ and can be calculated from the state probability vector $\boldsymbol{S}(\lambda)$:

$$\begin{aligned} P_0 = S_1(\lambda), \quad P_2 = S_3(\lambda) + S_4(\lambda), \\ P_1 = S_2(\lambda), \quad P_3 = S_5(\lambda), \quad P_4 = S_6(\lambda), \end{aligned}$$

where the index denotes the element of the state probability vector $\boldsymbol{S}(\lambda)$.

The state probability vector is calculated as

$$\boldsymbol{S}'(\lambda) \;\; = \;\; \boldsymbol{S}_0' e^{-\lambda 120} e^{\boldsymbol{T}\lambda 120} ,$$

where the initial state vector equals $\boldsymbol{S}_0 = (1,0,0,0,0,0)'$. The transition probability matrix is obtained from the water impurities and is equal to

$$\left(\begin{array}{cccccc} .04 & .01 & .95 & 0 & 0 & 0 \\ \frac{.04}{2} & \frac{.04+.01}{2} & \frac{.95}{2} & \frac{.01}{2} & \frac{.95}{2} & 0 \\ \frac{.04}{2} & \frac{.01}{2} & \frac{.04+.95}{2} & 0 & \frac{.01}{2} & \frac{.95}{2} \\ 0 & .04 & 0 & .01 & .95 & 0 \\ 0 & \frac{.04}{2} & \frac{.04}{2} & \frac{.01}{2} & \frac{.01+.95}{2} & \frac{.95}{2} \\ 0 & 0 & .04 & 0 & .01 & .95 \end{array} \right) .$$

Checklist assumptions

The model entails several assumptions. In this part we list all the assumptions to which the model is restricted. If possible, we mention how these assumptions can be assessed:

- The measurement error due to instrument noise is assumed independent and identically normally distributed. This can be assessed by using tests for normality and homoscedasticity.
- The number of isotope peaks for a peptide equals the observed number of peaks $m - 4$. If you find a violation against this assumption, e.g, only a shift of 2 Da occurred, this can be corrected by removing or adding isotope peaks in the structure of the model.
- Before labeling, the isotopic distribution for the labeled and unlabeled peptide are equal.
- Water impurities are known. This is measurable.
- Water impurities are constant during the reaction. This can be achieved by performing the labeling in excess of heavy-oxygen-water.
- Reaction time is known. This is measurable.
- The two oxygen reaction sites on the carboxyl-group are equally favorable, i.e., they have the same incorporation rate. A violation against this assumption will result in a joint spectrum with a pronounced peptide shift of 2 Da. The transition matrix $\boldsymbol{T}$ should be adjusted to account for this.
- Previous oxygen replacements do not influence the enzymatic reaction for future oxygen replacements, i.e., the incorporation rate and the transition matrix are static. This assumption is difficult to assess. Since, the model cannot account for this type of violation, the SSE will generally increase.
- As an initial condition, prior to labeling, we assume that the two oxygen atoms on the carboxyl-terminus are ^{16}O. If the assumption is violated, the correct distribution of oxygen isotopes on the carboxyl terminus can be imputed in $\boldsymbol{S}_0$.

Tables

Table 1: Simulation results for the peptide with mass 1000.5 Da: estimation of λ. The relative bias $\bar{b}$, empirical variance s^2_{emp}, and the average model-based variance s^2_{mb} should be multiplied by 10^{-4}.

Q	λ	$\bar{b}$	s^2_{emp}	s^2_{mb}	CI coverage
0.5	0.08	71.25	1985.85	2188.37	96.16
0.5	0.02	4.87	5.46	5.37	95.44
0.5	0.008	60.88	12.38	11.76	93.68
1	0.08	9.76	165.07	174.82	95.80
1	0.02	0.77	0.58	0.55	94.72
1	0.008	4.06	0.84	0.79	95.04
2	0.08	5.61	99.87	103.69	95.44
2	0.02	0.23	0.27	0.26	94.40
2	0.008	-2.52	0.70	0.68	94.60

Table 2: Parameter estimates based on six technical replicates for the tryptic bovine Cytochrome C peptides at mass 1167.61, 1455.66, and 1583.75 Da. The values should be multiplied by 10^4.

	CC1 (1167.61 Da)			CC2 (1455.66)			CC3 (1583.75 Da)		
	θ	$\hat{\theta}$	$\text{se}(\hat{\theta})$	θ	$\hat{\theta}$	$\text{se}(\hat{\theta})$	θ	$\hat{\theta}$	$\text{se}(\hat{\theta})$
H_1	-	7.4871	0.1147	-	2.4731	0.0120	-	2.2999	0.0061
H_2	-	7.3349	0.1141	-	2.2359	0.0119	-	2.2414	0.0061
H_3	-	6.3360	0.1107	-	2.2222	0.0118	-	2.1344	0.0060
H_4	-	7.1344	0.1134	-	2.4541	0.0120	-	2.3827	0.0062
H_5	-	4.8485	0.1063	-	1.9640	0.0117	-	1.8532	0.0059
H_6	-	6.1656	0.1101	-	2.4405	0.0120	-	2.4591	0.0062
σ^2	-	230.81	-	-	2.38	-	-	0.67	-

Figures

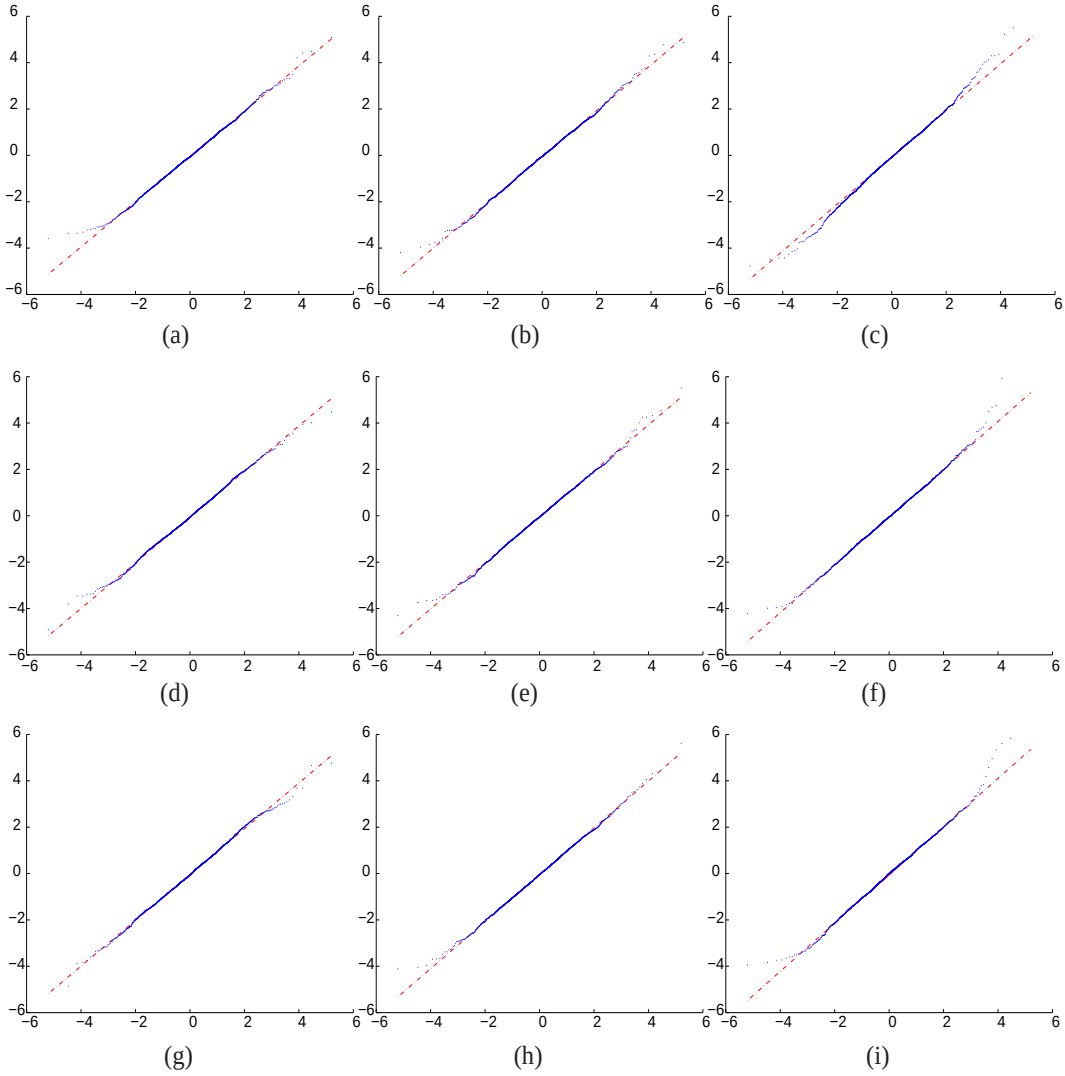
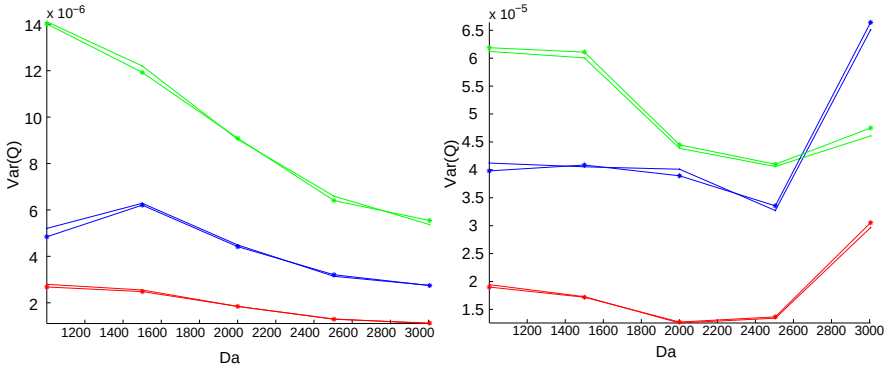
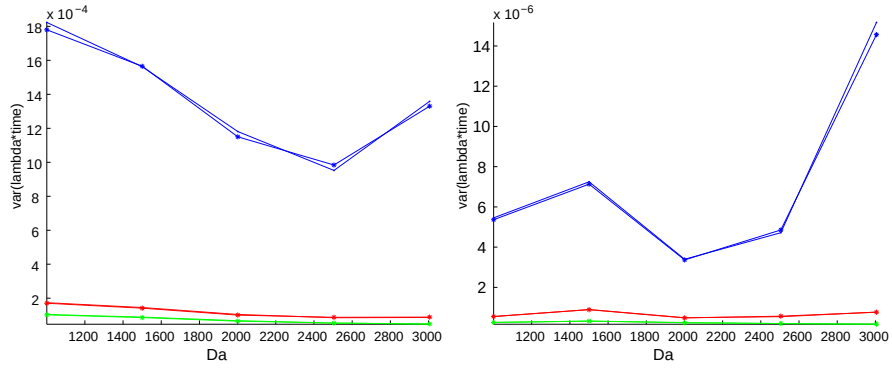


Figure 1: Q-Q-plots of the statistic in equation (16) for the relative abundance Q for the peptide with mass 1000.5 Da . The quantiles of the input sample (y-axis) are plotted against the quantiles of a t-distribution with 10 degrees of freedom (x-axis). Panels (a), (b) and (c) for $Q = 0.5$. Panels (d), (e) and (f) for $Q = 1$. Panels (g), (h) and (i) for $Q = 2$. Panels (a), (d) and (g) for $\lambda = 0.08$. Panels (b), (e) and (h) for $\lambda = 0.02$. Panels (c), (f) and (i) $\lambda = 0.008$.



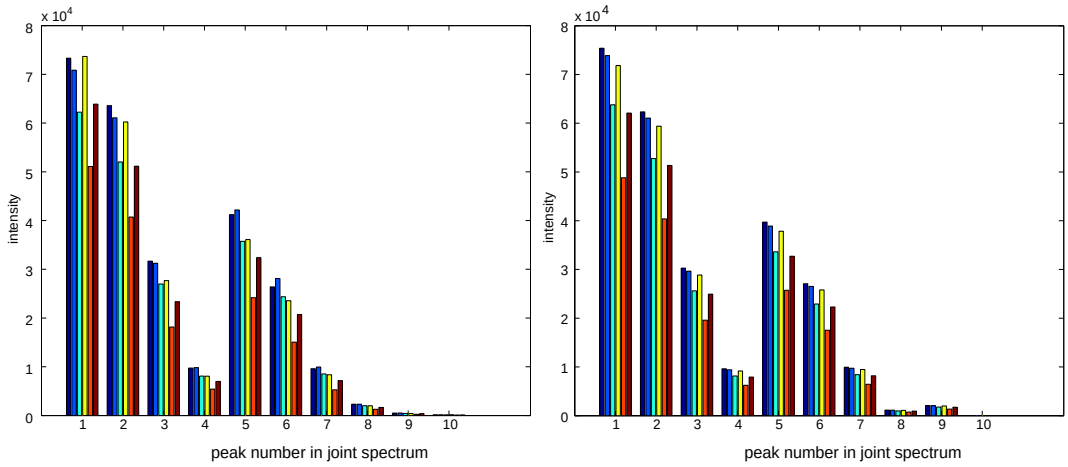
(a) (b)

Figure 2: Empirical variance (indicated by stars) and model-based variance (indicated by dots) for relative abundance Q equal to $1/2$, 1 and 2 (denoted by blue, red and green, respectively) based on two technical replicates. Panel (a): incorporation rate λ fixed at 0.08. Panel (b): incorporation rate λ fixed at 0.02. The duration τ of the enzymatic reaction is equal to 120 min.



(a) (b)

Figure 3: Empirical variance (indicated by stars) and model-based variance (indicated by dots) for incorporation rate λ with ion ratio Q equal to $1/2$, 1 and 2 (denoted by blue, red and green, respectively). Panel (a): incorporation rate λ fixed at 0.08. Panel (b): incorporation rate λ fixed at 0.02. The duration τ of the enzymatic reaction is equal to 120 min.



(a) (b)

Figure 4: Observed (panel (a)) and estimated (panel (b)) peak heights for the six replicated spectrum for the tryptic Bovine Cytochrome C peptide CC1 with mass 1167.61 Da and $Q = 0.33$. Note that the peaks are grouped per isotopic peak.

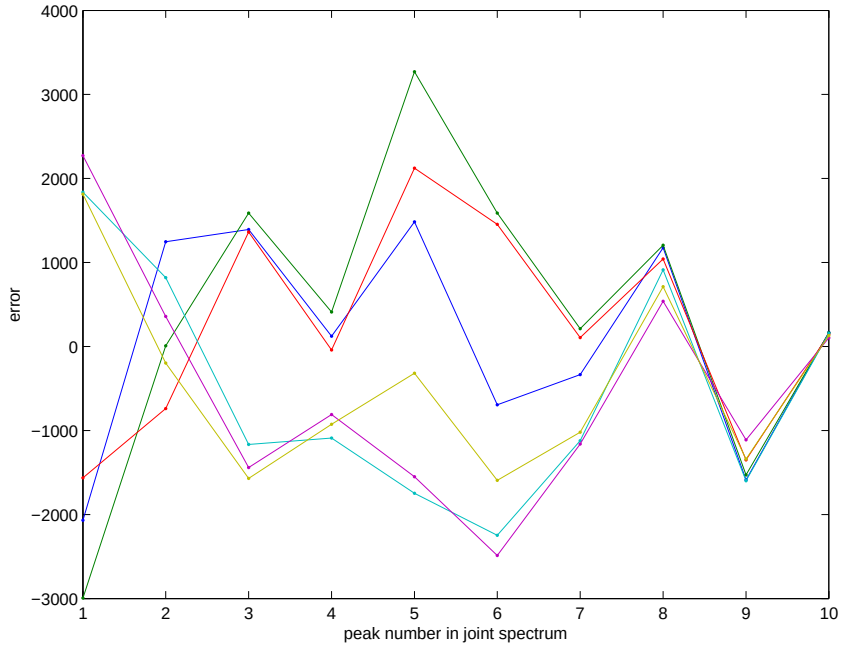
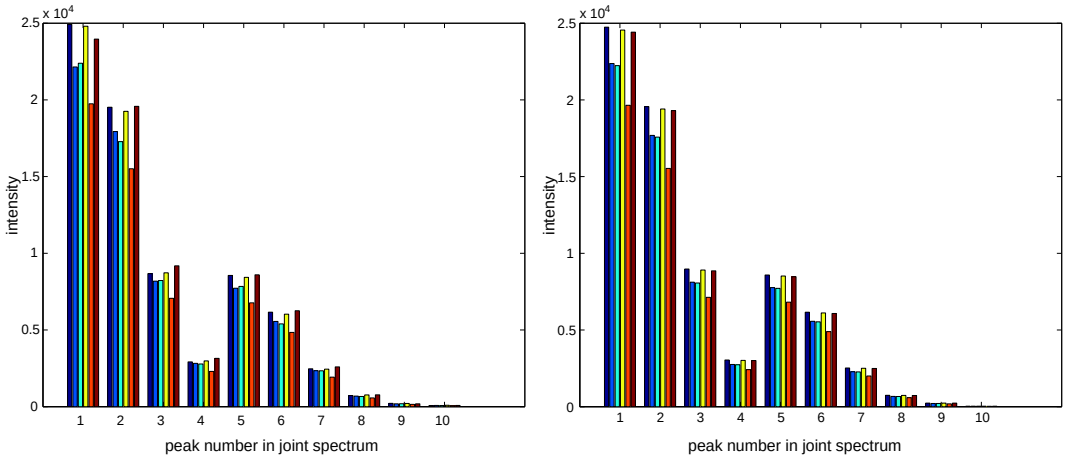


Figure 5: Residuals of peptide CC1 with mass 1167.61 Da and $Q = 0.33$ for the six technical replicates grouped per peak.



(a) (b)

Figure 6: Observed (panel (a)) and estimated (panel (b)) peak heights for the six replicated spectrum for the tryptic Bovine Cytochrome C peptide CC2 with mass $1455{\cdot}66$ Da and $Q = 0{\cdot}33$. Note that the peaks are grouped per isotopic peak.

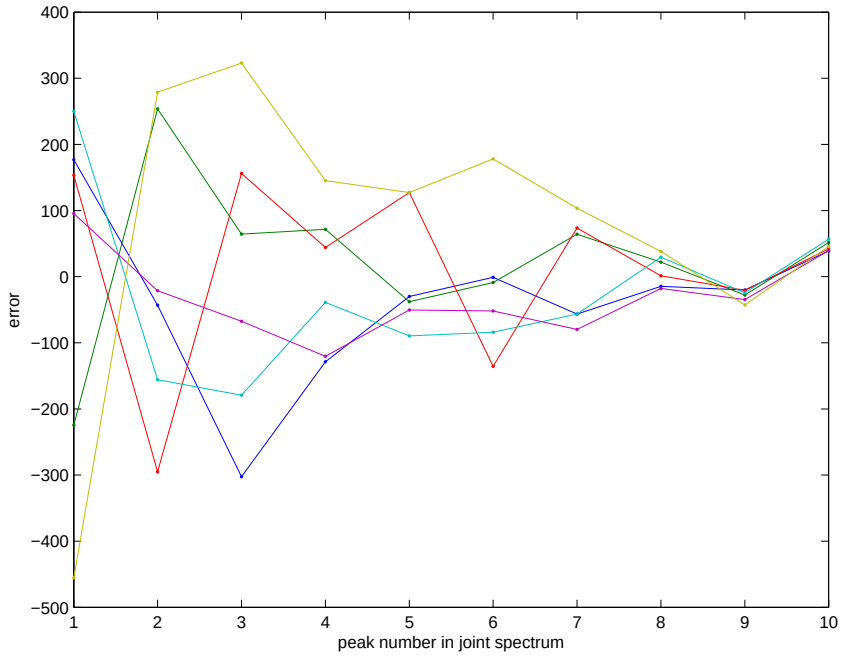
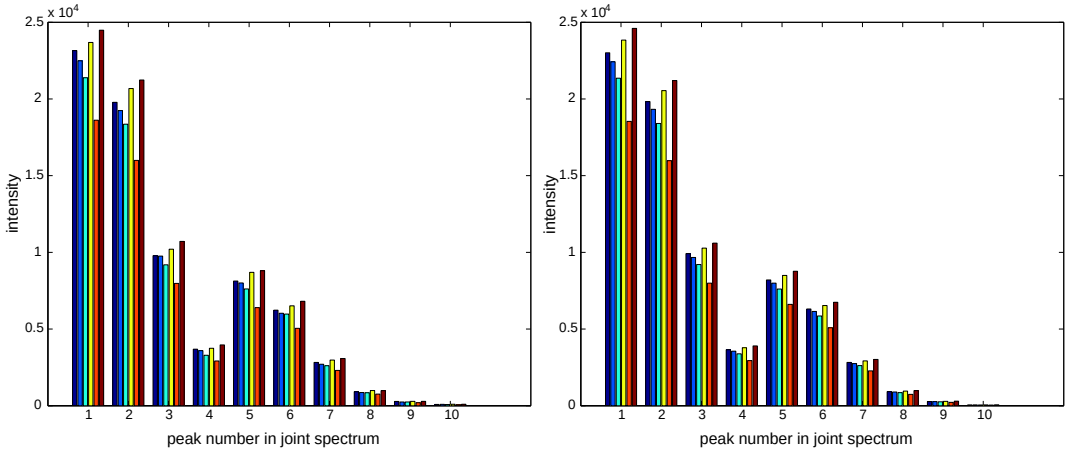


Figure 7: Residuals of peptide CC2 with mass 1455.66 Da and $Q = 0.33$ for the six technical replicates grouped per peak.



(a) (b)

Figure 8: Observed (panel (a)) and estimated (panel (b)) peak heights for the six replicated spectrum for the tryptic Bovine Cytochrome C peptide CC3 with mass 1583.75 Da and $Q = 0.33$. Note that the peaks are grouped per isotopic peak.

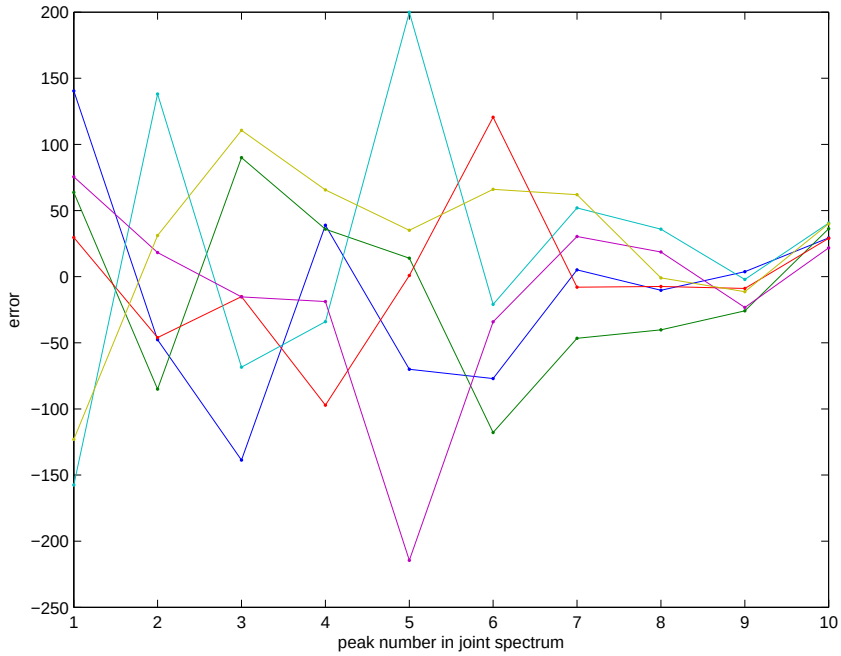
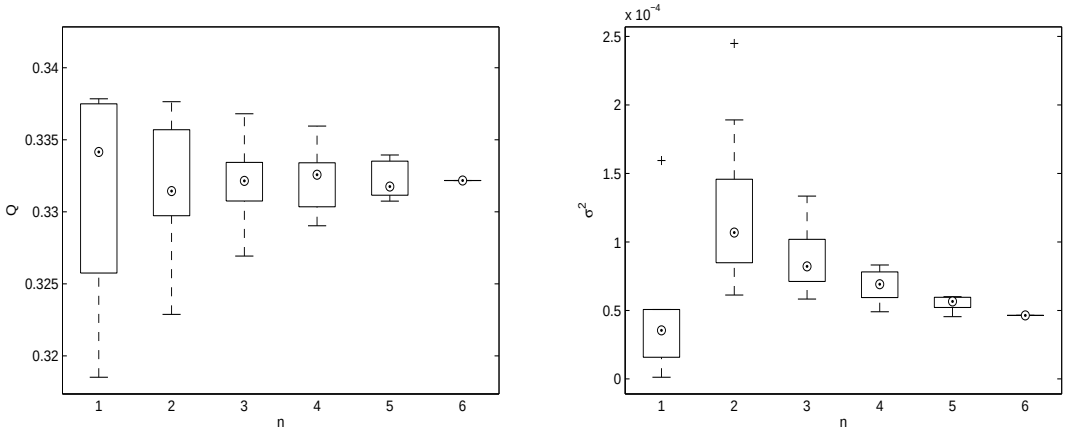
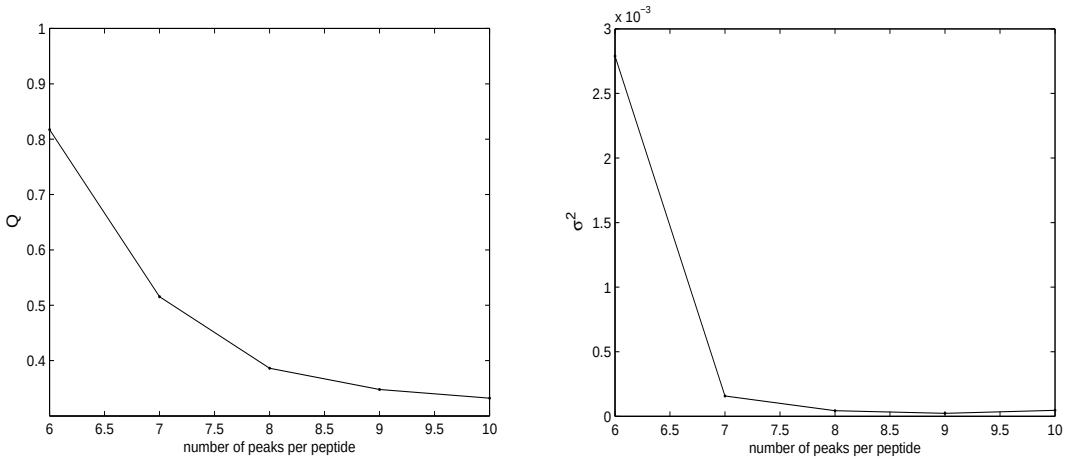


Figure 9: Residuals of peptide CC3 with mass 1583.75 Da and $Q = 0.33$ for the six technical replicates grouped per peak.



(a) (b)

Figure 10: Estimate (panel (a)) and variance (panel (b)) of parameter Q for peptide CC3 for all possible permutations of the six replicates for groups of $n = 1, \ldots, 6$. The x-axis indicates the number of replicates n. Similar results were obtained for peptide CC1 and CC2 (data not shown).



(a) (b)

Figure 11: Estimate (panel (a)) and variance (panel (b)) of parameter Q for peptide CC3 when leaving out the last peptide peaks. The x-axis indicates the number of observed peaks m. Similar results were obtained for peptide CC1 and CC2 (data not shown).

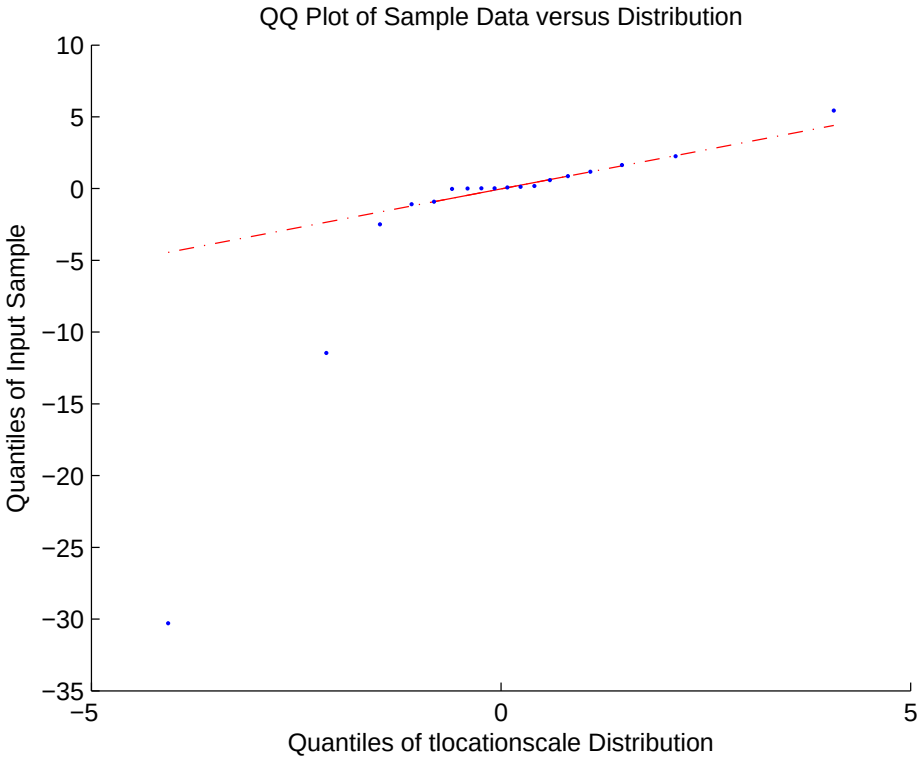


Figure 12: Q-Q-plots of the statistic in equation (16) for the relative abundance Q, reference intensity H and peptide-specific incorporation rate λ for peptide CC3. The quantiles of the input sample (y-axis) are plotted against the quantiles of a t-distribution with 47 degrees of freedom (x-axis). The three outlying points (left-hand side) correspond to the same outlying spectrum.

References

Eckel-Passow, J.E., Oberg, A.L., Therneau, T.M., Mason, C.J., Mahoney, D.W. et al. (2006). Regression analysis for comparing protein samples with ^{16}O / ^{18}O stable-isotope labeled mass-spectrometry. *Bioinformatics*, **22(22)**:2739-2745.

Hartley, H.O. (1961). The modified Gauss-Newton method for the fitting of non-linear regression functions by least squares. *Technometrics* **3**:269-280.

Higham, N. J. (2005). The scaling and squaring method for the matrix exponential revisited. *SIAM J. Matrix Anal. Appl.*, **26(4)**:1179-1193.

Johnson, K.L. and Muddiman, D.C (2004). A method for calculating $^{16}O/^{18}O$ peptide ion ratios for the relative quantification of proteomes. *American Society for Mass Spectrometry*, **15**:437-445.

López-Ferrer, D., Ramos-Fernández, A., Martnez-Bartolomé, S., Garca-Ruiz, P. and Vázquez, J. (2006). Quantitative proteomics using $^{16}O/^{18}O$ labeling and linear ion trap mass spectrometry. *Proteomics*, **6**:S4-S11.

Li, X., Yi, E.C., Kemp, C.J., Zhang, H. and Aebersold, R. (2005). A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Molecular & Cellular Proteomics*, **4**:1328-1340.

Mirgorodskaya, O.A., Kozmin, Y.P., Titov, M.I., Korner, R., Sonksen, C.P. and Roepstorff, P. (2000). Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using ^{18}O-labeled internal standards. *Rapid Communicatios in Mass Spectrometry*, **14**:1226-1232.

Miyagi, M. and Rao, K.C. (2007). Proteolic ^{18}O-labeling strategies for quantitative proteomics. *Mass Spectrometry Reviews*, **26**:121-136.

Pinheiro, J. C. and Bates, D. M. Mixed-Effects Models in S and S-PLUS. (2000). Springer.

Rao, K.C., Carruth, R.T. and Miyagi, M. (2005). Proteolic ^{18}O-labeling by peptidyl-lys metalloendopeptidase for comparative proteomics. *Journal of Proteome Research* **4**:507-514.

Rockwood, A.L. (1995). Relationship of Fourier transforms to isotope distribution calculations. *Rapid Communications in Mass Spectrometry*, **9**:103-105.

Senko, M.W., Beu, S.C. and McLafferty, F.W. (1995). Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distribution. *Journal of the American Society for Mass Spectrometry*, **6**:229–233.

Staes, A., Demol, H., Van Damma, J., Martens, L., Vandekerckhove, J. and Gevaert, K. (2004). Global differential non-gel proteomics by quantitative and stable labeling of tryptic peptides with oxygen-18. *Journal of Proteome Research*, **3**:786-791.

Storms, F.S., Heijden, R., Tjaden, U.R. and Greef, J. (2006). Considerations for proteolytic labeling-optimization of ^{18}O incorporation and prohibition of back-exchange. *Rapid Communications in Mass Spectrometry*, **20**:3491-3497.

Valkenborg, D., Assam, P., Thomas, G., Krols, L., Kas, K. and Burzykowski, T. (2007). Using a Poisson approximation to predict the isotopic distribution of sulphur-containing peptides in a peptide-centric proteomic approach. *Rapid Communications in Mass Spectrometry*, **21**:3387-3391.

Valkenborg, D., Jansen, I. and Burzykowski, T. (2008a). A model-based method for the prediction of the isotopic distribution of peptides. *Journal of the American Society for Mass Spectrometry*, **19(5)**:703-712.

Valkenborg, D., Thomas, G., Krols, L., Kas, K. and Burzykowski, T. (2009). A strategy for the prior processing of high-resolution mass spectral data obtained from high-dimensional combined fractional diagonal chromatography. *Journal of Mass Spectrometry*, **44(4)**:516-529.

Valkenborg, D., Van Sanden, S., Lin, D., Kasim, A., Zhu, Q. et al. (2008b). A cross-validation study to select a classification procedure for clinical diagnosis based on proteomic mass spectrometry. *Statistical Applications in Genetics and Molecular Biology*, **7(2)**:Article 12.

Yergey, J. A. (1983). A general approach to calculating isotopic distributions for mass spectrometry. *International Journal of Mass Spectrometry and Ion Physics*, **52**:337-349.

Ye, X., Luke, B., Andresson, T. and Blonder, J. (2009). 18O stable isotope labeling in MS-based proteomics. *Briefings in Functional Genomics and Proteomics*, **8**:136-144.