J. Am. Soc. Mass Spectrom. (2011) 22:499–507
DOI: 10.1007/s13361-010-0056-x

RESEARCH ARTICLE

A Bayesian Markov-Chain-Based Heteroscedastic Regression Model for the Analysis of ^{18}O-Labeled Mass Spectra

Qi Zhu,[1] Tomasz Burzykowski[2]

[1]Department of Electrical Engineering, ESAT/SCD Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, bus 2446, 3001 Heverlee, Leuven, Belgium
[2]Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Hasselt, Belgium

Abstract

To reduce the influence of the between-spectra variability on the results of peptide quantification, one can consider the ^{18}O-labeling approach. Ideally, with such labeling technique, a mass shift of 4 Da of the isotopic distributions of peptides from the labeled sample is induced, which allows one to distinguish the two samples and to quantify the relative abundance of the peptides. It is worth noting, however, that the presence of small quantities of ^{16}O and ^{17}O atoms during the labeling step can cause incomplete labeling. In practice, ignoring incomplete labeling may result in the biased estimation of the relative abundance of the peptide in the compared samples. A Markov model was developed to address this issue (Zhu, Valkenborg, Burzykowski. *J. Proteome Res*. **9**, 2669–2677, 2010). The model assumed that the peak intensities were normally distributed with heteroscedasticity using a power-of-the-mean variance funtion. Such a dependence has been observed in practice. Alternatively, we formulate the model within the Bayesian framework. This opens the possibility to further extend the model by the inclusion of random effects that can be used to capture the biological/technical variability of the peptide abundance. The operational characteristics of the model were investigated by applications to real-life mass-spectrometry data sets and a simulation study.

Key words: Biological/technical variability, Differential equations, Incomplete labelling, Isotopic distribution, MALDI-TOF MS, Markov chain, Mean-power-variance function, Peptides

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. Peptides are chains of amino acids and are composed of atoms of five chemical elements: carbon (C), hydrogen (H), nitrogen (N), oxygen (O), and sulphur (S). Because the chemical elements have different isotopes, peptides can have different isotopic variants, which differ with respect to their weights. For a peptide of a known chemical composition, the probabilities of occurrence of these variants are called the isotopic distribution. It follows that in a singly-charged high-resolution mass spectrum, a peptide produces a series of peaks that are separated by one mass-to-charge-unit (dalton, Da) that correspond to different isotopic variants of the peptide. These peaks are called isotopic peaks. Their relative heights correspond to the probabilities of the isotopic distribution of the peptide.

Electronic supplementary material The online version of this article (doi:10.1007/s13361-010-0056-x) contains supplementary material, which is available to authorized users.

Correspondence to: Qi Zhu; *e-mail:* qi.zhu@esat.kuleuven.be

Received: 19 August 2010
Revised: 10 December 2010
Accepted: 10 December 2010
Published online: 15 January 2011

The analysis of high-resolution matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry (MS) allows one to separate peptides, present in a sample, according to their mass to charge (*m/z*). It also provides a measure of abundance of the peptides. By comparing the protein abundances for different samples, differentially expressed proteins can be found. By analyzing the proteins, important information about, e.g., mechanisms of disease can be obtained.

When comparing two samples in the high-resolution MALDI-TOF MS, a labeling approach can be considered to reduce the between-spectra variability. The main idea of a labeling approach is similar to, e.g., two-channel cDNA microarrays, where the mRNAs of one sample are labeled with a green fluorescent dye and the mRNAs from the other sample are labeled with a red fluorescent dye. Afterwards, the samples are pooled together and processed simultaneously. The relative abundance between the mRNA molecules is calculated as the ratio of intensities observed after irradiating the fluorescent dyes. Likewise, in MS, the labeled peptides are labeled with a stable isotope, which results in a *m/z* shift and, thus, the labeled and unlabeled peptides are separated with respect to their *m/z* in a mass spectrum. A relatively new and powerful technique for stable isotope labeling is the enzymatic ^{18}O-labeling, where the two oxygen atoms in the carboxyl-terminus of a peptide are replaced with oxygen isotopes from heavy-oxygen-water. The labeling should lead, in ideal circumstances, to an increase of the *m/z* of the labeled peptide molecule by 4 Da. Figure 1 shows the symbolic spectra for two peptides before and after ^{18}O-labeling.

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 peaks observed for the peptide in the joint mass spectrum (see the right-hand side panel of Figure 1), 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 1, some isotopic peaks of the unlabeled peptide will still overlap with the monoisotopic peak of the labeled peptide. Thus, 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, due to water impurities, i.e., the presence of small quantities of ^{16}O- and ^{17}O-atoms, not all of the labeled peptides receive two ^{18}O-atoms. Instead, some may receive a combination of two of the three oxygen isotopes. For those molecules, the isotopic peaks will be shifted by multiples of 1 Da, causing an additional overlap with the peaks of the unlabeled sample. As a result, at the end of the enzymatic reaction, not all peptide molecules from Sample II may have been actually labeled with two ^{18}O-atoms. The isotopic peaks for these molecules will overlap with the peaks from Sample I.

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 from the unlabeled and labeled samples. In order to estimate the relative abundance of the peptide in the two samples, Valkenborg [2] and Zhu *et al.* [1] have proposed a regression approach similar in spirit to the method of Eckel-Passow *et al.* [3]. The approach allows for estimating the relative abundance while correcting for the presence of all the oxygen isotopes in the heavy-oxygen-water by using a flexible model. Moreover, the peptide's isotopic distributions are allowed to be estimated from the observed data.

In this paper, we show an alternative method. In particular, we formulate the model within the Bayesian framework. We use a variance function, which allows for the variance of the observed peak intensity to be equal to a power-of-the-mean function [1]. Using the Bayesian approach opens the possibility of incorporating prior information that could be helpful to analyze the data. In particular, such information exists for the isotopic distribution. Moreover, it allows the inclusion of random effects that can be used to capture the (between-spectra) technical and biological variability. The biological variability results from the variability of the abundance of a particular peptide in different biological samples. The technical variability results from the variability of the measurements of the abundance (intensity measurements) obtained for the same peptide in the same biological sample, but in different spectra.

The paper is organized as follows. "Experimental" presents a real-life data set, to which our model is applied. "Model formulation and practicalompl" gives a detailed description of the model implemented in the Bayesian framework with heteroscedastic residual error and random effects. In "Results and discussion", we evaluate the operational characteristics of the model by an application to real-life data sets and by a simulation study. Finally, in "Conclusions", some concluding remarks are given.

Experimental

We apply the developed method to a data set, which consists of 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 trypsin 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, what should result in the relative abundance ratio of $Q = 1/3$. 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 measurement procedure can be found in Staes *et al.* [4]. The process was repeated for the mixture resulting from mixing of one unit from the unlabeled part and three units from the labeled part, what should result in the relative abundance ratio of $Q = 3/1$.

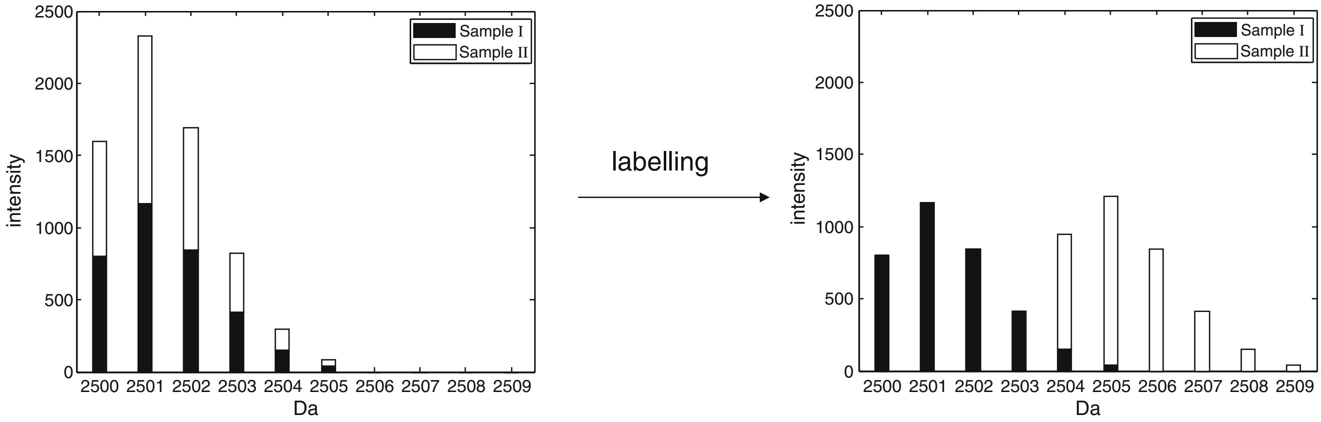


Figure 1. Effect of enzymatic ^{18}O-labeling in a mass spectrum in stick representation. Left panel: "sticks" can be seen as a representation of the distribution the isotopic variants of the peptide. Right panel: in ideal cases, labeling causes Sample II shifted by four Da (the figure was reproduced from Zhu *et al.* [1])

Thus, a total of 12 joint spectra, six for the $Q = 1/3$ and for the $Q = 3/1$ mixture, were available for the analysis.

We assume that, prior to the statistical analysis of a series of peaks observed in a MALDI-TOF spectrum and considered to be corresponding to a peptide, the spectrum was appropriately pre-processed. To this aim, we use the strategy proposed by Valkenborg *et al.* [5]. The pre-processing strategy extracts the information about the *m/z* 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.

We restrict the analysis to three bovine cytochrome *C* peptides, for which joint spectra of acceptable quality were obtained. The amino acid compositions of these peptides are as follows: peptide CC1 (*m/z* 1167.61 Da) - TGPNLHGLFGR; peptide CC2 (*m/z* 1455.66 Da) - TGQAPGFSYTDANK; peptide CC3 (*m/z* 1583.75 Da) - KTGQAPGFSYTDANK. Note that the molecules are protonated by the MALDI-procedure (*mH*$^{+}$). Therefore, the monoisotopic masses should be corrected by adding 1.00783 Da. Figure 2 presents, as an illustrative example, the observed spectra for the peptide at 1456.7 Da for both mixing experiments. Figure 3 presents the corresponding "stick" representation, for the same peptide.

The formulation of the model, based on which our analysis is carried out, was explained in full details by Zhu *et al.* [1]. In what follows, we demonstrate how we account for the random variability of the spectra by including the random effects in the model, together with a description of the practical implementation needed to this aim.

Model Formulation and Practical Implementation

In this section, we introduce the model formulation and practical implementation in the Bayesian framework. We extend the modeling approach, proposed by Valkenborg [2] (see Chapter 10) and Zhu *et al.* [1], by including random effects to account for technical/biological variability.

A Random-Effect(s) Model for the Replicated Joint Spectra

Consider a peptide, which has $l \geq 5$ isotopic variants (including the monoisotopic one). The enzymatic ^{18}O-labeling and mixing of this peptide with its unlabeled counterpart will result in an observed joint spectrum of $l+4$ peaks. The observed peak intensity y_{ij} in the ith replication of the mass spectra, where $j=1, 2, ...,$ denotes the position of the peak in the observed series of peaks in a joint spectrum, with $j=1$ referring to the monoisotopic peak of the peptide from Sample I, will be a function of the abundance of the unobserved isotopic variants of the peptide.

To model the observed peak intensities, we assume that

$$y_{ij} = \mu_{ij} + \varepsilon_{ij}, \tag{1}$$

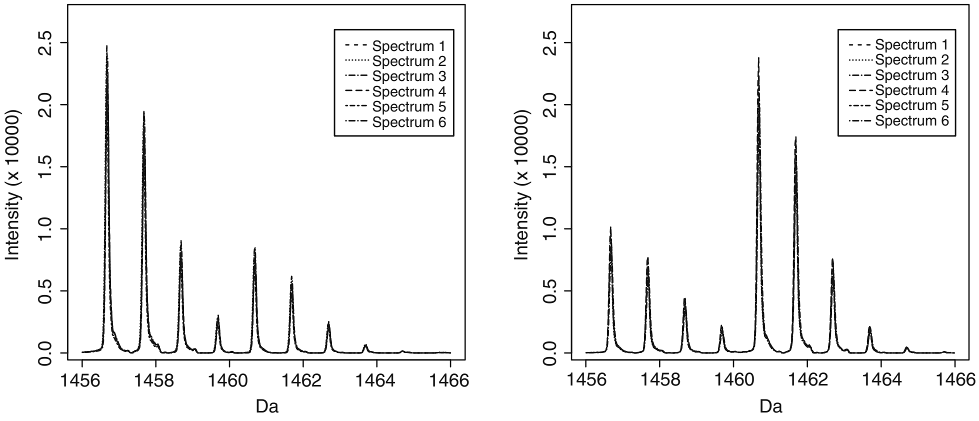
$$\text{where } \varepsilon_{ij} \sim N(0, \sigma^2 \mu_{ij}^{2\theta}), \tag{2}$$

and that ε_{ij}'s are independent. In (2), parameter θ is the power parameter for the variance function, to account for the (mean-dependent) heteroscedastic nature of the MS data. The power-of-the-mean variance function, shown in (2), was suggested as a suitable variance function (1), based on a real-life data exploration. The mean intensity in the model, μ_{ij}, of the jth peak in the ith spectrum is expressed as follows:

$$\mu_{ij} \equiv \mathrm{E}(y_{ij})$$

$$= \begin{cases} H_i R_j + Q_i H_i \sum_{k=0}^{\min(4,j-1)} P_k R_{j-k} & \text{if } 1 \leq j \leq l \qquad (3) \\ Q_i H_i \sum_{k=j-l}^{4} P_k R_{j-k} & \text{if } l+1 \leq j \leq l+4 \end{cases},$$

$$(4)$$



(a) 1456.7Da Q=0.33

(b) 1584.8Da Q=3

Figure 2. Graphical representation of the observed spectra of the six replications, for the peptide at 1456.7 Da

where H_i is the unobserved abundance of the peptide in Sample I, i.e., the unlabeled peptide sample, in the ith spectrum; and Q_i is the relative abundance of the peptide in Sample II (the labeled peptide sample), with respect to Sample I, in that spectrum. Both H_i and Q_i are assumed to be random:

$$H_i \sim N(H, \sigma_H^2), \tag{5}$$

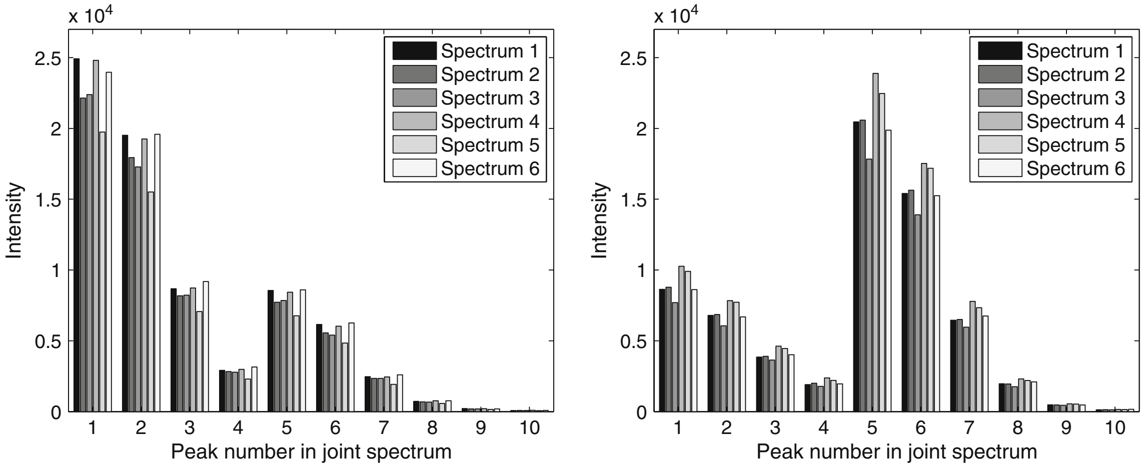
$$\text{and } Q_i \sim N(Q, \sigma_Q^2). \tag{6}$$

In (3), P_k is the *m/z* shift probability, calculated via a Markov-chain model. A brief description of the model is given in the next section. Parameters R_j are the *isotopic ratios*, representing the isotopic distribution. Figure 4 presents a symbolic representation of an isotopic distribution. Let h_1, h_2, h_3, etc., denote the probabilities of occurrence of, respectively, the first, second, third, etc. (with respect to the increasing mass) isotopic variants. We define the *isotopic ratio* of jth isotopic variant as:

$$R_j = h_j / h_1, \text{ where } j = 1, \ldots, l. \tag{7}$$

Instead of replacing the isotopic ratios by fixed values, obtained from an average approximation [3], we leave the ratios to be freely estimated from the data. This offers an advantage of accounting for the deviation of the true distribution around an average one. Finally, terms HQP_kR_{j-1-k} in (3) denote the contributions to the mean values of the observed peaks from the isotopic variants of the peptide from Sample II. Note that, for peaks l+1, ... , l+4, there are no contributions from the unlabeled peptide in Sample I. The isotopic ratios (7) are used for both the unlabeled and labeled peptide, because the ratios depend on the isotopic distribution, which is the same for both peptides.

The parameters of interest are Q, σ_Q^2 and σ_H^2, as they capture, respectively, the (mean) relative abundance of the peptide in the two samples, the biological and technical variability across the replicated mass spectra.



(a) 1456.7Da Q=0.33

(b) 1456.7Da Q=3

Figure 3. Stick representation of the spectra presented in Figure 2, for the peptide at 1456.7 Da. Bars of the same shade represent peaks from the same spectrum

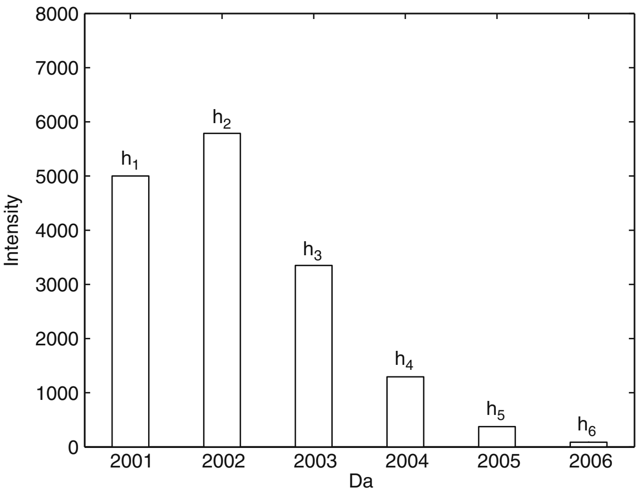


Figure 4. Graphical representation of the isotopic distribution

A Markov-Chain-Model for Enzymatic ^{18}O-Labeling

For the estimation of shift probability P_k, one can think of a Markov-chain-model [1, 2]. Denote the proportions of ^{16}O, ^{17}O, and ^{18}O atoms present in the heavy-oxygen water by p_{16}, p_{17}, and p_{18}, respectively, with $p_{16} + p_{17} + p_{18} = 1$. As a result of 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. The two reaction sites of a carboxyl terminus produce six possible isotope combinations:

$$\begin{aligned} X(1) &= (2,0,0), \quad X(3) = (1,0,1), \quad X(5) = (0,1,1), \\ X(2) &= (1,1,0), \quad X(4) = (0,2,0), \quad X(6) = (0,0,2), \end{aligned} \tag{8}$$

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

For different configurations *X(i)*, peaks corresponding to the isotopic distribution of a labeled peptide will shift with multiples of 1 Da. The resulting probability of a particular shift that follows from the probability distribution of the six possible configurations (states) of the carboxyl-terminus can then be expressed as follows:

$$\begin{aligned} P_0 &= P\{X(1)\}, \quad P_2 = P\{X(3)\} + P\{X(4)\}, \\ P_1 &= P\{X(2)\}, \quad P_3 = P\{X(5)\}, \quad P_4 = P\{X(6)\}, \end{aligned} \tag{9}$$

where P_i indicates the probability of the *m/z* shift of *i* Da (i=0, ..., 4). It should be noted that the *m/z* shifts are defined to be relative to a carboxyl-terminus, which contains two ^{16}O-atoms.

Let T denote the transition matrix for the transitions between the six states, which takes the following form

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

where p_{16} and p_{17} are assumed to be known. Row (i=1, ..., 6) and column (j=1, ..., 6) indices correspond to states X(1), ... , X(6). Element $[T]_{ij}$ of matrix T gives the probability to move from state *X(i)* to state *X(j)*.

Denote by λ the oxygen incorporation rate, which is the number of reactions per time unit, given a reaction time τ in heavy-oxygen water with impurities p_{16} and p_{17}. If we assume that the number of oxygen exchanges follows a Poisson distribution with oxygen incorporation rate λ, the shift probabilities can then be estimated from a Markov-chain-based model via a Poisson process. More specifically, the model takes the form: –1.0 cm

$$S'(\lambda) = S_0 e^{-\lambda\tau} e^{T\lambda\tau}, \tag{11}$$

where $S'(\lambda)$ is the vector containing the state probabilities for the isotope combination of the carboxyl-terminus of a peptide. Now, the shift probabilities, defined in (9), 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 $S_i(\lambda)$ denotes the ith element of the state probability vector $S(\lambda)$.

As a result, the estimation of the shift probabilities depends only on the parameter λ. Figure 5 shows the values

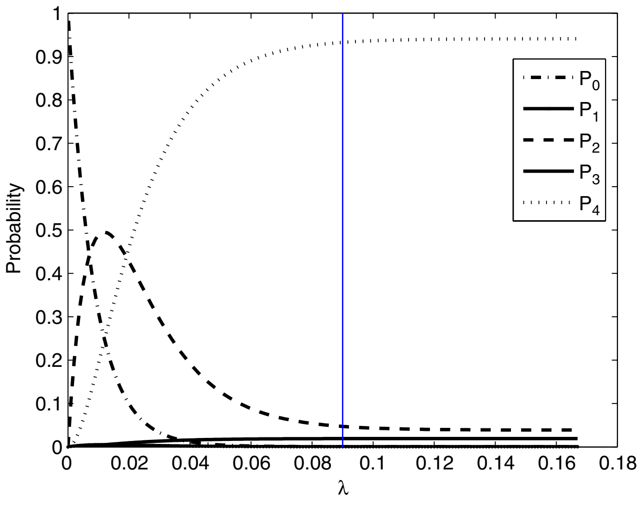


Figure 5. Shift probabilities P_{-0}, P_1, P_2, P_3 and P_4 in function of λ for an enzymatic reaction of 120 minutes with heavy-oxygen water impurities of P_{16}=2% and P_{17}=1%

of the *m/z* shift probabilities as a function of λ for a labeling reaction of τ=120 in heavy-oxygen water with impurities p_{16}=2% an p_{17}=1%. Note that when λ is greater than 0.09, the shift probabilities reach a "plateau". This indicates that when $\lambda \geq 0.09$ (and τ=120), the labeling stablizes, and different values of λ will lead to virtually the same values of shift probabilities.

Bayesian Model Implementation

In the following sections, the implementation of the Markov-chain-based model, using WinBUGS and JAGS, will be introduced.

Prior and Posterior Distributions

The analysis was performed using WinBUGS 1.4 through WBDiff (the interface of differential equations) and JAGS 1.0.3. It should be noted that all the parameters, except of θ, are positive. Thus, the logarithmic transformation can be used for these parameters to work with an unconstrained estimation approach. In addition, for λ to be estimable, it should be bounded in a [0, λ_0] interval (as indicated by the "plateau" in Figure 5). To constrain λ with an upper bound λ_0, a Box-Cox transformation can be considered:

$$\lambda = \frac{\lambda_0 \exp(\lambda')}{\exp(\lambda') + 1}.$$

For practical implementation, non-informative normal priors were defined on the logarithmic scale for parameters H, Q, and the isotopic ratio parameter, R_j. For parameter λ, a non-informative normal prior was used for λ'. For the variance parameters σ^2, σ_H^2, and σ_Q^2, non-informative gamma distributions were used for the inverse of these parameters. As θ can take any real value, a non-informative normal prior was defined on its original scale.

Since the variance function of the model, shown in (2), is dependent on the mean structure parameters, there are no closed form posterior distributions for the parameters. As a result, the posterior distributions need to be evaluated by numerical (sampling) methods, e.g., via Metropolis-Hasting algorithm with acception-rejection rules.

WinBUGS through WBDiff

Since the implementation of the model entails the estimation of Markov Chain transition probabilities through matrix exponential, the Bayesian model cannot be implemented directly through WinBUGS. However, WBDiff, which namely is a WinBUGS Differential Interface, makes such application in WinBUGS feasible. WBDiff is built for WinBUGS to do differential equations, and hence can handle matrix exponential as well. The software can be downloaded from http://www.winbugs-development.org.uk/wbdiff.html with user manual available from Lunn [6].

To implement the matrix exponential, shown in (11) via differential equations, let $\pi(t)$ denote the vector of the transition rate, given time t, such that $S(\tau) = S_0 e^{-\lambda\tau} e^{T\lambda\tau} = \frac{\partial \pi(t)}{\partial t} e^{-\lambda\tau}$. The vector $\pi(t) = (\pi_1(t), \pi_2(t), \ldots, \pi_6(t))$ contains transition rates related to the six states of oxygen combinations for the carboxyl-terminus, expressed in (8). Then the matrix exponential can be written down as the differential equations shown as follows:

$$\left\{ \begin{array}{lll}
\frac{\partial \pi_1(t)}{\partial t} & = \pi_1(t)T_{11}\lambda\tau + \pi_2(t)T_{21}\lambda\tau + \pi_3(t)T_{31}\lambda\tau & (13) \\
\frac{\partial \pi_2(t)}{\partial t} & = \pi_1(t)T_{12}\lambda\tau + \pi_2(t)T_{22}\lambda\tau + \pi_3(t)T_{32}\lambda\tau + \pi_4(t)T_{42}\lambda\tau + \pi_5(t)T_{52}\lambda\tau & (14) \\
\frac{\partial \pi_3(t)}{\partial t} & = \pi_1(t)T_{13}\lambda\tau + \pi_2(t)T_{23}\lambda\tau + \pi_3(t)T_{33}\lambda\tau + \pi_5(t)T_{53}\lambda\tau + \pi_6(t)T_{63}\lambda\tau & (15) \\
\frac{\partial \pi_4(t)}{\partial t} & = \pi_2(t)T_{24}\lambda\tau + \pi_4(t)T_{44}\lambda\tau + \pi_5(t)T_{54}\lambda\tau & (16) \\
\frac{\partial \pi_5(t)}{\partial t} & = \pi_2(t)T_{25}\lambda\tau + \pi_3(t)T_{35}\lambda\tau + \pi_4(t)T_{45}\lambda\tau + \pi_5(t)T_{55}\lambda\tau + \pi_6(t)T_{65}\lambda\tau & (17) \\
\frac{\partial \pi_6(t)}{\partial t} & = \pi_3(t)T_{36}\lambda\tau + \pi_5(t)T_{56}\lambda\tau + \pi_6(t)T_{66}\lambda\tau & (18)
\end{array} \right.$$

The differential equations in (18) can be implemented in WinBUGS using the WBDiff as an interface.

JAGS-Just Another Gibbs Sampler

Alternatively, the matrix exponential can be implemented in JAGS. In JAGS 1.0.3, matrix exponential is defined as an internal function mexp by loading the msm module (User manual available [7]).

Results and Discussion

In this section, we present results of an application of the model to the controlled experiment of the six replicated mass

Table 1. Results of the analysis of the data for peptide at 1168.6 Da (95% c.i.: 95% credible interval)

Parameter	Q = 1/3			Q = 3/1		
	TRUE	Median	95% c.i.	Median	95% c.i.	
H	–	69230	(57270, 81200)	–	22000	(14500, 28600)
	–	128700000	(39500000, 736800000)	–	45340000	(15640000, 209200000)
Q	0.3333	0.4852	(0.4087, 0.5703)	2.4	2.1364	(2.0307, 2.2426)
	–	0.0065	(0.0021, 0.0332)	–	0.0073	(0.0023, 0.0361)
$\lambda\tau$	–	5.3220	(4.6548, 6.2460)	–	19.2720	(16.0440, 19.8720)
σ	–	1.1476	(0.5704, 2.5538)	–	0.0999	(0.0541, 0.2940)
θ	–	0.7461	(0.6583, 0.8279)	–	0.9362	(0.8233, 1.0044)
R_2	0.6645	0.7617	(0.7167, 0.8133)	0.6645	0.6970	(0.6618, 0.7223)
R_3	0.2454	0.3048	(0.2798, 0.3303)	0.2454	0.2813	(0.2689, 0.2947)
R_4	0.0653	0.0554	(0.0482, 0.0643)	0.0653	0.0686	(0.0647, 0.0737)
R_5	0.0139	0.0095	(0.0074, 0.0119)	0.0139	0.0124	(0.0115, 0.0134)
R_6	0.0025	0.0014	(0.0009, 0.0020)	0.0025	0.0017	(0.0015, 0.0020)

spectra of bovine cytochrome *C* peptides. We also show results of a simulation study, undertaken to check the statistical properties of the proposed model.

Bovine Cytochrome C Data Sets

The model was applied to three peptide fragments (at 1168.6 Da, 1456.7 Da, and 1584.8 Da, respectively) of the replicated mass spectra of bovine cytochrome *C*. The proportions of water impurities of the heavy-oxygen water were assumed to be equal to p_{16}=2% and p_{17}=1%. The true values for the isotopic ratios R_j, j=1, ⋯, 5 were calculated via Fourier transform based on the atomic compositions of the three peptides [8]. As the duration of the experiment is not known, we estimate products $\lambda\tau$ instead of λ.

The model was fitted to the data by using WinBUGS 1.4 through WBDiff. Tables 1 and 2 show the results of fitting of the model, defined by (1)–(6) and (11)–(12), based on 10,000 samples after 10,000 burn-in. For illustrative purposes, the results are only shown for peptides with masses 1168.6 and 1456.7 Da, since the results of peptide at 1584.8 Da are quite similar to the ones for peptide at 1456.7 Da. It is important to mention that, despite the efforts to control the experiment, it appears that the achieved value of relative abundance Q was about 2.4, not 3. The value was estimated by using models for the analysis of ^{18}O-labeled mass spectra [1, 3]. This value was therefore assumed as a true relative abundance in Tables 1 and 2.

Several common characteristics can be observed in these tables. First, for each peptide, a considerable amount of technical variability, represented by σ_H^2, is worth noting. This demonstrates the advantage of using the ^{18}O-labeling strategy, since the variability can be removed from the comparison of the peptide abundance in the labeled and unlabeled samples. Moreover, compared with σ_H^2, the magnitude of between-spectra variability for the relative abundance, denoted as σ_Q^2, is negligible. This is because, for each of the peptide, the two samples from the six spectra are the same. Thus, in principle, they should produce exactly the same value for the relative abundance parameter.

Table 2. Results of the analysis of the data for peptide at 1456.7 Da (95% CI: 95% credible interval.)

Parameter	Q = 1/3			Q = 3/1		
	TRUE	Median	95% c.i.	Median	95% c.i.	
H	–	23000	(20780, 25010)	–	8877	(7472, 9787)
	–	4506000	(1503000, 23250000)	–	891000	(303500, 4687000)
Q	0.3333	0.3308	(0.3034, 0.3698)	2.4	2.3705	(2.3062, 2.4536)
	–	0.0041	(0.0015, 0.0177)	–	0.0055	(0.0018, 0.0272)
$\lambda\tau$	–	6.8532	(6.1824, 7.8924)	–	11.1600	(10.2480, 12.6720)
σ	–	0.3606	(0.1613, 0.8165)	–	0.4426	(0.1756, 1.3149)
θ	–	0.6944	(0.5961, 0.8021)	–	0.6444	(0.5182, 0.7595)
R_2	0.7933	0.7809	(0.7659, 0.7953)	0.7933	0.7775	(0.7684, 0.7865)
R_3	0.3567	0.3336	(0.3269, 0.3409)	0.3567	0.3432	(0.3380, 0.3486)
R_4	0.1166	0.0909	(0.0872, 0.0949)	0.1166	0.1014	(0.0988, 0.1039)
R_5	0.0306	0.0212	(0.0199, 0.0228)	0.0306	0.0245	(0.0234, 0.0255)
R_6	0.0068	0.0058	(0.0052, 0.0064)	0.0068	0.0070	(0.0066, 0.0075)

Table 3. Simulation results of the two settings – Mean estimate (M.Est.), mean relative bias (M.R.B.), empirical standard error S_{emp}, and model-based standard error S_{mb}

Parameter	Setting 1			Setting 2		
	M.Est.	M.R.B.	$S_{\mathrm{emp}}/S_{\mathrm{mb}}$	M.Est.	M.R.B	$S_{\mathrm{emp}}/S_{\mathrm{mb}}$
R_2	0.9110	0.0006	0.0034/0.0035	0.9095	-0.0010	0.0037/0.0060
R_3	0.4186	0.0100	0.0020/0.0020	0.4277	0.0320	0.0024/0.0037
R_4	0.1261	0.0022	0.0013/0.0013	0.1258	0.0002	0.0009/0.0013
R_5	0.0288	0.0059	0.0005/0.0006	0.0290	0.0114	0.0004/0.0006
R_6	0.0052	0.0029	0.0002/0.0002	0.0053	0.0157	0.0001/0.0002
μ_H	24030.5	0.0013	933.7/997.2	8080.1	0.0100	306.7/441.7
σ_H	2024.6	-0.0359	671.5/726.5	853.8	-0.0298	290.1/353.7
μ_Q	0.3460	0.0381	0.0984/0.1160	2.8887	-0.0371	0.2498/0.2686
σ_Q	0.0477	-0.0450	0.0129/0.0175	0.4762	-0.0475	0.1723/0.1690
σ	0.4533	0.1332	0.1828/0.2496	0.3563	-0.1093	0.1169/0.1813
θ	0.5955	-0.0076	0.0554/0.0567	0.6405	0.0675	0.0480/0.0763
$\lambda\tau$	7.9933	-0.0484	0.3317/0.4205	7.8722	-0.0628	0.0633/0.0761

For each of the peptides, especially for peptides with *m/z* 1456.7 Da and 1584.8 Da, the estimated values of θ from the two experiments with relative abundances 3/1 and 1/3, are very close to each other, by taking into account the precision expressed as the 95% credible interval. This suggests the chosen functional form of dependence of residual variance on the intensity was appropriate.

It is worth noting that, for the peptides with *m/z* 1584.8 Da and 1456.7 Da, the point estimates for Q and for the isotopic ratios R, are very close to the true values. Moreover, for the two peptides, the estimates of $\lambda\tau$, for different relative abundance values, are also very similar. For the peptide with *m/z* 1168.6 Da, the estimates for Q and R differ substantially from the true values. These patterns agree with the results obtained by [1].

Simulation Study

In this section, we present a simulation study of a setting with biological variability. The parameters in this simulation study, except of σ_Q, were chosen based on the values estimated for the peptide with *m/z* 1584.8 Da in the case study. The observed intensity values in the generated data sets were based on (1)–(6) and (11), (12), and truncated to zero if the summed intensity values appeared to be negative. To avoid numerical problems related to zero intensity values for the least abundant peaks, σ_Q was chosen as a compromise between the full representation of between-biological-sample variability (to be large enough) and the occurrence of numerical problems (to be small enough). In the simulations, two settings with two different relative abundances were considered, each with 100 data sets. For each data set, six biological replicates of mass spectra were assumed to be available. The two settings were:

$$\begin{array}{lllll} \text{Setting 1}: & Q = 1/3, & \sigma_Q = 0.05, & H = 24000, & \sigma_H = 2100, \\ \text{Setting 2}: & Q = 3, & \sigma_Q = 0.5, & H = 8000, & \sigma_H = 880. \end{array} \tag{19}$$

The other parameters were chosen as follows: σ=0.40, θ=0.60, $\lambda\tau$=8.4, M=1584.76 Da.

We chose the isotopic ratios to be the ratios from the average isotopic distribution estimated at *m/z* M= 1584.76 Da by a Poisson approximation [9]. The simulation was preformed using R2WinBUGS, the interface to call the application of WinBUGS 1.4 via R, with the discrete-time Markov-chain-based model implemented through WBDiff.

The results of the simulation study are presented in Table 3. The point estimates of the isotopic ratios, $\lambda\tau$, as well as mean relative abundance Q are very close to the true values. Regarding the parameters that reflect the technical and biological variability, i.e., σ_H and σ_Q, they are well-estimated with negligible bias. The estimates of power-of-the-mean variance function parameters θ and σ also correspond to their true values.

Conclusions

Several methods have already been proposed to analyze data from enzymatic ^{18}O-labeling experiments [3, 10–13]. Most of them, however, postulate the use of additional experimental steps, which is an important limitation. The method used in our analysis does not require such steps. It is similar in spirit to the approach developed by Eckel-Passow [3], but extends it mainly in three ways. First, it accounts for the possible presence of ^{17}O-atoms in the heavy-oxygen water. Second, instead of using an average isotopic distribution developed by Senko *et al.* [14], the method allows for the isotopic distribution to be estimated from the data and thus accounting for the possible deviation around the average isotopic distribution. Last but not least, the method is developed based on a unified modeling framework, by estimating all the parameters simultaneously, avoiding incomparability of parameters estimated in a multi-stage analysis.

We formulated the model proposed by Zhu *et al.* [1], in the Bayesian framework. Using the Bayesian approach allows for the incorporation of prior information that could be helpful to analyze the data. In particular, such information exists for the isotopic distribution. Moreover, it accounts for heteroscedastic residual error and technical/biological variability of peptide abundance in MS data by including the random effects. We assessed the performances of the model via both a real-life data application and a simulation study.

The application of the model to the bovine cytochrome *C* data set gives, in general, unbiased estimation, except for the peptide with *m/z* 1168.6 Da. The estimation bias for this peptide may be caused by the quality of MS measurements in the available spectra by some experimental factors unknown to us [1]. The results of the application of our method to simulated data, accounting for both technical and biological variability, confirm feasibility and satisfactory performance of the proposed modeling approach, under the correct model assumptions.

The approach can be extended in several ways. For instance, the inclusion of informative priors for the Bayesian model can be implemented. This should yield precision gain for the estimation of parameters. Moreover, the model can be formulated by considering the shape of the peaks by means of a suitable function instead of using the stick representation. These are the topics for further research.

Acknowledgments

The authors gratefully acknowledge the financial support from the IAP Research Network P6/03 of the Belgian Government (Belgian Science Policy).

References

1. Zhu, Q., Valkenborg, D., Burzykowski, T.: A Markov-chain-based heteroscedastic regression model for the analysis of high-resolution enzymatically ^{18}O-labeled mass spectra. *J. Proteome Res.* **9**(5), 2669–2677 (2010)
2. Valkenborg, D.: Ph.D. dissertation. *Statistical Methods for the Analysis of High-Resolution Mass Spectrometry Data.* I-BIOSTAT, Hasselt University, Belgium (2008)
3. Eckel-Passow, J., Oberg, A., Therneau, T., Mason, C., Mahoney, D., Johnson, K., Olson, J., Bergen III, H.: Regression analysis for comparing protein samples with ^{16}O/^{18}O stable-isotope labeled mass-spectrometry. *Bioinformatics* **22**, 2739–2745 (2006)
4. Staes, A., Demol, H., Van Damma, J., Martens, L., Vandekerckhove, J., Gevaert, K.: Global differential non-gel proteomics by quantitative and stable labeling of tryptic peptides with oxygen-18. *J. Proteome Res.* **3**, 786–791 (2004)
5. Valkenborg, D., Thomas, G., Krols, L., Kas, K., Burzykowski, T.: A strategy for the prior processing of high-resolution mass spectral data obtained from high-dimensional combined fractional diagonal chromatography. *J. Am. Soc. Mass Spectrom.* **44**, 516–529 (2009)
6. Lunn, D.: WinBUGS Differential Interface—Worked Examples. Imperial College of Medicine, London (2004)
7. Plummer, M.: *JAGS* version 1.0.3 manual 2004. http://www-ice.iarc.fr/~martyn/software/jags/jags_user_manual.pdf.
8. Rockwood, A.: Relationship of fourier transforms to isotope distribution calculations. *Rapid Commun. Mass Spectrom.* **9**, 103–105 (1995)
9. Breen, E.J., Hopwood, F.G., Williams, K.L., Wilkins, M.R.: Automatic Poisson peak harvesting for high throughput protein identification. *Electrophoresis* **21**, 2243–2251 (2000)
10. Rao, K., Carruth, R., Miyagi, M.: Proteolic ^{18}O-labeling by peptidyl-lys metalloendopeptidase for comparative proteomics. *J. Proteome Res.* **4**, 507–514 (2005)
11. Mirgorodskaya, O., Kozmin, Y., Titov, M., Korner, R., Sonksen, C., Roepstorff, P.: Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using ^{18}O-labeled internal standards. *Rapid Comun. Mass Spectrom.* **14**, 1226–1232 (2000)
12. López-Ferrer, D., Ramos-Fernández, A., Martnez-Bartolomé, S., Garca-Ruiz, P., Vázquez, J.: Quantitative proteomics using $^{16}O/^{18}O$ labeling and linear ion trap mass spectrometry. *Proteomics* **6**, S4–S11 (2006)
13. Ramos-Fernández, A., López-Ferrer, D., Vázquez, J.: Improved method for differential expression proteomics using trypsin-catalyzed ^{18}O-labeling with a correction for labeling efficiency. *Mol. Cell. Proteom.* **6**, 1274–1286 (2007)
14. Senko, M., Beu, S., McLafferty, F.: Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distribution. *J. Am. Soc. Mass Spectrom.* **6**, 229–233 (1995)