Regression analysis for comparing protein samples with \(^{16}\)O/\(^{18}\)O stable-isotope labeled mass spectrometry

J. E. Eckel-Passow\(^1\), A. L. Oberg, T. M. Therneau, C. J. Mason\(^2\), D. W. Mahoney, K. L. Johnson\(^2\), J. E. Olson\(^1\) and H. R. Bergen III\(^2\)

Division of Biostatistics, \(^1\)Division of Epidemiology, Department of Health Sciences Research and \(^2\)Mayo Proteomics Research Center, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA

Received on August 1, 2006; revised on August 23, 2006; accepted on August 26, 2006
Advance Access publication September 5, 2006
Associate Editor: Martin Bishop

ABSTRACT

Motivation: Using stable isotopes in global proteome scans, labeled molecules from one sample are pooled with unlabeled molecules from another sample and subsequently subjected to mass-spectral analysis. Stable-isotope methodologies make use of the fact that identical molecules of different stable-isotope compositions are differentiated in a mass spectrometer and are represented in a mass spectrum as distinct isotopic clusters with a known mass shift. We describe two multivariable linear regression models for \(^{16}\)O/\(^{18}\)O stable-isotope labeled data that jointly model pairs of resolved isotopic clusters from the same peptide and quantify the abundance present in each of the two biological samples while concurrently accounting for peptide-specific incorporation rates of the heavy isotope. The abundance measure for each peptide from the two biological samples is then used in downstream statistical analyses, e.g. differential expression analysis. Because the multivariable regression models are able to correct for the abundance of the labeled peptide that appear as an unlabeled peptide due to the inability to exchange the natural C-terminal oxygen for the heavy isotope, they are particularly advantageous for a two-step labeling procedure. We discuss how estimates from the regression model are used to quantify the variability of the estimated abundance measures for the paired samples. Although discussed in the context of \(^{16}\)O/\(^{18}\)O stable-isotope labeled data, the multivariable regression models are generalizable to other stable-isotope labeled technologies.

Contact: eckel@mayo.edu

1 INTRODUCTION

Differential proteomic technologies that allow global comparisons of protein profiles across two disease groups have enormous potential for identifying molecular markers that are able to predict disease diagnosis, prognosis or that can enhance treatment therapeutics. Global protein profiling requires the ability to detect and quantify all proteins that are present in a tissue or fluid sample at a single point in time. One of the primary objectives in differential proteomics is to estimate the difference in abundance of peptides or proteins across two or more disease groups of interest. Using mass-spectrometry technologies for this task entails (1) identifying all peptides associated with a protein and generating a list of all proteins in each sample and (2) quantifying the abundance of each detectable protein in each sample. This paper addresses the important task of quantification from mass-spectral data.

Stable-isotope labeling is a technique that allows two or more samples to be analyzed simultaneously in a single mass-spectrometry run. Comparing two or more samples in a single run versus comparing samples run independently (label-free) can limit variability due to issues related to chromatography, sampling variability, ion suppression, etc. Stable-isotope labeling relies on the fact that introducing a change in molecular mass via incorporation of isotopes does not change the chromatographic properties of a peptide. Herein, our focus is on \(^{16}\)O/\(^{18}\)O stable-isotope labeling, which involves trypsin catalyzed exchange of two C-terminal \(^{16}\)O atoms with \(^{18}\)O (Yao et al., 2001). Measuring the abundance of the signals from the \(^{16}\)O- and \(^{18}\)O-labeled peptides provides information on the relative abundances of the peptide in the two samples.

In \(^{16}\)O/\(^{18}\)O stable-isotope labeling both samples undergo enzymatic digestion, one of the samples in the presence of \(^{16}\)O water (unlabeled sample) and the other in \(^{18}\)O water (labeled sample). The natural catalytic activity of serine proteases (e.g. trypsin, Lys-C) can exchange both C-terminal oxygen atoms with an oxygen atom from water in the surrounding solution. For example, by performing trypsin digestion (which cleaves on the carboxy side of lysine or arginine residues) in \(^{18}\)O water both carboxy \(^{16}\)O atoms are replaced with \(^{18}\)O atoms.

There are two general protocols for applying the \(^{18}\)O label, which we refer to as the one-step and two-step labeling procedures. The one-step labeling procedure performs digestion of the unlabeled sample in \(^{16}\)O water and the labeled sample in \(^{18}\)O water; subsequently the two samples are pooled and subjected to chromatographic separation and mass-spectral analysis. In contrast, in the two-step labeling procedure both samples are initially digested in \(^{16}\)O water and then the labeled sample is exchanged with immobilized trypsin in \(^{18}\)O water; subsequently the two samples are pooled (Yao et al., 2003). The two-step process allows for a higher trypsin concentration without contamination. In the one-step protocol, the probability that an exchange occurs is equal to one; however, this is not true in the two-step protocol. In fact, the probability of an exchange in the \(^{18}\)O water for the two-step protocol may be <1, which our regression model accounts for.

\(^{1}\)To whom correspondence should be addressed.
While the isotopically-labeled and -unlabeled peptides from the two samples are not separated chromatographically, they are resolved in a mass spectrometer via a 2- or 4-Da mass shift that is induced by the presence of one or two $^{18}$O atoms in the labeled sample. An example of an isotopic distribution for an unlabeled peptide is provided in Figure 1a. In comparison, isotopic labeling with 90% pure $^{18}$O water results in a mixture of three possible isotopic distributions for each peptide: (1) unlabeled peptide molecules (denoted $^{16}$O or $^{18}$O$_0$) that are not able to exchange a naturally occurring C-terminal oxygen atom for the heavier $^{18}$O form and cannot be distinguished from the unlabeled sample (Fig. 1b: m/z = 956 and associated $^{13}$C isotopes), (2) singly-labeled peptide molecules (denoted $^{18}$O$_1$) that display a 2-Da mass shift from the unlabeled sample (Fig. 1b: m/z = 957 and associated $^{13}$C isotopes) and (3) doubly-labeled peptides (denoted $^{18}$O$_2$) that display a 4-Da mass shift from the unlabeled sample (Fig. 1b: m/z = 958 and associated $^{13}$C isotopes). The $^{18}$O incorporation rate differs from peptide-to-peptide according to factors such as the length of the peptide, the C-terminal sequence and the amount of back-exchange (Stewart et al., 2001; Schnell et al., 1996). In fact, the peptide-specific incorporation rate of the $^{18}$O label is estimated using the $^{18}$O$_1$ and $^{18}$O$_2$ components, which is discussed in detail below. For use in labeling experiments, the goal is to drive the reaction to completion and thus incorporation of two $^{18}$O atoms in the labeled sample.

The joint distribution of the unlabeled and labeled samples for a peptide is a set of three overlapping isotopic distributions that constitute 0 ($^{18}$O$_0$), 1 ($^{18}$O$_1$) or 2 ($^{18}$O$_2$) $^{18}$O incorporations (Fig. 1c). Quantification is performed by measuring all of the peak heights (abundances) in the isotopic distribution together with peak height information from the naturally occurring isotopes—mostly due to $^{13}$C. The naturally occurring isotopes result in overlapping peaks of the $^{18}$O$_0$, $^{18}$O$_1$ and $^{18}$O$_2$ labeled distributions, which complicates quantification. The joint distribution provided in Figure 1c is what we would expect from a pair of samples with an even (1:1) ratio utilizing 90% pure $^{18}$O water. The ratio is very difficult to quantify ‘by eye’, suggesting the need for automated algorithms. Furthermore, some peptides will have very low $^{18}$O incorporation rates; an incomplete reaction will give more abundance in the leftmost peaks of the distribution, further complicating the interpretation of a two-step experiment. Estimation of the incorporation rate is a part of our fitting procedure.

This paper addresses one step in the analysis of $^{16}$O/$^{18}$O data, fitting a model to peak height data such as shown in Figure 1c for quantification purposes. The regression framework coalesces and builds upon the work of Mirgorodskaya et al. (2000) and Johnson and Muddiman (2004). In contrast to methods that use a univariate approach to estimate the contribution of the unlabeled and labeled samples (Hicks et al., 2005; Johnson and Muddiman 2004), our method follows Mirgorodskaya et al. (2000) and uses a multivariable regression approach that concurrently models the joint isotopic distribution of the labeled and unlabeled samples. Our multivariable regression model differs from Mirgorodskaya et al. (2000) in two important aspects. First, our regression model does not require running the labeled sample independently in order to obtain the expected distribution of the labeled peptide that accounts for peptide-specific incorporation rates of the $^{18}$O label. Following the approach of Johnson and Muddiman (2004), we utilize the average amino acid averagine—which only requires a peptide’s molecular mass—to approximate the chemical composition. Second, the overall abundances for a peptide in the two samples are directly adjusted for the peptide’s incorporation rate of the $^{18}$O label. The incorporation rate is estimated directly from the multivariable regression model.

2 MATERIALS AND METHODS

Human transferrin and bovine serum albumin were trypsin digested together in either 90% $^{18}$O water or 100% $^{16}$O water and then mixed 1:1. Subsequently, the mixed sample was analyzed by liquid chromatography Fourier-transform ion cyclotron resonance (LC-FT-ICR). Note that the two proteins were present at approximately equal amounts in both the labeled and unlabeled samples and the ratio of labeled to unlabeled digests was equal to one. Additionally, the two proteins were digested in 90% $^{18}$O water and 100% $^{16}$O water and analyzed by LC-FT-ICR independently.
3 STATISTICAL METHODS

The abundance data can be parameterized in three different ways: (1) as a function of the concentrations of the peptide in the original samples, \( \theta_{1,1} \) and \( \theta_{1,2} \) (Mirgorodskaya et al., 2000), which are the original parameters of interest; (2) in terms of the amount of the compound that is finally labeled with 0, 1 or 2 \( ^{18}\text{O} \) atoms, which we will define using the vector \( \delta \); and (3) as the amount of compound that has undergone 0, 1 or 2 exchanges (each of which may or may not have successfully substituted an \( ^{18}\text{O} \) atom), which we will define using the vector \( \beta \). Another parameter, in common to all models, is the incorporation rate of the \( ^{18}\text{O} \) label into the peptide. We show that the three formulations are closely linked and we can pass from one to another as needed.

The input data for the algorithm reported here consists of a peak list that encompases a single peptide, consisting of mass and abundance. The creation of this list via other software tools is assumed. For the data examined in this paper, the peak detection routine required a minimum of three peaks in order to conclude that an isotopic cluster is potentially real and does not simply denote a set of noise peaks. Molecules with larger mass or larger abundance have more peaks than molecules with small mass or small abundance.

We will begin with an overview of the multivariable regression model that was proposed by Mirgorodskaya et al. (2000) in order to set up the notation for the two alternative regression models. Let \( y_{i0}, y_{i1}, \ldots, y_{i n} \) be a set of observed peak heights at a spacing of one Dalton that represents a joint isotopic cluster, as determined from some peak-detection procedure. Also, assume \( y_{0,i} \) corresponds to the \( ^{18}\text{O} \) (similarly, \( ^{18}\text{O}_0 \)) labeled (monoisotopic) peak.

Mirgorodskaya et al. (2000) described a multivariable regression model that jointly fits the unlabeled (Sample 1) and labeled (Sample 2) forms of a peptide \( c \) concurrently and produces estimates of the overall abundances, \( \theta_{1,c} \) and \( \theta_{2,c} \), corresponding to the two samples, respectively. The multivariable regression approach uses all of the peaks across the entire joint isotopic cluster.

The multivariable regression model can be written as

\[
E(y_{i}) = W_{c} \theta_{c}, \quad \text{where } E(y_{i}) = [y_{i0}, y_{i1}, \ldots, y_{in}]^{T}
\]

is a \( n \times 1 \) vector of experimentally observed isotopic abundances (peak heights) for the \( c \)-th peptide, \( W_{c} \) is a \( n \times 2 \) matrix of expected (theoretical) isotopic abundances for the \( c \)-th peptide in the unlabeled and labeled samples, and \( \theta_{c} \) is a \( 2 \times 1 \) vector of estimable parameters that denotes the abundances corresponding to Sample 1 and 2, respectively, for the \( c \)-th peptide.

To obtain the expected isotopic distribution of the labeled sample—and accounting for peptide-specific incorporation rates of the \( ^{18}\text{O} \) label—Mirgorodskaya et al. (2000) proposed running the labeled sample separately and the experimentally observed isotopic abundances are retained as input into the second column of the matrix of expected isotopic abundances \( W_{c} \). In contrast, the expected isotopic distribution for the unlabeled sample is determined theoretically and displayed in the first column of \( W_{c} \). The vector of estimable parameters \( \theta_{c} \) is the least squares solution

\[
\theta_{c} = (W_{c}^{T}W_{c})^{-1}W_{c}^{T}y_{c},
\]

and provides the values of ultimate interest; \( \theta_{1,c} \) and \( \theta_{2,c} \) denote the abundance of peptide \( c \) in Sample 1 and 2, respectively.

Our regression models make use of the multivariable regression framework of Mirgorodskaya et al. (2000). Ultimately, if a peptide’s incorporation rate is known to be \( s_{c} \), then the matrix \( W_{c} \) can be broken up into its respective parts, \( W_{c} = (X, S_{c}) \), where

\[
\begin{bmatrix}
1 & y_{0} & 0 & 0 \\
1 & y_{1} & 0 & 0 \\
1 & y_{2} & y_{0} & 0 \\
1 & y_{3} & y_{1} & 0 \\
1 & y_{4} & y_{2} & y_{0} \\
1 & y_{5} & y_{3} & y_{1} \\
\vdots & \vdots & \vdots & \vdots \\
1 & y_{n-1} & y_{n-2} & y_{n-3}
\end{bmatrix}
\quad \text{and } S_{c} =
\begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & (1 - s_{c})^2 \\
0 & 2s_{c}(1 - s_{c}) & 0 \\
0 & 0 & s_{c}^2
\end{bmatrix}
\] (1)

The matrix \( X \) contains the expected (theoretical) abundance distributions for the \( ^{18}\text{O}_0 \), \( ^{18}\text{O}_1 \) and \( ^{18}\text{O}_2 \) labeled isotopic clusters that by definition must each sum to one; the intercept term is represented by the column of ones. Specifically, \( \gamma_0, \gamma_1, \ldots \) denotes the expected abundance distribution, where \( \gamma_i \) is the abundance of the peptide that has \( i \) extra neutrons due to natural isotopes. The last column of \( S_{c} \) represents the probability that the given peptide, when processed in \( ^{18}\text{O} \) water, will finish with 0, 1 or 2 \( ^{18}\text{O} \) atoms, and is a function of both the purity of the \( ^{18}\text{O} \) water and the reaction kinetics of the peptide. The first row and column of \( S_{c} \) represents the intercept, and the second column represents the labeling of the sample processed in \( ^{18}\text{O} \) water, e.g. 100% of the product will have zero \( ^{18}\text{O} \) atoms.

The regression model of Mirgorodskaya et al. (2000) requires an a priori estimate of the peptide’s incorporation rate, \( s_{c} \), in order to obtain accurate estimates of the abundance of peptide \( c \) in the paired samples. Mirgorodskaya and colleagues indirectly estimate the incorporation rate by running the labeled sample separately and using the experimentally observed peak heights as the expected abundance distribution. However, running a labeled-only sample is not ideal for global biomarker-discovery studies where hundreds or thousands of peptides are being quantified in every sample and quite often the same peptide is not detectable in every sample or in multiple runs of the same sample. We propose two alternative regression models that can be used for quantification that adjust for a peptide’s incorporation rate without requiring additional mass-spectrometry runs of the labeled-only sample. Moreover, our regression models correctly estimate the abundance of the labeled sample that is represented as an \( ^{18}\text{O}_0 \) isotopic distribution due to the inability of a peptide to exchange a \( ^{18}\text{O}_0 \) atom during digestion. Thus, the models are particularly applicable to the two-step labeling approach. The effective incorporation rate, \( s_{e} \), and the overall sample abundances, \( \theta_{1,c} \) and \( \theta_{2,c} \), are estimated concurrently using all of the experimentally observed abundances that comprise the joint isotopic cluster. In order to introduce the required notation for the alternative regression models, we will first describe the derivation of a peptide’s effective incorporation rate, \( s_{e} \), and will then proceed with a description of the regression models.

3.1 Reaction kinetics

For practical reasons, the \( ^{18}\text{O} \) water utilized in \( ^{16}\text{O}^{18}\text{O} \) stable-isotope labeled experiments typically contains a known proportion of \( ^{18}\text{O} \) water. For example, the data shown in Figure 1c utilized 90% pure (enriched) \( ^{18}\text{O} \) water, \( p = 0.90 \). All of the examples will focus on 90% enriched \( ^{18}\text{O} \) water since that is the reactant most commonly used in our laboratory. In order for a peptide in the labeled sample to incorporate an \( ^{18}\text{O} \) atom, two events must occur. First, an exchange...
must take place, and second, an $^{18}$O atom must be chosen from the enriched $^{18}$O water for that exchange. Because the probability of an exchange is independent of the purity associated with the $^{18}$O water, the probability that a C-terminal oxygen ultimately ends up with an $^{18}$O atom at the end of the reaction time is

$$s_c = \text{Prob} (= 1 \text{ exchange occurred}) \cdot p$$

$$= (1 - e^{-K_c t}) \cdot p$$

$$= p_1 \cdot p$$

where $K_c$ denotes the reaction constant for trypsin with peptide $c$, $t$ denotes the reaction time, and $p$ denotes the purity of the $^{18}$O water. For either large $K_c$ or long $t$, the probability that at least one exchange occurs approaches one ($p_1 \rightarrow 1$) and $s_c = p$; for incomplete reactions $s_c < p$. Figure 2 provides the probability of ultimately obtaining 0, 1 or 2 $^{18}$O atoms as a function of the total expected number of exchanges $K_c t$. The value of $s_c$ is thought of as an 'effective $p$'; the spectral shape of a peptide with incomplete incorporation in $^{18}$O water of purity $p$ (i.e. $s_c < p$) is identical to the spectral shape for complete incorporation in $^{18}$O water of purity $s_c$.

### 3.2 Multivariable regression models

We propose two alternative regression models for quantification that adjust for a peptide’s incorporation rate using the information from all of the experimentally observed peak heights that comprise a joint isotopic cluster. The observed incorporation rate, $s_c$, and the overall sample abundances for each peptide, $\theta_{1'}$ and $\theta_{2'}$, are estimated concurrently using all of the experimentally observed abundances.

The first multivariable regression model (‘label-state model’) is in terms of the characteristics of the mixed sample, and in particular, the label states. The label-state model is defined as $E[y_{sc}] = X_{dsc}$ and models the experimentally observed peak heights directly, using only the design matrix $X$, defined in Equation (1). The chemical composition of peptide $c$ is estimated using averagine (Senko et al., 1995), and likewise, the expected abundance distribution is estimated using currently existing algorithms (e.g. Rockwood and Van Orden 1996). The design matrix $X$, typically spans 7–14 peaks, depending on the putative mass of the peptide. Any row of $X$, that accounts for < 0.1% of the mass is retained. For example, a peptide with mass $\sim 500$ Da would have a design matrix with eight rows whereas a peptide with mass $\sim 3000$ Da would have 12 rows. Note that for the data examined in this paper, the peak detection routine required a minimum of three peaks in order to conclude that an isotopic cluster is potentially real and does not simply denote a set of noise peaks. For cases where the number of rows in the design matrix is larger than the number of detected peak heights, a value of zero is recorded for the peaks that were not detected. Because the detection threshold is actually a value larger than zero, this may not be the optimal solution and current work entails further investigation of the lower limit of detection. The $4 \times 1$ vector of parameter estimates, $\delta_{sc}$, denote the abundance of the peptide in the mixed sample that was labeled with each of the label states: $\delta_{s0}$ denotes the intercept and $\delta_{s1}$, $\delta_{s2}$ and $\delta_{s3}$ denote the amount of the peptide in the mixed sample that was labeled with 0, 1 or 2 $^{18}$O atoms, respectively. The optimal solution is the least-squares estimate subject to the constraint $\delta_s \geq 0$; programs to compute the constrained least-squares estimate are widely available. The use of constraints is critical, as we find in practice that the constrained solution has better performance than the unconstrained one.

The parameter estimates from the label-state model are not of interest in themselves; however, the overall abundances from Sample 1 and 2, $\theta_{1'}$ and $\theta_{2'}$, are functions of the parameter estimates. Consider $\delta_{s1}$, the abundance of peptide $c$ in the mixed sample that was labeled with zero $^{18}$O atoms. Assuming 100% pure $^{18}$O water, then all of Sample 1 (unlabeled sample) was labeled with zero $^{18}$O atoms as well as any molecules of peptide $c$ in Sample 2 (labeled sample) that were unable to exchange either of their C-terminal oxygen atoms for an $^{18}$O atom. Thus, the expected value of $\delta_{s1}$ is $E(\delta_{s1}) = \theta_{1'} + (1 - s_c)^2 \theta_{2'}$. Similarly, the expected value of $\delta_{s2}$ and $\delta_{s3}$ are $E(\delta_{s2}) = 2s_c(1 - s_c)\theta_{2'}$ and $E(\delta_{s3}) = s_c^2 \theta_{2'}$. From the expected values, it can be seen that the values of ultimate interest ($s_c, \theta_{1'}$ and $\theta_{2'}$) are simple functions of the parameter estimates—there are three equations with three unknowns. If $\delta_{s1}$, $\delta_{s2}$ and $\delta_{s3}$ are estimates of $\delta_{s1}$, $\delta_{s2}$ and $\delta_{s3}$, solving the system of equations results in

$$\hat{s_c} = \frac{2\delta_{s3}}{2\delta_{s3} + \delta_{s2}}$$

$$\hat{\theta}_{2'} = \frac{\delta_{s2}}{2s_c(1 - s_c)} \left(\text{equivalently, } \hat{\theta}_{2'} = \frac{\delta_{s3}}{s_c^2}\right)$$

$$\hat{\theta}_{1'} = \hat{\delta}_{s1} - (1 - \hat{s_c})^2 \hat{\theta}_{2'}$$

The estimated incorporation rate, $\hat{s_c}$, is a function of $\hat{\delta}_{s2}$ and $\hat{\delta}_{s3}$, the estimated abundances of peptide molecules that were labeled with one and two $^{18}$O atoms. Important for the two-step labeling approach, $(1 - s_c)\theta_{2'}$ denotes the abundance of the labeled peptide that was unable to exchange either of the C-terminal oxygen atoms for the heavier $^{18}$O form. The label-state model is appealing because an a priori estimate of $s_c$ is not required to obtain estimates of $\theta_{1'}$ and $\theta_{2'}$ for each peptide. Instead, the value of $s_c$ is allowed to vary across peptides and is directly estimated from the experimentally
observed peak heights, which allows for automated quantitative analysis of $^{18}$O/$^{16}$O labeled spectra.

The second multivariable regression model ('exchange-rate model') is also in terms of the characteristics of the mixed sample; however, the exchange-rate model models the fraction of molecules that participated in an exchange. The exchange-rate model is defined as $E(\gamma) = W_p \beta_c$, where $W_p = (X,D)$.

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -p & (1-p)^2 \\ 0 & 0 & p & 2p(1-p) \\ 0 & 0 & 0 & p^2 \end{bmatrix}$$

and $p$ denotes the purity constant. The $4 \times 1$ vector of parameter estimates, $\beta_c$, denotes the abundance of the peptide in the mixed sample that had an exchange in 0, 1 or 2 of the C-terminal oxygen atoms. As an unconstrained problem, the parameter estimates from the label-states model ($\delta_c$) are equivalent to the parameter estimates from the exchange-rate model ($\beta_c$); $\delta_c = D \beta_c$. Thus, although the parameter estimates have different interpretations across the two regression models, they ultimately generate identical estimates of $\dot{s}_c, \dot{\theta}_{c1}$ and $\dot{\theta}_{c2}$ under ordinary least squares. As a constrained problem the exchange-rate model is superior, one can show that any solution to $\beta_c$ will also satisfy the non-negativity constraints on $\delta_c$, but not vice-versa. Additional constraints are $\dot{\theta}_{c1} \geq 0, \dot{\theta}_{c2} \geq 0$ and $\dot{s}_c \leq p$. All together, these limit the solution to the region above the surface shown in Figure 3. Most noticeably, as $\dot{\theta}_{c2}$ increases, $\dot{\theta}_{c1}$ must also increase in order for $\dot{\theta}_{c1} \geq 0$.

To solve the constrained problem, we first perform a constrained least-squares fit in terms of $\dot{\beta}_c$; the constrained fit is nearly as fast as the unconstrained least squares fit for a problem of this size. If the solution point lies above the surface shown in Figure 3, then we use the derivations in Equation (3), but in terms of $\dot{\beta}_c$:

$$\dot{p}_c = \frac{2 \dot{\beta}_c}{2 \dot{\beta}_c + \dot{p}_c}$$

(note $\dot{s}_c = p, p$ from equation 2),

$$\dot{\theta}_{c2} = \frac{\dot{\beta}_c}{2 \dot{p}_c(1 - \dot{p}_c)}$$

(equivalently, $\dot{\theta}_{c2} = \frac{\dot{\beta}_c}{\dot{p}_c^2}$)

If the constrained solution in terms of $\dot{\beta}_c$ lies below the surface shown in Figure 3, then we use the point on the surface directly above the target solution, that is, hold $\dot{\beta}_c$ and $\dot{\theta}_{c3}$ fixed at the target solution. Additionally, the estimated incorporation rate is constrained to $0.70 \leq \dot{s}_c \leq 0.90$ (not shown in Fig. 3). A lower threshold of 0.70 was chosen to improve the fit for low abundance peptides where estimation is difficult.

### 4 RESULTS

As an illustration, the constrained exchange-rate model was applied to the joint isotopic cluster presented in Figure 4a, which resulted in the following parameter estimates: $\delta_{00} = 103.9, \dot{\beta}_{c1} = 52141.5, \dot{\theta}_{c2} = 0$ and $\dot{\theta}_{c3} = 30206.4$. The parameter estimates imply that the abundance of the peptide in the mixed sample that had an exchange in 0, 1, or 2 of the C-terminal oxygen atoms was 52141.5, 0 and 30206.4, respectively (arbitrary units). Substituting the parameter estimates into Equation (4) results in an estimated incorporation rate of $\dot{s}_c = 0.90$, implying that the reaction went to
Fig. 5. The estimated ratios \( \left( \hat{\theta}_2/\hat{\theta}_1 \right) \) that were calculated using the constrained exchange-rate model versus the standard error of the estimated ratio. The horizontal reference line denotes the expected ratio.

![Graph](image)

Fig. 6. Distribution of the estimated incorporation rate \( \left( \hat{\xi}_i \right) \) for the molecules shown in Figure 5; vertical axis denotes percentage of molecules with the corresponding incorporation rate. The estimated incorporation rates were constrained to \( 0.70 \leq \hat{\xi}_i \leq 0.90 \).

![Graph](image)

completion for the corresponding peptide. The overall estimated abundance in the \( ^{18}\text{O} \)- and \( ^{16}\text{O} \)-labeled samples was \( \hat{\theta}_1 = 52141.5 \) and \( \hat{\theta}_2 = 30206.4 \), respectively, and the relative abundance of the labeled sample to the unlabeled sample is \( \hat{\theta}_2/\hat{\theta}_1 = 0.58 \), which is smaller than the true relative amount of 1. The error represents a combination of sample process variability, instrument variability, estimation error, etc. Additionally, if the theoretical distribution is incorrectly specified then measurement errors in the independent variables (i.e. matrix \( X_i \)) exist and results in biased parameter estimates and inflated model error variance (Neter et al., 1996; Myers, 1990). Therefore, if the theoretical distribution predicted by \textit{averagine} is incorrect, then the parameter estimates will be biased. The errors associated with using \textit{averagine} to approximate isotopic distributions have been discussed in the literature (Johnson and Muddiman; Senko et al., 1995).

Figure 4a demonstrates that the constrained exchange-rate model provides an adequate fit to the observed data. Plots of the residuals from the regression model versus \( m/z \) and versus the observed peak heights (abundances) show that there is no evidence that the residuals are systematically related to \( m/z \) or abundance (data not shown).

Clearly, the advantage of the proposed methods, in comparison to each sample in comparison to each chromatography run as is presently done. Particularly, incorporating information across runs will particularly benefit peptides that have a small number of peaks within any given run. Combining information across runs will particularly benefit peptides that have a small number of peaks within any given run. Combining information across runs will particularly benefit peptides that have a small number of peaks within any given run. Combining information across runs will particularly benefit peptides that have a small number of peaks within any given run. Combining information across runs will particularly benefit peptides that have a small number of peaks within any given run. Combining information across runs will particularly benefit peptides that have a small number of peaks within any given run. Combining information across runs will particularly benefit peptides that have a small number of peaks within any given run. Combining information across runs will particularly benefit peptides that have a small number of peaks within any given run. Combining information across runs will particularly benefit peptides that have a small number of peaks within any given run. Combining information across runs will particularly benefit peptides that have a small number of peaks within any given run.

As a further demonstration of the methods, the proposed approach was applied to a single chromatography run of the data discussed in Section 2, which consists of a mixture of sera that was mixed 1:1 and labeled using 90% pure \( ^{18}\text{O} \) water. Figure 5 displays the estimated ratio \( \left( \hat{\theta}_2/\hat{\theta}_1 \right) \) versus the standard error of the estimated ratio for each detectable molecule; the standard errors were calculated using the delta method. The estimated ratios shown in Figure 5 are biased down, that is, the estimated ratios tend to be smaller than the true value of 1. As stated previously, the bias can be due to multiple factors (variability due to sample processing and instrumentation, errors associated with \textit{averagine}, etc.). Figure 5 shows that the estimation error increases as the standard error increases. From our experience and others (Johnson and Muddiman, 2004), low-abundant molecules that exist near the noise threshold often produce unreliable ratio estimates. Figure 4b displays a molecule that exists near the noise level and demonstrates the difficulty associated with quantifying molecules that exist near the limit of detection; the estimated ratio for the molecule in Figure 4b is \( \hat{\theta}_2/\hat{\theta}_1 = 10.5 \).

The standard error of the ratio represents an estimate of variability and can be used as weights in downstream statistical analyses; molecules with small standard errors will have more weight in the statistical analysis than molecules with large standard errors. When reporting the estimated ratios to investigators a confidence interval associated with the estimated ratio is more useful than the standard error, especially since the appropriate confidence interval for a ratio is asymmetric.

Figure 6 displays the distribution of estimated incorporation rates for the molecules displayed in Figure 5; the estimated incorporation rates were constrained to \( 0.70 \leq \hat{\xi}_i \leq 0.90 \). Approximately 60% of the molecules had incorporation rates near 0.90, which implies that most of the molecules had complete incorporation of the \( ^{18}\text{O} \) label. Roughly 23% of the molecules had incorporation rates that were constrained to 0.70.

Future work consists of incorporating information across chromatography runs to estimate a single \( s_{2'} \), \( \theta_{1'} \) and \( \theta_{2'} \) for each sample in comparison to each chromatography run as is presently done. Particularly, incorporating information across chromatography runs of the same sample is expected to result in overall more stable estimates. For a given peptide, we expect \( \hat{\xi}_i \) and \( \hat{\theta}_2/\hat{\theta}_1 \) to remain constant across chromatography runs and thus incorporating information across runs is a reasonable approach. Combining information across runs will particularly benefit peptides that have a small number of peaks within any given run. Combining information across chromatography runs will require a single \( \beta_{10}, \beta_{11}, \beta_{22}, \) and \( \beta_{33} \) as well as a factor variable that has a level for each of the runs and thus the sample size will increase with a minimal increase in the number of estimated parameters.

5 DISCUSSION

We have discussed a simple multivariable regression approach for modeling \( ^{16}\text{O}/^{18}\text{O} \) stable-isotope distributions. The regression
model provides estimates of the contribution of each peptide from each of the samples that are paired together as well as an estimate of the incorporation rate for each peptide. Unlike Mirgorodskaya et al. (2000), we do not require running the labeled sample twice in order to estimate the expected distribution of peak heights for the labeled sample that accounts for peptide-specific incorporation rates of the \( ^{18}\text{O} \) label. Instead, we estimate the incorporation rate from the data at hand by reparameterizing the model described by Mirgorodskaya and colleagues. The proposed regression framework is particularly advantageous for a two-step labeling approach, which has the added complexity that some molecules from a peptide are not able to exchange either of their C-terminal \( ^{16}\text{O} \) atoms for a \( ^{18}\text{O} \) atom.

Multivariable regression analysis is a well-understood modeling technique that has a toolbox of inference diagnostics that are useful for evaluating the quantification process. The standard errors associated with the overall estimated abundances can be estimated via the delta method and then used to evaluate the precision of the estimates. We suggest using the standard errors as weights in downstream statistical analyses, e.g., differential expression analysis. From our experience, the standard errors are particularly helpful for experiments that are not able to employ a large sample size and thus where variability estimates are even more important.

In summary, the proposed regression models are particularly well suited for automated analysis of large-scale global proteomic studies. Least-squares regression is computationally quick and the ability to estimate the effective incorporation rate from the data at hand is ideal for high throughput quantification.

**ACKNOWLEDGEMENTS**

Funding was provided to Eckel-Passow by NCI grant R25 CA92049 and to Oberg by the Fraternal Order of Eagles Cancer Research Fund. This work was also partially supported by the Lustgarten Foundation. We are grateful to the reviewer for insightful comments, which resulted in an improved manuscript.

**Conflict of Interest:** none declared.

**REFERENCES**


Yao,X. et al. (2003) Dissection of proteolytic \( ^{16}\text{O} \) labeling: endoprotease-catalyzed \( ^{16}\text{O} \)-to-\( ^{18}\text{O} \) exchange of truncated peptide substrates. *J. Proteome Res.*, 2, 147–152.