A Method for Automatically Interpreting Mass Spectra of $^{18}$O-Labeled Isotopic Clusters*

Christopher J. Mason‡, Terry M. Therneau§, Jeanette E. Eckel-Passow§, Kenneth L. Johnson‡, Ann L. Oberg§, Janet E. Olson¶, K. Sreekumaran Nair||, David C. Muddiman**, and H. Robert Bergen III‡

$^{18}$O/$^{18}$O labeling is one differential proteomics technology among many that promises diagnostic and prognostic biomarkers of disease. Although the incorporation of $^{18}$O in the C-terminal carboxyl group during endoproteinase digestion in the presence of $H_2^{18}$O makes the process of labeling facile, the ease and effectiveness of label incorporation have in some regards been outweighed by the difficulties in interpreting the resulting spectra. Complex isotopic patterns result from the composition of unlabeled ($^{18}$O$_0$), singly labeled ($^{18}$O$_1$), and doubly labeled species ($^{18}$O$_2$) as well as contributions from the naturally occurring isotopes (e.g., $^{13}$C and $^{15}$N). Moreover, because labeling is enzymatic, the number of $^{18}$O atoms incorporated can vary from peptide to peptide. Finally, it is difficult to distinguish highly up-regulated from highly down-regulated or C-terminal peptides. We have developed an algorithm entitled regression analysis applied to mass spectrometry (RAAMS) that automatically, rapidly, and confidently interprets spectra of $^{18}$O-labeled peptides without requiring chemical composition information derived from product ion spectra. The algorithm is able to measure the effective $^{18}$O incorporation rate due to variable enzyme substrate specificity of the pseudosubstrate during the isotope exchange reaction and corrects for the $^{18}$O$_0$ abundance that remains in the labeled sample when using a two-step digestion/laboring procedure. We have also incorporated a method for distinguishing pure $^{18}$O$_0$ from pure $^{18}$O$_2$ peptides utilizing impure $H_2^{18}$O. The algorithm operates on centroided peak lists and is therefore very fast: nine chromatograms of, on average, 1,168 spectra and containing, on average, 6,761 isotopic clusters were interpreted in, on average, 45 s per chromatogram. RAAMS is fast enough (average, 38 ms/spectrum) to allow the possibility of performing information-dependent MS/MS on a chromatographic time scale on species exceeding predetermined ratio thresholds. We describe in detail the operation of the algorithm and demonstrate its use on datasets with known and unknown ratios. *Molecular & Cellular Proteomics 6:305–318, 2007.

Differential proteomics technologies promise diagnostic and prognostic biomarkers to reduce the burden of disease (1–4). $^{16}$O/$^{18}$O stable isotope labeling is one tool among many with which to quantify the relative expression differences of proteins between two biological samples (5). $^{16}$O/$^{18}$O labeling, in contrast to other popular differential labeling methodologies such as CAT, isobaric tag for relative and absolute quantitation (iTRAQ$^{\text{TM}}$)1, and stable isotope labeling of amino acids in cell culture (SILAC), (a) does not select for peptides containing certain amino acids (a limitation of ICAT), (b) does not require fragmentation spectra (tandem MS) to record an abundance ratio (a limitation of iTRAQ), and (c) is applicable to biological specimens such as serum, plasma, and cerebrospinal fluid (a limitation of SILAC) (6–8). In a typical $^{18}$O experiment, two samples (or two pools of samples) of a biological matrix such as serum are collected with one typically representing a normal or control population and the other representing a disease state, although other arrangements are possible. One sample is digested with an endoproteinase, typically trypsin (9), in the presence of $H_2^{18}$O, whereas the other sample is digested with the same enzyme in $H_2^{16}$O. The trypsin-catalyzed exchange of up to two $^{16}$O atoms for two $^{18}$O atoms at the C-terminal carboxyl group of the peptide provides a mass shift of 2 ($^{18}$O$_0$) or 4 ($^{18}$O$_2$) daltons. The two digests are then mixed and subjected to mass spectrometry where the isotopic clusters for each peptide appear as a set of peaks corresponding to the unlabeled ($^{18}$O$_0$) and labeled ($^{18}$O$_1$ and $^{18}$O$_2$) species. Measuring the heights of these peaks gives information about the relative abundances of the various peptides in the two samples.

Interpreting the resulting spectra, however, is not completely straightforward. Contributions from the artificially enriched $^{18}$O isotopes and naturally occurring isotopes (particularly $^{13}$C and $^{34}$S) combine to form complex, overlapping isotopic distributions especially for high molecular weight peptides where the monoisotopic peak is no longer the most abundant isotope. Schnolzer et al. (10) demonstrated that once the initial cleavage and exchange has occurred the proteolytic peptide forms a pseudosubstrate for the enzyme, causing the incorporation of a second $^{18}$O atom. However, different peptide pseudosubstrates have different reaction...
rates, $K_m$, with trypsin, which causes peptide-to-peptide variability in the incorporation of $^{18}\text{O}$. The reaction rate is determined by such factors as peptide length (10) and sequence composition (11).

Recent digestion protocols (11, 12) introduce a further complication because two digestion steps are utilized: a first digestion using solution-based trypsin in $\text{H}_2^{16}\text{O}$ and a second digestion using immobilized trypsin in $\text{H}_2^{18}\text{O}$. For peptides with high incorporation rates, the two-step procedure is effectively the same as the single step digestion; however, with low incorporation rate peptides, the possibility arises that a peptide molecule cleaved in $\text{H}_2^{16}\text{O}$ may never reassociate with trypsin in $\text{H}_2^{18}\text{O}$. Bantscheff et al. (13) observed that, when using the two-step digestion procedure, it is important to account for the peptide molecules with no $^{18}\text{O}$ atoms at their C termini that remain in the labeled sample. In addition to these complications, other factors that can affect the labeling process include back exchange and pH (14).

Several different methods are described in the literature for calculating $^{16}\text{O}/^{18}\text{O}$ ratios while accounting for variable incorporation. Yao et al. (5) used a ratio combining the experimental abundances of the $^{18}\text{O}_0$, $^{18}\text{O}_1$, and $^{18}\text{O}_2$ peaks along with the theoretical isotopic abundances of those peaks; chemical composition information derived from known peptide sequence or from product ion spectra was used to obtain the theoretical isotopic abundances. Johnson and Muddiman (15) removed the requirement for sequence information from the product ion spectra by utilizing the average amino acid average (16) to calculate approximate chemical compositions, by modeling the contributions of $^{13}\text{C}$ and $^{34}\text{S}$ with a power function, and by including this in the ratio calculation. Both of these calculations assume a single step digestion and do not account for $^{18}\text{O}_0$ abundance in the labeled sample.

At least two software implementations exist that perform ratio calculations; both have limitations. Halligan et al. (17) and Halligan et al. (15) ratio calculations. They required zoom scans, utilized the protein identification data to determine the monoisotopic mass, and assumed that this mass appeared experimentally. Qian et al. (19) utilized THRASH (20) to detect peptide pairs and then applied the ratio equations of Johnson and Muddiman (15). THRASH, however, uses a single theoretical isotopic distribution and must separately detect the $^{18}\text{O}_0$ and $^{18}\text{O}_2$ species; therefore it will have difficulty with more massive peptides or those with incomplete incorporation.

Our current work describes a completely automated method for locating and interpreting $^{18}\text{O}$-labeled isotopic clusters in parent ion chromatograms without requiring product ion or selected ion monitoring/zoom spectra. Central to our algorithm is the use of linear regression to simultaneously fit all peaks in the isotope cluster rather than just the peaks representing the $^{18}\text{O}_0$, $^{18}\text{O}_1$, and $^{18}\text{O}_2$ species. We use the residuals from the regression to compute a fitting score between the theoretical and experimental isotopic distributions that can be used to rank potential candidate biomarkers. Mirgorodskaya et al. (21) also used linear regression with $^{18}\text{O}$-labeled internal standards but measured the relative abundances of the $^{18}\text{O}_0$ and $^{18}\text{O}_2$ species for a given peptide in a separate experiment instead of allowing these abundances to vary in the regression.

We also describe a method of distinguishing highly up-regulated peptides from highly down-regulated or C-terminal peptides. When using highly enriched $\text{H}_2^{16}\text{O}$, peptides that have fully incorporated two $^{18}\text{O}$ atoms and that appear only in the labeled sample appear identical to peptides that have incorporated no $^{18}\text{O}$ atoms either because they were C-terminal (and thus not a substrate for trypsin) or because they were significantly down-regulated. By lowering the enrichment level of the $\text{H}_2^{18}\text{O}$ to, for example, 90%, a unique peak “signature” is generated that the algorithm can recognize to distinguish these two cases. Furthermore our algorithm accounts for the residual $^{18}\text{O}_0$ abundance that remains in the labeled sample from peptides with low incorporation rates; this is particularly relevant when utilizing the two-step digestion procedure.

In a manner similar to THRASH, the algorithm performs automated “reduction” of spectra. No assumption is made regarding monoisotopic mass, which is determined using an alignment procedure similar to that described by Senko et al. (16) and utilized by THRASH. However, THRASH operates directly on the raw spectral data; this allows it to detect species at very low signal to noise ratios but also makes it very slow. Instead we have focused on centroided peak data supplied by the vendor software from parent ion mass spectra for reasons of speed and because such data were readily available. The algorithm is indeed quite fast, interpreting a chromatogram of ~1,300 spectra in less than 1 min; THRASH might take several hours to process such data. The algorithm was developed using data from FT-ICR instrumentation but should be generally applicable to any instrument with sufficient resolving power.

We will describe in detail the operation of our algorithm, entitled regression analysis applied to mass spectrometry (RAAMS), and evaluate its performance using datasets with both known and unknown ratios.

MATERIALS AND METHODS

Datasets—Three datasets were used to demonstrate the effectiveness of the algorithm: two with known $^{18}\text{O}/^{16}\text{O}$ ratios and one with unknown ratios; these datasets are summarized in Fig. 1. The first two datasets consisted of purified proteins at known $^{18}\text{O}/^{16}\text{O}$ ratios but digested in $\text{H}_2^{18}\text{O}$ of different purities. In one dataset human transferrin and bovine serum albumin were trypsin-digested together in either 99% $\text{H}_2^{18}\text{O}$ or 100% $\text{H}_2^{16}\text{O}$ and then mixed in seven $^{18}\text{O}/^{16}\text{O}$ ratios (20, 10, 3, 1, 0.3, 0.1, and 0.05). Digests of the two proteins in either 99% $\text{H}_2^{18}\text{O}$ or 100% $\text{H}_2^{16}\text{O}$ were used to represent ratios of 0 and $\infty$, respectively. The second dataset was comprised of the same two proteins digested in 90% $\text{H}_2^{18}\text{O}$ or 100% $\text{H}_2^{16}\text{O}$ and mixed in the same ratios. Each sample was then analyzed by LC-FT-ICR. Thus, a dataset was comprised of nine chromatograms at nine different $^{18}\text{O}/^{16}\text{O}$ ratios. Note that the two pure proteins were present at approximately equal amounts in both the labeled and unlabeled samples; only the ratios of labeled to unlabeled digests of these proteins were varied. It is important to understand that these known ratios were generated for verification purposes only; a typical $^{18}\text{O}$ experiment would consist of a single mixed sample with a different (and unknown) $^{18}\text{O}/^{16}\text{O}$ ratio for each protein. These two datasets are available upon request (see below) in mzXML format.
**Experimental methodology.** Three datasets were prepared: the first two datasets consisted of purified proteins (human transferrin and serum albumin) mixed at known $^{18}O/^{16}O$ ratios. The third dataset consisted of human plasma from disease and control pools.

To verify that the algorithm was able to detect differences in complex samples a third dataset was prepared. This set consisted of two pools of human plasma each composed of five disease or five control patients. The disease pool was trypsin-digested in 90% $^{16}O$, whereas the control pool was digested in $^{18}O$. These pools were then combined, and strong cation exchange (SCX) chromatography was performed, collecting 40 1-min fractions. We examined by nano-LC-FT-ICR that single fraction with the highest UV absorbance response because it was more likely to provide a complex sample.

**Sample Preparation**—Human albumin, human apotransferrin, DTT, and iodoacetamide were purchased from Sigma. Sequencing grade trypsin for solution-based trypsin digestion was purchased from Promega (Madison, WI). Trypsin immobilized on agarose beads was purchased from Pierce. Disposable size exclusion columns (Micro Bio-Spin, P6 in Tris) were purchased from Bio-Rad. $^{18}O$-Enriched water (99 atom % enrichment) was purchased from Isotec (Miamisburg, OH). Solvents for LC-MS were purchased from Burdick and Jackson. Formic acid was from Fluka.

Two sample tubes, each containing 400 µg each of albumin and transferrin, were initially treated with DTT and alkylated with iodoacetamide. Samples were digested with solution-based trypsin in $^{16}O$, 10% acetonitrile overnight at 37 °C. After this initial digestion, a second digestion using immobilized trypsin was used to exchange respective isotope labels into each sample. A 45-µl volume of immobilized trypsin beads, previously washed with 50 mM Tris, was purchased from Bio-Rad. $^{18}O$-Enriched water (99 atom % enrichment) was purchased from Isotec (Miamisburg, OH). Solvents for LC-MS were purchased from Burdick and Jackson. Formic acid was from Fluka.

Two sample tubes, each containing 400 µg each of albumin and transferrin, were initially treated with DTT and alkylated with iodoacetamide. Samples were digested with solution-based trypsin in $^{16}O$, 10% acetonitrile overnight at 37 °C. After this initial digestion, a second digestion using immobilized trypsin was used to exchange respective isotope labels into each sample. A 45-µl volume of immobilized trypsin beads, previously washed with 50 mM Tris, was purchased from Bio-Rad. $^{18}O$-Enriched water (99 atom % enrichment) was purchased from Isotec (Miamisburg, OH). Solvents for LC-MS were purchased from Burdick and Jackson. Formic acid was from Fluka.

Two sample tubes, each containing 400 µg each of albumin and transferrin, were initially treated with DTT and alkylated with iodoacetamide. Samples were digested with solution-based trypsin in $^{16}O$, 10% acetonitrile overnight at 37 °C. After this initial digestion, a second digestion using immobilized trypsin was used to exchange respective isotope labels into each sample. A 45-µl volume of immobilized trypsin beads, previously washed with 50 mM Tris, was purchased from Bio-Rad. $^{18}O$-Enriched water (99 atom % enrichment) was purchased from Isotec (Miamisburg, OH). Solvents for LC-MS were purchased from Burdick and Jackson. Formic acid was from Fluka.

Two sample tubes, each containing 400 µg each of albumin and transferrin, were initially treated with DTT and alkylated with iodoacetamide. Samples were digested with solution-based trypsin in $^{16}O$, 10% acetonitrile overnight at 37 °C. After this initial digestion, a second digestion using immobilized trypsin was used to exchange respective isotope labels into each sample. A 45-µl volume of immobilized trypsin beads, previously washed with 50 mM Tris, was purchased from Bio-Rad. $^{18}O$-Enriched water (99 atom % enrichment) was purchased from Isotec (Miamisburg, OH). Solvents for LC-MS were purchased from Burdick and Jackson. Formic acid was from Fluka.

Two sample tubes, each containing 400 µg each of albumin and transferrin, were initially treated with DTT and alkylated with iodoacetamide. Samples were digested with solution-based trypsin in $^{16}O$, 10% acetonitrile overnight at 37 °C. After this initial digestion, a second digestion using immobilized trypsin was used to exchange respective isotope labels into each sample. A 45-µl volume of immobilized trypsin beads, previously washed with 50 mM Tris, was purchased from Bio-Rad. $^{18}O$-Enriched water (99 atom % enrichment) was purchased from Isotec (Miamisburg, OH). Solvents for LC-MS were purchased from Burdick and Jackson. Formic acid was from Fluka.

Two sample tubes, each containing 400 µg each of albumin and transferrin, were initially treated with DTT and alkylated with iodoacetamide. Samples were digested with solution-based trypsin in $^{16}O$, 10% acetonitrile overnight at 37 °C. After this initial digestion, a second digestion using immobilized trypsin was used to exchange respective isotope labels into each sample. A 45-µl volume of immobilized trypsin beads, previously washed with 50 mM Tris, was purchased from Bio-Rad. $^{18}O$-Enriched water (99 atom % enrichment) was purchased from Isotec (Miamisburg, OH). Solvents for LC-MS were purchased from Burdick and Jackson. Formic acid was from Fluka.

Two sample tubes, each containing 400 µg each of albumin and transferrin, were initially treated with DTT and alkylated with iodoacetamide. Samples were digested with solution-based trypsin in $^{16}O$, 10% acetonitrile overnight at 37 °C. After this initial digestion, a second digestion using immobilized trypsin was used to exchange respective isotope labels into each sample. A 45-µl volume of immobilized trypsin beads, previously washed with 50 mM Tris, was purchased from Bio-Rad. $^{18}O$-Enriched water (99 atom % enrichment) was purchased from Isotec (Miamisburg, OH). Solvents for LC-MS were purchased from Burdick and Jackson. Formic acid was from Fluka.

Two sample tubes, each containing 400 µg each of albumin and transferrin, were initially treated with DTT and alkylated with iodoacetamide. Samples were digested with solution-based trypsin in $^{16}O$, 10% acetonitrile overnight at 37 °C. After this initial digestion, a second digestion using immobilized trypsin was used to exchange respective isotope labels into each sample. A 45-µl volume of immobilized trypsin beads, previously washed with 50 mM Tris, was purchased from Bio-Rad. $^{18}O$-Enriched water (99 atom % enrichment) was purchased from Isotec (Miamisburg, OH). Solvents for LC-MS were purchased from Burdick and Jackson. Formic acid was from Fluka.

Two sample tubes, each containing 400 µg each of albumin and transferrin, were initially treated with DTT and alkylated with iodoacetamide. Samples were digested with solution-based trypsin in $^{16}O$, 10% acetonitrile overnight at 37 °C. After this initial digestion, a second digestion using immobilized trypsin was used to exchange respective isotope labels into each sample. A 45-µl volume of immobilized trypsin beads, previously washed with 50 mM Tris, was purchased from Bio-Rad. $^{18}O$-Enriched water (99 atom % enrichment) was purchased from Isotec (Miamisburg, OH). Solvents for LC-MS were purchased from Burdick and Jackson. Formic acid was from Fluka.
collected in full profile mode from m/z 350 to 1,900 at 100,000 resolving power (full width, half-maximum at m/z 400) using two microscans and a target population of $1 \times 10^5$ charges for automatic gain control (AGC) in the ICR cell. The linear ion trap performed MS/MS experiments on the three most abundant ions from the FT-ICR survey scans using 2 microscans, a precursor isolation width of 2 m/z units, 35% normalized collision energy, and 30 ms activation time. An automatic gain control target value of $1 \times 10^2$ was used for MS/MS experiments in the LTQ mass analyzer.

Software Implementation—Software was implemented in C++ on Linux and Mac OS X. Performance measurements were made on a computer with a 3-GHz Xeon processor running Linux 2.6. Software and datasets are available as open source upon request.

COMPUTATIONAL DATA INTERPRETATION

The algorithm takes as input the centroided peaks from a single spectrum as provided by the instrument vendor software and produces as output a list of isotope clusters detected therein along with their monoisotopic, neutral masses, their centroid retention times, and the ion abundances of both the unlabeled ($^{16}$O, sample A) and labeled ($^{18}$O, sample B) species. These two abundances, denoted $\theta_0$ and $\theta_1$, respectively, can be used to find differentially expressed peptides in either an absolute ($\theta_A - \theta_B$) or relative ($\theta_B/\theta_A$) scale. The estimated variance-covariance matrix of $\theta$ is also returned, allowing confidence intervals to be formed.

The algorithm operates in three steps. First, a list of all potential isotope clusters is generated based on peak spacing consistent with any allowed charge state. Some isotope peaks will belong to multiple potential isotope clusters. Second, the ion abundances of each of the potential clusters are fit to a theoretical isotopic model using linear regression, giving a best fit to each in terms of the relative amounts of $^{16}$O, $^{18}$O, and $^{18}$O for that peptide. Finally, the algorithm decides between potential isotope clusters that share peaks based on their residuals from the regression fits. Each step is described in further detail below and summarized in Table I.

Step One: Examining Peak Spacing—In the first step of the algorithm, potential isotope clusters are found with spacing approximately equal to $1/z$. To accomplish this, the data are divided into disjoint groups of peaks where a “group” is defined by a leading and trailing peak-free region of at least 3 m/z units. A distance matrix is computed between all pairs of peaks in this group. Charge states up to 6 are considered; elements in the distance matrix with values of $100235 \pm 2$.

2 The variables used are: $\theta_0$, predicted amount of material in sample A ($^{16}$O, unlabeled sample); $\theta_1$, predicted amount of material in sample B ($^{18}$O, labeled sample); $X$, matrix of theoretical isotopic abundances, see Fig. 2; $y$, column vector of experimentally observed peak heights, see Fig. 2; $\hat{y}$, column vector of predicted peak heights; $\beta$, column vector of parameter estimates with $\beta_1$, $\beta_2$, and $\beta_3$ corresponding to the amount of $^{16}$O, $^{18}$O, and $^{18}$O (zero, one, or two incorporations), respectively, and $\beta_4$ corresponding to the intercept; $p$, fractional purity of $H_2^{18}$O, in this work 90% $H_2^{18}$O ($p = 0.9$) is prepared by mixing $H_2^{18}$O and the vendor’s >99% $H_2^{18}$O; $s$, effective purity of $^{18}$O in the isotopic distribution of an individual peptide, higher s values indicate more incorporation of $^{18}$O.

TABLE I

| RAAMS algorithm overview |
|--------------------------|
| **Step 1:** Find sets of peaks that could be potential isotope clusters based on their spacing. Find all charge states ($1 \leq z \leq 6$) consistent with isotope spacing of $1/z$ and having at least 3 peaks. |
| **Step 2:** For each potential isotope cluster: |
| a) Form a matrix $X$ with theoretical isotopic abundances, generated using averagine (16) and Mercury (22), for incorporation of 0, 1, or 2 $^{18}$O atoms. |
| b) Form vectors $y$ with experimental peak abundances, each $y$ vector assuming a different peak is the monoisotopic mass. |
| c) Solve matrix equation using non-negative least squares. |
| d) Decide between possible $y$ vectors (monoisotopic masses) using residuals. |
| e) Compute and correct for incorporation rate using $^{18}$O/$^{16}$O abundances. |
| **Step 3:** Decide between possible charge states by computing a charge assignment score that combines goodness of fit for all potential isotope clusters that share common peaks. |

error, $1.00235/2 \pm$ error, $\ldots, 1.00235/6 \pm$ error are of interest. The average charge state spacing of 1.00235 is from Horn et al. (20). Error in charge state spacing is allowed to be as high as 40 ppm to account for partially unresolved peaks. For each of the possible charge states from highest to lowest, the algorithm identifies adjacent pairs of peaks that have the requisite spacing. For each such pair, in order of abundance, the cluster is extended by adding those peaks that best fit the spacing of the original pair. The algorithm stops walking outward when one of two conditions is met: (a) the next peak with appropriate spacing is greater than 3 m/z units away or (b) the average error in peak spacing for the whole run exceeds 40 ppm. In this way, a potential isotope cluster is found that has m/z spacings of approximately $1/z$ for a given charge state $z$. At least three peaks must be found for a potential isotope cluster to be passed to later stages of the algorithm. Isotope clusters resulting from low abundance species with mass less than $\sim 500$ Da may legitimately have only two peaks (the A ($^{18}$O) and A + 1 ($^{13}$C$_1^{18}$O$_1$) peaks with negligible $^{13}$C contribution) above the limit of detection, but these are very difficult to distinguish from noise spikes.

Three special cases converge to make it difficult to determine, based only on peak spacing, the correct charge state for a given run of peaks. (a) Noise peaks can interdigitate between real peaks. (b) Isotope clusters can legitimately overlap and share one or more peaks. (c) Low abundance peaks can fall below the signal to noise threshold of the peak detection algorithm. Examples of these are shown in Supplemental Fig. S1. To address the issue of noise peaks, legitimate peak sharing, and missing peaks, the algorithm attempts to fit all charge states that are consistent with the peak spacing data. The algorithm also tolerates up to two missing peaks internal to a candidate isotope cluster. In each of the three cases, the result is one or more isotope clusters with correct charge state assignments and one or more in-
correct assignments that share peaks with the correct assignments. We resolve between these correct and incorrect isotope clusters in step three of the algorithm (described below).

**Step Two: Examine Peak Abundances**—In step two of the algorithm, the candidate isotope clusters from step one are evaluated based on their abundances. The purpose of step two is to determine the contributions of the three label states (exchange of zero, one, or two oxygen atoms at the C terminus of the peptide) to the experimentally observed abundances of the isotope cluster by least squares fitting of a linear combination of expected isotopic abundances for each of these states. For each candidate isotope cluster, its charge state (determined in step one) and the mass of its first peak are used to compute an approximate neutral mass, and then an approximate chemical composition is determined using the average amino acid averagine, which has the chemical formula $C_{4.9384}H_{7.7583}N_{1.3577}O_{1.4773}S_{0.0417}$ (16). Optionally the algorithm can adjust up or down the number of sulfurs predicted by averagine and perform the steps below for these additional expected chemical compositions.

As shown in Fig. 2a, for each chemical composition, three theoretical isotopic distributions are generated using Mercury (22) that correspond to the three label states. The first theoretical isotopic distribution is that which would result if all the molecules of peptide had not exchanged either of their C-terminal oxygen atoms (the $^{18}O_0$ label state) and is shown as black bars in Fig. 2a. The second theoretical distribution is that which would result if all of the molecules of peptide exchanged exactly one of their C-terminal oxygen atoms for an oxygen atom from the $H_2^{18}O$ used for labeling (the $^{18}O_1$ label state). If the purity of the $H_2^{18}O$ were 100%, the $^{18}O_1$ distribution would be exactly the same as that of the $^{18}O_0$ label state except shifted to the right by 2 Da and thus would be indistinguishable. As will be explained later, $H_2^{18}O$ of 90% purity is used instead, resulting in the $^{18}O_1$ distribution shown with light gray bars in Fig. 2a. Similarly the third theoretical distribution ($^{18}O_2$ label state) would result if all of the peptides exchanged both of their C-terminal oxygen atoms for two oxygen atoms drawn from the 90% pure $H_2^{18}O$. The three theoretical isotopic distributions are arranged in the columns of a matrix, $X$, as shown in Fig. 2b (note that in this figure, the matrices are shown transposed). The $X$ matrix is scaled so that each column sums to 1. A column of ones is also added to represent the intercept. For a further discussion of the regression model, see Eckel-Passow et al. (23).

The experimental peak abundances are arranged in a column vector, $y$, also shown in Fig. 2b. Because it is not known beforehand which experimental peak is the monoisotopic mass (all $^{12}C$, all $^{14}N$, all $^{16}O$, etc.), a number of different $y$ vectors, corresponding to different alignments of theoretical isotopes to experimental peaks, are attempted. Indeed the monoisotopic mass may not even be detected experimentally as in the case of species highly up-regulated in the labeled sample. Fig. 2b shows only the $y$ vector of the correct align-
The result of this regression is a column vector $\beta$ with elements representing the amount of experimental abundance that can be accounted for by exchange of zero ($\beta_1$), one ($\beta_2$), or two ($\beta_3$) of the C-terminal oxygen atoms or by an abundance offset ($\beta_0$). The abundance parameters $\beta_{1-3}$ are constrained to be non-negative because they represent the amount of abundance in each of the three label states. By using a least squares fit, the abundances of all peaks in the isotope cluster are used in determining the abundances from sample A and sample B ($\theta_A$ and $\theta_B$; described below) instead of just the abundances of A, A + 2, and A + 4 peaks as in previous work (5, 15). Mirgorodskaya et al. (21) also used regression analysis for quantification with 18O-labeled internal standards but measured the relative abundances of the $^{18}$O$_1$ and $^{18}$O$_2$ species in a separate experiment rather than allowing them to vary in the regression.

The regression is performed for each possible alignment of theoretical distribution to experimental peaks. The alignment with the lowest residual sum-of-squares error in the fit is chosen as correct; this is similar in concept to the alignment procedure described by Senko et al. (16). In choosing the correct alignment, it is important to account for the fact that the isotopic distribution of complete incorporation of two $^{18}$O atoms closely resembles that of the natural abundance distribution and that the two distributions are thus indistinguishable. This means that very low residuals can be obtained by a partial match to only the rightmost peaks of a cluster (incorrectly assuming, for instance, that the experimental $^{18}$O$_2$ is the theoretical $^{18}$O$_2$) because this fit will have fewer peaks and therefore less error. To prevent this, four zeros are added to each end of the $X$ matrix when determining the correct alignment; these zeros (shown in gray in Fig. 2b) introduce high residuals for those problematic alignments at the tails of the distribution that would otherwise effectively ignore a large amount of experimental abundance. However, they also bias the parameter estimates. Therefore, once the correct alignment is found, a version of $X$ lacking these extra “isotopes” is fit to the $y$ vector for the correct alignment, and the resulting $\beta$ is used for all further calculations.

The result of this final non-negative least squares fitting is an estimate of the vector $\beta$ where elements 1–3 represent the abundances contributed by the exchange of zero, one, or two of the C-terminal oxygen atoms, respectively, and element 0 represents an intercept term. These values take into account and correct for the presence of both the naturally occurring isotopes ($^{13}$C, $^{34}$S, etc.) and also for the e.g. 90% impurity of the $^{18}$O$_2$ present in the $^{18}$O$_2$. However, to convert these values into the original sample abundances $\theta_A$ (unlabeled; $^{16}$O) and $\theta_B$ (labeled; $^{18}$O), we must take into account the fact that trypsin has different affinities for different peptide pseudosubstrates. Two events must occur for a peptide molecule to incorporate an $^{18}$O atom. 1) That peptide must react (exchange) with trypsin, and 2) an $^{18}$O atom must be chosen from the 90% $^{18}$O$_2$. The probability of a given peptide molecule having at least one interaction with trypsin is equal to $(1 - e^{-Kt})$ where $K$ is a reaction constant that varies from peptide to peptide and $t$ is reaction time; we denote this probability $p_c$. If a given peptide molecule reacts at least once with trypsin, the probability of obtaining an $^{18}$O atom is equal to the purity of the $^{18}$O$_2$, denoted $p$. Thus, for a given peptide molecule $c$, the probability of incorporating a C-terminal $^{18}$O atom is as follows.

$$s_c = (\text{Probability of at least 1 reaction})(\text{purity of water})$$
$$= (1 - e^{-Kt})p$$
$$= p_c p$$  

(Eq. 1)

Conversely the probability of not incorporating an $^{18}$O atom is $(1 - s_c)$. The probability $s_c$ can be thought of as an incorporation rate or “effective purity.”

There are four sets of molecules we must take into consideration: 1) molecules from sample A, which never contain $^{18}$O, 2) molecules with no exchange events in sample B, 3) molecules where exactly one of the C-terminal oxygen atoms has exchanged, and 4) molecules where both C-terminal oxygen atoms have exchanged. The probability of a molecule having no exchange events in sample B is zero. This is because, in the single step digestion procedure, the probability of a molecule having no exchange events in sample B is zero. This is because, in the single step digestion procedure, peptide cleavage and labeling both occur simultaneously in $^{18}$O$_2$. However, in the two-step digestion, there is the possibility that a peptide may be cleaved in $^{18}$O$_2$ but then never (re)react with trypsin in $^{18}$O$_2$. In this procedure, the observed abundance in the A peak will be a combination of abundances from unlabeled molecules in both sample A and sample B. Thus, $p_{c1}$ is equal to the abundance of molecules from sample A plus those molecules from sample B where neither of the C-terminal oxygen atoms has exchanged. $p_{c2}$ is equal to the abundance of molecules from sample A plus those molecules from sample B where only one of the C-terminal oxygen atoms has exchanged. The probability of this occurring is equal to the probability of having one $^{18}$O atom exchange and the other not exchange; there are two ways this could be done, and thus the final probability is $2p_{c1}(1 - p_{c3})$. $p_{c3}$ is equal to the abundance of molecules from sample B where both C-terminal oxygen atoms have exchanged as follows.

$$\beta_1 = \theta_A + (1 - p_c)^2 \theta_B$$  

(Eq. 2)

$$\beta_2 = 2p_c(1 - p_c) \theta_B$$  

(Eq. 3)

$$\beta_3 = p_c^2 \theta_B$$  

(Eq. 4)
Solving sequentially for $\rho_c$, $\theta_B$, and $\theta_A$ yields the following.

$$\rho_c = \frac{2\beta_3}{2\beta_3 + \beta_2} \quad (s_c = \rho_c \rho) \quad (\text{Eq. 5})$$

$$\theta_B = \frac{\beta_2}{2\rho_c(1 - \rho)} \quad (\text{Eq. 6})$$

$$\theta_A = \beta_1 - (1 - \rho_c)^2\theta_B \quad (\text{Eq. 7})$$

Thus, by assuming that the rate of incorporation of an $^{18}$O atom at the C terminus of a peptide is not influenced by the presence there of another $^{18}$O atom, we can use a function of the ratio of one to two incorporations of $^{18}$O (visible experimentally as the ratio $\beta_2/\beta_3$) to estimate the incorporation rate $s_c$. Then the amounts of material in control ($\theta_A$) and disease ($\theta_B$) samples can be corrected for the incorporation rate using our estimate of $s_c$. When the three parameters $\beta_{1,2,3}$ are non-zero, we calculate the ratio of material in sample B (labeled) to that in sample A (unlabeled) as $\theta_B/\theta_A$, making use of the correction for $s_c$ as shown above. This requires that we detect at least the $^{18}$O$_0$, $^{18}$O$_1$, and $^{18}$O$_2$ peaks to obtain estimates for $\beta_1$, $\beta_2$, and $\beta_3$ and that these estimates are not constrained to zero by the non-negative regression. In cases where at least one of the parameter estimates is constrained to zero, we directly use the ratio $(\beta_2/\beta_3)/\beta_1$. In both cases we constrain to a minimum or maximum ratio to avoid dividing by zero. We could also use $s_c$ to filter out those peptides that, due to a low incorporation rate, do not accurately report the true ratio of amounts. Fig. 2c demonstrates that the regression also provides predicted isotopic abundances, denoted $\hat{y}$ and shown as circles.

To evaluate the goodness of fit between the experimental and predicted isotopic abundances, we compute a fitting “score” as in Equation 8,

$$\text{Fitting score} = \frac{\sum(y - \hat{y})^2}{\sum y} \quad (\text{Eq. 8})$$

where the sum of the squared residuals from the fit is divided by the sum of experimental peak abundances. The denominator serves to normalize the error from the regression for comparison across isotope clusters with varying abundances and numbers of peaks. Lower scores are better in that clusters that more closely agree with their predicted abundances will have lower scores. Optimal scoring is an area of active investigation for our group.

**Step Three: Decide between Charge Assignments**—The third and final step of the algorithm further examines the candidates from step one that share common peaks. Recall that missing peaks, noise peaks, and overlapping isotope clusters introduce complications into the charge assignment process. Additional fits are attempted to decide which of these assignments are correct, and these additional fits result in isotope clusters that share peaks. Furthermore this peak sharing can be either legitimate (when two real isotope clusters overlap) or artifactual as in the case of noise peaks or missing peaks (see examples in Supplemental Fig. S1). A strict rule might be to never allow a peak to participate in more than one isotope cluster, but this would incorrectly eliminate some clusters that legitimately shared peaks.

Two competing factors must be accounted for in deciding between isotope clusters that share peaks. Isotope clusters with better (lower) fitting scores should be preferred over those with worse fitting scores. However, fitting score is influenced by the number of peaks in the isotope cluster: clusters with fewer peaks will have less error. This means that, for some isotope clusters, an incorrect assignment with lower charge state (and therefore fewer peaks) will have a better fitting score than a correct assignment with higher charge state. Therefore, we must also evaluate the amount of abundance that the isotope cluster accounts for out of the surrounding peaks. Obviously isotope clusters that use more of the surrounding peaks (i.e. a higher charge state) should be preferred over those that use fewer of them (i.e. a lower charge state).

We combine these two competing factors by computing a charge assignment score for each isotope cluster in the transitive closure of isotope clusters that share peaks (if an isotope cluster A shares peaks with B and B shares peaks with C, then ABC constitutes the transitive closure of isotope clusters that share peaks). We compute a charge assignment score for each isotope cluster in the transitive closure as the squared error from the fitting for this isotope cluster plus the squared abundances for all the other peaks in the transitive closure that were not used by that cluster (denoted $y'$) divided by the sum of abundances of all peaks in the transitive closure as follows.

$$\text{Charge assignment score} = \frac{\sum(y - \hat{y})^2 + \sum(y')^2}{\sum y + \sum y'} \quad (\text{Eq. 9})$$

We then rank each isotope cluster by charge assignment score from lowest to highest and examine each in turn in this order. Isotope clusters that share peaks and have charge states that are even multiples of each other are not allowed to co-exist. The algorithm decides among clusters that violate this sharing rule by picking the one with the best (i.e. lowest) charge assignment score. In this way, legitimate peak sharing is allowed (Supplemental Fig. S1b has overlapping clusters with $z = 2$ and $z = 3$, which are allowed), but the incorrect, duplicate clusters introduced by noise or missing peaks are removed.

**Evaluation of Alignment Correctness**—A key factor influencing the correctness of the ratios produced by the algorithm is whether it determines the correct monoisotopic mass, that is to say whether it determines the correct alignment between the theoretical and experimental isotopic distributions. When the algorithm incorrectly assigns the monoisotopic mass, it generally assigns either the $^{18}$O$_1$ or $^{18}$O$_2$ $(A + 2$ or $A + 4)$
isotope as the monoisotopic mass, introducing a shift in mass of 2 or 4 Da. Although there are indeed some species that differ by exactly a small, integer number of daltons, the most likely explanation for such a mass error is that the algorithm has incorrectly assigned the monoisotopic mass. Because peptides with masses less than ~4,000 daltons have readily detectable monoisotopic masses, determining the correct alignment is trivial using the chromatogram from the sample digested in H$_2$^{16}O. We used these two pieces of knowledge (±4-Da shift and correct monoisotopic mass from $^{16}$O sample) to verify the algorithm when fitting variable ratios.

We ran the algorithm, using a single theoretical isotopic distribution corresponding to the natural isotopic abundances for the chemical composition predicted by averagine, against the $^{16}$O chromatogram and then combined the appearance of species at multiple charge states, in multiple spectra chromatographically, and across each of the samples by using mass and retention time thresholds of 20 ppm and 120 s, respectively, as described previously for our label-free approach (26). (We also used thresholds in retention time to remove artifact species, such as polymer introduced by the chromatography, as described by Mason et al. (27).) For each species we determined a consensus monoisotopic mass across all isotopic clusters for that species as their abundance weighted average of monoisotopic mass.

Next we applied the algorithm to all nine chromatograms each of the 90 and 99% enrichment datasets but allowed the algorithm to fit unknown $^{18}$O/$^{16}$O ratios. We then used the same grouping procedure on this "unknown" data with one exception: we allowed two isotope clusters A and B to be grouped across multiple spectra and multiple charge states if mass(A) = mass(B) + n x 1.00235 Da ± 20 ppm where n varies from 0 to 4. We then rated the correctness of the algorithm in determining the monoisotopic mass for each cluster in the unknown data as whether its monoisotopic mass is the same (within 20 ppm) as the consensus monoisotopic mass from the $^{16}$O data.

RESULTS

Fig. 3 characterizes the ratios found in the chromatograms (see "Materials and Methods" for description of datasets) of purified proteins digested in 90% H$_2$^{18}O. There is one row for each of seven different, known ratios with all $^{18}$O at the top and all $^{16}$O at the bottom; the 10 and 0.1 ratios were omitted for clarity (the complete figure is available upon request). Fig. 3a shows scatter plots of the isotope clusters found, with fitting score (Equation 8) on the x axis (plotted on a log base 10 scale) versus predicted $^{18}$O/$^{16}$O ratio on the y axis (also log base 10 scale). The ratio expected in each row is printed in the left-hand margin and is indicated by the solid line; the dashed line indicates a ratio of 1. Lower scores represent better agreement between theoretical and experimental isotopic distributions. Points are shaded by the abundance of the most abundant peak in each isotope cluster. Between scores of ~0.1 and 0.5, very few isotope clusters of high abundance are observed, and their ratios vary widely from the expected.

Manual inspection of 50 isotope clusters with a range of abundances and with scores between 0.1 and 0.105 revealed that 42 of 50 were correctly assigned. However, at scores between 0.2 and 0.205, only 5 of 50 were correct. Based on this we have used a threshold value of 0.1, shown by a solid vertical line. Note that at high score (≈0.5) more abundant peaks are again observed; manual inspection revealed that many of these occur when noise peaks are separated (by random chance) from a highly abundant peak by exactly 1/z. This results in very high residuals for these incorrect fits.

One factor in the ability of the algorithm to determine the correct ratio is whether it can determine which of the peaks of an isotope cluster is the monoisotopic mass, particularly for peptides highly up-regulated in the labeled sample; put another way, it is important for the algorithm to determine the correct “alignment” between the theoretical and experimental isotopic distributions. The isotope clusters in Fig. 3a are colored green if the alignment was determined to be correct using the procedure described above and red if it was incorrect; isotope clusters whose correct alignment could not be determined because their mass and retention time could not be matched to the $^{16}$O sample are colored black.

Fig. 3b shows a histogram, for each of the nine expected ratios, of the number of isotope clusters (horizontal axis) that have a given ratio (vertical axis, which has the same scale as in a). Only those clusters with scores less than the threshold of 0.1 are shown. Stacked bars are colored red, green, or black to indicate alignment correctness (note: truncated bars are one-third the correct length). In general, the experimental ratios correspond well with the expected ones out to expected ratios of about 20 and 0.05; the distributions do broaden, however, at the more extreme ratios. Some bias toward $^{18}$O is observed in the ratios; reverse labeling experiments and/or normalization may be useful in correcting this.

Fig. 4 shows FT-ICR mass spectra of a single tryptic peptide of BSA, having been digested in (a) normal H$_2$^{16}O, (b) H$_2$^{18}O reported by the vendor as >99% pure, and (c) 90% pure H$_2$^{16}O. In Fig. 4b the majority of the peptide molecules have incorporated two $^{18}$O atoms, producing the expected 4-Da mass shift. As can be seen from comparing Fig. 4, a and b, the isotopic abundance distributions of $^{16}$O- and that of $^{18}$O-labeled peptides have a very similar shape (imagine shifting the isotopic distribution in Fig. 4a 4 daltons higher). Thus, when a peptide exists in high abundance in one sample but is at a level below the detection limit in the other, it is not possible to tell from which sample it originated. This determination depends in large part on the alignment step in the algorithm that attempts to determine which peak is the monoisotopic peak; if the wrong peak is decided upon, the algorithm will compute incorrect ratios. Moreover the C-terminal peptide of each protein in the sample will not incorporate $^{18}$O label because it is (typically) not a substrate for...
Comparison of Fig. 4, a and c, shows that this ambiguity can be partially resolved by using H$_2^{18}$O of less than 100% purity, resulting in a distinct pattern of peaks. Fig. 4c shows the same peptide digested in 90% H$_2^{18}$O. In this example, $\sim 1/2$ of the peptide molecules have incorporated only a single $^{18}$O atom, and a small fraction, $\sim 4\%$, have not incorporated any. Thus, utilization of 90% H$_2^{18}$O incorporates a signature to discriminate, for example, a C-terminal peptide that is unlabeled from a highly up-regulated, labeled ($^{18}$O$_2$) peptide.

Fig. 5 shows the advantage of using impure H$_2^{18}$O in distinguishing extreme ratios by depicting histograms of the number of correct versus incorrect alignments for each of the nine expected $^{18}$O/$^{16}$O ratios in both 99 and 90% H$_2^{18}$O. Across a range of abundance bins (reported as the base 10 logarithm of the abundance of the most abundant peak in each isotope cluster), the number of isotope clusters where the alignment was determined to be correct is shown in green, whereas the number with incorrect alignments are shown in red. The number of isotope clusters where the correct alignment could not be determined (see above for procedure) is shown in black. Only isotope clusters with scores less than the threshold of 0.1 are included. In Fig. 5a the digest was performed in 99% pure H$_2^{18}$O. At lower ratios (more $^{16}$O), the algorithm has difficulty determining the correct alignment, as shown by the large number of isotope clusters with incorrect alignments, because the expected distributions are indistinguishable to the algorithm. However, when the purity of the H$_2^{18}$O is dropped to 90%, as shown in Fig. 5b, the algorithm does better as indicated by the lower number of incorrect alignments at low $^{18}$O/$^{16}$O ratios. This is because the additional impurity causes a distinct pattern of peaks in the extreme $^{18}$O case where most of the abundance in the combined sample comes from the labeled sample (where the $^{18}$O/$^{16}$O ratio is high). The algorithm uses this fact to distinguish the two extremes of labeling during the alignment proc-

Fig. 3. Observed $^{18}$O/$^{16}$O ratios from nine chromatograms with known ratios. a, scatter plots of isotope clusters detected by the algorithm in seven LC-FT-ICR chromatograms of purified proteins trypsin digested in H$_2^{16}$O and in 90% H$_2^{18}$O and then mixed to form the expected $^{18}$O/$^{16}$O ratios indicated in the left margin (from all $^{18}$O at the top to all $^{16}$O at the bottom; note that the 10 and 0.1 ratios are omitted for clarity). Fitting score is plotted on the x axis, and the calculated $^{18}$O/$^{16}$O ratio is on the y axis (both log$_{10}$ scale). Lower scores indicate better fit; the vertical line indicates the threshold level of 0.1 used in this work. Points are colored according to whether the algorithm correctly determined the alignment between theoretical and experimental isotopic distributions (see text for automated procedure used to evaluate correctness): green indicates that the alignment was correct, red indicates incorrect, and black indicates that correctness could not be determined. The abundance of the most abundant peak in the isotope cluster is indicated by shade (log$_{10}$ scale shown at bottom). The very poor fits at the right-hand edge of the plot occur when noise peaks are separated (by random chance) from a highly abundant peak by exactly 1/z. b, histograms of the $^{16}$O/$^{18}$O ratio of those isotope clusters from a with fitting score <0.1. The y axis is the same as in a, and the stacked bars are colored to indicate alignment correctness as in a. Truncated bars (indicated by slashes) are one-third their correct length. Inf, infinity.
Fig. 4. Examples of the same peptide labeled in H$_2^{16}$O (a), >99% H$_2^{18}$O (b), and ~90% H$_2^{18}$O (c). Note that the small peaks at m/z ~654.4 and 654.9 are due to incomplete exchange and correspond to the $^{18}$O$_1$ and $^{18}$O$_1^{13}$C$_1$ species, respectively.

Fig. 5. Effect of impure H$_2^{18}$O on alignment correctness. a, histograms of abundances of the most abundant peak from isotope clusters found in nine LC-FT-ICR chromatograms of purified proteins trypsin-digested in H$_2^{16}$O and in 99% H$_2^{18}$O and then mixed to form the expected 18O/16O ratios indicated in the left margin (from all 18O at the top to all 16O at the bottom). Stacked bars are colored by the number of isotope clusters with correct (green), incorrect (red), and unknown (black) alignment between theoretical and experimental isotopic distributions. b, as in a but with 90% H$_2^{18}$O. By decreasing the purity of the water the performance of the algorithm in finding the correct alignment is improved at low 18O/16O ratios. Note that due to differences in chromatography, the total number of clusters is not comparable between a and b. Inf, infinity.
a and b at all ratios. Ratios closer to 1 are easier to detect at lower abundance because they are more likely to have all three of the A, A + 2, and A + 4 peaks. Work to determine the optimum level of impurity is ongoing. Note that the total number of isotope clusters is not equal between Fig. 5, a and b, due to differences in chromatography.

The variable incorporation rate of 18O into individual peptides is another factor that can influence the correctness of the ratios computed by the algorithm. Different peptides can have different incorporation rates because trypsin has different affinities for them. A peptide with a low incorporation rate will show an artificially low 18O/16O ratio because the labeled sample will contain an artificially high amount of 18O0 and 18O1, and an artificially low amount of 18O2. The effect of this is the same as if the purity of the H218O used for labeling were lower. Various efforts have been made to drive the reaction further toward complete incorporation of two 18O atoms (11, 28, 29). Increasing the concentration of trypsin in the labeling reaction will also increase the reaction rate. One commercially available kit (12) (Prolytica, Stratagene) achieves this by utilizing two digestion steps: a longer, high volume digestion in H218O with solution-based trypsin followed by a second, shorter digestion in H218O with immobilized trypsin. The use of immobilized trypsin allows for higher trypsin concentrations without increasing the number of autolytic fragments resulting from simply increasing the trypsin concentration. However, in the single step digestion procedure, if a peptide is seen, it must have had at least one opportunity for exchange; this is not true in the two-step digestion procedure. The algorithm attempts to correct for low incorporation rates by computing an effective purity, denoted s, using a function of the 18O/16O ratio. It then subtracts some abundance from the unlabeled sample and adds it to the labeled sample to account for peptide molecules with no 18O atoms that remain in the labeled sample (Equations 5–7).

This approach is demonstrated in Fig. 6, which shows a histogram of the distribution of the rate of incorporation of 18O, indicated by the effective purity sc, across isotope clusters found in the nine chromatograms with known ratios. Higher sc values indicate more efficient labeling, which would suggest higher affinity of trypsin for a given peptide. Three representative peptides are shown in a, b, and c in both the 18O sample (top) used to determine the correct monoisotopic mass and the all 18O sample (bottom), which should show complete incorporation; the predicted abundance distribution (p) is indicated by circles. Moving from right to left, Fig. 6c shows a peptide with a high sc value, indicating substantial incorporation of 18O. Fig. 6b shows a peptide at intermediate sc where less than half the molecules have incorporated one 18O atom and almost none have incorporated two. Finally in Fig. 6a a peptide with a low sc value is shown, indicating poor incorporation; indeed very few molecules have incorporated one 18O atom, and essentially none have incorporated two 18O atoms. Peptides with such a low incorporation rate are problematic because it is difficult to correctly estimate the incorporation rate when so little 18O is seen. In general, these peptides are simply not amenable to quantification.

Table II gives, for each of the nine chromatograms at different expected ratios, information about the number of clusters found and the number of species (peptides) identified by grouping those clusters using tolerances in mass and retention time. Only those isotope clusters with scores less than 0.2 are considered. Over 500 species were found even in a digest containing only two proteins; we believe that large
amounts of chymotryptic cleavage, nonspecific cleavage, post-translational modification, and salt adduction account for this high number of species (data not shown). We report the percent of total ion current (TIC) that could be determined to belong to peaks involved in isotope clusters detected by the algorithm both before and after grouping. On average, 61% of the ion current could be accounted for by the algorithm. The percentage after grouping reflects the fact that we require at least three isotope clusters to find a species; on average about 53% of TIC could be accounted for by the algorithm after grouping. Finally the algorithm took on average 45 s to process an 80-min chromatogram (of about 1,150 spectra) or about 26 spectra/s.

To give some sense of how the algorithm performs on more complex mixtures, Fig. 7 shows peptides found in a single SCX fraction of pooled human plasma that had been depleted of the six most abundant proteins. Isotope clusters detected by the algorithm that represented occurrences of peptides at multiple charge states and in multiple spectra chromatographically were combined using tolerances of 20 ppm and 120 s as described by Mason et al. (27). Fig. 7a plots the $^{18}\text{O}/^{16}\text{O}$ ratio of the most abundant ion for each peptide versus the fitting score of that ion. Lower scores indicate better fit, and larger ratios indicate species up-regulated in the disease sample. An example of a highly up-regulated peptide with an $^{18}\text{O}/^{16}\text{O}$ ratio of $\sim 50$ is shown in Fig. 7b where circles indicate the predicted isotopic abundances ($\hat{y}$). Fig. 7c shows a slightly down-regulated peptide ($^{18}\text{O}/^{16}\text{O} \sim 0.4$) with a very good fit (score, 0.007). Fig. 7d shows a peptide with a neutral mass of 4,224 Da; the algorithm calculates a $^{18}\text{O}/^{16}\text{O}$ ratio of 2.8. This isotopic distribution would be challenging to interpret manually or with previous ratio calculations. It should be noted that, as the mass of peptides increase, their isotopic distributions become dominated by $^{13}\text{C}$ isotopes, and they broaden considerably. Above $\sim 2,500$ Da, the naturally occurring $^{13}\text{C}$ makes it more difficult to detect the presence of the $^{18}\text{O}$ introduced by the labeling.

Parenthetically we note that averagine yields only an approximate chemical composition for a given neutral mass; the predicted chemical composition can be wrong, particularly in the number of sulfurs. We examined the effect of allowing the algorithm to vary the number of sulfurs up or down from that predicted by averagine and determined that the effect of this, although improving scores for some peptides, was minimal (data not shown).

**DISCUSSION**

We borrow many concepts from the THRASH algorithm (20), which we have used previously as part of a label-free biomarker discovery strategy (15). The chief distinction be-

![Fig. 7. Ratios observed in a complex sample.](image)

| Table II Performance of RAAMS on 9 chromatograms with known ratios |
|--------------------------|--------------------------|-------------------|-------------------|-------------------|-------------------|
| **Expected ratio** | **Isotope clusters** | **Species** | **Percentage of TIC** | **Total time Spectra/s** | **Before grouping** | **After grouping** |
|--------------------------|--------------------------|-------------------|-------------------|-------------------|-------------------|
| $\infty$ | 5,029 | 413 | 52 | 39 | 39.4 | 32.2 |
| 20.00 | 6,085 | 514 | 59 | 51 | 43.9 | 27.1 |
| 10.00 | 6,967 | 540 | 63 | 55 | 46.6 | 25.7 |
| 3.00 | 6,341 | 515 | 64 | 55 | 48.1 | 24.2 |
| 1.00 | 6,302 | 486 | 62 | 53 | 45.2 | 25.9 |
| 0.33 | 7,136 | 582 | 63 | 56 | 44.8 | 25.6 |
| 0.10 | 8,089 | 673 | 65 | 59 | 49.8 | 23.0 |
| 0.05 | 7,590 | 641 | 63 | 57 | 48.8 | 23.4 |
| 0.00 | 7,314 | 627 | 62 | 54 | 42.1 | 27.3 |
| **Average** | **6,761** | **555** | **61** | **53** | **45.4** | **26.1** |

*a Total ion current accounted for by algorithm.
tween this work and THRASH is that this work is amenable to 18O-labeled data where complex overlapping isotopic distributions preclude the single theoretical isotopic distribution used by THRASH. However there are other, possibly more subtle, differences we wish to point out.

Our algorithm makes use of centroided peak lists. This is in sharp contrast to THRASH, which does not perform a priori peak detection and instead directly examines the full-profile spectral data. Whereas we only attempt alignments that correspond to the apexes of peaks (typically 10–12 alignments per cluster), THRASH, in contrast, tries a series of several thousand alignments in very small m/z increments assuming each of the several thousand samples in the spectral region under examination may contain the monoisotopic mass. This makes THRASH much more robust in its ability to detect low abundance clusters with peaks below the signal to noise threshold of a peak picking algorithm but also makes it orders of magnitude slower. Also although the present work uses linear regression to find the best fit in a least squares manner in both m/z and abundance, THRASH iterates an overall abundance parameter to scale the theoretical isotopic distribution being aligned.

Our algorithm attempts all allowed charge states that are consistent with the peak data; THRASH, in contrast, uses a Fourier-Patterson charge state determination algorithm (30) to determine the prevailing charge state in a spectral region. THRASH only falls back to fitting all allowed charge states if the fit resulting from the charge state determined by the Fourier-Patterson is poor. Our method relies on the fact that performing the regression is very fast and that far fewer possible alignments are attempted. THRASH also depends on a subtraction step in which the predicted isotopic abundances are subtracted from the original spectrum. Experience has shown this subtraction to be problematic in that it often leaves residuals behind that may have spectral characteristics, resulting in the detection of artifactual isotope clusters with acceptable fitting scores. We avoid the subtraction by simply fitting the two clusters separately.

CONCLUSIONS

In conclusion, we have described and demonstrated the effectiveness of an algorithm that automatically interprets spectra of 18O-labeled isotope clusters. It utilizes linear regression to model the complex, overlapping isotopic distributions that result in these spectra. The algorithm allows for correction of peptide to peptide variability of incorporation rate, particularly when using a common two-step digestion/labelling procedure. It also provides for the use of impure H218O to distinguish strongly up-regulated peptides from strongly down-regulated and C-terminal peptides. The algorithm operates on centroided peak data and does not require fragmentation spectra; it is fast, interpreting an entire 1,500-spectrum chromatogram in less than a minute. This provides for the possibility that the algorithm could be used on line during acquisition to direct the instrument to perform MS/MS on peptides with interesting ratios. This could help ameliorate the fact that in a typical MS/MS experiment one must extensively separate complex samples or risk missing low abundance peptides due to the limited capacity of the instrument to perform fragmentation.

The algorithm was applied here to FT-ICR data but should be generally applicable to any instrument with sufficient resolving power. Also although it was designed for 18O data, it handles unlabeled isotopic distributions as a special case and could be used to interpret label-free experiments as well. It could also be easily adapted for other labeling moieties. In the future we hope to improve our scoring scheme, to incorporate knowledge of the expected distribution of the incorporation rate (s_i), and to incorporate protein identification data when available.

Acknowledgments—We thank Douglas Mahoney for helpful discussions and Keith Perrington for early work on a Javascript spectrum viewer used in examining the output of the algorithm. Rudi Chiarito refined the viewer and contributed to the software library (libmpmc) that underlies our work.

* This work was supported by NIDDK Grant R01DK 70179-2 and NCI Grants R33 CA105295 (to D. C. M.) and R25 CA92049 (to J. E. E.-P.) from the National Institutes of Health, the Fraternal Order of Eagles Cancer Research Fund, the Lustgarten Foundation for Pancreatic Cancer Research, the James M. Kemper Biomarker Research Fund, the Gordon C. and Elizabeth W. Gilroy Fellowship in Proteomic Research, and the Mayo Clinic College of Medicine. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

†† To whom correspondence should be addressed: 3-115 Medical Sciences Building, Mayo Clinic College of Medicine, 200 1st St. S. W., Rochester, MN 55902. Tel.: 507-558-0381; Fax: 507-284-8433; E-mail: bergen.bob@mayo.edu.

REFERENCES

1. Baker, M. (2005) In biomarkers we trust? Nat. Biotechnol. 23, 297–304
2. Diamandis, E. P. (2004) Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. Mol. Cell. Proteomics 3, 367–378
3. Ransohoff, D. F. (2005) Lessons from controversy: ovarian cancer screening and serum proteomics. J. Natl. Cancer Inst. 97, 315–319
4. Anderson, N. L., and Anderson, N. G. (2002) The human plasma proteome: history, character, and diagnostic prospects. Mol. Cell. Proteomics 1, 845–867
5. Yao, X., Freas, A., Ramirez, J., Demirev, P. A., and Fenselau, C. (2001) Proteinolytic 18O labeling for comparative proteomics: model studies with two serotypes of adenovirus. Anal. Chem. 73, 2836–2842
6. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat. Biotechnol. 17, 994–999
7. Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., and Pappin, D. J. (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol. Cell. Proteomics 3, 1154–1169
8. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture and its application to human protein expression profiling. Nat. Biotechnol. 20, 1151–1156

Molecular & Cellular Proteomics 6.2 317
in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol. Cell. Proteomics 1, 376–386
9. Hedstrom, L. (2002) Serine protease mechanism and specificity. Chem. Rev. 102, 4501–4524
10. Schnolzer, M., Jedrzejewski, P., and Lehmann, W. D. (1996) Protease-catalyzed incorporation of 18O into peptide fragments and its application for protein sequencing by electrospray and matrix-assisted laser desorption/ionization mass spectrometry. Electrophoresis 17, 945–953
11. Yao, X., Afonso, C., and Fenselau, C. (2003) Dissection of proteolytic 18O labeling: endoprotease-catalyzed 18O-to-16O exchange of truncated peptide substrates. J. Proteome Res. 2, 147–152
12. Stratagene (2004) ProlyticaTM 18O Labeling Kit Instruction Manual, Stratagene, La Jolla, CA
13. Bantscheff, M., Dumpelfeld, B., and Kuster, B. (2004) Femtomole sensitivity post-digest 18O labeling for relative quantification of differential protein complex composition. Rapid Commun. Mass Spectrom. 18, 869–876
14. Stewart, I. I., Thomson, T., and Figeys, D. (2001) 18O labeling: a tool for proteomics. Rapid Commun. Mass Spectrom. 15, 2456–2465
15. Johnson, K. L., and Muddiman, D. C. (2004) A method for calculating 16O/18O peptide ion ratios for the relative quantification of proteomes. J. Am. Soc. Mass Spectrom. 15, 437–445
16. Senko, M. W., Beu, S. C., and McLafferty, F. W. (1995) Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. J. Am. Soc. Mass Spectrom. 6, 229–233
17. Halligan, B. D., Slyper, R. Y., Twigger, S. N., Hicks, W., Olivier, M., and Greene, A. S. (2005) ZoomQuant: an application for the quantitation of stable isotope labeled mass spectrometry. Bioinformatics 22, 2739–2745
18. Mantel, N. (1969) Restricted least-squares regression and convex quadratic programming. Technometrics 11, 763–773
19. Lawson, C. L., and Hanson, R. J. (1974). Solving Least Square Problems, Prentice Hall, Englewood Cliffs, NJ
20. Johnson, K. L., Mason, C. J., Muddiman, D. C., and Eckel, J. E. (2004) Analysis of the low molecular weight fraction of serum by LC-dual ESI-FT-ICR mass spectrometer: precision of retention time, mass, and ion abundance. Anal. Chem. 76, 5097–5103
21. Senko, M. W., Beu, S. C., and McLafferty, F. W. (1995) Automated assignment of charge states from resolved isotopic peaks for multiply charged ions. J. Am. Soc. Mass Spectrom. 6, 52–56