RAPID COMMUNICATIONS IN MASS SPECTROMETRY
Rapid Commun. Mass Spectrom. 2008; **22**: 2248–2252
Published online in Wiley InterScience (www.interscience.wiley.com) DOI: 10.1002/rcm.3591

RCM

Isotope correction of mass spectrometry profiles

Günther Eibl, Katussevani Bernardo, Therese Koal, Steven L. Ramsay, Klaus M. Weinberger* and Armin Graber

Biocrates Life Sciences, Innsbruck, Austria

Received 30 December 2007; Revised 28 April 2008; Accepted 28 April 2008

Isotope correction of a profile is an important step in the analysis of mass spectrometry derived data. The problem is mathematically formulated as a system of linear equations which is general enough to include previous correction methods. For the solution of these equations when applied to the whole profile an efficient algorithm is developed. In experimental tests the resulting algorithm corrected the profile fast and successfully. Copyright © 2008 John Wiley & Sons, Ltd.

By using tandem mass spectrometry based technology platforms it is nowadays possible to extract, deliver and present rich information residing in metabolite networks with unprecedented speed, low cost, and high confidence. Modern mass spectrometer based methods allow rapid screening, identification, and quantification of up to 1000 metabolites in only a few hours from as little as 10 μL of sample, be it blood, sera, etc. Using a targeted approach, incorporating numerous internal standards information about metabolite reaction to disease and treatments can be provided. The ultimate goal is to use the measurements to identify metabolic biomarkers.

Isotope correction of a profile is one of the first steps in data analysis of mass spectrometry profiles. One of the goals is to remove the first isotope peaks (Fig. 1), which can lead to a higher number of reported metabolites showing the same effects as the monoisotopic peak in clinical studies. Additionally, isotope peaks with the same mass-to-charge ratio m/z as an internal standard can seriously deteriorate the quantification of the corresponding metabolite.

Isotope correction is often discussed in the literature regarding tracer experiments based on ^{13}C-labeled substrates.[1–3] Although similar methods are applied, their focus lies on the determination of mass isotopomers for metabolic flux analysis. Although other papers address tandem mass spectrometry, the methods proposed there are not general and fast enough to be suitable for our purpose.[4,5] All these methods have in common that they consider much fewer molecules and are thus not limited by storage or calculation time constraints. In our setting thousands of data points requiring efficient calculation and storage of the corresponding matrix are considered.

The paper is organized as follows: the following section presents a quite general formulation of the problem as solving a system of linear equations. In the section after that the method is presented. Effort had to be spent to achieve an efficient way of solving the problem arising from the decision to correct the whole profile. The key parts of the algorithm are formulated as pseudocode. In the Results section, successful operation of the algorithm was confirmed and the final section details the conclusions.

**Correspondence to*: K. M. Weinberger, Biocrates Life Sciences, Innrain 66, 6020 Innsbruck, Austria.
E-mail: klaus.weinberger@biocrates.com

FORMULATION OF THE PROBLEM AS A SYSTEM OF LINEAR EQUATIONS

The binomial distribution

We start considering a molecule consisting of n carbon atoms ignoring isotopes arising from other atoms for the moment. Each of the n carbon atoms is a ^{12}C-atom with probability $1 - p = 0.989$ and a ^{13}C isotope with a probability of $p = 0.011$ of approx. 1%. The probability $P(k; n, p)$ that k out of n carbon atoms are ^{13}C-atoms then follows a binomial distribution with parameters n and p, i.e.:

$$P(k; n, p) = Binom(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}. \qquad (1)$$

For small values of p the probability $P(k; n, p)$ quickly decreases with increasing k; the two most important terms are $P(0; n, p) = (1-p)^n$ and $P(1; n, p) = np(1-p)^{n-1}$. The latter term can be approximated by np due to the smallness of p. This formula can be applied to all other atoms, only p changes. The ^{18}O isotopes can be treated by setting $P(k; n, p)$ to zero for odd values of k and $P(k; n, p) = Binom(k/2; n, p)$ with $p = 1 - 0.998 = 0.002$.

Obtaining linear equations

Let $x(M)$ denote the total signal arising from a molecule with monoisotopic mass M consisting of n carbon atoms. We assume that for a given mass M the molecule is uniquely determined. The result should not be affected very much by departures from this assumption because the number of C- and O-atoms in molecules changes only slightly with mass. The fraction $P(k; n, p)$ of the molecules has k ^{13}C carbon atoms and gets observed at mass $M + k$ as signal $x(M + k, M) = P(k; n, p)x(M)$. The expected first isotope peak $x(M + 1, M)$ can easily be estimated using the approximation $x(M + 1, M)$ is approx. $npx(M)$, so for carbon the signal of the first isotope peak is approximately n% of the whole signal $x(M)$ arising from the molecule.

Conversely, the signal $y(M)$ measured at mass M consists of the monoisotopic signal $x(M, M)$ of the molecule with mass M, the first isotope peak $x(M, M - 1)$ of the molecule with

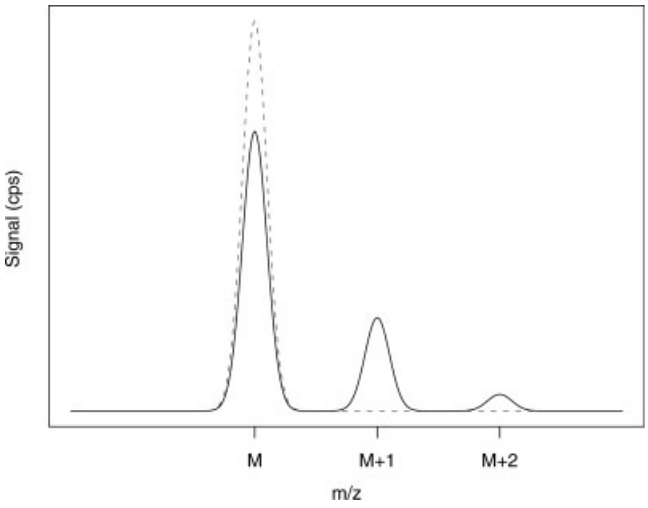

Figure 1. Effect of isotopes: the dashed line denotes the total signal *x* arising from a molecule with monoisotopic mass *M*, the solid line denotes the measured signal *y* having the monoisotopic peak at $m/z = M$, the first isotope peak at $m/z = M+1$ and the second isotope peak at $m/z = M+2$.

mass $M-1$, the second isotope peak $x(M, M-2)$ of molecules with mass M – 2, etc.

$$y(M) = \sum_{k\geq 0} x(M, M-k) \tag{2}$$

$$= \sum_{k\geq 0} P(k; n(M-k), p)x(M-K), \tag{3}$$

where $n(M - k)$ denotes the number of C-atoms of the corresponding molecules. Since $P(k; n, p)$ decreases so fast with k we can limit ourselves to the case $k = 0,\ldots..., K$; typically, K is set to 2.

Let $\tilde{S} = (1, \ldots, N)$ and $x = (x_{\tilde{s}})_{\tilde{s}\in\tilde{S}}$ denote the N molecules considered, $m = (m_{\tilde{s}})_{\tilde{s}\in\tilde{S}}$ and $n = (n_{\tilde{s}})_{\tilde{s}\in\tilde{S}}$ their corresponding masses and number of carbon atoms and let $y = (y_s)_{s\in S}$ for $S = (1,\ldots..., L)$ denote the L signals measured at masses $m = (m_s)_{s\in S}$. Then we can state the situation as a system of linear equations which we write in matrix formulation as:

$$Ax = y, \tag{4}$$

where A is defined as:

$$A_{s,\tilde{s}} = P(k, n(\tilde{s}), p) \quad \text{for} \quad m_s = m_{\tilde{s}} - k. \tag{5}$$

The goal of getting the signal x arising from individual molecules from the measured signal y then corresponds to solving the system of linear equations.

The system of equations stated above is quite general and not limited to the situation where profiles are considered. If more peaks are measured than molecules ($L > N$) the system of equations is usually not solvable. In this case one can obtain x by minimizing the residual $||Ax - y||$. This is a well-known problem in numerics and it is usually solved by using the pseudo-inverse matrix. A similar system of equation arises in Wahl *et al.*,[2] where an additional parameter b represents the background. Stated as a linear regression problem essentially $||Ax + b - y||$ is minimized. This paper focuses on isotope correction of small molecule profiles, whereas in protein informatics overlapping peptide signals in mass spectra are typically deconvoluted by correlation with theoretically predicted known or modeled peptide isotopic peak profiles. First, experimentally measured isotopic peak clusters are frequently used for ion charge state recognition since the initial charge assignment relies on the spacing of peaks in the m/z dimension. Then, intensities are exploited in a subsequent step to address the potential of overlapping signals from multiple peptides. Additionally, at higher masses the monoisotopic peak is not necessarily the most intense peak in the isotope distribution. While the isotope correction of peptides could be treated by a system of equations (the matrix would have more off-diagonal terms and the diagonal term would be smaller), the determination of the number of charge states is not included in this approach. Instead of solving a system of equations, appropriate methods for charge state determination are applied there.[6]

For as many peaks as signals ($L = N$) the system can be solved using Gaussian elimination. The elimination algorithm consists of the computationally expensive part where the matrix A is factorized into a product of an upper and a lower triangular matrix and the much less-expensive part of forward- and then backward-substitution. The situation $L = N$ is considered in Liebisch *et al.*[5] By sorting the molecules and the signals with regard to their masses, A takes the form of a lower triangular matrix and the system of equations can be solved in an intuitive way by forward-substitution without being forced to perform the time-consuming factorization part.

OUR ALGORITHM

As in a targeted approach a peak-finding algorithm is not necessary, we decided not to introduce one for isotope correction. Thus, we chose to use all the points of the signal y also as 'source' x, i.e. we set $L = N$, $x = y$, $\tilde{S} = S$ and $(m_{\tilde{s}})_{\tilde{s}\in\tilde{S}} = (m_s)_{s\in S}$. Furthermore, we have implicitly assumed $z = 1$. For the precursor scan with the second analyzer set to $m/z = 184$ we get signals at masses $m = 300, 300.1,\ldots..., 900$ resulting in 6000 data points. We obtain the number $n_{\tilde{s}}$ of C- and of O-atoms of the molecules with mass $\tilde{m}_{\tilde{s}}$ by selecting the molecule having the mass nearest to $\tilde{m}_{\tilde{s}}$ in a data file which provides the molecule name, the mass and the components of the molecule. Since the differences between subsequent masses in this data file are very small it can happen that the wrong molecule is chosen due to the experimental inaccuracy in determining the mass. However, the number of C- and O-atoms of molecules changes only slowly with mass so the resulting error should be small. Finally, we must subtract the components of the fragment selected by the second mass analyzer from the composition of the metabolites because we can only obtain a signal if the fragment is monoisotopic.

Since we wanted to correct the whole profile an efficient implementation was necessary. The important part is the determination and the storage of matrix A. Once we have obtained A, the system of equations can be solved using Gaussian elimination. The matrix A is far too big to be processed directly. For 6000 data points, A is a 6000×6000 matrix if all signals are non-zero and too large to be treated directly even when zeros are omitted. However, if all signals are sorted with respect to the corresponding mass, it becomes clear that A is lower-triangular and sparse having at most

DOI: 10.1002/rcm

$K+1$ entries in each column. Since A is lower-triangular with non-zeros in the diagonal and the number of equations equals the number of unknowns, A is invertible and thus the solution can be uniquely determined. Due to the size and sparsity of A we saved it as a list a where each entry consists of the four numbers (s, $\tilde{s}$, $A_{s,\tilde{s}}$, k).

Only molecules with mass $m_{\tilde{s}}$ slightly smaller than m_s affect the signals y_s through their kth isotope peak. We can use this fact to considerably restrict the search for these molecules and thus speed up the algorithm. For ^{18}O isotopes, we must search twice as far than for ^{13}C isotopes. This is taken into account using a factor c, which is set to 2 and 1, respectively. The conditions for an effect of molecule $\tilde{s}$ on molecule s are the following: the mass $m_{\tilde{s}}+ck$ must be near enough to m_s, and m_s must be bigger than $m_{\tilde{s}}$. With Δm defined as the minimum difference between subsequently measured masses we get:

$$0.5 \geq \mathrm{TOL} \geq |m_s - (m_{\tilde{s}} + ck)| \geq |\Delta m(s-\tilde{s}) - k| \qquad (6)$$

and $\tilde{s} < s$ (note that we have already sorted the masses). These two conditions result in the condition:

$$\tilde{s} \in \left[s - \frac{ck+0.5}{\Delta m}, s-1 \right] \qquad (7)$$

Typically, $\Delta m = 0.1$, TOL is typically set to 0.21. Points with zero signal are not saved. Thus, the mass $m_{\tilde{s}}+ck$ can be further away from m_s than the inaccuracy in measuring the mass. In this case no correction is performed.

Summarizing, we can write the algorithm for obtaining a in the following form:

for $s := N$ **to** 2 **do**
  append (s, s, $P_s(0; n_s, p)$) to the list a
  for $k := N$ **to** $K+1$ **do**

$$\tilde{S} := \left[\max\left\{ s - \frac{ck+0.5}{\Delta m}, 1 \right\}, s-1 \right]$$

$$\hat{s} := \min_{s \in \tilde{S}} |m_{\tilde{s}} - (m_s - ck)|$$

  if $|m_{\tilde{s}} - (m_s - ck)| \leq$ TOL **then**
    append $(s, \hat{s}, P_{\hat{s}}(k; n_{\hat{s}}, p))$ to the list a
  end if
  end for
  end for

Now the system of equations is solved using Gaussian elimination

$x := y$
for $s :=$ length(a) **to** 1 **do**
  if $a(s, 4) = 0$ **then**

$$x_{a(s,1)} := \frac{x_{a(s,1)}}{a(s,3)}$$

  end if
  if $a(s, 4) > 0$ **then**

$$x_{a(s,1)} := \max\{0, x_{a(s,1)} - a(s,3)x_{a(s,2)}\}$$

  end if
end do

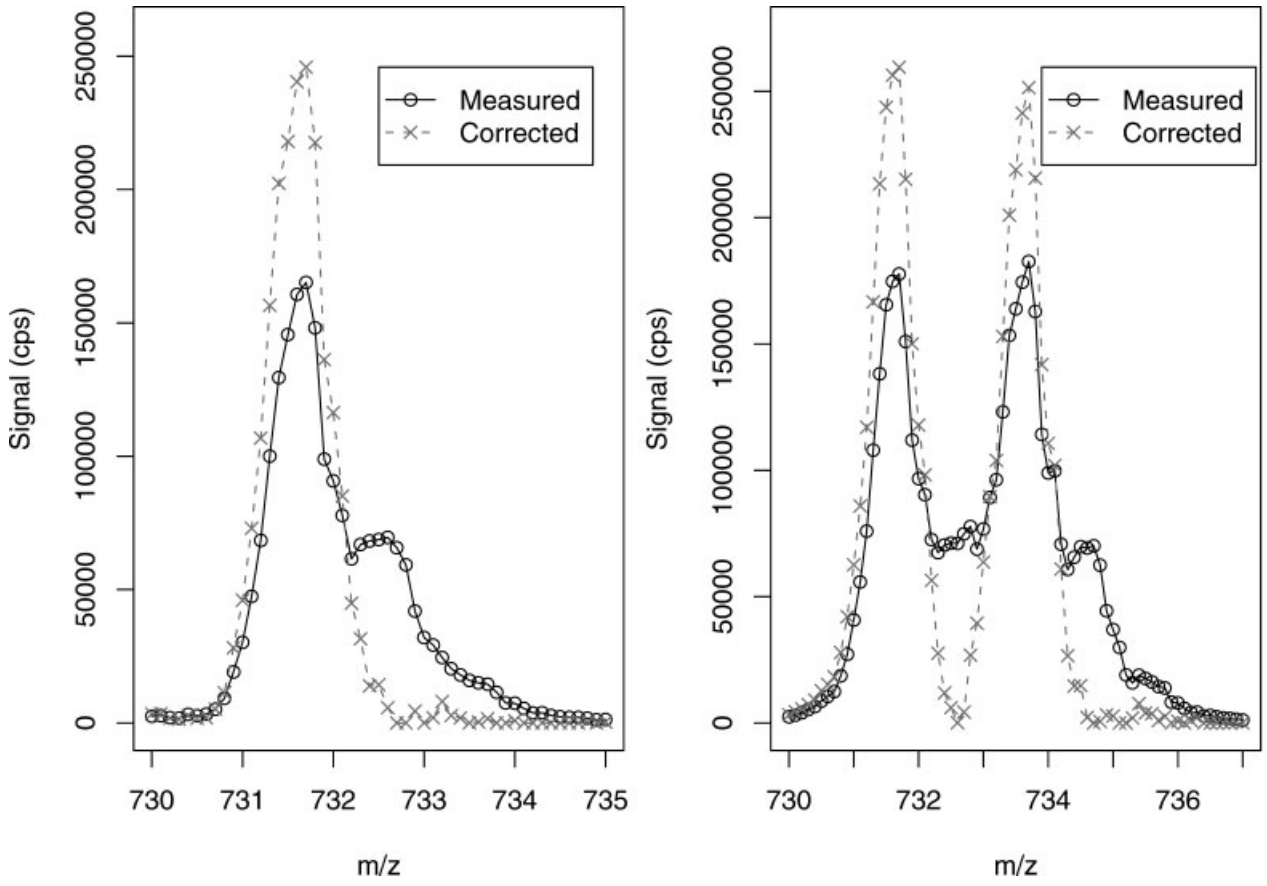

Figure 2. Left panel: successful removal of a pure isotope peak: the solid line is the measured signal y, the dashed line is the corrected signal x. Right panel: two lipids with same intensity and $\Delta m/z = 2$ (GPCho 32:1 and GPCho 32:0). The pure isotope peak between the two peaks is successfully removed.

DOI: 10.1002/rcm

Note that the last entries in a are the *smallest* masses, so, during Gaussian elimination, we really substitute in the *forward* direction. We avoid negative signals by simply setting negative values to zero. In another approach one could consider minimizing $||Ax-y||$ under the constraint that all signals x_s are non-negative. However, a trial using the package tsnnls[7] did not show big differences to our simpler approach.

The correction for ^{13}C and ^{18}O isotopes is done separately in subsequent steps. The right procedure treats both isotopes in one single step as communicated in van Winden *et al.*[1] The right matrix is the same as the product of the two single matrices; however, there the probabilities of isotopes $(P(k, n(\tilde{s}), p)$ are calculated at slightly different indices $\tilde{s}$ leading to slightly different values $n_{\tilde{s}}$. Due to the slow change of $n_{\tilde{s}}$ with $\tilde{s}$ the resulting difference is small and our procedure is a sufficiently good approximation to the precise procedure.

RESULTS

In all our experiments we analyzed lipids using tandem mass spectrometry (API4000QTrap). Low mass resolution (peak width = 1) was selected for the detection to maximize the mass spectrometric sensitivity for the lipids resulting in an improvement of the individual limits of detection. Working with low mass resolution also ensures proper operation of the algorithm even for broad peaks.

The left panel of Fig. 2 shows the analysis of a single lipid (GPCho 32:1, $m/z = 731.7$). As expected two peaks were measured instead of one single peak. The pure isotope peak at $m/z = 732.7$ can be successfully removed by our method. The right panel of Fig. 2 shows a tandem mass spectrum of two lipids (GPCho 32:1 and GPCho 32:0) with subsequent mass difference of 2 Da and concentrations of 25 μM for each of them. The pure isotope peak at $m/z = 732.7$ lying exactly between the two monoisotopic peaks could be removed even in low mass resolution detection mode demonstrating the suitability and robustness of the newly developed isotope correction method.

The correction should not only remove pure isotope peaks, it should also be able to keep peaks of analytes with low concentration which are buried under the first isotope peak of a analyte with high concentration. We selected two analytes with mass difference of 1 Da. While keeping the concentration of the analyte with higher mass SM (d18:1/18:0) ($m/z = 730.7$) fixed at 25 μM we increased the signal of the analyte with lower mass GPCho 32:2 ($m/z = 729.7$) from 0 μM up to 250 μM. The peak of the analyte with SM (d18:1/18:0) can be kept with good accuracy even when the first isotope peak of GPCho 32:2 is much higher (Fig. 3).

Figure 4 shows the result for the region $m/z \in [800,830]$ for an arbitrary sample. As expected (the condition of matrix A is low) no accumulation of errors can be seen. We were able to denote most of the peaks, only the peaks at $m/z = 814$ and $m/z = 816$ remain unknown. Pure isotope peaks have been removed successfully (dots) especially for the region $m/z \in [820,830]$. The correction of the lipid SM (d18:1/24:4)

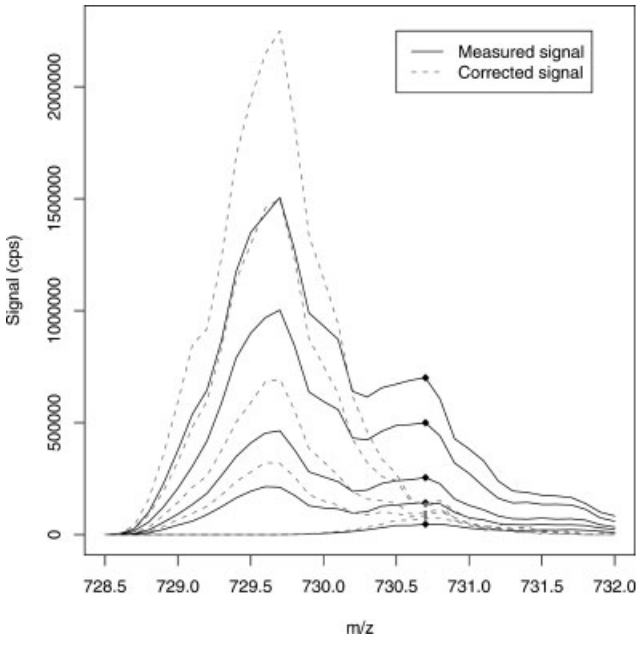

Figure 3. Correction of the signal of SM (d18:1/18:0) ($m/z = 730.7$) whose signal is buried under the first isotope peak of GPCho 32:2 ($m/z = 729.7$): the solid line is the measured signal y, the dashed line is the corrected signal x. The squares and dots denote the measured and corrected signals of the analyte. The concentrations of GPCho 32:2 are 0, 25, 50, 100 and 250 μM, respectively.

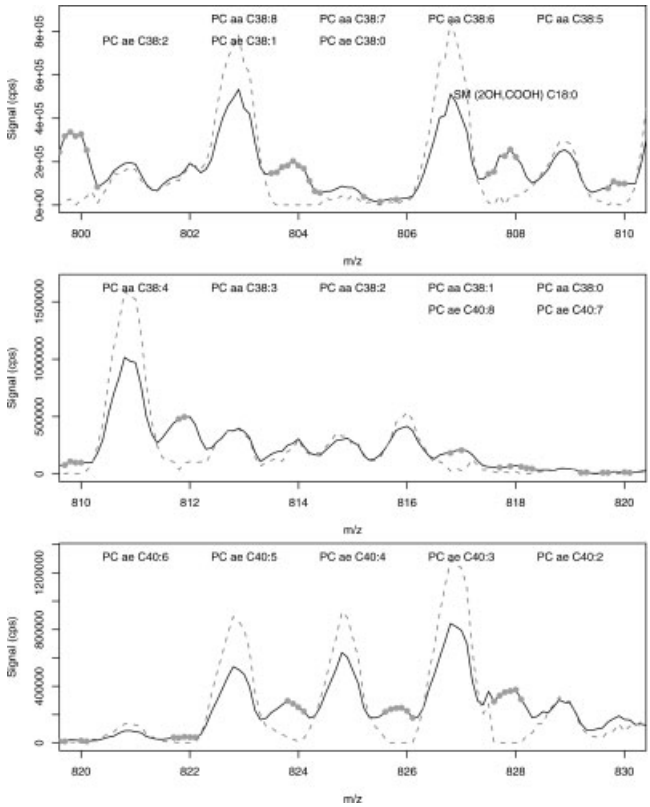

Figure 4. Isotope correction applied to an arbitrary sample for the region $m/z \in [800,830]$. The solid line is the measured signal y, the dashed line the corrected signal x and the dots are signals from pure isotope peaks which have been removed successfully. For numbers of the form xx:y, xx denotes the total chain length and y denotes the number of double bonds.

DOI: 10.1002/rcm

situated at $m/z = 807.6$ is especially important for our targeted approach. Due to isotope correction the intensity is decreased from 1.5×10^5 counts per second (cps) down to 1400 cps which is already below the noise level. This corresponds exactly to the case of the right panel of Fig. 2, where a pure isotope peak lying between two real peaks could be removed. Without isotope correction a peak for SM (d18:1/24:4) would have been falsely reported.

CONCLUSIONS

We formulate isotope correction mathematically as a system of linear equations. In contrast to previous approaches we correct the whole profile, which would be unfeasible without an efficient algorithm for the determination and solution of the equations. In experimental tests the resulting algorithm needed only 3 s for the correction of a profile using an implementation in R on a standard PC. Profiles were successfully corrected and especially pure isotope peaks could be removed.

REFERENCES

1. van Winden WA, Wittmann C, Heinzle E, Heijnen JJ. *Biotechnol. Bioeng.* 2002; **80**: 477.
2. Wahl SA, Dauner M, Wiechert W. *Biotechnol. Bioeng.* 2004; **1686**: 108.
3. Wittmann C, Heinzle E. *Biotechnol. Bioeng.* 1999; **62**: 739.
4. Han X, Gross RW. *Anal. Biochem.* 2001; **295**: 88.
5. Liebisch G, Lieser B, Rathenberg J, Drobnik W, Schmitz G. *Biochim. Biophys. Acta* 2003; **85**: 259.
6. Zhang X, Hines W, Adamec J, Asara JM, Naylor S, Regnier FE. *J. Am. Soc. Mass. Spectrom.* 2005; **16**: 1181.
7. Cantarella J, Piatek M. tsnnls: A solver for large sparse least squares problems with non-negative variables.

DOI: 10.1002/rcm