2010

Automated assignment of ionization states in broad-mass matrix-assisted laser desorption/ionization spectra of protein mixtures

Dariya I. Malyarenko
William & Mary

William E. Cooke
William & Mary, cooke@physics.wm.edu

Dennis M. Manos
William & Mary, dmanos@wm.edu

Christine L. Bunai
William & Mary

Recommended Citation

Malyarenko, D. I., Cooke, W. E., Bunai, C. L., & Manos, D. M. (2010). Automated assignment of ionization states in broad-mass matrix-assisted laser desorption/ionization spectra of protein mixtures. Rapid Communications in Mass Spectrometry, 24(1), 138-146.

RAPID COMMUNICATIONS IN MASS SPECTROMETRY
Rapid Commun. Mass Spectrom. 2010; **24**: 138–146
Published online in Wiley InterScience (www.interscience.wiley.com) DOI: 10.1002/rcm.4371

RCM

Automated assignment of ionization states in broad-mass matrix-assisted laser desorption/ionization spectra of protein mixtures

Dariya I. Malyarenko[1]*, William E. Cooke[2], Christine L. Bunai[1] and Dennis M. Manos[1,2]

[1]Department of Applied Science, College of William and Mary, Williamsburg, VA 23187-8795, USA
[2]Department of Physics, College of William and Mary, Williamsburg, VA 23187-8795, USA

Received 29 May 2009; Revised 15 September 2009; Accepted 6 November 2009

A computational technique is presented for the automated assignment of the multiple charge and multimer states (ionization states) in the time-of-flight (TOF) domain for matrix-assisted laser desorption/ionization (MALDI) spectra. Examples of the application of this technique include an improved, automatic calibration over the 2 to 70 kDa mass range and a reduced data redundancy after reconstruction of the molecular spectrum of only singly charged monomers. This method builds on our previously reported enhancement of broad-mass signal detection, and includes two steps: (1) an automated correction of the instrumental acquisition initial time delay, and (2) a recursive TOF detection of multiple charge states and singly charged multimers of molecular $[MH]^+$ ions over the entire record range, based on MALDI methods. The technique is tested using calibration mixtures and pooled serum quality control samples acquired along with clinical study data. The described automated procedure improves the analysis and dimension reduction of MS data for comparative proteomics applications. Copyright © 2009 John Wiley & Sons, Ltd.

Matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) has become an important tool for comparative proteomics because it can gently ionize large molecules, and usually produces a predominance of singly charged ions[1,2] (a gain in data reduction compared to electrospray ionization). This is important in profiling studies because the number of mass peaks is usually much larger than the number of independent samples that are measured. The presence of ionization noise, like multiply charged states, can complicate the statistical analysis for the discovery of diagnostic patterns[3–6] because not all of the peaks represent new information about the abundance of proteins or peptides within the sample. However, when MALDI-TOF acquisition is optimized for a broad mass-range survey,[3,7,8] many molecules are routinely observed with multiply charged states. Commercial packages with mass spectrometry (MS) instruments frequently lack the adequate analytical tools that can address this data redundancy. The unambiguous detection and assignment of ionization states requires the estimation of uncertainties for the peak locations and intensities over the full range of the TOF record.

In this paper, we present a computational algorithm for automated detection of ionization states, comprising multiply charged states and singly charged multimers, in the TOF domain (before m/z calibration) for broad-range MALDI data. This approach is tested on mass spectra of several mixtures of calibration standard proteins and of pooled serum samples acquired under experimental settings optimized in three different mass ranges.[8] These mass spectra were collected as a part of larger clinical studies for leukemia[3] and prostate cancer serum[8] samples. For example, in this work, we will present MALDI-TOF spectra (acquired along with the clinical data set) for a calibration mixture of the seven proteins which resulted in 26 different m/z peaks. Fifteen of these peaks are due to multiple charge states or homogeneous multimers of the seven proteins and another nine peaks are the ionization states of intermolecular combinations (heterogeneous multimers) of the three dominant proteins in the mixture.

In clinical data, such peaks, representing multiple copies of the same chemical information, should be combined or eliminated before one attempts a statistical classification of the mass spectra, but they do provide additional information about the MS instrument. Because the multiple charge states and multimer states have m/z values that are ratios of simple integers, a set of such peaks can be used to help determine the true zero time for a TOF spectrum, even without knowledge of the precise m/z value for any one member of the set. Moreover, because all such sets should yield the same zero time offset, this procedure should work even in very complicated spectra when one of the members of the set may be obscured by other peaks. We illustrate this by using our algorithm to analyze the peaks in a pooled blood serum mixture, which is typically used for quality control to characterize MALDI-TOF in clinical samples. We find that one-third of the peaks in this spectrum can be identified as belonging to the ionization states connected to other peaks (for an average of 2–3 peaks for the same protein).

*Correspondence to: D. I. Malyarenko, Department of Physics, College of William and Mary, Williamsburg, VA 23187-8795, USA.
E-mail: dimaly@wm.edu

Our computational algorithm automatically detects ionization states in the TOF domain (before m/z calibration) for broad-range MALDI data. The algorithm builds on enhanced signal detection sensitivity achieved by our previously described signal processing methods.[7–9] This enhancement accounts for the inherent TOF-dependent peak broadening and enables equally efficient detection and uncertainty estimation over the full range of the TOF record.[8] In the first step of our present algorithm, which we will describe in more detail in a later section, we find the lag associated with the maximum correlation between the observed peak positions and the locations predicted by assuming characteristic patterns of higher charge states for each singly charged species in a spectrum. This lag is then used to calculate a single shift, which we identify as the instrumental time delay correction, and this shift aligns all the predicted (z-scaled) peak positions with their observed locations to an accuracy that is better than the instrumental half-width. In the second step, the time delay corrected TOF peak locations are assigned to ionization states that are closest to their predicted locations (within one half-width). This assignment is done starting with the largest peaks, and proceeding in order of decreasing intensity.

Below, two examples of the application of this ionization state assignment algorithm are suggested and discussed in terms of reduced data redundancy by reconstruction of the molecular spectra and improved automated broad-range m/z calibration. The results are compared with available commercial alternatives. The ionization state identification reduces the data dimension by 30–50%. This benefits the statistical analysis of diagnostic patterns and improves the selection of candidates for ID experiments. The improved automated broad mass-range calibration is made possible by a combination of enhanced data processing (peak detection) and the use of auto-assigned ionization states as landmarks in otherwise sparse data regions. This recalibration, including the initial time delay correction, and the assignment of ionization states, enhances the mass accuracy of linear TOF records by 10-fold, or more, for higher masses. The fully automated nature of the procedure reduces human bias and is useful in high-throughput proteomic applications.[3,6,10,11]

EXPERIMENTAL

Mass spectra

Samples and MS acquisition

MALDI-TOF MS spectra for protein standard (PS and PR) calibration mixtures and pooled serum quality control (QC) samples were acquired as a part of large clinical studies for leukemia[3] and prostate cancer patients (ongoing research[8]). Sample preparation procedures are described in detail in Refs. 3 and 8, respectively. Briefly, protein standard samples were acquired from Bruker Daltonics (PS) and Ciphergen Biosystems (PR), and prepared according to the manufacturer's protocols. Protein standard mixtures were given the labels PS1: insulin, ubiquitin, cytochrome C, and myoglobin; PS2: trypsinogen, protein A, and albumin; and PR: beta-galactosidase, cytochrome C, myoglobin, aldehyde phosphate dehydrogenase (APDH), and albumin. A 1:1 mixture of PS1+PS2 was also used. QC samples were purified with C3 (hydrophobic affinity) beads, and IMAC-Cu (immobilized copper) and NP20 (normal phase) affinity capture surfaces using CLINPROT (Bruker) and BioMek (Beckman) robotic bio-processor platforms. The matrices alpha-cyano-4-hydroxycinnamic acid (CHCA) and sinapinic acid (SA) were used at 1:5 dilution with QC samples.

Acquisition was performed using Bruker Ultraflex III and Ciphergen PBS II TOF mass spectrometers. For the Ultraflex, three mass ranges were scanned: 1–20 kDa, 2–100 kDa, and 15–150 kDa, and one was scanned for the PBS: 1–150 kDa. The broad mass range acquisition optimization for the Ultraflex instrument was performed as discussed elsewhere.[8] Range-specific time-lag focusing and three dwell times were used: 1 ns or 2 ns (Ultraflex) and 4 ns (PBS). An initial time delay was automatically recorded by the instrument during acquisition. To assess systematic errors, several Ultraflex spectra were acquired one year apart for the same samples under similar instrumental settings. The delay values, dependent on the mass-range settings, were 0 μs for the PBS, and for the Ultraflex were: 15.177, 24.012 (in year 2), 28.056, 30.587 (in year 2), 30.949, and 103.977 μs. The spectra were obtained by laser desorption at 10 positions with 100 shots averaged per position for the Ultraflex, and 12 positions with 16 shots averaged for the PBS.

Signal processing

As a benchmark, Bruker flexAnalysis version 2.4 was used for signal processing and calibration of the Ultraflex spectra before automatic assignment (see below). Commercial signal processing included median baseline subtraction, Savitsky-Golay filter, and 'sums' peak detection. Currently, none of the available commercial algorithms accommodates the changing peak width of the data over a broad m/z-range. Thus, whenever the signal-to-noise ratio (SNR) thresholds were set to detect low-intensity heavy-mass peaks ($SNR < 4$), a large false detection rate (~40%) was observed for the early mass region. Furthermore, the commercial filtering and smoothing algorithms produced artifacts (multiplets) for higher masses (>20 kDa), complicating automated signal detection.

Our custom processing[8,9] included Gaussian, or exponential baseline de-trending, integrative down-sampling (IDS), optimal linear filtering (OLF) and trivial (first difference) peak detection.[12] Integrative down-sampling according to the peak-width dependence on TOF ensures that the signal width is constant over a full mass range and that the filtering does not produce artifacts (in contrast to flexAnalysis). Preprocessing by signal integration and filtering enhanced the sensitivity of the peak detection.[8,9] With the SNR threshold set to 3–4, no false signals were detected by this procedure for calibration mixtures. The peak-width-dependent down-sampling added a benefit in that the detected peak intensities (local maxima) directly represented the area under the peak, which was required for our downstream ionization state assignment algorithm.

TOF to m/z calibration

In the following sections italicized abbreviations will denote specific variables or measured quantities used by equations or algorithms in contrast to the same abbreviations used as

DOI: 10.1002/rcm

general terms. The universal calibration equation was assumed (according to the manufacturer's specifications) to be a quadratic in *TOF*:

$$\begin{aligned} TOF &= TD + (t_i - 1)DW, \\ m/z &= C_2 TOF^2 + C_1 TOF + C_0 \\ &= C_2\left(TOF + \tfrac{C_1}{2C_2}\right)^2 + C_0 - \tfrac{C_1^2}{4C_2} \end{aligned} \quad (1)$$

where *TD* is the initial instrumental acquisition delay, $t_i = 1,2..N$, is an indexed time-of-arrival for a signal in a TOF record of length *N*, and *DW* is the acquisition dwell time in nanoseconds (ns) associated with each signal, i.e., the time between successive indices (1, 2 and 4 ns for the data discussed in this paper). C_2, C_1, C_0 are the coefficients obtained through the least-squares fit to a set of known masses. For the acquired MS data, *TD* ranged from 0 to 110 μs (see '*MS acquisition*' section above), depending on which of the two instruments we used, and also on the mass range specified for acquisition. Both Ultraflex and PBS data were internally calibrated using *automated assignment* after signal processing. As described in the previous section, a large false detection rate in flexAnalysis frequently prevented automatic assignment of internal calibrants.

The calibration constants recorded by the flexAnalysis software for the linear *TOF* of Eqn. (1) produced an *m*/*z* axis significantly different from the axis directly exported from this software. In our communication with factory software engineers, they indicated that they use a proprietary calibration procedure that apparently constructs a *non-linear TOF* before applying the quadratic fit in Eqn. (1). Thus, the calibration coefficients that the flexAnalysis software provides cannot be applied to a *linear TOF*. Since there is no functionality in the factory software to export the re-calibrated *TOF* axis separately, the only direct way to compare to the instrumental calibration was to export the *m*/*z* axis from flexAnalysis after both *TOF* and *m*/*z* calibration was performed. For the PBS instrumental calibration,[7,9] the calibration constants provided by the instrument were applicable to the *linear TOF* of Eqn. (1). This required only an appropriate scaling of the time units from microseconds to nanoseconds and regrouping to the universal polynomial form in *TOF* (Eqn. (1)). Our custom auto-recalibration procedure described below also used Eqn. (1) with a *linear TOF*.

Ionization state assignment algorithm

Initial delay correction

The physics of TOF mass separation[13] ensures that the C_2 coefficient of the quadratic time term in Eqn. (1) is at least three orders of magnitude larger than the other terms. In fact, since the linear term can be absorbed as a time zero shift, the constant term typically represents an *m*/*z* correction of less than 1 Dalton. We observed this to be true in different mass ranges and under different mass focusing regimes both for the Ultraflex and the PBS instruments:

$$\begin{aligned} m/z &\approx C_2 TOF^2 + \varepsilon \\ TOF(z) &\approx \begin{cases} TOF(+1)/\sqrt{z} & z = +2, +3, .. \\ TOF(+1)\sqrt{w} & w = +2, +3, .. \end{cases} \end{aligned} \quad (2)$$

here ε is a small correction of the order $10^{-3}TOF$. Therefore, for low resolution, $M/\Delta M < 1000$, characteristic of broad-range mass spectra, the linearly inverse scaling of successive ionization states (which we call *z*-scaling, for charge states $[MzH]^{+z}$, $z = 1,2,3...$) in the *m*/*z* domain is well approximated by the inverse square-root scaling in *TOF* (Eqn. (2)). Note that this scaling, which only depends on the correct choice of a zero time as long as the constant term in Eqn. (1) is small, is independent of the quadratic coefficient. Thus, this *z*-scaling in the time domain only depends on the time of the true zero and is therefore relatively independent of the *m*/*z* calibration. For singly charged multimer ions $[wMH]^+$, formed by binding together *w* identical molecules (*w*-scaling), the corresponding scaling in *TOF* will be directly proportional to the $\sqrt{w}$.

This simple approximation allows prediction of the *TOF* positions for ionization states for a given $[MH]^+$ *before mass calibration*. When peaks associated with +1 charge states are present, the 'predicted' peak locations for their corresponding multimers (*w*) and multiply charged states (*z*), *TOF*(*w* or *z*), overlap some of the observed *TOF* signals. If the zero time is correct, and the constant term in Eqn. (1) is small, then only when the true distance is subtracted, will the cross-correlation between measured peak locations and those scaled by the $\sqrt{z}$ (Eqn. (2)) be a maximum. Thus, we can determine the true initial time delay as that which induces the maximum number of overlapping peaks. This single zero time correction (typically, $\delta TD < 0.03TD < 1$ μs) is sufficient to align ionization states with the observed peak locations to *better than the half-width accuracy* for the majority of charge states and multimers. Note that this one correction simultaneously aligns +2, +3, *etc*. states to their corresponding (*z*-scaled) +1 locations in a spectrum, and also aligns multimers $[wMH]^+$ to their corresponding singly charged (*w*-scaled) molecular states.

Iterative ionization state assignment

After choosing the optimal zero time shift, we proceeded to perform an ionization (charge and multimer) state assignment step by following a simple algorithm based on several empirical rules, which are valid for a MALDI process:[2] (1) the parent has the highest intensity in its ionization set; (2) the intensity of the charge state *z*+1 is lower than charge state *z*; (3) charge state *z*+2 may appear in the spectrum only if both *z* and *z*+1 are present. Starting from the most intense peaks in the list, the procedure checks for the presence of all possible charge states and multimers (*z* and *w*) within the measured *TOF* range that satisfy the above assumptions by looking for the peak locations within one *TOF* half-width of the predicted state:

$$\begin{aligned} TOF(z) &= TOF(+1)/\sqrt{z} \pm HW_{TOF(+1)}/\sqrt{z} \\ TOF(w) &= TOF(+1)\sqrt{w} \pm HW_{TOF(+1)}\sqrt{w} \end{aligned} \quad (3)$$

When detection is completed for a subset from a particular molecular ion, signals from this set are excluded from further assignment. The next highest intensity signal is assigned as a putative singly charged (+1) species of a new subset which is similarly subjected to the same iterative procedure. This process continues until no further subsets can be assigned. Thus, the peak list for a single spectrum is automatically assigned to a combination of charge-state and multimer

DOI: 10.1002/rcm

subsets, each subset consisting of products (z, w) coming from a particular +1 (molecular) ion.

Automated reconstruction of +1 spectrum and recalibration

The intensities of the corresponding assigned (z, w) ionization states were added to the appropriate +1 intensity to reconstruct a 'molecular' spectrum having lower redundancy and higher signal to noise than the original. These model spectra were created using calibrated peak positions and measured their intensities and widths, using one of several types of lineshapes. For Ultraflex spectra using CHCA matrix we chose symmetric Lorentzian lineshapes, for Ultraflex with SA matrix we used Gaussian forms, and for PBS spectra with SA matrix we used skewed half-Gaussian-half-Lorentzian. These were selected as best representations for these instruments based on our earlier work on linear filtering in signal processing.[9] Another separate option for molecular spectra was to use only the locations and intensities of the +1 peaks without addition of ionization states. This eliminates data redundancy without enhancing the signal to noise.

The assigned ionization state locations for $[MH]^+$ with known masses were used in a least-squares fit to a quadratic function relating m/z to TOF to determine the calibration coefficients according to Eqn. (1). The TOF positions in the equation were those corrected for initial time delay error, δTD, and used for automated assignment of ionization states. Our self-calibration algorithm was fully automated, and used only the 'known' molecular weights for the chemical components of the mixture (provided as inputs to the calibration routine by the user) and their auto-detected ionization states (computed). The most abundant internal calibrant m/z (within the full-width) was chosen as an 'anchor' to automatically match the mass list for calibrants with the detected +1 TOF positions. By changing the list of 'known' masses to include only low-mass (<20 kDa) or high-mass (>20 kDa) components for the broad-range spectra, piecewise calibration improved m/z precision.

RESULTS AND DISCUSSION

The automated ionization (charge and multimer) state assignment algorithm consisted of two steps: time delay correction and iterative ionization state assignment. To evaluate the algorithm performance and limitations, we applied it to broad mass-range MALDI-TOF data for protein mixtures generated under different experimental settings using two different instruments (see Experimental section).

Initial delay correction

Time delay correction is illustrated in Fig. 1, by alignment of 'predicted' ionization states with the observed peak positions for the mixture of seven calibration proteins (PS1+PS2).The data were processed, according to methods described in earlier papers,[7–9] to enhance signal detection sensitivity more than ten-fold, especially for the late arrival times (the PS2 components). After processing, 26 signals, spanning the full range of the TOF record, were detected[8,12] above $SNR = 3$. Without such prior processing, only four of the most abundant (true) PS1 peaks would have been detected with the same SNR threshold, and equally importantly, five small, additional noise peaks would have been erroneously auto-tagged as peaks.

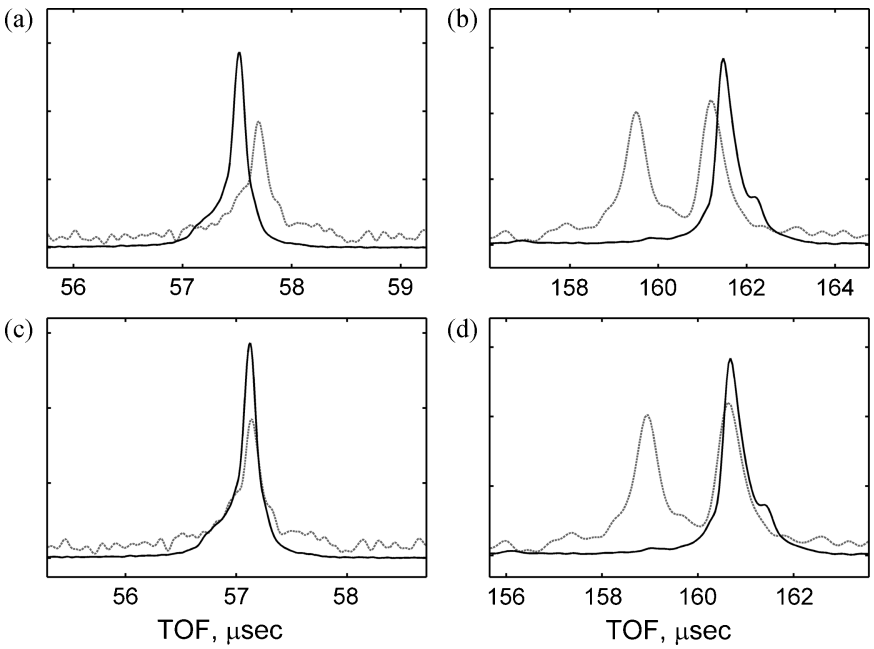

Figure 1. Zoom-in of the regions of the PS1+PS2 TOF spectrum: (a) an illustration of TOF shift for $z = 2$ charge state of ubiquitin (dashed) with respect to the scaled $TOF/\sqrt{2}$ position of the z = 1 charge state (solid); (b) an illustration of TOF shift for the myoglobin dimer $w = 2$ (dashed) with respect to the scaled $TOF\sqrt{2}$ position of the $w = 1$ state (solid). The lower panes (c, d) show the same figures as on top after correcting both by the same zero time shift.

DOI: 10.1002/rcm

The top two panes of Fig. 1 show the data before the initial time delay correction, where the z-scaled (solid trace) spectrum is shifted from the observed positions (dashed trace). On the bottom, the shift for $z = +2$ is corrected by recalibration of the initial time delay automatically detected by correlation. We have used the $z,w = 2$ scale for the *TOF* delay correction, since it usually shows the largest number of peaks contributing to correlation for MALDI spectra. The detected time delay correction that was applied was less than 3% of original instrumental initial time delay ($\delta TD = 0.84$ μs, $TD = 30.95$ μs). After correction, the predicted (z,w –scaled) and the observed peaks were aligned to better than one half-width precision. Note that a single correction is sufficient to align both the $z = +2$ and the $w = 2$ (dimer) states. In fact, a single delay correction was sufficient to align all ionization states over the *TOF* range from 20 to 200 μs (2–100 kDa) with accuracy better than one half-width. The highest ionization states observed for the PS1+PS2 data were $z = +3$ and $w = 2$. For the QC data collected under the same experimental settings as PS1+PS2, this procedure aligned charge states up to $z = +5$, which were observed for serum albumin.

The assumption of a constant delay correction with a quadratic *TOF* to m/z transformation (Eqn. (1)) is only approximately correct over a broad mass range, so the alignment of the signals with predicted positions is not 'exact' for all the peak locations. For broad-range data, the regions with the highest density of the z-scaled states will determine the best correction value. The majority of $z = +2$ peaks were observed between 50 and 100 μs. By changing the range of the peaks in the list, we determined that the actual optimal delay correction is slightly lower (0.56 μs) for early signals (*TOF* <100 μs) and slightly higher (1.12 μs) for late signals (*TOF* >100 μs) compared to the 'average' (full-range) $\delta TD = 0.84$ μs. However, due to the nearly quadratic dependence of peak-width on *TOF*,[7–9] using an 'average' time delay correction is adequate to align the predicted and the observed positions within a half-width tolerance over the entire mass range of a typical broad-range MALDI-TOF scan. This was also true for the PR mixture (PBS data), where the optimal broad mass (1–150 kDa) correction was 0.17 μs, while the best correction for later components only (>20 kDa) was 0.35 μs. For the PS1 (1–20 kDa) and PS2 (15–150 kDa) mixtures separately, the corresponding full-range time delay corrections were 0.75 and 0.96 μs, respectively. For the Ultraflex spectra taken under the same instrumental settings (except for the laser power) a year later, the 'average' corrections were 0.63, 0.73 and 1.2 μs for the PS1, PS1+PS2 and PS2, respectively. We note that these shifts do not appear to change much when the recorded instrumental zero time (*TD*) does not change, so this shift might be a systematic characteristic of the instrument electronics. If this systematic time shift is known or characterized for a specific instrument, then it can be set as a parameter in our algorithm. However, for the sake of generality, the automated correction calculation option is provided.

The top and bottom graphs in Fig. 2 compare, respectively, the early *TOF* (low mass) spectra for the PS1 and late *TOF* (heavy mass) spectra for PS2 versus the mixed PS1 and PS2

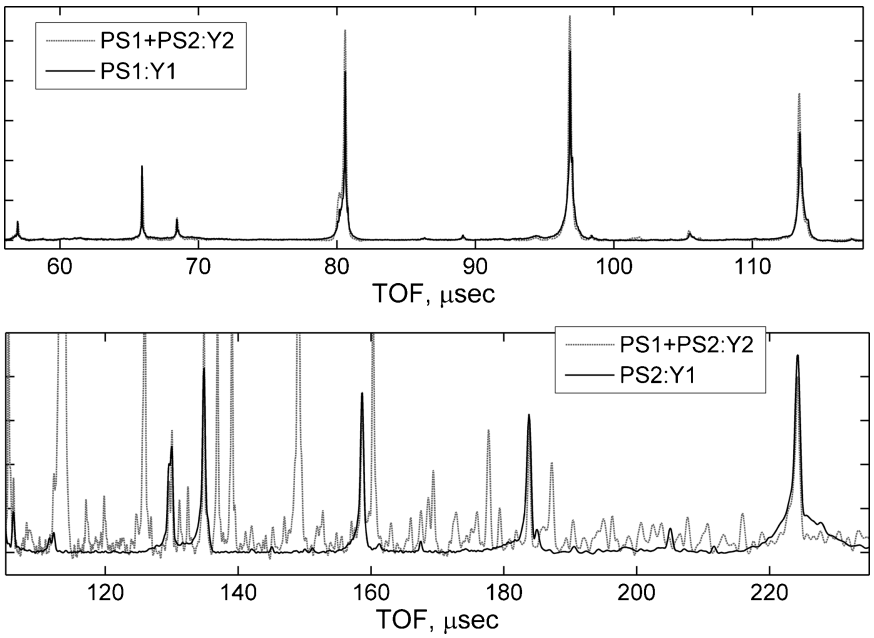

Figure 2. The top graph compares early *TOF* (low mass) spectrum for the PS1 (4 proteins) versus the one for the PS1+PS2 mixture (collected 1 later – Y2). PS1+PS2 data in the bottom pane is normalized to the albumin (right most) peak intensity. The lower trace shows that the combined PS1 and PS2 mixture (7 proteins) has much more structure than PS2 (3 proteins). This is likely due to formation of heteromolecular multimers in addition to the multiple ionization states of PS1 components, similar to those identified in the caption to Fig. 3. The lines that are common to the just two spectra are well aligned in *TOF* after the time delay correction.

DOI: 10.1002/rcm

after initial delay correction. Different, but complimentary, compositions facilitate verification of assignments of PS1+PS2 mixture components that result from individual contributions of PS1 or PS2 ions. Additionally, these spectra were taken approximately one year apart, so their excellent agreement in *TOF* can be taken as evidence of only small systematic errors in the method. Due to different experimental settings, the PS1 spectrum has higher resolution than the broad-range PS1+PS2, showing some salt adduct and neutral loss structure around major peaks. The lower trace shows that the combined PS1 and PS2 mixture (7 proteins) has much more structure than PS2 (3 proteins). This is likely to result from formation of heteromolecular inter-combinations along with multimers from PS1 components in the region from 120 to 220 μs. However, the lines that were common to the two spectra were well aligned in *TOF* after only using the time delay correction (before *m*/*z* calibration).

The precision with which these corrections can be calculated by our algorithm depends on the full width at half-maximum (*FWHM*) of the peaks, which is the minimal distance between the detectable peaks. The minimum *FWHM* in *TOF* depends on the instrumental focusing parameters and the dwell time, *DW*, which was smallest (1 ns) for the PS1 mixture, and largest (4 ns) for the PR. For the data in this paper, the δTD precision was about 0.01 μs for PS1, 0.03 for PS1+PS2, 0.02 for PR and 0.15 μs for PS2 spectra. Thus, the error due to the accuracy of the calculation is much smaller than that due to the approximation of a constant shift over the broad mass range (discussed above). Also, setting the longest correlation lag to <10% of the initial delay, *T*D ~10-100 μs, was sufficient to find the typically small correction shift. This helped reduce the calculation time as well as eliminated correlation contributions from random peak overlaps for large shifts not related to ionization scales.

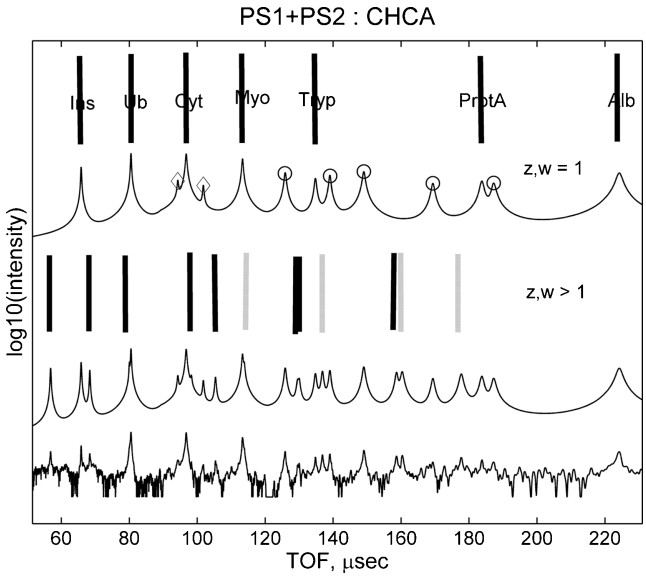

Figure 3. The top lines show the locations of the seven protein components of the PS1+PS2 mixture, labeled as follows: Ins –insulin, Ub – ubiquitin, Cyt – cytochrome C, Myo – myoglobin, Tryp – trypsinogen, ProtA – protein A, Alb –albumin. The curve below it is the reconstructed spectrum including only $z = w = 1$ states. The two diamonds mark detected peaks not associated with the proteins. The five circles mark peaks associated with the following singly charged heterogeneous intermolecular combinations (left to right): Ub + Cyt, Ub + Myo; Cyt + Myo, Ub + Cyt + Myo, and 2Ub + Cyt + Myo. The next unlabeled set of vertical lines marks the locations of $z > 1$ (solid) and $w > 1$ (dashed) peaks. The next traces show a computed spectrum including all detected peaks for comparison to the bottom trace, which shows the original data.

Iterative ionization state assignment

The results of the iterative ionization state assignment procedure are shown in Fig. 3 for the spectrum of the protein standard mixture PS1+PS2 (with CHCA). Before the iterative assignment of ionization states, the *TOF* axis was corrected by $\delta TD = -0.84$ μs shift (Eqn. (1)), after which, the algorithm converged quickly to automatically assign 26 observed signals into 14 ionization subsets, so half of the signals were eliminated as redundant ($z,w > 1$). In Fig. 3, the top lines show the *TOF* of the seven proteins included in the two mixtures, while the second trace shows a constructed signal including only the $z = 1$ and $w = 1$ peaks that were identified. Circles mark the five peaks identified, after *m*/*z* calibration, as heteromolecular combinations of the three most abundant PS1 proteins (see caption). Diamonds mark the two unidentified, low intensity ($SNR < 4$) peaks (possibly adducts or contaminants). The third graph down illustrates where the multimer ($w > 1$) and multiple charge states ($z > 1$) occur. The next trace represents a simulated spectrum using all of the detected peaks, while the bottom trace shows the actual data. Clearly, this spectrum has been greatly simplified, reducing to only the molecular $[MH]^{+1}$ peaks shown at the top.

The automated ionization state assignments were also performed for individual PS1 and PS2 spectra with CHCA (*TOF* shown in Fig. 2), where 4 of 14 peaks in PS1 and 10 of the 19 peaks in PS2 were found to be redundant ($z,w > 1$ states). Because between one-third and one-half (more for higher masses) of the detected peaks in a spectrum are typically identified as redundant after choosing only one parameter, namely, the time zero shift, our algorithm is not very sensitive to occasional peaks that might overlap peaks that are assigned to the ionization sets. Thus, in the pooled serum sample (Fig. 4), 25 of the 80 peaks were identified as redundant, so that even if a few were obscured by nearby peaks, they would not have much effect on the calculated zero time shift. This observation indirectly confirmed that the use of *z*-scale correlation for automatic detection of initial delay correction was warranted. On average, more than half of the ionization subsets assigned in experimental spectra by our automated procedure had a single $[MH]^{+1}$ member, as expected for MALDI processes. Between 30–60% of detected higher intensity +1 ions had detectible multiple ionization states $z,w > 1$ (e.g., 8 out of 14 for PS1+PS2, and 18 out of 55 for QC).

The automated ionization state assignment for the PS1+PS2 mixture was consistent in assigning states for the eight peaks that were in common with PS1 and the six peaks in common with PS2 mixtures (Fig. 2). It also correctly identified the charge states of the inter-combinations of the PS1 components that were not in the PS2 spectrum, and

DOI: 10.1002/rcm

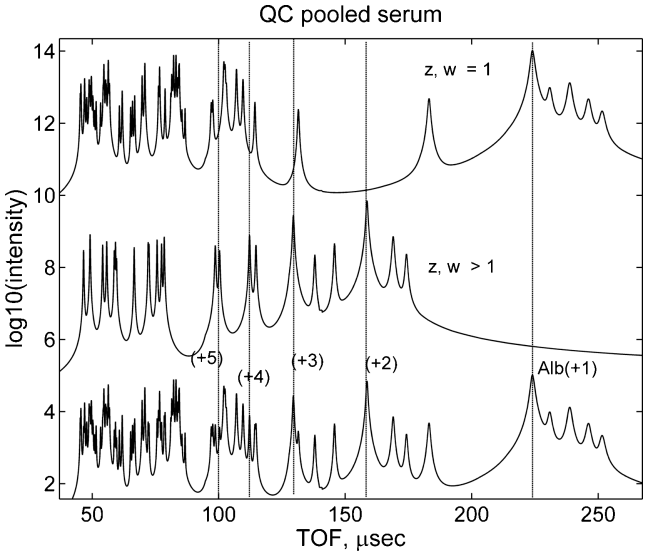

Figure 4. Automated ionization state assignment for the broad mass range QC pooled serum spectrum with CHCA matrix, acquired under experimental setting similar to the PS1+PS2 mixture (Fig. 3). The data is reconstructed from the peak list using Lorentzian peak model, and plotted on a log10 scale. The top graph shows the +1 spectrum, the second down is a combination of all $z,w>1$ states, and the bottom trace is the original reconstructed spectrum for the complete peak list. The five auto-assigned states of albumin (later used for internal calibration) are marked with the dashed vertical lines.

found peaks in PS2 distinct from those of PS1 (Fig. 2). We confirmed our automatic assignments for the PS1 mixture for m/z below 10 kDa by collecting high-resolution spectra that clearly show isotope structures consistent with $z>1$ charge states.[14] Supplementary Fig. 1 (see Supporting Information) illustrates this with a peak identified as ubiquitin, along with a side peak showing a neutral loss of H_2O or NH_4, and its doubly ionized form. The high-resolution data, taken in the reflectron mode, shows the isotopic structure of ubiquitin +2, at half-Da spacing, confirming the automated state assignment for the lower resolution spectra. The apparent 'lower' resolution observed in the reflectron mode for the myoglobin +2 peak preceding ubiquitin +1 was also consistent with a half-Da-spaced isotopic pattern. Similar experimental confirmation of the assignment for ionization states of the pooled serum spectrum is shown for a few low-mass structures in Supplementary Fig. 2 (see Supporting Information). Based on literature reports,[10] major high intensity peak clusters with associated charge states detected in our QC spectra[8] were identifiable as abundant serum proteins (e.g., albumin, transthyretin) and known forms of apolipoproteins (e.g., C-I, C-II, C-III). Some preliminary assignments, pending experimental verification, are shown in the Supporting Information (HUPO 2009 poster).

Although our automated ionization state assignment algorithm can assign the +1 state as compared to $z>1$ and $w>1$ peaks for the mixed-molecule clusters for the PS spectra (e.g., Fig. 3, circles), it does not currently automatically identify them with *specific* combinations of molecules in *TOF*. In principle, this can be accomplished by additional iteration after the ionization state assignment; but this has not yet been implemented. Typically, inter-combination clusters appeared only for highly abundant ions, and the probability of their formation was always lower than that of various states for the corresponding component molecules. In the clinical samples, for example, the levels of detection for the most abundant components of serum purified with beads were typically 5–10 times lower than for the protein standards. Thus no multi-molecular clusters were observed in those samples. We plan to perform a more systematic sensitivity study and summarize our results in future publications.

This ionization state assignment algorithm was also applied to the mass spectra of PS1, PR and leukemia QC samples using a SA matrix. With SA, the low-resolution TOF spectra were complicated by multiple adducts and neutral losses in addition to the ionization states. For instance, the PS1+SA spectrum had 41 peaks, almost three times more than PS1+CHCA (14). The scaling of distances for adducts and neutral losses in the spectra with SA matrix were consistent with the automated charge state assignments[14] (e.g., Supplementary Fig. 1, Supporting Information). Automated assignment of adducts and their ionization states in protein mixtures with the SA matrix has the potential to reduce spectrum dimensionality by 3–5-fold in addition to about 2-fold reduction already achieved by ionization state assignment for the protein components. However, we observed that the empirical rules we used successfully for ionization state intensities of protein components were not always applicable to adducts. In future work we plan to amend the algorithm to automatically detect adducts and neutral losses for the SA spectra.

Our iterative algorithm assigns ionization states for the *peaks within the half-width uncertainty closest to the predicted position* (Eqn. (3)). Since the resolution of peak detection in MS data is typically twice the half-width (full-width), the neighboring peaks are always separated by twice the threshold of the ionization state assignment for our algorithm. Thus, false detection/assignment from neighboring peak clusters is unlikely. For the assignment of ionization states for both QC and PS, at most one peak falls within the uncertainty range of each predicted position. However, without the δTD correction, for example when using the delay assigned by the instrument's program, the shift error for the predicted z-scale positions would be more than twice as high as the signal resolution, causing many ionization states *to be misassigned* in *TOF*. Resolution typically decreases with flight time, as higher mass peaks are broader. Using uncorrected *TD*s from the instrument was less important for low resolution PS2 data, especially at higher masses, where the peak width is higher, yielding misassignment error rates of 10–15%. However, for the low-mass region of PS1 data, where the resolution is high, the error rates associated with ionization state assignment made without correcting the instrumental value of *TD* could be as high as 50%.

Although, misassignments might be allowed to occur if an expected peak were completely missing or if a peak overlapped a peak of another origin, we did not observe this to be a problem for the broad-mass calibration data that we studied. This may reflect the general trend of decreasing

DOI: 10.1002/rcm

intensity with increasing ionization state, as assumed in our algorithm. When spectra of multiple replicates are available for the sample, additional confidence measures can be derived by calculating correlations between peaks across the sample replicates. In preliminary studies of replicate spectra (see Supporting Information: HUPO 2009 poster), we found that correlation is highest between closest charge states, z and $z+1$, but the present algorithm, developed for a *single spectrum*, did not provide such confidence measures.

Applications: dimension reduction and automated recalibration

The most useful application for our ionization state assignment algorithm is in reducing the number of variables and eliminating many variable correlations, thereby improving the statistical analysis in comparative proteomics and clinical research. Since clinical data sets contain a limited number (typically only hundreds) of patient samples, this reduction of the data redundancy improves sensitivity and specificity. Moreover, because these redundant peaks carry the same biological information (the amount of the original molecule) but are usually smaller and noisier than the $z=w=1$ peaks and are more likely to vary with different experimental conditions (e.g., laser power fluctuations, matrix variations), these peaks can be very problematic for many classifiers or diagnostic variable selection. For example, in a tree-based classifier, random ordering of peaks may suggest choosing one or another of a set of the redundant peaks as the start of a branch. Thus highly correlated peaks, such as these redundant sets, may make a classifier or variable selection methodology very unstable. However, if the intensity of the detected multiply charged peaks is added to the +1 peak intensity, this reduces the dimensionality of the spectrum, eliminating a set of highly correlated peaks, and improves the *SNR*. In other cases, it may be best to simply eliminate the redundant entirely, if their variation is too much greater than that of the $z=w=1$ peak. As shown in the example of Fig. 3, up to one-half of the detected peaks can be assigned to redundant ionization states, allowing a proportional dimension reduction. Reconstruction of the molecular spectrum is also useful to determine the molecular targets for protein identification. One of our future goals for such dimension reduction and reconstruction is to similarly include adducts and neutral losses.

As we have shown above, the time delay correction necessary for an automatic assignment of ionization states implies that the m/z calibration equation requires only one constant, which is easily determined from any known peak. Of course, the preference is to use the largest possible number of known signals in a given mass range, to generate the most trustworthy fit of the m/z calibration equation (Eqn. (1)), and the inclusion of charge states into the calibration over a broad range is known to improve calibration precision.[15] Thus, when the ionization states of m/z standards are known, then our algorithm can perform a fully automated internal calibration of TOF records over a broad mass range. Alternatively, the calibration can be performed piece-wise (which allows for a small m/z dependence of the calibration constants of Eqn. (2)). The detection of ionization states for internal standard components, which either can be added to, or may be present in clinical samples naturally increases the number of landmark peaks that can be used for calibration and alignment of data. Instrumental time delay correction is most important for improving the calibration of the low-mass, high-resolution data, while improved peak detection is more important for the high-m/z data, where the peak width changes considerably.

As an example, we used 15 of the 26 detected signals associated with ionization subsets of the seven molecular ions of known masses in PS1+PS2, to recalibrate the m/z axis and found at least a ten-fold greater precision than the flexAnalysis instrumental calibration above 20 kDa (Fig. 5). This recalibration also aligns peaks in replicate spectra to better than a peak half-width over the entire range from 2 to 70 kDa (Fig. 2). Much of the improvement achieved by this recalibration procedure is the result of the enhanced sensitivity of reliable signal detection associated with better signal processing.[8,9] Figure 5 shows that *auto-assignment* using the built-in instrumental algorithm results in large errors when compared to our recalibration procedure using multiple ionization states.[15] The addition of molecular clusters (as calibration peaks) may, in general, help enhance precision, especially in sparse regions of the spectrum. In the samples shown here, peaks from the mixture components adequately covered the entire mass range, so it was not necessary to add molecular clusters.

The symbols in Fig. 5 show errors for the actual peak positions detected by a simple first difference with *SNR* threshold[8,12] both for custom processed data and after the flexAnalysis. Thus, the m/z errors plotted in Fig. 5 have the same contribution from peak detection uncertainty for both methods. We have systematically analyzed the improvement in peak centroid precision after our preprocessing in Ref. 9. In Fig. 5, the lower precision of the peak location of a few of the closely overlapping peaks (e.g., at 33 kDa) is responsible for higher absolute error after calibration for both commercial

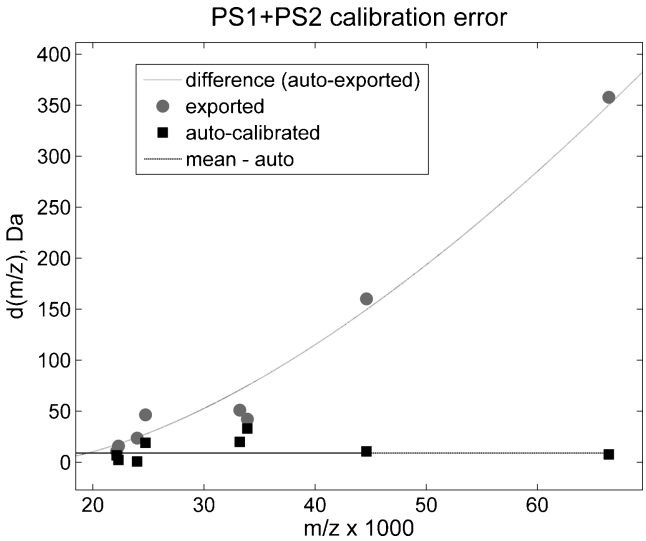

Figure 5. Mean errors in *m/z* using the auto-calibration method that auto-detects high-mass peaks (squares) versus the exported flexAnalysis *m/z*, based on the peaks auto-detected by flexAnalysis. The lines are plotted to guide the eye.

DOI: 10.1002/rcm

and custom procedures, but these close peaks did not have much effect on the actual calibration coefficients.

When 'automatic assignment' algorithms are compared, for the PS1 standard mixture, four major (+1) peaks are auto-assigned by flexAnalysis; the errors are comparable for our algorithm. However, for the PS1+PS2 mixture only four major (+1) peaks are auto-assigned by flexAnalysis. These are entirely PS1 components; none of the PS2 components are auto-detected by flexAnalysis, which produces large calibration errors in the >20 kDa range. We attribute this to the inability of the flexAnalysis signal processing algorithms (see Experimental section) to accommodate for the changing peak width over the broad range of *m/z*. For the PS2-only mixture, flexAnalysis auto-detected only one component, making quadratic auto-calibration impossible. Our algorithm had no such difficulty. As expected, 'manual' calibration in flexAnalysis using the same peak standards as detected by our automated algorithm provided comparable accuracy.

Our automated procedure was applied to a pooled serum spectrum (Fig. 4), where five charge states of albumin were readily detected and auto-assigned. This gave an automated recalibration over the range from 13 to 70 kDa. Our recalibration for PS1+PS2 protein standards was consistent with the *m/z* assignments for these albumin states in pooled serum, providing encouraging evidence of instrumental stability. Moreover, for pooled serum records acquired under similar instrumental conditions a year later, recalibration only required the *TOF* zero shift to reproduce the *m/z* calibration of the previous year to within one half-width precision. Note that this *TOF* recalibration requires only a single parameter correction, for instrumental acquisition delay (*TD*), which is automatically computed from the data record and applied to correct and assign peaks from the data record without the need for further human intervention or judgment.

CONCLUSIONS

We have described an algorithm that automatically detects and corrects for systematic initial delay shifts and assigns charge states and multimers under a wide range of instrumental MALDI-TOF settings. The algorithm performance was benchmarked using spectra for protein standard mixtures and pooled serum over a mass range of 1–150 kDa. The higher the data resolution, the more important is the delay correction for the correct assignment of the ionization states. Up to a half of the detected signals in a MALDI-TOF spectrum are associated with higher ionization states of the protein components, mainly doubly charged, but some carrying higher charge, and some representing molecular combinations such as dimers also appear. Additional sensitivity for molecular ion detection, along with a reduction of spectral redundancy, is achievable by adding the signal intensities of a set of charge states and multimers associated with a particular molecular species, using that sum of intensities to represent that species. Ionization state detection in *TOF* allows the ionization states to be used for an automated *m/z* calibration. When higher calibration precision is required over a broad mass range, such ionization states provide more precise piecewise re-calibration. Because they are entirely automated, and require no human intervention or judgment, the techniques described in this paper will be useful for high-throughput applications like comparative clinical proteomics[4,5,10,11] and mass spectrometry imaging.[6] We are presently extending this method for automated detection and assignment of ionization adducts (matrix, salts) and neutral losses, as well as for preliminary assessments of post-translational and metabolic protein modifications (see Supporting Information: HUPO 2009 poster). Note: Matlab scripts implementing the automated procedures described in this paper are available at no cost for academic users on request from the authors.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article.

Acknowledgements

This research was supported by NIH grant CA126118 from the Advanced Proteomics Platforms and Computational Sciences Program within the Clinical Proteomics Initiative of the National Cancer Institute. We wish to thank our collaborators from Eastern Virginia Medical School, Prof. John Semmes, for overseeing data collection experiments in his Proteomics Laboratory, and Dr. Lisa Cazares for experimental support. We would also like to acknowledge helpful discussions with Prof. Gene Tracy of the College of William and Mary and Dr. Sergei Dickler of Bruker Daltonics.

REFERENCES

1. Karas M, Hillenkamp F. *Anal. Chem.* 1988; **60**: 2299.
2. Zenobi R, Knochenmuss R. *Mass Spectrom. Rev.* 1998; **17**: 337.
3. Semmes OJ, Cazares LH, Ward MD, Qi L, Moody M, Maloney E, Morris J, Trosset MW, Hisada M, Gygi S, Jacobson S. *Leukemia* 2005; **19**: 1229.
4. Baumann S, Ceglarek U, Fiedler GM, Lembcke J, Leichtle A, Thiery J. *Clin. Chem.* 2005; **51**: 973.
5. Wilkes JG, Buzatu DA, Dare DJ, Dragan YP, Chiarelli MP, Holland RD, Beaudoin M, Heinze TM, Nayak R, Shvartsburg AA. *Rapid Commun. Mass Spectrom.* 2006; **20**: 1595.
6. Chaurand P, Norris JL, Cornett DS, Mobley JA, Caprioli RM. *J. Proteome Res.* 2006; **5**: 2889.
7. Malyarenko DI, Cooke WE, Adam BL, Malik G, Chen H, Tracy ER, Trosset MW, Sasinowski M, Semmes OJ, Manos DM. *Clin. Chem.* 2005; **51**: 65.
8. Gatlin-Bunai CL, Cazares LH, Cooke WE, Semmes OJ, Malyarenko DI. *J. Proteome Res.* 2007; **6**: 4517.
9. Malyarenko DI, Cooke WE, Tracy ER, Drake RR, Shin S, Semmes OJ, Sasinowski M, Manos DM. *Rapid Commun. Mass Spectrom.* 2006; **20**: 1670.
10. Hortin GL. *Clin. Chem.* 2006; **52**: 1223.
11. Rai AJ, Stemmer PM, Zhang Z, Adam BL, Morgan WT, Caffrey RE, Podust VN, Patel M, Lim LY, Shipulina NV, Chan DW, Semmes OJ, Leung HC. *Proteomics* 2005; **5**: 3467.
12. Morris JS, Coombes KR, Koomen J, Baggerly KA, Kobayashi R. *Bioinformatics* 2005; **21**: 1764.
13. Cotter RJ. *Time-of-Flight Mass Spectrometry*. ACS: Washington, DC, 1997; 326.
14. Malyarenko D, Bunai C, Tracy M, Nyalwidhe J, Cazares L, Manos D, Kuschner K, Tracy E, Cooke W. *Proc. 56th ASMS Conf. Mass Spectrometry and Allied Topics*, CDROM, June 2008.
15. Vera CC, Zubarev R, Ehring H, Hakansson P, Sunqvist BR. *Rapid Commun. Mass Spectrom.* 1996; **10**: 1429.

DOI: 10.1002/rcm