RAPID COMMUNICATIONS IN MASS SPECTROMETRY
Rapid Commun. Mass Spectrom. 2009; **23**: 3448–3456
Published online in Wiley InterScience (www.interscience.wiley.com) DOI: 10.1002/rcm.4272

Isotope pattern vector based tandem mass spectral data calibration for improved peptide and protein identification

Jingfen Zhang[1], Dong Xu[2], Wen Gao[3], Guohui Lin[4]* and Simin He[3]**

[1]Computer Science, University of Missouri, 110 Life Sciences Building, Columbia, MO 65211, USA
[2]Digital Biology, University of Missouri-Columbia Computer Science Department, 1201 East Rollins Road, Columbia, MO 65211, USA
[3]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China
[4]Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada T6G 2E8

Received 26 May 2009; Revised 1 September 2009; Accepted 6 September 2009

Tandem mass spectra contain noisy peaks which make peak picking for peptide identification difficult. Moreover, all spectral peaks can be shifted due to systematic measurement errors. In this paper, a novel use of an *isotope pattern vector* (IPV) is proposed for denoising and systematic measurement error prediction. By matching the experimental IPVs with the theoretical IPVs of candidate fragment ions, true ionic peaks can be identified. Furthermore, these identified experimental IPVs and their corresponding theoretical IPVs are used in an optimization process to predict the systematic measurement error associated with the target spectrum. In return, the subsequent spectral data calibration based on the predicted systematic measurement error enhances the data quality. We show that such an integrated denoising and calibration process leads to significantly improved peptide and protein identification. Different from the commonly employed chemical calibration methods, our IPV-based method is a purely computational method for individual spectra analysis and globally optimizes the use of spectral data. Copyright © 2009 John Wiley & Sons, Ltd.

Peptide and protein identification by *peptide mass fingerprinting* (PMF) and *tandem mass spectra* (MS/MS) plays an indispensable role in current proteomic research.[1–3] Despite many great efforts,[4–11] obtaining reliable identification results computationally remains a challenging problem due to two major sources of complexities.[3] One complexity relates to the issues in protein database searching such as unexpected peptide modifications, limited databases, and non-specific cleavages, all of which require more efficient identification algorithms and biological knowledge. The other complexity is due to the spectral data being usually blended with noise and often shifted because of unknown systematic measurement errors. In the literature, most of the efforts are to resolve the first source of complexities, while the second is much less studied. Nevertheless, effective denoising and calibration adjusting the systematic measurement error are important for peptide and protein identification, and they deserve equal research. In this paper, we propose a novel use of an *isotope pattern vector* (IPV) for denoising and systematic measurement error prediction. We demonstrate using real MS/MS datasets that the calibrated spectra by our method lead to significantly improved peptide and protein identification results.

**Correspondence to*: G. H. Lin, Department of Computing Science, University of Alberta, Edmonton, Alberta T6G 2E8, Canada.
E-mail: ghlin@cs.ualberta.ca
***Correspondence to*: S. M. He, Institute of Computing Technology, Chinese Academy of Sciences, P.O. Box 2704, Beijing 100080, China.
E-mail: smhe@ict.ac.cn

A typical MS/MS spectrum contains hundreds to thousands of mass-to-charge (m/z) peaks. Among these peaks, only tens are true peaks that come from common types of ionic peptide fragments, which are useful for peptide (protein) identification. The rest of them are noise peaks. Without an effective denoising process, noise peaks can impose a heavy computational burden for peptide identification (through either database search or *de novo* sequencing). Furthermore, they will also lead to many false peptides. To make things even worse, the systematic m/z measurement error, if not removed, will shift all the true m/z peaks, imposing a bigger challenge to the *de novo* sequencing algorithm and the database search. In severe cases, it can force the database search to have a large tolerance threshold, and thus degrades a highly accurate MS/MS instrument to a medium accurate one.[12]

To date, spectral data denoising and systematic measurement error correction have been well recognized as two key steps to improve the peptide and protein identification. Several experimental and computational methods have been proposed for these purposes. Roughly, existing denoising methods can be categorized as threshold filtering, de-isotoping and denoise transforming. Among these three categories, threshold filtering is probably the most straightforward. For instance, some methods select only peaks with intensities greater than a given threshold[4] or a computed threshold[13,14] for peptide identification; the other selects only a specific number of the most intensive peaks in the specified m/z intervals.[15] Since the intensity is not the fundamental attribute of true peaks, these

threshold filtering methods cannot thoroughly remove the noise, but may miss true peaks with low intensities. De-isotoping methods all assume an elemental composition $(C_6H_5NO)_n$[14,16] or $(C_{4.9384}H_{7.7583}N_{1.3577}O_{1.4773}S_{0.0417})_n$[17–19] for an observed peak with mass M. Based on the assumed elemental composition, they can calculate the theoretical distribution of the isotopic peaks and identify some of the true peaks. However, in general, these methods do not validate whether or not the base peak M is true, and the assumed elemental composition could be too crude to identify the complex convolution of isotopes. Consequently, the de-isotoping methods would inevitably miss a certain portion of fragment ions. Other denoising methods employ well-known data transformation techniques such as wavelets.[20] The success of these methods heavily relies on the internal parameters such as wavelet base functions, order, and the level of decomposition. In addition, some commercial software, such as ProteinLynx Global Server,[21] supply certain denoising and de-isotoping functionalities. But they can only deal with raw but not centroid spectral data.

Besides the noise issue, spectral data analysis is also confronted by the systematic measurement errors. These errors are caused during measurement by environmental factors such as temperature. The consequence is that all the true peaks in the spectra are shifted, which make the mismatches between true peaks and theoretical peaks too big for peptide identification. For example, it is observed that even with a careful 5-ppm-accuracy instrument-wise calibration, the systematic measurement errors of many time-of-flight (TOF) spectra could still be as high as 100 ppm, which makes the peptide identification problem more challenging. Several error correction (after instrument-wise calibration) methods have been developed, including internal standard reference (ISR),[22–24] external standard reference (ESR),[25–27] and computational methods.[28–30] Essentially, ISR adds certain reference materials with known quality into samples, and estimates the spectral measurement error using the difference between the measured mass and the theoretical mass of the references. ISR is in general accurate, but there is a risk of possible cross-pollution between references and samples, and high-intensity reference peaks can depress the true peaks. ESR measures the references and samples independently, and estimates the sample systematic measurement error using the reference systematic measurement error. ESR overcomes the above two disadvantages of ISR, but its error prediction accuracy is generally lower than ISR. One reason is that it is difficult to tune the two experimental conditions to be identical. Computational error correction methods do not depend on extra experiments. For example, commonly occurring background peptides from keratins or trypsin autolysis products have been proposed to be used similarly as the references in the ISR methods for error prediction.[28,29] Nonetheless, it is worth pointing out that the commonly occurring products may not occur in every spectrum. A result-driven method has been proposed to first analyze a set of spectra, and then to pick up highly reliable identification results as references to estimate the measurement error distribution for all the spectra.[30] This method highly depends on the reliable identification results and the estimated measurement error distribution might not represent the true error in individual spectra.

The drawbacks and the resultant performance limitation of existing methods motivated us to develop a novel use of isotope patterns for denoising and systematic measurement error correction purposes. Our way of using isotope patterns, to be presented in detail next, does not need to set up peak intensity thresholds or any other parameters, neither does it pre-assume any elemental composition for the ionic peaks. On the other hand, our method computationally predicts the systematic measurement errors for individual spectra through an optimization process to optimize globally the use of spectral data.

Most amino acids are composed of five elements of hydrogen (H), carbon (C), nitrogen (N), oxygen (O), and sulfur (S). All these five elements have stable isotope patterns in nature. For example, the two most important isotopes of hydrogen are H and D, which have relative abundance of 99.985% and 0.015%, respectively; the three most important isotopes of oxygen are ^{16}O, ^{17}O, and ^{18}O, which have relative abundance of 99.759%, 0.037%, and 0.204%, respectively. During the mass spectral data analysis, when the selection window in the mass spectrometer is properly set, the stabilities of elemental isotope patterns ensure that: (1) the isotopic peaks associated with a peptide or a fragment ion will co-occur with its monoisotopic peak in the same spectrum, and (2) the isotopic peaks will occur in a stable profile also. Such a co-occurrence can be used as a strong evidence that one ion is present in the spectrum. Furthermore, the elemental composition of the ion can be predicted using the stable isotope profile, as we will show in the following. These elemental compositions are then used to calculate the theoretical masses for the corresponding ions, respectively. Subsequently, the differences between experimental masses and their corresponding theoretical masses are used in a least-squares fit to best approximate the systematic measurement error for the target spectrum. With the estimated systematic measurement error, the spectrum is re-calibrated and we show that such re-calibrated spectra can lead to significantly improved peptide identification result.

In the next section, we present the concept of the *isotope pattern vector* (IPV) and the novel way to use IPV for denoising and systematic measurement error prediction. In the Results section, we demonstrate the success of our method by showing that, on all three experimental MS/MS datasets we tested, the calibrated spectra lead to significantly better peptide and protein identification results.

EXPERIMENTAL

Isotope pattern vector (IPV)

Given an ion P, let M denote its monoisotopic mass, and let M_k denote its k-th isotopic mass, with k extra neutrons, for $k = 1, 2, \ldots$. We define the *Isotope Pattern Vector* (IPV) to digitally describe the isotope profile of ion P as IPV $= (M, T_1, T_2, \ldots)$, where T_k $(k = 1, 2, \ldots)$ is the relative abundance of M_k with respect to M.

Theoretical IPV (tIPV)

Consider the abundance of the isotopes and the computational complexity and accuracy, we define the *theoretical*

DOI: 10.1002/rcm

IPV of ion P, tIPV $= (M_t, T_1, T_2)$, using only its first two isotopic masses. Here, M_t is the theoretical monoisotopical mass of the ion, and T_1, T_2 are the relative abundances of the first and second isotopical masses, M_1 and M_2, with respect to M_t. When ion P has a formula $Cn_1Hn_2Nn_3On_4Sn_5$, its tIPV $= (M_t, T_1, T_2)$ is calculated as follows:

$$\begin{aligned} M_t = {} & n_1 * 12 + n_2 * 1.0078 + n_3 * 14.0030 + n_4 * 15.9972 \\ & + n_5 * 31.9721, \\ T_1 = {} & n_1 * q_C + n_2 * q_H + n_3 * q_N + n_4 * q_{O1} + n_5 * q_{S1}, \\ T_2 = {} & n_4 * q_{O2} + n_5 * q_{S2} + \tfrac{1}{2}T_1^2 - \tfrac{1}{2}(n_1 * q_C^2 + n_2 * q_H^2 \\ & + n_3 * q_N^2 + n_4 * q_{O1}^2 + n_5 * q_{S1}^2), \end{aligned}$$

where q_C, q_H, q_N are the relative abundances of ^{13}C to ^{12}C, D to H, and ^{14}N to ^{15}N; and q_{O1}, q_{O2} (q_{S1}, q_{S2}) are the relative abundances of ^{17}O to ^{16}O and ^{18}O to ^{16}O (^{33}S to ^{32}S and ^{34}S to ^{32}S), respectively.

To predict the elemental composition $X = (n_1, n_2, n_3, n_4, n_5)^T$ from the matched eIPV $= (M_e, R_1, R_2)$ and tIPV $= (M_t, T_1, T_2)$, let $\delta_m = M_e - M_t$, $\delta_1 = R_1 - T_1$, and $\delta_2 = R_2 - T_2$. We have the following equations:

$$\begin{aligned} \delta_m = {} & n_1 * 12 + n_2 * 1.0078 + n_3 * 14.0030 + n_4 * 15.9972 \\ & + n_5 * 31.9721 - M_e, \\ \delta_1 = {} & n_1 * q_C + n_2 * q_H + n_3 * q_N + n_4 * q_{O1} + n_5 * q_{S1} - R_1, \\ \delta_2 = {} & n_4 * q_{O2} + n_5 * q_{S2} \\ & - \frac{1}{2}\left(n_1 * q_C^2 + n_2 * q_H^2 + n_3 * q_N^2 + n_4 * q_{O1}^2 + n_5 * q_{S1}^2\right) \\ & + (n_1 * q_C + n_2 * q_H + n_3 * q_N + n_4 * q_O + n_5 * q_S) \\ & * R_1 - \frac{1}{2}R_1^2 - R_2 + \frac{1}{2}\delta_1^2. \end{aligned}$$

We can approximate $[\delta_m \delta_1 \delta_2] = AX + B$ by omitting the residual $\frac{1}{2}\delta_1^2$ from the equation for δ_2, where A is a constant matrix and B is a constant vector as follows:

$$A = \begin{bmatrix} 12 & 1.0078 & 14.0030 & 15.9972 & 31.9721 \\ q_C & q_H & q_N & q_{O1} & q_{S1} \\ q_C R_1 - \frac{1}{2}q_C^2 & q_H R_1 - \frac{1}{2}q_H^2 & q_N R_1 - \frac{1}{2}q_N^2 & q_{O2} + q_{O1} R_1 - \frac{1}{2}q_{O1}^2 & q_{S2} + q_{S1} R_1 - \frac{1}{2}q_{S1}^2 \end{bmatrix} \text{ and } B = \begin{bmatrix} -M_e \\ -R_1 \\ -\frac{1}{2}R_1^2 - R_2 \end{bmatrix}.$$

Experimental IPV (eIPV)

In an MS/MS spectrum, peaks are characterized as (m/z, intensity) pairs, where m/z is the mass-to-charge of a peptide or fragment ion (x-axis) and intensity represents the absolute abundance of the ion (y-axis). For a triplet of isotopic peaks (p_0, p_1, p_2) associated with an ion P, where $p_k = (MZ_k, I_k)$ for $k = 0, 1, 2$, their (common) charge state z can be calculated from the distance between the values of MZ_k. We collect such triplets from the spectrum, allowing p_2 (or both p_1 and p_2) to be missing. For each triplet, we calculate the experimental monoisotopic mass of the ion as $M_e = MZ_0{}^* z - M_H{}^* (z-1)$, where $M_H = 1.0078$ is the proton mass. We define the *experimental* IPV (eIPV) for this ion as eIPV $= (M_e, R_1, R_2) = (M_e, I_1/I_0, I_2/I_0)$. Here I_1/I_0, I_2/I_0 measures the relative abundance of the first and second isotopic ion with respect to the monoisotopic ion, respectively.

Matching eIPV and tIPV

Given a pair of an eIPV $= (M_e, R_1, R_2)$ and a tIPV $= (M_t, T_1, T_2)$, their distance is defined as d(eIPV, tIPV) $= ((M_e - M_t)^2 + (R_1 - T_1)^2 + (R_2 - T_2)^2)^{1/2}$. Using this Euclidean distance measure, an eIPV can be used for both denoising and the systematic measurement error prediction purposes. Essentially, an eIPV comes more likely from an ion than from noise if it matches well with the ion's tIPV. Therefore, its distances to theoretical IPVs can be used to remove noise peaks. Furthermore, one can predict the elemental composition of the ion(s) from the match between eIPV and tIPV. Subsequently, the systematic measurement error in the spectrum can be estimated by comparing the measured masses from the spectral peaks and the theoretical masses from the elemental composition.

It follows that d^2(eIPV, tIPV) $= (\delta_m, \delta_1, \delta_2)^* (\delta_m, \delta_1, \delta_2)^T = X^T A^T A X + 2B^T A X + B^T B$. This way, searching for the tIPV becomes an optimization problem to minimize d^2(eIPV, tIPV). This minimization problem to find the best-fit elemental composition can be solved by some search methods, as reported previously.[31]

IPV-based denoising

The idea underlying the IPV-based denoising is that true isotopic peaks generally cluster together and have a stable distribution, while noise peaks occur rather randomly and independently. There are two main challenges: (1) to determine the tIPV upon observing an eIPV and (2) to separate overlapping peaks due to the mass difference between ions that matches to the mass difference between isotopes.

To solve the first problem, we follow the work done previously[31] to calculate the expected tIPV from the observed monoisotopic mass M_e and the statistical distribution of theoretical ions from a non-redundant database. For this purpose, we calculate the minimum, the mean, and the maximum values for T_1 and T_2 in tIPV for each ion. Then the *normalized deviations* from R_1 (defined as $\min\{|R_1 - T_{1,\min}|, |R_1 - T_{1,\max}|\}/T_{1,\text{mean}}$ when $T_{1,\min} \le R_1 \le T_{1,\max}$, or 0 otherwise) and R_2 (similarly defined) are calculated, respectively. These values are used to define the distance from the eIPV to the tIPV.

Overlapping peaks are difficult to separate. We resolve this issue by following a previously reported method.[32] Essentially, we summarize the overlapping peaks into several predominant types, and modify the above definitions of normalized deviations. For example, the most important type of overlapping peaks is the isotopic peaks of two of the same charged ions with 1 u mass difference. In our past experience, we found that such pairs of ions are often the water-loss and ammonia-loss ions of a common ion. Complex types of overlapping peaks could involve different-charged ions and even noise peaks.

DOI: 10.1002/rcm

IPV-based systematic measurement error prediction

The spectral data measurement error can normally be divided into two parts: random error and systematic error. Random errors follow a zero-mean normal distribution while the systematic errors are determined by the data measurement mechanism of the instrument. For example, the systematic measurement error of a TOF instrument can be expressed as a polynomial function (most of the time, a linear function is typically sufficient) in the theoretical masses of ions, and the ion-trap and Fourier transform ion cyclotron resonance (FT-ICR) instruments have some other specific error functions. Determination of the parameters in these error functions is rather challenging but certainly very important for spectral data recalibration. In this work, we focus on TOF spectra, and approximate the data measurement error using a linear distribution of systematic errors and a normal distribution of random errors. That is, $M_e - M_t = a^* M_t + \varepsilon$, where M_e is the observed mass, M_t is the theoretical mass, a is a constant (called the *relative systematic measurement error*), and $\varepsilon \sim N(0, \sigma^2)$. Our goal is to estimate the value of a for each spectrum by using the spectrum alone.

Following our IPV-based denoising step, we identify a series of eIPVs generated from true ions. For each of these experimental monoisotopic ion masses, we associate it with the predicted theoretical monisotopic ion mass through the optimization process described above. In this way, we have a series of pairs (M_e, M_t). The relative systematic measurement error a is determined by solving the following minimization problem: *minimize* $\sum_{i=1}^{k} ((M_{e,i} - M_{t,i}) - a * M_{t,i})^2$, where $M_{e,i}$ and $M_{t,i}$ are the i-th observed monoisotopic ion mass and the corresponding theoretical one, respectively. We use a least-squares fit method to solve this minimization problem.

Datasets

We used three experimental datasets to test our IPV-based method. Dataset A can be downloaded,[33] which includes 46 195 .dta files. These spectra are produced after analyzing five trypsin-digested gel regions, representative of the yeast proteome in triplicate (i.e., three samples, denoted as a, b, and c) by nanoscale microcapillary LC/MS/MS using quadrupole time-of-flight (Q-TOF) mass spectrometers. These 15 slices are labeled as 1a, 1b, 1c, 3a, 3b, 3c, 5a, ..., 9a, 9b, 9c, respectively. Dataset A is used to demonstrate the extent of improvement in peptide and protein identification through the IPV-based denoising process.

Dataset B contains 52 high-quality Q-TOF spectra with tryptic digestion peptides (whose C-terminus is either R or K). This dataset has been studied by Taylor and Johnson.[8]

Dataset C contains 114 tandem spectra selected from the production of a Q-TOF Ultima Global spectrometer for tryptic digestion peptides of eight proteins: Myoglobin (horse skeletal muscle), BSA (bovine serum albumin), fetuin (fetal calf serum type III), lysozyme (egg white), alpha-lactoalbumin (bovine), BCA (bovine milk), phosvitin (egg yolk), and ribonuclease B (bovine pancreas).[32] All three datasets were used to illustrate the performance of IPV-based systematic measurement error prediction.

RESULTS AND DISCUSSION

We use the peptide and protein identification results to demonstrate the performance of our IPV-based spectral data denoising and systematic measurement error prediction. We compare our results with the results achieved by standard runs of Mascot (version 2.1.02).[7] Mascot from Matrix Science is one of the most popular and powerful search engines that use mass spectral data to identify proteins from primary sequence databases.

IPV-based denoising results

On dataset A, Mascot was used to interpret the downloaded data (called *original data*). Our IPV-based denoising method was applied to the original data to identify true peaks from fragment ions. The identified peaks form the *IPV-processed data*, which was also fed to Mascot for peptide identification.

To estimate the false positive rate of the *peptide-spectral matches* (PSMs), we applied Mascot to search against a composite target-decoy database.[34,35] This database contains all yeast protein sequences in both forward and reverse orientations. The precision of the search is defined as the number of true positive PSMs (TP) divided by the sum of TP and the number of false positive PSMs (FP), i.e., precision = TP/(TP + FP). The Mascot search parameters were set to the same as Elias *et al*.[34] and a similar criterion for score filtering was set to achieve around 99% precision. Table 1 collects the numbers of spectra, peptides, and proteins that are selected/identified by Mascot using the original data and using the IPV-processed data, respectively. Keeping the similar ~1% false positive rate, there are 12.31%, 12.02% and 7.22% more spectra in the three samples, respectively, confidently interpreted by Mascot on the IPV-processed data than on the original data. Consequently, both protein and proteome coverages are improved after applying IPV-based denoising process (Table 1). Specifically, there are on average 11.64% more peptides and 6.56% more proteins confidently identified by Mascot after applying the IPV-based denoising process.

Comparing the two sets of Mascot search results for the three samples a, b, and c, there are 2792, 2634, and 2365 common spectra confidently interpreted by Mascot, respectively. On these common spectra, an average of 15.85% (41.45 vs. 35.78), 6.89% (41.24 vs. 35.28), and 13.04% (38.98 vs. 34.89)

Table 1. Mascot search results on dataset A using the original and the IPV-processed data

	Original data			IPV-processed data			Intersection			Union		
Sample	Spectra	Peptides	Proteins	Spectra	Peptides	Proteins	Spectra	Peptides	Proteins	Spectra	Peptides	Proteins
a	3167	2777	461	3557	3140	503	2792	2469	423	3932	3448	541
b	3011	2597	456	3373	2950	491	2634	2304	409	3750	3243	538
c	2798	2446	453	3000	2640	468	2365	2083	402	3433	3003	519

DOI: 10.1002/rcm

increase in Mascot scores are achieved on three samples, respectively, after the IPV-based denoising process. That is, the Mascot search results on these common spectra are significantly more reliable after the spectra are denoised by our method.

There are 375, 377, and 433 original spectra in the three samples, respectively, interpreted by Mascot but not after being denoised by the IPV; in the opposite direction, there are 765, 739, and 635 original spectra, respectively, which cannot be interpreted by Mascot but are interpreted after the IPV-based denoising. The distributions of Mascot scores on the identified peptides for the two sets of spectra are similar. The mean scores and standard deviations are (25.30, 8.56) and (25.98, 8.93), respectively. These results imply that the reliability of the Mascot search on these two sets of spectra is very close. Further statistics show that the average peptide lengths on these two sets are both 13; nevertheless, the number of peptides longer than 13 identified using the IPV-processed data is much larger than that using the original data (1330 vs. 657). In addition, there are a total of 772 precursors (parent peptides) with mass larger than 2100 Da interpreted. Among them, 276 (35%) large peptides could not be interpreted by Mascot without the IPV-based denoising. These results suggest a strong ability of IPV-based denoising for identifying larger peptides to provide more sequence information for protein identification.

We also compared the average numbers of peaks in spectra that can be interpreted by Mascot. For the spectra that can be interpreted using both the original data and the IPV-processed data, the average number of peaks is 530; for the spectra that can be interpreted using the original data only, the average number of peaks is 364; for the spectra that can be interpreted using the IPV-processed data only, the average number of peaks is 623. These numbers indicate that many spectral peaks would not be interpreted by Mascot without the IPV-based denoising process due to the rich isotopic information and the abundance of noise peaks. However, spectra with much fewer peaks are normally lack of isotopic information, and they would better be interpreted directly by Mascot.

Combining the two sets of identification results by Mascot on dataset A using both the original data and the IPV-processed data, there are 11 115 spectra interpreted (24.1% of the total), 9694 peptides identified, and 1598 proteins identified. The respective increases for the three samples from using the original data alone are 23.8%, 24.0%, and 16.6%. It is worth pointing out that there are still 75.9% spectra in dataset A not interpreted even with the IPV-based denoising process. Besides post-translational modifications, we suspected that the systematic measurement errors in these spectra are larger than the Mascot search parameter of 0.2 u. Our next section of systematic measurement error prediction results confirmed this.

IPV-based systematic measurement error prediction and recalibration results

In this section, we first report the accuracy of systematic measurement error prediction by the IPV-based process, and then examine the subsequent improvement on peptide and protein identification from the data recalibration. In systematic measurement error prediction, we use a linear function to approximate the error distribution, and our goal is to predict the *relative measurement error* (RME). For each spectrum in the three datasets that can be reliably interpreted by Mascot using the original data, we have both the observed mass and the theoretical mass for an ion. These masses are fit to a least-squares method for the minimization problem to obtain the RME. Such an obtained RME is taken as the *true RME* of the spectrum.

Without running Mascot, we can apply the IPV-based method to predict the systematic measurement error the spectrum. The achieved RME is the *predicted RME* of the spectrum. The differences between the predicted RMEs and the true RMEs on datasets B and C are plotted in Figs. 1 and 2

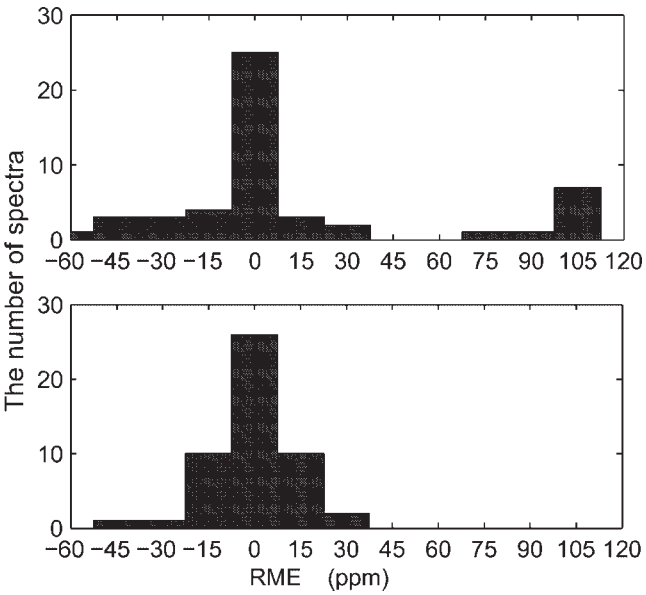

Figure 1. The true RME distribution (top) and the distribution of their difference to the IPV-based predicted RMEs (bottom) on dataset B.

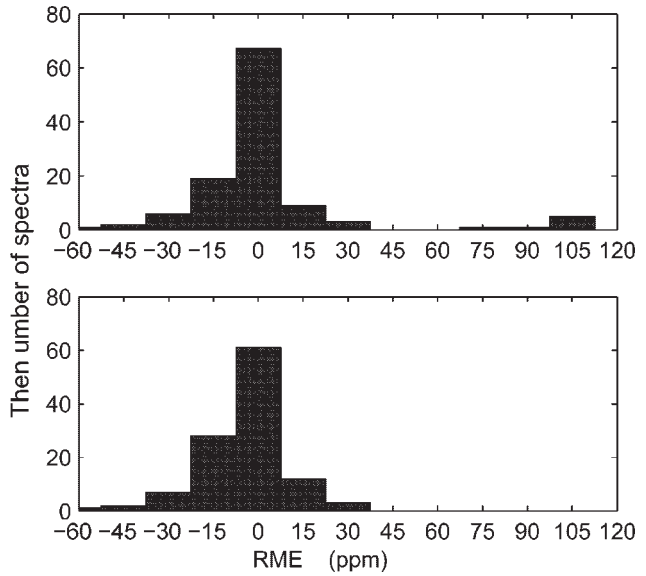

Figure 2. The true RME distribution (top) and the distribution of their difference to the IPV-based predicted RMEs (bottom) on dataset C.

DOI: 10.1002/rcm

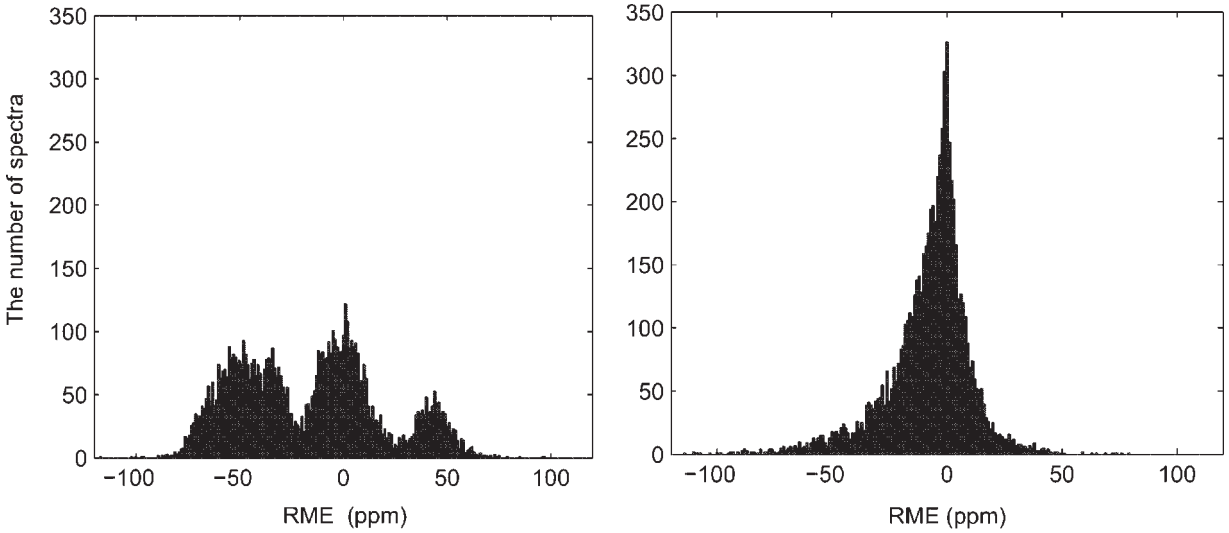

Figure 3. The distribution of the true RMEs (left) and the distribution of the difference between the IPV-based predicted RMEs and the true RMEs (right), on dataset A.

(bottom), respectively. In summary, there are 52 spectra in dataset B and the true RME range is (−56.226, 111.442) ppm (Fig. 1, top). The average difference between the predicted and the true RMEs is 9.604 ppm. Dataset C contains 114 spectra, whose true RME distribution is relatively more complex. The true RME range on dataset C is (−53.197, 103.114) ppm (Fig. 2, top), and the average difference between the predicted and the true RMEs is 10.279 ppm, slightly worse than dataset B. Nevertheless, the proportion of spectra whose RME difference is less than 30 ppm is 89.92%, and the proportion of spectra whose RME difference is less than 40 ppm increased to 94.96%.

For dataset A, the working condition of the instrument for collecting all spectra in three samples was reported stable. Therefore, their systematic measurement errors are expected to be stable too and to follow a normal distribution. The mean of this normal distribution would reflect the error level of the instrument. We selected 3427, 3244, and 2882 spectra from the three samples, respectively: (1) they are reliably interpreted by Mascot using the original data and (2) each containing more than five matched fragment ions. Using the observed and the theoretical masses of the matched ions, we calculated the true RME for each of these spectra. These true RMEs are plotted in Fig. 3 (left). They are also plotted in

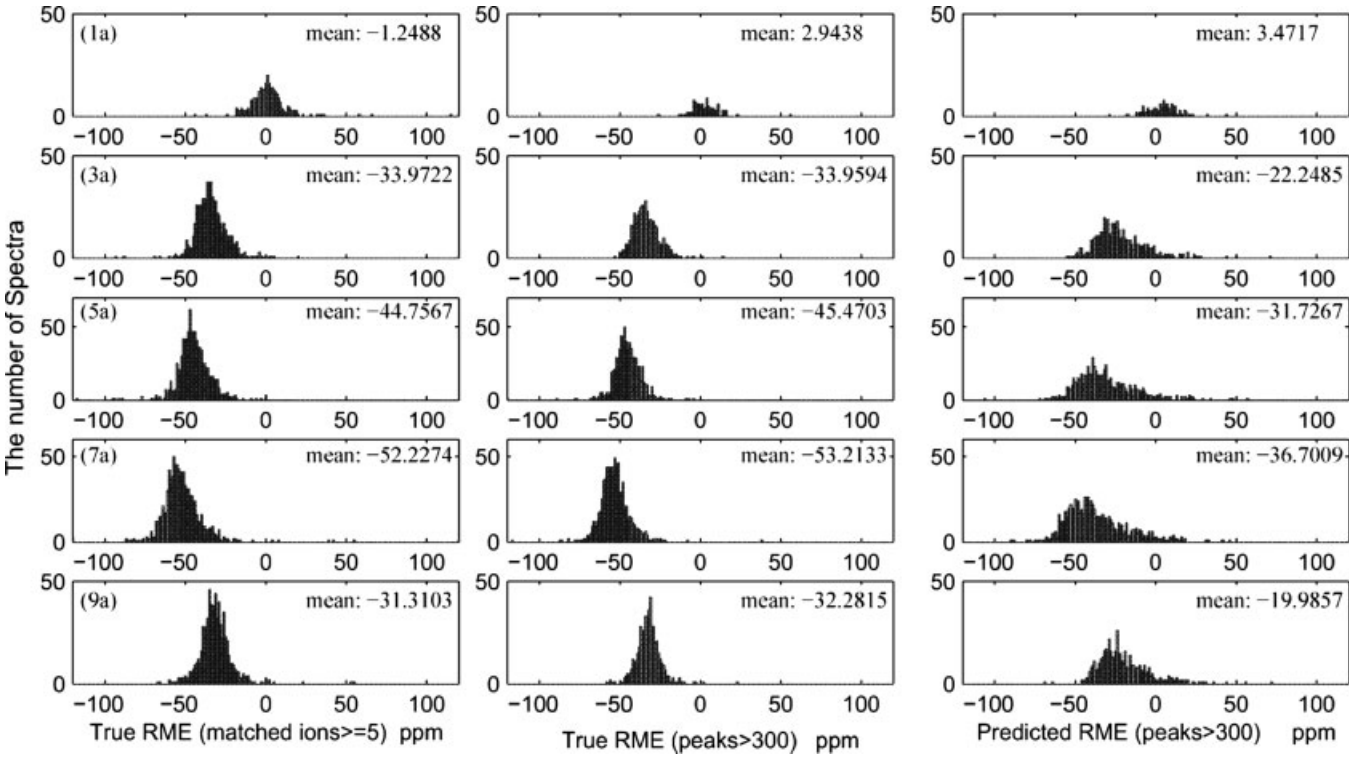

Figure 4. Separated in slices 1a–9a: the true RME distribution in all the spectra (first column), the true RME distribution in the spectra each containing at least 300 peaks (second column), and the distribution of the IPV-based predicted RMEs in the spectra each containing at least 300 peaks (third column).

DOI: 10.1002/rcm

RCM

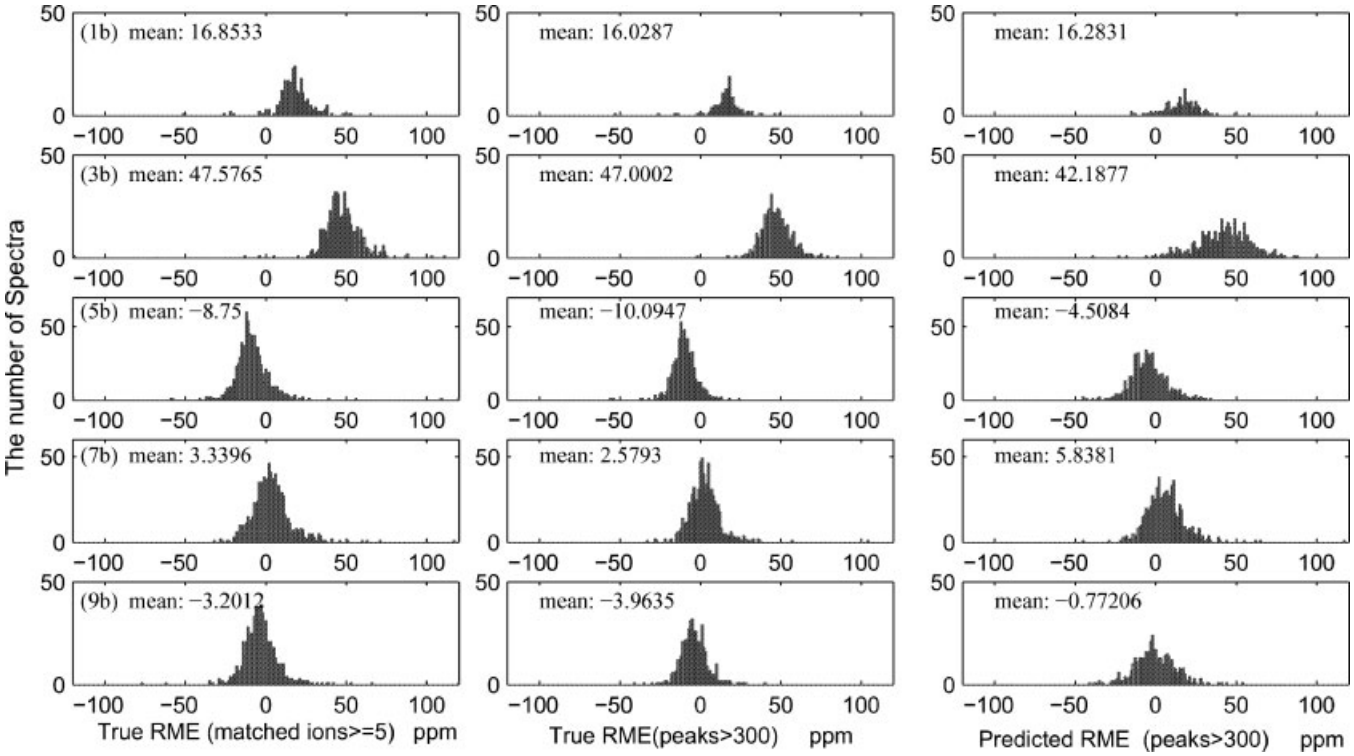

Figure 5. Separated in slices 1b–9b: the true RME distribution in all the spectra (first column), the true RME distribution in the spectra each containing at least 300 peaks (second column), and the distribution of the IPV-based predicted RMEs in the spectra each containing at least 300 peaks (third column).

the first column of Figs. 4–6, separated into 15 slices. Next, we ran our IPV-based method on spectra containing 300 peaks or more (there are 2499, 2461, and 2002 such spectra in the three samples, respectively) to their respective systematic measurement error. These predicted RMEs are plotted in the third column of Figs. 4–6, separated into 15 slices. The true RMEs for this subset of spectra are also plotted, for comparison purposes, in the second column of Figs. 4–6.

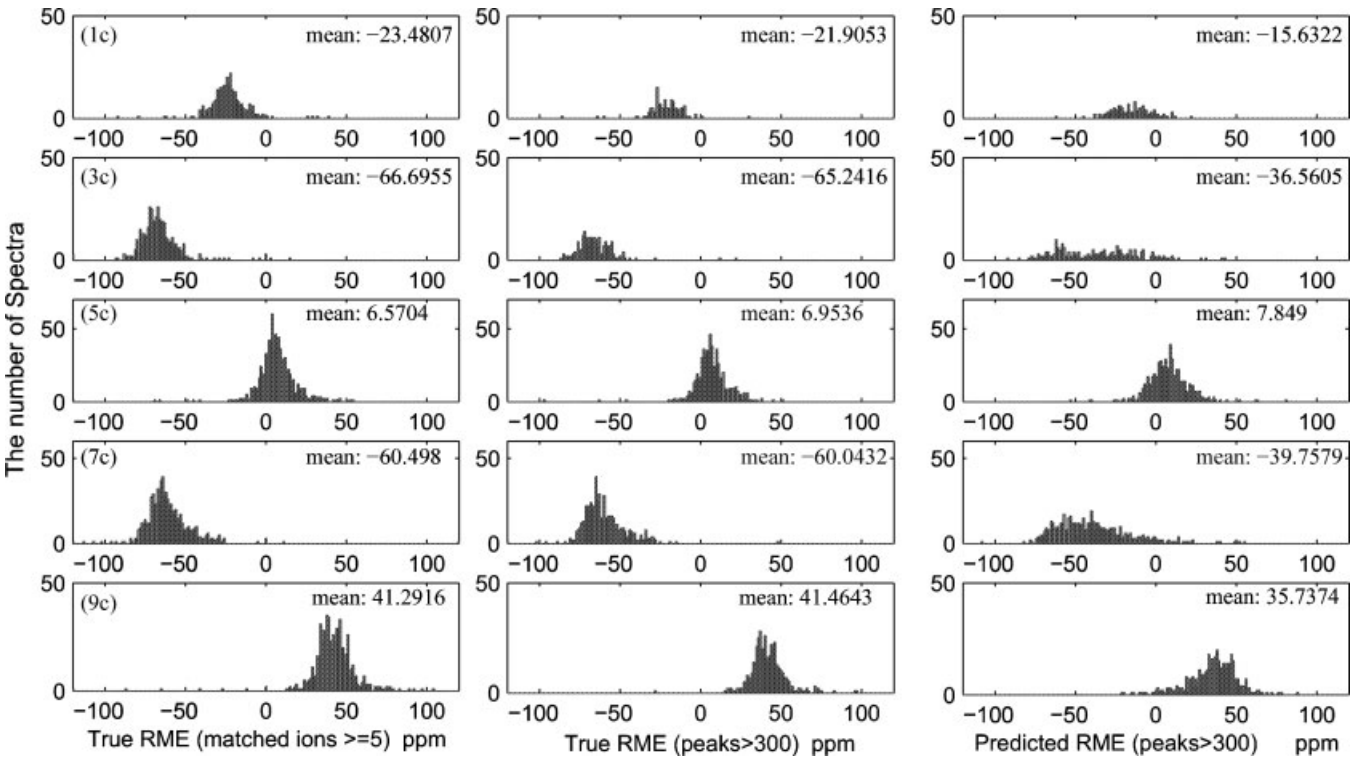

Figure 6. Separated in slices 1c–9c: the true RME distribution in all the spectra (first column), the true RME distribution in the spectra each containing at least 300 peaks (second column), and the distribution of the IPV-based predicted RMEs in the spectra each containing at least 300 peaks (third column).

DOI: 10.1002/rcm

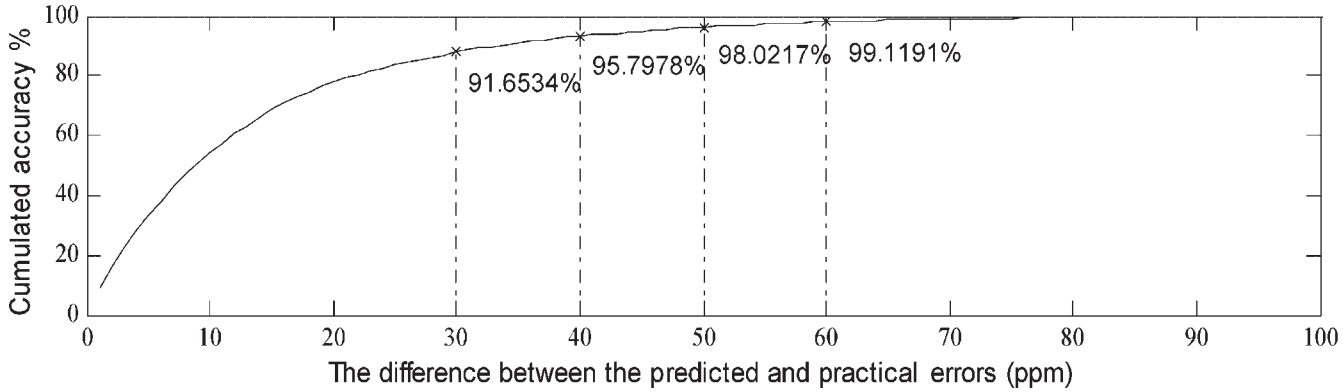

Figure 7. The cumulated accuracies of IPV-based systematic measurement error prediction.

It can be seen from Figs. 4–6 that the true RME distributions in the first two columns are very similar to each other. In fact, the average RME difference over all 15 pairs of mean RMEs is less than 2 ppm. That is, the RME distribution of the spectra with ≥300 peaks would largely reflect the RME distribution of all spectra in dataset A. In this sense, we believe it is good enough if we can accurately predict the RME distribution of the spectra containing ≥300 peaks. Across all 15 pairs of means of the predicted RMEs and the true RMEs for the spectra containing ≥300 peaks, the least difference is 0.25 ppm, the largest is 28.68 ppm, and the average is 10.11 ppm. The correlation coefficients between the three series of 15 mean RMEs are 0.9992, 0.9943, and 0.9936, respectively.

Noticeably in Figs. 4–6, the predicted RMEs in slice 3c are the farthest away from the corresponding true ones (the mean difference reached the largest value of 28.68 ppm). From the original data, we can claim that this set of spectra probably have the lowest quality. One reason is that the average number of peaks in the other 14 slices is 670, while in this slice it is only 473. This suggests that our IPV-based systematic measurement error prediction relies on more peaks to provide sufficient isotopic information. A second reason is the signal-to-noise ratio in slice 3c is much lower compared to the other 14 slices.

The RMEs of individual spectra are more important for the error distribution of a set of spectra. Individual RMEs can be used for data recalibration for subsequent peptide and protein identification. For each spectrum in dataset A containing ≥300 peaks, we calculated the difference between the predicted RME and the true RME. These differences are plotted in Fig. 3 (right). The plot follows a normal distribution $N(-7.9807, 18.7566^2)$ with a normality testing p-value less than 0.01. When separated into the three samples, the three normal distributions are $N(-13.2776, 18.5211^2)$, $N(-2.2252, 13.6991^2)$, and $N(-8.5004, 22.1124^2)$, respectively (all normality testing p-values less than 0.01). Moreover, among these 6925 spectra, as shown in Fig. 7, 91.65% of them have an error difference less than 30 ppm, 95.80% of them have an error difference less than 40 ppm, and 98.02% of them have an error difference less than 50 ppm. This tells that if we use the predicted RMEs to recalibrate the spectra, more than 95% of them will have error distributed in (−40, 40)ppm. Such recalibrated spectra can then be confidently interpreted by Mascot. Compared with the first column of Figs. 4–6 where the true RMEs are distributed in (−90, 80)ppm, our IPV-based systematic measurement error prediction and the subsequent recalibration have a major impact on improving data quality.

As discussed at the end of the last section, as a large number of spectra in dataset A (75.9% of the total) are not confidently interpreted by Mascot even after the IPV-based denoising process, we suspect that their systematic measurement errors might exceed the predefined error ranges for precursors and fragment ions in Mascot search, both of which are set to 0.2 u. These two error distributions are rather independent of each other, and we had no information on precursors in MS/MS spectra. In the experiment, we ran our IPV-based method to predict the systematic measurement errors for these difficult spectra. For each of them, if the recalibrated spectrum can be reliably interpreted by Mascot, we subsequently estimated its measurement error for precursors. We found that some precursor measurement errors are close to 100 ppm. This matched well with our suspicion that the 0.2 u precursor error range in default Mascot search is too small. Consequently, we set up more experiments to either increase the precursor error threshold from 0.2 u to 0.3 u, or use the IPV-based systematic

Table 2. The Mascot peptide and protein identification results on dataset A: default or increased precursor error threshold, without and with the IPV-based spectral data calibration

	$P_{0.2}$		$P_{0.3}$		$Q_{0.2}$		$Q_{0.3}$	
Sample	Peptides/Proteins		Increased Peptides/Proteins		Increased Peptides/Proteins		Increased Peptides/Proteins	
a	3140	503	47	23	64	23	86	24
b	2950	491	18	7	57	56	64	55
c	2640	468	68	21	56	27	123	62

DOI: 10.1002/rcm

measurement error prediction to recalibrate the spectra (the resultant data are called *calibrated data*). Specifically, we compared four sets of search results: (1) keep the precursor error threshold at 0.2 u, and the search result is denoted as $P_{0.2}$; (2) increase the precursor error threshold to 0.3 u, denoted as $P_{0.3}$; (3) keep the precursor error threshold at 0.2 u and apply the IPV-based calibration, denoted as $Q_{0.2}$; and (4) increase the precursor error threshold to 0.3 u and apply the IPV-based calibration, denoted as $Q_{0.3}$. In the three samples, the numbers of peptides and proteins confidently identified in all four experiments are collected in Table 2.

The results in Table 2 show that both increasing the precursor error threshold (to 0.3 u) and the IPV-based data recalibration improved the peptide and protein identification by Mascot search. In particular, the number of identified proteins increased significantly while keeping ~1% false positive rate, possibly because more peptides are confidently identified. More specifically, with a precursor error threshold of 0.3 u but no IPV-based data recalibration, the numbers of identified proteins increased 4.57%, 1.43%, and 4.49%, in the three samples, respectively. Applying the IPV-based data recalibration but keeping the precursor error threshold at 0.2 u, the numbers of identified proteins increased 4.57%, 11.41%, and 5.77%, in the three samples, respectively. Applying the IPV-based data recalibration and the increasing precursor error threshold to 0.3 u, the numbers of interpreted proteins increased 4.77%, 11.20%, and 13.25%, respectively. Comparing these numbers, one can conclude that the IPV-based systematic measurement error prediction and the subsequent data recalibration contribute more to the improvement of protein identification than increasing the precursor error threshold.

Availability

The method described in this paper has been incorporated into the program pQMass. pQMass and its documentation are available to download.[36]

Acknowledgements

JZ and DX are partially supported by the MU-Monsanto Program and an NSF grant NSF/ITRIIS-0407204. GL is partially supported by NSERC. SH is supported by the National Key Basic Research and Development Program (973) under Grant No. 2002CB713807 and the National High Technology Research and Development Program (863) under Grant Nos. 2007AA02Z315 and 2007AA02Z326. We thank Zuofei Yuan and Leheng Wang for assistance in software development.

REFERENCES

1. Yates JR III. *J. Mass Spectrom.* 1998; **33**: 1.
2. Aebersold R, Mann M. *Nature* 2003; **422**: 198.
3. Hernandez P, Müller M, Appel RD. *Mass Spectrom. Rev.* 2006; **25**: 235.
4. Eng JK, McCormack AL, Yates JR III. *J. Am. Soc. Mass Spectrom.* 1994; **5**: 976.
5. Mann M, Wilm M. *Anal. Chem.* 1994; **66**: 4390.
6. Danick V, Addona TA, Clauser KR, Vath JE, Pevzner PA. *J. Comput. Biol.* 1999; **6**: 327.
7. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. *Electrophoresis* 1999; **20**: 3551.
8. Taylor JA, Johnson RS. *Anal. Chem.* 2001; **73**: 2594.
9. Chen T, Kao M-Y, Tepel M, Rush J, Church J. *J. Comput. Biol.* 2001; **8**: 325.
10. Ma B, Zhang KZ, Hendrie C, Liang CZ, Li M, Doherty-Kirby A, Lajoie G. *Rapid Commun. Mass Spectrom.* 2003; **17**: 2337.
11. Fu Y, Yang Q, Sun R, Li D, Zeng R, Ling CX, Gao W. *Bioinformatics* 2004; **20**: 1948.
12. Zubarev R, Mann M. *Mol. Cell. Proteomics* 2007; **6**: 377.
13. Cannataro M, Guzzi PH, Mazza T. Veltri, Preprocessing, management, and analysis of mass spectrometry proteomics data. In *Workflows Management: New Abilities for the Biological Information Overflow – NETTAB 2005*, Naples, 5–7October, 2005.
14. Rejtar T, Chen HS, Andreev V, Moskovets E, Karger BL. *Anal. Chem.* 2004; **76**: 6017.
15. Grossmann J, Roos FF, Cieliebak M, Lipták Z, *et al. J. Proteome Res.* 2005; **4**: 1768.
16. Available: http://www.nitehawk.com/voyager_macros/.
17. Gentzel M, Kocher T, Ponnusamy S, Wilm M. *Proteomics* 2003; **3**: 1597.
18. Senko MW, Beu SC, McLafferty FW. *J. Am. Soc. Mass Spectrom.* 1995; **6**: 52.
19. Senko MW, Beu SC, McLafferty FW. *J. Am. Soc. Mass Spectrom.* 1995; **6**: 229.
20. Lange E, Gropl C, Reinert K, Kohlbacher O, Hildebrandt R. High-accuracy peak picking of proteomics data using wavelet techniques. In *Pacific Symposium on Biocomputing* 2006; **11**: 243–254.
21. Available: http://www.waters.com/WatersDivision/contentd.asp?watersit=EGOO-6QKNZV.
22. Nepomuceno AI, Muddiman DC, Bergen HR 3rd, Craighead JR, Burke MJ, Caskey PE, Allan JA. *Anal. Chem.* 2003; **75**: 3411.
23. Chalmers MJ, Quinn JP, Blakney GT, Emmett MR, *et al. J. Proteome Res.* 2003; **2**: 373.
24. Johnson KL, Mason CJ, Muddiman DC, Eckel JE. *Anal. Chem.* 2004; **76**: 5097.
25. Easterling ML, Mize ML, Amster JI. *Anal. Chem.* 1999; **71**: 624.
26. Moskovets E, Chen HS, Pashkova A, Rejtar T, *et al. Rapid Commun. Mass Spectrom.* 2003; **17**: 2177.
27. Syka JE, Marto JA, Bai DL, Horning S, *et al. J. Proteome Res.* 2004; **3**: 621.
28. Schulze WX, Mann M. *J. Biol. Chem.* 2004; **279**: 10756.
29. Graber A, Juhasz PS, Khainovski N, Parker KC, *et al. Proteomics* 2004; **4**: 474.
30. Lasonder E, Ishihama Y, Andersen JS, Vermunt AM, *et al. Nature* 2002; **419**: 537.
31. Zhang J, Gao W, Cai J, He S, *et al. IEEE/ACM Trans. Comput. Biology Bioinformatics* 2005; **2**: 217.
32. Zhang J, He S, Ling CX, Cao X, *et al. Rapid Commun. Mass Spectrom.* 2008; **22**: 1203.
33. Available: http://gygi.med.harvard.edu/pubs.
34. Elias JE, Hass W, Faherty BK, Gygi SP. *Nat. Methods* 2005; **2**: 667.
35. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. *Anal. Chem.* 2002; **74**: 5383.
36. Available: http://pfind.ict.ac.cn/pQMass.htm.

DOI: 10.1002/rcm