RAPID COMMUNICATIONS IN MASS SPECTROMETRY
Rapid Commun. Mass Spectrom. 2008; **22**: 1203–1212
Published online in Wiley InterScience (www.interscience.wiley.com) DOI: 10.1002/rcm.3488

PeakSelect: preprocessing tandem mass spectra for better peptide identification

Jingfen Zhang[1]*, Simin He[1], Charles X. Ling[2], Xingjun Cao[3], Rong Zeng[3] and Wen Gao[1]*

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China
[2]Department of Computer Science, University of Western Ontario, London, Ontario, Canada
[3]Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Science, Chinese Academy of Sciences, Shanghai 200031, China

Received 29 June 2007; Revised 13 December 2007; Accepted 3 February 2008

We present a new preprocessing method, *PeakSelect*, to improve the accuracy and efficiency of Tandem Mass-Spec peptide (protein) identification. The fundamental difference between noise and fragment ions in spectra is that ions have isotopes but noise does not. We propose a new and important concept of an Isotope Pattern Vector (IPV) which characterizes the isotope cluster of fragment ions. Then the noise and real peaks can be distinguished by the quantitative IPV values. *PeakSelect* first uses a new method of the Gaussian Mixture Model and Expectation-Maximization (EM) algorithm to find the base intensity level (baseline) in a spectrum. Then *PeakSelect* selects features based on the IPV and baseline, and constructs a decision tree to automatically classify the peaks into different categories such as noise, single ion peaks, and overlapping peaks. Experiments show that *PeakSelect* can help to reduce the Mascot searching time and increase the reliability of peptide identifications. In particular, *PeakSelect* performs well on complex spectra with a large number of peaks from large peptides, and supports more sequence identification than other well-known systems. Copyright © 2008 John Wiley & Sons, Ltd.

Mass spectrometric analysis and database search has been a well-known tool for peptide and protein identification.[1,2] During experiments, the peptides separated by liquid chromatography are fragmented and ionized by collision-induced dissociation (CID)[3] and the ions are measured by mass spectrometry for mass/charge ratios (m/z). Consequently, the peptides are identified (or sequenced) by these m/z values of ions in the tandem mass spectrum with a sequence database search.

Generally, a good quadrupole time-of-flight (Q-TOF)[4] spectrum of a peptide has 1000 to 5000 or more peaks, but only 1–5% of these peaks are 'real peaks' while the others are peaks corresponding to noise or the isotopes of fragment ions. Here, 'real peaks' are the monoisotopic peaks corresponding to the fragment ions in tandem mass spectra, such as the *b*-, *y*-, and *a*-types of fragment ions.[5–8] The numerous noise and isotopic peaks in tandem mass spectra can lead to a heavy computational cost in database search. Furthermore, the noise can cause either false negative or false positive peptide identifications since they may match with the theoretical ions of an irrelevant peptide sequence. To increase the accuracy of peptide identification and decrease the computation complexity, a preprocessing of tandem mass spectra should be introduced to distinguish the real peaks from noise and isotopic peaks in the spectrum before database searching.

The preprocessing has two major purposes: denoising and deconvolution of isotopic peaks (or deisotoping). The difficulties in the preprocessing include: (1) the quality of spectra is totally different. For example, the distributions of noise peaks in spectra are significantly different; (2) the intensity of many important ions (e.g., *b*-series and *y*-series ions) is very low, which is confounding with the noise peaks in intensity; (3) the convolution of isotopic peaks is complex, which makes it more difficult to distinguish the individual monoisotopic ions.

To date several methods have been proposed for the preprocessing of tandem mass data, including threshold filtering, deisotoping and denoise transforming. The threshold filtering is the most straightforward approach. As peaks with very small abundance values are unlikely to be real peaks, threshold filtering methods select peaks above a given threshold,[9] a specific number of the most intensive peaks in the specified m/z intervals,[10] or peaks above a computed intensity baseline.[11] However, as the abundance is not the fundamental attribute of real peaks, the filtering method cannot thoroughly remove the noise just depending on thresholds. The deisotoping methods[12–16] first calculate the theoretical isotopic pattern of an assumed elemental composition such as $n^*(C_6H_5NO)$[13] or $n^*(C_{4.9384}H_{7.7583}N_{1.3577}O_{1.4773}S_{0.0417})$,[15] and then the deviation from the actual data and the theoretical data will yield a hidden peak. Although these

*Correspondence to: J. Zhang and W. Gao, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China.
E-mail: zhangjingf@missouri.edu; wgao@ict.ac.cn
Contract/grant sponsor: National Key Basic Research and Development Program (973) of China; contract/grant number: 2002CB713807.
Contract/grant sponsor: CAS Knowledge Innovation Program, National High Technology Research and Development Program (863) of China; contract/grant number: 2007AA02Z315 and 2007AA02Z326.

methods can identify some isotope peaks and overlapping cases, the assumed elemental composition used to collapse the isotope pattern is too coarse to identify the complex convolution of isotopes. Consequently, this method would inevitably lead to loss of some important fragment ions. In the denoising mechanism, some well-known procedures such as wavelet transformation have been utilized to denoise the raw tandem mass spectrum.[12,17] However, the parameters such as the wavelet base functions, order, and level of decomposition would impact the potential spectra distortion. Nowadays, the preprocessing is still a challenging problem. Some commercial software such as ProteinLynx™ Global *Server* [18] also supplies preprocessing functionalities. Later, we will show that *PeakSelect* coupled with *Mascot* [19] performs better than *ProteinLynx* coupled with *Mascot*.

Real peaks in spectra differ from noise peaks in many aspects. At first, the fundamental difference is that ions have isotopes but noise does not. The theoretical isotope pattern of an ion is decided by its atomic component.[20,21] Thus, we can search a real peak by scanning its neighboring peaks and checking whether potential isotopic peaks exist and whether the experimental isotope pattern matches with the corresponding theoretical one. Secondly, most of the noise peaks are randomly produced by the mass spectrometer during CID and, as a result, the noise intensity roughly follows the normal distribution. Hence, real peaks can be distinguished from noise peaks by calculating the distribution of noise peaks. Finally, the features of the different convolutions should be analyzed to identify the peaks of individual monoisotopic ions involving convolutions of isotopic peaks.

Based on the above discussion, we present a new solution, *PeakSelect*, in this paper for mass spectra produced by quadrupole time-of-flight hybrid mass spectrometers, such as the QSTAR® XL Hybrid LC/MS/MS system[22] and the Q-Tof Ultima Global.[23] In contrast to the threshold filtering and denoise transforming, we use the Gaussian Mixture Model (GMM) to estimate the base intensity level of noise peaks (or baseline) and treat the baseline as one feature to distinguish noise and real peaks. Instead of the method of assumed elemental composition, we propose a key concept of an Isotope Pattern Vector (IPV) to characterize the isotope cluster of a fragment ion universally. In addition, we investigate the cases of overlapping isotope peaks before deisotoping. Then, we study the difference between noise, single fragment ions and overlapping ions to construct a decision tree to distinguish the peaks.

We apply *PeakSelect* on different datasets. The experimental results show that *PeakSelect* helps to reduce the Mascot searching time and increase the number of interpreted peptides and proteins at the same time. In addition, *PeakSelect* outperforms the preprocessing of ProteinLynx™ Global *Server* (version 2.0.5) by improving the sensitivity of Mascot searches.

The rest of this paper is organized as follows. In the following section we explain our algorithm in more detail. Then, in the next section we describe the datasets and demonstrate the experimental investigations. Finally, in the Conclusions, we discuss further developments.

METHOD

Our solution has three new contributions. At first, a key concept of an Isotope Pattern Vector (IPV) is proposed to digitally characterize the isotope cluster of a fragment ion universally. Thus the noise and real peaks can be distinguished by the quantitative IPV value. Secondly, a new method based on the Gaussian Mixture Model (GMM) and an Expectation-Maximization (EM) algorithm is used to find the base intensity level of noise peaks in spectra. Finally, after selecting the possible features based on the IPV and investigating the complex overlapping of isotope peaks, a decision tree is constructed to classify the peaks into different categories such as noise, single ion peaks and overlapping peaks. Therefore, all the potential monoisotopic masses of ions can be calculated.

Isotope pattern vector

As we know, each fragment ion has theoretical isotopes while noise does not. Hence, the concept of the isotope pattern vector (denoted as IPV) can not only distinguish the noise and real peaks, but can also describe the profile of the isotopes of an ion. Suppose that the monoisotopic mass of a fragment ion P (with molecular formula $C_{n1}H_{n2}N_{n3}O_{n4}S_{n5}$) is M, and its first four isotopes (i.e., with one, two, three and four extra neutrons, respectively) are P_1, P_2, P_3 and P_4. We define the isotope pattern vector of P as $IPV = (M, T_1, T_2, T_3, T_4, \Delta m_1, \Delta m_2, \Delta m_3, \Delta m_4)$, where T_k is the relative abundance of P_k with respect to P, and Δm_k are the mass differencees between P_k and P, for $k = 1$–4, respectively.

The theoretical *IPV* (denoted as *tIPV*) of a fragment ion can be deduced from its elemental component and the probability of the isotopes of each element. For simplicity, we just show T_1, T_2, Δm_1, Δm_2 of *tIPV* for a given formula $C_{n1}H_{n2}N_{n3}O_{n4}S_{n5}$ as follows:

$$M = (12, 1.0078, 14.0030, 15, 9972, 31, 9721) \times (n_1, n_2, n_3, n_4, n_5)^T \quad (1)$$

$$T_1 = n_1 q_C + n_2 q_H + n_3 q_N + n_4 q_{O1} + n_5 q_{S1}, \quad (2)$$

$$T_2 = n_4 q_{O2} + n_5 q_{S2} + \frac{1}{2} T_1^2 - \frac{1}{2}(n_1 q_C^2 + n_2 q_H^2 + n_3 q_N^2 + n_4 q_{O1}^2 + n_5 q_{S1}^2), \quad (3)$$

$$\Delta m_1 = (n_1 q_C \Delta C + n_2 q_H \Delta H + n_3 q_N \Delta N + n_4 q_{O1} \Delta O_1 + n_5 q_{S1} \Delta S_1)/T_1 \quad (4)$$

$$\begin{aligned} \Delta m_2 = {} & n_4 q_{O2} \Delta O_2 + n_5 q_{S2} \Delta S_2 + \frac{1}{2}(n_1 q_C \Delta C + n_2 q_H \Delta H \\ & + n_3 q_N \Delta N + n_4 q_{O1} \Delta O_1 + n_5 q_{S1} \Delta S_1)^2 \\ & - \frac{1}{2}(n_1 q_C^2 \Delta C + n_2 q_H^2 \Delta H + n_3 q_N^2 \Delta N + n_4 q_{O1}^2 \Delta O_1 \\ & + n_5 q_{S1}^2 \Delta S_1)/T_2 \end{aligned} \quad (5)$$

where q_C, q_H, q_N are relative abundances of ^{13}C to ^{12}C, D to H, and ^{15}N to ^{14}N, and q_{O1}, q_{O2} (q_{S1}, q_{S2}) are the ratios of ^{17}O to ^{16}O, ^{18}O to ^{16}O (^{33}S to ^{32}S, ^{34}S to ^{32}S), respectively. ΔC, ΔH, ΔN, ΔO_1, ΔO_2, ΔS_1, ΔS_2 are the corresponding mass differences between the monoisotope and isotopes.

DOI: 10.1002/rcm

The experimental isotope pattern (denoted as *eIPV*) of a fragment ion P can be calculated if the isotope ions are detected by a mass spectrometer. We characterize a peak in a mass spectrum in terms of (m/z, *intensity*), where *intensity* is the relative height of the peak. For a cluster of peaks (p_0, p_1, p_2, p_3, p_4) with the (m/z, *intensity*) pairs (Mz_k I_k), $k = 0$–4, the charge z can be calculated from the interval between Mz_k. For $k > 1$, the value of (Mz_k I_k) will be substituted by zero if p_k does not exist. After normalizing $z = 1$, the (m/z, *intensity*) pairs are converted into (M_k, I_k), where $M_k = Mz_k{}^*z - (z-1)^*1.0078$, $k = 0$–4, respectively. Then *eIPV* can be obtained by:

$$\begin{aligned} eIPV &= (M_0, R_1, R_2, R_3, R_4, \Delta m_1, \Delta m_2, \Delta m_3, \Delta m_4) \\ &= (M_0, \frac{I_1}{I_0}, \frac{I_2}{I_0}, \frac{I_3}{I_0}, \frac{I_4}{I_0}, M_1 - M_0, M_2 \\ &\quad - M_0, M_3 - M_0, M_4 - M_0) \end{aligned} \tag{6}$$

Considering the measure error of the mass spectrometer, the isotope peaks of a fragment ion should be observed and the experimental isotope pattern should match its theoretical isotope pattern.

Baseline identification

Most noise peaks are randomly produced by the mass spectrometer during CID. Generally, each mass spectrum exhibits a base intensity level of noise peaks (baseline) which varies across the m/z axis with different fractions. For example, the spectra of two peptides FTQKIFGGQNNSK and KSLLLSQILHK are shown in Figs. 1(a) and 1(b), respectively. It can be seen that the baseline in Fig. 1(a) is much higher than that in Fig. 1(b).

Intensity is one important factor to distinguish noise and real peaks. However, due to the variety of the baselines in different spectra and if there are very low *b*- and *y*-series ions in the spectra, the threshold filtering strategy cannot remove the noise without losing important real peaks. Here, we propose a more accurate method to identify the baseline. In fact, we can divide peaks into three classes: (a) low noise, in which peaks are almost noise and distribute uniformly around the m/z axis, (b) high real peaks, and (c) a mixture of high noise and low real peaks. Therefore, we utilize two baselines to divide these three classes of peaks: One is global baseline which depicts the up-bound of low noise, and the other is local baseline which is the low-bound of high real peaks.

Practically, we apply a Gaussian mixture model (GMM) in which the components represent the above three classes of peaks. We use the mean and standard deviation to characterize the base level of intensity, and calculate two kinds of baselines of global baseline and local baseline, noted as $I_{\text{baseline}} = (GI_{\text{mean}}, GI_{\text{deviation}}, LI_{\text{mean}}, LI_{\text{deviation}})$. The value of I_{baseline} is obtained by an Expectation-Maximization (EM) algorithm to estimate the parameters of the GMM. Note that we use the relative intensities of peaks in the spectrum. For example, Fig. 2 shows the total low peaks and two classes of low peaks in the spectra of FTQKIFGGQNNSK and KSLLLSQILHK, respectively. The calculated results of I_{baseline} are (1.458, 0.5903, 3.738, 0.986) and (0.611, 0.398, 2.397, 0.478), respectively, which are consistent with the observation of the noise in the spectrum.

Overlapping cases

Normally, an important ion, such as *b*-, *y*-, *a*-ions with high intensity in spectra, always has more than one isotope peak with a $(1/z)$ u interval in m/z value ($z = 1$–4) if no overlapping ions exist in its vicinity. In this case, the corresponding *eIPV* matches perfectly with the *tIPV*. Due to the complex overlapping of isotope peaks, it is difficult to distinguish the individual monoisotopes of ions. However, the overlapping ions cannot only be distinguished from noise peaks, but can also be split by the match score of *eIPV* and *tIPV*. We have investigated the isotope profiles in spectra and summarized some predominant types of overlapping patterns.

The most important overlapping is that the isotopic peaks of two ions with 1 u mass interval and with same charge overlap each other. The two ions are always the water-loss and ammonia-loss ions of an important ion. This overlapping pattern is shown in Fig. 3(a). In this case, the value of R_1 in the $eIPV = (M_e, R_1, R_2 \ldots)$ calculated from (p_0, p_1, p_2, $p_3 \ldots$) is far greater than the value of T_1 in the corresponding $tIPV = (M, T_1, T_2, \ldots)$. Similarly, some simple but important overlapping is that isotopic peaks of two ions with 3 u mass interval overlap each other and the pattern is shown in Fig. 3(b). In this case, the values of R_1 and R_2 in the *eIPV* calculated from (p_0, p_1, p_2, $p_3 \ldots$...) can match well with the values of T_1 and T_2 in *tIPV* while R_3 is far larger than T_3.

The most complex pattern involves different charge states and noise peaks. Because of different noise baselines and different m/z values of peaks, the same profile of peaks will correspond to different ion overlapping cases. For example, in Fig. 4(a), there is a singly charged Ion1 that overlaps with

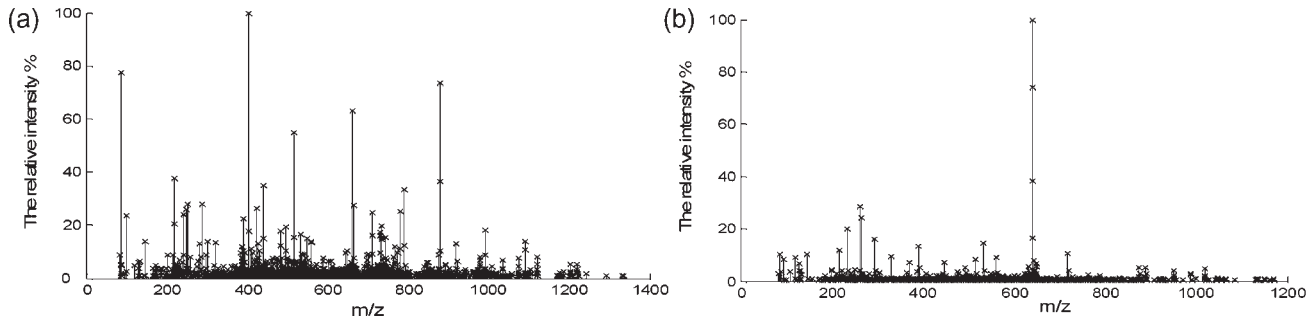

Figure 1. Tandem spectra of the peptides FTQKIFGGQNNSK (a) and KSLLLSQILHK (b). The baseline in (a) is much higher than that in (b).

DOI: 10.1002/rcm

RCM

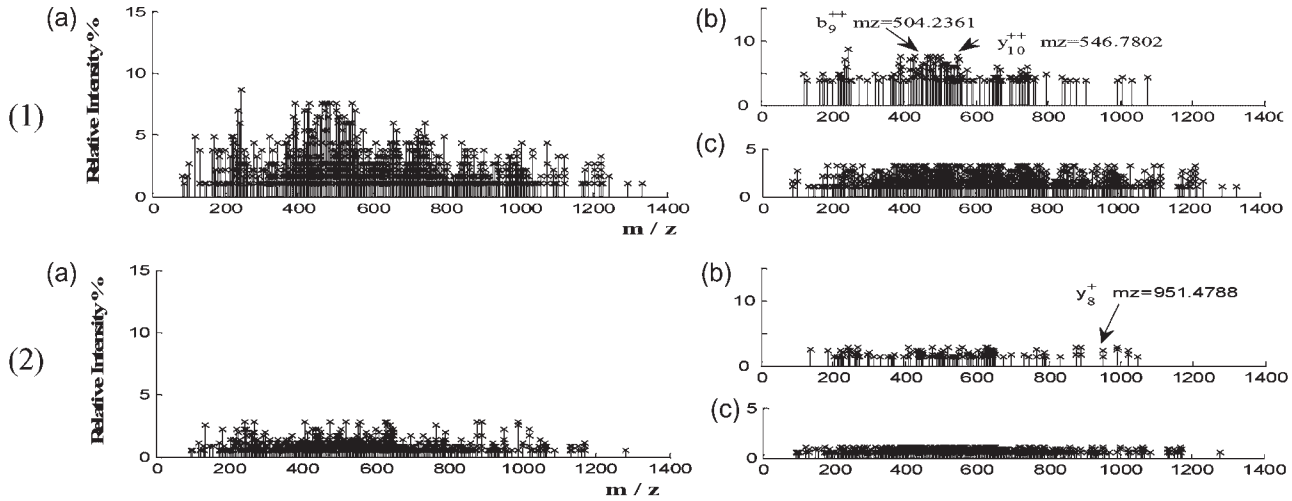

Figure 2. The low peaks in the spectrum of peptide FTQKIFGGQNNSK (1) and KSLLLSQILHK (2), in which (a) depicts the total peaks while (b) shows some real peaks such as b_9^{++}, y_{10}^{++} and y_8^+ which confused with high noise peaks and (c) shows the low noise peaks.

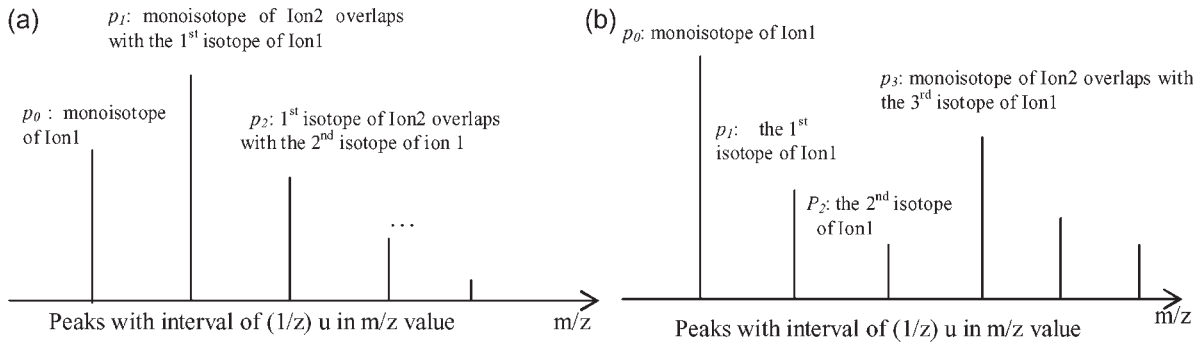

Figure 3. Profiles of the overlapping cases in which there are two ions with same charges and with 1 u and 3 u intervals in mass, shown in (a) and (b), respectively.

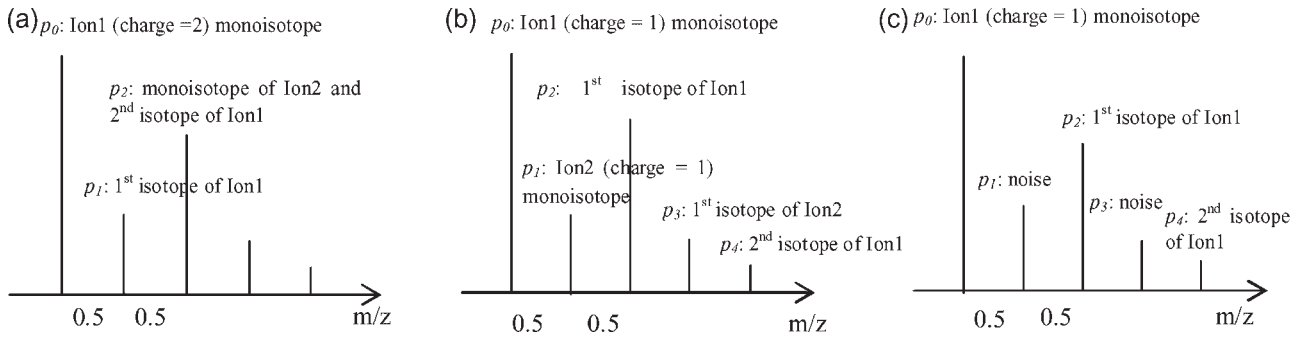

Figure 4. The same profile of peaks in intensity corresponds to three different ion overlapping cases. In (a), a singly charged Ion1 overlaps with another Ion2. In (b), it is not overlapping but two singly charged ions have a mass interval of 0.5 u. In (c), there is only a singly charged ion with some noise peaks.

another Ion2 and the charge of Ion2 can be identified by its following peaks. Here, a $eIPV = (M_e, R_1, R_2 \ldots)$ is calculated from $(p_0, p_1, p_2, \ldots)$ and the corresponding $tIPV = (M, T_1, T_2, \ldots)$, where R_1 matches well with T_1 but R_2 is far greater than T_2; in Fig. 4(b), no overlapping is found but two singly charged ions have 0.5 u mass interval as two perfect matches exist: one is between the *eIPV* and *tIPV* of (p_0, p_2, p_4) and the other is between the *eIPV* and *tIPV* of $(p_1, p_3, \ldots)$. In Fig. 4(c), there is only a singly charged ion (p_0, p_2, p_4) with some noise peaks because p_1 and p_3 are under the global baseline. Although p_4 is under the global baseline too, it can still be identified as the second isotope of p_0 from the value of *eIPV*.

DOI: 10.1002/rcm

Therefore, we use the above overlapping cases and try to find the rules from the point of view of the *IPV* to identify the monoisotopes of individual ions.

Feature selection

In this section, we investigate the difference between noise and fragment ions based on some selected features. The purpose is to construct a decision tree to classify the peaks based on the value of these features.

Distance between the peak's intensity and baseline

As mentioned in the section entitled 'Baseline identification' above, the intensity (represents the relative abundance) is an important factor to distinguish noise and real peaks. Peaks above *Local Baseline* are likely to be real peaks while peaks under *Global Baseline* are most likely to be noise and the peaks between *Local* and *Global Baseline* may be either noise or real peaks. We do not intend to exactly filter noise peaks by the baseline thresholds but to utilize the distance between a peak's intensity and baseline as one important feature to decide whether the peak is noise. Consider a peak with intensity I_{peak} and the global baseline and local baseline of $I_{\text{baseline}} = (GI_{\text{mean}}, GI_{\text{deviation}}, LI_{\text{mean}}, LI_{\text{deviation}})$, the value of F_{RA1} and F_{RA2} are treated as the first kind of feature and are calculated as follows:

$$F_{RA1} = (I_{\text{peak}} - GI_{\text{mean}})/GI_{\text{deviation}} \tag{7}$$

$$F_{RA2} = (I_{\text{peak}} - LI_{\text{mean}})/LI_{\text{deviation}} \tag{8}$$

Mass residue

There is a mass residue between the weight and nominal mass of an atom. For example, the nominal masses of C, H, N, O, S are (12, 1, 14, 16, 32) and the mass residues are (0, 0.007825, 0.00307, −0.00509, −0.02793), respectively. Since peptides are composed of C, H, N, O, S atoms, the mass residue of a peptide is subjected to its nominal mass. For each nominal mass, the range of its residue can be obtained by calculating all the theoretical fragment ions produced by trypic peptides from proteins of SWISS-PROT.

For a given ion with mass of M, the nominal mass $Nomi_M$ should be $Integ_M$, e.g., the integral part of M, or $Integ_M - 1$, and the mass residue Res_M will be $M - Integ_M$ or $M - Integ_M + 1$. According to the statistical results from proteins of SWISS-PROT, we obtain the residue range of $[Res_1, Res_2]$ for each integer. Since ions with one to five charges co-exist in a spectrum, the charge state of ions has to be recognized to determine the monoisotopic mass. For a peak with an m/z value of Mz, we calculate the corresponding $Nomi_M$ and Res_M supposing that it corresponds to an ion with z charges ($z = 1$–5). If the calculated Res_M is far from the theoretical residue range of $Nomi_M$, then the peak cannot correspond to a z-charged ion. Therefore, we use the following feature F_{Res}:

$$F_{\text{Res}} = \frac{sign[(\text{Res}_M - \text{Res}_1) * (\text{Res}_M - \text{Res}_2)] * min\{|\text{Res}_M - Res_1|, |\text{Res}_M - Res_2|\}}{|\text{Res}_2 - \text{Res}_1|} \tag{9}$$

which characterizes the distance between Res_M and the theoretical residue mass range $[Res_1, Res_2]$ of $Nomi_M$, where $sign(x) = 1$ if $x > 0$ else $sign(x) = 0$.

Distance between eIPV and tIPV

As discussed in the section entitled 'Isotope pattern vector' above, the isotope peaks of a fragment ion should be observed and the experimental isotope pattern should match its theoretical isotope pattern. In other words, for a given peak p_0, we can find its isotopic peaks and calculate the distance between its *eIPV* and *tIPV*. Then, these values of *IPV* are treated as another important kind of feature.

The elemental component of a fragment ion is unknown during the preprocessing and the exact value of *tIPV* cannot be calculated. However, the theoretical *tIPV* of an ion can be estimated by the expected (or mean) value.[20] For example, Fig. 5 shows the *IPV* values corresponding to some peaks. More detailed, for a group of peaks (646.2718, 647.2669, 648.2671), the minimal, mean and maximal values of T_1 in *tIPV* are $(T_{1min}, T_{1mean}, T_{1max}) = (0.3473, 0.3644, 0.3833)$ and the corresponding values of T_2 are $(T_{2min}, T_{2mean}, T_{2max}) = (0.0807, 0.0854, 0.1756)$, respectively.

The practical values (R_1, R_2) of *eIPV* are (0.4163, 0.2490) and they are coincident with the *tIPV*. Therefore, it can be accepted as a valid isotope group. However, another potential isotopic group of peaks starting from 646.2718 is (646.2718, 646.7721, 647.2669, 647.7984), and the calculated theoretical values of T_1 are (0.7002, 0.7283, 0.7594) and values of T_2 are (0.2794, 0.2967, 0.4013), but the (R_1, R_2) of *eIPV* are (0.1392, 0.2490), which are far from the theoretical range. Therefore, it can be considered as invalid. The same cases are for the other two groups starting from peak 648.7895.

In *PeakSelect*, we calculate the values of F_{P1} and F_{P2} as follows to characterize the distance between the practical and theoretical relative abundance in *IPV*, and treat them as important features to select valid peaks.

$$F_{P1} = \frac{sign[(R_1 - T_{1min}) * (R_1 - T_{1max})] * min\{|R_1 - T_{1min}|, |R_1 - T_{1max}|\}}{T_{1\text{mean}}}, \tag{10}$$

$$F_{P2} = \frac{sign[(R_2 - T_{2min}) * (R_2 - T_{2max})] * min\{|R_2 - T_{2min}|, |R_2 - T_{2max}|\}}{T_{2mean}} \tag{11}$$

Other features

Some other characters such as the assigned charge to peaks of a potential ion, the number of isotope peaks in the potential

DOI: 10.1002/rcm

RCM

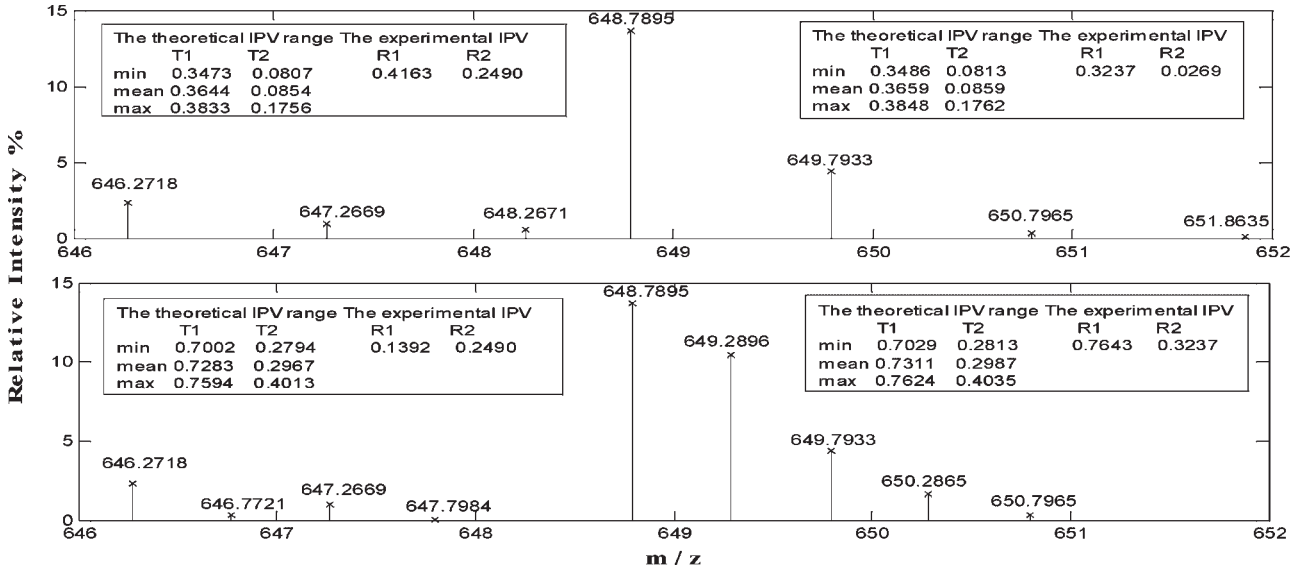

Figure 5. Potential isotopic peaks with different charges ($z = 1$ or 2). The numbers within the pane are the calculated values of *eIPV* and *tIPV*. The real peaks with correct charge state can be distinguished by the distance between the values of *eIPV* and *tIPV*.

isotope cluster, and the distance between the theoretical Δms of the isotopes and the measured Δms in *IPV*, etc., are also important to judge the validation and overlapping of peaks. For example, when the charge state of an ion is 2 or 3, there are always more than two isotope peaks of the ion existing in the spectrum while less than three isotope peaks exist when the charge state of an ion is 1. We use F_{charge}, F_{IsoNum}, F_{Merr1}, F_{Merr2} to represent these characters.

Classification and decision tree

We select some peaks corresponding to noise, a single ion and overlapped ions as training samples to verify the difference of the selected feature values, as described in the section entitled 'Feature selection' above. Specifically, in order to select training samples, we at first judge whether a peak is noise, or corresponds to an ion, or involves overlapped ions when the peptide sequence corresponding to the spectrum is known. Then, we investigate the difference in the feature values, and learn the rules from the training samples. Finally, we construct a decision tree (learnt by WEKA C4.5 toolbox) to classify the peaks into three classes: class 1, noise; class 2, real peaks corresponding to a single ion; and class 3, real peaks involving overlapping which includes five overlapping cases described in the section entitled 'Overlapping cases' above.

According to the rules of the decision tree, all of the peaks in a spectrum can be classified by the calculated values of features. Note that each peak will be classified into one and only one class. Since there would be different potential isotope clusters under a different charge z, a given peak p_0 is judged as noise if all of the potential isotope clusters are classified into class 1. For a given peak p_0 (with m/z value as Mz), if the potential isotope cluster under charge z is classified into class 2, then the monoisotopic mass $M = Mz^*z-(z-1)^*1.0078$ is selected to present a potential fragment ion. Furthermore, if peak p_0 is classified into class 3, there may be two monoisotopic masses obtained according to the overlapping cases. Finally, the masses corresponding to peaks which have been classified into class 2 and class 3 are selected prior to the database searching.

EXPERIMENTAL INVESTIGATIONS

In this section, we evaluate *PeakSelect* with the metrics of identification accuracy and search speed of *Mascot* search on the data after the process of *PeakSelect*. At first, we compare the performance of *PeakSelect* with the existing software ProteinLynx^{TM} Global *Server* version 2.0.5 (denoted as *ProteinLynx* for simplicity). Then we evaluate the performance of *PeakSelect* by applying it in a large-scale analysis of a yeast whole-cell lysate.

Comparing *PeakSelect* with *ProteinLynx*

In this section, we compare *PeakSelect* with *ProteinLynx* by their ability to improve the identification accuracy. The dataset includes spectra produced by a Q-Tof Ultima Global mass spectrometer from a tryptic digestion of peptides of eight proteins. They are 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), respectively.

At first, all data were converted from raw instrument output into the .dta format using software *ProteinLynx* without noise reduction and deisotoping. We denote these as raw .dta data. Secondly, the raw .dta data were preprocessed by *PeakSelect* and *ProteinLynx*. We used two groups of processing parameters in *ProteinLynx*, one is the default value and the other is selected manually so that the number of selected peaks by *ProteinLynx* is close to that selected by *PeakSelect*. The better search result from the two groups of parameters is selected as the comparing candidate of *ProteinLynx*.

DOI: 10.1002/rcm

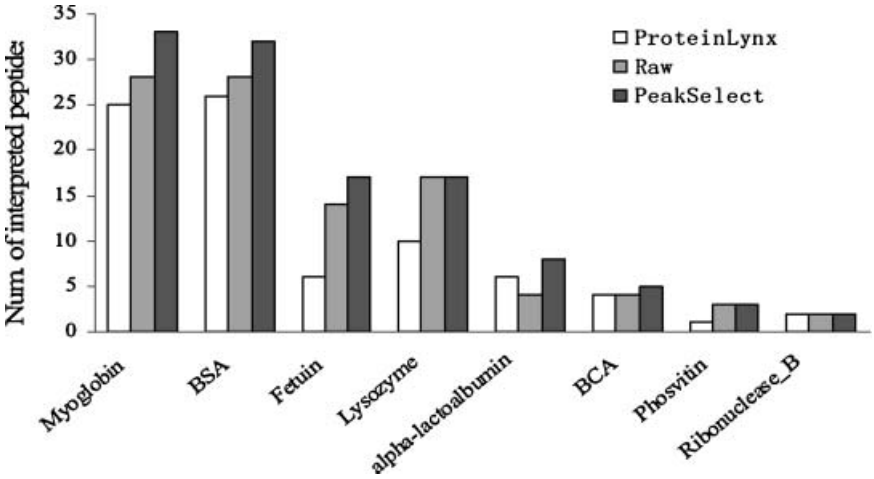

Figure 6. The *Mascot* search results of eight proteins on three kinds of spectra data: preprocessed by *ProteinLynx*, raw, and preprocessed by *PeakSelect*, respectively.

The on-line *Mascot* is chosen as the search engine to interpret these spectra, in which all cysteine residues were searched as carboxamidomethyl-cysteine and methionine residues were allowed to be oxidized. Up to 1 internal cleavage site was allowed for tryptic searches in the SWISS-PROT database. Parameters set for searches include use of monoisotopic atomic masses and a tolerance of 100 ppm for precursor and 0.2 Da for fragment ions. Here, one spectrum is treated as interpreted by *Mascot* if the *Mascot* score is no less than 20 and the expectation value less than 0.05.

As we know, the more the interpreted spectra, the higher the reliability of the identification. Therefore, we use the number of interpreted spectra to evaluate the performance. The search results are shown in Fig. 6. There are a total of 80, 117 and 100 interpreted peptides of the eight proteins from the three datasets: preprocessed by *ProteinLynx*, *PeakSelect* and raw .dta data, respectively. In other words, there is an average of 46% and 17% increased interpreted spectra in data preprocessed by *PeakSelect* than those by *ProteinLynx*. In fact, the searches on the data preprocessed by *PeakSelect* are more reliable since the quality of the score and expectation value (data have not shown here) is better than that on data of raw and preprocessed by *ProteinLynx*.

Performance of *PeakSelect* on large-scale yeast whole-cell lysate data

In this section, we investigate the performance of *PeakSelect* on large-scale data. The dataset can be downloaded from the Internet,[24] which includes 46195 .dta files produced after analyzing five trypsin-digested gel regions representative of the yeast proteome in triplicate by nanoscale microcapillary LC/MS/MS using QSTAR mass spectrometers.[25] Mascot version 2.1.02 is selected to interpret the downloaded data (we denote it as raw data in the rest of the paper) and the preprocessed data by *PeakSelect*.

According to the methods described by Elias *et al*.,[25] we use searches against a composite target-decoy database containing all yeast protein sequences in both forward and reverse orientations to estimate the false positive rate of peptide-spectral matches (or say PSMs). All of the search parameters are same as that in Elias *et al*.[25] In addition, we choose the same score filter criteria as described in Supplementary Table 1 in Elias *et al*.[25] to achieve around 99% precision or 1% false positive rate.

The *Mascot* searching time is decreased to 1/2 to 1/4 on the three samples after the preprocessing of *PeakSelect*. Tables 1 and 2 depict the number of interpreted tandem mass spectra, peptides and proteins from the raw and preprocessed data.

Table 1. Numbers of interpreted tandem spectra identified by *Mascot* from the raw and preprocessed data

	Raw .dta				Preprocessed .dta by *PeakSelect*			
	Selected[a]	FP[ab]	TP[ab]	Precision[ab]	Selected[a]	FP[ab]	TP[ab]	Precision[ab]
Sample 1	3,167	24	3,142	99.24%	3,557	44	3,513	98.32%
Sample 2	3,011	32	2,979	98.93%	3,373	42	3,331	98.75%
Sample 3	2,798	52	2,746	98.14%	3,000	52	2,948	98.27%
Mean	2,992	36	2,956	98.80%	3,310	46	3,264	98.61%

[a]To be selected, PSMs had to receive scores greater than or equal to those indicated in Supplementary Table 1 in Elias et al.[25]
[b]FP, estimated false positive identifications, and calculated by doubling the number of decoy hits; TP, estimated true positive identifications; Precision, TP/(TP+FP).

DOI: 10.1002/rcm

Table 2. Numbers of peptides and proteins identified from the raw and preprocessed data

	Peptides[a]				Proteins[a]			
	Raw-data	Preprocessed	Overlapped	Union[b]	Raw-data	Preprocessed	Overlapped	Union[b]
Sample 1	2777	3140	2469	3448	461	503	423	541
Sample 2	2597	2950	2304	3243	456	491	409	538
Sample 3	2446	2640	2083	3003	453	468	402	519
Mean	2607	2910	2285	3232	457	487	411	533

[a]The interpreted peptides and proteins are selected from the target hits but not from the decoy hits.
[b]The number of identified peptides and proteins combined from the raw and preprocessed data.

From the data in the columns 'Selected' and 'Precision' in Table 1, we can see that under the similar around 1% false positive threshold, there are 12.31%, 12.02% and 7.22% increased numbers of the interpreted spectra in three samples in the preprocessed data. Consequently, both protein and proteome coverage is improved after applying *PeakSelect*. For example, from the data in the columns 'Raw data' and 'Preprocessed' in Table 2, we can calculate that there are an average 11.64% and 6.56% increased numbers of interpreted peptides and proteins (selected from the target hits) after applying *PeakSelect*. The results show the great improvement of *PeakSelect* over the raw data.

On average, 2597 tandem spectra were confidently assigned by *Mascot* from both the raw and preprocessed data. However, *Mascot* scores on the preprocessed data are better than those on raw data. Of the 2792, 2634, and 2365 confidently co-assigned (or overlapped) spectra from three samples, the *Mascot* score on the preprocessed spectra are increased by 15.85% (i.e., (41.45 − 35.78)/35.78), 16.89% (i.e., (41.24 − 35.28)/35.28), and 13.04% (i.e., (38.98 − 34.89)/34.89) than those on the raw data. This shows the benefit of preprocessing. The detailed distribution of *Mascot* scores is depicted in Fig. 7.

There are 375, 377, and 433 interpreted spectra only from raw data of three samples, in which there are 1161 identifications from the target database while there are 765, 739, and 635 interpreted spectra only from the preprocessed data and 2085 identification. We note these as unoverlapped identifications. The distribution of *Mascot* scores on the 1161 and 2085 identification are similar. For example, the mean score and standard deviation are (25.30, 8.56) and (25.98, 8.93), respectively (see data in Fig. 8(a)). However, the number of peaks in spectra, the length of identified peptides and the mass of precursors are different.

Figure 8(b) depicts the difference in length of identified peptides. Both the mean peptide lengths are 13. However, the number of peptides larger than 13 residues in the preprocessed data is greater than those in raw data (1330 vs. 657). Since the mass of the precursor ([M+H]) is in proportion to peptide length, the difference in the mass of precursors is similar to that in peptide length. In fact, there are a total of 772 precursors with mass larger than 2100 u that are interpreted in which 276 are interpreted after the *PeakSelect* process. In other words, there are more than 35% larger peptides that cannot be interpreted without preprocessing. Therefore, *PeakSelect* can help to identify longer and larger peptides.

Figure 8(c) shows the distribution of peak number in spectra. For co-interpreted spectra, the mean number in raw data is 530. For the unoverlapped interpreted spectra in raw data, the mean peak number is 364. However, for the unoverlapped interpreted spectra in preprocessed data, the mean number is 623. This implies that a lot of spectra with a larger number of peaks cannot be interpreted without

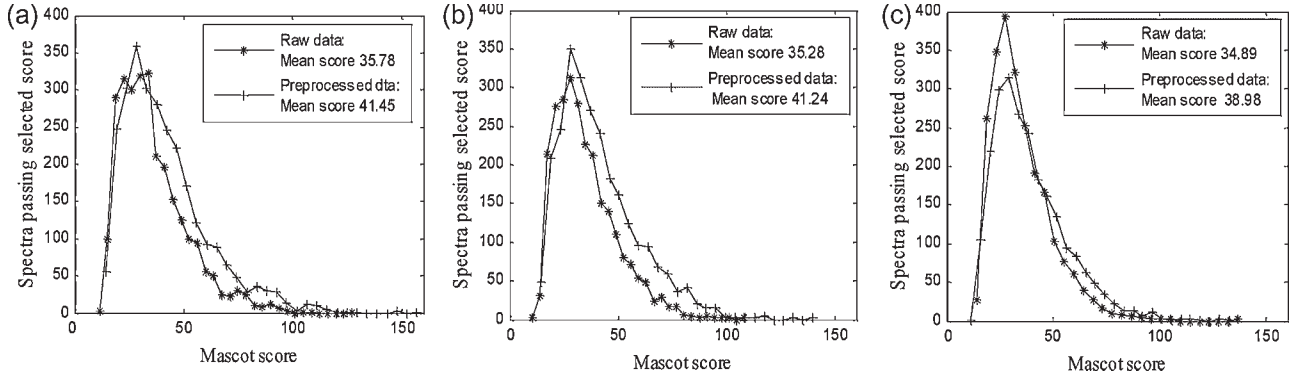

Figure 7. Distribution of *Mascot* scores on the spectra which are identified in both raw (*) and preprocessed (+) data from three samples are depicted in (a), (b) and (c), respectively. The curves represent the score distribution with interval of score of 5. The mean scores of the preprocessed data are increased 15.73%, 16.89% and 13.04% over that of the corresponding raw data in the three samples, respectively.

DOI: 10.1002/rcm

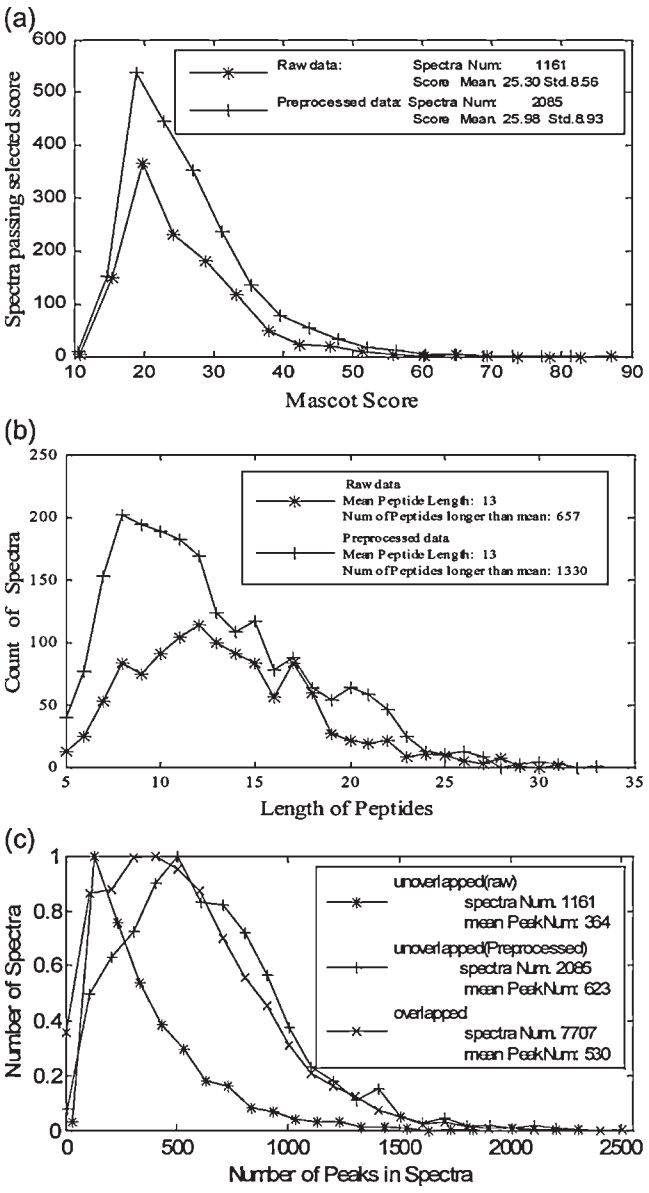

Figure 8. Information of the unoverlapped identifications. (a) Depiction of the distribution of the *Mascot* score in raw data and preprocessed data. (b) Length of the interpreted peptide on two datasets. There are 673 more identifications with more than 13 residues interpreted by the preprocessed spectra than those by the raw data. (c) Distribution of the spectra according to the number of peaks in them. The mean peak number of the preprocessed spectra, 623, is far larger than 364 the mean number of the raw data.

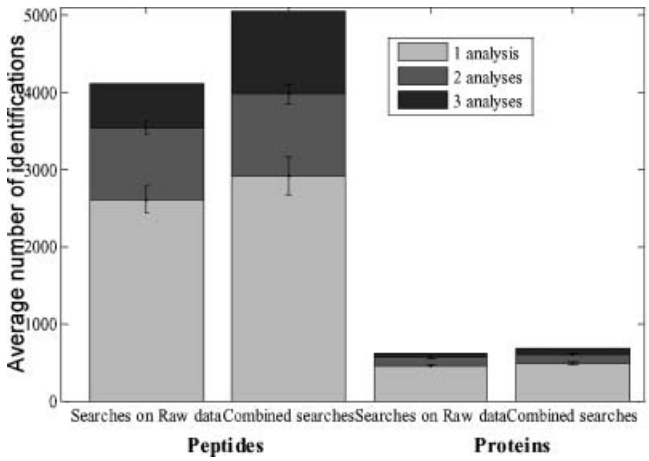

Figure 9. Error bars to show the maximum and minimum identifications for pairwise analyses, including searches on raw data and searches combining preprocessed data by *PeakSelect*. It seems very effective to combine *PeakSelect* to increase the identified information.

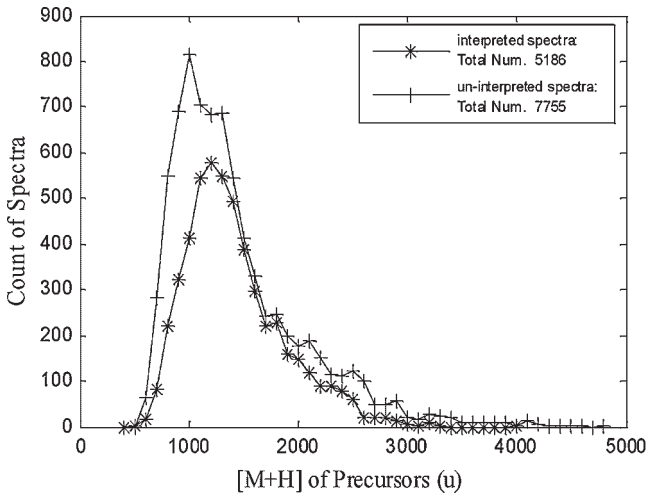

Figure 10. The interpreted and uninterpreted spectra which have more than 500 peaks. Many of the uninterpreted spectra have large [M+H] values. Since the mass error is almost linear to the mass of ions in TOF spectrometers, maybe the mass error is beyond the search parameter of 0.2 u in the uninterpreted spectra.

preprocessing because they have rich isotopic information and noise. However, *PeakSelect* can help to interpret these spectra.

If we introduce the process of *PeakSelect* and combine the Mascot search results, the average number of yielded peptides is 3232 and there is 11.07% increase compared to the number of 2910 just interpreted by raw spectra. For proteins, the number of combined yields is 537, with 10.27% increase compared to the number of 487 without *PeakSelect*. As discussed in Elias *et al.*,[25] the replicate analyses will increase the number of interpreted peptides and proteins. The error bars in Fig. 9 indicate the maximum and minimum identifications for pairwise replicate analyses under raw data and combining the preprocessed data. It seems very effective to combine *PeakSelect* to increase the identified information.

Although it can increase the identified peptides and proteins by combining the preprocessing of *PeakSelect*, there are still a large number of spectra uninterpreted. For example, Fig. 10 shows the interpreted and uninterpreted spectra which have more than 500 peaks. The data in Fig. 10 shows that many of the uninterpreted spectra have a large [M+H] value of precursor. They cannot be interpreted by

DOI: 10.1002/rcm

Mascot; besides the reason of post-translation modification, another important reason is that the measurement error of mass in the spectra is larger than the search parameters of 0.2 u. We will focus on recalibration of mass errors in spectra to improve the search results on the spectra with large precursors in the future.

CONCLUSIONS

We present a preprocessing method *PeakSelect* for mass spectra produced by a QqTOF-configured type of tandem mass spectrometer to increase the accuracy and reliability of database searching for peptide (protein) identification. Based on a new method of baseline identification and the natural isotopic information inherent in tandem mass spectra, we construct a decision tree to classify the noise and ion peaks in spectra. We present a comparison between *PeakSelect* and the preprocessing of ProteinLynx^{TM} Global *Server* (version 2.0.5). The experimental results show that *PeakSelect* performs much better than *ProteinLynx* by increasing the number of interpreted spectra and keeping higher *Mascot* scores and lower *Mascot* expectation values. In a large-scale analysis of yeast whole-cell lysate with QSTAR mass spectrometers, both peptide and protein coverages have been dramatically improved with the *PeakSelect* process.

Acknowledgements

This work was funded by the National Key Basic Research and Development Program (973) of China under Grant No. 2002CB713807 and CAS Knowledge Innovation Program, National High Technology Research and Development Program (863) of China under Grant Nos. 2007AA02Z315 and 2007AA02Z326. The authors thank Dr. She Chen of the National Institute of Biological Sciences, Beijing, for allowing us to use their Mascot server. The authors would also like to thank Bingpeng Ma and Xiaobiao Wang, Leheng Wang, Ruixiang Sun and Zuofei Yuan of the Institute of Computing Technology (CAS) for insightful discussions.

REFERENCES

1. Aebersold R, Goodlett DR. *Chem. Rev.* 2001; **101**: 269.
2. Aebersold R, Mann M. *Nature* 2003; **422**: 198.
3. Papayannopoulos IA. *Mass Spectrom. Rev.* 1995; **14**: 49.
4. Cotter R. *Time-of-Flight Mass Spectrometry*. ASC Professional Reference Books: Washington, DC, 1997.
5. Roepstorff P, Fohlman J. *Biomed. Mass Spectrom.* 1984; **11**: 601.
6. Johnson RS, Martin SA, Biemann K, Stults JT, Watson JT. *Anal. Chem.* 1987; **59**: 2621.
7. Falick AM, Hines WM, Medzihradszky KF, Baldwin MA, Gibson BW. *J. Am. Soc. Mass Spectrom.* 1993; **4**: 882.
8. Rouse JC, Yu W, Martin SA. *J. Am. Soc. Mass Spectrom.* 1995; **6**: 822.
9. Eng JK, McCormack AL, Yates JR III. *J. Am. Soc. Mass Spectrom.* 1994; **5**: 976.
10. Baginsky S, Cieliebak M, Gruissem W, Kleffmann T, Liptak Z, Mueller M, Penna P. *AuDeNS: A Tool for Automatic De Novo Peptide Sequencing*, Technical Report No. 383, Dept. of Computer Science, ETH Zurich, Switzerland.
11. Cannataro M, Guzzi PH, Mazza T, Veltri P. *Preprocessing, Management, and Analysis of Mass Spectrometry Proteomics Data*, NETTAB'05.
12. Rejtar T, Chen HS, Andreev V, Moskovets E, Karger BL. *Anal. Chem.* 2004; **76**: 6017.
13. http://www.nitehawk.com/voyager_macros/ http://www.appliedbiosystems.com/.
14. Gentzel M, Kocher T, Ponnusamy S, Wilm M. *Proteomics* 2003; **3**: 1597.
15. Senko MW, Beu SC, McLafferty FW. *J. Am. Soc. Mass Spectrom.* 1995; **6**: 52.
16. Senko MW, Beu SC, McLafferty FW. *J. Am. Soc. Mass Spectrom.* 1995; **6**: 229.
17. Lange E, Gropl C, Reinert K, Kohlbacher O, Hildebrandt R. *High-Accuracy Peak Picking of Proteomics Data Using Wavelet Techniques, PSB* 2006 Online Proceedings.
18. http://www.waters.com/WatersDivision/Contentd.asp?watersit=JLEY-69WTCR.
19. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. *Electrophoresis* 1999; **20**: 3551.
20. Hoefs J. *Stable Isotope Geochemistry*. Springer: Heidelberg, 1997.
21. Zhang JF, Gao W, Cai JJ, He SM, Zeng R, Chen S. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 2005; **2**: 217.
22. https://www2.appliedbiosystems.com/catalog/myab/StoreCatalog/products/ApplicationsHierarchy.jsp?hierarchyID=102& category1st=a88& category2nd=a89.
23. http://www.waters.com/WatersDivision/Contentd.asp?watersit=EGOO-6LLUNU.
24. http://gygi.med.havard.edu/pubs.
25. Elias JE, Hass W, Faherty BK, Gygi SP. *Nat. Methods* 2005; **2**: 667.

DOI: 10.1002/rcm