PeakSelect: preprocessing tandem mass spectra for better peptide identification

Jingfen Zhang1*, Simin He1, Charles X. Ling2, Xingjun Cao3, Rong Zeng3 and Wen Gao1*

1Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China
2Department of Computer Science, University of Western Ontario, London, Ontario, Canada
3Institute 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.
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. 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 also supplies preprocessing functionalities. Later, we will show that PeakSelect coupled with Mascot 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. 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 and the Q-ToF Ultima Global. In contrast to the threshold filtering and denoise transformation, 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, but can also describe the profile of the isotope cluster of a fragment ion universally. In addition, we study the difference between noise, isotope cluster of a fragment ion, and its first four isotopes (i.e., with one, two, three and four extra neutrons, respectively) as $P_1$, $P_2$, $P_3$ 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 $\Delta m_k$ are the mass differences between $P_k$ and $P$, for $k = 1–4$, respectively.

The theoretical IPV (denoted as $IPV$) 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 $T_1$, $T_2$, $\Delta m_1$, $\Delta m_2$ of $IPV$ for a given formula $C_{10}H_{22}N_{10}O_{13}S_{10}$ as follows:

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

(1)

$$T_1 = n_1qC + n_2qO + n_3qN + n_4qO_1 + n_5qS_1,$$

(2)

$$T_2 = n_4qO_2 + n_5qS_2 + \frac{1}{2} n_1^2$$

$$- \frac{1}{2} (n_1q^2 + n_2qH + n_3qN + n_4qO_1 + n_5qS_1).$$

(3)

$$\Delta m_1 = (n_1qC \Delta C + n_2qH \Delta H + n_3qN \Delta N + n_4qO_1 \Delta O_1 + n_5qS_1 \Delta S_1)/T_1$$

(4)

$$\Delta m_2 = n_4qO_2 \Delta O_2 + n_5qS_2 \Delta S_2 + \frac{1}{2} (n_1qC \Delta C + n_2qH \Delta H + n_3qN \Delta N + n_4qO_1 \Delta O_1 + n_5qS_1 \Delta S_1)^2$$

(5)

$$- \frac{1}{2} (n_1q^2 \Delta C + n_2qH \Delta H + n_3qN \Delta N + n_4qO_1 \Delta O_1 + n_5qS_2 \Delta S_2)/T_2$$

where $qC$, $qH$, $qN$ are relative abundances of $^{13}C$ to $^{12}C$, $D$ to $H$, and $^{15}N$ to $^{14}N$, and $\Delta O_1$, $\Delta O_2$, $\Delta S_1$, $\Delta S_2$ are the corresponding theoretical ones.

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.
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 ($M_{2k}$, $I_k$), $k = 0$–4, the charge $z$ can be calculated from the interval between $M_{2k}$. For $k > 1$, the value of ($M_{2k}$, $I_k$) will be substituted by zero if $p_k$ does not exist. After normalizing $z = 1$, the ($m/z$, intensity) pairs are converted into ($M_0$, $I_0$), where $M_0 = M_{2z}^z$ $z$ = ($z$ – 1)$^{1.0078}$, $k = 0$–4, respectively. Then eIPV can be obtained by:

$$
elPV = (M_0, R_1, R_2, R_3, R_4, \Delta m_1, \Delta m_2, \Delta m_3, \Delta m_4)$$

$$= (M_0, I_1, I_2, I_3, I_4, M_1 - M_0, M_2)$$

$$- M_0, M_3 - M_0, M_4 - M_0$$

(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 $I_{baseline} = (G_{baseline}, G_{deviation}, L_{baseline}, L_{deviation})$. 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 $eIPV = (M_0, R_1, R_2, \ldots)$ calculated from ($p_0$, $p_1$, $p_2$, $p_3$, $p_4$) 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).
another Ion2 and the charge of Ion2 can be identified by its following peaks. Here, a eIPV = (M, R1, R2, ... ) is calculated from (p0, p1, p2, ...) and the corresponding tIPV = (M, T1, T2, ..., ) where R1 matches well with T1 but R2 is far greater than T2; 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 (p0, p2, p4) and the other is between the eIPV and tIPV of (p1, p3, ... ). In Fig. 4(c), there is only a singly charged ion (p0, p2, p4) with some noise peaks because p1 and p3 are under the global baseline. Although p4 is under the global baseline too, it can still be identified as the second isotope of p0 from the value of eIPV.

Figure 2. The low peaks in the spectrum of peptide FTQKIFGGQNNSK (1) and KSLLSQILHK (2), in which (a) depicts the total peaks while (b) shows some real peaks such as b9++, y10++, and y8 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.
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.

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

which characterizes the distance between \(\text{Res}_M\) and the theoretical residue mass range \([\text{Res}_1, \text{Res}_2]\) of \(\text{Nomim}_{M}\), where \(\text{sign}(\cdot) = 1 \text{ if } x > 0 \text{ else } \text{sign}(\cdot) = 0\).

**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 the peak’s intensity and baseline as one important feature to decide whether the peak is noise. Consider a peak with intensity \(I_{\text{peak}}\) and the global baseline and local baseline of \(I_{\text{baseline}} = (G_{\text{mean}}, G_{\text{deviation}}, L_{\text{mean}}, L_{\text{deviation}})\), the value of \(F_{\text{R11}}\) and \(F_{\text{R12}}\) are treated as the first kind of feature and are calculated as follows:

\[
F_{\text{R11}} = (I_{\text{peak}} - G_{\text{mean}})/G_{\text{deviation}} \tag{7}
\]

\[
F_{\text{R12}} = (I_{\text{peak}} - L_{\text{mean}})/L_{\text{deviation}} \tag{8}
\]

**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 tryptic peptides from proteins of SWISS-PROT.

For a given ion with mass of \(M\), the nominal mass \(\text{Nomim}_{M}\) should be \(\text{Integ}_{M}, \text{e.g.},\) the integral part of \(M\), or \(\text{Integ}_{M} - 1\), and the mass residue \(\text{Res}_M\) will be \(M - \text{Integ}_{M}\) or \(M - \text{Integ}_{M} + 1\). According to the statistical results from proteins of SWISS-PROT, we obtain the residue range of \([\text{Res}_1, \text{Res}_2]\) 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 \(M_{z}\), we calculate the corresponding \(\text{Nomim}_{Mz}\) and \(\text{Res}_M\) supposing that it corresponds to an ion with \(z\) charges \((z = 1–5)\). If the calculated \(\text{Res}_M\) is far from the theoretical residue range of \(\text{Nomim}_{M}\) then the peak cannot correspond to a \(z\)-charged ion. Therefore, we use the following feature \(F_{\text{R2}}\):

\[
F_{\text{R2}} = \frac{\text{sign}((R_1 - T_{1\text{mean}}) * (R_1 - T_{1\text{max}})) * \text{min}(|R_1 - 1\text{mean}|, |R_1 - 1\text{max}|)}{T_{1\text{mean}}} \tag{10}
\]

\[
F_{\text{R2}} = \frac{\text{sign}((R_2 - T_{2\text{mean}}) * (R_2 - T_{2\text{max}})) * \text{min}(|R_2 - 2\text{mean}|, |R_2 - 2\text{max}|)}{T_{2\text{mean}}} \tag{11}
\]

**Other features**

Some other characters such as the assigned charge to peaks of a potential ion, the number of isotope peaks in the potential
isotope cluster, and the distance between the theoretical $Dms$ of the isotopes and the measured $Dm\sin IPV$, 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 $F_{\text{charge}}, F_{\text{IsoNum}}, F_{\text{Merr1}}, F_{\text{Merr2}}$ 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$^{\text{TM}}$ 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.
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, 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. 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., 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. In addition, we choose the same score filter criteria as described in Supplementary Table 1 in Elias et al. 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

<table>
<thead>
<tr>
<th></th>
<th>Raw .dta</th>
<th>Preprocessed .dta by PeakSelect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Selected</td>
<td>FP</td>
</tr>
<tr>
<td>Sample 1</td>
<td>3,167</td>
<td>24</td>
</tr>
<tr>
<td>Sample 2</td>
<td>3,011</td>
<td>32</td>
</tr>
<tr>
<td>Sample 3</td>
<td>2,798</td>
<td>52</td>
</tr>
<tr>
<td>Mean</td>
<td>2,992</td>
<td>36</td>
</tr>
</tbody>
</table>

*aTo be selected, PSMs had to receive scores greater than or equal to those indicated in Supplementary Table 1 in Elias et al.

bFP, estimated false positive identifications, and calculated by doubling the number of decoy hits; TP, estimated true positive identifications; Precision, TP/(TP+FP).
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

<table>
<thead>
<tr>
<th>Sample</th>
<th>Peptidesa</th>
<th>Proteinsa</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Raw-data</td>
<td>Preprocessed</td>
</tr>
<tr>
<td>Sample 1</td>
<td>2777</td>
<td>3140</td>
</tr>
<tr>
<td>Sample 2</td>
<td>2597</td>
<td>2950</td>
</tr>
<tr>
<td>Sample 3</td>
<td>2446</td>
<td>2640</td>
</tr>
<tr>
<td>Mean</td>
<td>2607</td>
<td>2910</td>
</tr>
</tbody>
</table>

*aThe interpreted peptides and proteins are selected from the target hits but not from the decoy hits.

*bThe number of identified peptides and proteins combined from the raw and preprocessed data.
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, including searches on raw data and searches combining preprocessed data by PeakSelect. It seems very effective to combine PeakSelect to increase the identified information.

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.
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™ 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


Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/rcm