Processing of tandem mass spectrometric data based on decision tree classification

Jingfen Zhang[1,2], Simin He[1], Jinjin Cai[1], Xingjun Cao[3], Ruixiang Sun[1], Yan Fu[1], Rong Zeng[3] and Wen Gao[1,2]

[1]Institute of Computing Technology, Chinese Academy of Sciences Beijing 100080, China;
[2]Graduate University of Chinese Academy of Sciences Beijing 100080, China;

[3]Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Science, Chinese Academy of Sciences Shanghai 200031, China.

We present a processing method for Q-TOF tandem mass spectra to increase the accuracy of database searching for peptide (protein) identification. Based on the natural isotopic information inherent in tandem mass spectra, we construct a decision tree after feature selection to classify the noise and ions' peaks. Furthermore, we recognize overlapping peaks to find the monoisotopic masses of ions for the following identification process. This processing increases the search speed and the reliability of peptide identifications.

Key words: Peptide identification; Q-TOF tandem mass spectra; Feature selection; Decision tree;

Introduction

Mass spectrometric analysis and database searching has been a well-known tool for peptide and protein identification [1]. During the experiment, the peptides separated from liquid chromatographers are fragmented and ionized by collision-induced dissociation (CID) and the ions are measured by mass spectrometer for mass/charge rations (*m/z*). Consequently, the peptides are identified (or sequenced) by these *m/z* values of ions in tandem spectrum with a sequence database searching.

Due to the variety of the fragment ions under CID and the existence of a large number of spectral noises, it is difficult to determine the sequence of a peptide from its tandem spectrum. Generally, a Q-TOF [2] spectrum of a peptide has 2000 to 8000 or more peaks, but only 1~5 percent of these peaks are real peaks. Here, real pecks correspond to the important and known fragment ions and are useful for the peptide identification. To increase the accuracy of peptide identification and decrease the computation complexity, the processing of MS/MS spectra is introduced before the database searching in order to select the peaks corresponding to fragment ions and minimize the number of the selected peaks.

To date several methods have been proposed for the processing of tandem data, including threshold filtering, denoise transforming and deisotoping. 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 or a specific number of the most intensive peaks in the specified *m/z* intervals [3, 4, 5, 6, 7]. As we know, the abundance is not the fundamental attribute of the real peaks. Many important b-series ions have very low abundance. In addition, for

various spectra, the quality, i.e., the intensity baseline of noises is totally different. Thus, using thresholds to remove the noise is not perfect. In denoising mechanism, some well known procedures such as wavelet transformation have been used to denoise the raw MS/MS spectrum [6]. However, the parameters such as the wavelet base functions, order, and level of decomposition would impact the potential spectra distortion by this procedure. In deisotoping, the isotopes are removed so that every fragment ion is represented only by one peak and the complexity of spectra is greatly reduced [6, 7]. Since peaks overlapping, i.e. two or more different ions have the confused isotopes masses, are observed frequently in spectra, deisotoping without identifying whether a peak corresponds to the monoisotope of one ion or the isotope of another ion leads to loss of some overlapped but important fragment ions.

To address the above issues, we present a new processing method for Q-TOF tandem mass spectra based on classification. At first, instead of the threshold filtering and denoise transforming, we use Gaussian Mixture Model (GMM) to estimate the baseline of noise and treat the baseline just as one feature to distinguish noise and real peaks. Secondly, a key concept of Isotope Pattern Vector (IPV) is introduced to characterize the isotope cluster of a fragment ion. The complex overlapping of isotope peaks are considered before deisotoping. Then we investigate the difference between noise, single fragment ions and overlapping ions based on features such as the baseline of noise and IPV, etc. Finally a decision tree is constructed to classify the peaks and the monoisotopic masses of all potential ions are calculated.

We next apply our processing on four different datasets and conduct extensive experiments to evaluate the specificity and sensitivity of classification. We also evaluate the effect of the processing on the speed and accuracy of the ***MASCOT*** [4] search and ***pFind*** [8] search. The experimental results show that the data processing approaches can increase the search speed and the reliability of peptide identifications.

Method

In this section, a new method is proposed to deal with Q-TOF tandem mass spectra to select the real peaks corresponding to fragment ions in the spectrum. Our solution has three new contributions: a Gaussian Mixture Model, a key concept of Isotope Pattern Vector (IPV) and a decision tree.

Firstly, a method based on Gaussian Mixture Model (GMM) and Expectation-Maximization (EM) algorithm to find the intensity baseline of noise of each spectrum is described. Secondly, the concept of Isotope Pattern Vector (IPV) is introduced to digitally characterize the isotope cluster of a fragment ion. Subsequently, the calculation of the theoretical and experimental IPV is introduced. Then the difference between noise, single fragment ions and overlapping ions based on some proposed features is investigated. Finally, a decision tree is constructed to classify the peaks into three categories and the monoisotopic masses of potential ions are calculated.

GMM for Intensity Baseline of Noise

The signal to noise ratio of precursor and imperfect laboratorial environment such as temperature shifts in the laboratory all should impact the quality of spectrum. Therefore, for various spectra, the intensity distribution of noise is different. For example, Figure.1 and Figure.2 show the spectra of peptide CCAADDKEACFAVEGPK and YLGYLEQLLR, respectively. It can be observed that the intensity baseline of noise peaks in fig.1 is much higher than that in fig.2.

The peaks corresponding to noises are random produced by the mass spectrometry during CID. Therefore the variable of intensity of noise obeys a normal distribution approximatively and a Gaussian mixture model can be established in

which a Gaussian represents the distribution of intensity of noise. Intuitively, the centroid of the Gaussian corresponding to noise is treated as the baseline. Practically, the mean and standard deviation are used to characterize the baseline of noise, noted as $I_{baseline}$ = (I_{mean}, $I_{deviation}$), and the value of $I_{baseline}$ are obtained by an Expectation-Maximization (EM) algorithm to estimate the parameters of the Gaussian mixture model. It is noted that we use the r**elative intensities** but not the absolute value of the intensities of peaks in spectrum. The highest value in intensity is 100%. Using MATLAB toolbox, the calculated results of (I_{mean}, $I_{deviation}$) for the data in Figure.1 and Figure.2 are (2.290144%, 0.350236%) and (1.012099%, 0.076899%), respectively. The calculated value is consistent with the observation of the noise in spectrum.

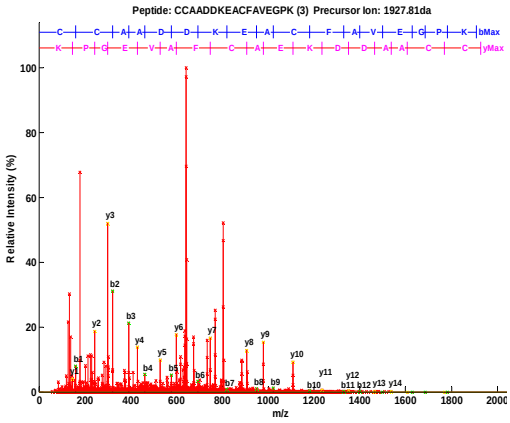

Figure1. MS/MS spectrum of peptide CCAADDKEACFAVEGPK in which the precursor holds 3 charges

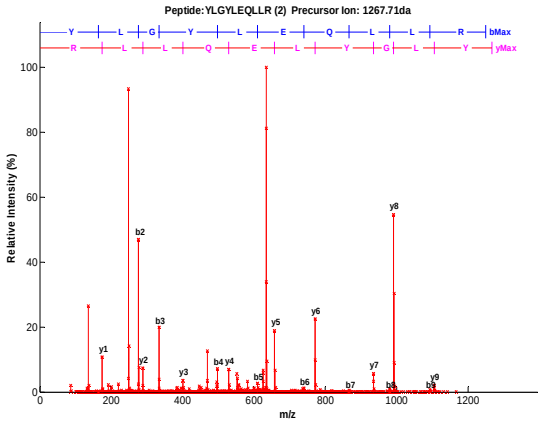

Figure2. MS/MS spectrum of peptide YLGYLEQLLR in which the precursor holds 2 charges

Isotope Pattern Vector (*IPV*)

Isotopes are elements that contain the same number of protons and electrons but differ in the number of neutrons in nucleus. The elements of H, C, N, O, and S have stable isotope distributions in nature [9]. Most proteins are composed of the above five elements, thereby, have relatively stable isotope patterns. We use isotope pattern vector (denoted as *IPV*) to digitally describes 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, Δm_1, Δm_2, Δm_3, Δm_4), where T_k is the relative abundance of P_k with respect to P, and Δm_k are the mass difference between P_k and P, for k=*1~4*, respectively.

Theoretical Isotope Pattern Vector (*tIPV*)

Since the five elements of H, C, N, O, and S have stable isotope distributions, the theoretical *IPV* (denoted as *tIPV*) of a fragment ion is definitely and can be deduced from its elemental component, i.e., from its molecular formulas. We assume that each extra neutrons of an atom in the peptide appears independently. Then the *tIPV* of for a given formula $C_{n1}H_{n2}N_{n3}O_{n4}S_{n5}$ can be deduced from the probability of the isotopes of each element. For simple, we just show the T_1, T_2, Δm_1, Δm_2 as follows:

$$M = (12.0000, 1.0078, 14.0030, 15.9949, 31.9721)\times(n_1, n_2, n_3, n_4, n_5)^{\mathrm{T}}, \quad (1)$$

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

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

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

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

where q_C, q_H, q_N are relative abundance 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 ratio 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 are the mass difference between ^{13}C and ^{12}C, D and H, and ^{14}N and ^{15}N, and ΔO_1, ΔO_2 (ΔS_1, ΔS2) are the mass difference between ^{17}O and ^{16}O, ^{18}O and ^{16}O (^{33}S and ^{32}S, ^{34}S and ^{32}S), respectively.

Experimental Isotope Pattern Vector (e*IPV*)

We can calculate the experimental isotope pattern (denoted as *eIPV*) of a fragment ion *P* if the isotope peaks of the ion are measured by mass spectrometer. We characterize an ion peak in mass spectrum in terms of (*m/z*, *intensity*), where *m*/*z* is the value of the mass to charge ratio and *intensity* is the relative height of the peak. Considering a group of isotope peaks (p_0, p_1, p_2, p_3, p_4) corresponding to an ion, the interval of the corresponding *m/z* values between p_0, p_1, p_2, p_3 and p_4 is around 1 da when the ion holds single charge while the interval is around 0.5 da when the ion holds double charges. In General, the interval is *1/z* da when the ion holds *z* charges. Contrariwise, the charge of an ion can be deduced by the *m/z* interval of the isotope peaks.

To calculate the value of *eIPV* for a fragment ion *P*, we find the corresponding isotope cluster peaks (p_0, p_1, p_2, p_3, p_4) with the (*m/z*, *intensity*) pairs (Mz_k I_k), *k*=0~4, in tandem spectrum and calculate the number of charge *z* from the interval between Mz_k. After normalizing *z*=1, the (*m/z*, *intensity*) pairs are converted to (M_k, I_k), where $M_k = Mz_k * z - (z\text{-}1)*1.0078$, *k*=0~4, respectively. Then *eIPV* can be obtained by:

$$eIPV = (M_0, I_1/I_0, I_2/I_0, I_3/I_0, I_4/I_0, M_1\text{-}M_0, M_2\text{-}M_0, M_3\text{-}M_0, M_4\text{-}M_0) \quad (6)$$

Feature Selection, Decision Tree and Classification

In this section, we investigate the difference between noise and fragment ions based on some proposed features, and construct decision tree to classify the peaks based on the value of the features.

At first, since peaks higher than noise baseline are more likely to be real peaks, it is necessary to find the noise baseline $I_{baseline}$ = (I_{mean}, $I_{deviation}$) of each spectrum. Secondly, each fragment ion has theoretical isotopes while noise does not have. Therefore, noise and real peaks can be distinguished based on the concept of *IPV*. Considering the measure error of mass spectrometer, the isotope peaks of a fragment ion should be observed and the experimental isotope pattern should match its theoretical isotope pattern. Thirdly, there are complex overlapping ions with different charge states and noisy data hence it is very important to recognize the charge state of fragment ions and the case of overlapping to determine all the monoisotopic mass of ions. Therefore, we select some features such as the charge state, the mass corresponding to the peak, the intensity distance between the peak and the intensity baseline of noise, and the distance

between the *eIPV* and *tIPV*, etc. Finally, we investigate the difference between noise and fragment ions, and learn the rules from some training samples and construct decision tree to classify the peaks: Class1: noise, Class2: real peaks but corresponding to single ion and Class3: real peaks corresponding to overlapping ions.

As described in above subsection, the interval of the value of *m/z* of the isotope peaks is around *1/z* da if the ion holds *z* charges. For a given peak p_0, we scan the spectrum and find out the overall groups of potential isotope peaks in tandem spectrum supposing three different charge states, i.e., *z*=1,2,3, and within a tolerance of 0.05/*z* da for the interval of *m/z* values. For example, there is a cluster peaks (p_0, p_1, p_2, p_3, p_4) with (*m/z*, *intensity*) pairs of (Mz_k, I_k), *k*=0~4 corresponding to a potential ion which holds *z* charges. Thus, the masses of the isotope of the ion (or the masses corresponding to the isotope peaks) can be calculated as $M_k = Mz_k*z - (z-1)*1.0078$ and the experimental *eIPV* can be calculated as (M_0, I_1/I_0, I_2/I_0, I_3/I_0, I_4/I_0, M_1-M_0, M_2-M_0, M_3-M_0, M_4-M_0). It is noted that if there is no peak at the *k*-th isotopic interval within given tolerance, then we set the virtual peaks $p_k, p_{k+1},..,p_4$ by setting the intensity I_j as zero, $j = k$~4. Therefore, it always can be obtained at least three groups of potential isotope peaks for a given peak p_0. Then it will be judged which one group corresponds to fragment ion by the following methods.

In the other hand, although the formula of a fragment ion is unknown during the processing, the theoretical *eIPV* of an ion can be estimated by the expected (or mean) isotope pattern of an average peptide of the given mass [10]. The average peptide is a peptide with an amino acid composition corresponding to the statistical distribution of amino acids in the non redundant database and the expected *tIPV*=(M_0, T_1, T_2, T_3, T_4, Δm_1, Δm_2, Δm_3, Δm_4) can be obtained. Therefore, we calculate the value of the features for each potential group of isotope peaks and obtain *V*=(M_0, *z*, $I_0-I_{mean}-3*I_{deviation}$, $I_0-I_{mean}+3*I_{deviation}$, I_1/I_0-T_1, I_2/I_0-T_2, I_3/I_0-T_3, I_4/I_0-T_4, $M_1-M_0-\Delta m_1$, $M_2-M_0-\Delta m_2$, $M_3-M_0-\Delta m_3$, $M_4-M_0-\Delta m_4$).

We select some peaks as training samples to observe the difference between the values of *V* corresponding to noise and that to real peaks. Specifically, we judge whether a peak is noise or it corresponds to an ion or it involves overlapped ions when the peptide sequence corresponds to the spectrum is known. There are four kinds of overlapping are considered as follows: case1: two ions with 1da difference in mass; case2: three consecutive ions with 1da difference in mass; case3: two ions with 3 da difference in mass; and case4: two ions with 2 da difference in mass which always confused with case1. The four profiles of the overlapping cases are showed in Figuer3. Then, we select three classes peaks corresponding to noise, single ion and overlapped ions, respectively. Finally, the decision tree to classify these peaks is constructed by using WEKA C4.5 toolbox.

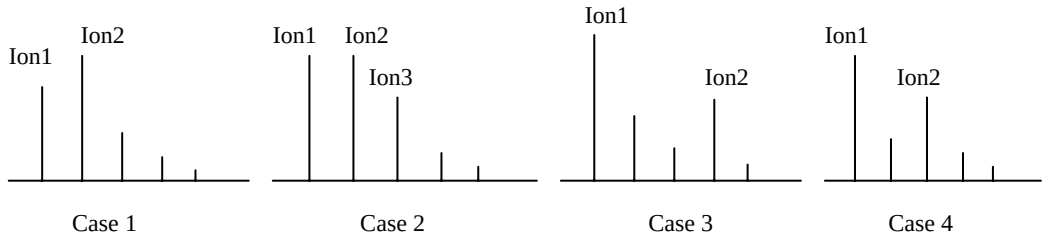

Figure3. Four profiles of the overlapping cases in which the Ion1, Ion2 and Ion3 represent the monoisotope of each ion involving overlapping.

According to the rules of the decision tree, all of the peaks in spectrum can be classified by the calculated values of *V* for its potential isotope peaks groups. It is noted that each peak will be classified to one and only one class. Specifically, a given peak p_0 is judged as noise if all of the values of *V* corresponding to the overall groups of potential isotope peaks are classified to Class1. For a given peak p_0 (with *m/z* value as *Mz*), if the value of *V* corresponding to the isotope peaks

group under charge *z* is classified to Class2, 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 to Class3, then there will be two or three monoisotopic masses can be obtained according to the overlapping cases. Finally, some masses corresponding to peaks which have been classified into Class2 and Class3 are selected prior to database searching.

Experimental Investigations

The experimental data sets we used are Q-TOF mass spectra and are given as follows including: 1) 54 spectra from tryptic digestion peptides supported by Dr. R. S. Johnson from the Immunex Corporation, Seattle, Washington, 2) 20 spectra of Glu-Fibrino peptide B, 3) 7 spectra of the tryptic peptides of bovine serum albumin protein, and 4) 9 spectra of the mixture of standard peptides measured during different time by the Research Centre for Proteome Analysis, Key Lab of Proteomics, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences. We note the four datasets as PepLutefisk, PepGFB, PepMix and PepBSA, repectively.

For performance metrics, we give some definitions as follow. At first, a peak is called real peak if its corresponding mass matches with a known theoretical ion; otherwise, it is called invalid peak. In these paper, the known theoretical ions include the predominant *a*-, *b*-, *y*- types of ions[12-13], immonium ion[14-15], and other less important ions such as *x*, *c*, *z* [12-13], internal fragment formed by a combination of the *a*- and *y*-type cleavage [14-15], and ions with a lost ammonia and a lost water[16]. It is noted that there are some peaks which really correspond to fragment ions but the corresponding masses can not match with any known theoretical ions because the knowledge of collision rules in CID is not complete at present. Consequently, the invalid peaks include noise peaks and some peaks corresponding to fragment ions but its ion type unknown to human beings. Secondly, it is said that there is a true positive if a real peak is classified correctly; otherwise it is said a false negative. Similarly, it is said that there is a true negative if an invalid peak is classified correctly; otherwise it is said a false positive. Finally, we use the *sensitivity* and *specifity* to measure the performance of classification. Here, *sensitivity* is defined as $TP/(TP+FN)$ and *specifity* is defined as $TN/(TN+FP)$, where *TP*, *FN* , *TN* and *FP* stand for the number of true positive, false negative, true negative, and false positive samples at classification, respectively.

In our experiment training samples of 900 cases and testing samples of 429156 cases are selected. The experimental results are summarized in Table 1. From the fourth column in Table 1, it is can be observed that the ratios of peak selection in four data sets are lower than 5%. The low selecting ratios can improve the speed of database searching greatly since the less the number of selected peaks, the simpler the computing of the subsequent identification process.

Table 1: The performance of the classification in processing

Data	The Num. of spectra	The Num. of total peaks/ The Num. of selected peaks	Ratio of Peak Selection	Sensitivity	Specifity
PepLutefisk	**54**	**89,256 / 3,721**	**4.168%**	**97.94%**	**99.06%**
PepGFB	**20**	**180,088 / 2,408**	**1.337%**	**97.77%**	**99.66%**
PepMix	**9**	**51,836 / 1,799**	**3.471%**	**93.68%**	**97.99%**
PepBSA	**7**	**18,720 / 789**	**4.215%**	**94.50%**	**97.76%**

As we know, it is the real peaks which make certain the identification of peptides. The more the real peaks are selected, the higher the accuracy of identification. Therefore the sensitivity of classification is very import for the identification.

The detailed results on *sensitivity* are depicted in Table 2. In the last column in Table 2, there are two kinds of false negative samples are given: one is the peaks corresponding to the predominant *a*-, *b*-, *y*- types of ions, and the other is the peaks corresponding to other less important types of ions. From the data in the last column, it can be observed that the former ***FN*** is much less than the later ***FN***, which means that the lost but important information in classification is few. Comparing with *sensitivity*, the *specifity* of processing is less important. The reason is: 1) the number of invalid peaks is related to the purity of testing samples and knowledge of collision rules in CID while the knowledge of collision rule is not sufficient and need improvement, hence the computing of *specifity* is not absolutely objective; 2) most peaks are of invalid peaks, a small number of classification error has little effect on the value of *specifity*.

Table 2: The detailed performance on *sensitivity* of the processing

Data	The Num. of selected peaks	The Num. of real peaks in spectra [a)]	The Num. of true positive (*TP*)	The Num. of false negative(*FN*) *a*-, *b*-, *y*- types of ions / other type of ions
PepLutefisk	**3,721**	**2909**	**2,849**	**11 / 49**
PepGFB	**2,408**	**1796**	**1,756**	**1 / 39**
PepMix	**1,799**	**775**	**726**	**9 / 40**
PepBSA	**789**	**379**	**358**	**3 / 18**

[a)] **peaks whose** corresponding masses match with known type of theoretical ion.

The purpose of processing is to improve the speed and accuracy of identification. We also evaluate the effect of the processing on the speed and accuracy of the ***pFind*** [8] search and ***MASCOT*** [4] search. In one hand, the experimental tests are performed with ***pFind***. The results show that under the same parameters of searching, the accuracy of identification is increase a little while speed of searching is improved up to 5~10 times. In the other hand, all the experiments are performed by submitting data to ***MASCOT*** through internet. Therefore, only accuracy level of search is compared while the testing of speed is not applicable. We submit two kinds of data to ***MASCOT***: the original spectra data and the spectra data after our process.

Comparing with the search results, we can see that: 1) If the peptide can be identified by the original data, i.e., the expected peptide sequence is listed at first position by ***MASCOT*** searching, it can be identified by the data after our processing, too. It means that the process does not destroy the data. 2) Comparing the search score including ***Score*** and ***Expect*** in ***MASCOT*** search results, there are 70% data (spectra) in which the scores for the data after our processing are much better than that for the original data. And 3) For some spectra, such as the spectrum of peptide QNCDQFEK(in which the amino acids ‘C’ is carbamidomethylated) and the spectrum of peptide DDPHACYSTVFDK, the query for the original data gives the expected sequence after the fifth position. However, the query for the processed data gives the correct answer at first position. Therefore, the searching after our process is more reliable. In the future research, we will focus on improving the *sensitivity* and *specifity* of the processing.

Acknowledgement

This work was funded by the National Key Basic Research and Development Program (973) of China under Grant No. 2002CB713807 and National Key Technologies R&D Program under Grant No.2004BA711A21. The authors thank Dr. R. Johnson for kindly providing the Q-TOF data. The authors would also like to thank Binpeng Ma and Xiaobiao Wang from the Institute of Computing Technology (CAS) for for insightful discussions

Reference

1. **Aebersold, R. and Mann, M. "Mass spectrometry-based proteomics", *Nature*, 2003, 422, 198-207**
2. **R. Cotter, "Time-of-Flight Mass Spectrometry", *ASC Professional Reference Books*, Washington, DC, 1997.**
3. **Eng, J. K., McCormack, A. L., Yates, J. R. III. "An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database", *J. Am. Soc. Mass Spectrom.*, 1994, 5, 976-989**
4. **Perkins,D.N., Pappin,D.J., Creasy,D.M. and Cottrell,J.S, "Probability-based protein identification by searching sequence databases using mass spectrometry data", *Electrophoresis*, 1999, 20, 3551-3567**
5. **Sacha Baginsky, Mark Cieliebak, Wilhelm Gruissem, Torsten Kleffmann, Zsuzsanna Liptak, Matthias Mueller and Paolo Penna, "AuDeNS: A Tool for Automatic De Novo Peptide Sequencing", *TECHNICAL REPORT* No. 383, ETH Zurich, Dept. of Computer Science**
6. **Tomas Rejtar, Hsuan-shen Chen, Victor Andreev, Eugene Moskovets, and Barry L. Karger, "Increased Identification of Peptides by Enhanced Data Processing of High-Resolution MALDI TOF/TOF Mass Spectra Prior to Database Searching", *Anal. Chem.*, 2004, 76, 6017-6028**
7. **Marc Gentzel, Thomas Kocher, Saravanan Ponnusamy Matthias Wilm, "Preprocessing of tandem mass spectrometric data to support automatic protein identification", *Proteomics*, 2003, 3, 1597–1610**
8. **Fu, Y., Yang, Q., Sun, R., Li, D., Zeng, R., Ling, C. X., Gao, W. "Exploiting the Kernel Trick to Correlate Fragment Ions for Peptide Identification via Tandem Mass Spectrometry", *Bioinformatics*, 2004, 20(12), 1948-1954**
9. **Jochen Hoefs, "Stable Isotope Geochemistry", *springer*, 1997**
10. **Jingfen Zhang, Wen Gao, Jinjin Cai, Simin He, Rong Zeng, and Runsheng Chen. "Predicting molecular formulas of fragment ions with isotope patterns in tandem mass spectra", *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2(3): in press, 2005***
11. **Taylor, J. A., Johnson, R. S. "Implementation and Uses of Automated De Novo Peptide Sequencing by Tandem Mass Spectrometry", *Anal.Chem.*, 2001, 73, 2594-2604**
12. **P. Roepstorff and J. Fohlman, "Proposal for a Common Nomenclature for Sequence Ions in Mass Spectra of Peptides," *Biomedical Mass Spectrometry*, 1984, 11(11), 601**
13. **R.S. Johnson, S.A. Martin, K. Biemann, J.T. Stults, and J.T. Watson, "Novel Fragmentation Process of Peptides by Collision-Induced Decomposition in a Tandem Mass Spectrometer: Differentiation of Leucine and Isoleucine," *Anal. Chem.*, 1987, 59(21), 2621-2625**
14. **A.M. Falick, W.M. Hines, K.F. Medzihradszky, M.A. Baldwin, and B.W. Gibson, "Low-Mass Ions Produced from Peptides by High-Energy Collision-Induced Dissociation in Tandem Mass Spectrometry," *J. Am. Soc. Mass Spectrometry*, 1993, 4(11), 882-893**
15. **I.A. Papayannopoulos, "The Interpretation of Collision-Induced Dissociation Tandem Mass Spectra of Peptides," *Mass Spectrometry Rev.*, 1995, 14(1), 49-73**
16. **J.C. Rouse, W. Yu, and S.A. Martin, "A Comparison of the Peptide Fragmentation Obtained from a Reflector Matrix-Assisted Laser Desorption-Ionization Time-of-Flight and a Tandem Four Sector Mass Spectrometer," *J. Am. Soc. Mass Spectrometry*, 1995, 6(9), 822-835**