J. Am. Soc. Mass Spectrom. (2012) 23:1209–1220
DOI: 10.1007/s13361-012-0380-4

RESEARCH ARTICLE

Investigation of Scrambled Ions in Tandem Mass Spectra. Part 1. Statistical Characterization

Nai-ping Dong, Yi-zeng Liang, Lun-zhao Yi

College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, People's Republic of China

Abstract

Scrambled ions have become the focus of recent investigations of peptide fragmentation. Here, an investigation of more than 390,000 high quality CID mass spectra is presented to explore the extent of scrambled ions in mass spectra and the possible fragmentation rules during scramble reactions. For the former, scrambled ions generally make up more than 10 % of mass spectra in number, although the abundances are less than 0.1 of the base peak. For the latter, relatively preferential re-opening sites were found for aliphatic residues Ala, Ile, Leu, and other residues such as Met, Gln, Ser, Phe, and Thr, whereas disfavored sites were found for basic residues Arg, Lys, and His, and Trp for both scrambled *b* and *a* ions. Similar preferential order in re-opening reaction was found in the reaction of losing internal residues when cleavage occurs at C-terminal side of 20 residues. However, when cleavage occurs at N-terminal side, Glu, Phe, and Trp become the most preferential sites. These results provide a deep insight into cleavage rules during scramble reactions for prediction of peptide mass spectra. Also, an additional investigation of whether scrambled ions could help discriminate false identifications from correct identifications was performed. Probing the number fraction of scrambled ions in falsely and correctly interpreted spectra and analyzing the correlation between scrambled ions and SEQUEST scores XCorr and *Sp* showed scrambled ions could at some extent help improve the discrimination in singly charged identifications, whereas no improvement was found for multiply charged results.

Key words: Scrambled ions, Statistical characterization, Tandem mass spectra, CID

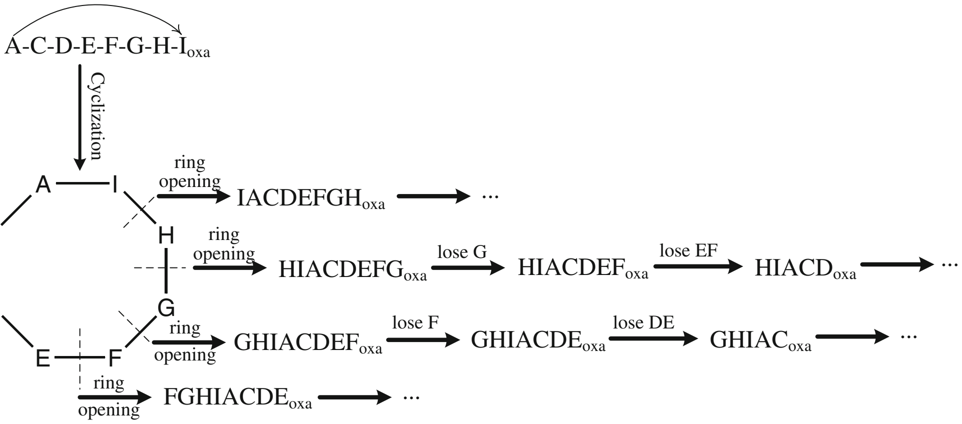
Introduction

In recent years, scrambled ions become the focus of attention in investigation of peptide fragmentation. So far, all scrambled ions reported are formed from *b* and *a* type ions (ion nomenclature is based on Roepstorff and Fohlman's [1] and Biemann's modifications [2]) and have generally been found in CID [3–6], CAD [7, 8] and HCD [7] tandem mass spectra. According to mass spectrometric and theoretical study of those ions, *b* or *a* ion scrambles via two-step reaction: cyclization and re-opening reaction. For the first step, a macrocycle is formed by nucleophilic attack of the free N-terminus on charged C-terminal oxazolone group (of *b* ions) or immonium group (of *a* ions). The macrocycle can be re-opened at any amide bond, leading to a sequence rearranged linear ion. Then amide bond cleavage occurs, losing C-terminal residues to form a scrambled ion. Apparently, the C-terminal residues lost at the second step are not the residues locate at C-terminus of primary sequence for the C-terminus is changed after the re-opening of the macrocycle. Thus, the scrambled ion can be considered as losing the internal residues from original *b* or *a* ions. These procedures are exemplified and illustrated in Scheme 1. As shown, Ala's N-terminal free amino group nucleophilically attacks C-terminus of the ion resulting in

Electronic supplementary material The online version of this article (doi:10.1007/s13361-012-0380-4) contains supplementary material which is available to authorized users.

Correspondence to: Yi-zeng Liang; *e-mail:* yizeng_liang@263.net

Received: 21 September 2011
Revised: 20 March 2012
Accepted: 22 March 2012
Published online: 27 April 2012



Scheme **1**. Permutation pattern of *b* ions ACDEFGHI^{+}

macrocycle cyclo-(ACDEFGHI). The macrocycle then opens up at any amide bond (except newly formed amide bond A–I), for example at bond

H–G, leading to linear sequence rearranged ion HIACDEFG^{+}. This rearranged ion occurs further degradation losing C-terminal residues to form scrambled ion such as HIACDEF^{+} (losing C-terminal G), HIACD^{+} (losing C-terminal EFG) etc., which scrambled the degradation of primary *b* ion. The cyclic structure of *b* ions have been proved by infrared multiple photon dissociation (IRMPD) [9] and ion mobility experiments [10]. In order to distinguish this reaction from conventional degradation of *b* or *a* ions, Harrison et al. named the scrambled ion non-direct sequence ion [4].

Although theoretical computation explicitly shows the pathways of scrambling of *a* and *b* ions, it is statistical investigation that would reveal the general fragmentation rules from a large number of MS/MS spectra generated by a large number of peptides with different amino acid compositions and numbers, as have been proven in discovery of peptide fragmentation behaviors. Thus, it would be valuable to employ statistical methods in obtaining fragmentation rules for scrambled ions. Until now, several systematic investigations have been provided. In these investigations, although not so large number of spectra was adopted, some preliminary rules were still revealed. Yagüe et al. [11] investigated anomalous fragmentation in eight peptide mass spectra and found the relatively more preferable elimination of aliphatic residues from scrambled *b* ions. Additionally, the authors found that the scramble reaction of *b* ions could be completed in activation time of 30 ms and eliminated by N-terminal acetylation, later confirmed by Jia et al. [8] and Harrison [6]. In a systematic work provided by Jia et al., some specific effects such as proline effect and Asn/Gln effect were found in the re-opening of macrocycles, whereas no predictive rules were found for cyclization reaction. Using model peptides with multiple alanines and one histidine located at a different position, scramble reaction was observed when histidine was near C-terminus of *b* ions and not recorded when histidine was near N-terminus [12]. In another work that investigated the influence of basic residues on scramble reaction, inhibition of formation of macrocycle was concluded for arginine containing *b* ions [13]. Also, preferential order of residues Asp, Glu, Lys, Asn, and Gln in re-opening reaction was also provided in that work. Most recently, a systematic investigation of acidic residues for *b* ion scrambling was carried out to examine the effect of single acidic residue and adjacent double acidic residues on the scramble reactions [14]. In this investigation, neither the presence nor the positions of acidic residues in peptides prevented *b* ion scrambling. Additionally, it is interesting to observe the dependence of preferential order of two acidic residues on collision energy, which may add to the uncertainty on evaluating the preferential order of residues during scramble reactions. However, all conclusions mentioned above should be verified by investigating a larger number of mass spectra to understand how general they are.

Besides the interests in fragmentation rules during scramble reactions, the most concerned for the study of scrambled ions is to find out what the extent of those ions is in tandem mass spectra and, consequently, if they really affect the results of peptide identification in qualitative proteomic research. Therefore, for the former, Saminathan et al. [10] examined 43 tryptic peptide mass spectra. The authors found 35 % of the spectra appeared as scrambled ions and averaged 8 % and 16 of the abundance on LCQ and QSTAR, respectively. For the latter, since scramble reaction of *a* or *b* ions means the peptides that could form same macrocyclic structure would produce similar mass spectra [14, 15], consideration of scrambled ions in peptide identifications might lead to ambiguous results. Recently, several works that more or less involved this problem were published. Saminathan et al. provided pioneer work to study the influence of integrating scrambled ion information into MASCOT search strategy on MASCOT search scores. The results were encouraging because no impact of scrambled ions was found on MASCOT identifications. Zubarev's group supported this conclusion by investigating a large number of CAD and HCD spectra in their communication [8], that sequence scrambling can be safely ignored in inspection of the possible factors leading to misidentification of peptide sequences.

In present work, we investigated more than 390,000 CID mass spectra extracted from human, *E. coli*, mouse, and yeast libraries in NIST Peptide Mass Spectra Libraries (release for 2011, http://peptide.nist.gov/) to mine possible fragmentation rules in scramble reactions of *b* or *a* ions. To achieve this, we re-annotated all CID mass spectra by including *y*, *b*, *a* ion series, precursor ion series, internal, and scrambled ions for the scrambled ions are out of consideration in the original annotation of the Libraries. After the re-annotation, a total number of more than 5.71 million scrambled ions were included in our dataset. In this dataset, more than 638,112 scrambled ions were extracted from singly charged peptide mass spectra. Since NIST Libraries are constructed by high quality peptide mass spectra, it would be beneficial to probe the extent and abundances of scrambled ions in tandem spectra. Furthermore, these spectra are valuable for confidently mining general fragmentation rules of the ions because they are produced by thousands of peptides with different sequences. In the last part of this work, we employed another standard data set to study whether the information of scrambled ions could improve or, at least, assist in improving the confidence of peptide identifications in protein sequence database search results.

Experimental

Data Collection and Processing

All data collection and analysis were performed on MATLAB (ver. 7.1.0, http://www.mathworks.com) by running programs written in-house. Mass spectra along with peptide sequences, charge states, and spectral annotations in the Libraries were read from NIST library files (.msp) by in-house programmed function *mspreader*. Since spectra in public downloadable Libraries were consensus spectra generated by clustering most similar spectra with highest database search scores, and the peaks in those consensus spectra were composed by peaks presented in the majority of replicate spectra with sufficient signal-to-noise ratio [15], we did not do any further spectral preprocessing such as denoising and de-isotoping for these spectra. Although the modification of residues such as oxidation of methionine is popular in peptides and proteins, we did not consider this situation in our investigation. The reason for this exclusion was to make the situation simple and keep all residues non-modified. Additionally, in order to make the investigation more confident, all multiple spectra in Libraries produced by same peptide with same charge state were removed. Lastly, total 399,830 spectra were included in our dataset with 33,187, 233,848, 113,558, and 19,237 spectra generated by singly, doubly, triply, and quadruply charged peptides, respectively. All peaks in a spectrum were normalized to the base peak.

Re-annotation of Mass Spectra

For re-annotation of mass spectra, we consulted the manual of NIST Libraries of Peptide Ion Fragmentation Spectra [16] and other references [10, 17]. In-house MATLAB programs *ionassign* and *annotationfilter* were written to annotate the spectra and filter absurd assignments (such as doubly or more charged b_2, a_2, or y_2 ions) automatically under set mass tolerance. The ions considered in this re-annotation procedure were: *y*, *b*, *a* type ions, precursor ions, and internal ions, along with their neutral losses, including water loss, ammonia loss, CO_2 loss (for *y* and precursor ions only), CO loss (for *a* ions and precursor ions, because losing CO from *b* ions leading to *a* ions), water added. Simultaneous losses were also considered in this procedure (but not for internal ions): water and ammonia loss, two water losses, two ammonia losses, water and CO loss (for precursor ions only), ammonia and CO loss (for precursor ions only). Lastly, since isotope peaks are commonly observed in peptide fragment ion spectra, the isotope counterparts of all ions mentioned above were predicted and assigned to spectral peaks with same mass tolerance.

Before assigning scrambled ions to mass spectral peaks, three hypotheses were set: (1) least length (defined by number of residues) of primary ions that form scrambled ions was 3 [18, 19]; (2) all scrambled ions were singly charged; and (3) scrambled ions formed via only one cyclization and re-opening step. Besides, since scrambled ions formed by ammonia loss of *a* ions have been found [20], water loss, water added, and ammonia loss of scrambled *b* and *a* ions were also considered. Because it is impossible to manually examine which sequence is correct for each mass spectral peak in our large dataset, we tried all possible rearrangement and residue losses for the primary sequence. Thus, a peak might be assigned by several ion names and one ion name might be composed by multiple sequences.

When all ion types were assured, the assignment was performed by MATLAB functions mentioned above under *m/z* tolerance of 0.8 Da. If there appeared multiple assignments for one peak or multiple sequences for one assignment, we retained all the assignments and sequences as the annotation of that fragment.

Statistical Investigation of Scrambled Ions

After re-annotating the mass spectra read from NIST Libraries, a statistical investigation was performed to explore the extent and fragmentation rules of scrambled ions. In order to extract peaks with high confident assignments of scrambled ions, three constraints were used: (1) no other ion type except scrambled ion was assigned to the peak; (2) the primary *b* or *a* ions that generated the scrambled ions must have been assigned to the peaks existing in the mass spectrum; and (3) the sequences for scrambled ions should not be the subsequences of respective primary sequences (for example, sequence CDE was considered as the subsequence of ion ACDEF$^+$, while ADEF was not); because these sequences may also be formed by internal fragmentation, this constraint dramatically reduced the annotation space of scrambled sequences (typically 40 % ~50 %). As there is high probability to produce random assignments to a peak because of too many permuted sequences generated by a given peptide, always several or even more than 10, these constraints could effectively reduce the random assignments. Moreover, as there was no specific *m/z* region found for scrambled ions, applying null hypothesis to estimate the extent of random assignments of those ions was inappropriate. Thus these rigorous constraints could assure high confident assignments. The investigation of the extent of scrambling reaction in mass spectra was subsequently operated by adopting all mass spectral peaks satisfying the above three constraints. We denoted this dataset as *dataset I*. For discovery of fragmentation rules of scramble reactions, however, only peaks annotated by single assignment with single sequence were adopted. Since it was impossible for us to filter all multiple assignments or multiple sequences to obtain determinate assignments or sequences for all peaks manually, such procedure could guarantee extracting unique re-opening and cleavage sites and obtaining high confident results. It should be mentioned here that because the mechanisms of scramble reactions occurred in doubly or more charged *b* or *a* ions are still unknown, and the primary *b* or *a* ions that produced scrambled ions could not be explicitly probed when ions with different charge states coexisted in a mass spectrum, we only investigated the spectra produced by singly charged precursor ions. Thus, 33,187 spectra with 94,930 scrambled *b* ions and 69,466 scrambled *a* ions were retained in this investigation. We denoted this as *dataset II*.

To probe the influence of scrambled ions on validation of protein database search results, *dataset III* was constructed by standard mass spectra data published by the Institute for System Biology (ISB) [21]. All spectra were read from SEQUEST mass spectral files (.dta). For each spectrum, peaks were sorted from highest intensity to lowest intensity and divided into two parts, and then median intensity value of the lower intensity part of that spectrum was calculated. Peaks with intensities below that median intensity value were considered as noise and filtered from that spectrum. All remaining peaks were annotated by MATLAB programs, which have been used in re-annotation of NIST peptide mass spectra under the *m/z* tolerance of 0.3 Da. In order to avoid falsely assigning isotope peaks to scrambled ions, which is always the case in annotation of tandem mass spectral peaks produced by spectrometers such as QTOF, a function named *isotopedistvalid* was programmed in-house to detect the isotopic distribution of each fragment ion. Once a possible isotopic distribution was detected, mercury algorithm [22] was employed to predict theoretical isotopic distributions of the sequences assigned to the monoisotopic peak. A similarity between the two distributions was then calculated by using the following formula:

$$S = 1 - \frac{\sum_i |I_e(i) - I_t(i)|}{\sum_i I_e(i) + I_t(i)}$$

where $I_e(i)$ and $I_t(i)$ represented the *i*th normalized peak intensity (normalized to the highest peak) in experimental and theoretical isotopic distributions. In the present work, an isotopic distribution detected by the program was accepted as real distribution of the fragment ion except the corresponding similarity score S was less than 0.9. If multiple isotopic distributions with different lengths (defined as number of peaks) were validated for single annotated peak, the longest isotopic distribution was retained. But if all isotopic distributions had the same length, the distribution with highest similarity score was retained. With the information of isotopic distribution, some misassignments were then corrected. Additionally, if several peaks were found sharing the same monoisotopic assignments, the peak with smallest *m/z* difference to the calculated *m/z* was assigned. Finally, scrambled ions satisfying the constraints mentioned above were collected for further analysis. For distinguishing true identifications from false identifications, spectra matched to the peptides that belonged to the 18 standard proteins and 15 identified contaminants reported by the original publication were considered as correct interpretations; otherwise were considered as false.

Results and Discussions

Scrambled Ions in CID Mass Spectra

The fraction and intensity distribution of scrambled ions in CID mass spectra are shown in Figure 1. It is surprising to find the general appearance of scrambled ions in CID mass spectra, as is shown in Figure 1a. According to Figure 1a, scrambled ions make up on average 13.25 %, 11.81 %, 7.95 %, and 3.75 % of mass spectra in number for singly, doubly, triply, and quadruply charged peptides, with maximum fraction to 49.71 %, 47.95 %, 29.26 %, and 20.51 %, respectively. It has been generally observed that large singly charged *b* ions could easily produce scrambled ions via head-to-tail cyclization and re-opening reactions [19]; such large fractions of scrambled ions in singly and doubly charged peptide CID mass spectra confirm this general observation, for most peptides in our dataset contain 10 or more amino acids. However, when peptides are triply or quadruply charged, the fraction of scrambled ions in number is dramatically decreased (Figure 1a). In previous investigations, multiply charged peptides have been demonstrated to be dissociated more efficiently than singly or doubly charged peptides to produce sequential informative mass spectra with multiply charged fragments [3, 23, 24].

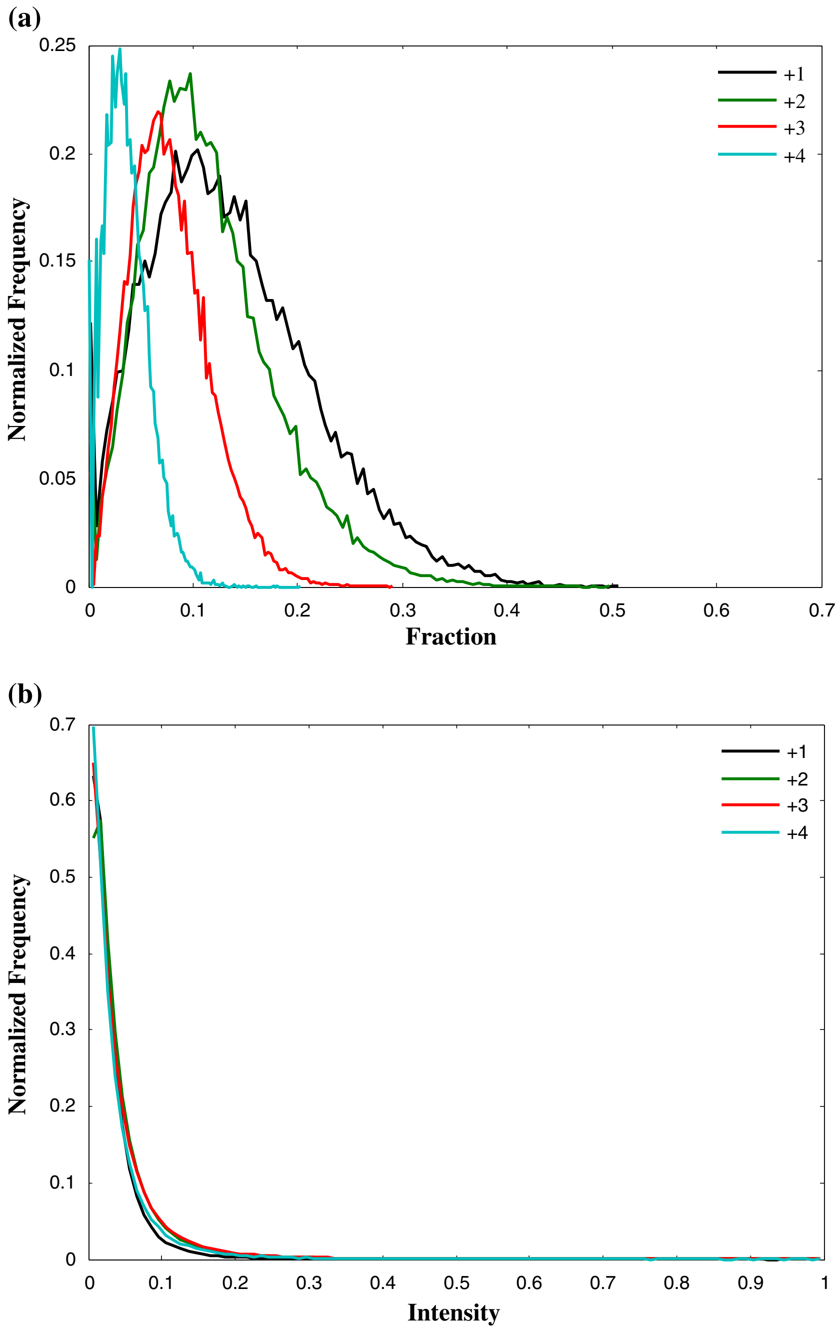


Figure 1. Distributions of fraction **(a)** and intensity **(b)** of scrambled ions at different charge states. Fraction of scrambled ions is calculated by number of peaks assigned to scrambled ions versus total number of peaks in that spectrum

This, as a result, may prohibit the scrambled reactions of *b* or *a* ions. But this should be carefully concluded because the spectra generated by highly charged peptides are much more complicated; that is, a peak could be assigned by different ion types or ions with different charge states. In fact, the extent of multiple assignments in triply and quadruply charged peptide mass spectra is much higher than singly and doubly charged (0.4612±0.0008, 0.5391±0.0016 versus 0.1760±0.0022 and 0.3281±0.0008 for Mouse Library). Consequently, a much less number of peaks was exclusively assigned to scrambled ions.

The distribution of abundances of scrambled ions in CID mass spectra is shown in Figure 1b. This distribution greatly extends the knowledge of scrambled ions in tryptic peptide mass spectra observed in previous work, in which only 43 singly and doubly charged tryptic peptides were used [10]. Although scrambled ions are generally observed, the intensities are generally below 0.1 of the base peak, and averaged normalized intensities among four charge states are 0.0327, 0.0451, 0.0463, and 0.0399, respectively. Additionally, there are about 3692 spectra (0.92 % of the total spectra) contain high intensity scrambled ions (normalized intensity ≥0.8). Unfortunately, since each of these high intensity scrambled ions was assigned by several different sequences, we could not target the exact cleavage site.

It has been observed that peptide length (defined as number of residues) could significantly impact peptide fragmentation [25–27]. The trends of fraction in number as well as averaged intensity of scrambled ions as a function of peptide length at different charge states were also investigated, and the results are shown in Figure 2. It is interesting to find nearly linear incremental trends for the fraction of scrambled ions in spectra when peptide length is below 16 in singly and doubly charged peptides. When the length of peptides is less than 30, fraction of scrambled ions increases along with peptide length among all four charge states. Yet, when lengths of peptides are more than 30, the trends become opposite, that the fraction is decreased, especially for triply charged ones. The reason for the increasing trends can be due to the preferential formation of scrambled ions for larger *b* and *a* ions.

In order to prove this explanation, the distribution of length of *b* and *a* ions, which generated scrambled ions,

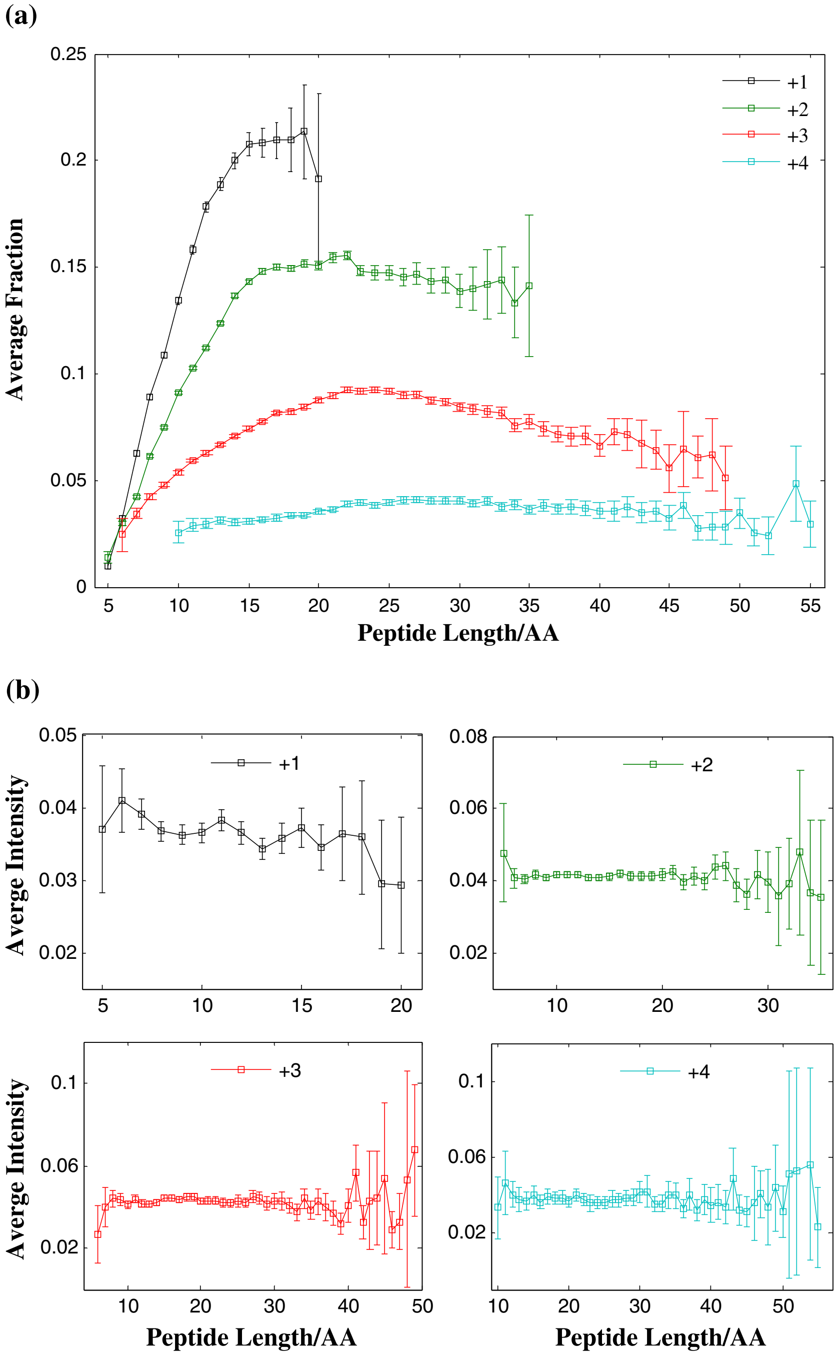


Figure 2. Fraction in number **(a)** and normalized intensity **(b)** of scrambled ions as a function of peptide length for spectra generated by singly (+1), doubly (+2), triply (+3) and quadruply (+4) charged peptides

versus length of precursor ion, was analyzed using *dataset II*, and the results are shown in Figure S-1 in the Supporting Information. It can be found that most *b* and *a* ions that formed scrambled ions accounted for 70 % to 90 % of their singly and doubly charged precursor peptide sequence and 50 % to 90 % of their triply and quadruply charged precursor peptide sequence in length. Previous theoretical investigation has found that larger *b* ions would adopt macrocyclic structures, even when only macrocyclic structure was identified for b_8 ion [19]. Thus, large *b* or *a* ions (generally larger than b_8) could be expected to cyclize more easily and, thereby, have high probability to form scrambled ions. However, the reason for the decreased trends when peptide length is larger than 30 could not be easily understood. One reason may be due to the more complex spectra generated by large peptides, leading to much more random assignments (or multiple assignments) than small peptides. This might also be the reason causing the dramatically decreased fraction of scrambled ions in quadruply charged spectra. Figure 2b shows the relationship between the intensity of scrambled ions and peptide length in tandem mass spectra. From the figure, a quite different conclusion could be drawn from the information of number fraction of scrambled ions. That is, the intensity of scrambled ions is independent of peptide length as well as charge state of precursor ions, being constant at ~0.04. Nevertheless, a relative positive correlation between length of primary ion and intensity of respectively formed scrambled ions is found (see Figure S-2 in the Supporting Information), supporting the trends that larger *b* or *a* ions are easier to form scrambled ions. Unfortunately, it is difficult to pick out primary ions and respective scrambled ions to show the correlation of intensities between both ions because probing the real intensity of primary ion from tandem mass spectral data only is impossible.

Residue Specific Reaction

During the years, scramble reactions of *b* and *a* type ions have been affirmed by several techniques, whereas the fragmentation rules behind the reactions are still lacking. In this section, fragmentation rules for re-opening and residue loss reaction are described using *dataset II*. It should be mentioned here that although scrambled *a* ions may be formed from primary *b* ions, it is impossible to trace the explicit primary ions (*b* or *a* ions) that produced those scrambled *a* ions. Thus, all ions assigned to scrambled *a* ions are hypothesized to be formed from *a* ions only and analyzed separately from scrambled *b* ions.

Figure 3a and b show the fraction of residues occurring re-opening reaction and internal residues loss in *b* ions. For re-opening at N-terminal and C-terminal side of 20 residues, the highest probability appears at residue Cys. However, the large error of the probability indicates that such probability should be further verified because only 28 peptides in our dataset contain unmodified Cys. Previous investigation it was observed that Ala was the favored loss site in scramble reactions [28]. The relative high fraction of occurring re-opening reaction at N-terminal and C-terminal side of Ala in Figure 3a confirms this observation. Besides the preferential re-opening reaction that happened at N-terminal and C-terminal side of Ala, other aliphatic residues Gly, Ile, Leu, and Val also show high fraction to occur the reaction, which have also been observed previously [11]. The reason behind these preferential reactions has not been fully understood since theoretical investigations are still lacking. It has been proposed that re-opening reaction needs attack of amide oxygen to the adjacent carbonyl carbon whose corresponding nitrogen is protonated [5], thus the residues with small side chains would be much more favorable because there is less steric hindrance between the two residues during the attack. Therefore, aliphatic residue sites become preferential as the bulk of aliphatic side chains are much smaller than side chains of other residues that contain functional groups or cyclic structures. Although the preferential sites for occurring re-opening reaction were found for aliphatic residues, a bit higher probability was observed for Met, Gln, Thr, and Tyr. This may be due to the effect of their side chains, as the side chains of residues Lys, Gln, Glu, Asp and Asn have been demonstrated to apparently influence the re-opening of the macrocyclic b_5^+ ions [28]. Nevertheless, it is surprising to find high probability of occurring re-opening reaction at N-terminal side of Phe for the side chain of Phe is phenyl group that can provide large hindrance.

Contrary to the favored ring re-opening sites of aliphatic residues and Met, Gln, Thr, and Tyr, the least favored sites are attributed to Arg, Lys, His, and Trp. The apparent prohibition of Arg and less activity of His in formation of macrocyclic *b* ions have been reported previously [12, 13], whereas the exact mechanism for this inhibition was not provided. Since basic residues could sequester protons at their side chains, the low fraction of occurring re-opening reaction at their sites may indicate the requirement of proton located at the amide bond during the reaction. The low fraction of occurrence re-opening reaction at N-terminal and C-terminal side of Trp may be due to the steric hindrance for Trp is the only amino acid containing two cyclic structures at its side chain. However, it is not clear why the relative low probability of occurring re-opening reaction occurred at N-terminal side of acidic residues Glu and Asp, whereas much higher probability occurred at C-terminal side of the residues.

When macrocyclic structure of *b* ion is re-opened at any amide bond except the newly formed one, new C-terminal residues could be lost as neutral molecule. The probability of occurring cleavage at N-terminal and C-terminal side of 20 residues is calculated and shown in Figure 3b. The situation becomes much more complex compared with the re-opening reaction in *b* ions. When cleavage occurs at C-terminal side, Ala, Met, Gln, Ser, and Thr have the highest probability whereas Arg, Lys, and His still have the lowest probability. When cleavage occurs at N-terminal side of residues, surprisingly, probability of occurring cleavage at Glu, Phe, Trp, and Tyr sites show up, but Ala becomes one of the sites

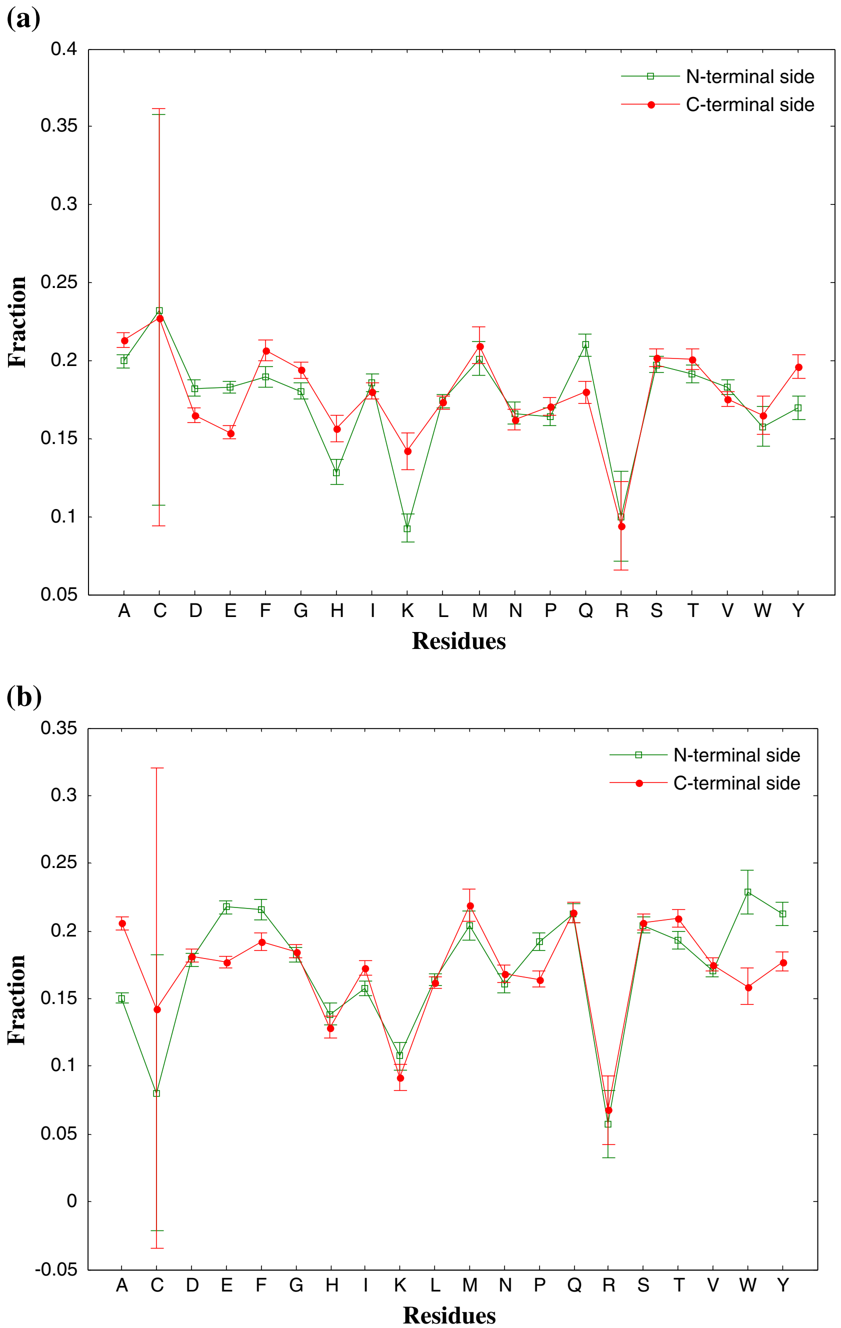


Figure 3. Fraction of 20 residues that occur reopening reaction **(a)** and losing internal residues **(b)** at N-terminal and C-terminal side in *b* ions. The single letter codes of residues listed along the lowermost raw correspond to the residues. The fraction value for each residue is calculated by number of that site occurring cleavage versus total number of corresponding residues in the dataset

that has the lowest probabilities. Additionally, no specific effect, such as proline effect, glutamic acid effect, etc., is found during the cleavage. The reason for this inconsistency with previous observations [8, 11–13] is unknown. One possible explanation is that since ion scrambled via two-step reactions, the transition structure formed by re-opening of macrocycle may be different from the structure for conventional peptide bond dissociation and, hence, caused the anomalous fragmentation rules. Another reason may be due to the random formation of scrambled ions in large dataset. That is, some possible rules may be inferred using dataset with a few spectra because scramble reactions are formed by limited and controlled peptide sequences. But for a large number of peptide sequences with different combinations of amino acids, the exclusive cleavage or prohibition occurred at some specific residue sites in one sub-dataset may not appear in another sub-dataset, which then results in different fragmentation rules.

In order to explore the influence of those preferential reactions occurring at aliphatic residue sites or Met, Gln, Thr, and Tyr and inhibited reactions at basic residue sites on scrambled ion intensity, 20 residues are divided into six groups:

(1) aliphatic group: Ala, Gly, Leu, Ile, and Val; (2) acidic group: acidic residues Asp and Glu; (3) basic group: basic residues His, Lys, and Arg; (4) cyclic side chain group: residues Phe, and Trp; (5) Pro; and (6) functional side chain group: the remaining residues Cys, Met, Asn, Gln, Ser, Thr, and Tyr. Pro was considered separately because of its particular structure and specific effect in peptide fragmentation. We calculated the average intensity for scrambled *b* and *a* ions whose primary ion sequences are missing all residues belonging to any one of the six groups. For comparison, the average intensity for absence of each residue was also calculated. The results are shown in Figure 4 and Figure S-3 in the Supporting Information. From the analysis, the lowest intensities of scrambled ions are found for the group that absents aliphatic residues, whereas highest intensities are found for the groups that absent basic residues and Pro, supporting the observations in Figure 3. But surprisingly, averaged intensity is significantly reduced when Lys was absent from peptides (0.0380±0.0003 when Lys was excluded from group (3) versus 0.0228±0.0014 when Lys was included). This abnormal tendency further supports the requirement of mobile protons in scrambled reactions for protonated side chain of Lys could nucleophilically attack the adjacent amide carbon to induce the amide bond cleavage [29]. However, the Arg acts conversely to Lys, which may be due to the special fragmentation behavior of Arg when lacking easily mobilizable proton (charge state of scrambled ion was assumed to be +1) [30]. The detailed probing of 20 absent residues implies that Ala, Ile, Leu, and Val as well as Lys containing peptides have the highest probability of scrambling reactions to occur (Figure S-3 and S-4). It is surprising to find that absence of Pro in scrambled ions could apparently increase the average intensity of the ions, indicating Pro in *b* ions may also prohibit scrambled reactions. As N-terminal cleavage to Pro is preferential in peptide fragmentation [31–33] and also observed in neutral loss of internal residues in scrambled reactions [7], the reason may be due to the side chain of Pro, which participates in a ring to the amide bond that requires high energy in forming macrocyclic structure of *b* ions containing Pro.

From the above analysis, preferential order of residue sites occurring scrambled reactions could thus be assumed. However, it should be noted that the exact order could not be quantified by statistical analysis only. According to Figure 3, in re-opening reaction, residue sites Met, Gln, Ser, Phe, and Thr are more preferential than aliphatic residue sites Ala, Gly, Ile, Leu, and Val, all of which are more preferential than Trp and Tyr, and the least favored sites are found for acidic residues Asp and Glu, and basic residues His, Lys, and Arg. When cleavage reaction occurs, the preferential order becomes much more complex, whereas the Met, Gln, Ser, Phe, and Thr sites are still the most preferential sites and basic residues are the least favored. Additionally, from Figure 4 and Figure S-3 in the Supporting Information, the preferential residue environment order for scrambled reactions can also be proposed, that Lys and aliphatic residues containing peptides are the most favored and His, Arg, and Pro containing peptides are the least favored. Unfortunately, no predominant or exclusive cleavage is found.

For scrambled *a* ions, probability of occurring re-opening and internal residue loss reaction at 20 residue sites are shown in Figure S-2 in the Supporting Information. As can be seen from the figure, the trends of preferential sites during the reactions are similar to *b* ions. In re-opening reaction, C-

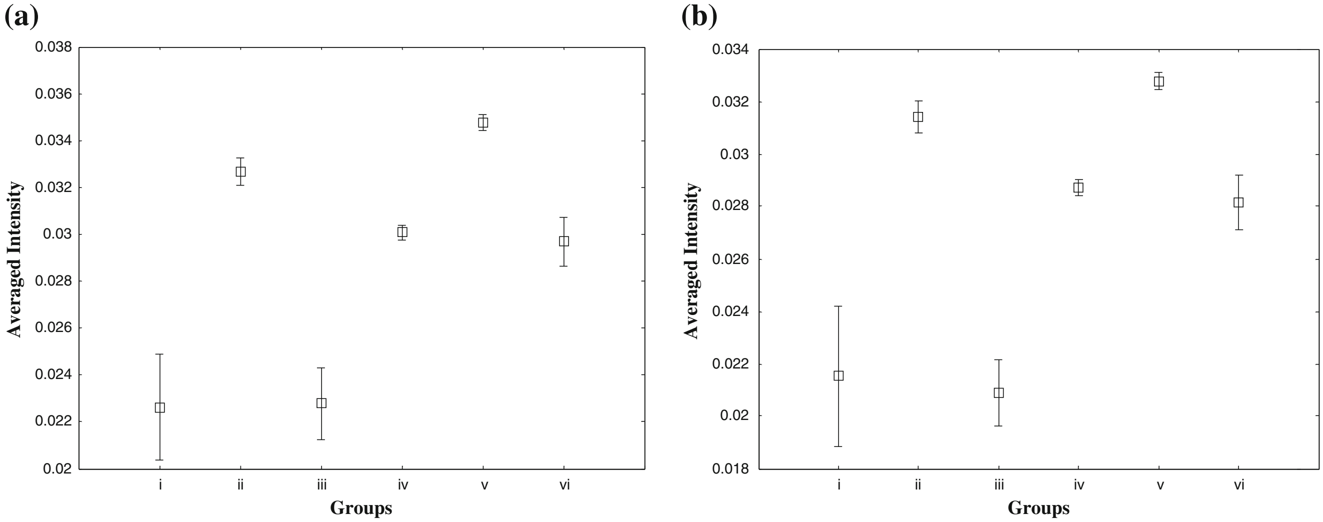


Figure 4. Average intensity of scrambled *b* (a) and *a* (b) ions when a group of residues is absent in peptides. Serial number along X axis denote: i) aliphatic group: Ala, Gly, Leu, Ile and Val; ii) acidic group: Asp and Glu; iii) basic group: His, Lys and Arg; iv) cyclic side chain group: Phe, Tyr, and Trp; v) Pro; vi) functional side chain group: Cys, Met, Asn, Gln, Ser, and Thr

terminal side of aliphatic residues Gly, Ile, and Val are preferential, as well as Asp Glu, Phe, and Met, and conversely, C-terminal side of basic residues Arg, Lys, His, and Trp are the most disfavored sites. When reaction occurs at N-terminal side, probability of Phe, Gly, and Met are reduced. Comparing the trends of preferential order of residues during scramble reaction in *b* and *a* ions, similar trends are found in re-opening and cleavage reactions, indicating that both type ions may react via same mechanisms after cyclization.

Are Scrambled Ions Helpful in Validation of Peptide Identifications?

Since formation of scrambled ions implies that peptide sequences that can form same cyclic structure would produce very similar spectrum [14, 15] or, at least, add complexity to mass spectra, it is important to evaluate the impact of the ions on peptide identifications. Here we used standard protein mix dataset (*dataset III*) with known false or random identifications (all are denoted as false identifications below) and correct identifications to find out whether the information of scrambled ions assigned in false and correct identifications could help validate database search results. Because this standard protein mix dataset is obtained from eight mass spectrometry platforms, it would be helpful to do the comparison among different platforms. In this analysis, we only used four platforms with four different mass analyzers (linear ion trap, 3D ion trap, Q-TOF, MALDI-TOF-TOF), which differ from mass accuracy, resolution, and identifications [34].

The fractions of number of scrambled ions in mass spectra as a function of peptide length for correct and false identifications on four platforms are shown in Figure 5 and Figure S-6 in the Supporting Information. As can be seen from the figures, significant difference of average fraction of scrambled ions between correct and false identifications is found for singly charged peptides, whereas nearly identical trends are found for doubly and triply charged peptides. The reason for nearly identical trends for multiply charged correct and false identifications may be due to the lower fraction of scrambled ions in those spectra than singly charged results. These trends may also be caused by the phenomenon that *m/z* distribution of *y*, *b*, *a* type ions and scrambled ions are completely mixed together in confidently identified dataset, resulting in a large fraction of scrambled ions generated in false identifications also. Analysis of SEQUEST [35] scores XCorr and *Sp* versus fraction of scrambled ions shows stronger positive correlations for correct identifications (see Figure S-7 in the Supporting Information), whereas sum intensity of scrambled ions versus score XCorr or *Sp* do not show any obvious trend (data not shown). These observations indicate that scrambled ions could at some extent improve the ability in discriminating correct identifications from false identifications in singly charged peptides, whereas no such ability was found for doubly and triply charged ones. In order to validate this conclusion, a boosting tree [36] method was performed with variables SEQUEST outputs XCorr, *Sp*, difference between highest and second normalized XCorr scores (dCn), precursor mass difference (dM), rank of *Sp* (R_{Sp}), fraction of predicted fragment ions matched (ions). For comparison, the same boosting tree with additional variable number fraction of scrambled ions in mass spectrum was also constructed. The result confirms the above observations that improved discrimination ability is found for singly charged peptides

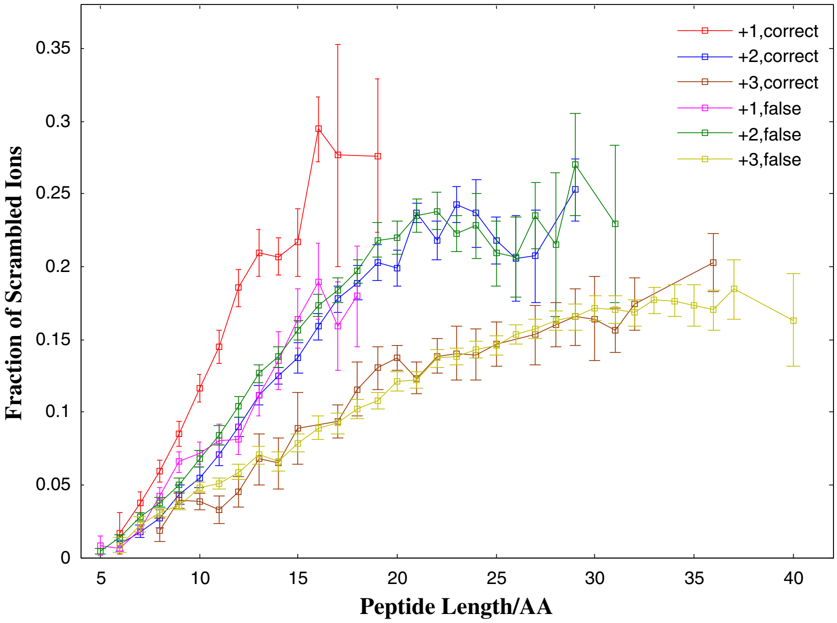


Figure 5. Number fraction of scrambled ions as a function of peptide length for false and true identifications at different charge states. The data were generated by 3D ion trap platform

only (misclassification rates 0.1419, 0.0876, and 0.0607 versus 0.1598, 0.0866, and 0.0622 for data generated by singly, doubly, and triply charged precursor ions respectively by using 3D ion trap data). These observations render the scrambled ions noninformative in validation of peptide identifications after performing protein database search, especially for multiply charged results. However, because scrambled ions generally make up more than 10 % of a mass spectrum (see Figure 1a), the conclusion that scrambled ions are useless during peptide identifications should still be carefully made, for such large extent would significantly impact the spectrum match scores during database search as well as de novo sequencing scores, as has been proven by investigating the changes of spectral match scores after removing scrambled ions from tandem mass spectra in our laboratory (data not shown).

In addition to the investigation of the relationship between scrambled ions and database search scores, comparison between sequential and scrambled ions at different peptide lengths in correct and false identifications was also made. The results are shown in Figure S-8 in the Supporting Information. As expected, no significant difference of the number fraction of scrambled ions in the spectrum is found between correct and falsely identified results at different charge states. The reason why average fraction of scrambled ions is higher for false identifications than correct identifications in some datasets may be due to the random assignments. For sequential ions, however, higher fractions are found for correct identifications in most datasets. It is interesting to find that most correct identifications have more sequential ions than scrambled ions, whereas for false identifications, the fraction values of sequential ions are too low and may intersect with fraction values of scrambled ions (see Figure S-8B, S-8G–S-8I). This finding may, to some extent, be used to validate peptide identifications, although still no significant improvement is found after adding the fraction difference between scrambled ions and sequential ions into boosting tree classification procedure.

Most recently, an investigation of masses of permuted sequence ions in high and low confidence peptide identification datasets was performed [37]. Although different characterization statistic and definition of permuted sequence ion from previous work [37] were used, the trends for sequential and permuted ions along peptide length are well consistent. Moreover, trends derived from the datasets generated on different platforms in our investigation expanded the knowledge of permuted sequence ions (i.e., scrambled ions) over previous investigations.

Conclusion

Much attention has been paid to scrambled ions in the study of peptide fragmentation in recent years, and deep insights into these ions, such as fragmentation pathways of cyclization and re-opening reactions, have been gained. In order to obtain fragmentation rules of scrambled ions, many systematic investigations were published. However, the exploration of a large number of scrambled ions to mine confident fragmentation rules is still lacking. Here, we presented an investigation of more than 390,000 spectra extracted from NIST Libraries to evaluate the extent of scrambled ions in CID mass spectra and mine fragmentation rules during re-opening reactions and neutral loss of internal residues. From this investigation, scrambled ions were generally observed and made up more than 10 % of CID tandem mass spectra in number, although the intensities were very low (less than 0.1 of the base peak). Additionally, the majority of scrambled ions were formed from large *b* or *a* ions, supporting the previous observation that large *b* ions adopted macrocyclic structures. In the exploration of fragmentation rules for re-opening reactions, relatively preferential sites were found for aliphatic residues Ala, Ile, and Leu as well as Met, Gln, Ser, Phe, and Thr, whereas disfavored sites were found for basic residues Arg, Lys, and His. This might indicate the requirement of protons that are located at the reaction sites. A disfavored reaction was also found for Trp site, indicating the steric hindrance could also affect the reactions. Furthermore, preferential reaction occurring at aliphatic residues (Ala, Ile, and Leu) and other residues (Met, Gln, Ser, Phe, and Thr as well as Lys) reveals the effect of side chains on re-opening reaction, that for the former, small bulk of side chains in aliphatic residues facilitates the attachment of the amide oxygen to the adjacent carbonyl carbon, and for the latter, specific side chains may participate in re-opening reaction to induce macrocycle breakage. For neutral losses of internal residues, anomalous preferential cleavage sites were observed. However, no predominant rules were found in this statistical analysis, indicating scramble reactions could occur at any amide bond of certain primary sequence.

In the last part of present work, the influences of scrambled ions on peptide identification results obtained by four tandem mass spectrometry platforms were also investigated. As have been demonstrated, poor relationship was found between SEQUEST scores (XCorr and *Sp*) and fraction of scrambled ions. In order to find out whether the scrambled ions could help validate SEQUEST search results, the number fraction of the ions in correct and false identifications was investigated. Furthermore, a boosting tree model was constructed to quantitatively evaluate the discriminant ability. The results showed that scrambled ions could to some extent help distinguish correct identifications from false identifications in singly charged peptides, whereas no improvement was observed for doubly and triply charged results.

Acknowledgments

The authors thank Liang-Xiao Zhang and Hong-dong Li for their kind discussion. This work was financially supported by the National Nature Foundation Committee of PR China (grants 20875104, 21075138, and 21105129), and the Fundamental Research Funds for the Central Universities (2010QZZD010 and 2011QNZT053).

References

1. Roepstorff, P., Fohlman, J.: Proposal for a Common Nomenclature for Sequence Ions in Mass Spectra of Peptides. *Biomed. Mass Spectrom.* **11**, 601 (1984)
2. Biemann, K.: Contributions of Mass Spectrometry to Peptide and Protein Structure. *Biomed. Environ. Mass Spectrom.* **16**, 99–111 (1988)
3. Tang, X.J., Thibault, P., Boyd, R.K.: Fragmentation Reactions of Multiply-Protonated Peptides and Implications for Sequencing by Tandem Mass Spectrometry with Low-energy Collision-Induced Dissociation. *Anal. Chem.* **65**, 2824–2834 (1993)
4. Harrison, A.G., Young, A.B., Bleiholder, C., Suhai, S., Paizs, B.: Scrambling of Sequence Information in Collision-Induced Dissociation of Peptides. *J. Am. Chem. Soc.* **128**, 10364–10365 (2006)
5. Bleiholder, C., Osburn, S., Williams, T.D., Suhai, S., Van Stipdonk, M., Harrison, A.G., Paizs, B.: Sequence-scrambling fragmentation pathways of protonated peptides. *J. Am. Chem. Soc* **130**, 17774–17789 (2008)
6. Harrison, A.G.: Peptide Sequence Scrambling through Cyclization of b (5) Ions. *J. Am. Soc. Mass Spectrom.* **19**, 1776–1780 (2008)
7. Jia, C., Qi, W., He, Z.: Cyclization Reaction of Peptide Fragment Ions During Multistage Collisionally Activated Decomposition: an Inducement to Lose Internal Amino-acid Residues. *J. Am. Soc. Mass Spectrom.* **18**, 663–678 (2007)
8. Goloborodko, A.A., Gorshkov, M.V., Good, D.M., Zubarev, R.A.: Sequence Scrambling in Shotgun Proteomics is Negligible. *J. Am. Soc. Mass Spectrom.* **22**, 1121–1124 (2011)
9. Erlekam, U., Bythell, B.J., Scuderi, D., Van Stipdonk, M., Paizs, B., Maitre, P.: Infrared Spectroscopy of Fragments of Protonated Peptides: Direct Evidence for Macrocyclic Structures of b5 Ions. *J. Am. Chem. Soc.* **131**, 11503–11508 (2009)
10. Saminathan, I.S., Wang, X.S., Guo, Y., Krakovska, O., Voisin, S., Hopkinson, A.C., Siu, K.W.: The Extent and Effects of Peptide Sequence Scrambling via Formation of Macrocyclic B Ions in Model Proteins. *J. Am. Soc. Mass Spectrom.* **21**, 2085–2094 (2010)
11. Yagüe, J., Paradela, A., Ramos, M., Ogueta, S., Marina, A., Barahona, F., Lopez de Castro, J.A., Vazquez, J.: Peptide Rearrangement During Quadrupole Ion Trap Fragmentation: Added Complexity to MS/MS Spectra. *Anal. Chem.* **75**, 1524–1535 (2003)
12. Bythell, B.J., Knapp-Mohammady, M., Paizs, B., Harrison, A.G.: Effect of the His on the Cyclization of b Ions. *J. Am. Soc. Mass Spectrom.* **21**, 1352–1363 (2010)
13. Molesworth, S.P., Van Stipdonk, M.J.: Apparent Inhibition by Arginine of Macrocyclic b Ion Formation from Singly Charged Protonated Peptides. *J. Am. Soc. Mass Spectrom.* **21**, 1322–1328 (2010)
14. Atik, A.E., Yalcin, T.: A Systematic Study of Acidic Peptides for b-Type Sequence Scrambling. *J. Am. Soc. Mass Spectrom.* **22**, 38–48 (2011)
15. Chen, X., Tirado, M., Steill, J.D., Oomens, J., Polfer, N.C.: Cyclic Peptide as Reference System for b Ion Structural Analysis in the Gas Phase. *J. Mass Spectrom.* **46**, 1011–1015 (2011)
16. http://peptide.nist.gov/docs/NIST_PepLib_08.pdf. Accessed 28 Sept 2010
17. Khatun, J., Ramkissoon, K., Giddings, M.C.: Fragmentation Characteristics of Collision-Induced Dissociation in MALDI TOF/TOF Mass Spectrometry. *Anal. Chem.* **79**, 3032–3040 (2007)
18. Molesworth, S., Osburn, S., Van Stipdonk, M.: Influence of Size on Apparent Scrambling of Sequence During CID of b-type Ions. *J. Am. Soc. Mass Spectrom.* **20**, 2174–2181 (2009)
19. Chen, X., Yu, L., Steill, J.D., Oomens, J., Polfer, N.C.: Effect of Peptide Fragment Size on the Propensity of Cyclization in Collision-Induced Dissociation: Oligoglycine b(2)–b(8). *J. Am. Chem. Soc.* **131**, 18272–18282 (2009)
20. Vachet, R.W., Bishop, B.M., Erickson, B.W., Glish, G.L.: Novel Peptide Dissociation: Gas-Phase Intramolecular Rearrangement of Internal Amino Acid Residues. *J. Am. Chem. Soc.* **119**, 5481–5488 (1997)
21. Klimek, J., Eddes, J.S., Hohmann, L., Jackson, J., Peterson, A., Letarte, S., Gafken, P.R., Katz, J.E., Mallick, P., Lee, H., Schmidt, A., Ossola, R., Eng, J.K., Aebersold, R., Martin, D.B.: The Standard Protein Mix Database: a Diverse Data Set to Assist in the Production of Improved Peptide and Protein Identification Software Tools. *J. Proteome Res.* **7**, 96–103 (2008)
22. Rockwood, A.L., Van Orden, S.L., Smith, R.D.: Rapid Calculation of Isotope Distributions. *Anal. Chem.* **67**, 2699–2704 (1995)
23. Wysocki, V.H., Tsaprailis, G., Smith, L.L., Breci, L.A.: Mobile and localized protons: a framework for understanding peptide dissociation. *J. Mass Spectrom.* **35**, 1399–1406 (2000)
24. Ishikawa, K., Nishimura, T., Koga, Y., Niwa, Y.: Role of Coulomb Energy in Promoting Collisionally Activated Dissociation of Multiply Charged Peptides Formed by Electrospray Ionization. *Rapid Commun. Mass Spectrom.* **8**, 933–938 (1994)
25. Kapp, E.A., Schutz, F., Reid, G.E., Eddes, J.S., Moritz, R.L., O'Hair, R.A., Speed, T.P., Simpson, R.J.: Mining a Tandem Mass Spectrometry Database to Determine the Trends and Global Factors Influencing Peptide Fragmentation. *Anal. Chem.* **75**, 6251–6264 (2003)
26. Elias, J.E., Gibbons, F.D., King, O.D., Roth, F.P., Gygi, S.P.: Intensity-Based Protein Identification by Machine Learning from a Library of Tandem Mass Spectra. *Nat. Biotechnol.* **22**, 214–219 (2004)
27. Huang, Y., Tseng, G.C., Yuan, S., Pasa-Tolic, L., Lipton, M.S., Smith, R.D., Wysocki, V.H.: A Data-Mining Scheme for Identifying Peptide Structural Motifs Responsible for Different MS/MS Fragmentation Intensity Patterns. *J. Proteome Res.* **7**, 70–79 (2008)
28. Molesworth, S., Osburn, S., Van Stipdonk, M.: Influence of Amino Acid Side Chains on Apparent Selective Opening of Cyclic b5 Ions. *J. Am. Soc Mass Spectrom.* **21**, 1028–1036 (2009)
29. Yalcin, T., Harrison, A.G.: Ion Chemistry of Protonated Lysine Derivatives. *J. Mass Spectrom.* **31**, 1237–1243 (1996)
30. Bythell, B.J., Suhai, S., Somogyi, A., Paizs, B.: Proton-Driven Amide Bond-Cleavage Pathways of Gas-Phase Peptide Ions Lacking Mobile Protons. *J. Am. Chem. Soc.* **131**, 14057–14065 (2009)
31. Hunt, D.F., Yates, J.R.I.I.I., Shabanowitz, J., Winston, S., Hauer, C.R.: Protein Sequencing by Tandem Mass Spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **83**, 6233–6237 (1986)
32. Vaisar, T., Urban, J.: Probing the Proline Effect in CID of Protonated Peptides. *J. Mass Spectrom.* **31**, 1185–1187 (1996)
33. Huang, Y., Triscari, J.M., Tseng, G.C., Pasa-Tolic, L., Lipton, M.S., Smith, R.D., Wysocki, V.H.: Statistical Characterization of the Charge State and Residue Dependence of Low-Energy CID Peptide Dissociation Patterns. *Anal. Chem.* **77**, 5800–5813 (2005)
34. Domon, B., Aebersold, R.: *Mass Spectrometry and Protein Analysis. Science* **312**, 212–217 (2006)
35. Yates, J.R.I.I.I., Eng, J.K., McCormack, A.L., Schieltz, D.: An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *Anal. Chem.* **67**, 1426–1436 (1994)
36. Freund, Y., Schapire, R.E.: A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **55**, 119–139 (1997)
37. Yu, L., Tan, Y., Tsai, Y., Goodlett, D.R., Polfer, N.C.: On the Relevance of Peptide Sequence Permutations in Shotgun Proteomics Studies. *J. Proteome Res.* **10**, 2409–2416 (2011)