analytical chemistry

ARTICLE

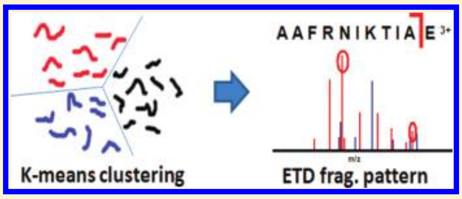
pubs.acs.org/ac

# Statistical Analysis of Electron Transfer Dissociation Pairwise Fragmentation Patterns

Wenzhou Li,[†] Chi Song,[‡] Derek J. Bailey,[§,∥] George C. Tseng,[‡] Joshua J. Coon,[§,∥,⊥] and Vicki H. Wysocki*,[†]

[†]Department of Chemistry and Biochemistry, University of Arizona, Tucson, Arizona 85721, United States
[‡]Department of Biostatistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, United States
[§]Department of Chemistry, University of Wisconsin, Madison, Wisconsin 53706, United States
[∥]Genome Center of Wisconsin, University of Wisconsin, Madison, Wisconsin 53706, United States
[⊥]Department of Biomolecular Chemistry, University of Wisconsin, Madison, Wisconsin 53706, United States

*Supporting Information*



**ABSTRACT:** Electron transfer dissociation (ETD) is an alternative peptide dissociation method developed in recent years. Compared with the traditional collision induced dissociation (CID) b and y ion formation, ETD generates c and z ions and the backbone cleavage is believed to be less selective. We have reported previously the application of a statistical data mining strategy, K-means clustering, to discover fragmentation patterns for CID, and here we report application of this approach to ETD spectra. We use ETD data sets from digestions with three different proteases. Data analysis shows that selective cleavages do exist for ETD, with the fragmentation patterns affected by protease, charge states, and amino acid residue compositions. It is also noticed that the $c_{n-1}$ ion, corresponding to loss of the C-terminal amino acid residue, is statistically strong regardless of the residue at the C-terminus of the peptide, which suggests that the peptide gas phase conformation plays an important role in the dissociation pathways. These patterns provide a basis for mechanism elucidation, spectral prediction, and improvement of ETD peptide identification algorithms.

Tandem mass spectrometry based peptide and protein identification involves the dissociation of peptide or protein ions to generate fragment ions. A conventional dissociation method is collision induced dissociation (CID), in which peptide precursor ions collide with inert gas atoms or molecules and dissociate. CID typically results in fragmentation along the peptide backbone at the amide bonds, producing predominantly N-terminal b and C-terminal y ions. It is widely known that the CID fragmentation patterns are highly dependent on the sequence of the peptide and the amino acid (AA) residue composition. Preferential cleavage, for example, is expected at the N-terminus of proline in the presence of a mobile proton or the C-terminus of aspartic acid when no mobile proton is available.[1−3] Many studies show that understanding these fragmentation patterns can potentially improve the interpretation of CID spectra as well as peptide and protein identifications.[4−8] An example of this is our recently reported peptide identification algorithm SQID,[9] which incorporates intensity statistics from a large CID data set and shows improved performance compared with several popular algorithms that do not strongly consider intensity.

Electron transfer dissociation (ETD),[10] similar to electron capture dissociation (ECD),[11] has gained popularity because of its ability to retain post-translational modifications and produce distinct c and z ion types compared with the b and y ions produced by CID. The electron transfer and dissociation, which cleaves the N−$C_\alpha$ bond, involve the formation of an aminoketyl radical and the backbone cleavage is believed to be less selective than CID with no strong cleavage preferences.[12] To date, several statistical studies have been published to examine the underlying fragmentation trends, e.g., Savitski and co-workers analyzed the pairwise fragmentation trends of ECD spectra of 14 967 tryptic peptide dications and found that the preference is complementary to CID;[13] Chalkley and co-workers characterized the frequency of observing different ion types in ETD in terms of protease used and charge states.[12] These studies have provided valuable information for understanding ETD mechanism as well as interpreting ETD spectra. However, no study has been done to examine the fragmentation trends for large data sets of ETD spectra using more advanced statistical techniques.

Our group has reported previously application of a statistical data mining strategy, penalized K-means clustering, to discover fragmentation patterns for CID,[3,14] and in the research reported here, we apply K-means clustering to ETD for fragmentation pattern discovery. Several ETD data sets collected by the Coon group at the University of Wisconsin-Madison, with sequences assigned to spectra by OMSSA, were subjected to analysis: one with 11 954 unique peptides produced by Lys-C digestion, one with 12 042 unique peptides produced by Glu-C digestion, and one with 6 423 high-resolution spectra of unique peptides produced by trypsin digestion. Mass spectra for all

**Received:** September 1, 2011
**Accepted:** October 24, 2011
**Published:** October 24, 2011

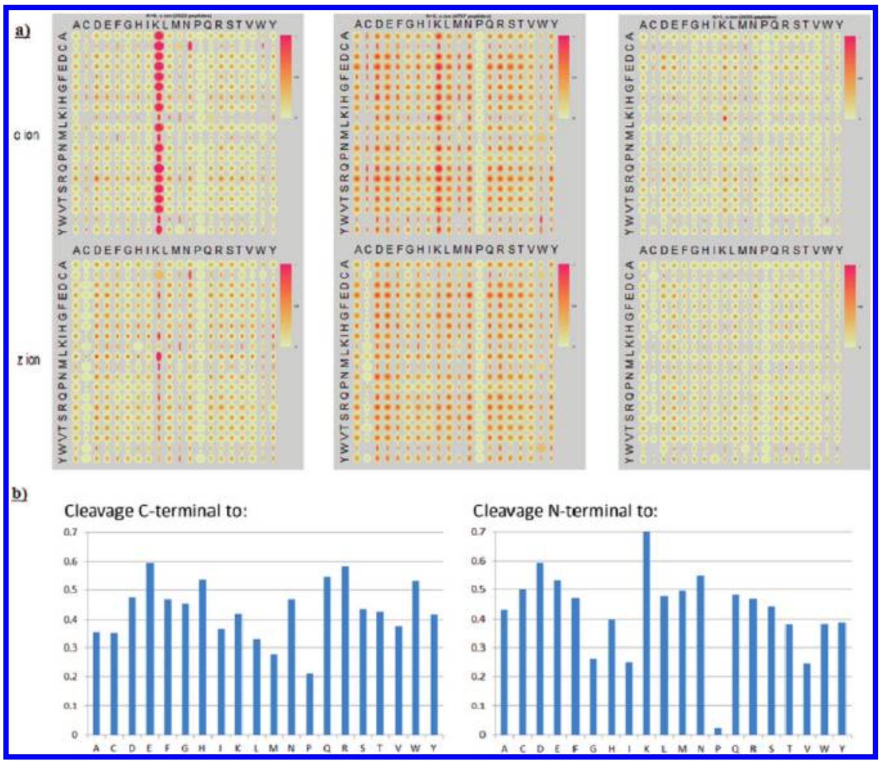


**Figure 1.** (a) Quantile maps for the three clusters obtained by K-means clustering of 11 954 spectra of Lys-C digested unique peptides. Two quantile maps are plotted for each cluster, one for c ions (top) and one for z ions (bottom). (b) Quantified cleavage preference in cluster 2. The left graph of b represents the cleavage C-terminal to a certain residue and the right graph represents the cleavage N-terminal to a certain residue. Cleavage preference (*y* axis) is represented by the probability for a certain amino acid pair to have strong cleavages (ref 9 for detailed calculation), e.g., cleavages C-terminal to E, H, N, Q, R, and W are relatively strong, and cleavages N-terminal to G, I, and V are rather weak.

**Table 1. Peptide Charge and Length Distribution for Lys-C Digestion ETD Data Set and Corresponding Clusters**

| | | all (11 954 peptides) | cluster 1 (3 522 peptides) | cluster 2 (4 737 peptides) | cluster 3 (3 695 peptides) |
|---|---|---|---|---|---|
| charge | 2 | 14% | 36% | 3% | 8% |
| | 3 | 45% | 38% | 60% | 33% |
| | 4 | 29% | 21% | 29% | 36% |
| | 5 and more | 12% | 5% | 9% | 22% |
| average charge | | 3.4 | 3.0 | 3.5 | 3.7 |
| average length | | 17 | 14 | 14 | 21 |
| sequence with internal Lys | | 28% | 13% | 31% | 38% |
| sequence with internal Arg | | 63% | 54% | 66% | 66% |
| fragmentation patterns | | N/A | very strong X-K cleavage ($C_{n-1}$ ion) | moderate cleavage for selected residue pairs | no cleavage preference |

data sets were obtained using an LTQ-Orbitrap (Thermo Fisher Scientific, San Jose, CA) to achieve high resolution and high mass accuracy; MS/MS of the Lys-C and Glu-C data sets were measured using the LTQ front end of the instrument (low resolution) and the tryptic data set was obtained by using the orbitrap as a high-resolution analyzer for product ions. The normalized fragment intensity for cleavage at each amino acid pair was extracted from each spectrum. As an example, c and z ions were identified from the spectrum of the $MH_2^{2+}$ ion of the peptide AAEDVAK and were then normalized to the most abundant peak among all c and z ions in that spectrum (highly charged fragments will also be included depending on the precursor ion charge). For c ions, the normalized intensities of c1, c2, c3, c4, c5, and c6 ions were associated with AA pairs A-A, A-E, E-D, D-V, V-A, and A-K, respectively, which correspond to the cleavage sites. After the information was collected for all the spectra in the data set, a matrix was created for c ions containing 400 AA combinations (20 AA × 20 AA; all cysteines in these data sets are carbamidomethylated, so "Cys" in this report are actually carbamidomethylated Cys), and each combination includes a number of normalized intensity values. The same procedure was performed for z ions and both c and z data were used together for clustering. The relationship between AA pairs and normalized intensity can be visualized by quantile maps[15] as shown in Figure 1, in which the left column represents the N-terminal residue of the pairwise cleavage site and the top row represents the C-terminal residue of the pair of cleaving amino acids. The horizontal dimension of each spot

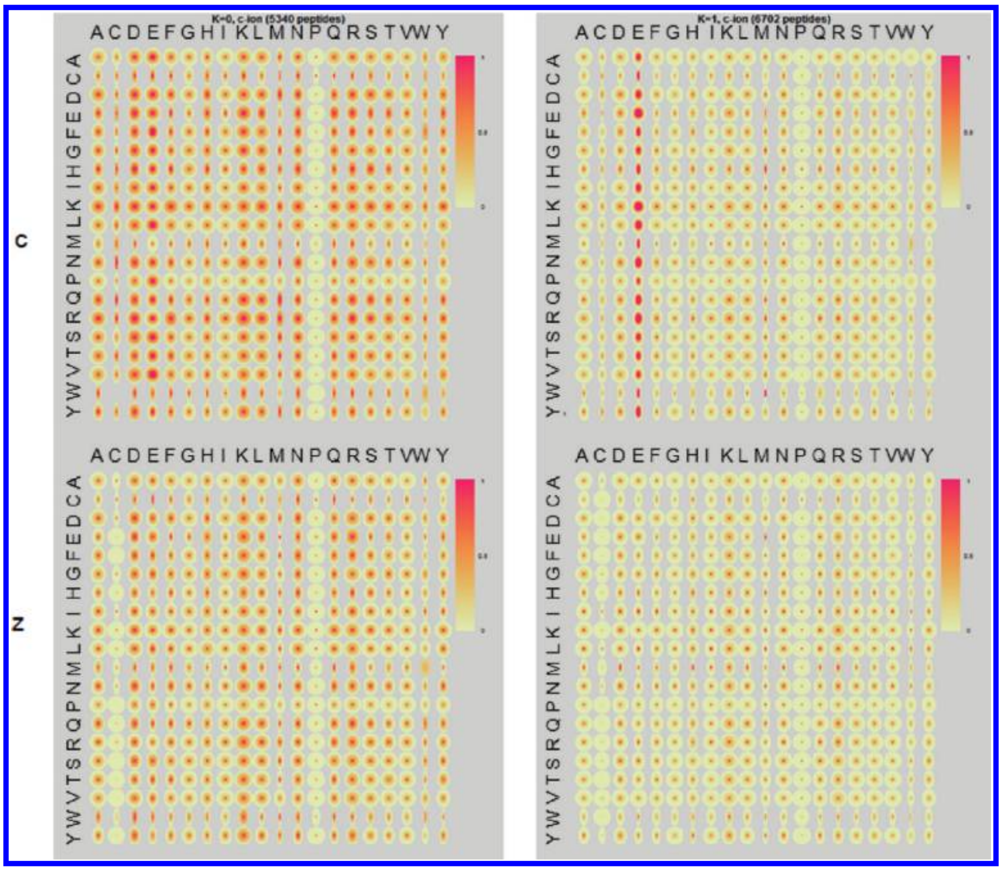


**Figure 2.** Quantile maps for the two clusters obtained by K-means clustering from 12 042 spectra of Glu-C digested unique peptides. Two quantile maps are plotted for each cluster, one for c ions (top) and one for z ions (bottom).

**Table 2. Peptide Charge and Length Distribution for Glu-C Digestion ETD Data Set and Corresponding Clusters**

| | | all (12 042 peptides) | cluster 1 (5 340 peptides) | cluster 2 (6 702 peptides) |
|---|---|---|---|---|
| charge | 2 | 1% | 0% | 1% |
| | 3 | 45% | 32% | 56% |
| | 4 | 43% | 57% | 33% |
| | 5 and more | 10% | 11% | 10% |
| average charge | | 3.6 | 3.8 | 3.5 |
| average length | | 16.2 | 15.3 | 17 |
| sequence with internal E | | 38% | 33% | 42% |
| fragmentation patterns | | N/A | moderate cleavage for selected residue pairs | very strong cleavage at X-E |

is proportional to the number of instances of a given pair, with a wider spot meaning more occurrences and thus higher confidence. In each spot, 10 quantiles of intensities of the entire distribution are plotted on circles using gradient colors. The darkness of the color represents the normalized intensity. A full dark spot represents high intensities for all occurrences in the distribution (e.g., A-K c-ion in the upper left cluster of Figure 1). A spot with a small dark dot in the center and white in the surrounding area represents a bimodal distribution (i.e., a portion of cleavage intensities are high (dark center) but others are low (lighter colored surrounding area), e.g., A-H c-ion in the upper left cluster of Figure 1).

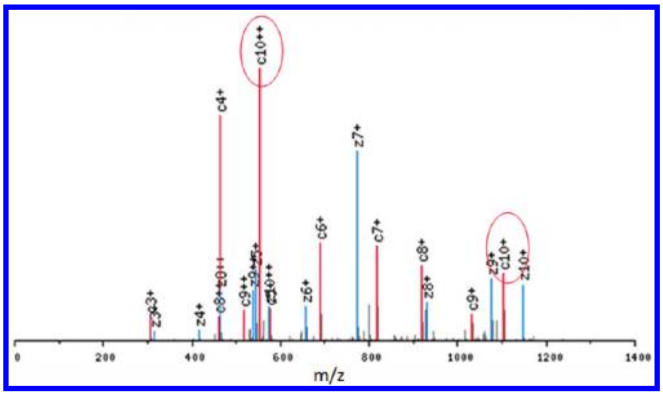


**Figure 3.** ETD spectrum of 3+ AAFRNIKTIAE. A strong $c_{10}^{2+}$ ion was observed which corresponds to the cleavage before the C-terminal residue Glu. Complementary $c_4/z_7$ ions correspond to the cleavage between Arg and Asn.

The K-means clustering algorithm partitions all spectra into K clusters based on the pairwise cleavage behaviors, with the principle that the peptides within a given cluster fragment as similarly as possible to each other and as differently as possible from those peptides in other clusters. More specifically, each peptide spectrum is plotted in a 400 dimensional space with each dimension representing the cleavage intensity from a certain AA combination; then the space is tentatively and repeatedly separated into K parts until the sum squared distance of each spectrum to its centroid is minimized. This approach allows

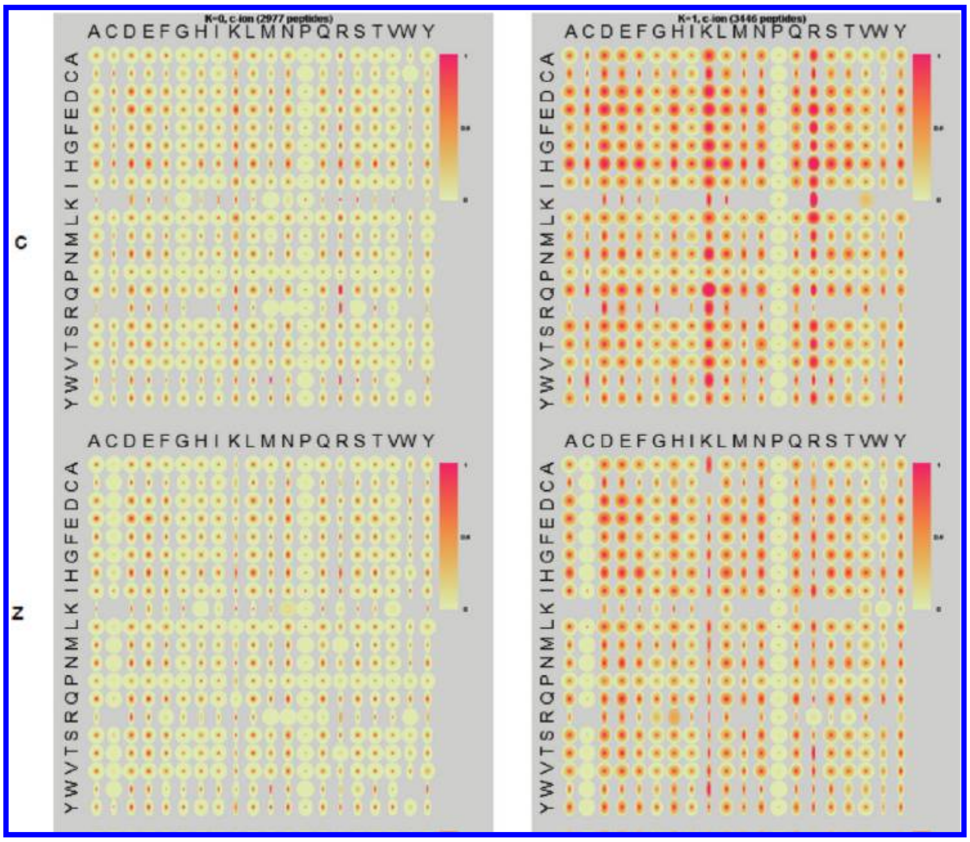


**Figure 4.** Quantile maps for the two clusters obtained by K-means clustering of 6 423 high-resolution spectra of unique tryptic peptides.

the extraction of independent patterns that were previously mixed. One drawback is that one must choose the number of clusters "K". In this work we produce multiple sets of clusters and choose the optimal K that produces distinct clusters without obvious subclustering. After the clustering, a CART (Classification And Regression Tree) program is used to extract sequence features (charge, length, mobile proton, etc.) for each cluster, so that the relationship between the sequence features and fragmentation behaviors can be established. A complete list of features considered can be found in the Supporting Information.

## ■ LYS-C DIGESTION

Clustering of the Lys-C digested peptides resulted in three clusters with distinct fragmentation behaviors (Figure 1a): (1) cluster 1, a cluster with extremely strong cleavage N-terminal to Lys (the majority are $c_{n-1}$ ions with "$n$" indicating the total number of residues; only 13% of the peptides have internal Lys); (2) cluster 2, a cluster with moderate cleavage preference for certain residues (see middle cluster of Figure 1 and detailed cleavage probability in Figure 1b); and (3) cluster 3, a cluster with more uniform cleavages. CART analysis showed that the separation mainly depends on charge and length (Table 1). Peptides in cluster 1 are lower charged and shorter, with 36% doubly charged peptides, 38% triply charged peptides, and an average length of 14. Cluster 2 peptides are the same length as those in cluster 1 but have higher average charge (3.5 versus 3.0). Peptides in cluster 3 are longer and more highly charged, with an average length of 21 and average charge of 3.7. With the consideration of the fragmentation patterns in each cluster, it can be seen that the backbone cleavage selectivity decreases with increasing charge states and length. This may indicate that the selective cleavage is charge or radical directed. For a lower charged Lys-C peptide, preferred charge locations may fragment with higher priorities; as the charge increases, there are more charged locations and thus more cleavable sites along the peptide backbone, so that the selectivity decreases.

Besides the features mentioned above, it is also observed that cleavage N-terminal to Pro is prohibited (notice the light color of the Pro column in the clusters and the low intensity in Figure 1b right), which is expected because the ring structure of Pro prohibits peptide cleavage even if the N−C$\alpha$ bond cleaves. In addition, the z ions from the cleavage N-terminal to carbamidomethylated Cys are generally missing due to a neutral loss of 90 from the side chain.[16,17] When the loss of 90 is considered, the missing z ion column can be recovered. This phenomenon suggests that ETD search engines should use the mass corresponding to a neutral loss of 90 when cleavage occurs N-terminal to carbamidomethylated Cys.

## ■ GLU-C DIGESTION

Glu-C digested peptides, which are mainly triply and quadruply charged, separated into two main clusters of behaviors. The first cluster (5 340 peptides, Figure 2 left) shows moderate cleavage preferences at various locations, which is similar to cluster 2 of the Lys-C digested peptides. The other distinct cluster (6 702 peptides, Figure 2 right) shows very strong cleavage N-terminal to Glu. Though 42% of these peptides have internal Glu, 98% of these X-E cleavages are $c_{n-1}$ ions involving no internal Glu. Table 2 summarizes the charge and length

Table 3. Peptide Properties for Tryptic ETD Data Set and Corresponding Clusters

| | | all (6 423 peptides) | cluster 1 (2 977 peptides) | cluster 2 (3 446 peptides) |
|---|---|---|---|---|
| charge | 2 | 0% | 0% | 0% |
| | 3 | 79% | 68% | 89% |
| | 4 | 18% | 27% | 11% |
| | 5 and more | 3% | 5% | 0% |
| average charge | | 3.2 | 3.4 | 3.1 |
| average length | | 16.4 | 19.6 | 13.6 |
| Lys ending | | 60% | 58% | 61% |
| Arg ending | | 40% | 41% | 39% |
| sequence with internal Lys | | 10% | 10% | 9% |
| sequence with internal Arg | | 15% | 15% | 12% |
| fragmentation patterns | | N/A | no cleavage preference | strong X-K and X-R cleavage |

distributions for the separation. It can be seen that the cluster with strong X-E cleavages (Figure 2, right) are relatively lower in charge (3.5 versus 3.8) but a little longer (17 versus 15). Figure 3 is an ETD spectrum showing an example of the enhanced cleavage N-terminal to Glu. This looks very similar to X-K cleavage in the Lys-C data set, with both X-K and X-E cleavage generating strong $c_{n-1}$ ions.

## TRYPSIN DIGESTION

The final spectral data set subjected to clustering corresponds to 6423 unique tryptic peptides. All peptides in this high-resolution tryptic data set have three or more charges, with 79% triply charged and 19% quadruply charged peptides. Two clusters were achieved through clustering: (1) a cluster with uniform cleavages; (2) a cluster with moderate cleavage preferences at various locations, including strong cleavage at the N-terminus of Lys and Arg in the c ions. As expected from the cleavage patterns in Figure 4, CART analysis (Table 3) shows that peptides in the first cluster are generally longer (20 versus 14) and slightly more highly charged (3.4 versus 3.1), while cluster 2 peptides are shorter and lower charged. The low percentage of internal Lys and Arg strongly indicates the preference for the $c_{n-1}$ ion, which is in agreement with the observation in Lys-C and Glu-C data sets. Note that strong preferential cleavage at Arg is seen only in this data set, where Arg (or Lys) occupies the C terminal positions.

Results from the three data sets indicate that $c_{n-1}$ is a preferred cleavage site for Lys-C, Glu-C, and tryptic peptides but cannot indicate whether the preference is simply due to a position effect or the fact that the peptides are ending with the specific basic and acidic residue Lys, Arg, and Glu. To clarify the issue, we analyzed the spectra of 550 peptides that do not end with Lys, Arg, and Glu. These are nonspecifically cleaved peptides from the Lys-C, Glu-C, and trypsin data sets. Figure 5 shows the distributions of 550 peptides: the $c_{n-1}$ ion was found to be the most intense peak among all c ions. It can be clearly seen that the cleavage intensity decreases as the distance from the C-terminus increases, and the $c_{n-1}$ ion is significantly stronger than the other c ions. This observation unequivocally indicates that the cleavage preference in ETD is highly affected by the residue position, which is possibly determined by the gas phase precursor structure as suggested by Moss and co-workers using model peptides.[18] As the charge increases, the structures of the peptides will change, and this position effect will diminish (Figure 5b), in agreement with the Lys-C clustering results. Although each cluster with a strong $c_{n-1}$ cleavage (first cluster of Figures 1a, and second cluster of Figures 2 and 4) contains some 4+ ions, which are not significantly intense, the overall cluster pattern is dominated by the greater percentage of and extremely strong $c_{n-1}$ cleavage, exhibited by the precursors with three charges. Note that the observed position effect might also be, at least partially, a result of the increasing probability for the c ion to hold charge as length and basicity increase; however, in this report, we simply use position to describe this phenomenon.

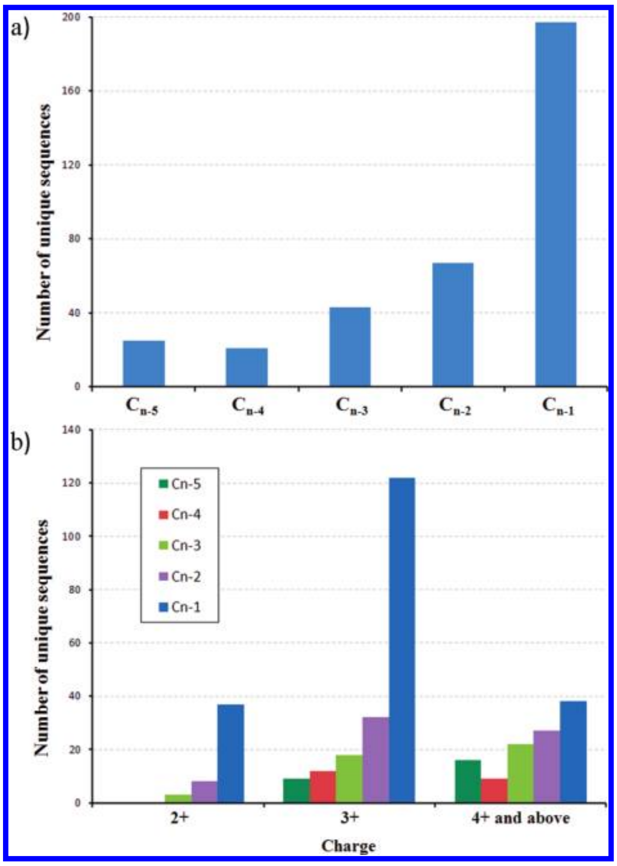


Figure 5. (a) Bar graph indicating the numbers of unique sequences (550 total, not ending with K, R, and E) in which the $c_{n-i}$ ion is the most intense peak among all c ions ($n$ is the peptide length, $i$ means the $i$th residue to the C-terminal of the peptide). (b) Distributions in terms of charge states.

Table 4. Summary of Observed ETD Fragmentation Patterns

| fragmentation pattern | peptide property |
|---|---|
| strong $c_{n-1}$ ion | lower charged |
| moderate preference, e.g., enhanced: E, H, N, Q, R, W-X, suppressed: X-G, I, V | lower charged |
| no obvious cleavage preference | higher charged |
| no cleavage for X-P (ring) and z ion from X-(C + 57) (neutral loss) | all peptides |

Table 4 summarizes the ETD fragmentation patterns observed by applying the K-means clustering method to Lys-C, Glu-C, and tryptic data sets. The patterns highly depend on charge state. At higher charges states, the cleavage is less selective with no dominant preferential cleavage. At lower charge states, there are very strong $c_{n-1}$ ions and moderate preferred cleavages involving certain residues, such as enhanced cleavages C-terminal to E, H, N, Q, R, and W, and suppressed cleavages N-terminal to G, I, and V. Though these trends are not phenomenal enough to be unequivocally described as dominating cleavages, many of them can also be observed in the ECD statistics published previously.[13] In addition, limited cleavage occurs to the N-terminus of Pro, which is expected due to the ring structure, and the z ions from the N-terminal cleavage of carbamidomethylated Cys are always missing, due to the neutral loss of 90 from the side chain.[16,17] We also examined the hydrogen transfer products in ETD, and the data are shown in the Supporting Information. Strong c-1 radical ions, formed after hydrogen transfer, are also observed corresponding to cleavage N-terminal to Lys and Glu, for Lys-C and Glu-C peptides. All these patterns could be used directly for ETD fragment intensity prediction, and, at the same time, provide guidance to clarify the underlying dissociation mechanisms. The results will be incorporated into our intensity based algorithm, SQID,[9] to improve ETD peptide identification.

## ■ ASSOCIATED CONTENT

**S Supporting Information.** Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION

### Corresponding Author

*E-mail: vwysocki@email.arizona.edu.

## ■ ACKNOWLEDGMENT

The research was supported by NIH Grant R01GM051387 to V.H.W. and Grant R01GM080148 to J.J.C.

## ■ REFERENCES

(1) Wysocki, V. H.; Tsaprailis, G.; Smith, L. L; Breci, L. A. *J. Mass Spectrom.* **2000**, *35* (12), 1399–406.

(2) Huang, Y.; Triscari, J. M.; Tseng, G. C.; Pasa-Tolic, L.; Lipton, M. S.; Smith, R. D.; Wysocki, V. H. *Anal. Chem.* **2005**, *77* (18), 5800–5813.

(3) Huang, Y.; Tseng, G. C.; Yuan, S.; Pasa-Tolic, L.; Lipton, M. S.; Smith, R. D.; Wysocki, V. H. *J. Proteome Res.* **2008**, *7* (1), 70–79.

(4) Gibbons, F. D.; Elias, J. E.; Gygi, S. P.; Roth, F. P. *J. Am. Soc. Mass Spectrom.* **2004**, *15* (6), 910–2.

(5) Havilio, M.; Haddad, Y.; Smilansky, Z. *Anal. Chem.* **2003**, *75* (3), 435–444.

(6) Zhang, Z. *Anal. Chem.* **2004**, *76* (14), 3908–3922.

(7) Elias, J. E.; Gibbons, F. D.; King, O. D.; Roth, F. P.; Gygi, S. P. *Nat. Biotechnol.* **2004**, *22* (2), 214–9.

(8) Narasimhan, C.; Tabb, D. L; Verberkmoes, N. C.; Thompson, M. R.; Hettich, R. L.; Uberbacher, E. C. *Anal. Chem.* **2005**, *77* (23), 7581–93.

(9) Li, W.; Ji, L.; Goya, J.; Tan, G.; Wysocki, V. H. *J. Proteome Res.* **2011**, *10* (4), 1593–1602.

(10) Syka, J. E. P.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 9528–9533.

(11) Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. W. *J. Am. Chem. Soc.* **1998**, *120*, 3265–3266.

(12) Chalkley, R. J.; Medzihradszky, K. F.; Lynn, A. J.; Baker, P. R.; Burlingame, A. L. *Anal. Chem.* **2010**, *82* (2), 579–84.

(13) Savitski, M. M.; Kjeldsen, F.; Nielsen, M. L.; Zubarev, R. A. *Angew. Chem., Int. Ed.* **2006**, *45*, 5301–5303.

(14) Tseng, G. C. *Bioinformatics* **2007**, *23*, 2247–2255.

(15) Tseng, G. C. *Comput. Stat. Data Anal.* **2010**, *54*, 1124–1137.

(16) Sun, R.; Dong, M.; Song, C.; Chi, H.; Yang, B.; Xiu, L.; Tao, L.; Jing, Z.; Liu, C.; Wang, L.; Fu, Y.; He, S. *J. Proteome Res.* **2010**, *9* (12), 6354–6367.

(17) Xia, Q.; Lee, M. V.; Rose, C. M.; Marsh, A. J.; Hubler, S. L.; Wenger, C. D.; Coon, J. J. *J. Am. Soc. Mass Spectrom.* **2011**, *22* (2), 255–264.

(18) Moss, C. L.; Chung, T. W.; Cerovsky, V.; Turecek, F. *Collect. Czech. Chem. Commun.* **2011**, *76* (4), 295–309.