research articles Journal of proteome research

A Data-Mining Scheme for Identifying Peptide Structural Motifs Responsible for Different MS/MS Fragmentation Intensity Patterns

Yingying Huang,[‡,§] George C. Tseng,[⊥] Shinsheng Yuan,[||,†] Ljiljana Pasa-Tolic,[#] Mary S. Lipton,[#] Richard D. Smith,[#] and Vicki H. Wysocki*[,‡]

Department of Chemistry, University of Arizona, Tucson, Arizona 85721, Department of Biostatistics and Department of Human Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, Department of Statistics, University of California, Los Angeles, California 90095, and Pacific Northwest National Laboratory, Richland, Washington 99352

Received February 26, 2007

Although tandem mass spectrometry (MS/MS) has become an integral part of proteomics, intensity patterns in MS/MS spectra are rarely weighted heavily in most widely used algorithms because they are not yet fully understood. Here a knowledge mining approach is demonstrated to discover fragmentation intensity patterns and elucidate the chemical factors behind such patterns. Fragmentation intensity information from 28 330 ion trap peptide MS/MS spectra of different charge states and sequences went through unsupervised clustering using a penalized K-means algorithm. Without any prior chemistry assumptions, four clusters with distinctive fragmentation patterns were obtained. A decision tree was generated to investigate peptide sequence motif and charge state status that caused these fragmentation patterns. This data-mining scheme is generally applicable for any large data sets. It bypasses the common prior knowledge constraints and reports on the overall peptide fragmentation behavior. It improves the understanding of gas-phase peptide dissociation and provides a foundation for new or improved protein identification algorithms.

Keywords: data mining • cluster analysis • K-means algorithm • penalized K-means algorithm • CART • statistical analysis • quantile map • peptide • MS/MS • intensity • ion trap • CID • fragmentation pattern • dissociation pattern • pairwise cleavage • Xxx-Zzz • cleavage pair • Fisher information • Pro • Gly • Asp • Glu • Arg • Lys

Introduction

Tandem mass spectrometry with collision-induced dissociation (CID) has become one of the most powerful techniques in proteomics because of its unparalleled sensitivity and robustness in identifying thousands of proteins from small quantities of complex samples.[1] Computer algorithms make it possible to derive useful information from the enormous amounts of data acquired from practical studies.[2] Different algorithms have different scoring methods to compare the likelihood of a certain peptide sequence candidate matching a given spectrum. As different as these approaches may appear, current readily available algorithms either totally ignore or only minimally take into account the different chemical properties of the side chains of the 20 amino acid (AA) residues. While algorithms using this overly simplified random cleavage model are able to identify a significant number of peptides, many other peptides cannot be identified. Even though intensity patterns of the fragment ions from a given peptide under the same experimental settings are highly reproducible, current readily available algorithms do not usually take advantage of this available information because these patterns are not very well understood. Thus, if one could understand at the molecular level how and why different peptide fragmentation patterns exist, one might make algorithms that better identify the correct sequence candidate by using the fragment ion intensities.

Earlier research by the authors and others on peptide fragmentation mechanisms relied on model peptides and focused on several specific residues including aspartic acid (Asp or D),[3,4] histidine (His or H),[5] and proline (Pro or P).[6] These studies provided a framework[7] as well as chemical "rules" used in sorting larger sets of acquired spectra in more recent investigations.[8–15] While these analyses offer insight into the fragmentation process, they are constrained by the chemical assumptions on which they are based and do not provide a complete overview. The number of spectra used by various data-mining approaches is often limited by the availability of correctly identified spectra, which are mostly from doubly charged tryptic peptides,[9–11,13,14] and the specific chemistry mechanism under study.[8,10,12,13,15] Because the chemical factors behind the dissociation process are so complex, many studies had to simplify the problem by focusing on just those peptides that contain very specific structural motifs and ignore others

[*] To whom correspondence should be addressed. E-mail: vwysocki@email.arizona.edu.

[‡] University of Arizona.

[§] Author is currently with Thermo Fisher Scientific in San Jose, CA 95134.

[⊥] University of Pittsburgh.

[||] University of California.

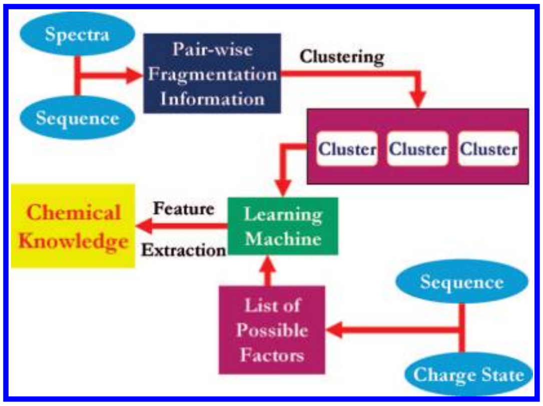
[†] Author is currently with Institute of Statistical Science, Academia Sinica, 128, Academia Rd. Sec. 2, Taipei 115, Taiwan.

[#] Pacific Northwest National Laboratory.

10.1021/pr070106u CCC: $37.00

Published on Web 12/04/2007

Scheme 1. Data-Mining Scheme to Identify Peptide Structural Motifs Responsible for Different MS/MS Fragmentation Intensity Patterns



that may contain "outliers" or odd behavior.[8,11,12] Even in the studies of peptides with very specific structural motifs using either model peptides or larger sets of acquired spectra, great variability of the fragmentation statistics is still common.[8,11,13,15] A description of the overall fragmentation behavior of peptides from all charge states and without restrictions on the residue content is not yet available. Our current understanding of unimolecular dissociation of peptides is still not sufficient to fully predict what a CID fragmentation spectrum will look like given the sequence and charge state.[16,17]

Reported here is a systematic knowledge mining scheme shown in Scheme 1 to (1) bypass constraints and obtain an overview of the fragmentation intensity patterns from peptides of all kinds, (2) understand the chemistry behind these patterns, and (3) provide a solid foundation for algorithms to use the fragment ion intensity from an unknown MS/MS spectrum to help identify the correct sequence candidate with high confidence. This data-mining scheme mainly involves two steps: cluster analysis and feature extraction. Cluster analysis[18] utilizes clustering techniques to find clusters of spectra with similar dissociation patterns without prior chemical assumptions. Feature extraction analyzes the corresponding sequences and charge states of the resulting clusters of peptides and inputs them to a learning machine to identify the structural motifs within the clusters. This study represents our effort to bypass the prior knowledge constraints: Instead of first sorting the spectra into subsets based on the several known chemical factors that may influence spectra, then looking for differences in the resulting patterns, we start searching for the different patterns directly, allowing those spectra sharing the same patterns to gather together, and then let the patterns lead us to the chemical motifs. Through this data-mining scheme, a detailed overview of the peptide fragmentation behaviors of a large spectral data set can be obtained, along with elucidation of the interplay between the underlying chemical factors that lead to different fragmentation patterns.

The heterogeneity in the data set creates great challenges for such a data-mining study. For cluster analysis, the first major challenge is the great variability of length and residue content of the peptides. Some examples of such variability are that the same residue may be repeated many times in one peptide while many others residues are missing and that two peptides may have no residues in common. The second major challenge comes from the low-mass and high-mass cutoffs in the experimental spectral acquisition. While the high-mass cutoff is fixed at *m/z* 2000 for our data set, the low-mass cutoff varies with the precursor ion *m/z*, because the data are acquired from traditional 3D ion traps.[19] Therefore the fragmentation information for each peptide bond in each peptide from our spectra, as in many practical proteomics data sets, is often not complete. This inevitable drawback of ion trap CID data leads to massive missing information from a statistical standpoint. In addition, the low-mass resolution of ion trap data that leads to ambiguous peak assignment will also make accurate assessment of the intensity patterns more difficult. Therefore, it is most challenging to find a working model for cluster analysis distance measurements (a determination of the degree of similarity between two spectra, or one spectrum and one cluster) that can tolerate all the heterogeneity in the data but remain sensitive enough to probe the different existing patterns. For feature extraction (e.g., identification of structural motifs within each cluster), the same challenge exists because of the the variability in peptide length and residue content that leads to the "nonstandard chemical features", meaning that the features may not exist in every peptide, and when they do, they may have great variability. Additionally, these features usually intertwine and our knowledge is limited on how they affect each other when multiple features are present. In order to overcome all of the above challenges, it is crucial to find powerful statistical methods and good models to apply the methods.

Methods

Cluster Analysis: Penalized K-Means. a. Data and Dissimilarity Measure. This study used the same 28 330 ion trap peptide MS/MS spectra of unique charge state and sequence described in a previous publication.[8] Of these, 25.3% are singly charged, 62.3% are doubly charged, and 12.3% are triply charged. The lengths of the peptides range from 5 to 55 AA residues, with the median and average both at 16 residues. The majority (94.7%) of these peptides end in either Arg or Lys, and 65.9% of them do not have any internal Arg or Lys. The pairwise fragmentation intensity map from b^+ and y^+ ions (also described in a previous publication[8]) is defined as the fragmentation pattern for one or a set of spectra. For example, singly charged b and y ions are first identified from the spectrum of 2+ peptide AAEDVAK and are then normalized to the most abundant peak among all b^+ and y^+ ions in that spectrum. Then the normalized intensities for b ions and y ions are cataloged by the pair of AA residues at the fragmentation site (A-A, A-E, E-D, D-V, V-A, A-K). By this method, the b^+ and y^+ information from each spectrum in our 28 330 spectral set is represented by a matrix that contains 800 variables (20 AA × 20 AA × 2 ion types: b's and y's), but only a few of the matrix cells contain values. This matrix of 800 variables is used as a representation of the fragmentation pattern for a given peptide. Similarly, the fragmentation pattern from a group of peptides can be represented by a matrix of 800 histograms (bar graphs of probability vs intensity of cleavage), and these results can be visualized by the quantile maps presented in Figure 1 (see Methods section d and Results section).

The goal of our cluster analysis is to partition all spectra into K clusters (sets of spectra) based on the fragmentation patterns in the spectra. The desired output is that the spectra within a given cluster display fragmentation patterns as similar as possible to each other and as different as possible from those spectra in other clusters. Therefore, the first step in the cluster analysis is to define the distance between two spectra or one

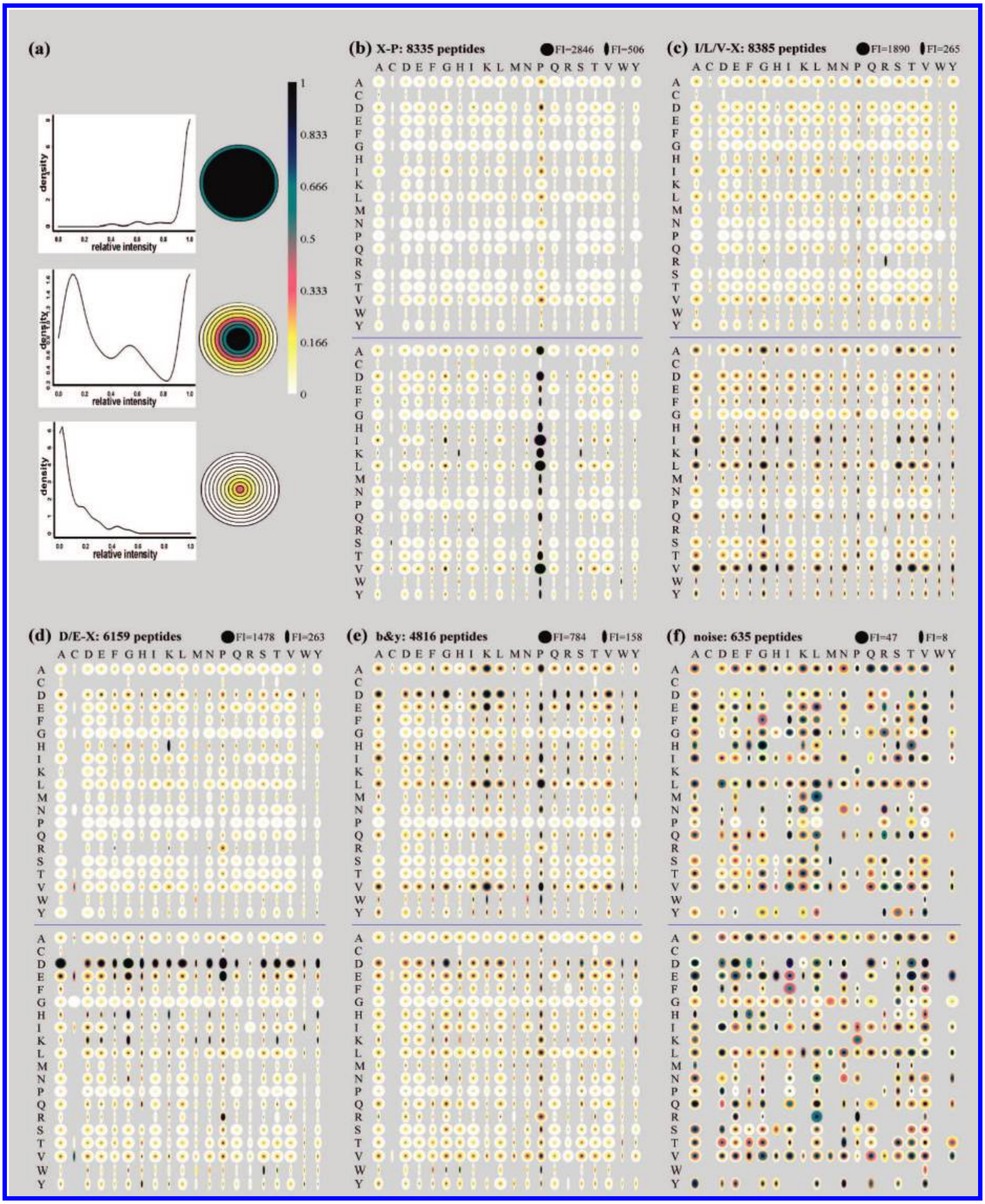


Figure 1. Quantile maps for the five clusters obtained by penalized K-means from 28 330 spectra of unique sequence and charge state. Two quantile maps are plotted for each cluster, one for singly charged **b** ions (top) and one for singly charged **y** ions (bottom). Single letter codes of the N-terminal residue Xxx of a cleavage pair are listed along the left-most columns and those of the C-terminal residue Zzz are listed along the top rows. Color represents intensity of cleavage, and the horizontal dimension of each circle represents the Fisher information of such pairwise cleavages, with illustrations shown at the top right corners. See Methods for details. (a) Three examples of quantile maps, their corresponding density functions, and the color gradient; (b) cluster **X-P**; (c) cluster **I/L/V-X**; (d) cluster **D/E-X**; (e) cluster **b&y**; (f) noise (outlier) cluster

spectrum and one cluster based on the dissimilarity between their fragmentation patterns. Intuitively, distance is a value that measures the similarity of a spectrum or a cluster to a characterized cluster of spectra, with smaller numbers indicating greater similarity. If, for example, a known cluster shows strong cleavage at every Xxx–Zzz pair of amino acids and a test spectrum shows dominant cleavage at one pair of residues such as Xxx–Pro, the “distance” between the test spectrum and the known spectrum will be large. Euclidean distance (i.e., sum of squares of the differences in each dimension) between two given peptide spectra is often not computable because each peptide spectrum is represented as an 800-dimensional observation, but only 1%–5% of the dimensions (fragmentation intensities for a given pairwise Xxx–Zzz cleavage) have values and these values often do not overlap. However, the distance between a peptide and a large set (cluster) of peptides is usually computable. For a large set (e.g., >500) of randomly selected peptides, the means of all 800 dimensions are usually observable. The distance of a peptide to a set of peptides can then be defined as the Euclidean distance of the one peptide to the mean of the peptide set (see Supporting Information 1 for formulation details). An adjusted Euclidean distance is used, which accounts for missing values but does not penalize them. Only cleavage pairs observed in both patterns from the same ion type can contribute to the distance calculation. The distance is then normalized to account for variable numbers of contributing cleavage pairs. In a similar manner, the distance between two large clusters can also be calculated.

b. K-Means vs Penalized K-Means. Once distances between fragmentation patterns are defined, cluster analysis can be performed to group peptides that have similar fragmentation patterns. Due to the fact that the distance between two peptides is often unmeasurable, many popular clustering algorithms that require computation of pairwise distances, such as hierarchical clustering or self-organizing maps, are inapplicable. The K-means clustering algorithm[20] only necessitates calculation of distances of a peptide to a cluster of peptides (i.e., the cluster center) and can assign data points into a predefined number (K) of clusters such that the sum of the squared distances of each point to its cluster center is minimized.

The K-means clustering algorithm has been widely applied in many scientific fields.[20] A major deficiency of K-means is that it assigns all data into clusters and does not allow peptide spectra to be excluded (treated as "noise" without being clustered). This is particularly an issue for our data because of the existence of a few spectra that are dissimilar from the majority of the spectra: Some spectra only have pairwise fragmentation intensity values recorded for a very limited number of cleavage pairs because they have shorter sequences (e.g., less than 7 AA residues) or incomplete fragmentation information because of the m/z cutoffs as described in the Introduction. Therefore, it is better to exclude these spectra, which will be defined as "noise" spectra, from clusters to avoid dilution or weakening of the cluster patterns by inclusion of spectra that are not similar in fragmentation to the majority of spectra in the cluster. To achieve this, an extended form of K-means, penalized K-means,[21] is applied, because it has better tolerance for noise by adding a tuning parameter λ. Spectra with squared distances to all cluster centers larger than λ (i.e., far away from all discovered cluster centers) are considered noise and will be assigned to the noise set. Intuitively, a smaller λ will produce tighter clusters, but more peptides will be assigned to the noise set. The selection of K and λ will be discussed in the next subsection. The detailed mathematical formulation of K-means and penalized K-means, as well as a test example comparing their effectiveness using two data sets known from previous study to have different fragmentation behaviors, are available in Supporting Information 1. It is shown by the test example that penalized K-means is more powerful and appropriate in assigning peptides into clusters by their fragmentation patterns than traditional K-means.

c. Selection of the Optimal K and λ A resampling technique similar to that used in Tibshirani and Walther,[22] or Tseng and Wong,[23] was used to select the optimal K and λ for penalized K-means. Given a fixed K and λ, clustering on all data was performed, and a cluster ID for each spectrum was obtained. This result is named "raw-clustering". After the raw-clustering, additional clustering called "subsample-judged clustering" was performed in which 10 subsamples were randomly generated such that each contains 70% of the spectra from the whole data set. Clustering on each of the 10 subsamples was performed, and the cluster ID for each spectrum was regenerated. The ideal selection of K and λ should give the same cluster ID for each spectrum in the raw-clustering and the 10 subsample-judged clusterings. To quantify this effect, a term called "clustering concordance rate" (fraction of spectra that fall into the same cluster in the raw-clustering and the 10 subsample-judged clustering steps) is calculated for various K and λ (see Supporting Information 2). The optimum K and λ are chosen as large as possible to obtain large but tight clusters, while the concordance rate remains high (i.e., 0.7–0.8). This subsample-judged clustering also can identify those peptides that display stable fragmentation patterns among the whole population. The collection of peptides with stable fragmentation patterns under the optimized K and λ, i.e., those that repetitively partition into the same cluster during raw-clustering and subsample-judged clustering, is most useful for the feature extraction step.

d. Visualization for the Clustering Result: Quantile Map. A visualization tool, called Quantile Map, was developed (written in C++) to present the fragmentation pattern from a given cluster of spectra.[8] Each Quantile Map is composed of a 20x20 matrix of small quantile maps that describe the distributions of cleavage intensities for individual AA pairs. Single letter codes of the N-terminal residue Xxx are listed along the left most columns, and those of the C-terminal residue Zzz are listed along the top rows. Figure 1a gives three examples of the quantile maps for certain Xxx–Zzz cleavages and their corresponding density functions. Ten concentric doughnuts showing the 5%, 15%, . . . , 95% quantiles (from outside inward) of the distribution are plotted for a certain Xxx–Zzz cleavage. Strengths of intensities are represented by the gradient of color shown on the right of Figure 1a. As shown in the bottom of Figure 1a, a uniform weak cleavage at a particular Xxx–Zzz pair (based on the total number of instances of this cleavage from the whole data set) will result in a light round spot if there are a large number of instances of this cleavage. An example of such a uniform weak cleavage would be Ala–Ala (A-A) cleavage in the y-ion (bottom) maps of Figure 1b and 1d. A uniformly strong cleavage at a particular Xxx–Zzz (top of Figure 1a) will result in a dark round spot if there are a large number of instances of this Xxx–Zzz cleavage in the data set (e.g., the Ile-Pro (I-P) y-ion cleavage (bottom) in Figure 1b). For a given cluster, two Quantile Maps, one for **b** ions (top) and one for **y** ions (bottom) are plotted. In some cases where the distribution of cleavage intensities for a given AA pair contains statistically insignificant information (e.g., large variation and small number of observations), the visual effect of such cell is reduced by adjusting its horizontal width. A statistical threshold called the Fisher information (FI) threshold was employed to determine whether cleavage abundance information for a particular residue pair should be shown when the number of occurences of that pair is limited. Fisher information of the distribution in each cell is computed and corresponds to the number of instances n of a particular residue combination divided by the variance of the fragmentation intensity. (See Supporting Information 3.) The horizontal width of the circles is proportional to their corresponding Fisher information. (Thus for a given cell, *color* represents the distribution of the cleavage intensity and the *width* represents the significance of this observation, with a larger *n* and a smaller variance resulting in a larger FI.) A dark vertical ellipse, for example, represents strong uniform cleavage at that particular Xxx–Zzz pair but relatively few instances of that particular Xxx–Zzz in the data set. For each cluster, the Fisher information at 95% quantile and 5% quantile are illustrated on the top right corner of its b-ion quantile map. These values serve as scaling factors to assist the reader in understanding the plots and scale with the statistical significance of the observed cleavage intensity of a particular residue pair indicated in the maps. For example, for Ile-Pro (I-P) or Ala-Pro (A-P) or Pro-Ala (P-A) or any other residue pair indicated by a full circle in the y-ion map of Figure 1b, a large number of examples of those cleavages exist in the data set, with FI approximately 2846. For Ala-Cys (A-C) or Cys-Ala (C-A) or Glu-Cys (E-C) or any other residue pair indicated by a thin ellipse in the y-ion map of (b), a smaller number of examples of that cleavage exist in the data set, with FI of approximately 506. A missing ellipse (just gray background) represents an Xxx–Zzz pair for which there were not enough instances of a particular Xxx–Zzz cleavage to give statistically meaningful results, i.e., the data did not meet the FI threshold.

Feature Extraction: CART. After clustering spectra by penalized K-means, feature extraction is performed to identify important chemical motifs and their priorities in determining how peptides fragment in the gas phase, with the implication of underlying chemical mechanisms. The purpose of this feature extraction step is different from a general classification problem (supervised machine learning), where classification error rate is the main concern and weighing factors are often used on different features in order to obtain the optimum result. Among many methods considered, classification and regression tree (CART)[24,25] with a suitably chosen feature space appears to be the most appropriate method.

The original CART is a greedy algorithm, in which the algorithm loops through each variable, searching for the best cut-point that divides the data into two groups with the largest decrease in the total loss (increase in the imparity scores) of all children nodes. Once the best variable is chosen as the split condition, each resulting divided subset is then sent back to CART for the next iteration. The algorithm stops when a chosen loss reduction requirement can no longer be met. The original CART searches only for the best feature at the current level, and that may not necessarily result in the best overall tree. The computational time to perform exhaustive searches for the best overall tree is beyond current computational capability. To accommodate this, a local expansion of CART (local-CART) was developed. Written in R[25] (using the package “rpart” with parameters cp=0.005, maxdepth=4, minsplit=100), the local-CART uses the top five ranking split conditions at each node instead of one. Each condition is assigned to the current node, and CART is applied to its two child nodes. Among the five trees obtained, the one with the smallest total loss is chosen and its condition assigned to the current node. Subsequently, each divided subset goes through similar “looking ahead” iterations, resulting in a decision tree in closer proximity to the true best overall tree.

Results

Cluster Analysis. a. Optimizing Parameters. The penalized K-means[21] algorithm partitioned the 28 330 spectra into clusters so that the spectra within a given cluster display fragmentation patterns as similar as possible. Selection of two parameters is required for this algorithm: the number of clusters K and the threshold λ. Using clustering concordance rate as an indication of the stability of the clustering results from various K (K=2–7) and λ (λ=75, 100, 125), the optimum values were found at K=4 + noise and λ=100 (see Methods and Supporting Information 2). *Only the intensity information from singly charged b and y fragment ions was used as the spectral information.*

b. Four Clusters of Spectra that Display Distinct Fragmentation Behaviors. The most stable and most chemically significant patterns were found using K=4 + noise, λ=100, with a pattern deemed chemically significant if the pattern had been previously noted by researchers in the field or if CART (see below) led to a chemical motif that provided a chemical explanation for the fragmentation. Figures 1b–1f show the five sets of Quantile Maps that describe the fragmentation behavior of b^+ and y^+ ions from the five clusters obtained using penalized K-means. The four regular clusters will each be named by their associated dominant features.

Cluster **X-P** (Figure 1b) shows dominant selective cleavage N-terminal to Pro (X-P cleavage, where X is any amino acid residue) in **y** ions (column labeled “P” in the lower map, which is composed of dark circles indicating strong cleavages). Cleavages at other residue pairs are generally strongly suppressed. The **b**-ion map also shows stronger cleavage at X-P, but the overall intensity is far less than that in the **y**-ion map. Cluster **I/L/V-X** (Figure 1c) displays scattered cleavages from many AA pairs in **y** ions, with stronger cleavage at the C-terminal side of the aliphatic residues. The cleavage intensity strength follows the order I ≈ V > L > A. Cleavage C-terminal to glutamine (Q-X) is also relatively strong. Cleavage C-terminal to asparagine (N-X), the other residue that also contains an amide side chain, is not strong and does not show the same cleavage propensity as Q-X. Cluster **D/E-X** (Figure 1d) shows very strong selective cleavage C-terminal to Asp (D-X) in **y** ions. Cleavage C-terminal to Glu (E-X) is also relatively strong, along with cleavage C-terminal to His (H-X) and Lys (K-X). The **b**-ion map, just as the **b**-ion maps in the previous two clusters, shows trends similar to those of the **y** ions, but with much less overall intensity. Cluster **b&y** (Figure 1e) is the only cluster that displays more intense fragmentation of the **b** ion than that of **y** ions. Strong cleavages at D-X, E-X, and X-P are found in both **b** and **y** ions. Strong cleavages C-terminal to aliphatic residues I, L, and V, as well as N-terminal to Lys (X-K), are observed in **b** ions. The last set of Quantile Maps shown in Figure 1f represents the cleavage patterns in the “noise” data set. A total of 635 spectra that cannot be clustered into the above four clusters are gathered here. In this noise cluster, cleavages of various intensities are found for almost all AA pairs in both **b** and **y** ions, as indicated by the colorful circles.

Several cleavages show zero or very low intensities in the four main clusters. These cleavage sites are G-X and P-X in both **b** and **y** ions. Cleavages at S-X and T-X are rather weak in all Quantile Maps except **y** ions from cluster **I/L/V-X**.

Distance measurements between the cluster centers were performed to illustrate the similarity between the clusters (Table 1). A larger number means less similarity. The farther the distance to the noise cluster, the more selective the fragmentation pattern is. The distance between the four regular cluster centers to the noise cluster center follows the order **X-P** > **D/E-X** > **b&y** > **I/L/V-X**. As expected, the distance between

Table 1. Distance between Different Cluster Centers (a smaller number indicates greater similarity)

	X-P	**I/L/V-X**	**D/E-X**	**b&y**	noise
X-P	0	25.7	22.0	37.4	102.1
I/V/L-X	25.7	0	26.1	34.3	66.5
D/E-X	22.0	26.1	0	34.0	85.0
b&y	37.4	34.3	34.0	0	59.9
noise	102.1	66.5	85.0	59.9	0

any pair of the four regular clusters is less than the distance of any of the four main clusters to the noise set.

Chemistry behind the Fragmentation Patterns. Following the cluster analysis, CART analysis, together with separate analyses on the sequence and charge state of the corresponding peptides, was performed to elucidate the chemical motifs behind the different fragmentation patterns observed. Only those spectra that show stable fragmentation patterns via subsample-judged clustering (see Methods) were included. The numbers of spectra from each cluster that were sent for CART analysis are shown in Table 2.

a. The Decision Tree. The features in our local-CART are first generated on the basis of the basic attributes of peptides (e.g., m/z value, charge state, length, position of a residue, number of occurrence of a residue, distances of the residue to N-terminus and C-terminus) as well as current known factors from the literature that can influence peptide dissociation (e.g., the mobility of proton,[3,7] a measure of how readily proton(s) is transferred intramolecularly after activation to induce charge-directed cleavage). The features were then evolved empirically through multiple analyses in this study; for example, after proton mobility was found to be important, an additional related definition of proton mobility based on Kapp et al.[13] was added. It is important to note here that while it is true that if a specific feature is not included in CART, its direct effect will not be measured, it is not totally impossible to probe unknown features using this strategy. If an unknown feature is composed of several existing features, it will be shown by CART as the grandchild (or great-grandchild, etc.) nodes with the existing features as discriminators compounded at various levels. If an unknown feature is related to one or several existing features, their relationships will be reflected in CART. For example, if the gas-phase basicity of the peptide is important in defining fragmentation patterns, but it is not included in CART, then the occurrences of R, K, and H will become significant in CART because it is known that the occurrences of R, K, and H will greatly affect the gas-phase basicity of peptides. If, however, there are truly unknown features that can influence peptide fragmentation patterns and are totally unrelated to any of the features considered in this study, then they are beyond the capability of the current feature extraction methods.

The features included in the final tree are listed in Table 3. The optimum decision tree that can best describe the priority and the interactions of different features in determining the fragmentation patterns (or in statistical terms, the tree that shows the best improvement of total misclassification rate) is shown in Figure 2.

b. Three Most Significant Factors. Proton mobility and the positions of proline and arginine are found to be the three factors most significant (Figure 2) in determining the peptide fragmentation patterns from Figure 1. Among these three, proton mobility, as defined in Table 3, was found at the root level of the optimum decision tree. Figure 3a shows the distribution of the proton mobility among the four regular clusters, illustrating how proton mobility separates clusters **X-P** and **I/L/V-X** from **D/E-X** and **b&y**. While the curves of proton mobility from **X-P** and **I/L/V-X** maximize around 1, the curves from **D/E-X** and **b&y** maximize around 0.25. Analyses of the charge state distribution (Figure 3b) and the basic residue content (see Supporting Information Table 1) among different clusters further support the argument that peptides from clusters **X-P** and **I/L/V-X** are from peptides that have mobile proton(s), and peptides from clusters **D/E-X** and **b&y** are from peptides that have relatively localized proton(s). (See Supporting Information 4 for a detailed discussion.) Also included in the final CART was another measurement of the number of added protons relative to the number of basic residues–H^+_Mob_Kapp–based on the work by Kapp et al.[13] The results show that H^+_Mob_Kapp is important but ranks second to H^+_Mob (Figure 2a) in determining the fragmentation patterns.

Following the yellow arrow from the root node in the decision tree (Figure 2a), which stands for the “True” condition for H^+_Mob$\geq$0.5, the relative position of Pro (POS_P, see Table 3) becomes the next most prominent discriminator for clustering peptides whose proton mobility is greater than or equal to 0.5. Figure 3c, which plots the distributions of the relative position of Pro among different clusters, shows that cluster **X-P** has a significantly different distribution than those from other clusters. The majority of the proline residues in **X-P** are located toward the middle 80% length of the sequence, while the opposite trend is observed in **I/L/V-X**, with Pro located near the ends of the sequence or absent. Compared to the two clusters **X-P** and **I/L/V-X**, the proline residues in clusters **D/E-X** and **b&y** are distributed more evenly. The residue analysis (Supporting Information Table 1) shows that cluster **X-P** is the most proline-rich cluster; on average, each peptide from **X-P** contains 1.5 Pro, while that number for other clusters ranges from 0.6 to 0.8.

Following the blue arrow (“False” condition for H^+_Mob$\geq$0.5) from the root node in the decision tree (Figure 2) are peptides whose added protons are relatively localized. For these peptides, the position of Arg is the most influential factor, as indicated by the appearance of two related discriminators: DistC_R (the average distance of Arg to the C-terminus) and POS_R (the relative position of R in the sequence; see Table 3). Figure 3d shows that the distribution of the relative position of Arg in **b&y** is different from those in other clusters: a significantly higher percentage of the peptides in **b&y** have Arg either at the N-terminus or present in the N-terminal-half of the sequence (as indicated by POS_R<0.5). Conversely, only 54% of the peptides in **b&y** have a C-terminal Arg, which is significantly lower than those from other clusters. When Arg is close to the N-terminus of the peptide (e.g., in the case of missed tryptic cleavages), fragmentation patterns change from showing only dominant **y** ions to showing both strong **b** and **y** ions.

Discussion

This study reports the overall peptide collision-induced dissociation behavior among a large set of ion trap spectra acquired in real-world practical proteomics studies (i.e., the spectra were not acquired from model peptides but from protein digests of lysed bacteria under investigation).[8] With no presorting of the spectra and no chemical assumptions made prior to the cluster analysis, the penalized K-means algorithm is able to partition the behaviors of 28 330 spectra of unique

Table 2. Sequence Analysis of the Peptides that Show Stable Fragmentation Patterns from Different Clusters, Where Stable Fragmentation (Stable Peptides) Refers to those Peptides that Partition to the *Same* Cluster throughout the Raw-Sampling and 10 Subsampling Steps (See Methods)

			peptide length		extreme nonselective[a]		short sequences[b]	
	no. of total peptides	no. of stable peptides	mean	Std. Dev.	no. of peptides	% population	no. of peptides	% population
X-P	8335	7149	18.5	19.2	62	0.9%	55	0.8%
I/L/V-X	8385	7364	15.0	15.9	1035	14.1%	230	3.1%
D/E-X	6159	4964	15.5	16.5	168	3.4%	214	4.3%
b&y	4816	3869	12.9	14	457	11.8%	463	12.0%
noise	635	531	8.2	8.7	263	49.5%	209	39.4%

[a] Definition of "extreme nonselective cleavage": Denote α, β, γ, δ, ε as the normalized intensities of the five most abundant **b** and **y** ions from a peptide, and $\alpha \geq \beta \geq \gamma \geq \delta \geq \varepsilon$. According to our normalization scheme, $\alpha = 1$. The peptide is considered to display "extreme nonselective cleavage" if all of the followings conditions are met: $\beta \geq 0.8$ and $\gamma \geq 0.7$ and $\delta \geq 0.6$ and $\varepsilon \geq 0.5$. [b] The phrase "short sequence" refers to any peptide that has equal or less than 6 AA residues.

Table 3. Features in the Final CART

name	description
M_Z	mass-to-charge ratio of the peptide precursor ion
Z	charge state of the peptide
N	number of occurrences of residue X in the peptide, where X equals any of the 20 AA residues
H$^+$_Mob	Z – (# of Arg) – 0.5 * (# of Lys + # of His)
H$^+$_Mob_Kapp	**1**: Z<#Arg, **2**: Z>#Arg AND Z<(#Arg+#Lys+#His), 3: Z>(#Arg+#Lys+#His)
Length	number of total residues in the peptide
POS_X	relative position of residue X in the peptide, where X equals any of the 20 AA residues; it is calculated as [(the residue position of X − 1)/(length of the peptide − 1)][a]
DistN_X	Average distance (defined by the # of residues) of residue X to the N-terminus of the peptide, where X equals to any of the 20 AA residues
DistC_X	average distance of residue X to the C-terminus of the peptide, where X equals any of the 20 AA residues

[a] Definition of relative (fractional) position of a residue is useful because the peptides investigated are of different lengths. The residue position of X is defined as the position of X in a peptide string starting from the N-terminus, e.g., the residue position of P in AVLPK is 4, and the relative position of P is 0.75. If multiple X residues are present in the sequence, then an average residue position is taken, e.g., the residue position of P in PVLPK is [(1 + 4)/2]=2.5, and the relative position of P is [(2.5–1)/(5–1)]=0.375. With this definition, if X is at the N-terminus, the relative position of X is zero, POS_X=0, and if X is at the C-terminus, POS_X=1. (Note: If a peptide does not contain X, it will also have a POS_X=0.) A smaller POS_X indicates that X is closer to the N-terminus.

sequence and charge state into four distinct clusters that are chemically significant. Structural motifs responsible for different fragmentation patterns are obtained on the basis of the results from CART, plus additional analyses of the sequences and charge states of the peptides in different clusters. Five dominant end points of the CART decision tree are circled with dashed lines to guide the discussion below.

If the added proton(s) is mobile (H+_Mob ≥0.5) and Pro is present and not close to either terminus, X-P cleavage dominates the fragmentation pattern (cluster **X-P**, Figure 1b), as illustrated by the circled end points of Figure 2d, which indicates 4587 and 2101 spectra with these characteristics. This fragmentation pattern usually occurs in longer peptides (mean length of 18.5 AA, Table 2). The fragmentation pattern observed in **X-P** shows the greatest difference from that in the noise cluster (Table 1). While the charge dependence of X-P cleavage corroborates data from our previous publication,[8] the position dependence of X-P cleavage supports previous arguments that X-P cleavage is assisted by locally sterically favored backbone and side-chain conformations.[8,10]

If the added proton(s) is mobile but Pro is either absent or close to the termini and Arg is less than five residues away from the C-terminus, cleavages C-terminal to branched aliphatic residues become the most abundant cleavages (cluster **I/L/V-X**, Figure 1b) along with a prominent population of nonselective cleavage behavior (Table 2). This is illustrated by the circled end point in Figure 2c, which corresponds to 5978 spectra of doubly charged peptides. The order of cleavage intensity C-terminal to aliphatic residues (I ≈ V > L > A) correlates well with the order of steric hindrance at the peptide bond imposed by residue side chains: Val and Ile have β-branched side chains, Leu has a γ branched side chain, and Ala has only a methyl group as its side chain. This evidence, together with the observation of zero or very low cleavage intensities at G-X and P-X among all clusters, further supports the argument that conformational influence imposed by the side chain can assist or hinder peptide fragmentation by favoring a particular intramolecular nucleophilic attack and/or a particular charge solvation structure.[8,10]

If the added proton(s) is relatively localized, the presence of Arg in a sequence at the C-terminus became crucial in determining the fragmentation pattern. A peptide with a number of Arg in the sequence that is equal or greater than the number of ionizing proton(s) will have proton(s) localized at Arg. If Arg, Asp, and/or Glu are present in the sequence AND Asp is at least one residue away from the C-terminal Arg OR Glu is at least three residues away from the C-terminal Arg (two circled boxes indicating 1019 and 398 spectra in Figure 2b), cleavage C-terminal to acidic residues dominates the fragmentation pattern (cluster **D/E-X**, Figure 1d). The position requirements of Asp and Glu with regard to Arg suggests possible interaction between the side chain of the acidic residue and the side chain of Arg. Such interaction has been proposed to shut down selective cleavage at D-X and E-X.[26] The fact that Glu has a longer side chain than Asp, and Glu needs to be further away from Arg than Asp, as shown by CART, also supports such an interaction. If a sequence did not satisfy the above requirements of Asp and Glu, its fragmentation pattern will be relatively nonselective (cluster **b&y**, the box in Figure 2b indicating 473 spectra on the bottom left). The observation that all b-ion maps except for the one from the **b&y** cluster show trends similar to those of their corresponding y-ion maps but with much less overall intensity correlates well with the fact that the majority of the peptide population in our data set end with R or K. However, if Arg is absent from the C-terminus and close to the N-terminus, intense **b** ions can often be

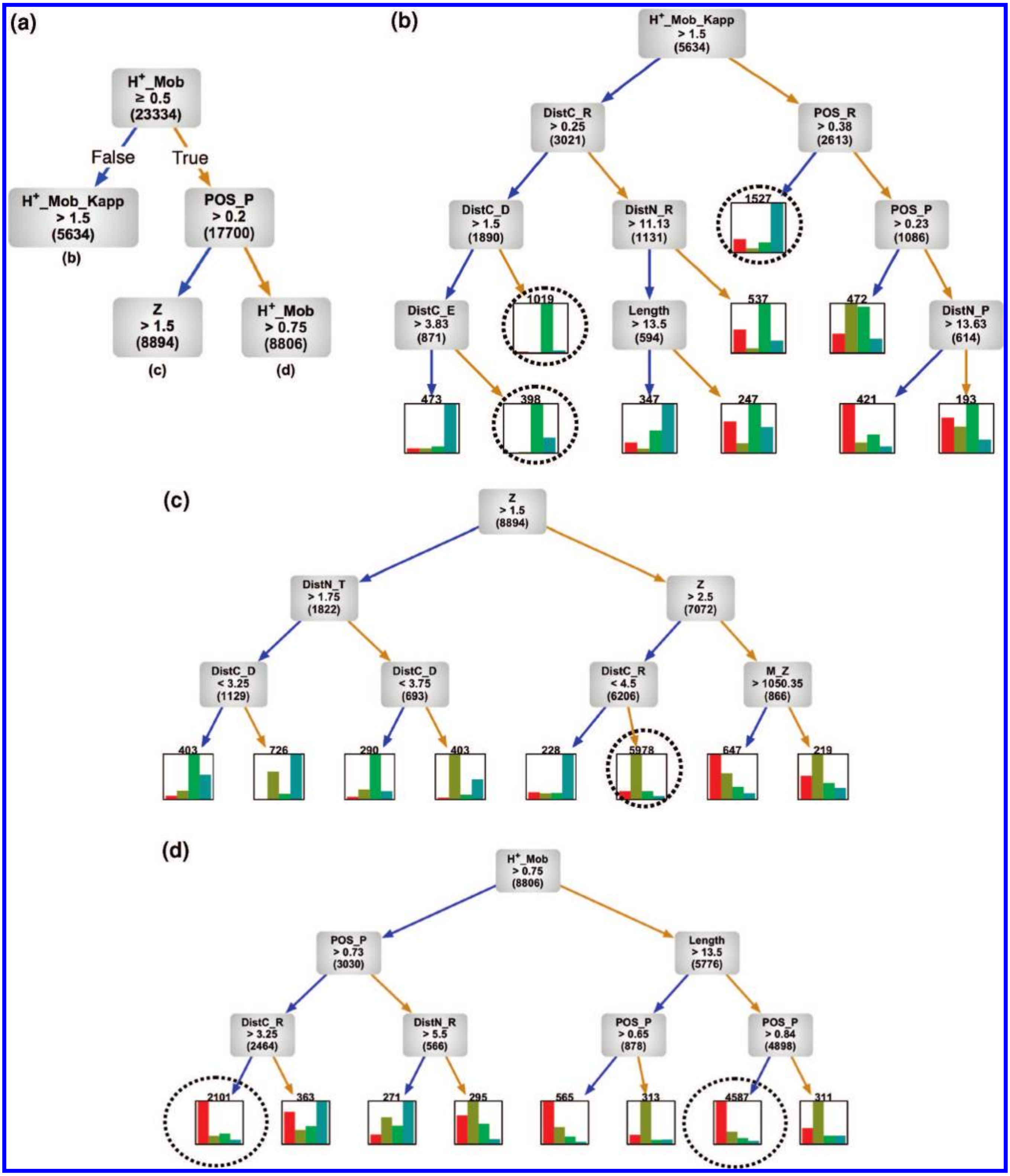


Figure 2. Final decision tree from CART: (a) the main tree (b, c, d) branches from (a). Yellow arrows (right side) follow "True" condition, while blue arrows (left side) follow "False" condition. Numbers in parentheses in each node represent the # of spectra at that node. Within each bar graph, the height of each bar represents the percentage of spectra that fall into such cluster. Red (far left) is for **X-P** cluster, gold (second from left) is for **I/L/V-X** cluster, green (third from left) is for **D/E-X** cluster, and blue (far right) is for **b&y** cluster.

observed (cluster **b&y**, circled box indicating 1527 spectra in Figure 2b). In addition to Arg, the position of His may also contribute to the intense **b**-ion patterns (Figure 3e), which correlates with a previous study[5] on model peptides. A large percentage of peptides in cluster **b&y** have His close to the N-terminus, while cluster **I/L/V-X** shows the opposite trend.

Among those peptides classified as noise, half of the peptides display "extreme nonselective cleavage" (multiple abundant

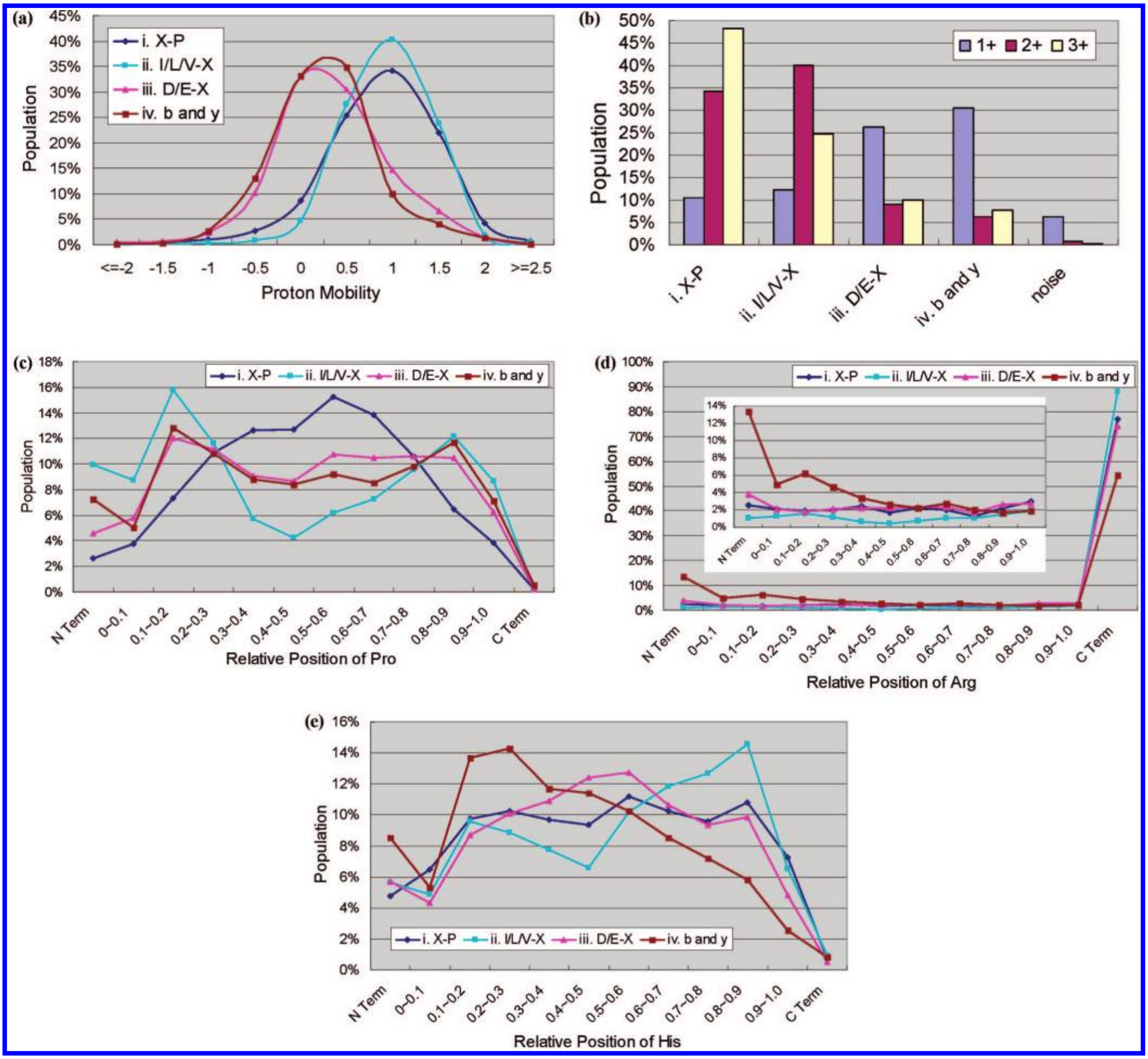


Figure 3. Distribution of (a) proton mobility, (b) charge states, (c) relative position of Pro, (d) relative position of Arg (an enlarged plot for Pos_R=0–0.9 is shown as an inset), and (e) relative position of His for the four clusters.

fragment ions, see Table 3); ~40% of the "noise" peptides have fewer than 7 AA residues, producing limited fragmentation information for the algorithm to use; and 75% of the peptides in the noise cluster are singly charged, which was reported to have the lowest percentage of identifiable intensity and are more likely to display nonbackbone cleavages than peptides of other charge states.[8]

Conclusions and Future Directions

The data-mining scheme proposed from this study can be readily applied to peptide MS/MS data sets acquired from other instruments and fragmentation methods, as long as such data sets are large enough and of high quality. In the last two years, linear ion traps[27] have emerged as a replacement for traditional 3D traps as the proteomics workhorse because of the increased ion capacity and improved trapping efficiency. However, peptide CID fragmentation patterns in the linear ion trap remain very similar to those from the 3D traps. Coupling the linear ion trap to mass analyzers that have high mass accuracy and resolution, such as the orbitrap[28,29] or the ICR,[30,31] allows CID MS/MS spectra to be acquired along with high mass accuracy measured on the precursor ions, or even on the fragment ions, if desired. Electron transfer dissociation (ETD)[32] has just recently been introduced to ion traps. This dissociation method can fragment multiply charged peptides that show very limited fragmentation in CID and often with contiguous fragment ion series that allow unambiguous sequence identification. The ability to alternate different dissociation methods for the same precursor ion will allow us to better identify a peptide and understand its gas-phase dissociation behavior from a new altitude. It is the authors' wish to see the data-mining scheme proposed in this paper applied to other data sets to derive new knowledge of peptide dissociation. Intensity profiles from different clusters of peptides along with the corresponding decision tree from such study create a solid foundation for new or improved sequencing algorithms to utilize the intensity information from MS/MS spectra.

Acknowledgment. The work results from a fruitful collaboration made possible by Institute of Pure and Applied Mathematics (IPAM), funded by NSF. The framework of this study was built during the IPAM Spring 2004 Program on "Proteomics: Sequence, Structure, Function", which the first three authors attended. Y.H. was supported by a 2003 Pfizer Graduate Research Fellowship in Analytical Chemistry. G.C.T. is supported by NIH grant KL2RR024154-01. S.Y. was supported by NSF grants DMS-9803459 and 0104038. Research at PNNL was supported by the U.S. DOE, Office of Biological and Environmental Research. PNNL is operated by Battelle Memorial Institute for the U.S. DOE under contract DE-AC06-76RLO 1830. Research in the Wysocki Laboratory was supported by NIH grant GM R0151387 to V.W.

Supporting Information Available: This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422* (6928), 198–207.

(2) Fenyo, D. A software tool for the analysis of mass spectrometric disulfide mapping experiments. *Comput. Appl. Biosci.* **1997**, *13* (6), 617–618.

(3) Dongre, A. R.; Jones, J. L.; Somogyi, A.; Wysocki, V. H. Influence of peptide composition, gas-phase basicity, and chemical modification on fragmentation efficiency: Evidence for the mobile proton model. *J. Am. Chem. Soc.* **1996**, *118* (35), 8365–8374.

(4) Gu, C.; Tsaprailis, G.; Breci, L.; Wysocki, V. H. Selective gas-phase cleavage at the peptide bond C-terminal to aspartic acid in fixed-charge derivatives of Asp-containing peptides. *Anal. Chem.* **2000**, *72* (23), 5804–5813.

(5) Tsaprailis, G.; Nair, H.; Zhong, W.; Kuppannan, K.; Futrell, J. H.; Wysocki, V. H. A mechanistic investigation of the enhanced cleavage at histidine in the gas-phase dissociation of protonated peptides. *Anal. Chem.* **2004**, *76* (7), 2083–2094.

(6) Vaisar, T.; Urban, J. Probing the proline effect in CID of protonated peptides. *J. Mass Spectrom.* **1996**, *31* (10), 1185–1187.

(7) Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A. Mobile and localized protons: a framework for understanding peptide dissociation. *J. Mass Spectrom.* **2000**, *35* (12), 1399–1406.

(8) Huang, Y. 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.* **2005**, *77* (18), 5800–5813.

(9) Tabb, D. L.; Smith, L. L.; Breci, L. A.; Wysocki, V. H.; Lin, D.; Yates, J. R. Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal. Chem.* **2003**, *75* (5), 1155–1163.

(10) Huang, Y.; Triscari, J. M.; Pasa-Tolic, L.; Anderson, G. A.; Lipton, M. S.; Smith, R. D.; Wysocki, V. H. Dissociation behavior of doubly-charged tryptic peptides: Correlation of gas-phase cleavage abundance with Ramachandran plots. *J. Am. Chem. Soc.* **2004**, *126* (10), 3034–3035.

(11) Huang, Y.; Wysocki, V. H.; Tabb, D. L.; Yates, J. R. The influence of histidine on cleavage C-terminal to acidic residues in doubly protonated tryptic peptides. *Int. J. Mass Spectrom.* **2002**, *219* (1), 233–244.

(12) Breci, L. A.; Tabb, D. L.; Yates, J. R.; Wysocki, V. H. Cleavage N-terminal to proline: Analysis of a database of peptide tandem mass spectra. *Anal. Chem.* **2003**, *75* (9), 1963–1971.

(13) Kapp, E. A.; Schutz, F.; Reid, G. E.; Eddes, J. S.; Moritz, R. L.; O'Hair, R. A. J.; Speed, T. P.; Simpson, R. J. Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Anal. Chem.* **2003**, *75* (22), 6251–6264.

(14) 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.* **2004**, *22* (2), 214–219.

(15) Tabb, D. L.; Huang, Y.; Wysocki, V. H.; Yates, J. R. Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* **2004**, *76* (5), 1243–1248.

(16) Zhang, Z. Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* **2004**, *76* (14), 3908–3922.

(17) Paizs, B.; Suhai, S. Fragmentation pathways of protonated peptides. *Mass Spectrom. Rev.* **2005**, *24* (4), 508–548.

(18) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer-Verlag: New York, 2001.

(19) Watson, J. T., *Introduction to Mass Spectrometry*, 3rd ed.; Lippincott-Raven Publishers: Philadelphia, PA, 1997.

(20) Hartigan, J. A.; Wong, M. A. A K-means clustering algorithm. *Appl. Statistics* **1979**, *28* (1), 100–108.

(21) Tseng, G. C. Penalized and weighted K-means for clustering with scattered objects and prior information in high-throughput biological data. *Bioinformatics* **2007**, *23*, 2247–2255.

(22) Tibshirani, R.; Walther, G. Cluster validation by prediction strength. *J. Comput. Graphical Statistics* **2005**, *14* (3), 511–528.

(23) Tseng, G. C.; Wong, W. H. Tight clustering: A resampling-based approach for identifying stable and tight patterns in data. *Biometrics* **2005**, *61* (1), 10–16.

(24) Breiman, L.; Friedman, J. J.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; CRC Press: Boca Raton, 1984.

(25) Ihaka, R.; Gentleman, R. R. A language for data analysis and graphics. *J. Comput. Graphical Statistics* **1996**, *5* (3), 299–314.

(26) Bailey, T. H.; Laskin, J.; Futrell, J. H. Energetics of selective cleavage at acidic residues studied by time- and energy-resolved surface-induced dissociation in FT-ICR MS. *Int. J. Mass Spectrom.* **2003**, *222* (1–3), 313–327.

(27) Schwartz, J. C.; Senko, M. W.; Syka, J. E. P. A two-dimensional quadrupole ion trap mass spectrometer. *J. Am. Soc. Mass Spectrom.* **2002**, *13* (6), 659–669.

(28) Makarov, A. Mass Spectrometer. U.S. Patent 5,886,346, 1999.

(29) Hu, Q.; Noll, R. J.; Li, H.; Makarov, A.; Hardmanc, M.; Cooks, R. G. The Orbitrap: a new mass spectrometer. *J. Mass Spectrom.* **2005**, *40*, 430–443.

(30) Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S. Fourier transform ion cyclotron resonance mass spectrometry: A primer. *Mass Spectrom. Rev.* **1998**, *17* (1), 1–35.

(31) Marshall, A. G. H.; Hendrickson, C. L. Fourier transform ion cyclotron resonance detection: principles and experimental configurations. *Int. J. Mass Spectrom.* **2002**, *215*, 59–75.

(32) Syka, J. E. P.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci.* **2004**, *101* (26), 9528–9533.

PR070106U