Anal. Chem. **2003,** *75,* 1155−1163

Statistical Characterization of Ion Trap Tandem Mass Spectra from Doubly Charged Tryptic Peptides

David L. Tabb,[†] Lori L. Smith,[‡] Linda A. Breci,[‡] Vicki H. Wysocki,[‡] Dayin Lin,[§] and John R. Yates, III*

SR11 Department of Cell Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037

Collision-induced dissociation (CID) is a common ion activation technique used to energize mass-selected peptide ions during tandem mass spectrometry. Characteristic fragment ions form from the cleavage of amide bonds within a peptide undergoing CID, allowing the inference of its amino acid sequence. The statistical characterization of these fragment ions is essential for improving peptide identification algorithms and for understanding the complex reactions taking place during CID. An examination of 1465 ion trap spectra from doubly charged tryptic peptides reveals several trends important to understanding this fragmentation process. While less abundant than y ions, b ions are present in sufficient numbers to aid sequencing algorithms. Fragment ions exhibit a characteristic series-specific relationship between their masses and intensities. Each residue influences fragmentation at adjacent amide bonds, with Pro quantifiably enhancing cleavage at its N-terminal amide bond and His increasing the formation of b ions at its C-terminal amide bond. Fragment ions corresponding to a formal loss of ammonia appear preferentially in peptides containing Gln and Asn. These trends are partially responsible for the complexity of peptide tandem mass spectra.

Tandem mass spectrometry (MS/MS) of peptides is a central technology for proteomics, enabling the identification of thousands of peptides from a complex mixture.[1−4] This increasingly widespread technique relies upon the fragmentation of peptides by collision-induced dissociation (CID), but the chemistry behind the fragmentation process is complex and not comprehensively understood.[5−8]

Peptides undergo CID after they are isolated from other ions by their mass-to-charge (m/z) ratios. Peptides in an acidic solution are introduced to the vacuum of the mass spectrometer via electrospray ionization.[9] The peptide ions are accelerated during CID, leading to more energetic collisions with the ion trap's inert gas molecules. The mobile proton model[10] describes how the added internal energy causes the ionizing proton(s) on each peptide to transfer intramolecularly until one destabilizes a peptide bond, resulting in the cleavage of that bond and the production of two fragments. While more energetic techniques may cleave many classes of bonds within the peptide structure, low-energy CID preferentially breaks the amide bonds. Once the fragment ions are produced, the mass spectrometer records their m/z ratios in a tandem mass spectrum.

Determining the sequence of a peptide from its tandem spectrum is complicated by the variety and variability of the fragment ions produced. Cleavage of amide bonds results in b and y ions[11,12] (see Figure 1). b ions may fragment further to produce a ions.[13] If only these three ions were produced for every amide bond in a 10-residue peptide, the fragment ion spectrum would contain 27 peaks. This ideal spectrum differs from experimental spectra as a result of several causes. First, a subset of the expected fragment ions may not be present. Second, fragment ions may undergo internal rearrangements, subsequent fragmentation, or both to yield other types of ions. Additionally, ions may be present in multiple charge states. Taken together, these influences may complicate interpretation of tandem mass spectra.[14]

* Corresponding author: (phone) 858 784-8876; (fax) 858 784-8883; (e-mail) jyates@scripps.edu.

[†] Current address: Department of Genome Sciences, University of Washington, Seattle, WA 98195.

[‡] Current address: Department of Chemistry, University of Arizona, Tucson, AZ 85721-0041.

[§] Current address: Waters Corp., 6747-A Sierra Court, Dublin, CA 94568.

(1) Washburn, M. P.; Wolters, D.; Yates, J. R., III. *Nat. Biotechnol.* **2001**, *19*, 242−247.

(2) VerBerkmoes, N. C.; Bundy, J. L.; Hauser, L.; Asano, K. G.; Razumovskaya, J.; Larimer, F.; Hettich, R. L.; Stephenson, J. L., Jr. *J. Proteome Res.* **2002**, *1*, 239−252.

(3) Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. *J. Proteome Res.*, in press.

(4) Florens, L.; Washburn, M. P.; Raine, J. D.; Anthony, R. M.; Grainger, M.; Haynes, J. D.; Moch, J. K.; Muster, N.; Sacci, J. B.; Tabb, D. L.; Witney, A. A.; Wolters, D.; Wu, Y.; Gardner, M. J.; Holder, A. A.; Sinden, R. E.; Yates, J. R.; Carucci, D. J. *Nature* **2002**, *419*, 520−526.

(5) O'Hair, R. A. J. *J. Mass Spectrom.* **2000**, *35*, 1377−1381.

(6) Schlosser, A.; Wolf, D. L. *J. Mass Spectrom.* **2000**, *35*, 1382−1390.

(7) Polce, M. J.; Ren, D.; Wesdemiotis, C. *J. Mass Spectrom.* **2000**, *35*, 1391−1398.

(8) Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A. *J. Mass Spectrom.* **2000**, *35*, 1399−1406.

(9) Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. *Science* **1989**, *246*, 64−71.

(10) Dongre, A. R.; Jones, J. L.; Somogyi, A.; Wysocki, V. H. *J. Am. Chem. Soc.* **1996**, *118*, 8365−8374.

(11) Roepstorff, P.; Fohlman, J. *Biomed. Mass Spectrom.* **1984**, *11*, 601.

(12) Johnson, R. S.; Martin, S. A.; Biemann, K. *Anal. Chem.* **1987**, *59*, 2621−2625.

(13) Yalcin, T.; Csizmadia, I. G.; Peterson, M. R.; Harrison, A. G. *J. Am. Soc. Mass Spectrom.* **1996**, *7*, 233−242.

(14) Hunt, D. F.; Yates, J. R. III; Shabanowitz, J.; Winston, S.; Hauer, C. R. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 6233−6237.

10.1021/ac026122m CCC: $25.00
Published on Web 01/30/2003

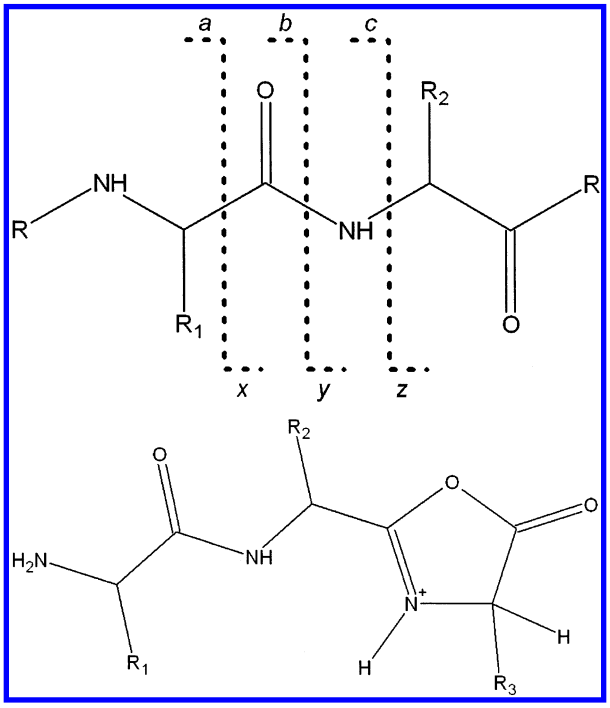

Figure 1. Peptide bond cleavage. Low-energy CID primarily cleaves peptide bonds, resulting in b ions (which contain the N-terminus and the atoms to the left of the dotted line) and y ions (which contain the C-terminus and the atoms to the right of the dotted line). b ions (pictured) generally take on an oxazolone structure, which may subsequently fragment to produce smaller b ions or lose carbon monoxide to form a ions. The remaining possible backbone ions (c, x, z) do not typically form under low-energy conditions.

Several computer programs have been created to match database peptide sequences to tandem mass spectra. SEQUEST,[15] Mascot,[16] and MS-Tag[17] implement sequence database search algorithms. The aim of this approach is to find the peptide sequence in a database that best explains the fragment ions present in a spectrum. Candidate sequences are found in the database on the basis of intact peptide masses, and the complete or partial spectra expected to result from the fragmentation of these candidate peptides are generated and compared to the experimental spectrum.

De novo sequencing algorithms, such as Lutefisk[18,19] and SHERENGA,[20] attempt to infer peptide sequences given only the information in each spectrum. These algorithms identify pairs of peaks that are separated by an amino acid's mass. If a series of such pairs can be found, a portion of the peptide's sequence may be identified, and the portions of the sequence for which no fragment ions are observed can be inferred.

Due to the lack of information available about fragmentation mechanisms, most algorithms rely on inaccurate or simplistic models of these spectra. Generating correct m/z values for fragment ion peaks is relatively straightforward. Most algorithms, however, do not implement intensity models capable of predicting intense peaks or absent ones. As a result, the information present in the second dimension of experimental spectra is not being exploited.

Previous efforts to statistically analyze fragment ion spectra have had limited success. Van Dongen and co-workers assembled a collection of 138 peptides for their analysis but focused on singly charged precursor ions and used high-energy CID.[21] Dančik and co-workers studied a collection of low-energy CID spectra while developing the SHERENGA algorithm, but the analysis assumed that all ions from a series would exhibit similar characteristics.[20] A promising recent effort applied a kinetic model to simulate fragment ion spectra as a function of several mechanisms.[22] Such studies have begun to establish a statistical foundation for future algorithms.

This research attempts to identify statistical trends in fragment spectrum peak intensity and to put these trends into chemical context. The relationship of fragment ion peak intensity to ion series origin and fragment mass will be explored first. Next, the influence of amino acid residues on neighboring amide bond cleavages will be evaluated. Finally, the link between amino acid composition and neutral loss fragmentation will be tested.

EXPERIMENTAL SECTION

Preparation of Yeast Proteome for MS Analysis. A culture of *Saccharomyces cerevisiae* (strain 1560) was grown to an optical density at 600 nm of 1.2 in 2 L of YPD media. Cells were lysed and proteins were extracted and digested according to a protocol similar to that described for the insoluble fraction by Washburn et al.[1] Briefly, the total yeast lysate was centrifuged and the pellet portion was dissolved in 90% formic acid and treated with cyanogen bromide (Sigma, St. Louis, MO) overnight while the soluble portion was directly reduced with tris(2-carboxyethyl)phosphine (Pierce Chemical Co., Rockford, IL) and alkylated with iodoacetamide (Sigma). For the CNBr-treated fraction, the pH of the solution was brought up to 8.0 with cold 30% NH_4OH and saturated $(NH_4)_2HCO_3$. The peptides of the pellet fraction were reduced and alkylated like those of the soluble fraction. For both soluble and pellet fractions, proteins were digested sequentially with endoproteinase Lys-C and trypsin. The resulting peptides were desalted once using SPEC solid-phase extraction C18 pipet tips (Ansys Diagnostics, Inc., Lake Forest, CA).

Analysis of Proteome by MudPIT. Multidimensional protein identification technology (MudPIT) was used to determine the protein content of this complex mixture.[23] The soluble fraction was divided and analyzed twice, while the pellet fraction was analyzed in a single experiment. The liquid chromatography columns were constructed at the time of use. A piece of fused-silica capillary (100-μm i.d./363-μm o.d.) (Polymicro) was pulled to have an opening of 5 μm. A biphasic LC column was prepared by packing the capillary with AQUA C18 particles (Phenomenex,

(15) Eng, J. K.; McCormack, A. L.; Yates, J. R., III. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.
(16) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell J. S. *Electrophoresis* **1999**, *20*, 3551−3567.
(17) Clauser, K. R.; Baker, P. R.; Burlingame, A. L. *Anal. Chem.* **1999**, *71*, 2871−2882.
(18) Taylor, J. A.; Johnson, R. S. *Rapid Commun. Mass. Spectrom.* **1997**, *11*, 1067−1075.
(19) Taylor, J. A.; Johnson, R. S. *Anal. Chem.* **2001**, *73*, 2594−2604.
(20) Danè×c1k, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. *J. Comput. Biol.* **1999**, *6*, 327−342.
(21) van Dongen, W. D.; Ruijters, H. F. M.; Luinge, H.-J.; Heerma, W.; Haverkamp, J. *J. Mass Spectrom.* **1996**, *31*, 1156−1162.
(22) Zhang, Z. *Proc. 50th Am. Soc. Mass Spectrom.*, Orlando, FL, 2002; Paper TPE-126.
(23) Link, A. J.; Eng, J. K.; Schieltz D. M.; Carmack, E., Mize G. J., Morris D. R.; et al. *Nat. Biotechnol.* **1999**, *17*, 676−682.

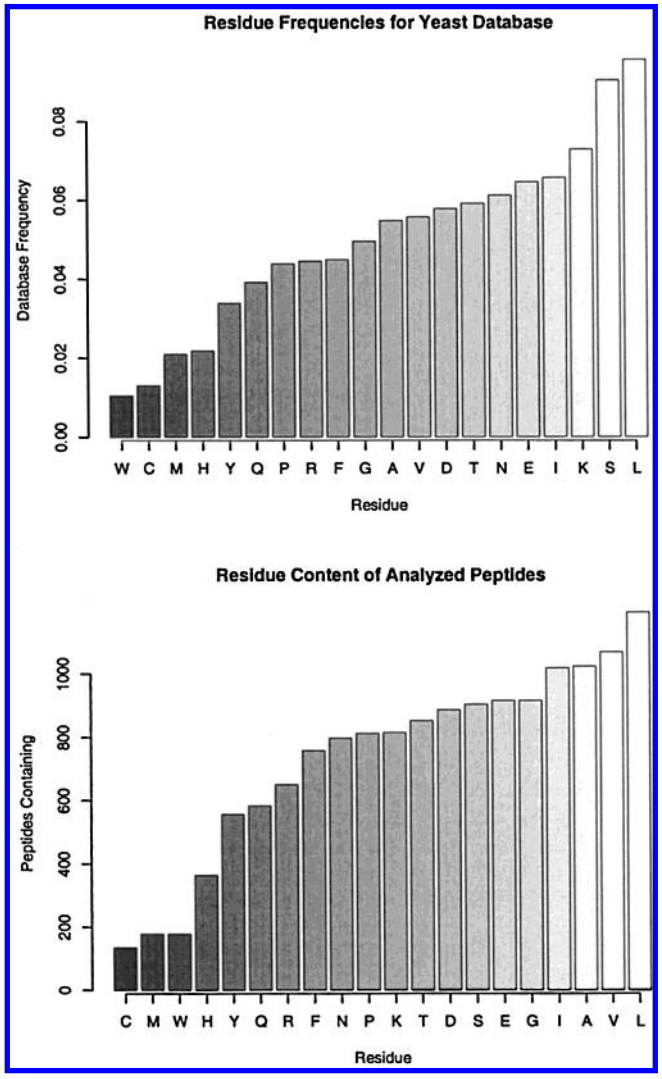

Figure 2. Residue frequencies and content. The peptides included in this analysis have a somewhat modified composition with respect to the sequence database by which they were identified. The residues identified most often among the peptides had alkyl side chains. The six rarest residues in the identified peptides were the same as those seen least often in the sequence database. Cysteine was present in only 134 peptides, and Met and Trp were each present in 178 peptides.

Torrance, CA) followed by Whatman SCX beads (VWR) using a high-pressure cell.[24] The yeast peptides were loaded onto the column using the same pressure cell.

The columns were positioned to elute directly into the ion source on a custom-built stage.[25] An Agilent HP1100 LC pump (Palo Alto, CA) produced a flow of 100 μL/min, which was split to produce a flow rate of 200−300 nL/min. There were 12 LC cycles in the completely automated LC/LC/MS/MS analysis. The initial cycle used no salt in order to collect data on peptides that bypassed the SCX media. Subsequent cycles used 4%, 8%, 10%, 12%, 15%, 20%, 30%, 40%, 50%, 75%, and 100% concentrations of 500 mM ammonium acetate. Each reversed-phase gradient used an "A" solvent of 5% acetonitrile/0.1% formic acid and a "B" solvent of 80% acetonitrile/0.1% formic acid. The gradients themselves lasted 100 min, ramping from 88% A/12% B to 45% A/55% B. The final cycle of the MudPIT increased to a maximum of 30% A/70% B over 110 min to elute the most hydrophobic peptides from the column.

Two different Thermo Finnigan LCQ Deca ion trap mass spectrometers (San Jose, CA) were used for the soluble fraction samples, and one of the pair was used for analyzing the pellet sample. XCalibur instrument control software, version 1.2, handled the instruments. The software's "dynamic exclusion" feature reduced the extent to which high-abundance peptides were sampled preferentially over low-abundance ones, thus increasing the instrument's effective sensitivity. A normalized collision energy of 35% was applied to the peptides, a setting that typically fragments all of the precursor peptide ions.

Preliminary Informatics. The SEQUEST algorithm[15] was used to process the spectra against the yeast open reading frame database.[26] The program was not configured to search specifically for tryptic peptides, so all peptides from the yeast genome were considered in assessing the identifications. The algorithm was configured to assume that all cysteine residues had been modified by reduction and alkylation. The configuration allowed matching to database peptides with masses 3 Da higher and 3 Da lower than the observed peptide mass.

The DTASelect algorithm[27] filters SEQUEST results and assembles protein-level information from peptide data. Its default settings indicated that more than 2500 proteins were identified in the assembled results of the three MudPIT analyses. The program's filtering capacity was used in two stages to isolate identifications for further analysis. The first filtering step was designed to isolate peptides coming from reliably identified proteins. SEQUEST's primary score for each peptide identification is the XCorr, a measure of how well the theoretical spectrum cross-correlates to the observed spectrum. XCorr cutoffs were found that would retain the top 10% of singly and triply charged peptides and 25% of the doubly charged peptides. The XCorr cutoffs for each charge state corresponding to these percentages were as follows: 1.699 (+1), 2.290 (+2), 3.083 (+3). Only spectra for which the first sequence match scored at least 8% better than than second were retained (corresponding to a SEQUEST-assigned "DeltCN" of 0.08). When multiple copies of a particular spectrum were found, DTASelect retained only the one with the highest XCorr. Two different sequence identifications were required for any protein to be retained. Cumulatively, this filtering reduced the set of 129 282 original spectra to 6417.

The second filtering step was designed to reduce the pool of identifications to the subset targeted for statistical analysis. DTASelect's charge filters isolated the doubly charged peptides remaining after the first round of filtering. Of these, only the peptides with more than 50% of expected ions appearing were retained. In addition, peptides were required to have sequences ending in arginine or lysine, with no internal arginine or lysine residues. The second round of filtering selected a final set of 1465 spectra.

A new algorithm, entitled DaughterDB, created databases describing these spectra.[28] The program creates three reports.

(24) Lin, D.; Alpert, A. J.; Yates, J. R., III. *Am. Genomic/Proteomic Technol.* **2001**, *1*, 38−46.
(25) http://fields.scripps.edu/mudpit/
(26) http://genome-www.stanford.edu/Saccharomyces/
(27) Tabb, D. L.; McDonald, W. H.; Yates, J. R. III. *J. Proteome Res.* **2002**, *1*, 21−26.
(28) http://fields.scripps.edu/DaughterDB/

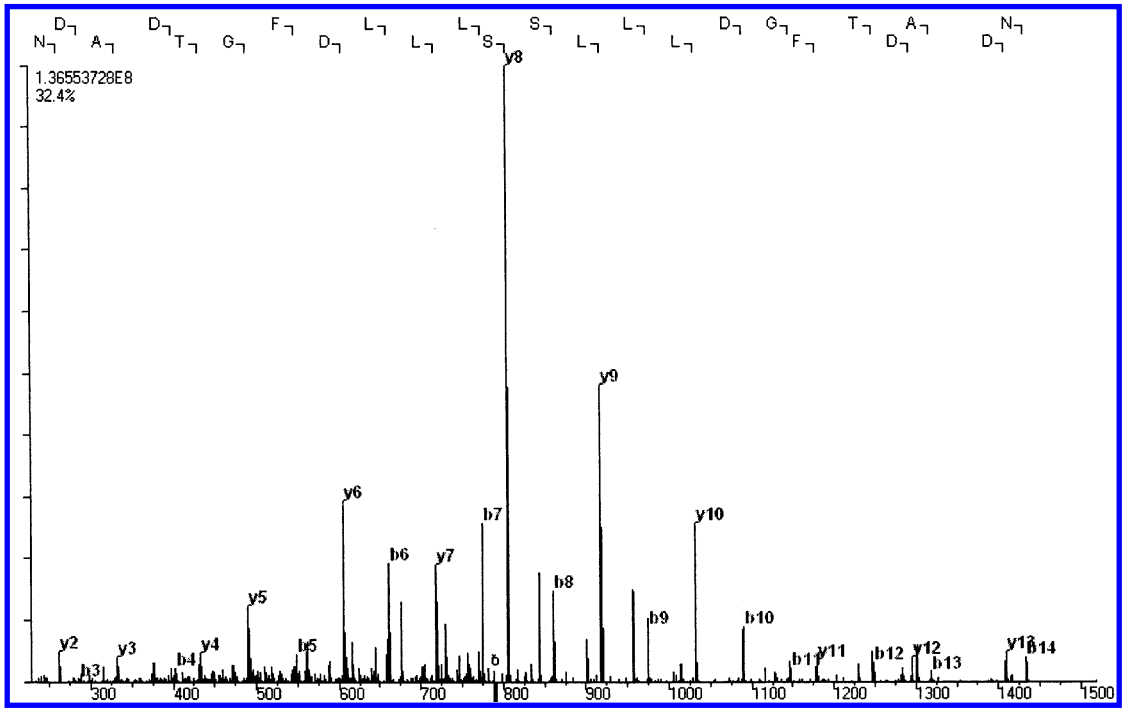

Figure 3. Mass spectrum of the sequence AVDDFLLSLDGTANK. The y5 ion corresponds to the intensity of the median y ion for all spectra in this analysis. The base peak in the above spectrum, which was identified as y8, embodies 9% of the spectrum's total intensity. Identified b and y ions account for 32.4% of this spectrum's total intensity.

The first includes a row of information for each spectrum, including the number of peaks appearing in the spectrum and the percentage of total intensity from singly charged fragment ions. The second report includes a row for each theoretical fragment ion expected within the scan range of each spectrum, describing the most intense matching ion in the experimental spectrum within 0.75 m/z to either side of the calculated fragment ion position. Ion intensities were reported as a fraction of the sum of the spectrum's intensities in order to normalize the intensity differences between spectra (see Figure 3 for examples). Although DaughterDB enumerates both singly and doubly charged fragment ions, only the singly charged fragments were included. Fragment ions of fragment ions are not identified by DaughterDB. The energy applied to the peptide precursors was sufficient to remove these ions' peaks completely from the fragment ion mass spectra. The third DaughterDB report is similar to the second but enumerates pairs of consecutive fragment ions that are both within the scan range. These data were analyzed statistically in a free statistics package called R,[29] which is based upon Insightful's S and S-Plus software.

RESULTS AND DISCUSSION

Peptide Diversity and Representation. The 1465 doubly charged tryptic peptides selected for this study had a median length of 15 residues. The middle 50% of peptide lengths ranged from 13 to 18 residues; the shortest peptide was 7 residues, and the longest was 27 residues. The median peptide mass was 1643 Da. The middle 50% of the masses ranged from 1397 to 1945 Da, while the full range stretched from 843 to 2747 Da. The median XCorr assigned by SEQUEST was 3.64. The middle 50% of XCorrs ranged from 3.12 to 4.28, with a minimum 2.29 (the cutoff applied by DTASelect) to a maximum of 6.63.

Figure 2 compares the composition of the full *S. cerevisiae* database to the peptide content of the identifications included in this analysis. Each peptide was required to terminate in a basic residue; 650 (44%) had C-terminal Arg residues, while 815 (55%) ended in Lys. Amino acid residues with alkyl side chains (Leu, Val, Ala, Ile) were the four most common residues found in the peptides.

The cross-correlation scores produced by SEQUEST for this set of peptides correlated to both peptide mass (correlation coefficient $r = 0.59$) and sequence length ($r = 0.61$). DaughterDB's report included several measures coordinating SEQUEST's match to the peptide. SEQUEST scores correlated most closely ($r = 0.66$) to the length of the longest contiguous series of singly charged fragment ions for each spectrum. SEQUEST appears to give higher scores to longer sequences and more massive peptides, especially favoring spectra with long contiguous series of matching fragment ions.

Series-Specific Characterization. The cleavage of a peptide at an amide bond in low-energy CID results in several series of ions (see Figure 1). The b and y ions directly result from cleavage, while a ions result from the formal loss of carbon monoxide from b ions. While b and y ions are found at the vast majority of locations where they are predicted, a ions are less common (see Table 1). The energy of fragmentation in low-energy CID is insufficient to break the bond between the α-carbon and the carbonyl, and so x ions are not typically produced; in this study, they are included only as a measure of background noise. In these spectra, predicted x ions can be matched to observed peaks 23% of the time, indicating that noise peaks in the spectra may be contributing substantially to the percentages of identified ions. Ions predicted to fall outside the observed scan range were excluded from this analysis; because the spectra were collected on an ion trap mass spectrometer, low-m/z peaks were truncated

(29) Hornik, K. The R FAQ. http://www.ci.tuwien.ac.at/~hornik/R/.

Table 1. Peptide Fragment Ion Series Comparison

	number found[a]	% found[b]	individual intensity[c] (%)	spectrum intensity[d] (%)
a	6863	40	0.20	1.98
b	14823	84	0.47	8.59
y	15729	90	1.03	21.82
x	3971	23	0.15	0.75

[a] Number of predicted ions for which a peak was observed within 0.75 of predicted m/z. [b]Comparison of the number of ions found to the number expected within the scan range for each ion series. [c] Median intensity for the ions identified for each series (leaving out ions that could not be matched to an observed peak). [d]Average percentage of each spectrum's intensity that can be accounted for by the ions from each series. The x ion series is included as a measure of noise; these ions are not expected to form in low-energy CID.

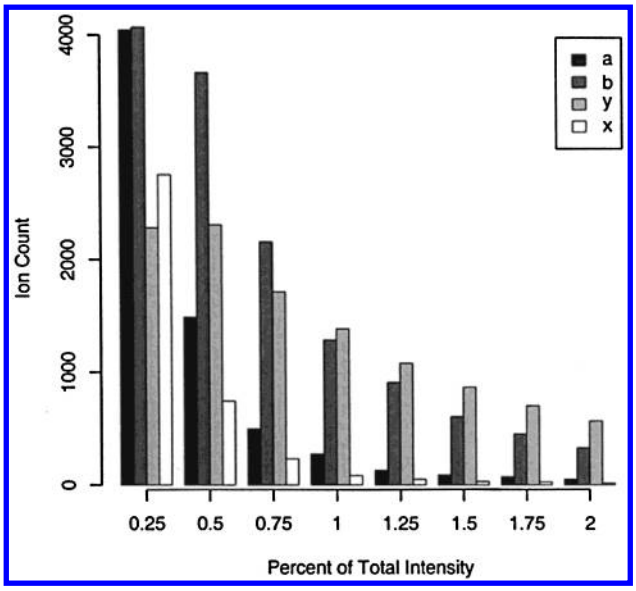

Figure 4. Intensity distributions of ion species. Small peaks are more common than intense ones for each series, but the more intense the peak, the more likely it is to represent an ion from the y series. The distribution of peak intensities extends beyond the most intense category in this graph: 31% of y peaks are more intense than 2% of the spectrum's summed intensity, as compared to 9% of b peaks, 3% of a peaks, and 1% of background (x) peaks.

from the spectrum. On average, the smallest m/z in each spectrum was 30% of the precursor's m/z.

In this spectral collection, y ions were found only slightly more often than b ions, but their peaks were typically more than twice as intense. The distributions of each series' intensities are shown in Figure 4. The percentage of intensity in each spectrum accounted for by b and y ions was calculated. The median percentage was 28.8%, with the range between the 25th and 75th percentiles spanning from 23.1% to 35.9%. See Figure 3 for a sample spectrum in which 32.4% of the intensity was accounted for by b and y ions. In its search for each fragment ion, DaughterDB counts only the largest peak found within 0.75 m/z of the calculated position. As a result, peaks representing isotopic variants of the fragment ions will not be counted as part of the identified intensity for each spectrum. Low-intensity peaks can account for a sizable portion of the spectrum's intensity because they are numerous. The remaining peaks may include unfragmented precursor ion, neutral losses from the fragment or precursor ions, or other rearrangement ions.

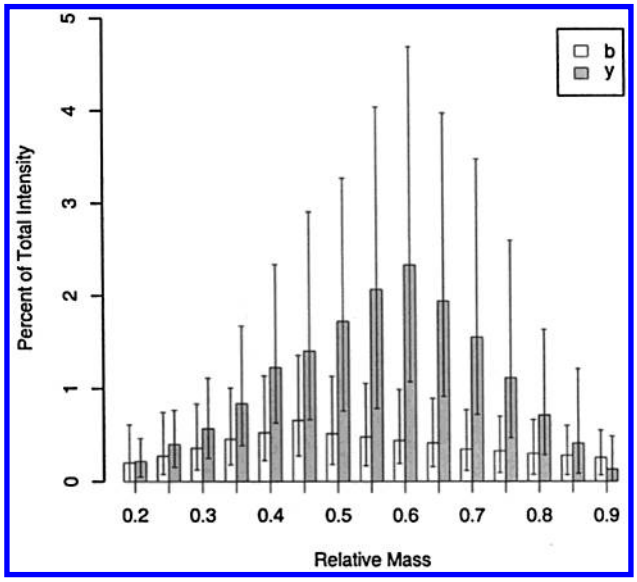

Figure 5. Fragment ion peak height versus relative mass. Peak intensities are related to the relative masses of the fragment ions they represent. The horizontal axis gives masses of fragments as a proportion of precursor mass. The bar shows the intensity of the median peak for the collection of ions in a particular relative mass bin, with a line extending above and below to show the 75th and 25th percentile intensities. Missing peaks are assigned intensities of zero. The y series shows a distinct peak at ~60% of precursor mass, while b peaks crest at 45%.

Figure 5 illustrates the relationship between fragment peak intensity and relative mass. The relative mass for each fragment ion is its mass divided by that of the intact peptide. The peaks for the y series reach a distinct peak in intensity when the fragments are approximately two-thirds the mass of the precusor ion. The b series peaks are highest at ~45% the mass of the precursor. The a series is excluded from this figure because fewer than half of these ions can be associated with spectral peaks outside the region ranging from 30% to 50% the mass of the precursor. The regions where the peaks are most intense are also the regions in which the intensity variation is greatest.

Residue-Specific Behavior. (1) N-Bias by Residue. Individual amino acid residues influence which of the two adjacent amide bonds (N-terminal or C-terminal) break in the process of collision-induced dissociation. The structure of Pro, for example, prevents the cleavage of the peptide bond C-terminal to the residue by hindering the attack of the N-terminal carbonyl[30] (see Figure 6). The extent to which each residue directionally enhances cleavage was measured by comparing the intensities of fragment ion peaks adjacent to each residue. The intensity of the C-terminal fragment peak was subtracted from the intensity of the N-terminal fragment peak. These differences were divided by the sum of the two peak intensities to yield the "N-bias". A subset of the fragment peaks was calculated to fall within the scan range but did not appear in the tandem mass spectra. Each of these peaks was assigned an intensity of zero and included in the analysis. If neither peak was observed for a pair, it was excluded. If either of the two fragment ions fell outside the scan range, the pair was excluded from the analysis.

Examples from the spectrum shown in Figure 3 may clarify this measure. Three y ion peak pairs show the presence of Leu

(30) Loo, J. A.; Edmonds, C. G.; Smith, R. D. *Anal. Chem.* **1993**, *65*, 425−438.

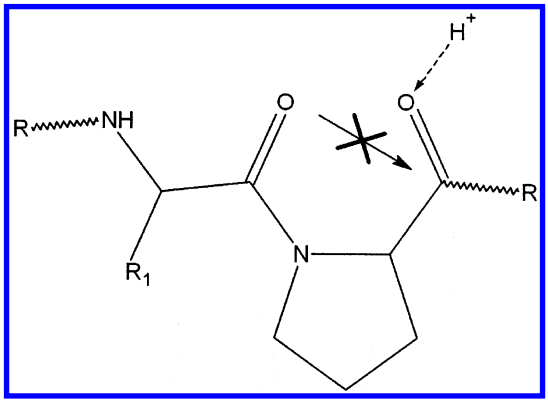

Figure 6. Unusual fragmentation of Pro. Because Pro's side chain forms a ring to its nitrogen, attack on its carbonyl carbon by the preceding carbonyl oxygen would result in a strained 5−5 bicyclic ring. Cleavage to its C-terminus is reduced, and cleavage to its N-terminus is encouraged, yielding a large differential between the fragment ion peaks adjacent the residue.

residues. In all three cases, the lower m/z peak is more intense than the higher m/z peak (y6 > y7, y8 > y9, and y9 > y10). For y ions, the higher m/z peak of each pair corresponds to the N-terminal cleavage. All three examples of Leu in this spectrum yield a negative N-bias (or positive "C-bias"); the C-terminal peaks are more intense than the N-terminal peaks. This spectrum's results for Leu are in line with the median behavior observed for the broader set of spectra, which showed an N-bias of −0.33 for this residue.

For b ions, the three residues with the highest N-bias were Pro, Gly, and Ser (see Figure 7). At the median, Pro's N-bias was 1.0; the peak representing the C-terminal cleavage product did not appear in the spectrum. Gly and Ser were more moderate influences on fragmentation, with N-biases of 0.47 and 0.35, respectively. Showing the opposite impact was His, which yielded a C-bias of −0.64. The phenomenon, however, was quite variable; the mean interquartile difference for b ion N-biases was 0.86. The variability for residues did not appear to bear a relationship to sample size; the interquartile differences for Cys, Met, and Trp were not substantially different than for the other residues.

Figure 7 shows the N-bias measurements for y ions. The highest N-bias residues for y ions were Pro (0.93), Gly (0.59), and Ser (0.35). Although His was the strongest C-bias influence in b ion intensities, it appeared to have little effect in the y ion series. A C-bias was found for Ile (−0.45), Val (−0.44), and Leu (−0.33). The average interquartile difference for y ion N-biases was 0.74. The variability may have decreased relative to the b series due to the greater intensity (and thus signal-to-noise ratio) of y series peaks. A visual inspection of measured N-biases and cleavage position did not reveal a relationship between the two.

Information from the literature of peptide dissociation mechanisms may explain the strong N-bias calculated for Pro cleavage and the strong C-bias calculated for His cleavage in b ions. Pro is the only cyclic amino acid of the 20 commonly occurring amino acids. It has been noted in the literature that there is a strong preference for Pro to cleave at its N-terminal side, whether the product formed is the b ion or y ion.[30] Because b ions formed at most residues are accepted to have oxazolone structures (see Figure 1)[31−33] and because the b ion that would be formed at the C-terminal side of Pro would require a transition state involving formation of an unstable strained 5−5 bicyclic ring (see Figure 6),[6,8,34] cleavage typically occurs at the N-terminal side of Pro rather than the C-terminal side. This selective cleavage also provides information on the mechanism of the cleavage to form b and y ions. If the mechanisms of formation of b and y ions have no common intermediates, one would not expect that both the b and y ions would show a strong N-bias. The strong N-bias shown in Figure 7 is indirect evidence that a coupled b−y cleavage mechanism[35] occurs for Pro; ring closure to form an oxazolone N-terminal to Pro leads to an ion−molecule complex that can dissociate to either a b or a y ion, depending on whether a proton transfer to the Pro occurs before the two fragments separate.

The C-bias for fragmentation at His has not been reported directly in the literature for a large body of compounds although mechanistic data have been reported that are consistent with this

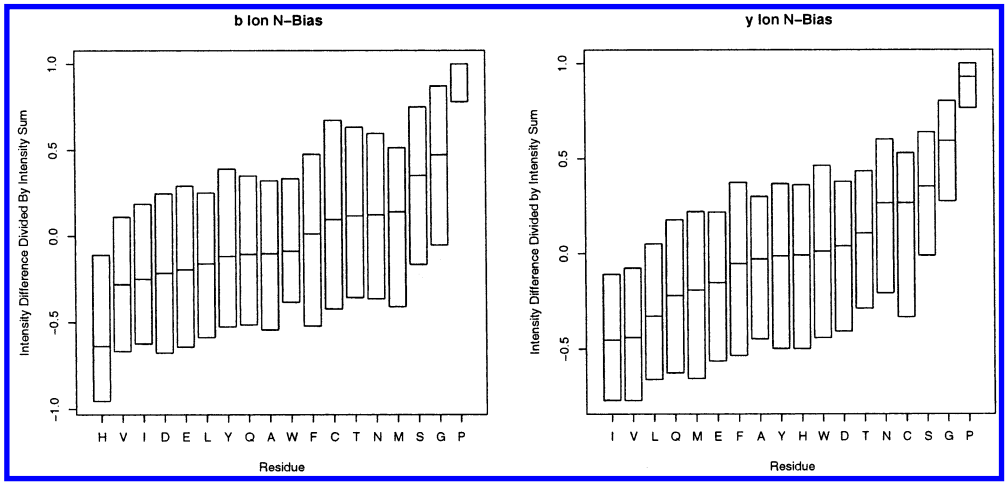

Figure 7. Ratio of intensity differences to intensity sums for individual residues. N-bias measures the extent to which each residue directs local fragmentation. The statistic measures the difference between the N-terminal peak intensity and the C-terminal peak intensity, normalizing this difference by the sum of the two peak intensities. Residues with N-biases of greater absolute value impact local fragmentation to a greater extent. The median bias for each residue is marked by the line across each box, and the upper and lower edges of the boxes represent the 75th and 25th percentiles, respectively. The most distinctive bias toward N-terminal fragmentation is that of Pro. A smaller N-bias appears for Gly and Ser. Hydrophobics Ile, Leu, and Val show a bias toward C-terminal cleavage in y ions, and His shows a pronounced bias toward C-terminal cleavage in b ions.

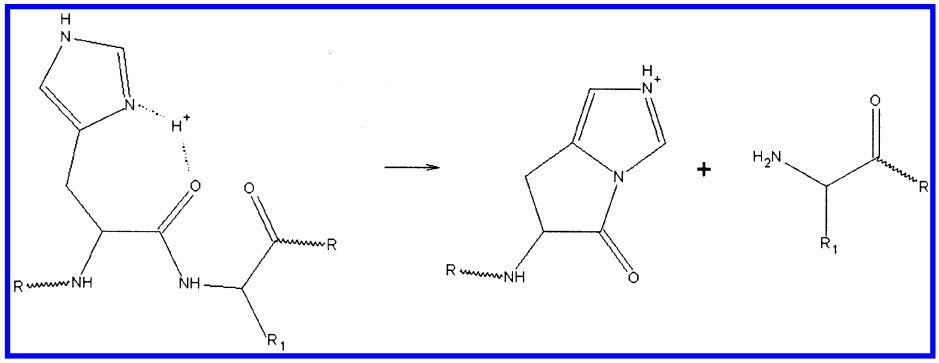

Figure 8. Histidine cleavage. Histidine can form unusual b ions. In normal fragment ion formation, the carbonyl N-terminal to a residue attacks the carbonyl at the residue's C-terminus, resulting in an intermediate that produces a b ion with a single, five-membered ring. The side chain of His may short-circuit that process by attacking its own carbonyl, yielding a b ion with a double ring structure.

bias. Wysocki and co-workers suggested that fragmentation at His occurs via formation of a b ion structure resulting from the side chain's attack on the backbone carbonyl (see Figure 8).[8] O'Hair and co-workers calculated the stability of this 5−5 ring bicyclic structure for the N-acetylated methyl ester of His and found that this structure is more stable than the corresponding protonated oxazolone.[36] A C-bias for cleavage at His that is greater than that at the other residues in Figure 7 is consistent with the suggestion that the His side chain produces a unique b ion structure not formed by the other residues in the analysis.[8,37] To test the hypothesis that other basic side chains may attack the neighboring carbonyl, a set of peptides including internal Arg and Lys residues was produced. The C-bias exhibited by Arg was comparable to that of His (data not shown), indicating that its side chain may also be capable of nucleophilic attack on the carbonyl oxygen, as suggested by Glish.[38]

(2) Neutral Loss by Residue. Neutral losses from fragment ions can complicate tandem mass spectra. A loss of ammonia either before or after fragmentation may reduce fragment ion masses by 17 mass units (often called "*" ions). The median intensity among observed b-17 ions was 0.24% of the spectrum's intensity, corresponding to approximately half the intensity of the median b ion. A total of 55.5% (9744) of predicted ions matched to a peak. The median intensity for observed y-17 ions accounted for 0.18% of the spectrum's intensity; the average y peak is 5 times the size of its ammonia loss peak. Of the predicted ammonia losses, 44.7% (7850) could be matched to observed peaks.

To determine the influence of residue content on the production of these neutral loss ions, the fragment ions associated with neutral loss peaks more intense than the observed median were segregated from the other fragment ions of the series. The residue content of these ions was compared to the residue content of fragment ions of the same type. The b5 ion of ASGEIVSINQINEAHPTK, for example, has the sequence ASGEI. If this ion showed a neutral loss of 17 mass units more intense than the median, this b5 ion would contribute to the composition ratios of Ala, Ser, Gly, Glu, and Ile. The results for Ala may clarify this process. The residue is found in 49.3% of predicted y ions (8650 of 17 538). Of the y ions with intense identifiable ammonia loss peaks, 48.3% (1894 of 3925) contain Ala. The frequency of Ala content for ions with substantial ammonia losses is not substantially different from the frequency of Ala content for all ions. There appears to be no relationship between Ala content and neutral loss of ammonia. The graphs in Figures 9 and 10 show the ratios of these percentages, so that a ratio of 1.0 indicates that a particular residue was found neither more nor less often than expected.

For b ions, the result shows that prominent ammonia-losing fragment ions are 32% more likely to contain Asn and 7% more likely to contain Gln than b ions in general (see Figure 9). On the other extreme, His (−30%) and Pro (−29%) are underrepresented in b ions that lose ammonia. The reliability for Met, Trp, and Cys may be limited by the number of peptides containing these residues (178, 178, and 134, respectively).

The case of y ions shows similarities to and differences from that of b ions (see Figure 9). Fragments that are 17 Da less massive than y ions are 28% more likely to contain Asn and 26% more likely to contain Gln than y ions in general. Ions from the y series that lose ammonia are 16% more likely to contain His, the opposite effect seen in b ions. The effect of His was observed to depend on the C-terminal residue of the peptide; fragment ions ending in Lys and containing His were more likely to produce neutral losses than those ending in Arg (data not shown). Pro content remains lower (−9%) in ammonia losing y ions, but not to the extent seen in b ions.

Fragment ions can also lose water molecules, a mass shift of 18 Da (sometimes referred to as "°" ions). Although ion trap mass spectrometers putatively feature unit mass resolution, some amount of conflation between ammonia and water loss ion peaks is to be expected. The median intensity of observed b-18 ions was 0.15% of the spectrum's intensity, with 66% (11 612) of calculated b-18 ions matching to a peak. The corresponding intensity for y-18 ions was 0.10%, with 51% (9025) of these ions matching to a peak. The intensity medians for ions showing a loss of 18 mass units

(31) Arnott, D.; Kottmeier, D.; Yates, N.; Shabanowitz, J.; Hunt, D. F. *Proc. 42nd ASMS Conf. Mass Spectrom. Allied Topics*, Chicago, IL, 1994; p 470.

(32) Nold, M. J.; Wesdemiotis, C.; Yalcin, T.; Harrison, A. G. *Int. J. Mass Spectrom. Ion Processes* **1997**, *164*, 137−153.

(33) Yalcin, T.; Khouw, C.; Csizmadia, I. G.; Peterson, M. R.; Harrison, A. G. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 1164−1174.

(34) Vaisar, T.; Urban, J. *J. Mass Spectrom.* **1996**, *31*, 1185−1187.

(35) Paizs, B.; Lendvay, G.; Vékey, K.; Suhai, S. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 525−533.

(36) Farrugia, J. M.; Taverner, T.; O'Hair, R. A. J. *Int. J. Mass Spectrom.* 2001, 209, 99−112.

(37) Farrugia, J. M.; O'Hair, R. A. J.; Reid, G. E. *Int. J. Mass Spectrom.* 2001, 210, 71−87.

(38) Vachet, R. W.; Asam, M. R.; Glish, G. L. *J. Am. Chem. Soc.* **1996**, *118*, 6252−6256.

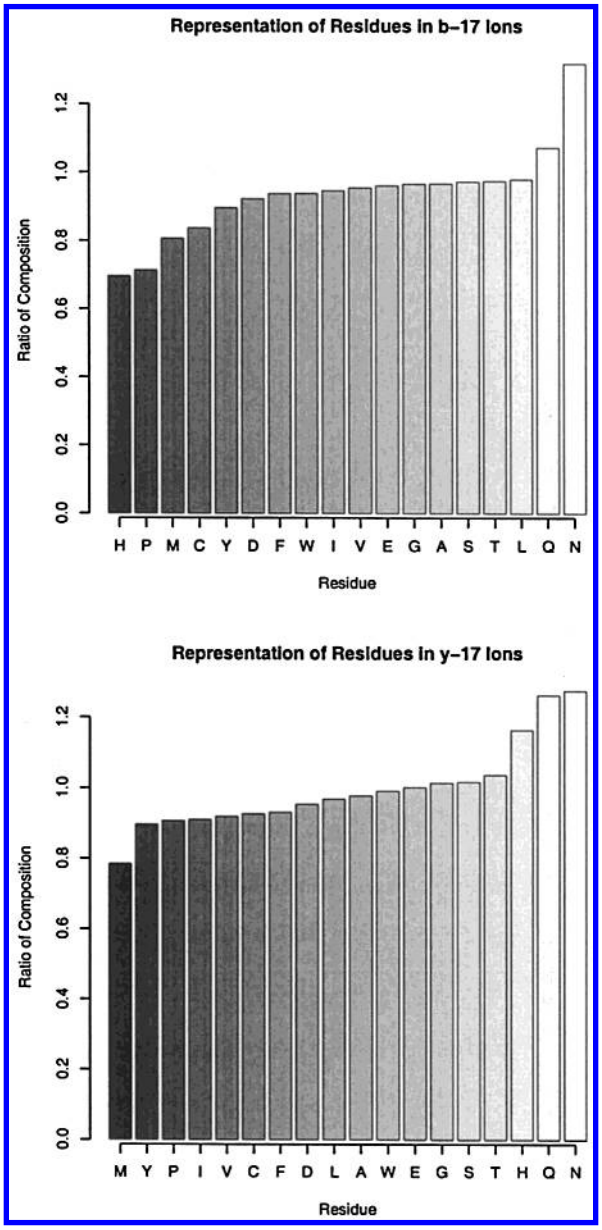

Figure 9. Composition ratio versus the residues in b-17 and y-17 ions. Fragment ions that exhibit prominent loss of ammonia are more likely to contain Asn or Gln than fragment ions in general. Ammonia loss is enhanced by His for y ions, but His suppresses the loss for b ions. Neutral loss of ammonia may be diminished by the presence of Pro and Met. The vertical axis shows the ratio between the sequence composition of the ions displaying intense loss peaks and the sequence composition of fragment ions from the appropriate series.

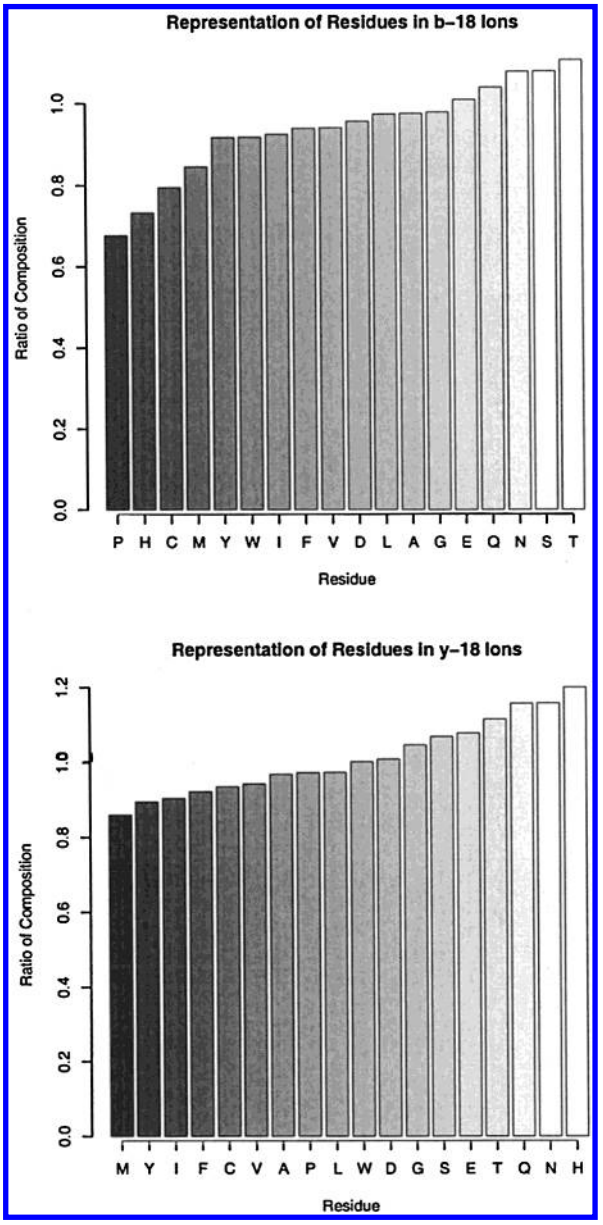

Figure 10. Composition ratio versus the residues in b-18 and y-18 ions. Fragment ions that exhibit prominent loss of water are more likely to contain Ser, Thr, or Glu. As seen in ammonia loss, His appears to enhance loss in y ions but diminish it in b ions. The mass accuracy of the ion trap, combined with the window for identifying peaks in DaughterDB, may result in the misidentification of some ammonia loss peaks as water loss peaks, thus resulting in Asn and Gln's ranking above.

are well below the corresponding medians for the ions exhibiting 17 mass unit losses, though a larger proportion of these calculated ions match to observed peaks.

The differences in residue content for water loss ions do not appear to be as marked as those in ammonia losses. Intense b-18 ions are 11% more likely to contain Thr and 8% more likely to contain Ser than b fragment ions in general (see Figure 10). Asn and Gln are also more common among these ions (8% and 4%, respectively). The appearance of these residues for both water and ammonia losses may have resulted from the limited mass accuracy of the ion trap, which can yield m/z measurement errors sufficient to cause DaughterDB to identify a particular peak as both a water loss and an ammonia loss. Another point of similarity between both classes of neutral loss is that both b-18 ions and b-17 ions diminish by the content of Pro or His.

Compositions of y-18 ions do not appear to be substantially different from those of y-17 ions (see Figure 10). Again, His is prominent in ions losing water (an increase of 20%). As with ammonia losses, His content's contribution to water neutral losses was observed to be dependent upon the presence of Lys at the C-terminal residue of the peptide. Asn and Gln are also common (16% each). Thr and Ser, which are found preferentially in b-18 ions, are among the residues increasing the presence of y-18 ions, but they do not stand out substantially.

Peptide dissociation trends reported in the literature can help explain and support the dependence of neutral losses of water and ammonia from sequence ions on amino acid residue composition. The formation of sequence ions competes with neutral loss ion formation due to side-chain interactions. Research has corroborated the relationship of Asn and Gln content to b-17 and y-17 ion formation (see Figure 9). The roles of Ser and Thr content in the formation of b-18 ions and of His in the formal loss of water from y ions (see Figure 10) have also been explored.

The reported loss of ammonia from Asn and Gln side chains in peptides corresponds to reported results for individual amino acids. O'Hair and co-workers[37] reported the fragmentation patterns for a variety of protonated *N*-acyl amino acid methyl esters, including Asn and Gln. Both of these ions underwent the neutral loss of ammonia readily, the source of which could not be from the N-terminus due to N-acylation. The only remaining possible sources of the ammonia loss were the side chains of Asn and Gln residues.

Ballard and Gaskell[39] investigated the neutral loss of water from singly protonated peptide ions and provided evidence for three dehydration reactions through ^{18}O-labeling studies: (i) loss of water from the C-terminal carboxylic acid or acidic residues, (ii) loss of water from the side-chain hydroxyl group of Ser or Thr residues, and (iii) loss of water from the amide carbonyl of the peptide bond. Their study concluded that a single peptide may follow any of several competing dehydration pathways. Due to the variety of peptide sequences included in the statistics reported here, some residue-specific trends may be masked by the presence of other amino acids. All tryptic peptides included in this study contained a C-terminal carboxylic acid and several amide carbonyls in peptide bonds, complicating the determination of which pathway led to the formation of observed water loss ions.

Water loss may precede the formation of fragment ions. Separate ^{18}O-labeling MS/MS experiments of Thr-containing peptides and corresponding ab initio calculations for these molecules have indicated that the neutral loss of water occurs from the precursor ion $[M + H^+]$ exclusively from the side-chain hydroxyl group.[40] The interaction of Ser's side chain with the backbone is a potential explanation for the neutral loss of water in low-energy CID MS/MS experiments. The currently accepted mechanism for protein splicing could be considered a solution-phase counterpart to this mechanism in which the hydroxyl oxygen of the Ser side chain attacks the carbonyl carbon to form a five-membered intermediate.[41] The presence of His in a peptide sequence has also been reported to enhance the neutral loss of water from sequence ions, perhaps through a neighboring group pathway involving induction of water loss by the nucleophilic side chain on the adjacent protonated carbonyl.[36] The presence of His in peptides could promote the neutral loss of water (as seen in Figure 10).

(39) Ballard, K. D.; Gaskell, S. J. *J. Am. Soc. Mass Spectrom.* **1993**, *4*, 477−481.
(40) van Dongen, W. D.; de Koster, C. G.; Heerma, W.; Haverkamp, J. *Rapid Commun. Mass Spectrom.* **1995**, *9*, 845−850.
(41) Noren, C. J.; Wang, J. M.; Perler, F. B. *Angew. Chem., Int. Ed.* **2000**, *39*, 450−466.

CONCLUSION

The intensities of fragment ion peaks are a function of their fragment types, the locations of cleavage within the peptide, and the residues that are adjacent to the cleavage. The residue content of fragment ions, in turn, plays a role in the formation of secondary fragment ions such as ions formed by the neutral loss of water or ammonia. While intensity may be quite variable, particularly for the most prominent peaks in spectra, the information encoded in fragment ion intensity should be used in algorithms that process these spectra. The trends observed in fragment ion intensity can be developed into improved models of peptide fragmentation.

Existing algorithms for identifying database peptides from tandem mass spectra employ simple fragmentation models to generate spectra for comparison to observed spectra. SEQUEST, for example, models all y ions to be a uniform intensity, with b ions a uniform lower intensity. The above results establish that even spectra that SEQUEST can correctly identify show distinctive patterns in intensity for each ion series and marked differences in fragmentation neighboring particular amino acid residues. An algorithm that uses these trends in intensity as part of its fragmentation model should offer improved accuracy in peptide identification, correctly identifying spectra that deviate from simple fragmentation models.

The challenge of inferring sequence directly from tandem mass spectra (de novo rather than from a database) requires that all information present in the spectrum be used optimally. The measures calculated in this research can be applied directly to this problem. Instead of relying solely upon m/z data to consider sequence possibilities, algorithms based on improved fragmentation models can use the information present in recorded fragment ion intensity to aid the process of sequence validation.

This analysis is limited to spectra for doubly charged peptides resulting from a complete trypsin digest. While this class comprises the identifications most common in proteomic experiments, other classes of peptides are also significant and produce distinct intensity patterns. Triply charged peptide spectra, in particular, differ in having doubly charged fragment ions intermixed with the singly charged ones. The locations of basic residues within the peptide sequence instead of or in addition to the terminal basic residue of tryptic peptides can also change the pattern of intensities observed in a spectrum. Characterizing these classes of spectra will help highlight the mechanisms by which peptides fragment and set the stage for a second generation of sequence identification software.

ACKNOWLEDGMENT

D.L.T. was supported by a National Science Foundation Graduate Research Fellowship. D.L. was supported by National Institutes of Health Grant R33 CA81665-04.

Received for review September 10, 2002. Accepted November 15, 2002.

AC026122M