

LETTERS

Intrinsic Amino Acid Size Parameters from a Series of 113 Lysine-Terminated Tryptic Digest Peptide Ions

Stephen J. Valentine, Anne E. Counterman, Cherokee S. Hoaglund-Hyzer, and David E. Clemmer*

Department of Chemistry, Indiana University, Bloomington, Indiana 47405

Received: September 30, 1998; In Final Form: January 14, 1999

Cross sections for mixtures of tryptic digest peptide ions formed by electrospray ionization have been measured by a new ion mobility/time-of-flight mass spectrometry technique. Analysis of a series of 113 peptides containing 5–10 residues and having a single lysine group located at the C-terminal end show that cross sections are largely dependent upon the amino acid composition of each peptide. Reduced cross sections (which take into account differences in mass) are found to correlate with the fractions of nonpolar or polar aliphatic residues. Average intrinsic contributions to size for individual amino acid residues (referred to as intrinsic size parameters) have been obtained by solving a system of equations that relates the 113 reduced cross sections to the occurrence frequency of each residue within the different sequences. These parameters fall into families according to the physical sizes and chemical properties of the amino acids; contributions to cross section from nonpolar residues are significantly larger than those from polar groups. Calculated cross sections that are obtained by combining intrinsic size factors with peptide sequences are remarkably accurate: >90% of calculated values are within 2% of experimental measurements.

Recently mass spectrometry-based techniques have been used to examine the conformations and folding of peptides and proteins in the absence of solvation effects.^{1,2} Information about the structural similarities of gas-phase and native conformations would illuminate intrinsic factors that influence folding.³ To date, cross sections for about a dozen protein and peptide ion systems have been reported.^{2,4} We have recently developed a new ion mobility/time-of-flight method that makes it possible to measure cross sections and mass-to-charge (m/z) ratios for mixtures of ions in a single experiment.⁵ Using this approach, we have accumulated a database containing cross sections for 660 different peptide ions. Here, we report results for a subset of 113 peptides having the general form $[(Xxx)_n\text{Lys} + \text{H}]^+$, where $n = 4-9$ and Xxx is any naturally occurring amino acid except Lys, Arg, His, or Cys.⁶ Analysis of the results for these

related peptide sequences shows that there are correlations of amino acid composition with cross section. The influence of each amino acid on cross section has been quantified in terms of an intrinsic size parameter by solving a system of equations that relates amino acid occurrence frequency to cross section for different peptide sequences in the array of data. Intrinsic size parameters can be combined with sequence information to accurately calculate cross sections for a remarkably large number of the peptides (90% of calculated values are within ~2.0% of experiment). These results demonstrate an important first step in prediction of structure from sequence in a simplified gas-phase environment which may complement efforts to predict structure from sequence in condensed phase.⁷

A consideration in the design of the database was that peptide cross sections should reflect elements of intrinsic structure

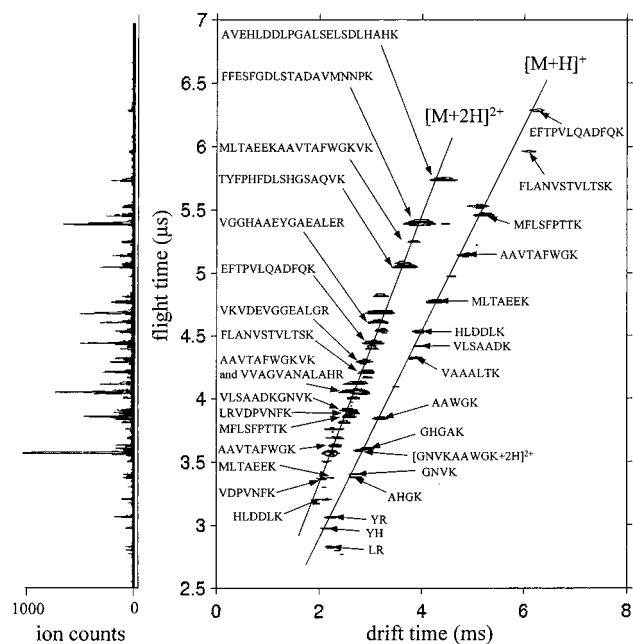


Figure 1. Contour plots of nested drift-time (bottom axis) and flight time (left axis) data for a mixture of peptide ions that were formed by electrospray ionization of a tryptic digest of the α - and β -chains of bovine hemoglobin. The drift time axis has been scaled to a He pressure of 2.000 Torr. Contours are shown on a 30 point scale which removes all features that contain fewer than 30 ion counts. Peak labels correspond to sequences that have been assigned to fragments that are expected from tryptic digestion based on comparison of measured m/z ratios (from ion flight times) to calculated molecular weights. The drift times (used to determine cross sections) are taken as the maximum peak height in all cases. The plot on the left shows a time-of-flight distribution obtained by compression of the drift time axis.

relevant to common protein sequences. To accomplish this, mixtures of peptides were generated by tryptic digestion of 34 common proteins such as cytochrome *c*, myoglobin, and albumin.⁸ The peptide mixtures were electrosprayed⁹ into the gas phase and analyzed using an ion mobility/time-of-flight mass spectrometry method that allows drift times (mobilities) and flight times (m/z ratios) for all components to be determined in a single experiment as described previously.⁵ The mobility of each peptide ion depends on its average collision cross section and its charge state.¹⁰ Compact conformers have smaller cross sections than elongated ones.¹¹ An example data set is shown as a contour plot in Figure 1 for peptides generated by digestion of α - and β -chains of hemoglobin. Combined drift time and flight time data for tryptic digests typically show families of $[M + H]^+$ and $[M + 2H]^{2+}$ ions.¹⁰ For species having similar m/z , $[M + 2H]^{2+}$ peptides usually have higher mobilities than $[M + H]^+$. Peptide sequences and collision cross sections are determined from the measured flight times (in the mass spectrometer) and drift times (in the ion mobility instrument),⁵ respectively. Only data from peaks that are unambiguously assigned to expected digest fragments are included in the database. The single data set shown in Figure 1 yields cross sections for 33 different ions. The complete database, which will be reported elsewhere,¹² contains information about the digested protein, peptide sequence, molecular weight, and cross section for 660 different ions, including 420 singly protonated peptides ranging in size from 2 to 15 residues and 240 doubly protonated peptides with 4–24 residues.

Figure 2 shows cross sections for the related series of 113 $[(Xxx)_n\text{Lys} + H]^+$ peptides. Plots of other subsets as

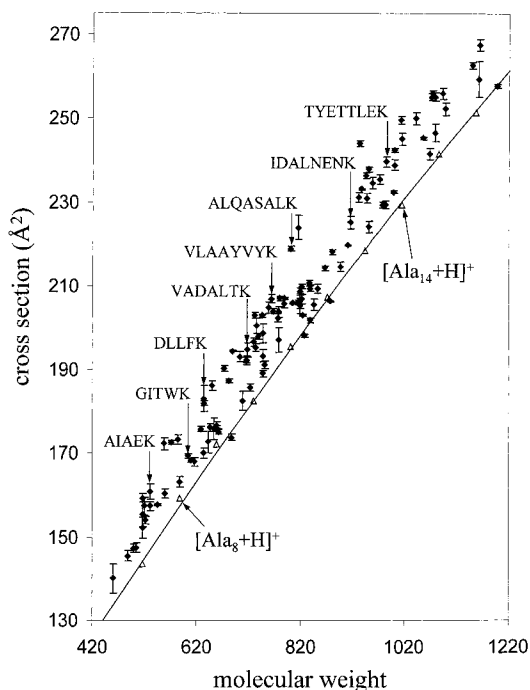


Figure 2. Cross sections for 113 $[(Xxx)_n\text{Lys} + H]^+$ peptides (solid diamonds). Uncertainties are shown as error bars and correspond to one standard deviation (when three or more measurements have been made) or to a range (when only two values are available). Labels show several sequences (selected randomly) for specific points. The open triangles show cross sections that were obtained for singly protonated polyalanine peptides containing 7–16 alanine residues. The solid line shows a polynomial best fit to the polyalanine data that is used in the determination of reduced cross sections. See text for discussion.

C-terminal arginine peptides appear similar. As a general trend, cross sections increase with molecular weight. However, values for sequences having similar molecular weights but different amino acid sequences vary by up to 10–15% over the molecular weight studied. To account for differences in cross section that arise due to differences in mass, we have normalized the $[(Xxx)_n\text{Lys} + H]^+$ cross sections with respect to values measured for a series of singly protonated polyalanine peptides with 7–16 residues (also shown in Figure 2). Polyalanine peptides of this size have roughly spherical (globular) conformations where the protonated N-terminal amino group is self-solvated by a large portion of the peptide chain (primarily through contacts with electronegative backbone carbonyl groups).¹³ Nearly all of the C-terminal lysine peptides are larger than a polyalanine of comparable mass. We define a reduced cross section as the value for each peptide divided by the cross section of polyalanine at an identical molecular weight (determined from a polynomial fit to the polyalanine data).

Trends regarding the influence of specific residue types upon cross section can be seen by plotting the reduced cross sections as a function of the fraction of different types of amino acids that are present in each sequence. Figure 3a shows a plot of reduced $[(Xxx)_n\text{Lys} + H]^+$ cross sections against the fraction of peptide mass corresponding to nonpolar aliphatic (Ala, Ile, Leu, Met, Val) residues. A positive correlation of the reduced cross section with the mass fraction of nonpolar residues is observed. The opposite trend is observed for the polar aliphatic (Asp, Glu, Asn, Gln, Ser, Thr) residues. This type of analysis can be extended by relating the frequency of occurrence of each amino acid to the reduced cross section in a system of 113 equations. In this case, the intrinsic contributions to size of each

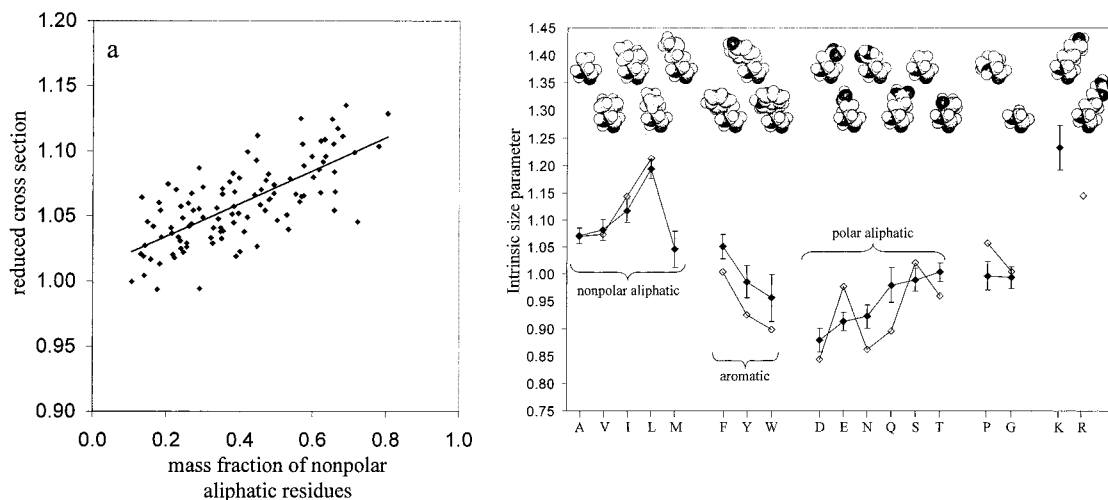


Figure 3. (a) Reduced cross sections for the $[(Xxx)_n\text{Lys} + \text{H}]^+$ peptides that are shown in Figure 2 as a function of the mass fraction of nonpolar aliphatic residues. Here, the mass fraction is defined as the sum of mass contributions from nonpolar aliphatic residues divided by the neutral peptide mass. The solid line represents a fit to the data, given by $\Omega_j(\text{exp})/\Omega_j(\text{PA}) = 0.130 m_{\text{np}} + 1.008$ ($R^2 = 0.531$), where $\Omega_j(\text{exp})/\Omega_j(\text{PA})$ is the reduced cross section and m_{np} is the mass fraction of nonpolar residues for each j peptide. (b) Intrinsic size parameters extracted for individual amino acids from a database of 113 Lys-terminated peptides (solid diamonds) and 38 Arg-terminated peptides (open diamonds). No parameters are given if the amino acid occurred in fewer than five different sequences. Uncertainties correspond to one standard deviation about the mean values that are determined by solving the system of reduced cross-section equations (see text for discussion). Atomic structures for amino acids including the peptide backbone are also shown.

amino acid can be written as an unknown parameter p_i which is related to the reduced cross sections, $\Omega_j(\text{exp})/\Omega_j(\text{PA})$, by

$$\frac{\sum_i n_{ij} p_i}{\sum_i n_{ij}} = \frac{\Omega_j(\text{exp})}{\Omega_j(\text{PA})}$$

where n_{ij} corresponds to the number of times an amino acid i occurs in each sequence j . The system of equations is solved for the 17 p_i parameters (amino acids except for Cys, Arg, and His) by a linear algebra regression method.¹⁴ The resulting best-fit average size parameters for each amino acid in the 5–10 residue Lys-terminated peptides are displayed in Figure 3b. Contributions to peptide size from individual residues vary by as much as ~40% and appear to fall into groups according to the chemical nature of the amino acid. A similar analysis of 38 analogous Arg-terminated peptides ($[(Xxx)_n\text{Arg} + \text{H}]^+$) from the database yields values for average size parameters (Figure 3b) that are similar to those determined from the lysine data for almost all amino acids.

Several trends regarding the intrinsic contributions of amino acids to cross section are apparent. First, the largest contributions come from the nonpolar Ala (1.07 ± 0.01), Val (1.08 ± 0.02), Ile (1.12 ± 0.02), and Leu (1.19 ± 0.02) residues. Contributions from polar groups such as Asp (0.88 ± 0.02), Glu (0.91 ± 0.02), and Asn (0.92 ± 0.02) are much smaller. The different behaviors of these residues can be understood by considering differences in long-range interactions of the different residue types. Dipole-dipole interactions between the Asp, Glu, and Asn residues or with the polar backbone should increase packing and decrease cross sections. Long-range interactions between these residues and the charged lysine will also cause conformers to contract. Second, contributions to cross section (especially for aliphatic chains) depend on the physical sizes of the side chains. An increase in the length of these aliphatic side chains leads to a larger intrinsic contribution to cross section. This can be

observed in the ~3–10% increases in intrinsic sizes observed for the following pairs of homologous side chains: Ile and Leu > Val, Glu > Asp, and Gln > Asn. Finally, some differences may also arise from differences associated with the residue-helium collision dynamics, a factor that can have a pronounced effect on the calculation of collision cross sections.^{15,16}

Large polar aromatic groups [Trp (0.96 ± 0.04) and Tyr (0.99 ± 0.03)] contribute surprisingly little to size; the nonpolar aromatic Phe (1.05 ± 0.02) residue is slightly larger but still smaller than Ala (1.07 ± 0.01). The intrinsic size parameter of Met is particularly interesting because it is the smallest of the nonpolar aliphatic groups even though the side chain is comparable in length to that of Leu (Figure 3b). We have grouped Met as a nonpolar aliphatic residue because it is normally characterized as such in solution.¹⁷ The intrinsic size parameter obtained from these studies suggests that in the gas phase Met should be grouped as a large polar residue. The relatively large polarizability of the sulfur atom and small dipole moment should undergo relatively long-range charge-induced dipole interactions that cause the system to contract. Finally, lysine (1.23 ± 0.04) is substantially larger than all other values, presumably because this residue is only located at the end of the peptide. The arginine residue, located at the same end position, displays a similar trend in that system.

The above determination of intrinsic size factors implicitly assumes that the influence of different residues on structure is similar for different sequences. It is possible to test the general applicability of these parameters by combining values with peptide sequences to calculate cross sections. For example, from the Asp (0.88), Leu (1.19), Phe (1.05), and Lys (1.23) parameters, we calculate a reduced cross section for DLLFK of 1.11. This gives a calculated cross section of 184 \AA^2 , in excellent agreement with experiment ($183 \pm 3 \text{ \AA}^2$). The comparison of the calculated reduced cross sections with the reduced experimental results for all 113 of the 5–10 residue C-terminal lysine peptides are shown in Figure 4a. These simple parameters accurately capture variations in the physical sizes

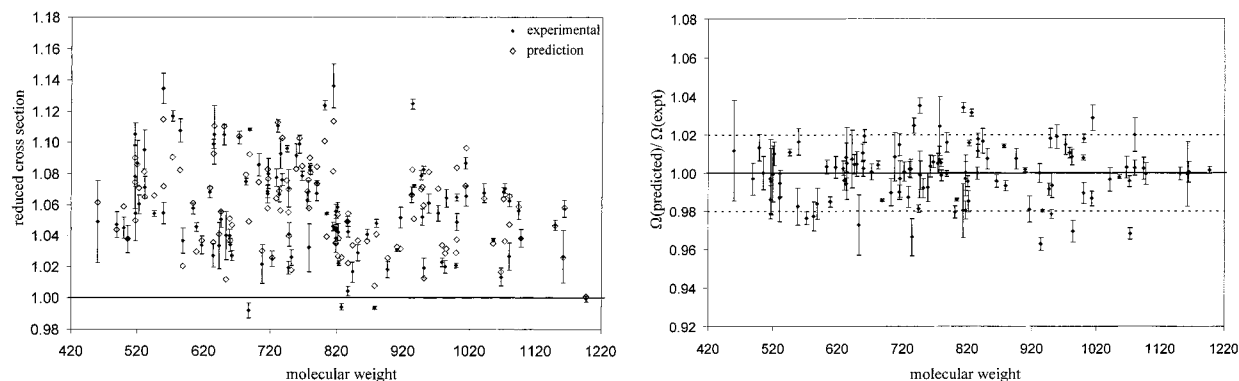


Figure 4. (a) Comparison of experimental reduced cross sections (solid diamonds) and values that have been calculated (open diamonds) from sequences using the intrinsic amino acid size parameters shown in Figure 3b for the $[(Xxx)_n\text{Lys} + \text{H}]^+$ peptides. See text for discussion. (b) Ratio of calculated to experimental cross sections for all 113 peptides. A prediction that is in perfect agreement with experiment has a ratio of 1.00. The dotted lines that are displayed show that 98 of the 113 peptides (90%) are predicted to within 2.0% of experiment.

of most of the different peptide sequences. The ratio of predicted to experimental cross sections for every peptide is shown in Figure 4b. Ninety-eight (~90%) of the calculated values fall within 2.0% of the corresponding experimental value.¹⁸ This result shows that amino acid composition is an important factor in establishing cross section.

Several studies have recently shown that amino acid sequence also plays an important role in establishing some structures. Cassidy has shown that the position of basic lysine groups in a series of lysine-doped glycine peptides (dodecamers) influences the rate of proton-transfer reactions.¹⁹ Hill and co-workers have used high-resolution ion mobility methods to separate a mixture of two four-residue peptides that vary in sequence.²⁰ Jarrold and co-workers have shown that the position of the charged basic residue can stabilize or destabilize the helix dipole in lysine-doped alanine peptides.²¹ Other studies from our lab also show that the position of charged residues as well as other groups can influence cross section. We interpret the remarkable agreement between calculated and experimental cross sections in these small peptides as a reflection of the importance of composition rather than sequence in establishing cross section. It seems likely that many of the deviations between predicted and experimental results (Figure 4) arise from differences in structure, which must depend on sequence. We are currently investigating simple sequence parameters that take into account residue positions in order to increase the accuracy of calculated values.

In summary, we have shown that it is possible to extract average intrinsic contributions to cross sections from individual amino acid residues by analyzing a series of C-terminal lysine peptide ions. Cross sections in this family of sequences depends largely upon the chemical nature and physical sizes of the amino acid side chains and to a lesser extent upon amino acid sequence. In the gas phase, long-range charge-dipole and dipole-dipole interactions involving polar groups cause the peptide to contract; cross sections for sequences with large numbers of polar groups are expected to be relatively small. Larger contributions are observed for residues with nonpolar side chains (presumably because charge-induced dipole and other interactions are weaker for these residues). The Met residue (normally considered a nonpolar residue in solution) appears to be best classified as a polar residue in the gas phase.

An important aspect of approaches for prediction of structure from sequence involves the interplay between homology (recognition of similarities between the sequence and known structures), methods that account for the physical and chemical interactions, and large amounts of structural data.^{7d,22} The type of structural data reported here should complement the under-

standing of surface accessibility, packing, and hydrophobic interactions from condensed-phase studies²³ and aid in refinement of theoretical methods. Predictions of cross section from sequence may also find analytical applications as a means of assigning peaks in new ion mobility methods that utilize gas-phase strategies for a rapid separation step for mass spectrometry.¹⁰

Acknowledgment. This work was supported by grants from the NSF (CAREER, Grant No. CHE-9625199) and NIH (Grant no. 1R01GM55647-01), with additional PRF support (Grant no. ACS-PRF 31859-G4).

References and Notes

- (1) Examples of the different techniques that have been used can be found in the following: Winter, B. E.; Light-Wahl, K. J.; Rockwood, A. L.; Smith, R. D. *J. Am. Chem. Soc.* **1992**, *114*, 5897–5898. Covey, T. R.; Douglas, D. J. *J. Am. Soc. Mass Spectrom.* **1993**, *4*, 616–623. Wood, T. D.; Chorush, R. A.; Wampler, F. M.; Little, D. P.; O'Connor, P. B.; McLafferty, F. W. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 2451–2454. Gross, D. S.; Schnier, P. D.; Rodriguez-Cruz, S. E.; Fagerquist, C. K.; Williams, E. R. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 3143–3148. Kaltashov, I. A.; Fenselau, C. C. *Proteins* **1997**, *27*, 165–170. Reimann, C. T.; Sullivan, P. A.; Axelsson, J.; Quist, A. P.; Altmann, S.; Roepstorff, P.; Velazquez, I.; Tapia, O. *J. Am. Chem. Soc.* **1998**, *120*, 7608–7616.
- (2) (a) Shelimov, K. B.; Jarrold, M. F. *J. Am. Chem. Soc.* **1997**, *119*, 2987–2994. (b) Valentine, S. J.; Anderson, J.; Ellington, A. E.; Clemmer, D. E. *J. Phys. Chem. B* **1997**, *101*, 3891–3900.
- (3) Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 2426–2427.
- (4) von Helden, G.; Wytenbach, T.; Bowers, M. T. *Science* **1995**, *267*, 1483–1485. Clemmer, D. E.; Hudgins, R. R.; Jarrold, M. F. *J. Am. Chem. Soc.* **1995**, *117*, 10141–10142. Wytenbach, T.; von Helden, G.; Bowers, M. T. *J. Am. Chem. Soc.* **1996**, *118*, 8355–8364. Shelimov, K. B.; Clemmer, D. E.; Hudgins, R. R.; Jarrold, M. F. *J. Am. Chem. Soc.* **1997**, *119*, 2240–2248. Valentine, S. J.; Counterman, A. E.; Clemmer, D. E. *J. Am. Soc. Mass Spectrom.* **1997**, *8*, 954–961. Wytenbach, T.; Bushnell, J. E.; Bowers, M. T. *J. Am. Chem. Soc.* **1998**, *120*, 5098–5103. Counterman, A. E.; Valentine, S. J.; Srebalus, C. A.; Henderson, S. C.; Hoaglund, C. S.; Clemmer, D. E. *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 743–759.
- (5) Hoaglund, C. S.; Valentine, S. J.; Sporleder, C. R.; Reilly, J. P.; Clemmer, D. E. *Anal. Chem.* **1998**, *70*, 2236–2242.
- (6) On the basis of the relative basicities of the different amino acid residues, we expect peptides to be protonated at the C-terminal lysine residue. We cannot rule out the possibility of salt-bridged structures as proposed recently, see: Schnier, P. D.; Price, W. D.; Jockusch, R. A.; Williams, E. R. *J. Am. Chem. Soc.* **1996**, *118*, 7178–7189.
- (7) (a) Matthews, B. W. *Annu. Rev. Phys. Chem.* (Rabinovitch, B. S., ed.) **1976**, *27*, 493–523. (b) Chou, P. Y.; Fasman, G. D. *Adv. Enzymol. Relat. Areas Mol. Biol.* **1978**, *47*, 45–148. (c) Kite, J. *Structure in Protein Chemistry*; Garland Publishing: New York, 1995. (d) Benner, S. A.; Cannarozzi, G.; Gerloff, D.; Turcotte, M.; Chelvanayagam, G. *Chem. Rev.* **1997**, *97*, 2725–2843.
- (8) Digests were prepared by addition of 150 μL of 0.2 mg/mL of trypsin (Sigma, sequencing grade) solution in 0.2 M ammonium bicarbonate (EM Science) to 0.5 mL of a 20 mg/mL solution of protein. After incubation

for 20 h at 37 °C, the trypsin was filtered from the digest using a microconcentrator (microcon 10, Amicon, Inc.) and the peptides that remained were lyophilized. For a discussion of general techniques, see: *Protein Sequencing: A Practical Approach*; Findlay, J. B. C.; Gelsow, M. J., Eds.; IRL Press, Oxford, 1989; p 43.

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

(10) Valentine, S. J.; Counterman, A. E.; Hoaglund, C. S.; Reilly, J. P.; Clemmer, D. E. *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 1213–1216.

(11) von Helden, G.; Hsu, M. T.; Kemper, P. R.; Bowers, M. T. *J. Chem. Phys.* **1991**, *95*, 3835–3837.

(12) Counterman, A. E.; Clemmer, D. E. Work in progress.

(13) Samuelson, S. O.; Martyna, G. J. *J. Chem. Phys.* in press. Hudgins, R. R.; Jarrold, M. F. Private communication. Clemmer, D. E. Unpublished results.

(14) Leon, S. J. *Linear Algebra with Applications*, 3rd ed.; Macmillan: New York, 1990; p 208–211.

(15) Shvartsburg, A. A.; Jarrold, M. F. *Chem. Phys. Lett.* **1996**, *261*, 86.

(16) Chen, Y.-L.; Collings, B. A.; Douglas, D. J. *J. Am. Soc. Mass Spectrom.* **1997**, *8*, 681.

(17) Voet, D.; Voet, J. G. *Biochemistry*, 2nd ed.; Wiley: New York, 1995; p 58–60.

(18) Technically, the comparison of our calculated and experimental cross sections is a retrodiction. That is, the experimental data were used to derive the parameters for the calculated cross sections. In some cases (which we have now included in the 113 peptides shown), peptide cross sections were recorded after initial model parameters (based on fewer peptide cross sections) had been derived. The level of accuracy from these bona fide predictions was near that shown in Figure 4. Other reports of the application of these parameters (and others) to other peptide types from the database are forthcoming from our laboratory: Valentine, S. J.; Counterman, A. E.; Clemmer, D. E. Work in progress.

(19) Cassady, C. J. *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 716–723.

(20) Asbury, G. R.; Wu, C.; Siems, W. F.; Hill, H. H. *Pittcon*, 1998, New Orleans, LA, oral presentation of abstract no. 568.

(21) Hudgins, R. R.; Ratner, M. A.; Jarrold, M. F. *J. Am. Chem. Soc.* **1998**, *120*, 12974–12975.

(22) Dudek, M. J.; Ramnarayan, K.; Ponder, J. W. *J. Comput. Chem.* **1998**, *19*, 548–573.

(23) Richards, F. M. *J. Mol. Biol.* **1974**, *82*, 1–14. Privalov, P. L.; Gill, S. J. *Adv. Protein Chem.* **1988**, *39*, 191–235. Lee, C.; Subbiah, S. *J. Mol. Biol.* **1991**, *217*, 373–388. Hunt, N. G.; Gregoret, L. M.; Cohen, F. E. *J. Mol. Biol.* **1994**, *241*, 214–225. Pascarella, S.; De Persio, R.; Bossa, F.; Argos, P. *Proteins* **1998**, *32*, 190–199. Koretke, K. K.; Luthey-Schulten, Z.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 2932–2937.