Determination of Sequence-Specific Intrinsic Size Parameters from Cross Sections for 162 Tripeptides

Amy E. Hilderbrand and David E. Clemmer*

Department of Chemistry, Indiana University, Bloomington, Indiana 47405

Received: February 11, 2005; In Final Form: April 13, 2005

Ion mobility and mass spectrometry techniques have been used to measure cross sections for 162 tripeptide sequences (27 different sets of six sequence isomers). The isomers have the general forms ABC, ACB, BAC, BCA, CAB, and CBA, where A corresponds to the amino acids Asp, Glu, or Gly, B corresponds to Lys, Arg, or Leu, and C corresponds to Phe, Tyr, or Ser. From these data, we derive a set of size parameters for individual amino acids that reflect the position of the amino acid in the sequence. These sequence-specific intrinsic size parameters (SSISPs) are used to retrodict cross-section values for the 162 measured sequences and to predict cross sections for all remaining tripeptide sequences (567 different sequences) that are comprised of these residues. In several types of peptide compositions, the position of the amino acid in the sequence has a significant impact on the parameter that is derived. For example, the sequence-specific intrinsic size parameter for leucine in the third position of a peptide (SSISP(Leu_3)) is ∼10% larger than SSISP(Leu_1). On average, cross sections that are derived using SSISPs provide a better representation of the experimental value than those derived from composition only intrinsic size parameters, derived as described previously (Valentine et al. *J. Phys. Chem.* **1999**, *103*, 1203). Finally, molecular modeling techniques are used to derive some insight into the origin of cross-section differences that arise from sequence variation.

Introduction

The ability to produce peptide and protein ions by mass spectrometry (MS) techniques[1,2] has made it possible to study conformations in the gas phase, where structure is defined only by intramolecular interactions.[3] Over the last decade, a range of different systems have been examined. These include small peptides, containing only a few amino acids,[4−8] synthetic peptides with defined structures,[9−12] several different proteins,[13−18] and studies of conformation as a function of charge state[19−22] as well as highly ordered noncovalent complexes[23−25] and studies to understand extremely large systems.[26,27]

There are a number of motivations for studying the structures of small peptides in the gas phase.[28,29] In solution, the structures of a number of alanine-rich helices[30] and other helical sequences have been studied in detail, and the dominant factors that influence helix formation have been discussed.[31] However, much less is known about the structures of nonhelical sequences.[32] In part, the dearth of information arises because such structures are often highly dynamic on the time scales of available experimental measurements. Even relatively defined motifs, such as helices, may fray at the ends, making it difficult to characterize structures.[33] In the gas phase, experiment and theory suggest that in many cases the removal of solvent may stabilize some types of structures, making it possible to study conformations for extended times.

From a practical point of view, experimental studies of small systems (in the absence of solvent) are readily complemented by detailed quantum chemical and molecular modeling calculations.[4,34,35] Thus, although a gas-phase measurement does not provide information that is capable of defining atomic coordinates, it is sometimes possible to obtain detailed insight from the interplay between experiment and theory.[4,34] Additionally, syntheses of short sequences (<20 residues) are usually relatively straightforward, making it possible to iteratively design specific motifs.[9,12] The combination of combinatorial chemistry[36] and MS analysis makes it possible to sample the conformations of many sequence types rapidly.

In this paper, we report ion mobility/MS measurements of cross sections for 162 singly protonated tripeptide sequences. These sequences were synthesized combinatorially as six different libraries, having the general forms ABC, ACB, BAC, BCA, CAB, and CBA, where the letters A, B, and C represent amino acids in the sequence. Each library contains 27 tripeptides of different mass. In total, there are six isomer forms of each of the 27 amino acid compositions. Of the 20 naturally occurring amino acids, only 9 (chosen as representatives of chemical type) were used in the synthesis. As described below, from the experimental results and some molecular modeling studies, we find trends in cross sections that are influenced by the position and sequence of some amino acids. From these trends, we were encouraged to develop the first set of intrinsic size parameters that take into account the position of the amino acid within the amino acid sequence. We refer to these as sequence-specific intrinsic size parameters (SSISPs). In concept, these parameters are analogous to the composition only intrinsic size parameters (COISPs) that we derived previously[37,38] and other principle component analyses used to predict mobilities;[39,40] however, in practice, the inclusion of sequence appears to provide a much better means of calculating cross sections for most small peptides.

An understanding of how the position of an amino acid in a peptide sequence influences cross section is important for a number of reasons. In a recent study, Srebalus-Barnes et al. used the intrinsic sizes of amino acids to correct for the contributions

* Author to whom correspondence should be addressed. E-mail: clemmer@indiana.edu.

10.1021/jp050761u CCC: $30.25
Published on Web 05/19/2005

of different side chains (situated at different i and $i + 4$ positions, where i denotes the location of the amino acid position in the sequence) in gas-phase helices and globules.[12] In a similar series of studies, Counterman et al. used a similar approach to determine intrinsic amino acid volumes, in an attempt to understand how solvation influences the packing of different amino acids.[41] In addition to such fundamental studies, recent advances in the field of proteomics have combined MS and tandem MS data (precursor ion and fragment ion information) with database searching algorithms to assign peptide (and protein) sequences.[42] In many cases, the applied assignment constraints yield a number of viable assignments; in most, the information is not sufficient to reliably make an unambiguous assignment. The SSISPs that we describe below should be directly useful as an additional constraint for such approaches.

Experimental Section

Synthesis of Six Tripeptide Libraries. Six tripeptide libraries were synthesized using standard solid-phase mix and split protocols[43] and Fmoc (fluorenylmethoycarbonyl) peptide chemistry.[44] The libraries were prepared so that each is randomized over three sets of three amino acid residues. By the synthesis of these libraries in this manner, each library contains the same set of masses, but each differs in sequence such that in total there are six isomer forms of each amino acid composition. Specifically, nine amino acids were utilized in the synthesis, Asp, Glu, and Gly at position A, Arg, Leu, and Lys at position B, and Phe, Ser, and Tyr at position C. The amino acid sets (A, B, and C) were arranged in all possible combinations, resulting in the following libraries: ABC, ABC, BAC, BCA, CAB, and CBA. At the start of the synthesis, equal molar quantities of resins containing the three C-terminal residues for a particular library were mixed in a nitrogen-agitated reaction vessel with dimethylformamide (DMF) prior to the removal of the N-terminal Fmoc protecting group with 20% piperidine in DMF. The resulting resin mixture was split into equal portions and placed into separate reaction vessels for the addition of the next amino acid residues. Amino acid coupling reactions were performed by the addition of preactivated benzotriazol-1-yloxy (OBt) esters of each amino acid being added to the Fmoc-deprotected resin. Activated amino acid OBt esters were generated by the reaction of the Fmoc amino acids (4.0 equiv) with 2-(1-H-benzotriazol-1-yl)-1,1,3,3-tetramethyluronium hexafluorophosphate (HBTU) (3.9 equiv) and 0.4 M *N*-methylmorpholine in DMF (4 equiv) in the reaction vessel containing the resin. After the final N-terminal amino acid residues were coupled, the Fmoc protecting group was removed from the N-terminus of the peptide. Each of the six libraries was cleaved from the solid-phase support, and the side chain protecting groups were removed using a trifluoroacetic acid/phenol/water/thioanisole/ethanedithiol solution (82.5:5:5:5:2.5 by volume). The resin was filtered from each of the solutions, and the peptides were precipitated in ether. The precipitates were washed several times with ether, dried, dissolved in an aqueous solution (30% acetic acid), and lyophilized. The following preloaded Wang resins (Novabiochem) were used: Fmoc-Asp(OtBu), Fmoc-Glu(OtBu), Fmoc-Gly, Fmoc-Arg(Pmc), Fmoc-Leu, Fmoc-Lys(Boc), Fmoc-Phe, Fmoc-Ser(tBu), and Fmoc-Tyr(tBu). The following N-α-Fmoc-protected amino acids (Novabiochem) were used: Fmoc-Asp(OtBu)-OH, Fmoc-Glu(OtBu)-OH, Fmoc-Gly-OH, Fmoc-Arg(Pmc)-OH, Fmoc-Leu-OH, Fmoc-Lys(Boc)-OH, Fmoc-Phe-OH, Fmoc-Ser(tBu)-OH, and Fmoc-Tyr(tBu)-OH.

Determination of Peptide Ion Cross Sections. Ion mobility techniques have been reviewed elsewhere.[45] A high-resolution ion mobility/time-of-flight instrument (described previously)[46,47] was used to measure peptide ion mobilities. Cross sections are determined from ion mobility distributions, which are recorded as follows. Peptide ions were produced by electrospraying solutions of library peptides (0.25 mg mL^{-1} in 49:49:2, water/acetonitrile/acetic acid) into the source region of the instrument. Ions are guided into the front of the 58.29 cm drift tube that is operated at a pressure of ~140 Torr and with an uniform electric field of 154.40 V cm^{-1}. The time required for short pulses (150 μs) of ions to drift across the drift region was recorded using a home-built acquisition system.[48] Generally, highly charged ions experience greater drift force than lower charge state ions; for ions of the same m/z ratio, those that are larger experience more collisions with the buffer gas and have lower mobilities than those that are more compact.[45] After drifting through the helium buffer gas, the ions are pulsed into a reflectron time-of-flight mass spectrometer where flight time distributions are recorded. Mass-to-charge ratios are determined using a standard multipoint calibration. As described previously, we refer to this as a nested approach. The typical ion mobility resolving power ($t_D/\Delta t_D$ where Δt_D is the full width half-maximum (fwhm) of the peak) is ~80 to 200. The resolving power of the time-of-flight mass spectrometer ($m/\Delta m$ where Δm is the fwhm of the peak) in this instrument is on the order of 1000 to 1500.

Once the drift time and m/z for an ion has been determined, a collision cross section can be calculated using eq 1[49]

$$\Omega = \frac{(18\pi)^{1/2}}{16}\frac{ze}{(k_B T)^{1/2}}\left[\frac{1}{m_I}+\frac{1}{m_B}\right]^{1/2}\frac{t_D E}{L}\frac{760}{P}\frac{T}{273.2}\frac{1}{N} \quad (1)$$

where t_D corresponds to the drift time and E, L, T, and P correspond to the electric field strength, drift tube length, buffer gas temperature, and pressure, respectively; m_I and m_B correspond to the masses of the ion and buffer gas, respectively. The other terms k_B, ze, and N correspond to Boltzmann's constant, the charge on the ion, and the neutral number density, respectively.

Determination of Amino Acid Size Parameters. We have previously shown that if we ignore the amino acid sequence and consider only composition, then intrinsic size parameters for individual amino acids (referred to in this paper as COISPs) can be calculated using a system of equations described by eq 2.[37,38]

$$\frac{\sum_i n_{ij} p_i}{\sum_i n_{ij}} = \frac{\Omega_j(\text{exp})}{\Omega_j(\text{PA})} \quad (2)$$

In this equation, Ω_j(exp) is the experimental cross section of peptide j, and Ω_j(PA) is a normalization factor derived from the cross sections of polyalanine peptides, as described previously. The term $\Omega_j(\text{exp})/\Omega_j(\text{PA})$ is referred to as a reduced cross section,[37,38] and it is used because it removes differences in size that arise because of differences in mass; n_i corresponds to the number of times that the size parameter p_i occurs for peptide j. Values of Ω_j(PA) were determined by extrapolating from a polynomial fit for the cross sections of polyalanine containing three to seven residues ($-1.274 \times 10^{-5}(\text{MW})^2 + 1.885 \times 10^{-1}(\text{MW}) + 46.462$).[50] Here, only three-residue peptides are considered such that $\sum_i n_{ij} \equiv 3$.

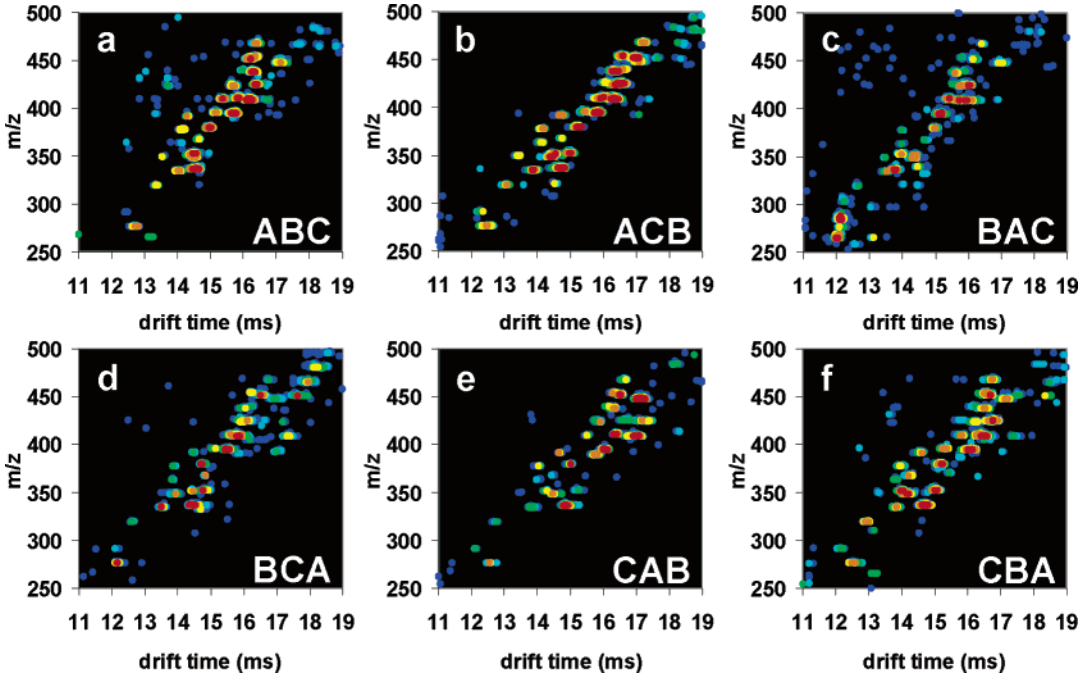

Figure 1. Two-dimensional plots of *m*/*z* versus drift time for the six, 27-component libraries. Each library contains the same *m*/*z* ratios but a different sequence isomer. The amino acids Asp, Glu, or Gly correspond to A, B corresponds to Arg, Leu, or Lys, and C corresponds to Phe, Ser, or Tyr. These three sets of amino acids were randomized over the three positions to create six libraries (a) ABC, (b) ACB, (c) BAC, (d) BCA, (e) CAB, and (f) CBA. From these plots, a drift time can be extracted for each of the 162 peptides (27 sets of six sequence isomers), and the cross sections can be calculated. The colors indicate increasing intensity from blue to red.

In this paper, all cross sections that were predicted or retrodicted were done so using size parameters that were calculated from the 162 tripeptide cross sections that were measured in this experiment. Values for the COISPs were obtained by solving a system of 162 equations for 9 variables, the nine amino acids used. Size parameters that include the location of the peptide in the sequence, the SSISPs, were obtained by solving the same system of 162 equations for 27 variables; in this case, the nine amino acids at each of the three positions in the peptide. A least-squares solution was obtained for the variables in each set of equations, and the uncertainties correspond to one standard deviation about the mean.

Molecular Modeling and Cross-Section Calculations. Molecular modeling techniques and cross-section calculations were employed to provide some insight about possible structures (and trends associated with variations in sequence) for ~10 sequences. Extended trial conformations of peptides were generated using the suite of programs available with the Insight II[51] molecular modeling software incorporating the AMBER force field. Each of these structures was taken through two different annealing processes. The first heated the molecules to 1000 K over 2 ps, equilibrated at 1000 K for 2 ps, and then cooled to 300 K over 1 ps. This process was repeated 100 times and the five lowest-energy structures were taken to a second round of annealing where they were heated to 500 K over 2 ps, equilibrated at 500 K for 2 ps, and then cooled to 300 K over 1 ps. This was repeated 100 times for a total of 500 structures. In the second, the temperature was increased to 500 K over 2 ps, equilibrated at 500 K for 2 ps, and then cooled to 300 K over 1 ps. This process was repeated 100 times and the five lowest-energy structures were taken to a second round of annealing where they were heated to 500 K over 2 ps, equilibrated at 500 K for 2 ps, and then cooled to 300 K over 1 ps. This was repeated 100 times for a total of 500 structures. This process yielded a total of 1200 possible structures for each of the trial conformations. The five lowest energies from each of the annealing steps were compared, and the five lowest-energy structures from these 60 structures were chosen as the representative lowest-energy structures. Trajectory collision cross sections were then calculated for each of the five lowest-energy structures using the MOBCAL program developed by Jarrold and co-workers.[52,53] These cross sections were compared to the experimental cross section, and the one with the smallest difference and within 2% of the experimental value was chosen as the representative structure.

Results and Discussion

Analysis of 162 Peptides from Six Libraries. Figure 1 shows two-dimensional representations of narrow regions of ion mobility/mass spectra data that were obtained upon analyzing each of the six library forms: ABC, ACB, BAC, BCA, CAB, and CBA. The drift time and *m*/*z* data that are shown correspond to the region where we would expect to find the singly protonated, $[M + H]^+$, form of each tripeptide. From visual inspection of these data, we are able to identify peaks with *m*/*z* values that are identical to those expected for each of the 27 synthesized peptides, as $[M + H]^+$ ions (in each of the library). In other regions of the data set (not shown), there is evidence for higher charge states (primarily $[M + 2H]^{2+}$) of some peptide sequences; additionally, there are peaks for some multimers (primarily $[2M + H]^+$ and $[2M + 2H]^+$) that can be assigned in some cases to specific sequences. All peaks that are not ascribable to the $[M + H]^+$ peptides of interest are easily distinguished from singly protonated peptides due to the fact that they are found in a different region of the two-dimensional dataset. Some additional features, observed in the region shown, have *m*/*z* values that cannot be attributed to the expected $[M + H]^+$ library peptides. Such peaks may be due to products that are not expected from the synthesis. We have previously commented about such species.[54] For the remainder of this work, we focus only on the $[M + H]^+$ forms of the 162 peptide sequences that we aimed to synthesize for this study.

TABLE 1: Measured Cross Sections for 162 Tripeptides

	residue[a]				cross section of specific library sequence[c]					
no.	A	B	C	*m*/*z*[b]	ABC	ACB	BAC	BCA	CAB	CBA
1	Gly	Leu	Ser	276.2	103.7	101.5	100.0	99.4	102.4	102.4
2	Gly	Lys	Ser	291.2	101.8	100.8	100.3	98.6	98.6	99.7
3	Gly	Arg	Ser	319.2	108.9	106.7	103.6	103.1	104.2	106.0
4	Asp	Leu	Ser	334.2	114.8	113.4	111.0	110.2	112.8	113.1
5	Gly	Leu	Phe	336.2	118.9	120.5	113.2	118.0	121.4	120.6
6	Glu	Leu	Ser	348.2	118.5	117.9	118.0	113.9	118.4	115.7
7	Asp	Lys	Ser	349.2	110.3	109.6	110.6	110.5	110.1	108.6
8	Gly	Lys	Phe	351.2	117.4	119.0	118.4	118.4	116.9	114.6
9	Gly	Leu	Tyr	352.2	118.1	122.7	115.0	121.4	124.4	123.2
10	Glu	Lys	Ser	363.2	115.1	115.6	115.0	113.1	115.0	113.4
11	Gly	Lys	Tyr	367.2	119.6	120.4	121.3	121.3	118.0	116.4
12	Asp	Arg	Ser	377.2	115.5	115.9	115.0	113.1	115.0	114.2
13	Gly	Arg	Phe	379.2	122.2	124.5	123.2	120.6	122.8	124.3
14	Glu	Arg	Ser	391.2	116.6	120.8	118.7	118.3	119.8	119.4
15	Asp	Leu	Phe	394.2	128.5	129.4	124.6	126.5	131.4	131.4
16	Gly	Arg	Tyr	395.2	124.0	126.4	124.6	123.5	122.0	125.4
17	Glu	Leu	Phe	408.2	132.2	133.5	129.8	129.5	139.2	134.7
18	Asp	Lys	Phe	409.2	125.5	130.1	127.2	127.6	126.9	129.5
19	Asp	Leu	Tyr	410.2	129.6	130.8	126.8	128.4	134.4	134.3
20	Glu	Lys	Phe	423.2	128.1	134.6	129.8	131.0	130.2	132.8
21	Glu	Leu	Tyr	424.2	134.1	135.3	131.6	131.7	141.5	137.3
22	Asp	Lys	Tyr	425.2	127.7	131.9	127.5	129.9	136.6	137.7
23	Asp	Arg	Phe	437.2	133.3	134.2	128.6	130.6	134.0	135.4
24	Glu	Lys	Tyr	439.2	130.3	136.4	131.2	132.8	132.5	132.8
25	Glu	Arg	Phe	451.2	132.6	139.0	133.4	135.1	135.1	136.5
26	Asp	Arg	Tyr	453.2	133.7	135.6	130.1	132.8	133.2	135.4
27	Glu	Arg	Tyr	467.2	134.0	140.9	135.6	137.3	135.8	136.9

[a] The three sets of amino acids (A, B, and C) were varied over the three positions of the peptides. This gave 27 sets of six sequence isomers. [b] Mass-to-charge values for the $[M + H]^+$ ion of the six peptide isomers containing residues A, B, and C. [c] Measured cross sections for the 27 peptides measured in each library are listed below their library residue sequence. All cross sections are in Å^2.

To determine the experimental cross sections, we identified drift times associated with the peak maxima for each of the $[M + H]^+$ peptides. In essentially all cases, the drift time distribution of the *m*/*z* region for a single peptide is dominated by a single sharp peak. The shape of this peak suggests that a single conformation dominates the distribution or that structural changes within a peptide are rapid with respect to experimental time scales such that the peak corresponds to an average of structures that are sampled. Cross sections for the dominate feature of each of the 162 peptide sequences are provided in Table 1. For this set of sequences, values range from 99.4 Å^2 for $[LSG + H]^+$ ($m/z = 276.1$) to 141.5 Å^2 for $[YEL + H]^+$ ($m/z = 424.2$). From many other experiments (and calibrations with known systems), we anticipate relative uncertainties are on the order of ±1%.

Trends in Cross Sections for Different Compositions and Sequences. For this relatively small (and highly defined) group of tripeptide sequences, it is possible to identify many different types of trends that are found as one amino acid is substituted for another or positioned at a different location in the sequence. In this section, we provide only a brief discussion of a few trends that are observed for only a few sequences. One of us (A.E.H.) has considerable interest in many additional trends that are observed; these are discussed elsewhere.[55]

Figure 2 shows cross sections for 24 selected sequences that contain a Leu residue at the first, second, or third position. In the set of sequences used for this comparison, the Leu residue has been combined with an aromatic (Phe or Tyr) and an acidic (Glu or Asp) residue. Thus, there are four sets of six sequence isomers, having the amino acid compositions: Asp, Leu, Phe; Glu, Leu, Phe; Asp, Leu, Tyr; and Glu, Leu, Tyr. The plot that is shown illustrates several trends that are found with changing composition and sequence of the above sets of amino acids.

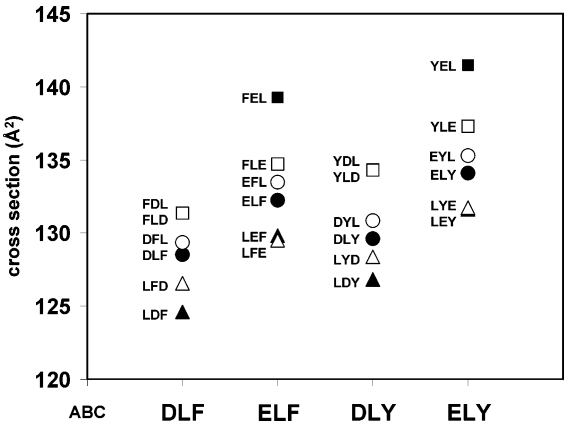

Figure 2. Cross sections of four sets of sequence isomers are plotted so that trends in these data can be identified. Each sequence isomer is identified by a symbol pertaining to its sequence, ABC, ACB, BAC, BCA, CAB, and CBA, where for the plotted set of peptides A corresponds to Asp or Glu, B corresponds to Leu, and C corresponds to Phe or Tyr. Several trends are seen for these sets of isomers.

For example, consider the extreme cross sections within each isomer set. Eight of the sequences shown in the figure have a Leu residue in the first position: Leu-Asp-Phe, Leu-Phe-Asp, Leu-Glu-Phe, Leu-Phe-Glu, Leu-Asp-Tyr, Leu-Tyr-Asp, Leu-Glu-Tyr, and Leu-Tyr-Leu. Within each of the four sets of isomers, these peptides have the smallest cross sections. For example, for the isomer set Leu-Asp-Phe, Leu-Phe-Asp, Asp-Leu-Phe, Asp-Phe-Leu, Phe-Asp-Leu, and Phe-Leu-Asp, the cross sections for the peptides containing Leu in their first position are $\Omega([LDF + H]^+) = 124.6$ Å^2 and $\Omega([LFD + H]^+) = 126.5$ Å^2 while cross sections for all other sequences are larger: $\Omega([DLF + H]^+) = 128.5$ Å^2, $\Omega([DFL + H]^+) = 129.4$ Å^2, $\Omega([DLF + H]^+) = 131.4$ Å^2, $\Omega([DLF + H]^+) = 131.4$ Å^2. In contrast, the eight peptides in the figure that have a Phe or Tyr (an aromatic residue) located in the first position have the maximum cross sections within the four isomers sets. For example, in this same set of isomers a cross section of 131.4 Å^2 was recorded for the Phe-Asp-Leu and Phe-Leu-Asp sequences, which is the largest value within this set. This trend, in which peptide isomers with aromatic residues in the first position have the largest cross section within the isomer set, extends to all four of the isomer sets that are shown in Figure 2.

As we considered these systems in more detail, it became clear that many peptides can fall into more than one specific trend. A good example of this involves the peptides Phe-Asp-Leu, Phe-Glu-Leu, Tyr-Asp-Leu, and Tyr-Glu-Leu. Although these peptides are considered within the trend involving aromatic residues in the first position (discussed above); additionally, each of these sequences has Leu as the third (C-terminal) residue and is the largest peptides within its set of isomers. Thus, one could also explain the trend in the sequence sets that we have chosen for Figure 2 by considering only the position of the Leu residue. In the end, as one considers the long history related to predicting protein structure from the propensities of amino acids to be found in specific types of structures,[56] it is clear that many additional factors influence structure, even in these very small systems. To some extent, we were surprised that many types of trends could be drawn from this relatively small set of sequences.

Trends in Structure That Are Observed from Molecular Modeling Calculations. To obtain more insight about the types of conformations that are present, we examined a number of sequences using molecular modeling techniques. Figure 3 shows

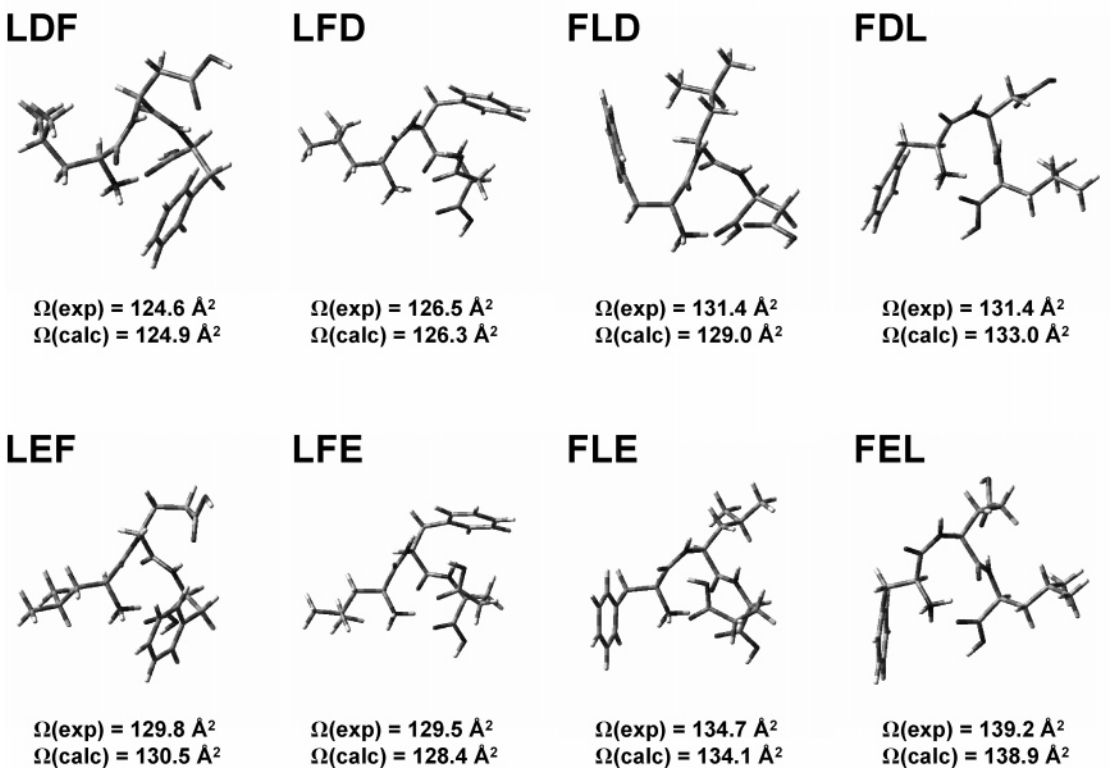

Figure 3. Structures from molecular modeling of eight three-residue peptides. These peptides all contain Leu and Phe residues with either Asp or Glu acidic residues. The four peptides at the top of the figure contain Asp, and the four peptides at the bottom of the figure contain Glu. These peptides were modeled to determine structural differences due to amino acid interactions within peptide isomers.

typical results for eight different sequences: Leu-Asp-Phe, Leu-Phe-Asp, Phe-Leu-Asp, Phe-Asp-Leu, Leu-Glu-Phe, Leu-Phe-Glu, Phe-Leu-Glu, and Phe-Glu-Leu. In this case, we show a relatively low-energy conformation that has a calculated cross section that is in close agreement with experiment. Upon examining these structures, we were struck by the fact that although the calculated and experimental cross sections are in good agreement, there is no obvious trend in the structure that explains the trends that we discussed above for positioning Leu at the first or last position or an aromatic residue in the first position. For example, although the structures for the Leu-Asp-Phe and Leu-Glu-Phe sequences (having Leu in the first position) are more compact than the Phe-Asp-Leu and Phe-Glu-Leu structures (with Leu in the last position), this does not appear to arise because of obvious differences in Leu packing. In both cases, the Leu residue appears to extend away from the peptide core and appears to have few conformational restrictions. Similarly, there is no obvious trend in structure that explains the experimental result in which an aromatic residue in the first position leads to the largest cross section.

Instead, the significant structural element that is apparent is associated with the position and differences in size associated with the Asp and Glu acidic residues. In both cases, if the acidic group is positioned as the third (C-terminal residue), then the peptide favors structures in which both the C-terminal carboxylic acid group and the carboxylic acid side chain interact with the protonated N-terminal amino group. Sequences in which the acidic residue is located in the second position appear to restrict the interaction of the side chain with the N-terminus; instead, these favor interactions between the ends of the peptides. In all of the sequences that we have studied, it appears that relatively compact conformations are favored due to interactions along the peptide with the protonated amino terminus (similar to charge-solvation structures reported previously).[10,50]

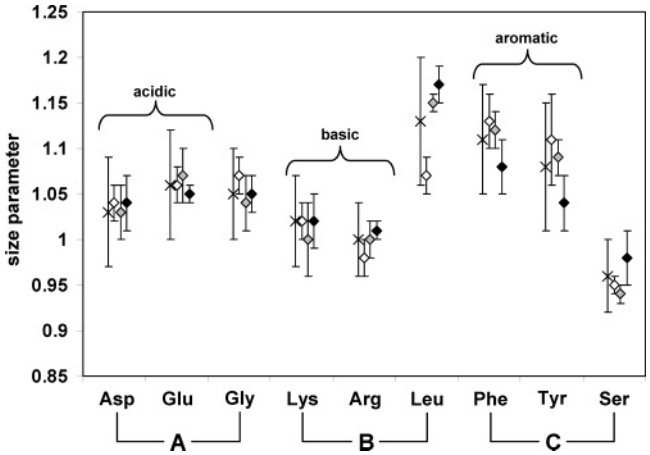

Figure 4. Size parameters were solved from sets of linear equations obtained from cross-section measurements of 162 peptides (27 sets of six sequence isomers). The symbol × corresponds to the composition only size parameters (COISPs) calculated from these data for each of the nine amino acids. The diamonds correspond to sequence-specific size parameters (SSISPs) calculated for each of the nine amino acids in each of the three positions. The white diamonds correspond to SSISPs in the N-terminal position of the peptide, the gray diamonds correspond to the SSISPs in the center position of the three-residue peptide, and the black diamonds correspond to the SSISPs in the C-terminal position of the peptide. Error bars indicate one standard deviation about the mean.

Determination of Sequence-Specific Intrinsic Size Parameters. Details involving the calculation of composition only and sequence-specific intrinsic size parameters, COISPs and SSISPs, were discussed above. Figure 4 shows the values obtained upon applying eq 2 (with the experimental cross sections in Table 1) for COISPs and SSISPs for the nine amino acids contained in these peptides. In this plot, a size parameter of 1.00 indicates that the amino acid of interest has the same overall influence

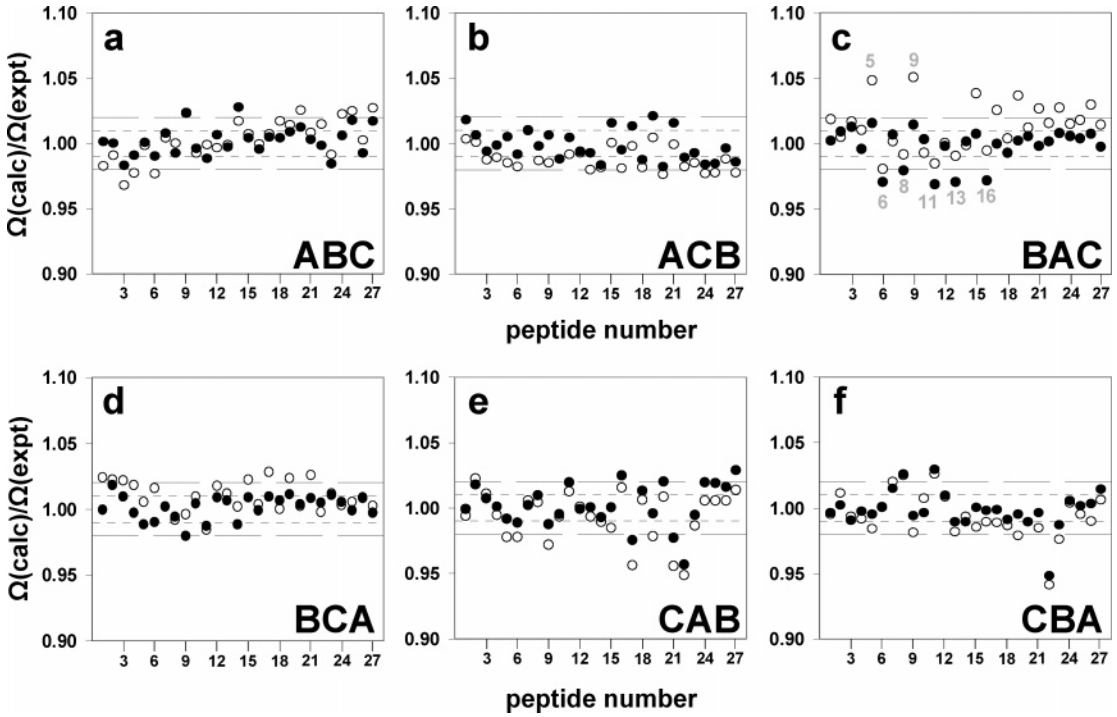

Figure 5. Plots of the cross section calculated using size parameters divided by the experimental cross section for each of the 27 peptides in each of the six libraries. The closed circles correspond to the cross sections calculated with SSISPs, and the open circles correspond to cross sections calculated using COISPs. The calculated cross sections for each of the peptides were determined by summing the size parameters (either SSISPs or COISPs) for each of the amino acids, dividing by 3 (the number of amino acids in the peptide), and multiplying by the cross section of polyalanine at the same molecular weight as the peptide. The short dashed lines indicate ±1% from the experimental value, and the long dashed lines indicate ±2% from the experimental value.

on cross section as an alanine residue within small singly protonated polyalanine peptides, which favor roughly spherical globular conformations.[37,38]

A number of points regarding Figure 4 are noteworthy. First, it is interesting to compare the COISP values obtained from these tripeptides with values that were derived previously from 5−10 residue tryptic peptides.[37,38] For the most part, the values derived from tripeptides are significantly different than those derived previously. The only exception is for the Leu residue, which has a relatively large COISP (1.13 ± 0.07 from tripeptides and 1.19 ± 0.02 from the tryptic peptides). The values of COISP(Arg) = 1.00 ± 0.04 and COISP(Lys) = 1.02 ± 0.05 from the tripeptides are significantly smaller than the values of 1.23 ± 0.04 and 1.27 ± 0.07 from tryptic peptides. We note that molecular modeling studies of the Lys- and Arg-containing tripeptide sequences require tightly packed conformations to reproduce the experimental data. A number of cases (e.g., EFR and RFE) required salt bridge charge assignments to create very compact conformations that have calculated cross sections that are in agreement with experiment.

Second, the uncertainties of the COISPs for tripeptides are much larger than those obtained for tryptic peptides. For example, the uncertainty of COISP(Leu) = 0.07 from tripeptides and 0.02 from tryptic peptides. Overall, the large uncertainties associated with these sequence isomers suggests that the influence on the peptide cross section associated with the position of the amino acid in the sequence cannot be captured from the average of all sequences. Indeed, when fewer values are averaged over specific positions in the sequence, the uncertainty is reduced substantially. For example, SSISP(Leu_1) = 1.07 ± 0.02, SSISP(Leu_2) = 1.15 ± 0.01, and SSISP(Leu_3) = 1.17 ± 0.02. Moreover, the decrease in the uncertainties obtained for SSISPs compared with COISPs is common to all of the nine amino acids determined here.

Finally, in some cases it appears that the position of the amino acid in the sequence has a large impact on the peptide cross section. Again, we consider the Leu residue where the contribution to the cross section when located at the third position is ∼10% larger than when located at the first position. In other cases, such as the acidic and basic residues (and also Gly), the impact of the position within the sequence appears to be much smaller.

As before,[37,38] it is possible to combine size parameters to calculate cross sections. Strictly speaking, these values should be called retrodictions (rather than predictions) since the experimental values, which are being calculated from size parameters, were used to generate the size parameters. Below, we present bona fide predictions for sequences that have not been measured yet. It is interesting to compare the quality of retrodictions that are obtained when experimental cross sections are calculated using the new SSISPs with the values obtained using the new COISPs obtained here.

Figure 5 shows a comparison of calculated and experimental cross sections using both methods to calculate cross sections for all 162 peptides. The values that are shown correspond to the ratio of the calculated cross sections (using either the COISP or the SSISP values). In this case, a ratio of 1.00 indicates that the parameters have exactly captured the experimental value. As the values deviate from 1.00, the values have been either overestimated or underestimated with respect to the experimental cross section. Overall, this figure shows that (1) most cross sections can be predicted using either COISPs or SSISPs to within ∼2% of the experiment and (2) the SSISP values do a better job than the COISP values of retrodicting the experimental cross sections (89% within 2% of experiment for SSISPs, compared with 75% for COISPs).

By examination of these data in more detail, it is possible to gain some insight into which sequences are accurately estimated

and which are not. To do this, it is necessary to understand the layout of Figure 5 in more detail. The figure is designed to be used along with the information in Table 1. The designation of parts a−f corresponds to the different sequences that are tabulated: ABC, ACB, BAC, BCA, CAB, and CBA, respectively. The individual points in each plot correspond to peptides that are plotted in the same order as they are tabulated. For example, from the combined information, we know that the open and solid circles, correspond to the fifth peptide in the BAC library (Figure 5c), which is comprised of the amino acids Gly, Leu, and Phe, in the sequence Leu-Gly-Phe. In this case, the COISP calculation (shown as the open circle) significantly overestimates the cross section for $[LGF + H]^+$ by 4.9%; however, the SSISPs yield a value (the solid circle) that is within 1.6% of the experiment. Similarly, the COISP cross section calculated for the ninth peptide in this library, corresponding to the Leu-Gly-Tyr sequence, is 5.1% greater than experiment, where as the SSISP cross section is within 1.5% of experiment. From these values as well as others that we have examined in detail, it appears that peptides having a Leu residue in the first position are typically overestimated using the COISPs. This is consistent with the discussion that we provided above (and the data in Figure 4) in which we outlined the finding that the position of Leu in the sequence has a large influence on its contribution to cross section (in these tripeptides).

In a number of cases, the COISP parameters appear to do a better job than the SSISP parameters. For example, Figure 5c also shows that peptides 6 (Leu-Glu-Ser), 8 (Lys-Gly-Phe), 11 (Lys-Gly-Tyr), 13 (Arg-Gly-Phe), and 16 (Arg-Gly-Tyr), which were calculated with SSISPs, have predicted cross sections that are significantly less than the experimental values. It is interesting that with the exception of the Leu-Glu-Ser sequence the rest of these peptides have a central Gly residue with a basic residue in the first position and an aromatic residue in the third position. Although this is a relatively small subset of sequences, the fact that the retrodictions in which sequence is included are not as good as composition only values suggests that other higher-order considerations may be important. As the size of the database increases, it will be interesting to keep track of those sequences in which retrodictions involving sequence arise as these may provide clues about what factors are required to understand these types of sequences in more detail.

Estimating Cross Sections for All 729 Tripeptides Comprised of the Nine Amino Acids Studied Here. The SSISPs that were determined above can be used to calculate 729 tripeptide cross sections (all possible with the nine amino acids considered here). This group includes calculated values for the 162 cross sections that were measured as well as prediction of 567 new tripeptides that contained the nine amino acids residues. Figure 6 shows a plot of the calculated cross sections for singly protonated monomer peptides, $[M + H]^+$. Over this range of amino acid types, for a given *m*/*z* value cross sections vary by as much as ~20%. Within the uncertainty of the calculation, we do not expect all of the peptide isomers to be resolved within the present experimental resolving powers or uniquely identified. However, in many cases, these peptides may be uniquely resolved from one another and possibly identified based on a combination of *m*/*z* information and SSISP prediction. As an example, at an *m*/*z* value of 392.3 three sequence isomers were predicted: Leu-Leu-Phe, Leu-Phe-Leu, and Phe-Leu-Leu. From the calculation of their cross sections using SSISPs, their cross sections were found to be 130.1, 132.4, and 136.1 Å^2. In this case, within the uncertainty of the calculation, we anticipate that Phe-Leu-Leu should be distinguished from Leu-Leu-Phe and Leu-Phe-Leu.

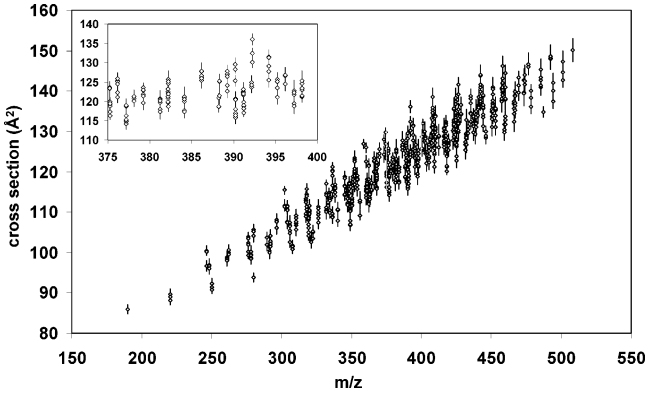

Figure 6. Prediction and retrodiction of a total of 729 tripeptides using SSISPs determined from cross sections of 27 sets of six sequence isomers (162 peptides). Error bars indicate the propagation of error determined using the error of each of the parameters used to calculate a predicted cross section. The inset shows an enlarged region from molecular weight 375 to 400 and from cross section 110 to 140 Å^2.

Summary and Conclusions

Ion mobility/MS techniques have been used to measure cross sections for six combinatorial libraries that were synthesized to produce 27 unique tripeptide compositions. The cross sections for 162 $[M + H]^+$ ions are reported. None of these cross sections had been reported previously. A unique feature of this study is that we have designed the synthesis so that there are six different isomer forms of each of the 27 compositions. This makes it possible to examine how amino acid sequence influences cross section. A number of interesting trends in how the position of an amino acid influences cross section can be observed. We discussed two trends: the observation that sequences containing a Leu residue in the first position have cross sections that are smaller than sequences in which this residue is located as the C-terminal residue, and the observation that aromatic residues located in the first position yield peptides with larger cross sections than their isomeric counterparts having aromatic residues located in the second or third position. We noted that while it is interesting to consider these trends these sequence trends alone are not unique explanations that are required to explain our data. That is, the cross section comes about from all possible interactions.

Molecular modeling studies of the positional dependence of the cross sections of Leu-containing peptides show no clear trend associated with the Leu residue. Instead, these calculations suggest that structural differences arise when an acidic residue is located in the second or third position of these sequences. That is, it appears that the ability of the acidic side chain to interact efficiently with the protonated amino terminus influences the type of conformation (and cross section). This type of interaction is analogous to charge-solvated structures that have been discussed previously.

An important product of this work is the first set of sequence-specific intrinsic size parameters that can be used to predict cross sections. These parameters appear to provide a significantly better approach for calculating cross sections from amino acid sequence, compared with parameters that are based only on the composition of the peptide. For this set of 162 tripeptides, 89% of cross sections calculated using SSISPs were within 2% of the experimental cross section measured. For this same set of tripeptides using composition only size parameters, only 75%

were within 2% of the experimental cross section. The approach for calculating cross sections is general, and a prediction of cross sections for all possible tripeptides, comprised of the nine amino acids studied here (729 total), was made. This prediction suggests that the maximum variability in cross sections (for a single mass) is ~20%. Thus, the ability to predict values within 2% of experiment should reduce some of the ambiguities that arise with database assignments of MS and tandem MS data.

Finally, we note that it should be relatively straightforward to extend this type of information for other types of sequences, different charge states, and larger peptide sizes. To some extent, it is probably possible to estimate sequences in which an Ile is substituted for Leu simply because these residues appear to affect cross section in a very similar way.[37,38,50] Other libraries can be used to develop parameters for remaining amino acids. We are currently examining several additional libraries that are aimed to extend this approach to larger sizes.

Acknowledgment. This work was funded by grants from the National Institutes of Health (1R01-GM59145-03) and the National Science Foundation (CHE0078737). The authors thank Dr. Stephen J. Valentine, Dr. Anne E. Counterman, and Dr. Catherine A. Srebalus-Barnes for many helpful discussions.

References and Notes

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

(2) Karas, M.; Hillenkamp, F. *Anal. Chem.* **1988**, *60*, 2299−2301.

(3) For example, see: Clemmer, D. E.; Jarrold, M. F. *J. Mass Spectrom.* **1997**, *32*, 577−592. Hoaglund-Hyzer, C. S.; Counterman, A. E.; Clemmer, D. E. *Chem. Rev.* **1999**, *99*, 3037−3079. Jarrold, M. F. *Annu. Rev. Phys. Chem.* **2000**, *51*, 179−207. Wyttenbach, T.; Bowers, M. T. *Top. Curr. Chem.* **2003**, *225*, 207−232.

(4) Wyttenbach, T.; von Helden, G.; Bowers, M. T. *J. Am. Chem. Soc.* **1996**, *118*, 8355−8364.

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

(6) Campbell, S.; Rodgers, M. T.; Marzluff, E. M.; Beauchamp, J. L. *J. Am. Chem. Soc.* **1995**, *117*, 12840−12854.

(7) Wyttenbach, T.; Liu, D.; Bowers, M. T. *Int. J. Mass Spectrom.* **2005**, *240*, 221−232.

(8) Wu, C.; Siems, W. F.; Klasmeier, J.; Hill, H. H., Jr. *Anal. Chem.* **2000**, *72*, 391−395.

(9) Hudgins, R. R.; Mao, Y.; Ratner, M. A.; Jarrold, M. F. *Biophys. J.* **1999**, *76*, 1591−1597.

(10) Counterman, A. E.; Clemmer, D. E. *J. Am. Chem. Soc.* **2001**, *123*, 1490−1498.

(11) Kinnear, B. S.; Jarrold, M. F. *J. Am. Chem. Soc.* **2001**, *123*, 7907−7908.

(12) Srebalus-Barnes, C. A.; Clemmer, D. E. *J. Phys. Chem. A* **2003**, *107*, 10566−10579.

(13) Suckau, D.; Shi, Y.; Beu, S. C.; Senko, M. W.; Quinn, J. P.; Wampler, F. M.; McLafferty F. W. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 790−793.

(14) 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.

(15) Valentine, S. J.; Anderson, J. G.; Ellington, A. D.; Clemmer, D. E. *J. Phys. Chem. B* **1997**, *101*, 3891−3839.

(16) Jarrold, M. F. *Acc. Chem. Res.* **1999**, *32*, 360−367.

(17) 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.

(18) Covey, T.; Douglas, D. J. *J. Am. Soc. Mass Spectrom.* **1993**, *4*, 616−623.

(19) Williams, E. R. *J. Mass Spectrom.* **1996**, *31*, 831−842.

(20) Valentine, S. J.; Clemmer, D. E. *J. Am. Chem. Soc.* **1997**, *119*, 3558−3566.

(21) Freitas, M. A.; Hendrickson, C. L.; Emmett, M. R.; Marshall, A. G. *Int. J. Mass Spectrom.* **1999**, *185/186/187*, 565−575.

(22) Badman, E. R., Hoaglund-Hyzer, C. S.; Clemmer, D. E. *Anal. Chem.* **2001**, *73*, 6000−6007.

(23) Cooks, R. G.; Zhang, D. X.; Koch, K. J.; Gozzo, F. C.; Eberlin, M. N. *Anal. Chem.* **2001**, *73*, 3646−3655.

(24) Counterman, A. E.; Clemmer, D. E. *J. Phys. Chem. B* **2001**, *105*, 8092−8096.

(25) Hodyss, R.; Julian, R. R.; Beauchamp, J. L. *Chirality* **2001**, *13*, 703−706.

(26) Sobott, F.; Hernández, H.; McCammon, M. G.; Tito, M. A.; Robinson, C. V. *Anal. Chem.* **2002**, *74*, 1402−1407.

(27) Loo, J. A. *Int. J. Mass Spectrom.* **2000**, *200*, 175−186.

(28) Clemmer, D. E.; Jarrold, M. F. *J. Mass Spectrom.* **1997**, *32*, 577−592.

(29) Barran, P. E.; Polfer, N. C.; Campopiano, D. J.; Clarke, D. J.; Langridge-Smith, P. R. R.; Langley, R. J.; Govan, J. R. W.; Maxwell, A.; Dorin, J. R.; Millar, R. P.; Bowers, M. T. *Int. J. Mass Spectrom.* **2005**, *240*, 273−284.

(30) Fiori, W. R.; Millhauser, G. L. *Biopolymers* **1995**, *37*, 243−250. Millhauser, G. L.; Stenland, C. J.; Bolin, K. A.; van de Ven, J. M. *J. Biomol. NMR* **1996**, *7*, 331−334.

(31) For example, see: Scholtz, J. M.; Baldwin, R. L. *Annu. Rev. Biomol. Struct.* **1992**, *21*, 95. Baldwin, R. L. *Biophys. Chem.* **1995**, *55*, 127. Chakrabartty, A.; Baldwin, R. L. *Adv. Protein Chem.* **1995**, *46*, 141. Rohl, C. A.; Baldwin, R. L. *Methods Enzymol.* **1995**, *46*, 1. Baldwin, R. L.; Rose, G. D. *Trends Biochem. Sci.* **1999**, *24*, 26.

(32) Struthers, M. D.; Cheng, R. P.; Imperiali, B. *Science* **1996**, *271*, 342−345. Struthers M.; Ottesen, J. J.; Imperiali, B. *Folding Des.* **1998**, *3*, 95−103.

(33) Miick, S. M.; Martinez, G. V. Fiori W. R.; Todd, A. P.; Millhauser, G. L. *Nature* **1992**, *359* 653−655.

(34) Shvartsburg, A. A.; Liu, B.; Jarrold, M. F.; Ho, K. M. *J. Chem. Phys.* **2000**, *112*, 4517−4526.

(35) Tuckerman, M. E.; Martyna, G. J. *J. Phys. Chem. B* **2000**, *104*, 159−178.

(36) Merrifield, B. *Science* **1986**, *232*, 341−347.

(37) Valentine, S. J.; Counterman, A. E.; Hoaglund-Hyzer, C. S.; Clemmer, D. E. *J. Phys. Chem. B* **1999**, *103*, 1203−1207.

(38) Valentine, S. J.; Counterman, A. E.; Clemmer, D. E. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 1188−1211.

(39) Shvartsburg, A. A.; Siu, K. W. M.; Clemmer, D. E. *J. Am. Soc. Mass Spectrom.* **2001**, *12*, 885−888.

(40) Mosier, P. D.; Counterman, A. E.; Jurs, P. C.; Clemmer, D. E. *Anal. Chem.* **2002**, *74*, 1360−1370.

(41) Counterman, A. E.; Clemmer, D. E. *J. Am. Chem. Soc.* **1999**, *121*, 4031−4039.

(42) For examples of database searching algorithms, see: *SEQUEST* (ThermoQuestCorp.) ftp://ftp.ncifcrf.gov; *MASCOT* (Matrix Science) http://www.matrixscience.com.

(43) Lebl, M.; Krchnak, V. *Methods Enzymol.* **1997**, *289*, 336−392.

(44) Wellings, D. A.; Atherton, E. *Methods Enzymol.* **1997**, *289*, 44−67.

(45) For reviews of ion mobility studies, see: Hagen, D. F. *Anal. Chem.* **1979**, *51*, 870−874. St. Louis, R. H.; Hill, H. H. *Crit. Rev. Anal. Chem.* **1990**, *21*, 321−355. von Helden, G.; Hsu, M. T.; Kemper, P. R.; Bowers, M. T. *J. Chem. Phys.* **1991**, *95*, 3835−3837. Jarrold, M. F. *J. Phys. Chem.* **1995**, *99*, 11−21. Clemmer, D. E.; Jarrold, M. F. *J. Mass Spectrom.* **1997**, *32*, 577−592. Liu, Y.; Valentine, S. J.; Counterman, A. E.; Hoaglund, C. S.; Clemmer, D. E. *Anal. Chem.* **1997**, *69*, 728A−735A.

(46) 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.

(47) Srebalus, C. A.; Li, J.; Marshall, W. S.; Clemmer, D. E. *Anal. Chem.* **1999**, *71*, 3918−3927.

(48) Hoaglund, C. S.; Valentine, S. J.; Sporleder, C. R.; Reilly, J. P.; Clemmer, D. E. *Anal. Chem.* **1998**, *70*, 2236−2242.

(49) Mason, E. A.; McDaniel, E. W. *Transport Properties of Ions in Gases*; Wiley: New York, 1988.

(50) Henderson, S. C.; Li, J.; Counterman, A. E.; Clemmer, D. E. *J. Phys. Chem. B* **1999**, *103*, 8780−8785.

(51) *Insight II 2000*; Accelrys Software: San Diego, CA, 2001.

(52) Mesleh, M. F.; Hunter, J. M.; Shvartsburg, A. A.; Schatz, G. C.; Jarrold, M. F. *J. Phys. Chem.* **1996**, *100*, 16082−16086.

(53) Shvartsburg, A. A.; Jarrold, M. F. *Chem. Phys. Lett.* **1996**, *261*, 86−91.

(54) Srebalus, C. A.; Li, J.; Marshall, W. S.; Clemmer, D. E. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 352−355.

(55) Additional trends are described in Hilderbrand, A. E. Ph.D. Thesis, Indiana University, Bloomington, IN, 2005, work in progress and manuscript in preparation.

(56) For example, see: Matthews, B. W. *Annu. Rev. Phys. Chem.* **1976**, *27*, 493−523. Chou, P. Y.; Fasman, G. D. *Adv. Enzymol. Relat. Areas Mol. Biol.* **1978**, *47*, 45−148. Kite, J. *Structure in Protein Chemistry*; Garland Publishing: New York, 1995. Benner, S. A.; Cannarozzoi, G.; Gerloff, D.; Turcotte, M.; Chelvanayagam, G. *Chem. Rev.* **1997**, *97*, 2725−2843.