Bull. Chem. Soc. Ethiop. **2001**, 15(2), 157-165.
Printed in Ethiopia

ISSN 1011-3924

PROTEINS *IN VACUO*. A MORE EFFICIENT MEANS OF CALCULATING ORIENTATIONALLY-AVERAGED COLLISION CROSS SECTIONS OF PROTEIN IONS

C.T. Reimann*

Department of Analytical Chemistry, Lund University, P.O. Box 124, SE-221 00 Lund, Sweden

(Received August 31, 2001; revised December 26, 2001)

ABSTRACT. With the aim of understanding solvent effects in protein folding, unfolding, stability and dynamic behavior, studies of protein ions in vacuo have become popular in recent years. One experimental descriptor which gives a general overview of ionic structure is the orientationally-averaged collision cross section σ_{avg}, which is obtained from ion drift mobility (IDM) and other kinds of measurements. In modelling protein structures in vacuo with molecular dynamics simulations, it is necessary to calculate σ_{avg} for a plurality of model structures for comparison with experiments. The collision cross section is sensitive to the roughness (concavity) of the protein surface because of the possibility of multiple collisions during an encounter between a given bath gas particle and the protein. Calculations of σ_{avg}, though in principle straightforward, are time consuming. In the work presented below, it was investigated whether a more efficient calculation scheme can be employed without sacrificing too much accuracy. In the new scheme, atomic-scale protein surface granularity is smoothed out by a collected-atoms approach, while large-scale concavity of the protein is essentially preserved.

KEY WORDS: Proteins in vacuo, Orientationally-averaged collision cross sections of protein ions, Modelling protein structures in vacuo, Ion drift mobility, Molecular dynamics simulations

INTRODUCTION

Recently, a number of research groups have been garnering information on the conformation of protein and peptide ions *in vacuo* [1, 2]. Such studies are intended to provide new insight into the role of solvent in protein folding and stability. At present, it can be suggested that protein ions *in vacuo* can both unfold and relax [3], as well as form structures which are as compact as the known native structures in solution phase. Unfortunately, current experiments on proteins *in vacuo* do not yield detailed structural information. Instead, such experiments provide low resolution information of large-scale shape features. Ion-drift mobility (IDM) [4], kinetic energy loss [5], and other gas-phase collision techniques [6] are providing a wealth of data on some large-scale conformational features of gas-phase proteins. These three techniques characterize a conformation in terms of one number — the orientationally-averaged collision cross section σ_{avg}.

*Corresponding Author. Tel.: +46 46-222-8172; Fax: +46 46-222-4544;
E-Mail: curt.reimann@analykem.lu.se

To complement the experiments, some efforts have been made to generate plausible protein ion structures *in vacuo* by molecular dynamics (MD) simulations [4, 7-12]. The obtained structures are characterized in great detail using global descriptors. "Standard" global descriptors include radius of gyration R_{gyr}, principal moments of inertia $\{I_i\}$, and root-mean-square deviations (RMSD) from a standard structure. When comparing MD simulations with the results of experiments, other global descriptors can be calculated, *e.g.* conformer lengths and the computationally more demanding σ_{avg} [13]. By comparison with available experimental data, such global descriptors can be used to assess models for protein ions *in vacuo*, where the similarity of global descriptors can be taken to imply similarity at a finer (but experimentally inaccessible) level of detail. Since the greatest wealth of data has so far been provided by IDM, the focus of the present work has been on computation of σ_{avg} (or, more accurately, the momentum collision integral) for model protein conformers.

The most general and time-consuming way of calculating σ_{avg} for a specific conformation is a full trajectory method (Figure 1a) taking into account both long-range and short-range (collisional) interactions between bath gas atoms and the protein atoms [14]. Aside from hard collisions which resemble simple bounces (bath particle *i* in Figure 1a) even particles passing outside the hard-spheres contact radii can be deflected substantially (bath particle *iii* in Figure 1a). To save computer time, a simpler exact hard-spheres scattering (EHSS) model for calculating collision cross sections can be employed [13]. This method takes into account scattering and multiple collisions of a bath gas particle with atoms on the protein surface but lumps all the details of the interaction potentials into hard-sphere contact distances R_{coll} appropriate for the interaction of the protein and bath atom types at the relevant temperature and relative impact speed range (Figure 1c). For the full trajectory method, a given bath gas atom interacts with a number of protein atoms simultaneously, while for the EHSS method a given bath gas atom interacts with at most one protein atom at a time (compare *i, ii* and *iii* in Figure 1a,c and note differences in the trajectories). In both cases, if bath gas particles incident at a certain impact parameter and protein orientation on infinitesimal transverse area *dA* are deflected through angle χ, then the contribution of that area to the collision cross section for that protein orientation is $d\sigma = dA[1 - \cos \chi]$. The angle χ is defined in Figure 1b in terms of the path of the bath particle prior to and after completing interactions with the protein. *dA* can be thought of as the cross sectional area per bath particle "bombarded" by the bath particles. To obtain σ_{avg} it is necessary to integrate $d\sigma$ over all approaches between bath gas atoms and the protein which yield significant momentum transfer. This means averaging over both protein orientation and impact parameter. (As implemented so far these methods assume that the protein is perfectly rigid and much heavier than the bath gas particle.)

In previous work [15], it has been noted that when the contact surfaces of a collection of atoms merge so as to approximate a simple object such as a cylinder, calculations of σ_{avg} for such atomic clusters can be simplified by using the approximate form. Also, in MD simulations, it is fairly common to treat atom combinations like —CH, —CH_2, and —CH_3 in the "united atoms" approximation in which each combination is replaced by a single "atom" with appropriately expanded van der Waals radius. A similar approach is taken in the work presented below. Instead of treating a protein atomistically, proximate and related atoms are lumped together and replaced by a ball of radius $R_{coll,i} = \sqrt{\sigma_{avg,i}/\pi}$ where $\sigma_{avg,i}$ is the collision cross section of only those atoms in lump *i,* isolated in space, computed by any suitable method. Then, σ_{avg} is calculated for the collection of balls. Since the number of balls is much less than the original number of atoms, considerable computing time should be saved using the EHSS method. Below are presented results of a test of this idea for a set of 75 residue proteins.

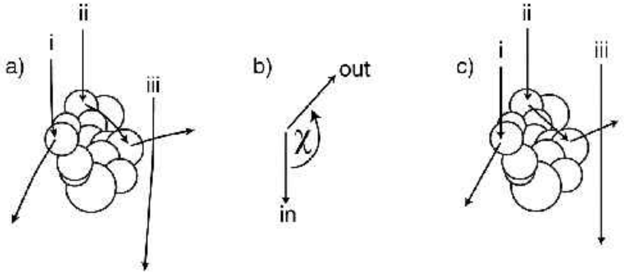

Figure 1. Schematic of interactions of bath gas particles with a protein. Bath gas particles are dimensionless points while the protein atoms are represented by appropriate contact spheres. *i, ii* and *iii* are example encounters of bath gas particles with the protein. a) Full trajectory approach: both long-range attraction and close, bounce-like collisions are apparent. b) Definition of scattering angle χ in terms of the overall change in the path of the bath gas particle after all interactions have ended. c) Hard-spheres approach wherein a gas bath particle interacts with at most one protein atom at a time. The trajectories are different in detail but the contact spheres are selected to give as representative an overall result as possible for quantities like collision cross section which depend on the outcome of one or more scattering events.

METHODS

For any assembly of atoms interacting with a He bath gas, the exact hard-spheres scattering method was carried out using contact/collision radii R_{coll} = 0.220, 0.265, 0.250, 0.245, 0.290 nm, respectively for H, C, N, O, and S. Where necessary, the program MOLMOL [16] was employed to add missing non-polar hydrogens to downloaded coordinate sets from the protein data base (PDB, Ref. [17]). The calculations used a working protocol of an average of three runs of 500 randomly-oriented snapshot cross section calculations apiece (with different random number seeds for each run). An implementation of EHSS was tested on the C_{60} fullerene structure. For fullerene, the projected hard-spheres cross section — the average union of the projected collisional cross sectional areas of the atoms [18] — calculated using a He-C hard-sphere contact distance of 0.286 nm, should equal the EHSS cross section calculated using a hard-sphere contact distance of 0.281 nm [13]. The results were in agreement to within 0.2%. For the 75-residue protein 1ghc, all-atom EHSS calculations using the employed protocol had a 95% chance of deviating no more than 0.4% from the best average obtained using more extended calculation protocols. However, the aim of the study was to explore whether condensation of the protein atoms into simple groupings of atoms would lead to faster calculations of σ_{avg} without sacrificing too much accuracy. Such a method will here be referred to as the collected-atoms approximation.

There are many ways to group atoms in a protein. A favored way would be to generate standard subsets of atoms (indexed by i) so that it is not necessary to always recalculate the cross section $\sigma_{avg,i}$ of each subset. In the scheme employed in the present work, the standard atom groupings employed were: the N-terminus up to the first C_α and its hydrogen; "connective" elements consisting of —(C=O)-(NH)—; amino acid sidechains; and the C-terminus. For a collection of proteins, all standard groupings in turn were submitted to a calculation of all-atoms EHSS cross section. For each distinct type of atom grouping, for example all tested alanine sidechains, an average cross section $\overline{\sigma}_{avg,i}$ was determined and converted to a radius according to $R_{coll,i} = \sqrt{\overline{\sigma}_{avg,i}/\pi}$. So, in this very simple model, even if alanine sidechains have some variability in their conformation from protein to protein, only a single representative collision radius was extracted. Finally, once the $R_{coll,i}$ values are calculated, in a given structure balls of those radii were placed at the corresponding geometrical centers of the atom groupings. The collection of these larger balls — the collected-atoms representation — was then subjected to the EHSS method for calculation of the collision cross section.

Figure 2 shows collision surfaces for the protein 1ghc. The all-atom approach is shown in Figure 2a, and the collected-atom approach is shown in Figure 2b using results summarized below and in the Table. The structure of the overall collision surface has a finer grain in Figure 2a but gross surface concavity is still well apparent when the collected atom spheres are used in Figure 2b.

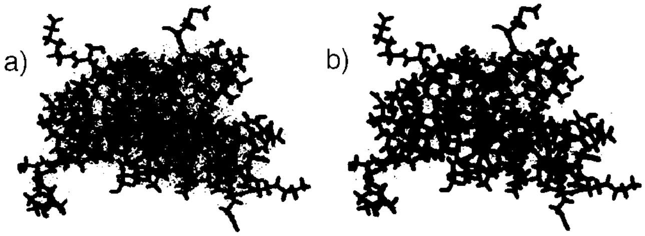

Figure 2. Hard-spheres collision surfaces of protein 1ghc represented by dotted texture rendered by MOLMOL [16]. a) All-atoms approach normally employed; b) collected-atoms approach considered in the present work. The surface granularity is finer in a) but gross surface concavity is preserved in b).

Selection of protein models from the PDB was guided by the procedure of Arteca and Tapia [19]. Consideration of condensed-phase structures is reasonable, as under suitable conditions gas-phase protein ions [4] and ionized supermolecular protein assemblies like viruses [20] can maintain native conformations. The focus of the study reported here was on proteins of sequence length 75 and with X-ray structures refined to better than 3 Å, as well as on a few non-native 75-residue subsets carved out of a larger protein, hen egg lysozyme. However, effective collision radii for key atom groupings were obtained based on a more extended set of proteins with 75, 129, 212, and 316 amino acids.

RESULTS AND DISCUSSION

The first task was to calculate the cross sections of all the available examples of each type of atom grouping and investigate the results for consistency. An obvious choice was to compare results obtained for proteins of different chain lengths. Due to space limitations only examples of the results can be given here.

Consider the sidechain of alanine (ALA). Distributions of sidechain $\sigma_{avg,i}$ are shown in Figure 3a. Using the Student's t-test as well as the Kolmogorov-Smirnov test [21] the distribution obtained from 75-residue proteins is likely (P<1%) to be different from the distributions obtained from either 129-, 212-, or 316-residue proteins, while the latter distributions could be the same (P = 3-60%). Yet, 95% of the $\sigma_{avg,i}$ values lie in a relative range of only 1% and the biggest difference in mean between different data sets for ALA is only 0.2%. Thus, it is justifiable to assume that $\sigma_{avg,i}$ is 0.2635 nm^2 for ALA. (Likewise, for the sidechain glycine (GLY), containing only one atom, $\sigma_{avg,i}$ is 0.1521 nm^2).

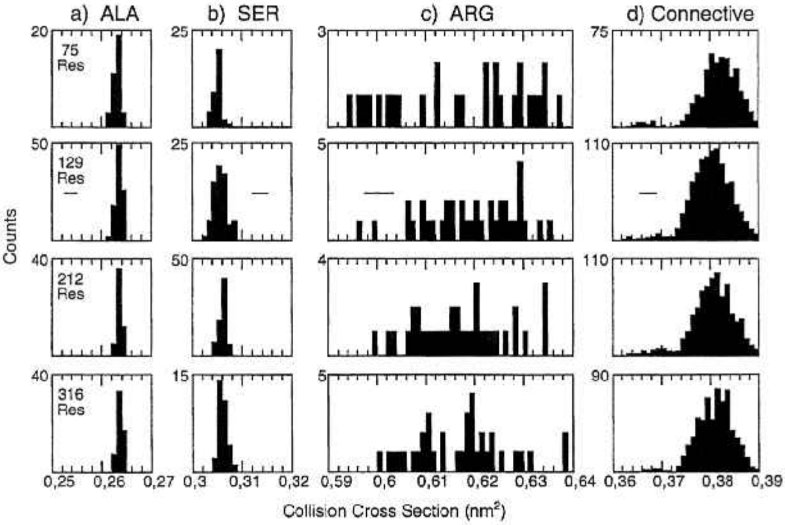

Figure 3. Histograms of all-atom exact hard-spheres scattering cross sections shown column-wise for: a) alanine sidechain; b) serine sidechain; c) arginine sidechain; and d) connective elements. The histograms are shown row-wise for proteins consisting of 75, 129, 212 and 316 residues. The bars indicate a 1% relative range of collision cross section in each column.

Consider the sidechain of serine (SER). Distributions of sidechain $\sigma_{avg,i}$ are shown in Figure 3b. Statistical tests again reveal substantial differences between results obtained from different sized proteins. Here a 1.5% range of $\sigma_{avg,i}$ contains 95% of the observed $\sigma_{avg,i}$

values. At a 1.5% range it can seem uncomfortable to use a single value for $\sigma_{avg,i}$, 0.3058 nm^2, but in fact mean values differ by at most 0.3% between the different data sets. The situation with bulky sidechains which are likely to be on the surfaces of proteins has less favorable aspects. Consider arginine (ARG), Figure 3c. On the one hand, statistical tests cannot distinguish the distributions. On the other hand, the spread in data is 6% containing 95% of the $\sigma_{avg,i}$ values. Using a single value for $\sigma_{avg,i}$, 0.6176 nm^2, will seem quite risky. Nevertheless, even for ARG, the maximum difference in the mean between two data sets is only 0.4%.

A special issue is the connective elements, as there are so many of them, see Figure 3d. The main portion of the distribution peaks at 0.381 nm^2, while a small peak appears at 0.37 nm^2. The latter is associated with connective elements leading into proline (PRO) which simply lack the backbone amide hydrogen. Considering the main portion of the distribution, 95% of all the $\sigma_{avg,i}$ values lie within a relative range of 5%, while means do not differ by more than 0.4%.

Mean values of EHSS cross sections obtained from data sets of the different sized proteins differ from each other by more than 0.5% for the N-terminus, isoleucine (ILE), lysine (LYS), cysteine (CYS), asparagine (ASN), glutamine (GLN), histidine (HIS), and tryptophan (TRP). Also, as shown in Table 1**,** for ILE, MET, LYS and ARG, the distributions have relative widths exceeding 5%. Again it would seem risky to take standard values for $\sigma_{avg,i}$, but it was decided to do so anyway as a test of a relatively simple methodology which does not require recalculating $\sigma_{avg,i}$ for every conformer. Consequently, for the set of 75-residue proteins, the N- and C-termini, the amino acid sidechains and the connective elements were replaced by larger spheres according to $\bar{\sigma}_{avg,i}$ in Table 1 and $R_{coll,i} = \sqrt{\bar{\sigma}_{avg,i}/\pi}$. Then the EHSS method was applied to both the all-atoms representation and the collected atoms representation of each protein.

Table 1. Summary of exact hard-spheres scattering cross sections calculated for the standard collected atom groupings described in the text. Given in each case is a grand average value $\bar{\sigma}_{avg,i}$ from four sets of proteins comprising 75, 129, 212 and 316 residues. "95% Range" is the percentage range in each case which encompasses 95% of all the $\sigma_{avg,i}$ values.

Atom groupings	$\bar{\sigma}_{avg,i}$ (nm^2)	95% Range
N-TER	0.300	1.61%
CON	0.381	4.88%
C-TER	0.290	1.78%
GLY	0.152	N.A.
ALA	0.264	0.96%
PRO	0.414	1.87%
VAL	0.423	1.93%
ILE	0.487	5.20%
LEU	0.483	2.55%
MET	0.496	5.01%
PHE	0.559	2.66%

Atom groupings	$\bar{\sigma}_{avg,i}$ (nm^2)	95% Range
ASP	0.389	2.31%
GLU	0.461	2.79%
LYS	0.550	6.44%
ARG	0.618	6.29%
SER	0.306	1.52%
CYS	0.359	3.26%
THR	0.385	2.74%
ASN	0.406	2.58%
GLN	0.476	3.42%
HIS	0.505	2.96%
TYR	0.589	2.65%
TRP	0.662	2.55%

The chief result of this study is shown in Figure 4. Plotting cross sections calculated using the all-atoms method *versus* cross sections calculated using the collected-atoms (CA) method leads to a straight line relation of the form $\sigma_{avg,all-atoms} = 0.494 + 0.978\sigma_{avg,CA}$; $R^2 = 0.998$ with RMS deviation of 0.56%. Since the protocol for all-atoms EHSS is accurate to about 0.4% the collected-atoms method at present is probably accurate to slightly better than 0.7-1%. As implemented, the EHSS method carried out in the collected-atoms representation is completed about twice as rapidly as the all-atom EHSS.

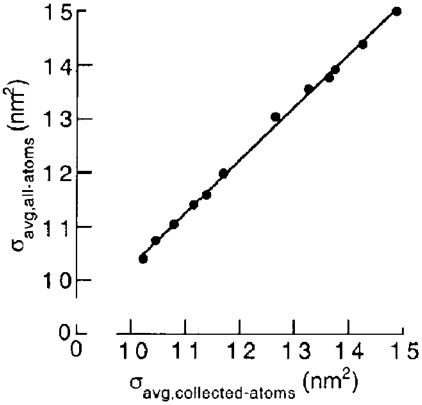

Figure 4. Plot of all-atom EHSS cross sections $\sigma_{avg,all-atoms}$ against EHSS scattering cross sections calculated in the collected atoms representation $\sigma_{avg,collected-atoms}$. The line is the result of curve fitting.

Figure 5 shows the contribution to σ_{avg} of bath-protein encounters characterized by different numbers of bounces (*i.e.,* different numbers of collisions between a given bath gas particle and the protein atoms during the interaction), and also the average value of $1 - \cos\chi$ for different numbers of bounces. For the two examples shown, 1ghc and 1tvs, use of the collected-atoms representation slightly exaggerates the single-bounce contribution to σ_{avg}, while the multi-bounce contribution is slightly underestimated. The bounce-resolved average deflection, expressed as the average value of $1 - \cos\chi$ for different numbers of bounces, is very similar between the collected-atoms approach and the all-atom EHSS approach. At this level of approximation, use of the collected-atoms representation leads only to subtle differences in average behavior and thus appears to be a useful approximation.

In the EHSS implementation, for each protein orientation, a number of bath gas particles encounter the protein, and the successive collisions or bounces of each bath gas particle are followed until the particle departs permanently from the protein. While a typical *atom* has a collision cross section of about 0.22 nm^2, the typical *cluster* of collected atoms considered here has a collision cross section of about 0.3-0.6 nm^2. Thus, when using the collected-atoms representation, acceptable accuracy should be attainable with about two times fewer bath gas particles encountering the overall cross sectional area of the protein than what is used in the usual all-atoms EHSS approach. This expectation was borne out in the present study.

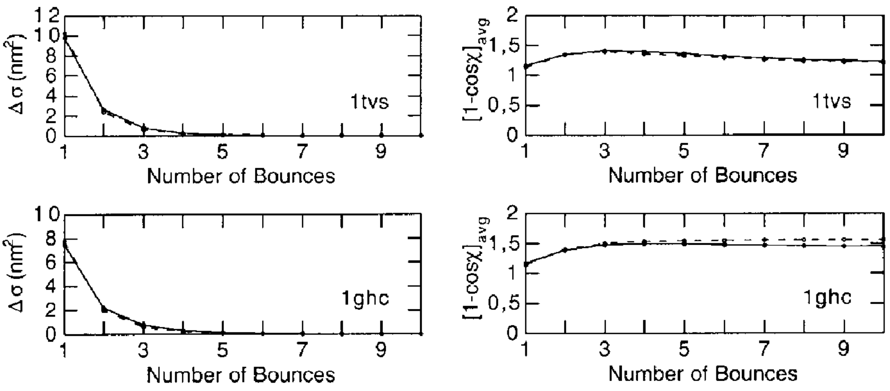

Figure 5. For the proteins 1tvs and 1ghc, the contribution to the scattering cross section of scattering events involving different numbers of hard collisions with the protein (left); and the scattering angle, expressed as $1-\cos\chi$, characteristic of scattering events involving different numbers of hard collisions with the protein (right). Solid line and closed circles: all-atoms representation. Dashed line and open circles: collected atoms representation.

CONCLUSIONS

A strategy of calculating the exact hard-spheres scattering cross section using clumps of atoms approximated as spheres gives acceptable accuracy and improves the computation speed by a factor of about four compared to an all-atoms approach used previously.

Considering proteins with 75 residues interacting with bath gas particles such as helium, the overwhelming majority of bath gas particles undergo either one or two collisions with the protein. Though the bounce trajectories are altered in detail by the collected atoms approach, overall the result is much the same. When the protein is represented by clumps of atoms — the collected-atoms representation — rather than by the atoms directly, so that the protein is simulated by about ten times fewer but correspondingly bigger spheres, there is enough granularity left to yield roughly the same result. One may say that at the level of approximation represented by the exact-spheres scattering method for calculating the collision cross section, the finer atomic level of the molecular surface granularity is not sensed to any great extent.

According to the calculations presented above, the proteins are characterized by a certain roughness. The collision cross sections are roughly 25% greater than the average projected areas of the proteins, and deflection angles χ are on average 110 to 120°.

In future work, it will be desirable to study the extensibility of the methodology presented above to larger proteins as well as to grossly non-native structures. In addition, one should search for a simple means of obtaining the cross section of each atom grouping actually appearing in a given conformation, rather than just relying on the mean value of the collision cross section for each type of group, *i.e.* customize the calculation rapidly to each conformation.

It will also be of interest to work with Gaussian ("fuzzy") shape descriptors [22] as a means of modelling the effects of molecular fluctuations on the cross section calculations.

ACKNOWLEDGEMENT

Special thanks to Sven Oscarsson for help with computer resources, and to a referee for pointing attention to Ref. [22].

REFERENCES

1. Hoaglund-Hyzer, C.S.; Counterman, A.E.; Clemmer, D.E. *Chem. Rev.* **1999**, 99, 3037.
2. Jarrold, M.F. *Ann. Rev. Phys. Chem.* **2000**, 51, 179.
3. Jarrold, M.F. *Acc. Chem. Res.* **1999**, 32, 360.
4. Mao, Y.; Woenckhaus, J.; Kolafa, J.; Ratner, M.A.; Jarrold, M.F. *J. Am. Chem. Soc.* **1999**, 121, 2712.
5. Nesatiy, V.; Chen, Y.-L.; Collings, B.A.; Douglas, D.J. *Rap. Commun. in Mass Spectrom.* **1998**, 12, 40.
6. Jørgensen, T.J.D.; Andersen, J.U.; Hvelplund, P.; Sørensen, M. *Intl. J. Mass Spectrom.* **2001**, 207, 31.
7. Reimann, C.T.; Velázquez, I.; Tapia, O. *J. Phys. Chem. B* **1998**, 102, 2277.
8. Reimann, C.T.; Velázquez, I.; Tapia, O. *J. Phys. Chem. B* **1998**, 102, 9344.
9. Reimann, C.T.; Velázquez, I.; Bittner, M.; Tapia, O. *Phys. Rev. E* **1999**, 60, 7277.
10. Velázquez, I.; Reimann, C.T.; Tapia, O. *J. Phys. Chem. B* **2000**, 104, 2546.
11. Arteca, G.A.; Velázquez, I.; Reimann, C.T.; Tapia, O. *Chem. Phys. Lett.* **2000**, 327, 245.
12. Mao, Y.; Ratner, M.A.; Jarrold, M.F. *J. Phys. Chem. B* **1999**, 103, 10017.
13. Shvartsburg, A.A.; Jarrold, M.F. *Chem. Phys. Lett.* **1996**, 261, 86.
14. Mesleh, M.F.; Hunter, J.M.; Shvartsburg, A.A.; Schatz, G.C.; Jarrold, M.F. *J. Phys. Chem.* **1996**, 100, 16082.
15. Jarrold, M.F.; Constant, V.A. *Phys. Rev. Lett.* **1991**, 67, 2994.
16. Koradi, R.; Billeter, M.; Wüthrich, K. *J. Mol. Graph.* **1996**, 14, 51.
17. http://www.rcsb.org/pdb/
18. Wyttenbach, T.; von Helden, G.; Batka, J.J.J.; Carlat, D.; Bowers, M.T. *J. Am. Soc. Mass Spectrom.* **1997**, 8, 275.
19. Arteca, G.A.; Tapia, O. *J. Chem. Inf. Comput. Sci.* **1999**, 39, 642.
20. Siuzdak, G. *J. Mass Spectrom.* **1998**, 33, 203.
21. Press, W.H.; Teukolsky, S.A.; Vetterling, W.T.; Flannery, B.P. *Numerical Recipes in Fortran 77*; Cambridge University Press: Cambridge; **1999**.
22. Grant, J.A.; Gallardo, M.A.; Pickup, B.T. *J. Comput. Chem.* **1996**, 17, 1653.