JCTC Journal of Chemical Theory and Computation

On the Use of low-resolution Data to Improve Structure Prediction of Proteins and Protein Complexes

Marco D'Abramo,[†] Tim Meyer,[†] Pau Bernadó,[‡] Carles Pons,[§,⊥] Juan Fernández Recio,[§] and Modesto Orozco*[,†,||,⊥]

Molecular Modeling and Bioinformatics Unit, IRB-BSC Joint Research Program in Computational Biology, Institute for Research in Biomedicine Josep Samitier 1-5, Barcelona 08028, Spain and Barcelona Supercomputing Center, Jordi Girona 29, Barcelona 08034, Spain, Structural and Computational Biology Program, Institute for Research in Biomedicine Josep Samitier 1-5, Barcelona 08028, Spain, Life Sciences Department, Barcelona Supercomputing Center, Jordi Girona 29, Barcelona 08034, Spain, Departament de Bioquímica i Biología Molecular, Facultat de Biología, Universitat de Barcelona, Avgda Diagonal 645, Barcelona 08028, Spain, and National Institute of Bioinformatics, Parc Científic de Barcelona, Josep Samitier 1-5, Barcelona 08028, Spain

Received June 16, 2009

Abstract: We present a systematic study of the ability of low-resolution experimental data, when combined with physical/statistical scoring functions, to improve the quality of theoretical structural models of proteins and protein complexes. Particularly, we have analyzed in detail the "extra value" added to the theoretical models by: electrospray mass spectrometry (ESI-MS), small-angle X-ray scattering (SAXS), and hydrodynamic measurements. We found that any low-resolution structural data, even when (as in the case of mass spectrometry) obtained in conditions far from the physiological ones, help to improve the quality of theoretical models, but not all the coarse-grained experimental results are equally rich in information. The best results are always obtained when using SAXS data as experimental constraints, but either hydrodynamics or gas phase CCS data contribute to improving model prediction. The combination of suitable scoring functions and broadly available low-resolution structural data (technically easier to obtain) yields structural models that are notably close to the real structures.

Introduction

The prediction of the three-dimensional structures of proteins based only on the knowledge of their sequences has been the "Holy Grail" of computational biology for many years. Proteins with very close homologues of known structure can now be safely modeled with a quite good global quality,[1,2] but there are still many proteins for which homologues cannot be found in structural databases, and they have to be modeled based on more risky threading or ab initio methods.[2] Recent versions of these programs, combined with suitable scoring functions, are able to provide ensembles of reasonable solutions among which the real one is hidden. Unfortunately, as CASP experiments have demonstrated,[3] it is not always easy to detect the best solution among many reasonable ones. The situation is even more challenging in the prediction of protein−protein complexes, especially in those cases where an important degree of structural distortion in the monomers is required for the assembly. Although recent CAPRI (http://www.ebi.ac.uk/msd-srv/capri) experiments have demonstrated that, at least in some cases, it is possible to produce

* Corresponding author. E-mail: modesto@mmb.pcb.ub.es.
[†] Molecular Modeling and Bioinformatics Unit, Institute for Research in Biomedicine and Barcelona Supercomputing Center.
[‡] Structural and Computational Biology Program, Institute for Research in Biomedicine.
[§] Life Sciences Department, Barcelona Supercomputing Center.
[||] Universitat de Barcelona.
[⊥] National Institute of Bioinformatics.

10.1021/ct900305m CCC: $40.75
Published on Web 09/28/2009

theoretical models of sufficient quality for accurate biological and functional annotation, there are still too many cases for which even the best protein−protein docking codes suggest structural models that are very far from the reality.[4] Errors in the prediction of the structure of proteins or protein complexes might lead to the design of irrelevant experiments and to the formulation of erroneous functional hypothesis.

Due to the complexity of predicting the structure of proteins and protein complexes based only on theoretical methods, several authors have supported the use of experimental data to restrain the accessible space to be sampled by the theoretical algorithms. In this field, the combination of physical methods like molecular dynamics (MD) with high-resolution techniques, such as X-ray or NMR spectroscopy, has been very fruitful for many years and is considered the gold standard for protein structure determination.[5] Unfortunately, high-resolution methods are not always easy to apply in a high-throughput mode for proteins of moderate or large size, and MD is not the best technique to predict the novo protein structures due to CPU requirements and to force-field uncertainties. Thus, alternative methods need to be designed to combine methods for the prediction of the structure of proteins and protein complexes with easy-to-obtain low-resolution structural data.

Evolutionary analysis, through the detection of correlated mutations,[6] the evolutionary trace method,[7] or the use of environment-specific substitution tables,[8] is a source of low-resolution data that can be easily incorporated into structure prediction to discard unlikely models.[9−11] Site-directed mutagenesis has been another source of data on the location of individual residues in different protein regions or on the relative positioning of pairs of residues in the three-dimensional structure.[12,13] A similar type of information can be obtained by using covalent cross-linkers coupled, for example, to mass spectroscopic measures.[14] However, all these techniques have obvious caveats: (i) evolutionary analysis from multiple sequence/structure alignments is prone to error, and it requires a massive amount of data that is not always available, and (ii) site-directed mutagenesis and cross-linking experiments are technically complex and (as correlated mutations) provide only local information of a few selected residues, not on global three-dimensional structure. Cryo-electron tomography is a very promising technique able to provide detailed structural information of the sample but not fast and cheap enough, yet, to be easily used in a high-throughput context. We believe that more general structural information on proteins and complexes can be gained from other low-resolution biophysical methods, such as the small-angle X-ray scattering (SAXS) hydrodynamic radius (R_h) estimations derived from hydrodynamic measurements or the apparent charge or collision cross-sections (CCS) determined in mass spectroscopy experiments. All these methods are simple and fast and can be performed at a high-throughput regime, making them ideal for proteome-scale or cell-scale determination of proteins or protein−protein complexes. Unfortunately, none of them is able to provide by themselves unambiguous three-dimensional structural models of proteins. In this paper, we provide a systematic analysis on the robustness of these low-resolution data and on their ability to enrich structural predictions of theoretical models of protein and protein−protein complexes.

Methods

Calculation of Synthetic Low-Resolution Structural Data. We have explored three experimental observables providing low-resolution structural information on proteins: (i) the collision cross-section (CCS), (ii) the small-angle X-ray scattering spectra (SAXS), and (iii) the hydrodynamic radius (R_h). CCS is experimentally derived from the time-of-flight of protein ions in a spectrometer drift tube in the presence of inert gases under electrospray vaporization conditions and provides information on the effective area of collision of the inert molecules with a protein under vacuum conditions.[15] SAXS profiles are experimentally obtained from the analysis of the elastically scattered X-rays by particles in solution and provide low-resolution (>15 Å) structural information of them, essentially referring to their size and shape.[16−18] Finally, the hydrodynamics radii also provide information on the overall size of the proteins determined by its self-diffusion coefficients via the Stokes−Einstein relationship.[19] For benchmark purposes, low-resolution structural data were simulated here from known experimental structures or from MD trajectories to probe the robustness of the data to dynamics and/or environmental effects (see below).

A subset of CCS values were evaluated using: (i) the most accurate (but computationally demanding) trajectory method (TM),[23] where the colliding ions are treated as a collection of atoms, each one represented by a 12−6−4 potential (i.e., including a realistic treatment of long-range interactions between the ion and the buffer gas, which have been found to significantly affect the CCS), and the orientationally averaged collision integral is determined by averaging over all possible collision geometries and (ii) the faster but less accurate projection approximation (PA), as implemented in the sigma software,[31] which essentially finds the average "shadow" as a trial conformer is rotated through all possible orientations, disregarding the details of the scattering process. The good correlation between PA and TM values ($r = 0.99$, data not shown) allowed to extrapolate accurate TM-CCS values by applying an empirical correction factor of 1.3 to the PA-CCS estimates. The SAXS curves were simulated by means of the CRYSOL program[22] using default parameters, and the HydroNMR[20] software package was used to calculate R_h using a value of 3.3 Å as the atomic element radius.[21] The calculation of low-resolution parameters from known experimental data raises some concerns that need to be considered before evaluating the information load contained in these data: (i) in the case of CCS, experimental measures are recorded in the gas phase, and it is not clear how well gas-phase structures represent solution ensembles,[24,25] (ii) Dynamics effects are expected to introduce non-negligible changes of different magnitude, and (iii) low-resolution experimental measures are always prone to errors and to uncertainties that cannot be neglected. Thus, as a first step in our study, we checked the goodness of gas-phase electrospray mass spectrometric (MS-ESI) experiments as sources of structural information on the solution structure.

This was done by performing extended (0.1 μs long) MD simulations in gas-phase conditions for different proteins: bovine pancreatic trypsin inhibitor (1BPI), Cytochrome c6 (1LS9), egg-white lysozyme (1LYS), and ubiquitin (1UBQ) as well as protein−protein complexes: (1ACB, 1CBW, 1CSE, and 3TGI). Protein charges were determined from their empirical relationship to the solvent accessible surface area (SASA),[26] charges were placed at most favorable positions using an iterative titration algorithm[27] which yields a quite symmetric charge distribution (less careful titration procedures, leading to local charge concentration, are likely to produce artifactual unfoldings of protein structure even during short MD simulations). Previous calculations on a large set of protein folds[27] demonstrated that the trajectories are not very sensitive to alternative (if reasonable) choices of titrable site, which means that the ensemble of conformations sampled for the most stable charge configuration (for a given charge state) is a good representative of the ensemble of conformations found experimentally (where different distributions of charge might coexist). In order to check the importance of the total charge in the structure, we performed additional simulations using slightly different charge states for three proteins for which CCS was experimentally determined for different charge states (1UBQ, 1LS9 and 1LYS; see data on http://www.indiana.edu/~clemmer/Research). For one of the proteins (1UBQ), calculations were repeated using a large charge density to check whether or not the protein structure will be stable when very heavily charged.

MD simulation protocols described elsewhere[27] were used to obtain vacuum trajectories for the different protein systems. In order to guarantee that results were not contaminated from force-field uncertainties, protein simulations were repeated using three of the most popular force fields (AMBER-parm03,[28] OPLSAA,[29] and CHARM22[30]). These calculations did not reveal any force-field dependent artifacts, and accordingly, more costly simulations for protein complexes were performed considering only the parm03 force field. Snapshots from the 0.1 μs trajectories were collected every 2 ps and used for structural analysis and for the computation of the CCS. All gas-phase simulations were carefully checked to verify the lack of structural distortions in the gas phase, which could create doubts on the validity of the CCS measures as descriptors of protein structure in solution.

The robustness of low-resolution measures to protein flexibility was analyzed by running MD simulations for the different proteins and protein complexes (see Table 1) in aqueous solution using the TIP3P water model[32] with periodic boundary conditions and with particle mesh corrections[33] in the isothermal ($T = 300$ K) isobaric ($P = 1$ atm) ensemble. Trajectories were collected using AMBER parm03 for 100 ns (10 for complexes) after 1 ns of equilibration, following the procedure noted elsewhere.[27] MD ensembles were collected and used to predict the R_h, CCS, and SAXS profiles, which were then compared to those obtained by a single structure (experimental or the MD-averaged one). Experimental data, when used to determine a single structure, have an associated error related to the technique itself and to the fluctuations in the structure. To estimate an upper limit for this magnitude, we used the standard deviations (σ) in the parameters, as provided by MD simulations, which undoubtedly represent an overestimation of the experimental uncertainty, i.e., values outside the $\pm 3\sigma$ interval for CCS ($\pm 2\sigma$ for $<\chi>_{SAXS}$ and R_h), with respect to those corresponding to the experimental structures were disregarded. This choice means that the test of enrichment below represents then a lower limit of accuracy for the proposed methodology.

Empirical Scoring Functions. We have complemented the use of low-resolution structural data by introducing the empirical scoring functions developed to detect local errors in the structures of protein monomers and protein complexes, which are probably not detected in low-resolution structural data, as those used in this paper. For this purpose, ProSA-II[34] was used to study monomeric proteins, and pyDock[35] was used to analyze protein complexes. ProSA-II is a diagnostic tool that is based on the use of residue−residue potentials of mean force derived from the statistical analysis of all available protein structures. The scoring function in pyDock is composed of "soft-truncate" van der Waals (with 0.1 weighting factor; AMBER parameters; with maximum values of +1.0 kcal/mol), Coulombic electrostatics (with distance-dependent dielectric constant, AMBER charges with the Coulombic term truncated to ± 1.0 kcal/mol), and accessible surface-area-based desolvation energy with atomic solvation parameters (ASP) previously optimized for rigid-body docking. This scoring scheme has shown top performance in the scorer experiment at the most recent CAPRI competition. Before pyDock scoring, incomplete side chains have been automatically rebuilt with SCWRL 3.0.[36]

Set of Predicted Structures. The ability of low-resolution data (supplemented by empirical scoring functions) to improve model prediction was first tested using all the model sets submitted to the CASP7 competition. For each model, CCS, SAXS curve, R_h, and PROSA-II, Z-scores[34] were computed. Models showing one or more of the four observables (CCS, $<\chi>_{SAXS}$, R_h, and Z_{score}^{PROSA}) in severe disagreement (see above) with the corresponding experimental structure were discarded, and a new prediction set was calculated. In order to generate the prediction curves of the protein complexes, all the available models for predictors found in the CAPRI experiment web site (http://www.ebi.ac.uk/msd-srv/capri) were used. The general protocol for structure observable prediction was as above, with the only difference that we used the energy-based scoring function provided by pyDock instead of that in PROSA-II. In all cases, refined predictions were compared to a background model obtained by repeating the procedure but setting to infinite level of noise.

Results and Discussion

Effect of Structural Flexibility on Low-Resolution Data. MD simulations were used to generate ensembles of conformations for different proteins and protein complexes and to check the robustness of the different low-resolution data to structural fluctuations. Average results in Table 1

Table 1. MD Simulations for the Different Proteins and Protein Complexes[a]

simulation[b]	RMSd Å	std dev	TmScore (Å)	std dev	R_g (Å)	std dev	R_h (Å)	std dev	CCS (Å^2)
1UBQ P03	4.2	0.3	1.4	0.1	11.8	0.1	17.0	0.1	1 054
1UBQ C22	4.4	0.2	1.9	0.1	11.7	0.1	17.0	0.2	1 043
1UBQ ON2	4.5	0.2	1.9	0.1	11.8	0.1	17.0	0.1	1 051
1UBQ P99 T3P	1.5	0.4	1.2	0.1	11.8	0.1	17.3	0.2	1 084
1UBQ X-ray					11.6		17.1		972
1UBQ expt									1 050[c]
1BPI P03	2.4	0.2	1.4	0.1	11.7	0.1	16.1	0.2	919
1BPI C22	2.6	0.2	2.0	0.2	11.5	0.1	16.1	0.3	912
1BPI ON2	2.8	0.2	2.0	0.1	11.4	0.1	15.9	0.3	916
1BPI P99 T3P	1.2	0.2	1.4	0.1	11.0	0.1	15.9	0.2	977
1BPI X-ray					11.3		16.2		797
1BPI expt									900[c]
1LYS P03	2.8	0.1	1.9	0.1	13.7	0.1	19.3	0.1	974
1LYS C22	3.5	0.2	2.1	0.1	13.8	0.1	19.6	0.1	988
1LYS ON2	4.4	0.2	2.3	0.1	13.7	0.1	19.6	0.1	996
1LYS P99 T3P	1.0	0.2	1.1	0.1	14.2	0.1	20.2	0.1	1 088
1LYS X-ray					13.9		19.8		1 031
1LS9 P03	2.4	0.1	1.7	0.1	12.0	0.1	17.4	0.2	761
1LS9 C22	3.6	0.3	2.1	0.1	12.2	0.1	17.5	0.2	777
1LS9 ON2	2.9	0.2	2.0	0.1	12.2	0.1	17.6	0.2	778
1LS9 P99 T3P	1.3	0.3	1.5	0.2	12.4	0.1	18.1	0.2	820
1LS9 X-ray					12.0		17.7		716[c]

complexes	RMSd (Å)	std dev	TmScore (Å)	std dev	R_g (Å)	std dev	R_h (Å)	std dev	CCS (Å^2)
1ACB X-ray					19.1		27.2		1 842
1ACB T3P	1.6	0.2	1.3	0.3	19.4	0.1	26.8	0.2	1 916
1ACB vac	3.0	0.1	2.5	0.3	17.4	0.1	25.6	0.2	1 772
1CBW X-ray					18.8		27.4		1 797
1CBW T3P	1.8	0.5	1.5	0.8	19.2	0.1	27.0	0.3	1 917
1CBW vac	4.7	0.2	2.1	0.1	17.1	0.0	26.0	0.2	1 799
1CSE X-ray					19.1		27.2		1 858
1CSE T3P	1.2	0.2	0.8	0.1	19.4	0.1	26.8	0.2	1 906
1CSE vac	2.9	0.3	1.8	0.2	18.5	0.1	25.6	0.2	1 782
3TGI X-ray					17.2		25.8		1 749
3TGI T3P	1.3	0.2	1.1	0.1	18.8	0.1	26.4	0.3	1 809
3TGI vac	5.8	0.4	3.2	0.2	18.1	0.1	25.2	0.3	1 770

[a] The code after the PDB name indicates the force field used in the protein MD simulations (P03 = AMBER-parm03, C22 = CHARM22, and ON2 = OPLSAA). T3P and vac denotes simulations in water using the TIP3P water model and in vacuum conditions, respectively. [b] Charge states used in the MD simulations were +6 for 1UBQ, +6 for 1BPI, +8 for 1LYS, and +7 for 1LS9. Where available, the corresponding experimental CCS values for the same charge state used in the simulation are reported. [c] Experimental CCS values were taken from ref 41a for 1UBQ, from ref 41b for 1BPI, and from ref 41d41d for 1LS9.

and specific examples in Figure 1 demonstrate that flexibility introduced by MD simulation does not disrupt the native conformation but leads to a non-negligible oscillation in structure. Differences in the oscillation of RMSd and TmScores indicate that most of the structural variation found along the trajectories is localized in loops, which is expected to be very mobile in aqueous solution, but the remaining regions of proteins are quite stable after 0.1 μs (see Figure 1). Very interestingly, the total number of residue−residue contacts is also very stable along the trajectory (Figure 2), but individual residue−residue contacts are much more labile. In fact, less than 50% of the "native" contacts appearing in the MD-averaged structure are present during the entire trajectory, while the rest are in fast interchange (Figure 2). Interestingly, while MD and experimental structures are very close (see Supporting Information, Figure S1), only 70−80% of the experimental native contacts are conserved more than 10% of the time in our trajectories. All these results suggest that some caution is necessary in the interpretation of the concept of native contact, since large portions of proteins behave like liquids or melted solids,[37] with many residues being quite promiscuous. On the contrary, the global structure seems quite insensitive to flexibility effects, suggesting that the overall structure descriptors, which can be easily obtained from low-resolution experiments, might be also quite robust to structural oscillation in water at room temperature. This is confirmed by inspection of Figures 3 and 4, which illustrate that all the low-resolution structural descriptors considered here — CCS, R_h, and SAXS curves — are very stable and show only moderate variations along the trajectory. All of these findings strongly suggest that these structural descriptors can be safely used as structural restraints to derive an average protein (or protein complex) structure.

Reliability of Structural Data Obtained in the Gas Phase. As noted above, a second topic of concern is whether or not structural data in the gas phase (like CCS) reflect the structural properties in solution. In order to analyze this point, we performed extended MD simulations of isolated proteins and of protein complexes in the gas phase, using simulation conditions similar to those of electrospray mass spectroscopy.[27] As previously found for other proteins,[27,38−40] vaporization preserves surprisingly well (at least in the submicrosecond time scale) the structure of proteins, at both the local

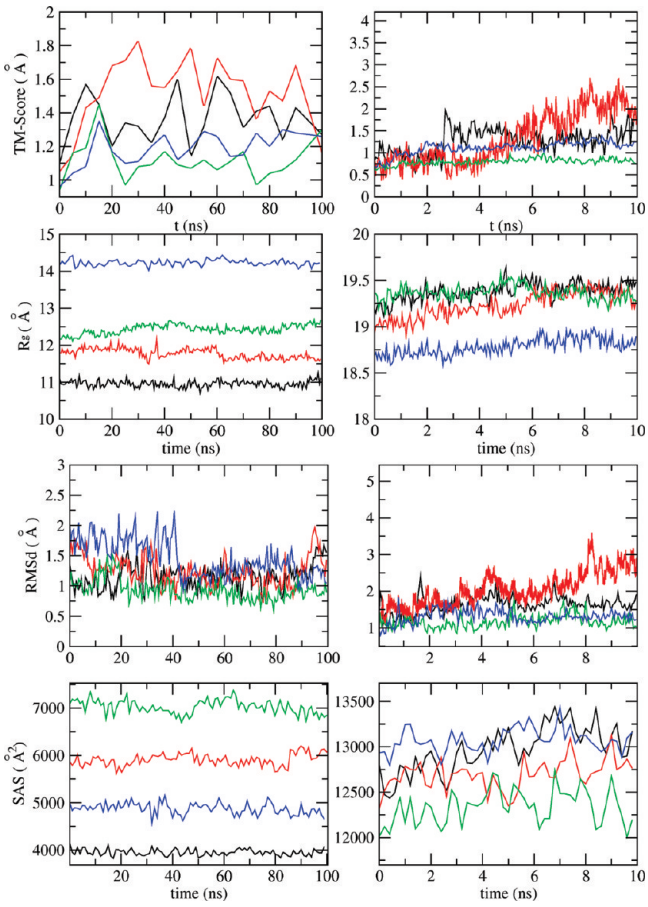

Figure 1. Variation of different structural descriptors along MD trajectories of some proteins and protein complexes. Top to bottom: TmScore, radii of gyration, root-mean-square derivation, and solvent accessible surface. Left: monomeric proteins (green: 1LYS, red: 1LS9, blue: 1UBQ, and black: 1BPI). Right: protein complexes (green: 1CSE, red: 1CBW, blue: 3TGI, and black: 1ACB).

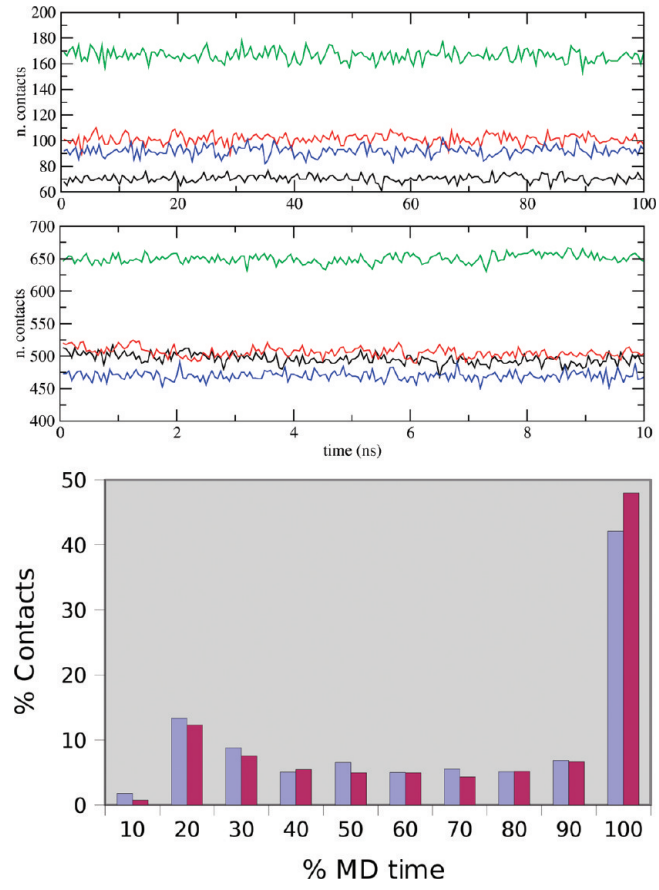

Figure 2. Oscillation in the total number of residue−residue contacts along the trajectories of proteins (top panel; color code in Figure 1) and protein complexes (bottom panel; color code in Figure 1). The histogram in the bottom corresponds to the distribution of residue contacts (blue protein monomers and red protein complexes) according to their persistence in trajectory (from 10 to 100%). In all cases, a residue is considered in contact when the Cα-Cα distance is lower than 7 Å in MD trajectories of proteins and protein complexes.

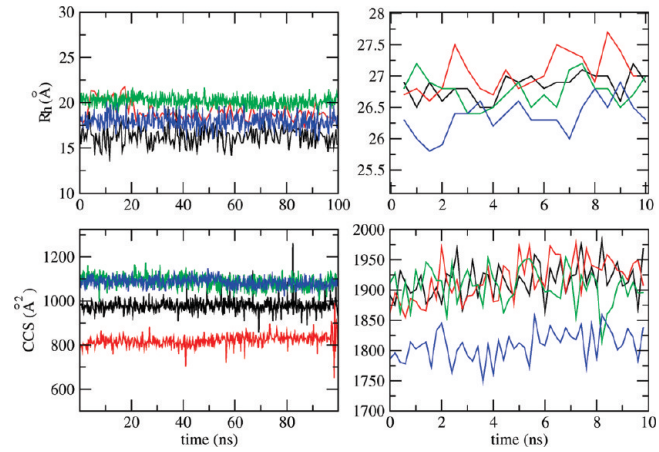

Figure 3. Oscillation of R_h (top plots) and CCS (bottom plots) versus time for some protein monomers (left) and protein-complexes (right). Color code as in Figure 1.

and global levels (see Figures 5 and 6 and Supporting Information, Figure S1). Particularly, global structural descriptors are always well maintained irrespective of the force field used for the simulations (Figure 6), suggesting that gas-phase data derived from electrospray experiments contain structural information that could be used to understand the structure of proteins in solution. This is confirmed when analyzing the CCS obtained from MD trajectories in the gas phase, which are only slightly smaller than those derived from trajectories in solution (in average 6%, see Figure 7). The agreement between the MD gas phase CCS and the CCS derived from the X-ray structure is good (6% overestimation), in fact better than obtained between X-ray CCS and MD-solution CCS (aqueous simulations overestimate by 12% the CCS expected for the protein in X-ray structure), suggesting that the crystal lattice is compressing slightly the protein with respect to the dilute aqueous conditions and confirming that in vacuo structural information can be used to obtain structural insights for the protein in solution. Two additional points are worth noting here: (i) the excellent agreement found between the available experimental (gas phase) and MD-computed values (Figure 7), which reinforces the confidence in our simulations and (ii) the smaller standard deviation in the structural descriptors associated to gas phase simulations, which confirms the idea that in the gas phase the structure of the protein is rigidified, reducing the impact of flexibility (Figure 7).

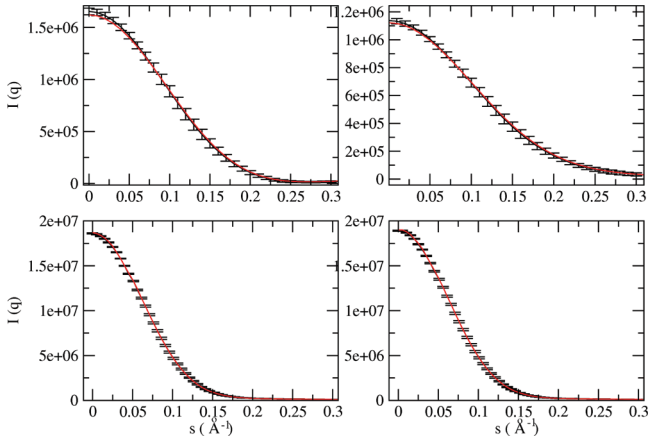

Figure 4. Superposition of SAXS curves obtained for protein structures sampled during MD trajectories of two proteins (left) and two protein complexes (right). The line in red corresponds to the curve simulated for a single average structure, while the line in black (with standard deviations) corresponds to the averaged spectra. Note that in the region richer in information (*s* around 0.1), the black and red lines are equivalent.

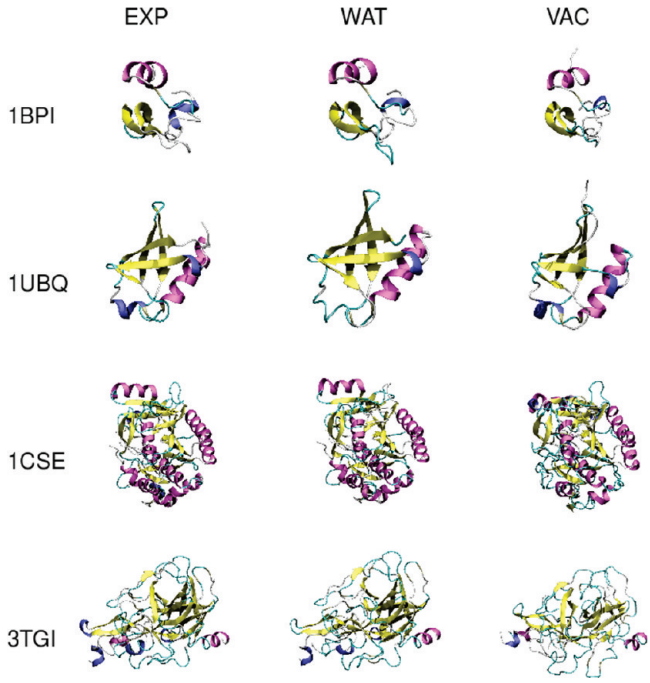

Figure 5. Representation of the experimental and MD-averaged structures of selected monomeric proteins (1BPI and 1UBQ) and protein complexes (1CSE and 3TGI). In all cases, the experimental structure, the MD-averaged one in aqueous solution, and the MD-averaged structure obtained after extended simulation in the gas phase are displayed.

In order to check whether or not results were too dependent on the total charge assigned to the protein, we collected additional trajectories for ubiqutin, cytochrome C and lysozime using a range of charge states around the expected optimum one (see above). Results, reported in the Supporting Information, Figure S2, strongly suggest that within the region of interest (i.e., close to the expected charge state), the protein structure is quite insensitive to changes in total charge, in good agreement with the experimental data[41] (see Supporting Information, Figure S2). For one of the proteins

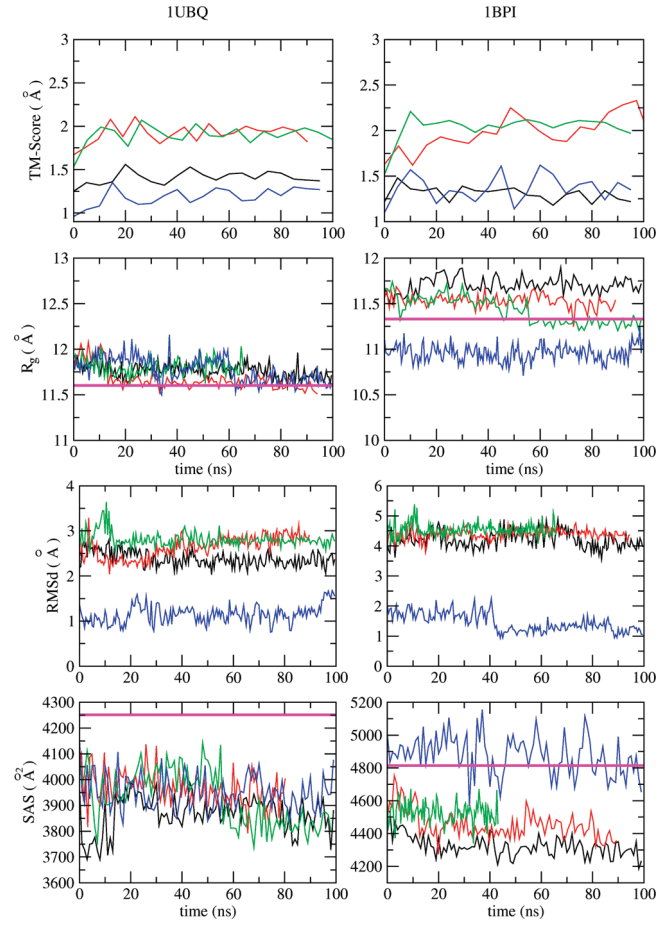

Figure 6. Global structural descriptors (TmScore, gyration radii, RMSd, and SAS) for proteins in the gas phase for 1BPI and 1UBQ obtained from MD simulations in the gas phase with CHARM22 (green), OPLSAA (red), and AMBER-parm03 (black) force fields. The reference results for the simulation in solution are displayed in blue, and the values derived from the experimental structure (crystal) are shown as straight lines in magenta.

(ubiquitin), for which a very large positive charge was considered, protein unfolding was observed (see Supporting Information, Figure S2), in agreement with results reported by other groups.[41d] The fact that, in these extreme conditions, the MD-simulated CCS was smaller than that of the experimentally found[41a] is not unexpected, and simply reflects that the trajectory was too short to reproduce a complete unfolding, which, according to experimental measures, is expected to happen in the millisecond time scale.[41d,25b] In any case, it seems that, unless extreme charge conditions are considered, the gas phase CCS remains close to those expected for solution structures.

MD simulations suggest that protein complexes are stable in the gas phase for long periods of time, and that the sampled gas-phase structures are not far away from the standard sampled structures in aqueous solution (see Figure 5). This is clearly shown in different metrics displayed in Figure 8, which demonstrate that the conclusions derived for proteins are also valid for noncovalent protein complexes and that despite non-negligible local distortions the overall shapes of the protein complexes are well preserved in the gas phase, at least in the submicrosecond time scale. Not

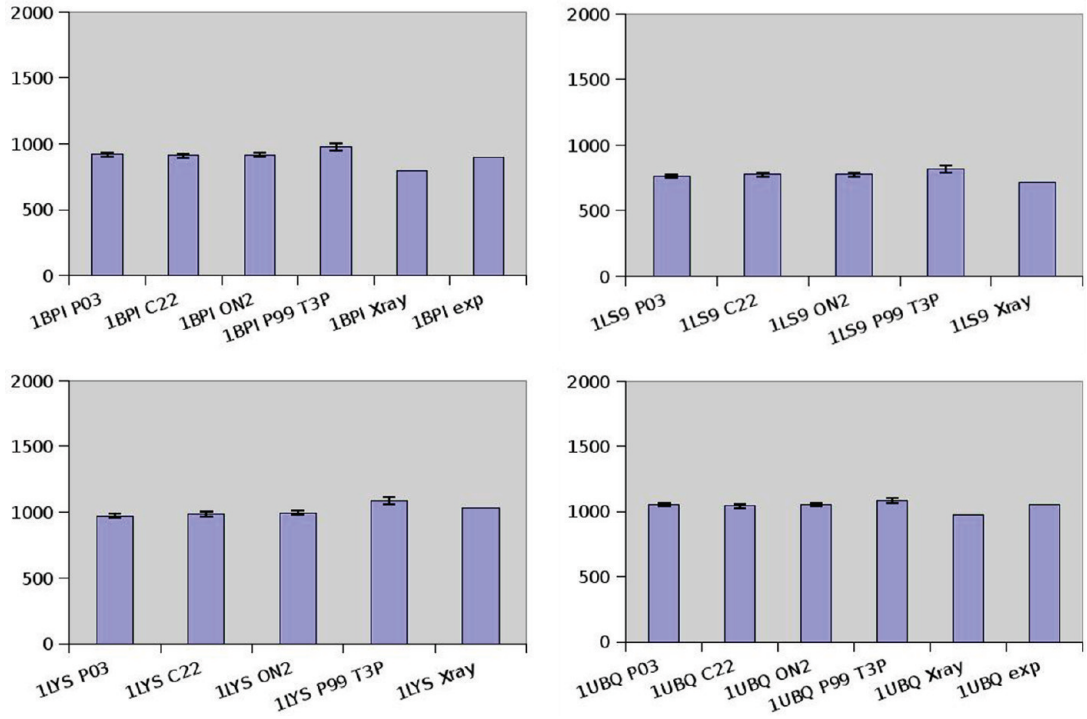

Figure 7. CCS (in Å^2) for four small proteins, as determined from MD simulations in the gas phase considering different force-fields. Results, obtained from MD simulations in solution, of crystal structure and, when available, experimental ESI measures are displayed for comparisons. Standard deviations associated to the different MD averages are shown.

surprisingly then, there is a close correspondence between the CCS obtained from MD simulations in the gas phase and those obtained in aqueous trajectories (Figure 9). In fact, the agreement with X-ray values is better for gas-phase simulations (3% underestimation) than for solution trajectories (5% overestimation), mimicking the results found for isolated proteins and showing that, while proteins in pure water expand with respect to crystal conditions, they compress when moving to the gas phase. In summary, MD simulations strongly suggest that gas phase experiments like MS-ESI produce valuable data on the structure of protein and protein complexes in solution.

Information Load in Low-Resolution Structural Data. It is clear that the amount of information in the low-resolution data considered here is modest, and that these techniques alone are unable to unambiguously determine the three-dimensional structure of proteins or protein complexes. Thus, the R_h gives only information on the shape of a molecule perpendicular to an external field, is quite ambiguous in terms of structure definition, and is prone to artifacts for proteins with tails. The CCS is a magnitude related to the molecular surface in conditions far from the physiological ones, and the information derived is then unable to differentiate between different conformations displaying similar molecular surface. Finally, the SAXS spectrum contains, in principle, all possible structural information on the protein, but interpretation of the SAXS spectrum is difficult, and ambiguous assignments are often derived. However, despite all their limitations, these methods are quite powerful to discard erroneous solutions that can be obtained from modeling techniques. This is shown when scanning the CASP7 deposited model for monomeric proteins. Thus, if we randomly select five models (for each target) from those deposited in the CASP7 database, we have a random probability around 50% of choosing one of the good solutions (we considered good solutions those with a (GDT -5) $>$ GDT$_{\text{best solution}}$; GDT being a score function defined by CASP7 evaluators).[42] This probability sharply increases to 70% (CCS and R_h) and 80% (SAXS) when low-resolution structural data are used to clean up unrealistic models, see Figure 10. However, the largest enrichment is obtained when low-resolution structural data, which provide global information on the protein structure, are combined with statistical potentials like ProSA, which detect local errors undetectable in low-resolution data. Thus, when combining ProSA with CCS and R_h restrains, the chances of finding the good model among five randomly chosen increases to 86% and to more than 98%, if ProSA is combined with SAXS spectra (see Figure 10). In summary, low-resolution data, especially SAXS spectra, combined with standard statistical potentials dramatically increases the possibilities to find good structural models when using standard protein-modeling tools.

The same analysis performed for protein complexes, using now the CAPRI database, provides qualitatively similar results. Thus, we have a random probability around 30% of finding an acceptable prediction (as defined by CAPRI evaluators),[43] when only five random models are selected. The chances increase to 45 (R_h), 55 (CCS), and 75% (SAXS) when low-resolution structural data are used as restrains (Figure 11). Again the chances increase when an empirical potential, such as pyDock, is used to detect local errors and an experimental low-resolution data are used to detect global structural errors. Thus, when pyDock is combined with CCS, the chances of finding an acceptable structure in five

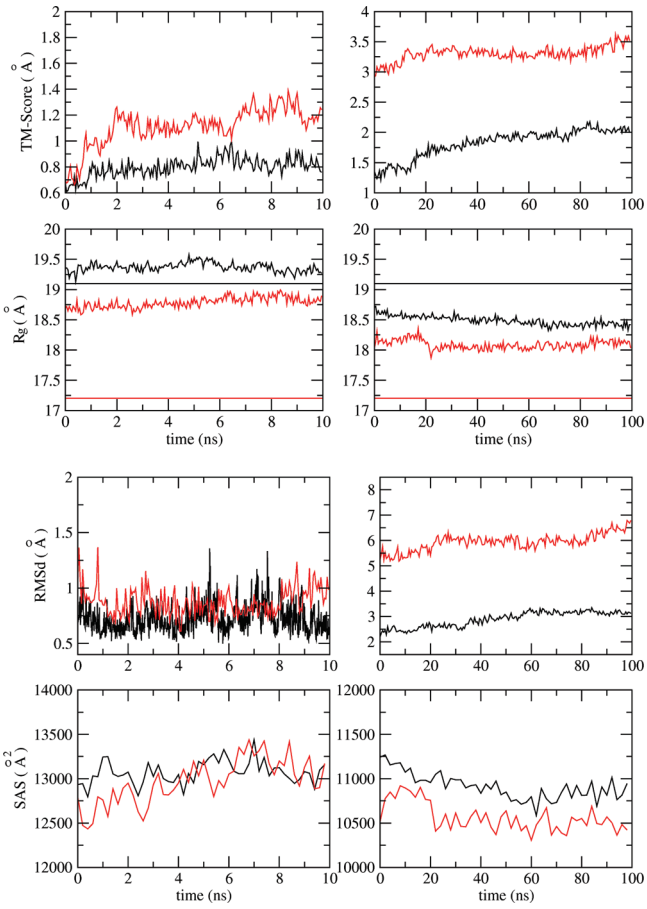

Figure 8. Global structural descriptors (TmScore, gyration radii, RMSd, and SAS) for protein complexes (1CSE and 3TGI) obtained from MD simulations in the gas phase (right panels). The reference results for the simulation in solution are displayed on the left panels, and the values derived from the experimental structure (crystal) are shown as straight lines.

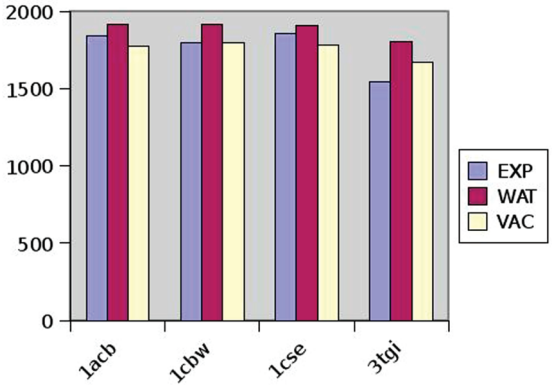

Figure 9. CCS (in Å^2) of different protein complexes as determined from MD simulations in the gas phase, MD simulations in solution, and crystal structures.

randomly selected ones increase to more than 70%, the chances increasing to 94% if PyDock is combined with SAXS spectra. In summary, even for the very difficult case of protein complex prediction, low-resolution structural data, especially when combined with statistical potentials, dramatically increases the chances of selecting good structural

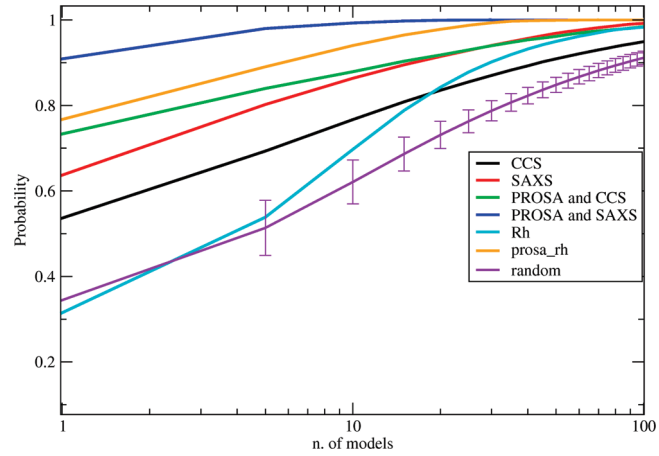

Figure 10. Enrichment curves for the prediction of the best structural model of monomeric proteins obtained using data from the CASP7 experiment.

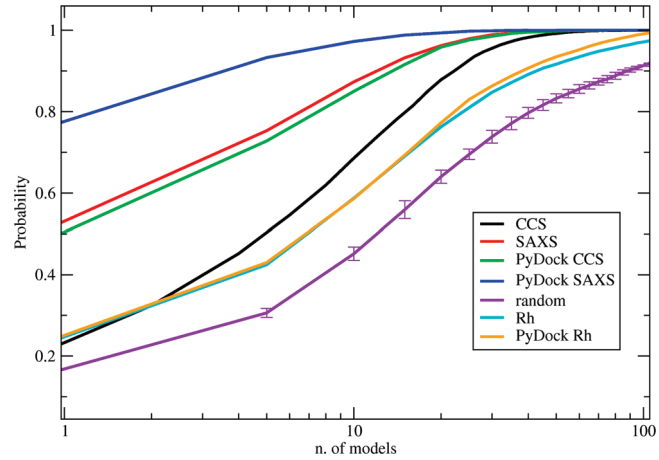

Figure 11. Enrichment curves for the prediction of the best structural model of protein complexes obtained by using data from the CAPRI experiment.

models from the ensemble of solutions provided by standard modeling techniques (see Figure 11).

Conclusions

Low-resolution structural data are very robust and quite insensitive to oscillations due to the intrinsic flexibility of proteins. Global descriptors of proteins, like the collision cross-section, are quite robust to dramatic changes in the environment, allowing then the use of structural information derived from very hostile conditions, such as the gas phase, to gain insight into the structure of the proteins (and protein complexes) in aqueous solutions. Importantly, despite the reduced information content existing in hydrodynamic radii, collision cross sections, or SAXS curves, these magnitudes can be used to detect errors in theoretical structural models that were accepted by scoring procedures in leading modeling tools. The global result is an overall improvement in the quality of the final suggested models of proteins and of protein−protein complexes. The improvement is especially important when the low-resolution data are combined with empirical potentials, such as ProSA or pyDock. Finally, the analysis of CASP7 and CAPRI experiments demonstrates that SAXS is the richest source of structural

information of the three considered here, while the hydrodynamic radius (even useful) provides a more reduced amount of information. Overall, our results demonstrate that low-resolution structural data, in general, is easy to obtain and can be efficiently used to increase the chances of success in the prediction of the three-dimensional structure of proteins and protein complexes, even when obtained in conditions quite far from the physiological ones.

Acknowledgment. This work has been supported by the Spanish Ministry of Education and Science (BIO2006-01602, CONSOLIDER Project in Escience, and BIO2008-02882), the Spanish Ministry of Health (COMBIOMED network), the Fundación Marcelino Botín, and the National Institute of Bioinformatics. All calculations were performed in the MareNostrum supercomputer at the Barcelona Supercomputing Center.

Supporting Information Available: Figures show the stability of secondary structures and the comparison between experimental and simulated CCS values. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Marti-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sanchez, R.; Melo, F.; Sali, A. *Annu. Rev. Biophys. Bio.* **2000**, *29*, 291.

(2) Simons, K. T.; Strauss, C.; Baker, D. *J. Mol. Biol.* **2001**, *306*, 1191.

(3) Kryshtafovych, A.; Fidelis, K.; Moult, J. *Proteins* **2007**, *69*, 194.

(4) Lensink, M. F.; Mendez, R.; Wodak, S. J. *Proteins* **2007**, *69*, 704.

(5) Karplus, M.; Kuriyan, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6679.

(6) Pazos, F.; Helmer-Citterich, M.; Ausiello, G.; Valencia, A. *J. Mol. Biol.* **1997**, *271*, 511.

(7) Lichtarge, O.; Yamamoto, K. R.; Cohen, F. E. *J. Mol. Biol.* **1997**, *274*, 325.

(8) Chelliah, V.; Chen, L.; Blundell, T. L.; Lovell, S. C. *J. Mol. Biol.* **2004**, *342*, 1487.

(9) Aloy, P.; Querol, E.; Aviles, F. X.; Sternberg, M. J. *J. Mol. Biol.* **2001**, *311*, 395.

(10) Chelliah, V.; Blundell, T. L.; Fernandez-Recio, J. *J. Mol. Biol.* **2006**, *357*, 1669.

(11) Juan, D.; Pazos, F.; Valencia, A. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 934.

(12) Matsumoto, R.; Sali, A.; Ghildyal, N.; Karplus, M.; Stevens, R. L. *J. Biol. Chem.* **1995**, *270*, 19524.

(13) Xu, L. Z.; Sanchez, R.; Sali, A.; Heintz, N. *J. Biol. Chem.* **1996**, *271*, 24711.

(14) Andrea, S. *J. Mass. Spectrom.* **2003**, *38*, 1225.

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

(16) Petoukhov, M. V.; Svergun, D. I. *Curr. Opin. Struct. Biol.* **2007**, *17*, 562.

(17) Putnam, C. D.; Hammel, M.; Hura, G. L.; Tainer, J. A. *Q. Rev. Biophys.* **2007**, *40*, 191.

(18) Koch, M. H.; Vachette, P.; Svergun, D. I. *Q. Rev. Biophys.* **2003**, *36*, 147.

(19) Tanford, C. In *Physical Chemistry of Macromolecules*; J. Wiley and Sons: New York, 1961.

(20) Garcia de la Torre, J.; Huertas, M. L.; Carrasco, B. *J. Magn. Reson.* **2000**, *147*, 138.

(21) Bernado, P.; Garcia de la Torre, J.; Pons, M. *J. Biomol. NMR* **2002**, *23*, 139.

(22) Svergun, D.; Barberato, C.; Koch, M. H. J. *J. Appl. Crystallogr.* **1995**, *28*, 768.

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

(24) Shelimov, K. B.; Clemmer, D. E.; Hudgins, R. R.; Jarrold, M. F. *J. Am. Chem. Soc.* **1997**, *119*, 2240.

(25) (a) Jarrold, M. F. *Annu. Rev. Phys. Chem.* **2000**, *51*, 179. (b) Breuker, K.; McLafferty, F. W. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 18145.

(26) Kaltashov, I. A.; Mohimen, A. *Anal. Chem.* **2005**, *77*, 5370.

(27) Meyer, T.; de la Cruz, X.; Orozco, M. *Structure* **2009**, *17*, 88.

(28) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, T.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999.

(29) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474.

(30) MacKerell, A. D.; Bashford, D.; Bellott; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586.

(31) Scarff, C. A.; Thalassinos, K.; Hilton, G. R.; Scrivens, J. H. *Rapid Commun. Mass. Spectrom.* **2008**, *22*, 3297.

(32) William, L. J.; Jayaraman, C.; Jeffry, D. M.; Roger, W. I.; Michael, L. K. *J. Chem. Phys.* **1983**, *79*, 926.

(33) Darden, T.; York, D.; Pederson, L. *J. Chem. Phys.* **1993**, *98*, 10089.

(34) Sippl, M. J. *Proteins* **1993**, *17*, 355.

(35) Cheng, T. M.; Blundell, T. L.; Fernandez-Recio, J. *Proteins* **2007**, *68*, 503.

(36) Canutescu, A. A.; Shelenkov, A. A.; Dunbrack, R. L., Jr. *Protein Sci.* **2003**, *12*, 2001.

(37) Rueda, M.; Ferrer-Costa, C.; Meyer, T.; Perez, A.; Camps, J.; Hospital, A.; Gelpi, J. L.; Orozco, M. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 796.

(38) Bothner, B.; Siuzdak, G. *ChemBioChem* **2004**, *5*, 258.

(39) Jarrold, M. F. *Annu. Rev. Phys. Chem.* **2000**, *51*, 179.

(40) Patriksson, A.; Marklund, E.; van der Spoel, D. *Biochemistry* **2007**, *46*, 933.

(41) (a) Valentine, S. J.; Counterman, A. E.; Clemmer, D. E. *J. Am. Soc. Mass Spectrom.* **1997**, *8*, 954. (b) Shelimov, K. B.; Clemmer, D. E.; Hudgins, R. R.; Jarrold, M. F. *J. Am. Chem. Soc.* **1997**, *119*, 2240. (c) Valentine, S. J.; Anderson, S. J.; Ellington, A. E.; Clemmer, D. E. *J. Phys. Chem. B* **1997**, *101*, 3891. (d) Segev, E.; Wyttenbach, T.; Bowers, M. T.; Gerber, R. B. *Phys. Chem. Chem. Phys.* **2008**, *10*, 3077.

(42) CASP7; http://predictioncenter.org/casp7. Accessed Sept 4, 2009.

(43) CAPRI; http://www.ebi.ac.uk/msd-srv/capri. Accessed Sept 8, 2009.

CT900305M