Supplementary Material

Enhanced Conformational Space Sampling Improves the Prediction of Chemical Shifts in Proteins

P.R.L. Markwick, C.F. Cervantes, B.L. Abel, E.A. Komives, M. Blackledge, and J.A. McCammon

Full References:

2(c) Shen, Y.; Lange, O.; Delaglio, F.; Rossi, P.; Aramini, J.M.; Liu, G.; Eletsky, A.; Wu, Y.; Singarapu, K.K.; Lemak, A.; Ignatchenko, A.; Arrowsmith, C.H.; Szyperski, T.; Montelione, G.T.; Baker, D.; Bax, A. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 4685-4690.

16) Case, D.A.; Darden, T.A.; Cheatham, T.E.; Simmerling, C.L.; Wang, J.; Duke, R.E.; Luo, R.; Crowley, M.; Ross, W.C.; Zhang, W.; Merz, K.M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossvary, I.; Wong, K.F.; Paesani, F.; Vanicek, J.; Wu, X,; Brozell, S.R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D.H.; Seetin, M.G.; Sagui, C.; Babin, V.; Kollman, P.A. **2008**, AMBER10, University of California, San Francisco.

Methods

Accelerated Molecular Dynamics:

The details of the accelerated molecular dynamics method have been discussed previously in the literature.[1,2] The essential idea behind accelerated molecular dynamics is to define a reference, or 'boost energy', E_b, which is fixed above the minimum of the potential energy surface. At each step in the AMD simulation, if the potential energy of the system lies below this boost energy, a continuous, non-negative bias is added to the actual potential. If the potential energy is greater than the boost energy, it remains unaltered. This results in a raising and flattening of the potential energy landscape, decreasing the magnitude of the energy barriers between low energy states, and therefore enhancing the escape rate from one low energy conformational state to another, whilst maintaining the essential details of the underlying potential energy surface. The extent to which the potential energy surface is modified depends on the difference between the boost energy and the actual potential. Explicitly, the modified potential, V*(r), is defined as:

$$V^*(\vec{r}) = V(\vec{r})$$

if the potential energy, V(**r**), is equal to or greater than the boost energy, and

$$V^*(\vec{r}) = V(\vec{r}) + \Delta V(\vec{r})$$

if the potential energy is less than the boost energy.

The energy modification, or bias, is given by:

$$\Delta V(\vec{r}) = \frac{(E_b - V(\vec{r}))^2}{\alpha + (E_b - V(\vec{r}))}$$

The extent of acceleration (ie. how aggressively we enhance the conformational space sampling) is determined by the choice of the boost energy and the acceleration parameter, α. Conformational space sampling can be enhanced by either increasing the boost energy, or decreasing the acceleration parameter. In the present work, the extent of conformational space sampling was controlled by systematically increasing the boost energy using a fixed acceleration parameter. During the course of the simulation, if the potential energy is modified, the forces on the atoms are recalculated for the modified potential. The use of the bias potential defined above ensures that the derivative of the modified potential will not be discontinuous at points where $V(\mathbf{r})=E_b$. It should be noted that if the boost energy is set too large for a given acceleration parameter, the modified potential energy surface becomes iso-energetic, resulting in a random walk through phase space. In the present work, we made considerable efforts to avoid this 'over-acceleration' regime: Comparison of the order parameters at different acceleration levels reveals that even at the most aggressive acceleration level employed in this work, the observed enhanced conformational space sampling is restricted to well defined regions of the protein, an observation that is not concurrent with a random walk. Furthermore, we performed additional AMD simulations using a slightly larger acceleration parameter, and observed similar conformational space sampling.

One of the favorable characteristics of AMD is that it yields a canonical average of an observable, so that thermodynamic and other equilibrium properties of the system can be accurately determined. The corrected canonical ensemble average of the system is obtained by re-weighting each point in the configuration space on the modified potential by the strength of the Boltzmann factor of the bias energy, $exp[\beta\Delta V(\mathbf{r})]$ at that particular point. Whilst AMD represents a robust free energy sampling method, obtaining an accurate estimate of the time-scale of the observed conformational space sampling is extremely difficult: The application of the bias potential has a strong effect on the transmission coefficient. The only way to avoid this problem would be to set the boost energy below the entire transition state region, thereby preserving the transition-state theory formalism. However, considering that proteins possess a highly rugged and structured potential energy landscape, setting the boost energy below the entire transition state region on the protein PES, would not allow acceleration over the larger energy barriers that separate the conformational sub-states.

Simulation Details:

The simulation protocol described here has been previously presented elsewhere.[3] The X-ray crystal structure of the protein IκBα(67-206) (PDB code 1NFI) was placed in a periodically repeating box with 8,000 water molecules and 8 Na+ counter-ions. Initially a set of five standard classical MD simulations were performed. In each case, the system was brought to thermodynamic equilibrium at 300K, 1 bar pressure using a Langevin thermostat with a collision frequency of 3-ps^{-1} and a Berendsen weak-coupling pressure-stat. For each of the five simulations a different random seed generator was employed and, after equilibration, a 10-ns production MD simulation was performed under periodic boundary conditions with a time-step of 1-fs. Electrostatic interactions were treated using the Particle Mesh Ewald (PME) method[4] with a direct space sum limit of 10-A. The ff99SB force-field[5] was used for the solute residues and the TIP4P water force-field was employed for the solvent molecules. All simulations were performed using the AMBER10 code.[6] These initial five 10-ns MD simulations provided the starting point for the biased potential accelerated molecular dynamics simulations discussed below, provide an estimate of the average dihedral angle energy V(dih) and were also used as a control set in order to assess the improvement in the calculation of the NMR observables on enhancing the conformational space sampling.

For IκBα, a "dual boost" AMD methodology[7] was employed whereby, in addition to the acceleration that is applied across the torsional terms of the force-field, a fixed background acceleration was also applied across the entire potential. This background potential is weak with acceleration parameters: $\alpha(tot)$=(0.2* total number of atoms in system in kcal/mol) and $E_b(tot) - V(tot)=\alpha(tot)$. In total, twenty dual-boost accelerated molecular dynamics simulations were performed at five different torsional acceleration levels. The torsional acceleration parameter, $\alpha(dih)$, in each case was fixed at 120-kcal/mol and the acceleration level was controlled by varying the boost energy, $E_b(dih)$. The boost energy for the five acceleration levels was set at 240, 360, 480, 600 and 720-kcal/mol above the dihedral angle energy (estimated from the average dihedral angle energy from the unbiased 10-ns MD simulations). Each AMD simulation was performed for 10,000,000 steps (the equivalent of a 10-ns standard MD simulation) and the physical conditions, force-fields and all other simulation parameters employed were identical to those described above for the five standard 10-ns MD control set. The atomic coordinates and necessary energetic data (such as the magnitude of the bias potential) were saved across each trajectory every 5 steps for analysis. All AMD simulations were performed using an in-house modified version of the AMBER10 code.

For each AMD simulation, the corrected canonical ensemble was determined by performing a free-energy weighting protocol: The strict Boltzmann re-weighting criterion described above in the discussion of the AMD methodology was relaxed, as it was found that this approach over-estimated the population of the low-energy states, a problem that has been previously discussed in the literature.[8] Instead, an initial free-energy "pre-pruning" for each AMD ensemble was

performed in which the high energy structures (some 80% of the trajectory) were stripped out and the remaining 20% was used to perform a clustering analysis. The clustering protocol employed was based on QR-factorization and principal components analysis, and a series of short (5.5-ns) classical MD simulations were seeded from the resulting cluster. The initial 0.5-ns was discarded and a MM/PBSA analysis[9] was performed on the seeded MD simulations in order to obtain the relative free energies. In this way, the AMD simulations were primarily employed to obtain enhanced conformational space sampling, while the free-energy statistics are provided by the MM/PBSA analysis. Using the seeded MD simulation trajectories and the associated MM/PBSA free-energy statistics, multiple free-energy weighted molecular ensembles were constructed at each acceleration level.

SHIFTX[10] was employed to obtain chemical shifts for all $^{1}H^{N}$, ^{15}N, $^{13}C\alpha$, $^{13}C\beta$ and $^{13}C'$ nuclei. Chemical shifts were calculated for each member of the free-energy weighted ensemble and then averaged. The resulting predicted chemical shift data was then averaged over all ensembles at a given acceleration level. In this way, we obtain both a time- and ensemble-averaged representation of the chemical shift data. As discussed below, the chemical shift averaging procedure that we have employed is important and so we find it appropriate to describe the averaging procedure in more detail:. At each acceleration level, we have performed 20 AMD simulations. For each of these AMD simulations, MM/PBSA-based free energy weighted molecular ensembles were constructed and the chemical shifts were calculated for each member of the ensemble and linearly averaged. The same weight was applied to each structure as the free energy weighting procedure has already been performed whilst generating the ensemble. This initial averaging procedure accounts for the temporal average. In order to obtain both a time- and ensemble-averaged prediction of the chemical shifts, the resulting time-averaged chemical shifts from each ensemble were then averaged over all free-energy weighted molecular ensembles generated at the same acceleration level. This averaging procedure is also linear: The 20 AMD simulations performed at each acceleration level do not sample exactly the same conformational space, and yet each of these twenty trajectories is treated as being equally valid, as they are all generated under exactly the same conditions.

The accuracy of the predicted chemical shifts was assessed by calculating the root-mean-square difference (RMSD) to the corresponding experimental data. The RMSDs for each nucleus ($^{1}H^{N}$, ^{15}N, $^{13}C\alpha$, $^{13}C\beta$ and $^{13}C'$) were then summed to give a cumulative RMSD, a single value that was used to identify the optimal conformational space sampling associated with the optimal (torsional) acceleration level, which was found to be E_b(dih)-V(dih) = 600 kcal/mol with α(dih)= 120 kcal/mol. As we mention in the paper, this optimal torsional acceleration level also produced the best representation of the experimental RDC data.[3]

The chemical shift averaging procedure that we have employed is important: As we have stated in the paper, no individual molecular ensemble obtained at the optimal acceleration level gives a chemical shift prediction that is as good as the trajectory averaged result, an observation that is

not only true for the calculation of chemical shift data, but also for other NMR observables such as RDCs and scalar J-couplings.[11] The averaging procedure represents an attempt to over-come the severe statistical-mechanical sampling problems which are always present when trying to calculate time- and ensemble-averaged NMR observables: Each molecular ensemble generated at a given acceleration level represents a single brute-force simulation over an extended (but unknown) time-period. Chemical shift averaging over a single free-energy weighted ensemble provides a representation of the temporal average. However, the twenty AMD simulations performed at a given acceleration level do not sample exactly the same conformational space. Based on the seminal work of Frauenfelder,[12] our understanding of long-scale dynamics in proteins such as IκBα is that the system sits in a well-defined energy minimum and on slow time-scales undergoes stochastically mediated dynamic excursions to higher energy, low-populated conformational states. The AMD approach increases the frequency of these dynamic excursions. After free energy weighting of each trajectory, we find that the low energy regions of conformational space are well-conserved in each individual AMD trajectory, whilst the higher energy, low populated regions of conformational space are differentially sampled. Therefore, the second averaging procedure (averaging the predicted chemical shift data obtained from the different molecular ensembles constructed at each acceleration level) provides both a time- and ensemble average representation of the chemical shift data.

Calculation of Order Parameters:

Figure 2 (upper panel) in the paper compares experimental NH spin relaxation order parameters to those obtained from the molecular ensembles generated at the optimal acceleration level for the reproduction of the chemical shift data. The theoretical order parameters, S^2, were calculated within the framework of the Lipari-Szabo formalism.[13] The structures for each ensemble were superposed by mass-weighted backbone root-mean-square fitting to the average structure (residues 75-201), and the order parameters were calculated as:[14]

$$S^2 = \frac{1}{2}\left[3\sum_{i=1}^{3}\sum_{j=1}^{3}\left\langle \mu_i \mu_j \right\rangle^2 - 1 \right]$$

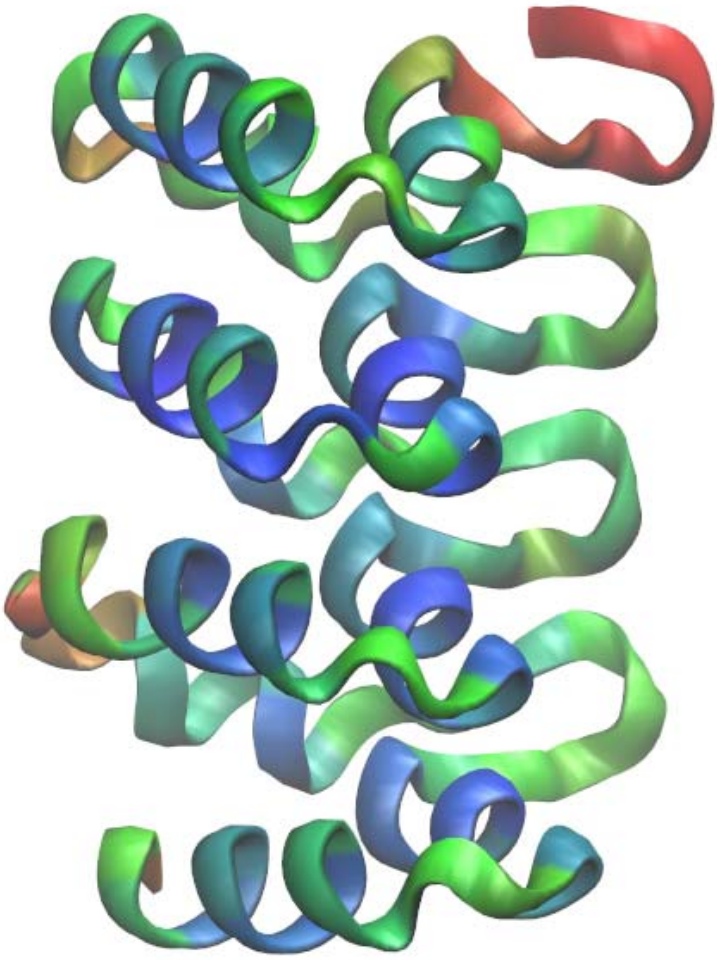
where μ(i) are the Cartesian coordinates of the normalized inter-nuclear vector of interest (N-H) and i,j=x,y,z. The order parameters presented in Figure 2 are trajectory averaged. The NH order parameters calculated from the molecular ensembles generated at the optimal acceleration level for the reproduction of the chemical shift data are shown in Figure S1 using a temperature spectrum varying from 0.0 (red) to 1.0 (blue) superposed on the X-ray crystal structure.

Software:

All AMD and standard MD simulations were performed using an in-house modified version of the AMBER10 simulation suite.[6] Mass-weighted root mean square fitting to the average structure (residues 75-201) was performed using the 'ptraj' subroutines in AMBER10 and the

order parameters were calculated using an in-house code as described above. The chemical shifts were calculated using the SHIFTX chemical shift prediction algorithm.[10]

Figure S1: NH Order parameters calculated from molecular ensembles generated at the optimal acceleration level for the reproduction of the chemical shift data.



Chemical Shift Results for IκBα:

Figure S2A: Comparison of Experimental and Predicted Cα Chemical Shifts in IκBα. Predicted chemical shifts from the X-ray crystal structure are shown as black circles. Predicted chemical shifts from the trajectory averaged RDC-optimal AMD ensembles are shown as red circles.

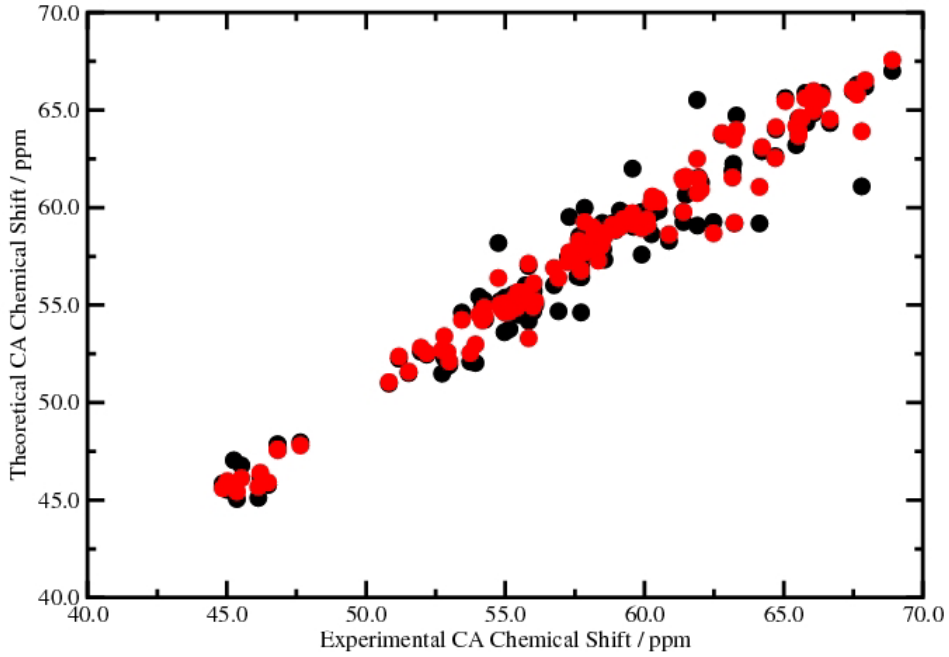


Figure S2B: Predicted Cα Chemical Shift RMSDs for each residue in IκBα. The results for the X-ray crystal structure (black line) are compared to those obtained from the trajectory averaged RDC-optimal AMD ensembles (red line).

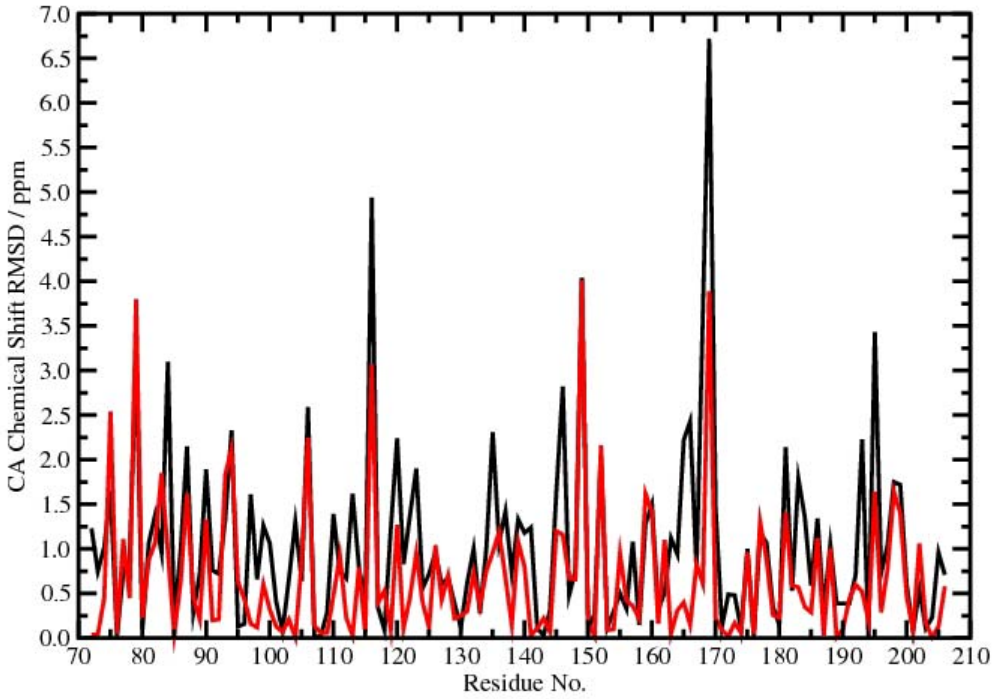


Figure S3A: Comparison of Experimental and Predicted Cβ Chemical Shifts in IκBα. Predicted chemical shifts from the X-ray crystal structure are shown as black circles. Predicted chemical shifts from the trajectory averaged RDC-optimal AMD ensembles are shown as red circles.

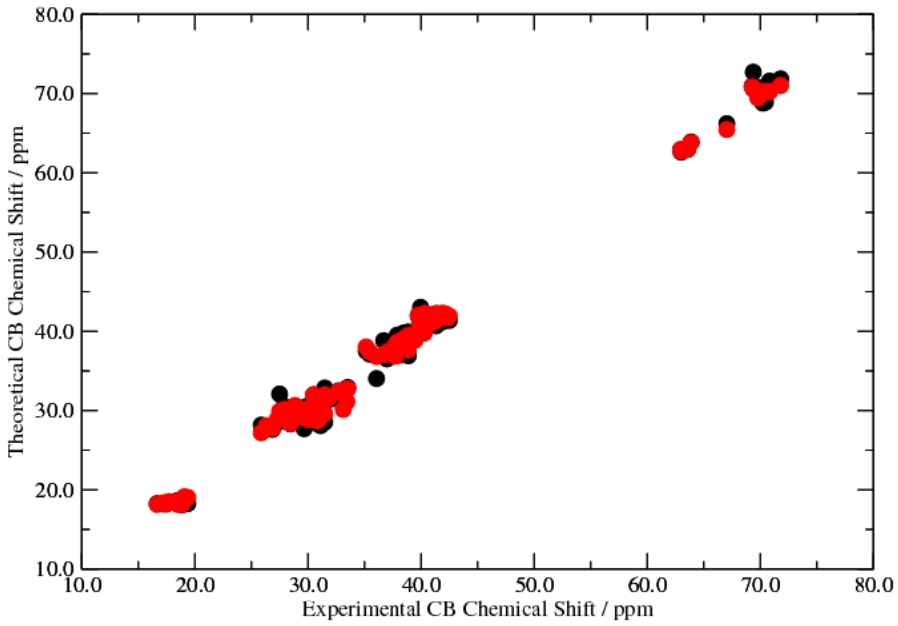


Figure S3B: Predicted Cβ Chemical Shift RMSDs for each residue in IκBα. The results for the X-ray crystal structure (black line) are compared to those obtained from the trajectory averaged RDC-optimal AMD ensembles (red line).

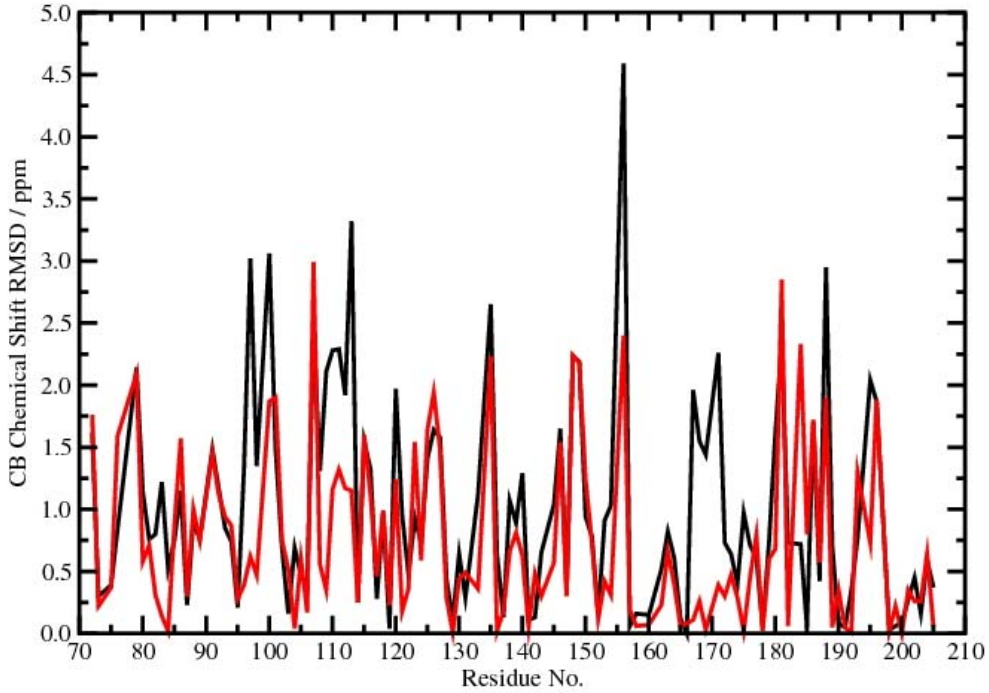


Figure S4A: Comparison of Experimental and Predicted C' Chemical Shifts in IκBα. Predicted chemical shifts from the X-ray crystal structure are shown as black circles. Predicted chemical shifts from the trajectory averaged RDC-optimal AMD ensembles are shown as red circles.

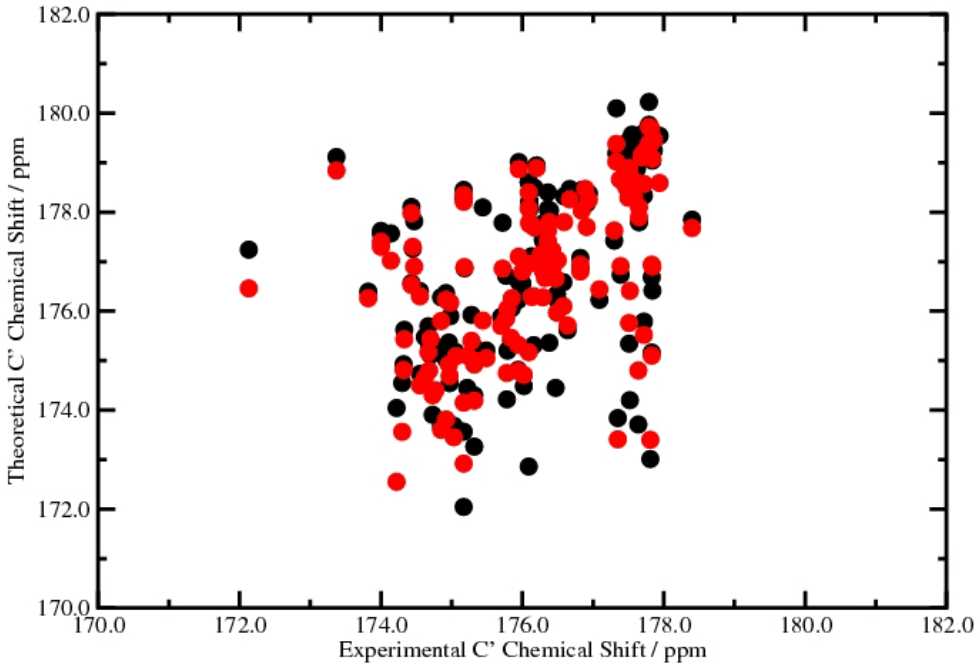


Figure S4B: Predicted C' Chemical Shift RMSDs for each residue in IκBα. The results for the X-ray crystal structure (black line) are compared to those obtained from the trajectory averaged RDC-optimal AMD ensembles (red line).

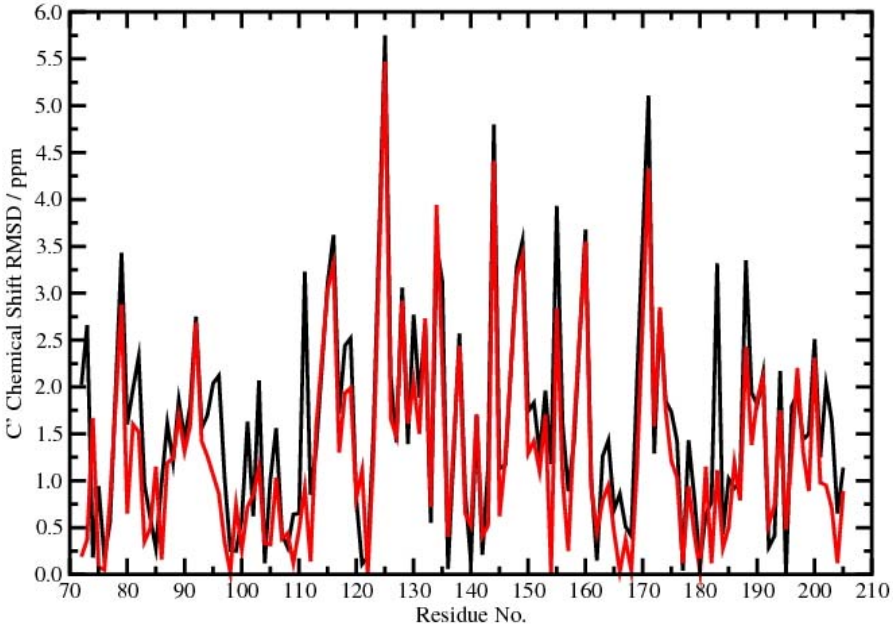


Figure S5A: Comparison of Experimental and Predicted H^N Chemical Shifts in IκBα. Predicted chemical shifts from the X-ray crystal structure are shown as black circles. Predicted chemical shifts from the trajectory averaged RDC-optimal AMD ensembles are shown as red circles.

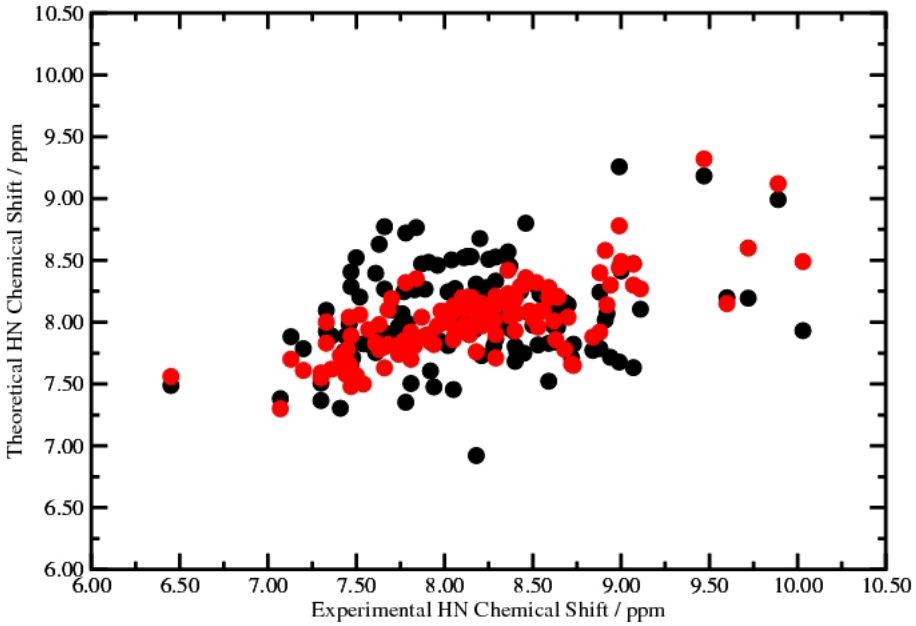


Figure S5B: Predicted H^N Chemical Shift RMSDs for each residue in IκBα. The results for the X-ray crystal structure (black line) are compared to those obtained from the trajectory averaged RDC-optimal AMD ensembles (red line).

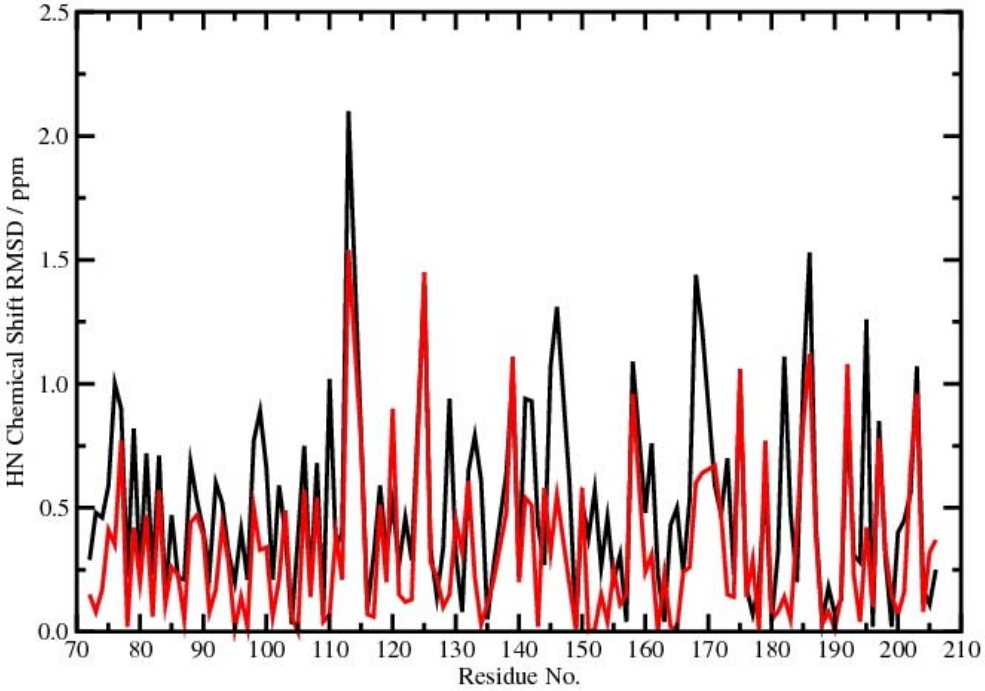
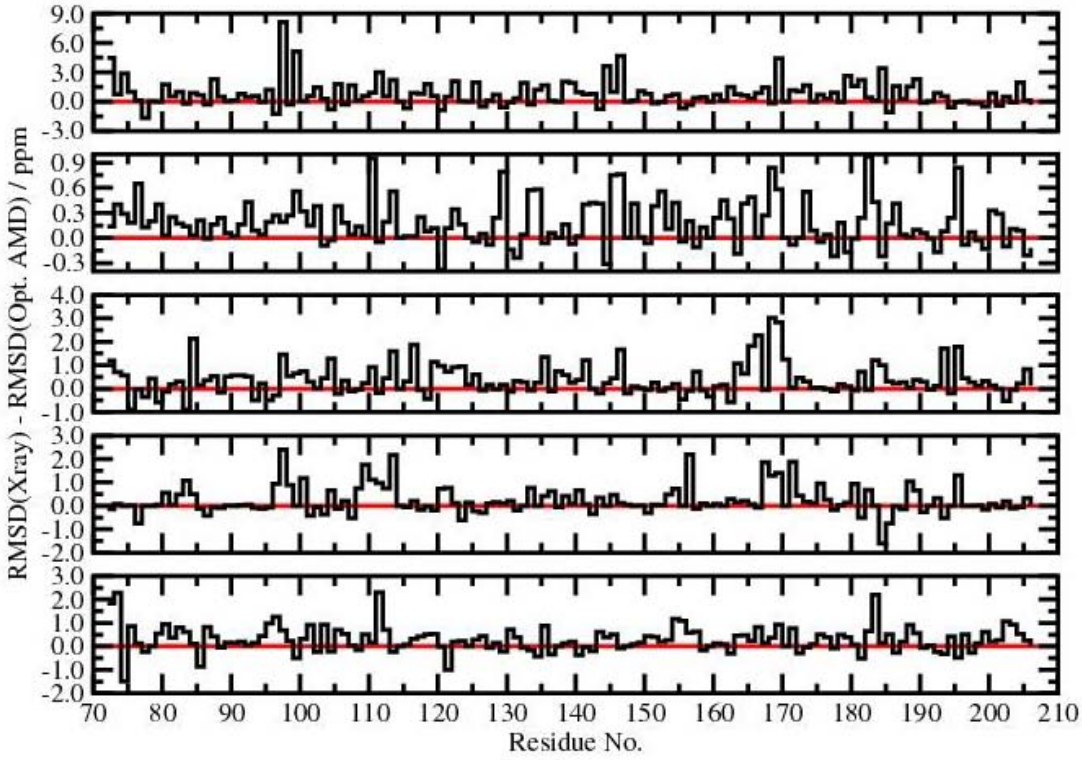


Figure S6: Difference between the predicted chemical shift RMSD for the X-ray crystal structure [RMSD(X-ray)] and that obtained from the trajectory averaged molecular ensembles generated at the optimal acceleration level [RMSD(Opt. AMD)]. Each panel represents a different type of nucleus: From top to bottom, results are shown for nuclei N, H^N, Cα, Cβ and C'. Large positive values indicate a substantial improvement in the predicted chemical shifts obtained from the enhanced conformational space sampling compared to the static X-ray crystal structure. The red line (the zero line) is plotted as a reference. In addition to the general observation that there is a net global improvement in all the chemical shift predictions, there is a strong correlation between those regions which exhibit enhanced conformational space sampling and those regions that show significant improvements in the predicted chemical shifts. This is most obvious for the amino-nitrogen and the Cα and Cβ chemical shifts. For example, for the amino-nitrogen, we observe significant improvements in the predicted chemical shifts for residues 72-75 (in the flexible N-terminal tail), 87, 97, 99, 102, 110-113, 144-149, 169-173 and 180-184 and smaller improvements in the predicted chemical shifts are also observed for those residues that lie in close proximity to these.



Tabulated Results for Ubiquitin:

Table ST1: Tabulated results for ubiquitin.

Structure(s)	RMSD $^{1}H^{N}$/ppm	RMSD ^{15}N/ppm	RMSD $^{13}C\alpha$/ppm	RMSD $^{13}C\beta$/ppm	RMSD $^{13}C'$/ppm	Cumulative RMSD/ppm
1UBQ	0.45	2.45	0.98	1.05	1.02	5.95
Standard 5ns MD	0.52	2.88	0.89	1.05	1.07	6.41
Opt. AMD	0.46	2.31	0.87	1.00	0.99	5.63
EROSII	0.46	2.69	0.89	1.01	1.02	6.07

Chemical shift results for ubiquitin (residues 1-71) are summarized in Table ST1: Predicted chemical shift RMSDs are presented for the X-ray crystal structure (1UBQ), a set of standard 5-ns MD simulations, a set of molecular ensembles obtained at the acceleration level that best reproduced the available experimental RDC and scalar J-coupling data,[11] and the EROSII ensemble[15] (the product of a rigorous ensemble-averaged restrained MD protocol using an extensive set of experimental RDC datasets and over 2800 NOE distance restraints). The results presented for the standard 5-ns MD and the optimal AMD simulations are trajectory averaged. The chemical shift results obtained for ubiquitin are in many ways very similar to those presented for IκBα: The molecular ensembles generated at the acceleration level that best reproduced the experimental RDC data also produced the best predicted chemical shifts, a result which is identical to that observed for IκBα. We observe an improvement of 0.78 ppm in the cumulative chemical shift RMSD from 6.41 ppm for the trajectory averaged standard 5-ns MD simulations to 5.63 ppm for the trajectory-averaged optimal AMD molecular ensembles. This improvement in the cumulative RMSD is smaller than that observed for IκBα (1.52 ppm), a result which is commensurate with the fact that ubiquitin exhibits substantially less molecular flexibility than the IκBα system: A comparison of the NH order parameters obtained from experimental spin-relaxation data and those calculated from molecular ensembles at the RDC-optimal acceleration level for ubiquitin revealed that only six residues (8-11 and 46-48) showed substantial enhanced conformational space sampling on μs-ms time-scales.[11] The same NH order parameter comparison for IκBα is shown in Figure 2 (upper panel) in the paper, revealing substantially more slow time-scale dynamics. Similarly, the backbone RMSD to the average

structure for the optimal AMD ensembles is substantially larger for IκBα than for ubiquitin. For both IκBα and ubiquitin, the most significant improvement in the predicted chemical shift data is obtained for the ^{15}N nuclei. The best predicted chemical shift results are obtained for the optimal AMD ensembles, closely followed by the 1UBQ structure and the EROSII ensemble.

It is noticeable that the X-ray crystal structure for ubiquitin (1UBQ) provides a remarkably good prediction of the chemical shift data. The cumulative chemical shift RMSD for 1UBQ is 5.95 ppm, which is substantially better than the result obtained for the IκBα X-ray crystal structure (8.19 ppm). We consider that there are a variety of reasons for this: First, it has been recognized that the quality of predicted chemical shift results is strongly correlated to the resolution of the X-ray crystal structure.[10] A simple explanation for this observation is that the higher resolution affords a more accurate representation of the time- and ensemble-averaged structure. The X-ray crystal structure (1UBQ) has a resolution of 1.8 angstroms compared to the resolution of the X-ray crystal structure of IκBα (2.7 angstroms). The higher resolution of the ubiquitin structure could therefore account, at least in part, for the lower chemical shift RMSD value. Secondly, a static X-ray crystal structure will provide a much better representation of the time-ensemble average structure for a rigid system than for a flexible system. As we have discussed above, ubiquitin shows considerably less dynamics and therefore less molecular flexibility than IκBα and this could also account, at least in part, for the lower cumulative chemical shift RMSD value for 1UBQ. Ubiquitin was used as part of the training set for the development of the SHIFTX program. The IκBα system was certainly not used as part of the training set for this chemical shift prediction algorithm and, to the best knowledge of the authors, no homologous ankyrin repeat protein was present in the training set. In light of this, it is understandable that the static X-ray crystal structure of ubiquitin will provide more accurate chemical shift predictions than IκBα.

Finally, we would also like to note that the atomic coordinates for IκBα(67-206) were obtained from an X-ray crystal structure of the IκBα/NF-κB complex. Unfortunately, no X-ray crystal structure is available for IκBα in its free-state. We certainly do not ignore the possibility that the structure of IκBα in the free-state might be slightly different from that observed in the complex. However, we would like to point out that the highly flexible regions of IκBα, where we see the most prestigious improvement in the predicted chemical shifts, are found on the opposite side of the molecule to the binding interface and play no direct role in the binding interaction. Although the issue concerning to what extent the local structure of IκBα is affected at the binding interface with Nf-κB is somewhat ambiguous, we do not believe that it in any way detracts from the principal results of this study.

REFERENCES:

[1] D. Hamelberg, J. Mongan and J.A. McCammon, *J. Chem. Phys.* **2004**, *120*, 11919.

[2] D. Hamelberg and J.A. McCammon, *J. Am. Chem. Soc.* **2005**, *127*, 13778.

[3] C.F. Cervantes, P.R.L. Markwick, S-C. Sue, J.A. McCammon, H.J. Dyson, E.A. Komives, *Biochemistry* **2009**, *48*, 8023.

[4] T.E. Cheatham, J.L. Miller, T. Fox, T.A. Darden and P.A. Kollman, *J. Am. Chem. Soc.* **1995**, *117*, 4193.

[5] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, *Proteins:Struct. Funct. and Bio-informatics*, **2006**, *65*, 712.

[6] D.A. Case *et. al.* **2008**, AMBER10, University of California, San Francisco. (see above for complete reference).

[7] D. Hamelberg, C.A. de Oliveira and J.A. McCammon, *J. Chem. Phys.* **2007**, *127*, Art. No. 155102.

[8] T. Shen and D. Hamelberg, *J. Chem. Phys.* **2008**, *129*, Art. No. 034103.

[9] I. Massova and P.A. Kollman, *J. Am. Chem. Soc.* **1999**, *121*, 8133.

[10] S. Neal, A.M. Hip, H. Zhang and D.S. Wishart, *J. Biomol. NMR*, **2003**, *26*, 215.

[11] P.R.L. Markwick, G. Bouvignies, L. Salmon, J.A. McCammon, M. Nilges and M. Blackledge. *J. Am. Chem. Soc.* (accepted).

[12] H. Frauenfelder, S.G. Sligar and P.G. Wolynes, *Science*, **1991**, *254*, 1598.

[13] G. Lipari and A. Szabo, J. Am. Chem. Soc. 1982, 104, 4546.

[14] I. Chandrasekhar, G.M. Clore, A. Szabo, A.M. Gronenborn and B.R. Brooks, *J. Mol. Biol.* **1992**, *226*, 239.

[15] O.F. Lange, N.A. Lakomek, K.F. Walter, C. Fares, G.F. Schroeder, K.F. Walter, S. Becker, J. Meiler, H. Grubmuller, C. Griesinger and B.L. de Groot, *Science* **2008**, *320*, 1471.