Molecular dynamics ensemble refinement of the heterogeneous native state of NCBD using chemical shifts and NOEs

Elena Papaleo[1,2], Carlo Camilloni[3,4], Kaare Teilum[1], Michele Vendruscolo[3] and Kresten Lindorff-Larsen[1]

[1] Structural Biology and NMR Laboratory, Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, DK-2200 Copenhagen, Denmark

[3] Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom

Corresponding Author

Kresten Lindorff-Larsen

Email address: lindorff@bio.ku.dk

Present Addresses

[2] Computational Biology Laboratory, Danish Cancer Society Research Center, Copenhagen, Denmark

[4] Department of Biosciences, University of Milano, Milano, Italy

ABSTRACT

Many proteins display complex dynamical properties that are often intimately linked to their biological functions. As the native state of a protein is best described as an ensemble of conformations, it is important to be able to generate models of native state ensembles with high accuracy. Due to limitations in sampling efficiency and force field accuracy it is, however, challenging to obtain accurate ensembles of protein conformations by the use of molecular simulations alone. Here we show that dynamic ensemble refinement, which combines an accurate atomistic force field with commonly available nuclear magnetic resonance (NMR) chemical shifts and NOEs, can provide a detailed and accurate description of the conformational ensemble of the native state of a highly dynamic protein. As both NOEs and chemical shifts are averaged on timescales up to milliseconds, the resulting ensembles reflect the structural heterogeneity that goes beyond that probed e.g. by NMR relaxation order parameters. We selected the small protein domain NCBD as object of our study since this protein, which has been characterized experimentally in substantial detail, displays a rich and complex dynamical behaviour. In particular, the protein has been described as having a molten-globule like structure, but with a relatively rigid core. Our approach allowed us to describe the conformational dynamics of NCBD in solution, and to probe the structural heterogeneity resulting from both short- and long-timescale dynamics by the calculation of order parameters on different time scales. These results illustrate the usefulness of our approach since they show that NCBD is rather rigid on the nanosecond timescale, but interconverts within a broader ensemble on longer timescales, thus enabling the derivation of a coherent set of conclusions from various NMR experiments on this protein, which could otherwise appear in contradiction with each other.

INTRODUCTION

Molecular dynamics (MD) simulations have the potential ability to provide an accurate, atomic-level description of the conformational ensembles of proteins and their macromolecular complexes (Lindorff-Larsen et al., 2005; Dror et al., 2012; Perilla et al., 2015). Nevertheless, simulations are limited by both the accuracy of the physical models (force fields) and the precision due to conformational sampling (Mobley, 2012; Esteban-Martín, Fenwick & Salvatella, 2012). To overcome these problems, it is possible to bias the simulations using experimental data as structural restraints taking into account the inherent averaging in the experiments (Lindorff-Larsen et al., 2005; Camilloni et al., 2012; Lehtivarjo et al., 2012; Pitera & Chodera, 2012; Camilloni & Vendruscolo, 2014; Ravera et al., 2016). In this way, the experimental data can be included as a system-specific force-field correction, that combines the two sources of information using Bayesian statistics or the maximum entropy principle (Pitera & Chodera, 2012; Roux & Weare, 2013; Cavalli, Camilloni & Vendruscolo, 2013; Boomsma, Ferkinghoff-Borg & Lindorff-Larsen, 2014; White & Voth, 2014; Olsson et al., 2014; MacCallum, Perez & Dill, 2015; Hummer & Köfinger, 2015; Bonomi et al., 2016, 2017). Among the many techniques that can be used to probe structure and dynamics of proteins, NMR spectroscopy stands out as being able to provide a number of different parameters that are sensitive to protein dynamics over different timescales, as well as to probe the "average structure" in solution.

Previously, replica-averaged simulations have provided a wealth of information about the dynamical ensembles that proteins can attain in solution (Lindorff-Larsen et al., 2005; Tang, Schwieters & Clore, 2007; Fenwick et al., 2011; Camilloni et al., 2012; Ángyán & Gáspári, 2013; Camilloni, Cavalli & Vendruscolo, 2013a,b; Islam et al., 2013; Vögeli et al., 2014; Camilloni & Vendruscolo, 2014). Exploiting improvements in the accuracy and speed of predicting protein NMR chemical shifts from protein structure, (Kohlhoff et al., 2009; Han et al., 2011; Li & Brüschweiler, 2012) it is now possible to combine experimental chemical shifts with molecular simulations to study protein structure and dynamics (Wishart & Case, 2001; Cavalli et al., 2007; Shen et al., 2008; Wishart et al., 2008; Robustelli et al., 2009, 2010; Boomsma et al., 2014). In particular, chemical shifts can be used as replica-averaged structural restraints to determine the conformational fluctuations in proteins (Camilloni et al., 2012; Camilloni, Cavalli & Vendruscolo, 2013a,b; Kannan et al., 2014; Kukic et al., 2014; Krieger et al., 2014). By using experimental data as a "system specific force field correction" (Boomsma, Ferkinghoff-Borg & Lindorff-Larsen, 2014) such experimentally-restrained simulations remove some of the uncertainty associated with imperfect force fields and sampling (Tiberti et al., 2015; Löhr, Jussupow & Camilloni, 2017).

Previously, we developed a dynamic-ensemble refinement (DER) approach for determining simultaneously the structure and dynamics of proteins by combining distance restraints from nuclear Overhauser effect (NOE) experiments, dynamical information from relaxation order parameters and MD simulations (Lindorff-Larsen et al., 2005). Similarly, it has been demonstrated that accurate ensembles of conformations that represent longer timescale dynamics can be obtained from residual dipolar couplings (Lange et al., 2008; De Simone et al., 2009, 2015). These applications have, however, relied on a type of data (relaxation order parameters or residual dipolar couplings) that may not be readily available.

We therefore sought to extend this approach to study conformational variability using more commonly available data, thus making the DER method more generally applicable. We thus focus on using NMR chemical shifts and NOEs as these are both commonly available and are averaged over long, millisecond timescales. We demonstrate the potential by describing the structural heterogeneity of a highly dynamic protein. Our method relies on supplementing the sparse experimental data with the experimentally-validated CHARMM22* force field (Piana, Lindorff-Larsen & Shaw, 2011), which provides a relatively accurate description of the subtle balance among the stability of the different secondary structure classes, and which has been shown to provide a good description of many structural and dynamical aspects related to protein structure (Shaw et al., 2010; Lindorff-Larsen et al., 2012a,b; Piana, Lindorff-Larsen & Shaw, 2012; Papaleo et al., 2014; Rauscher et al., 2015). Our hypothesis was that using a more accurate force field would make it possible to determine an accurate ensemble from less information-rich experimental data. In particular, though chemical shifts in principle contain very detailed information, this information is difficult to extract using current methods.

As object of our study we selected NCBD (the Nuclear Coactivator Binding Domain) of CBP (CREB Binding Protein), a 59-residue protein domain that has been experimentally characterized in substantial detail. Experiments on NCBD have revealed a rich and complex dynamical behaviour of the protein in solution (Demarest et al., 2004; Ebert et al., 2008; Kjaergaard, Teilum & Poulsen, 2010; Kjaergaard, Poulsen & Teilum, 2012; Kjaergaard et al., 2013). For a protein of its size, NCBD displays surprisingly broad NMR peaks, suggestive of conformational heterogeneity with relatively slow interconversion between different states. Nevertheless, it was possible to assign both backbone and side chain chemical shifts and determine a number of conformationally-averaged inter-nuclear distances, including a few long-range contacts, via NOE experiments (Ebert et al., 2008; Kjaergaard, Teilum & Poulsen, 2010; Kjaergaard, Poulsen & Teilum, 2012). NMR relaxation experiments suggest that the protein, at least on the nanosecond timescale, is relatively rigid (Kjaergaard, Poulsen & Teilum, 2012). NCBD forms complexes with several other proteins, where it intriguingly folds into remarkably different tertiary structures (Demarest et al., 2002; Qin et al., 2005). For example, the structure of NCBD in complex with ACTR (Demarest et al., 2002) and certain other partners (Waters et al., 2006; Lee et al., 2010) resembles the average structure populated by NCBD in the absence of binding partners (Figure 1), whereas the structure of NCBD is markedly different when bound to the protein IRF-3 (Qin et al., 2005). Thus, the dynamical properties of NCBD, and its ability to adopt different conformations, appear crucial for its diverse biological functions.

Our results show that a dynamic ensemble refinement that combines NOEs, chemical shifts and the CHARMM22* force field provides a rather accurate description of the structural dynamics of the ground state structure of NCBD. We show via cross-validation with independent NMR data that all three components (the two sources of experimental information and the force field) contribute to the overall accuracy. The ensemble that we obtained reveals a relatively broad distribution of conformations, reflecting the conformational heterogeneity of NCBD on the millisecond timescale. Further, we quantified the level of structural fluctuations that would be measured by relaxation experiments and demonstrate that, on the nanosecond timescale, NCBD is more rigid, thus helping to reconcile earlier conflicting views of this protein.

MATERIALS AND METHODS

Ensemble generation. MD simulations were performed using *Gromacs 4.5*, (Pronk et al., 2013) coupled to a modified version of *Plumed 1.3*, (Bonomi et al., 2009) and using either the CHARMM22* (Piana, Lindorff-Larsen & Shaw, 2011) or CHARMM22 (MacKerell, et al., 1998) force fields. As starting structure for most simulations we used the first conformer from a previously determined NMR structure of free NCBD as deposited in the PDB entry 2KKJ (Kjaergaard, Teilum & Poulsen, 2010). To evaluate the effect of our choice of the initial structure, we also performed one simulation starting from an alternative NCBD conformation (PDB entry: 1ZOQ, chain C) (Qin et al., 2005). Missing residues in 1ZOQ (compared to 2KKJ) were rebuilt by *Modeller 9.11* (Fiser & Šali, 2003).

The protein was embedded in a dodecahedral box containing 8372 TIP3P water molecules (Jorgensen et al., 1983) and simulated using periodic boundary conditions with a 2 fs timestep and LINCS constraints (Hess et al., 1993). Production simulations were performed in the NVT ensemble with the Bussi thermostat (Bussi, Donadio & Parrinello, 2007) using a pre-equilibrated starting structure for which the volume was selected based on a short NPT simulation. NaCl was added to a concentration of ~20 mM to reproduce the experimental conditions at which chemical shifts and NOEs were determined (Kjaergaard, Teilum & Poulsen, 2010). The van der Waals and short-range electrostatic interactions were truncated at 9 Å, whereas long-range electrostatic effects were treated with the particle mesh Ewald method (Essmann et al., 1995).

We carried out MD simulations with replica-averaged experimental restraints using 1, 2, 4 or 8 replicas (Table S1 gives an overview of the simulations that were performed). The use of replica-averaged restrained simulations enables us to use different equilibrium experimental observable as a restraint in MD simulation in a way that minimises the risk of over restraining because replica-averaging is a practical implementation of the maximum entropy principle. As a control we also performed a simulation that was not biased by any experimental restraints (i.e. an unbiased simulation). To examine the role played by each of the different types of experimental data, we also performed simulations in which we included different combinations of the experimental restraints: chemical shifts only (CS), NOEs only (NOE), and both chemical shifts and NOEs (CS-NOE). In the simulations, each replica was evolved through a series of simulated annealing (SA) cycles between 304 and 454K for a total duration of 0.6 ns per cycle. We only used structures from the 304K portions of the simulations for our analyses.

Chemical shifts for the backbone atoms (Cα, C', Hα, H and N) and Cβ CS (deposited in BMRB entry 16363) were used as restraints (with the exception of the Cβ of glutamines). The resulting dataset includes 54 Cα, 37 Cβ, 52 Hα and 48 C', H and N chemical shifts, respectively. The backbone chemical shifts cover most of the NCBD sequence with the exception of the first four to six N-terminal residues, depending on type of chemical shifts. The Cβ chemical shifts for the first seven N-terminal and last five C-terminal residues, as well as for some residues of the loops connecting the α-helices, are missing with few exceptions.

During the structure determination protocol, chemical shifts were calculated by *CamShift* (Kohlhoff et al., 2009) for all the nuclei for which an experimental value is available and then averaged over the replicas of the replicas. The resulting average over the replicas was compared with the experimental value, and the ensemble as a whole restrained using a harmonic function with a force constant of 5.2 kJ $mol^{-1}ppm^{-2}$ (Camilloni et al., 2012; Camilloni, Cavalli & Vendruscolo, 2013a). At the higher temperatures, T, explored during the simulated annealing, the force constant was scaled by a factor of (304 K/T). The value of the force constant was chosen roughly to match the calculated chemical shifts to experiments within the uncertainty of the *CamShift* predictor; the experimental uncertainty of the chemical shifts is negligible in comparison.

NOE restraints were obtained by 455 NOE-derived distance intervals (Kjaergaard, Teilum & Poulsen, 2010) (BMRB entry 16363) of which 46 were long-range (i.e. separated by more than 4 residues). The proton-proton distances, r, were calculated and averaged as r^{-6} over the replicas (Tropp, 1980; Lindorff-Larsen et al., 2005). We used a flat-bottomed harmonic function implemented in *Gromacs* to restrain the calculated averaged distances within the experimentally-derived intervals. We used a variable force constant for the NOE-restraints during the SA cycles, allowing the protein to sample more diverse structures in the high-temperature regime and thus to decrease the risk of getting trapped in local minima. Force constants of 1000, 20 and 125 kJ mol^{-1} nm^{-2} were used for the 304K phase, a heating phase (from 304K to 454K) and cooling phase (from 454K to 304K), respectively.

In short, in the replica-averaged simulations we calculated at each step and for each replica-conformation the atomic distances that were measured by the NOE experiments and the back-bone chemical shifts. These calculated single-conformer values were then averaged (linearly for the shifts and using r^{-6} averaging for the distances) to determine the replica-averaged values, which were then compared to the experimentally determined values. Thus, the simulations penalize deviations between the calculated ensemble averages and experimental values but allow fluctuations of individual structures. In this way, the simulations are biased so as to agree with the experimental data as a whole, while allowing individual conformations to take on conformations whose NMR parameters differ from the experimentally derived averages.

To examine the role of the force field used in our approach, we compared the results from two different force fields belonging to the same family (CHARMM). These force fields mostly differ for the main-chain dihedral angle potential, as well a few parameters for certain side chains. Further, in a previous comprehensive evaluation of protein force fields it, was demonstrated that these two force fields resulted in very different levels of agreement between simulations and experiments (Lindorff-Larsen et al., 2012a), making it possible for us to evaluate the importance of force field accuracy in restrained simulations.

Unbiased simulations for the calculation of fast-timescale order parameters. We also performed 28 independent unbiased MD simulations, each 50 ns long, at 304K and with the same computational setup as the restrained simulations, but without any restraints. As starting points, we selected seven different structures from each of the four replicas obtained in the CS-NOE-4 ensemble (Table S1). In particular, the seven structures were selected from the SA cycles after convergence (i.e. at SA cycles 65, 75, 85, 95, 100, 110, 125). We calculated fast timescale order parameters, which correspond to those measured by NMR relaxation measurements, from these 28 unbiased simulations using a previously described approach (Maragakis et al., 2008). In particular, we calculated bond-vector autocorrelation functions (independently from each simulation) including both internal motions and overall tumbling of NCBD. The resulting correlation functions were then averaged over the 28 simulations and subsequently fitted globally to a Lipari-Szabo model (Lipari & Szabo, 1982) to yield relaxation order parameters. To calculate order parameters that report on the long-timescale motions we first aligned the full ensemble and then calculated order parameters as ensemble averages (Maragakis et al., 2008).

Analyses of convergence and cross validation. We used two different methods to examine the convergence of our simulations. First, we used the ENCORE ensemble comparison method (Lindorff-Larsen & Ferkinghoff-Borg, 2009; Tiberti et al., 2015) to quantify the overlap between the structural ensembles. The latter is based on clustering the structures using affinity propagation (setting the "preference value" in the clustering to 12) and subsequent comparison of the ensembles by calculating the Jensen-Shannon (JS) divergence between pairs of ensembles by comparing how they populate the different clusters. For additional details, please confer to original descriptions of the method (Lindorff-Larsen & Ferkinghoff-Borg, 2009; Tiberti et al., 2015). As an alternative method, we calculated the Root Mean Square Inner Product (RMSIP) over the first 10 eigenvectors obtained from a principal component analysis of the covariance matrix of atomic (C_α-atoms) fluctuations (Amadei, Linssen & Berendsen, 1993).

To cross-validate our ensembles we calculated the chemical shifts of side chain methyl hydrogen and carbon atoms using *CH3Shift* (Sahakyan et al., 2011) (both ^{1}H and ^{13}C shifts) and *PPM* (Li & Brüschweiler, 2012) (only ^{1}H shifts) and compared to the previously determined experimental side chain chemical shifts. In particular, we compared the calculated side chain chemical shifts with the experimental values (deposited in BMRB entry 16363) using a reduced χ^2 metric. In this metric, the square deviation between the calculated and experimental values were normalized by the variance of the chemical shift predictor (for each type of chemical shift) and the total number of chemical shifts, so that low numbers indicate good agreement between experimental and calculated chemical shifts.

RESULTS AND DISCUSSION

Convergence of the simulations. Before assessing the accuracy of the different structural ensembles that we generated, we first ensured that the simulated annealing protocol allowed us to obtain converged ensembles that represent the dynamical properties encoded in the experimental restraints and the molecular force field. To quantify convergence of the ensembles, we calculated two different measures of the overlap between the subspaces sampled by different simulations.

First, we used a previously described approach (Lindorff-Larsen & Ferkinghoff-Borg, 2009; Tiberti et al., 2015), which is based on a quantification of the extent to which the different ensembles mix during conformational clustering, to calculate the Jensen-Shannon (JS) divergence between the ensembles (Figure 2). A JS divergence of zero is evidence of identical ensembles, and it has previously been observed that a JS divergence in the range of 0.1-0.3

represents similar ensembles (Lindorff-Larsen & Ferkinghoff-Borg, 2009; Tiberti et al., 2015). We expect that in a converged replica-averaged simulation that the different replicas should populate equally the different structural basins. With this in mind, we calculated the JS divergence between two replicas in a simulation restrained by NOEs and chemical shifts (Figure 2, black line). We find that after approximately ~30 cycles of simulated annealing the two replicas have covered approximately the same conformational space with the JS divergence stabilizing around 0.2-0.3 with the fluctuations in the JS-divergence representing the stochastic nature of the simulations. Thus, we decided to discard the first 45 simulated annealing cycles from all the simulations. As an alternative measure of ensemble similarity we also calculated the Root Mean Square Inner Product (Hess, 2002) (RMSIP) with very similar results. In particular, the similarity of the two replicas converge to an RMSIP value greater than 0.83 (here RMSIP=1 is expected for fully overlapping ensembles).

As a second, perhaps even more stringent, test of convergence we also examined whether two simulations with the same number of replicas and experimental restraints, but initiated from substantially different starting structures, converge to similar ensembles. Indeed, we find that simulations initiated from two distinct structures of NCBD (Table S1) converge to similar ensembles when the first 45 cycles are discarded as initial equilibration (Figure 2, grey line). Thus, based on these two tests we concluded that our sampling protocol allows us to obtain structural ensembles that represent the force field and restraints employed.

Assessment of the accuracy of the NCBD ensembles. Once we had assessed the convergence of the simulations, we analysed the different ensembles to evaluate their accuracy. To do so, we back-calculated experimental parameters that were not used as restraints and compared them with the experimental values. As our different simulations employed different sets of experimental restraints, not all experimental data can be employed for validation purposes. For example, while the NOEs can be used to evaluate the quality of an ensemble obtained using CS-restraints, they can obviously not be used to validate an ensemble that was generated using those NOEs as restraints.

We first examined whether the CS or NOE restraints alone are sufficient to increase the accuracy in the description of the conformational ensemble of NCBD. We thus compared unbiased simulations with simulations biased by either CS or NOEs by cross-validation with the measured NOEs and CS, respectively.

We back-calculated NOEs from the inter-proton distances and observed substantial violations (some greater than 2 Å) in both unbiased and CS ensembles (Figure S1) independently of the number of replicas used for the averaging. To determine the origin of these discrepancies we calculated intramolecular contacts between side chains, and observed an overall decrease in these (from 27 in the previously-determined NMR ensemble, to 14 and 17 in unbiased and CS-restrained, respectively). More specifically we found a loss of inter-helical contacts between helices α1 and α2 in the simulations, in agreement with our finding of several long-range NOEs that are violated in these ensembles.

These results demonstrate that the CS-restraints and MD force field, as implemented here, are not sufficient to provide a fully accurate description of the conformational ensemble of NCBD. Similarly, we found that back-calculation of backbone chemical shifts from the unbiased simulation and, to a lesser extent a NOE-restrained ensemble, resulted in deviations from experiments. We therefore decided to determine conformational ensembles that combine the information of the NOEs, chemical shifts and force field in replica-averaged simulations (CS-NOE) aiming to provide a more accurate structural ensemble of NCBD than possible via the application of just one of the two classes of restraints. We also assessed the influence of the choice of force field since we expected that a more accurate ensemble could be obtained with the relatively limited amounts of experimental data when using a more accurate force field. Thus, we compared simulations using either the CHARMM22 force field (CS-NOE-4-C22 simulation), or a more recent and accurate force field variant, CHARMM22* (CS-NOE simulations).

As both the NOEs and backbone chemical shifts were used as restraints they cannot be used for validation of these ensembles. Instead, we turned to side-chain methyl chemical shifts for a comparison and validation of the different ensembles. Methyl-containing residues, for which the chemical shifts are available, cover the entire protein structure and are thus excellent probes of both local structure (^{13}C methyl chemical shifts, which are mostly dependent on the rotameric state) and long-range contacts (^{1}H methyl chemical shifts). The methyl chemical shifts were predicted by *CH3Shift* (Sahakyan et al., 2011) and the resulting values compared to experiments, separating the contributions from ^{13}C and ^{1}H. We then calculated χ^2_{red} thus taking into account the inherent uncertainty of the chemical shift predictions (Sahakyan et al., 2011).

As also indicated by the calculation of NOEs and backbone chemical shifts, we find that the side chain chemical shifts predicted from the unbiased simulation (green line in Figure 3) deviates substantially from experiments. The introduction of backbone chemical shift restraints (CS ensembles, orange line in Figure 3) provides a better structural ensemble than the force field alone, especially for ^{13}C methyl chemical shifts and when averaged over 2 or 4 replicas. We also calculated the chemical shifts from NOE-derived ensembles, obtained with or without replica-averaging. Surprisingly, we find that the ensembles obtained using NOEs as replica-averaged restraints (NOE, magenta line in Figure 3) perform slightly worse than the CS ensemble. Thus, when evaluated in this way, ensembles derived by MD refinement using either backbone chemical shifts or NOEs do not increase accuracy compared to the ensemble deposited in the PDB.

By combining the NOEs, chemical shifts and the CHARMM22* force field we were, however, able to obtain even more accurate ensembles, in particular when averaging over four replicas, as assessed by the ability to predict side chain ^{13}C and ^{1}H methyl chemical shifts (Figure 3). Interestingly we find that not only the experimental data but also the CHARMM22* force field contributes to the improved agreement with the experimental data. Indeed, when we employ both chemical shift and NOE-based restraints in simulations averaged over 4 replicas, but replacing the CHARMM22* force field by an earlier, less accurate variant of the same force field (CHARMM22; CS-NOE-4-C22) (Lindorff-Larsen et al., 2012a) we find that the accuracy decreases dramatically. Calculations of ^{1}H methyl chemical shifts using *PPM* (Li & Brüschweiler, 2012) instead of *CH3Shift* demonstrate that the conclusions are robust to the method for calculating the chemical shifts (Figure S2). Similarly, calculations of the chemical shifts using the ensemble generated from the alternative starting structure (CS-NOE-2-1ZOQ) resulted in essentially the same agreement with the experimental data as when simulations were initiated from the 2KKJ structure (Figure 3), confirming the conclusions from the convergence analysis described above (Figure 2). The CS-NOE-4 ensemble, which we found to provide the most accurate representation of the free state of NCBD in solution, is shown in Figure 4. It is a relatively broad ensemble of conformations, where the three helical regions are maintained overall, but differ in the lengths and relative positions of the three α-helices.

Small Angle X-ray scattering (SAXS) measurements have been carried out for NCBD in solution (Kjaergaard, Teilum & Poulsen, 2010) and previously been compared to simulation-derived ensembles of NCBD (Knott & Best, 2012; Naganathan & Orozco, 2013). We thus calculated the radius of gyration (R_g) using *CRYSOL* (Svergun, Barberato & Koch, 1995) for the various ensembles. In all cases we find that the average R_g values are in the range of 13.7Å – 14.9Å. These values are comparable to that obtained previously from simulations (13.7 Å) (Knott & Best, 2012) but lower than the values estimated from a Guinier analysis of the experimental data (~16.5 Å) or an ensemble-optimization method (18.8 Å) (Kjaergaard, Teilum & Poulsen, 2010). We note, however, that the experimental values also include contributions from a ~8% population of unfolded protein that is not captured by our simulations. Although a detailed understanding is lacking for the role of solvation on the SAXS properties of partially disordered proteins we, however, expect that the discrepancy between experiment and simulation should be ascribed to remaining force field deficiencies. Indeed, overly large compaction of proteins is a common problem of most atomistic force fields (Piana, Klepeis & Shaw, 2014) though recent work suggests that, at least for fully disordered proteins, that modified protein-water interactions can improve accuracy (Nerenberg et al., 2012; Best, Zheng & Mittal, 2014; Henriques, Cragnell & Skepö, 2015; Mercadante et al., 2015; Piana et al., 2015). We also note that while the force field used here (CHARMM22*) in certain cases has been shown to produce too compact structures, (Piana et al., 2015) in other cases it appears to perform quite well (Rauscher et al., 2015). We expect that resolving these issues will require both further force field developments (Best, 2017) as well as improved methods for comparing experiments and SAXS experiments (Hub, 2018).

A unified view of NCBD dynamics. While the broad peaks and sparse NOEs are suggestive of a rather dynamic protein, previous NMR relaxation measurements of side chain dynamics found relatively high order parameters ($S^2_{\mathrm{relaxation}}$) comparable to values found in well-ordered proteins (Kjaergaard, Poulsen & Teilum, 2012). To shed light on this apparent discrepancy and to assess whether our relatively broad structural ensemble is compatible with mobility on different timescales, we calculated S^2 values representing different timescales.

To mimic the dynamics probed in relaxation experiments we selected 28 structures from each of the 4 replicas of the CS-NOE-4 ensemble sampled at seven different SA steps. Starting from each of these conformations we performed 50 ns of unbiased MD simulation (in total 1.4 μs, Figure S3), and from each simulation we calculated the autocorrelation functions of the N-H bond vectors (without removing the overall rotational motion of the protein). These correlation functions were subsequently averaged and fitted to the Lipari-Szabo model to estimate the $S^2_{\mathrm{relaxation}}$ values, which report on the nanosecond dynamics of the protein (Figure 5, black line). The results show a relatively rigid ensemble on the ns timescale attested by high order parameters throughout most of the polypeptide backbone.

To quantify the backbone dynamics on the longer timescales that may influence both the NOE and chemical shifts (but which the relaxation measurements would not be sensitive to) we defined and calculated "$S^2_{\mathrm{chemical\ shift}}$"-values from the structural variability in the ensemble after aligning the structures. These S^2 values include contributions also from any millisecond-timescale motions that might be present in the ground state of NCBD. As internal and overall motions cannot be decoupled, the results of such calculations will depend on how the ensemble is aligned. In our calculations we chose *THESEUS* (Theobald & Steindel, 2012) as the least biased method to align the structures (Figure 4). These order parameter calculations reveal a broader distribution of conformations with additional, longer-timescale dynamics evident both in loop regions and the C-terminal region, even though relatively high S^2 values are found in the regions of secondary structures (Figure 5, grey line).

A similar analysis of side chain motions suggests even greater differences in motions present on relaxation and chemical shift timescales. In particular, we find that, for methyl-bearing side chains, $S^2_{\text{chemical shift}}$-values are on average lower than $S^2_{\text{relaxation}}$-values by 0.4 compared to an average difference of 0.2 for the backbone amides. Finally, we note that although both calculated $S^2_{\text{chemical shift}}$-values and $S^2_{\text{relaxation}}$-values correlate strongly with the experimentally determined side chain $S^2_{\text{relaxation}}$-values (Spearman correlation coefficient of 0.9 and 0.8, respectively), a more quantitative analysis is hampered by several issues including (i) the presence of a small population of unfolded protein in the experiments, (ii) the difficulty in appropriate model selection of the calculated correlation functions, (iii) the well-known observation of too-fast rotational motions of proteins in the TIP3P model that we used and (iv) uncertainties in the parameterization of the rotational motions in the experimental analyses. We note, however, the potential complications that arise from the fact that the $S^2_{\text{chemical shift}}$-values were obtained from simulations with an experimental bias, whereas the $S^2_{\text{relaxation}}$-values were obtained from simulations starting from such a biased ensemble, but performed with the standard CHARMM22* force field.

Taken together, however, our calculations of order parameters demonstrate that NCBD may be described as a semi-rigid protein on fast-timescales, but with additional dynamics in the backbone and—in particular—side chains on timescales longer than the rotational correlation time of the protein, as also previously suggested (Kjaergaard, Poulsen & Teilum, 2012).

CONCLUSIONS

We have presented an application of the dynamic-ensemble refinement method to study the native state dynamics of NCBD. In the original implementation of DER we combined NMR relaxation order parameters with NOEs in MD simulations (Lindorff-Larsen et al., 2005). This approach was here extended to the combination of chemical shifts and NOEs to make it more generally applicable. In particular, our results show that it is possible to combine NOEs, backbone chemical shifts and an accurate MD force field into replica-averaged restrained simulations, and that all three components add substantially to the accuracy of the resulting NCBD ensemble.

NMR structures are typically obtained by combining distance information from NOE measurements with *in vacuo* simulations, in certain cases with subsequent refinement by short, MD simulations in explicit solvent. Further, the inherent ensemble averaging of the experimental data is typically not exploited explicitly. In this way, standard NMR structures can provide highly accurate models of the "average structure" of a protein, but only little information about the conformational heterogeneity around this average.

Replica-averaged MD simulations make it possible to obtain structural ensembles that match the experimental data according to the principle of maximum entropy (Pitera & Chodera, 2012; Roux & Weare, 2013; Cavalli, Camilloni & Vendruscolo, 2013; Boomsma, Ferkinghoff-Borg & Lindorff-Larsen, 2014; White & Voth, 2014; Olsson et al., 2014). In such calculations prior information, here in the form of a molecular mechanics force field, is biased in a minimal fashion to agree with the experimental data. Thus, to obtain an accurate ensemble, such simulations require an accurate force field, an efficient sampling approach as well as sufficient experimental information. Our results show that, at least in the case of the small, but relatively mobile protein NCBD, it is possible to perform such simulations when NOEs are supplemented by the information available in the backbone chemical shifts and a well-parameterized molecular force field. The application of the experimentally-derived structural restraints helps overcome at least some of the deficiencies in force field accuracy and also improves sampling of the relevant regions of conformational space.

Our approach also allowed us to probe the structural heterogeneity arising from both short- and long-timescale dynamics by the calculation of order parameters. In the case of NCBD we found that this protein can be described as a relatively rigid protein domain on a fast timescale, as attested by the high relaxation order parameters that, nevertheless, displays additional motions in both the backbone and side chains on longer timescales. This situation is reminiscent of the molten globule state of apomyoglobin, that also displays restricted motions on the nanosecond timescale but with greater motions on a slower timescale (Eliezer et al., 2000; Meinhold & Wright, 2011). The current study also provides the groundwork for further studies on NCBDs intricate conformational dynamics, and the relationship to ligand binding (Dogan et al., 2012; Zijlstra et al., 2017). Given the importance of understanding and quantifying protein dynamics, in particular on long timescales, we expect that our approach, which uses only commonly available data, and possible combined with novel algorithms for enhancing sampling (Bonomi et al., 2016; Bonomi, Camilloni & Vendruscolo, 2016), will have a wide range of applications.

Acknowledgements

We would like to thank Magnus Kjaergaard, Wouter Boomsma, Matteo Tiberti and Peter Wright for fruitful discussion and comments.

Funding Sources

E.P. and K.L.-L. were supported by a Hallas-Møller stipend from the Novo Nordisk Foundation. The project was also supported by the Danish e-Infrastructure Cooperation HPC Grant 2013. We also acknowledge that the results of this research have been achieved using the PRACE Research Infrastructure Resource Curie (France, 7th PRACE Tier0, NMRFUNC).

Supporting Information

Supporting figures with the data on back-calculation of NOEs in MD-unbiased and MD-CS ensembles are reported in Figure S1. The deviation between experimental and calculated side-chain ^{1}H chemical shifts calculated with PPM software are shown in Figure S2. The main-chain RMSD profiles of the 28 unbiased constant temperature MD simulations of NCBD are reported in Figure S3.

REFERENCES

Amadei A., Linssen AB., Berendsen HJ. 1993. Essential dynamics of proteins. *Proteins* 17:412–25. DOI: 10.1002/prot.340170408.

Ángyán AF., Gáspári Z. 2013. Ensemble-based interpretations of NMR structural data to describe protein internal dynamics. *Molecules* 18:10548–10567. DOI: 10.3390/molecules180910548.

Best RB. 2017. Computational and theoretical advances in studies of intrinsically disordered proteins. *Current Opinion in Structural Biology* 42:147–154. DOI: 10.1016/j.sbi.2017.01.006.

Best RB., Zheng W., Mittal J. 2014. Balanced Protein − Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *Journal of Chemical Theory and Computation* 10:5113–5124.

Bonomi M., Branduardi D., Bussi G., Camilloni C., Provasi D., Raiteri P., Donadio D., Marinelli F., Pietrucci F., Broglia RA., Parrinello M. 2009. PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications* 180:1961–1972. DOI: 10.1016/j.cpc.2009.05.011.

Bonomi M., Camilloni C., Cavalli A., Vendruscolo M. 2016. Metainference: A Bayesian inference method for heterogeneous systems. *Science Advances* 2:e1501177–e1501177. DOI: 10.1126/sciadv.1501177.

Bonomi M., Camilloni C., Vendruscolo M. 2016. Metadynamic metainference: Enhanced sampling of the metainference ensemble using metadynamics. *Scientific Reports* 6:31232. DOI: 10.1038/srep31232.

Bonomi M., Heller GT., Camilloni C., Vendruscolo M. 2017. Principles of protein structural ensemble determination. *Current Opinion in Structural Biology* 42:106–116. DOI: 10.1016/J.SBI.2016.12.004.

Boomsma W., Ferkinghoff-Borg J., Lindorff-Larsen K. 2014. Combining experiments and simulations using the maximum entropy principle. *PLoS computational biology* 10:e1003406. DOI: 10.1371/journal.pcbi.1003406.

Boomsma W., Tian P., Frellsen J., Ferkinghoff-Borg J., Hamelryck T., Lindorff-Larsen K., Vendruscolo M. 2014. Equilibrium simulations of proteins using molecular fragment replacement and NMR chemical shifts. *Proceedings of the National Academy of Sciences* *of the United States of America* 111:13852–7. DOI: 10.1073/pnas.1404948111.

Bussi G., Donadio D., Parrinello M. 2007. Canonical sampling through velocity rescaling. *The Journal of chemical physics* 126:14101. DOI: 10.1063/1.2408420.

Camilloni C., Cavalli A., Vendruscolo M. 2013a. Assessment of the Use of NMR Chemical Shifts as Replica-Averaged Structural Restraints in Molecular Dynamics Simulations to Characterise the Dynamics of Proteins. *The journal of physical chemistry. B*. DOI: 10.1021/jp3106666.

Camilloni C., Cavalli A., Vendruscolo M. 2013b. Replica-Averaged Metadynamics. *Journal of Chemical Theory and Computation* 9:5610–5617. DOI: 10.1021/ct4006272.

Camilloni C., Robustelli P., De Simone A., Cavalli A., Vendruscolo M. 2012. Characterisation of the conformational equilibrium between the two major substates of RNase A using NMR chemical shifts. *Journal of the American Chemical Society* 134:3968–3971.

Camilloni C., Vendruscolo M. 2014. Statistical mechanics of the denatured state of a protein using replica-averaged metadynamics. *Journal of the American Chemical Society* 136:8982–91. DOI: 10.1021/ja5027584.

Cavalli A., Camilloni C., Vendruscolo M. 2013. Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. *The Journal of chemical physics* 138:94112. DOI: 10.1063/1.4793625.

Cavalli A., Salvatella X., Dobson CM., Vendruscolo M. 2007. Protein structure determination from NMR chemical shifts. *Proceedings of the National Academy of Sciences of the United States of America* 104:9615–20. DOI: 10.1073/pnas.0610313104.

Demarest SJ., Deechongkit S., Dyson HJ., Evans RM., Wright PE. 2004. Packing, specificity, and mutability at the binding interface between the p160 coactivator and CREB-binding protein. *Protein science* 13:203–10. DOI: 10.1110/ps.03366504.

Demarest SJ., Martinez-Yamout M., Chung J., Chen H., Xu W., Dyson HJ., Evans RM., Wright PE. 2002. Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. *Nature* 415:549–53. DOI: 10.1038/415549a.

Dogan J., Schmidt T., Mu X., Engström Å., Jemth P. 2012. Fast association and slow transitions in the interaction between two intrinsically disordered protein domains. *The Journal of biological chemistry* 287:34316–24. DOI: 10.1074/jbc.M112.399436.

Dror RO., Dirks RM., Grossman JP., Xu H., Shaw DE. 2012. Biomolecular simulation: a computational microscope for molecular biology. *Annual review of biophysics* 41:429–52. DOI: 10.1146/annurev-biophys-042910-155245.

Ebert M-O., Bae S-H., Dyson HJ., Wright PE. 2008. NMR relaxation study of the complex formed between CBP and the activation domain of the nuclear hormone receptor coactivator ACTR. *Biochemistry* 47:1299–308. DOI: 10.1021/bi701767j.

Eliezer D., Chung J., Dyson HJ., Wright PE. 2000. Native and Non-native Secondary Structure and Dynamics in the pH 4 Intermediate of Apomyoglobin. *Biochemistry* 39:2894–2901. DOI: 10.1021/BI992545F.

Essmann U., Perera L., Berkowitz ML., Darden T., Lee H., Pedersen LG. 1995. A smooth particle mesh Ewald method. *The Journal of Chemical Physics* 103:8577. DOI: 10.1063/1.470117.

Esteban-Martín S., Bryn Fenwick R., Salvatella X. 2012. Synergistic use of NMR and MD simulations to study the structural heterogeneity of proteins. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 2:466–478. DOI: 10.1002/wcms.1093.

Fenwick RB., Esteban-Martín S., Richter B., Lee D., Walter KFA., Milovanovic D., Becker S., Lakomek NA., Griesinger C., Salvatella X. 2011. Weak Long-Range Correlated Motions in a Surface Patch of Ubiquitin Involved in Molecular Recognition. *Journal of the American Chemical Society* 133:10336–10339. DOI: 10.1021/ja200461n.

Fiser A., Šali A. 2003. MODELLER: Generation and Refinement of Homology-Based Protein Structure Models. *Methods in Enzymology* 374:461–491. DOI: 10.1016/S0076-6879(03)74020-8.

Han B., Liu Y., Ginzinger SW., Wishart DS. 2011. SHIFTX2: significantly improved protein chemical shift prediction. *Journal of biomolecular NMR* 50:43–57. DOI: 10.1007/s10858-011-9478-4.

Henriques J., Cragnell C., Skepö M. 2015. Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment. *Journal of Chemical Theory and Computation* 11:3420–3431. DOI: 10.1021/ct501178z.

Hess B. 2002. Convergence of sampling in protein simulations. *Physical review. E, Statistical, nonlinear, and soft matter physics* 65:31910.

Hess B., Bekker H., Berendsen H., Fraaije J. 1993. LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry* 12:1463–1472. DOI: 10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

Hub JS. 2018. Interpreting solution X-ray scattering data using molecular simulations. *Current Opinion in Structural Biology* 49:18–26. DOI: 10.1016/J.SBI.2017.11.002.

Hummer G., Köfinger J. 2015. Bayesian ensemble refinement by replica simulations and reweighting. *The Journal of chemical physics* 143:243150. DOI: 10.1063/1.4937786.

Islam SM., Stein RA., McHaourab HS., Roux B. 2013. Structural refinement from restrained-ensemble simulations based on EPR/DEER data: application to T4 lysozyme. *The journal of physical chemistry. B* 117:4740–54. DOI: 10.1021/jp311723a.

Jorgensen WL., Chandrasekhar J., Madura JD., Impey RW., Klein ML. 1983. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* 79:926. DOI: 10.1063/1.445869.

Kannan A., Camilloni C., Sahakyan AB., Cavalli A., Vendruscolo M. 2014. A conformational ensemble derived using NMR methyl chemical shifts reveals a mechanical clamping transition that gates the binding of the HU protein to DNA. *Journal of the American Chemical Society* 136:2204–7. DOI: 10.1021/ja4105396.

Kjaergaard M., Andersen L., Nielsen LD., Teilum K. 2013. A Folded Excited State of Ligand-Free Nuclear Coactivator Binding Domain (NCBD) Underlies Plasticity in Ligand Recognition. *Biochemistry*:130201143825004. DOI: 10.1021/bi4001062.

Kjaergaard M., Poulsen FM., Teilum K. 2012. Is a malleable protein necessarily highly dynamic? The hydrophobic core of the nuclear coactivator binding domain is well ordered. *Biophysical journal* 102:1627–35. DOI: 10.1016/j.bpj.2012.02.014.

Kjaergaard M., Teilum K., Poulsen FM. 2010. Conformational selection in the molten globule state of the nuclear coactivator binding domain of CBP. *Proceedings of the National Academy of Sciences of the United States of America* 107:12535–40. DOI: 10.1073/pnas.1001693107.

Knott M., Best RB. 2012. A preformed binding interface in the unbound ensemble of an intrinsically disordered protein: evidence from molecular simulations. *PLoS computational biology* 8:e1002605. DOI: 10.1371/journal.pcbi.1002605.

Kohlhoff KJ., Robustelli P., Cavalli A., Salvatella X., Vendruscolo M. 2009. Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. *Journal of the*

American Chemical Society 131:13894–5. DOI: 10.1021/ja903772t.

Krieger JM., Fusco G., Lewitzky M., Simister PC., Marchant J., Camilloni C., Feller SM., De Simone A. 2014. Conformational recognition of an intrinsically disordered protein. *Biophysical journal* 106:1771–9. DOI: 10.1016/j.bpj.2014.03.004.

Kukic P., Camilloni C., Cavalli A., Vendruscolo M. 2014. Determination of the Individual Roles of the Linker Residues in the Interdomain Motions of Calmodulin Using NMR Chemical Shifts. *Journal of molecular biology* 426:1826–1838. DOI: 10.1016/j.jmb.2014.02.002.

Lange OF., Lakomek N-A., Farès C., Schröder GF., Walter KFA., Becker S., Meiler J., Grubmüller H., Griesinger C., de Groot BL. 2008. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science (New York, N.Y.)* 320:1471–5. DOI: 10.1126/science.1157092.

Lee CW., Martinez-Yamout MA., Dyson HJ., Wright PE. 2010. Structure of the p53 transactivation domain in complex with the nuclear receptor coactivator binding domain of CREB binding protein. *Biochemistry* 49:9964–71. DOI: 10.1021/bi1012996.

Lehtivarjo J., Tuppurainen K., Hassinen T., Laatikainen R., Peräkylä M. 2012. Combining NMR ensembles and molecular dynamics simulations provides more realistic models of protein structures in solution and leads to better chemical shift prediction. *Journal of biomolecular NMR* 52:257–67. DOI: 10.1007/s10858-012-9609-6.

Li D-W., Brüschweiler R. 2012. PPM: a side-chain and backbone chemical shift predictor for the assessment of protein conformational ensembles. *Journal of biomolecular NMR* 54:257–65. DOI: 10.1007/s10858-012-9668-8.

Lindorff-Larsen K., Best RB., Depristo MA., Dobson CM., Vendruscolo M. 2005. Simultaneous determination of protein structure and dynamics. *Nature* 433:128–32. DOI: 10.1038/nature03199.

Lindorff-Larsen K., Ferkinghoff-Borg J. 2009. Similarity measures for protein ensembles. *PloS one* 4:e4203. DOI: 10.1371/journal.pone.0004203.

Lindorff-Larsen K., Maragakis P., Piana S., Eastwood MP., Dror RO., Shaw DE. 2012a. Systematic validation of protein force fields against experimental data. *PloS one* 7:e32131. DOI: 10.1371/journal.pone.0032131.

Lindorff-Larsen K., Trbovic N., Maragakis P., Piana S., Shaw DE. 2012b. Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *Journal of the American Chemical Society* 134:3787–91. DOI: 10.1021/ja209931w.

Lipari G., Szabo A. 1982. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *Journal of the American Chemical Society* 104:4546–4559. DOI: 10.1021/ja00381a009.

Löhr T., Jussupow A., Camilloni C. 2017. Metadynamic metainference: Convergence towards force field independent structural ensembles of a disordered peptide. *The Journal of Chemical Physics* 146:165102. DOI: 10.1063/1.4981211.

MacCallum JL., Perez A., Dill KA. 2015. Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proceedings of the National Academy of Sciences of the United States of America* 112:6985–90. DOI: 10.1073/pnas.1506788112.

MacKerell, AD., Bashford D., Dunbrack, RL., Evanseck JD., Field MJ., Fischer S., Gao J., Guo H., Ha S., Joseph-McCarthy D., Kuchnir L., Kuczera K., Lau FTK., Mattos C., Michnick S., Ngo T., Nguyen DT., Prodhom B., Reiher WE., Roux B., Schlenkrich M., Smith JC., Stote R., Straub J., Watanabe M., Wiórkiewicz-Kuczera J., Yin D., Karplus M. 1998. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *The Journal of Physical Chemistry B* 102:3586–3616. DOI: 10.1021/jp973084f.

Maragakis P., Lindorff-Larsen K., Eastwood MP., Dror RO., Klepeis JL., Arkin IT., Jensen MØ., Xu H., Trbovic N., Friesner RA., Palmer AG., Shaw DE. 2008. Microsecond Molecular Dynamics Simulation Shows Effect of Slow Loop Dynamics on Backbone Amide Order Parameters of Proteins. *The Journal of Physical Chemistry B* 112:6155–6158. DOI: 10.1021/jp077018h.

Meinhold DW., Wright PE. 2011. Measurement of protein unfolding/refolding kinetics and structural characterization of hidden intermediates by NMR relaxation dispersion. *Proceedings of the National Academy of Sciences of the United States of America* 108:9078–83. DOI: 10.1073/pnas.1105682108.

Mercadante D., Milles S., Fuertes G., Svergun DI., Lemke E a., Gräter F. 2015. Kirkwood-Buff approach rescues over-collapse of a disordered protein in canonical protein force fields. *The Journal of Physical Chemistry B*:150601163019003. DOI: 10.1021/acs.jpcb.5b03440.

Mobley DL. 2012. Let's get honest about sampling. *Journal of computer-aided molecular* *design* 26:93–5. DOI: 10.1007/s10822-011-9497-y.

Naganathan AN., Orozco M. 2013. The conformational landscape of an intrinsically disordered DNA-binding domain of a transcription regulator. *The journal of physical chemistry. B* 117:13842–50. DOI: 10.1021/jp408350v.

Nerenberg PS., Jo B., So C., Tripathy A., Head-Gordon T. 2012. Optimizing solute-water van der Waals interactions to reproduce solvation free energies. *The journal of physical chemistry. B* 116:4524–34. DOI: 10.1021/jp2118373.

Olsson S., Vögeli BR., Cavalli A., Boomsma W., Ferkinghoff-Borg J., Lindorff-Larsen K., Hamelryck T. 2014. Probabilistic Determination of Native State Ensembles of Proteins. *Journal of Chemical Theory and Computation* 10:3484–3491. DOI: 10.1021/ct5001236.

Papaleo E., Sutto L., Gervasio FL., Lindorff-Larsen K. 2014. Conformational Changes and Free Energies in a Proline Isomerase. *Journal of Chemical Theory and Computation* 10:4169–4174. DOI: 10.1021/ct500536r.

Perilla JR., Goh BC., Cassidy CK., Liu B., Bernardi RC., Rudack T., Yu H., Wu Z., Schulten K. 2015. Molecular dynamics simulations of large macromolecular complexes. *Current Opinion in Structural Biology* 31:64–74. DOI: 10.1016/j.sbi.2015.03.007.

Piana S., Donchev AG., Robustelli P., Shaw DE. 2015. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *The Journal of Physical Chemistry B* 119:5113–23. DOI: 10.1021/jp508971m.

Piana S., Klepeis JL., Shaw DE. 2014. Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Current opinion in structural biology* 24:98–105. DOI: 10.1016/j.sbi.2013.12.006.

Piana S., Lindorff-Larsen K., Shaw DE. 2011. How robust are protein folding simulations with respect to force field parameterization? *Biophysical journal* 100:L47-9. DOI: 10.1016/j.bpj.2011.03.051.

Piana S., Lindorff-Larsen K., Shaw DE. 2012. Protein folding kinetics and thermodynamics from atomistic simulation. *Proceedings of the National Academy of Sciences of the United States of America* 109:17845–50. DOI: 10.1073/pnas.1201811109.

Pitera JW., Chodera JD. 2012. On the Use of Experimental Observations to Bias Simulated Ensembles. *Journal of Chemical Theory and Computation* 8:3445–3451. DOI:

10.1021/ct300112v.

Pronk S., Páll S., Schulz R., Larsson P., Bjelkmar P., Apostolov R., Shirts MR., Smith JC., Kasson PM., van der Spoel D., Hess B., Lindahl E. 2013. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* 29:845–54. DOI: 10.1093/bioinformatics/btt055.

Qin BY., Liu C., Srinath H., Lam SS., Correia JJ., Derynck R., Lin K. 2005. Crystal structure of IRF-3 in complex with CBP. *Structure* 13:1269–77. DOI: 10.1016/j.str.2005.06.011.

Rauscher S., Gapsys V., Gajda MJ., Groot BL De., Grubmüller H. 2015. Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field : A Comparison to Experiment Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field : A Comparison to Experiment. *Journal of Chemical Theory and Computation* 11:5513–5524. DOI: 10.1021/acs.jctc.5b00736.

Ravera E., Sgheri L., Parigi G., Luchinat C. 2016. A critical assessment of methods to recover information from averaged data. *Physical chemistry chemical physics : PCCP* 18:5686–701. DOI: 10.1039/c5cp04077a.

Robustelli P., Cavalli A., Dobson CM., Vendruscolo M., Salvatella X. 2009. Folding of small proteins by Monte Carlo simulations with chemical shift restraints without the use of molecular fragment replacement or structural homology. *The journal of physical chemistry. B* 113:7890–6. DOI: 10.1021/jp900780b.

Robustelli P., Kohlhoff K., Cavalli A., Vendruscolo M. 2010. Using NMR chemical shifts as structural restraints in molecular dynamics simulations of proteins. *Structure* 18:923–33. DOI: 10.1016/j.str.2010.04.016.

Roux B., Weare J. 2013. On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *The Journal of chemical physics* 138:84107. DOI: 10.1063/1.4792208.

Sahakyan AB., Vranken WF., Cavalli A., Vendruscolo M. 2011. Structure-based prediction of methyl chemical shifts in proteins. *Journal of biomolecular NMR* 50:331–46. DOI: 10.1007/s10858-011-9524-2.

Shaw DE., Maragakis P., Lindorff-Larsen K., Piana S., Dror RO., Eastwood MP., Bank JA., Jumper JM., Salmon JK., Shan Y., Wriggers W. 2010. Atomic-level characterization of the structural dynamics of proteins. *Science* 330:341–6. DOI: 10.1126/science.1187409.

Shen Y., Lange O., Delaglio F., Rossi P., Aramini JM., Liu G., Eletsky A., Wu Y., Singarapu KK., Lemak A., Ignatchenko A., Arrowsmith CH., Szyperski T., Montelione GT., Baker D., Bax A. 2008. Consistent blind protein structure generation from NMR chemical shift data. *Proceedings of the National Academy of Sciences of the United States of America* 105:4685–90. DOI: 10.1073/pnas.0800256105.

De Simone A., Aprile FA., Dhulesia A., Dobson CM., Vendruscolo M. 2015. Structure of a low-population intermediate state in the release of an enzyme product. *eLife* 4. DOI: 10.7554/eLife.02777.

De Simone A., Richter B., Salvatella X., Vendruscolo M. 2009. Toward an accurate determination of free energy landscapes in solution states of proteins. *Journal of the American Chemical Society* 131:3810–1. DOI: 10.1021/ja8087295.

Svergun D., Barberato C., Koch MHJ. 1995. CRYSOL – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *Journal of Applied Crystallography* 28:768–773. DOI: 10.1107/S0021889895007047.

Tang C., Schwieters CD., Clore GM. 2007. Open-to-closed transition in apo maltose-binding protein observed by paramagnetic NMR. *Nature* 449:1078–82. DOI: 10.1038/nature06232.

Theobald DL., Steindel PA. 2012. Optimal simultaneous superpositioning of multiple structures with missing data. *Bioinformatics* 28:1972–1979. DOI: 10.1093/bioinformatics/bts243.

Tiberti M., Papaleo E., Bengtsen T., Boomsma W., Lindorff-Larsen K. 2015. ENCORE: Software for quantitative ensemble comparison. *PLoS computational biology* 11:e1004415.

Tropp J. 1980. Dipolar relaxation and nuclear Overhauser effects in nonrigid molecules: The effect of fluctuating internuclear distances. *The Journal of Chemical Physics* 72:6035. DOI: 10.1063/1.439059.

Vögeli B., Orts J., Strotz D., Chi C., Minges M., Wälti MA., Güntert P., Riek R. 2014. Towards a true protein movie: a perspective on the potential impact of the ensemble-based structure determination using exact NOEs. *Journal of magnetic resonance* 241:53–9. DOI: 10.1016/j.jmr.2013.11.016.

Waters L., Yue B., Veverka V., Renshaw P., Bramham J., Matsuda S., Frenkiel T., Kelly G., Muskett F., Carr M., Heery DM. 2006. Structural diversity in p160/CREB-binding protein coactivator complexes. *The Journal of biological chemistry* 281:14787–95. DOI: 10.1074/jbc.M600237200.

White AD., Voth GA. 2014. Efficient and Minimal Method to Bias Molecular Simulations with Experimental Data. *Journal of Chemical Theory and Computation* 10:3023–3030. DOI: 10.1021/ct500320c.

Wishart DS., Arndt D., Berjanskii M., Tang P., Zhou J., Lin G. 2008. CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. *Nucleic acids research* 36:W496-502. DOI: 10.1093/nar/gkn305.

Wishart DS., Case DA. 2001. Use of chemical shifts in macromolecular structure determination. *Methods in enzymology* 338:3–34.

Zijlstra N., Dingfelder F., Wunderlich B., Zosel F., Benke S., Nettels D., Schuler B. 2017. Rapid Microfluidic Dilution for Single-Molecule Spectroscopy of Low-Affinity Biomolecular Complexes. *Angewandte Chemie* 129:7232–7235. DOI: 10.1002/ange.201702439.

FIGURE LEGENDS

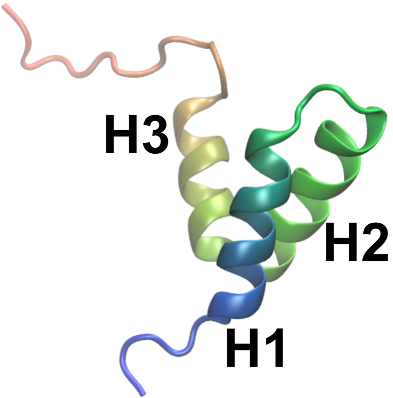
Figure 1. A previously determined structural model of the conformation of NCBD in solution. The structure is shown as a cartoon (PDB entry: 2KKJ) with the protein coloured from the N- to the C-terminal (blue to red). The three α-helices are labelled. The goal of this work is to provide an ensemble of structures that represent the conformational fluctuations associated with this average conformation.

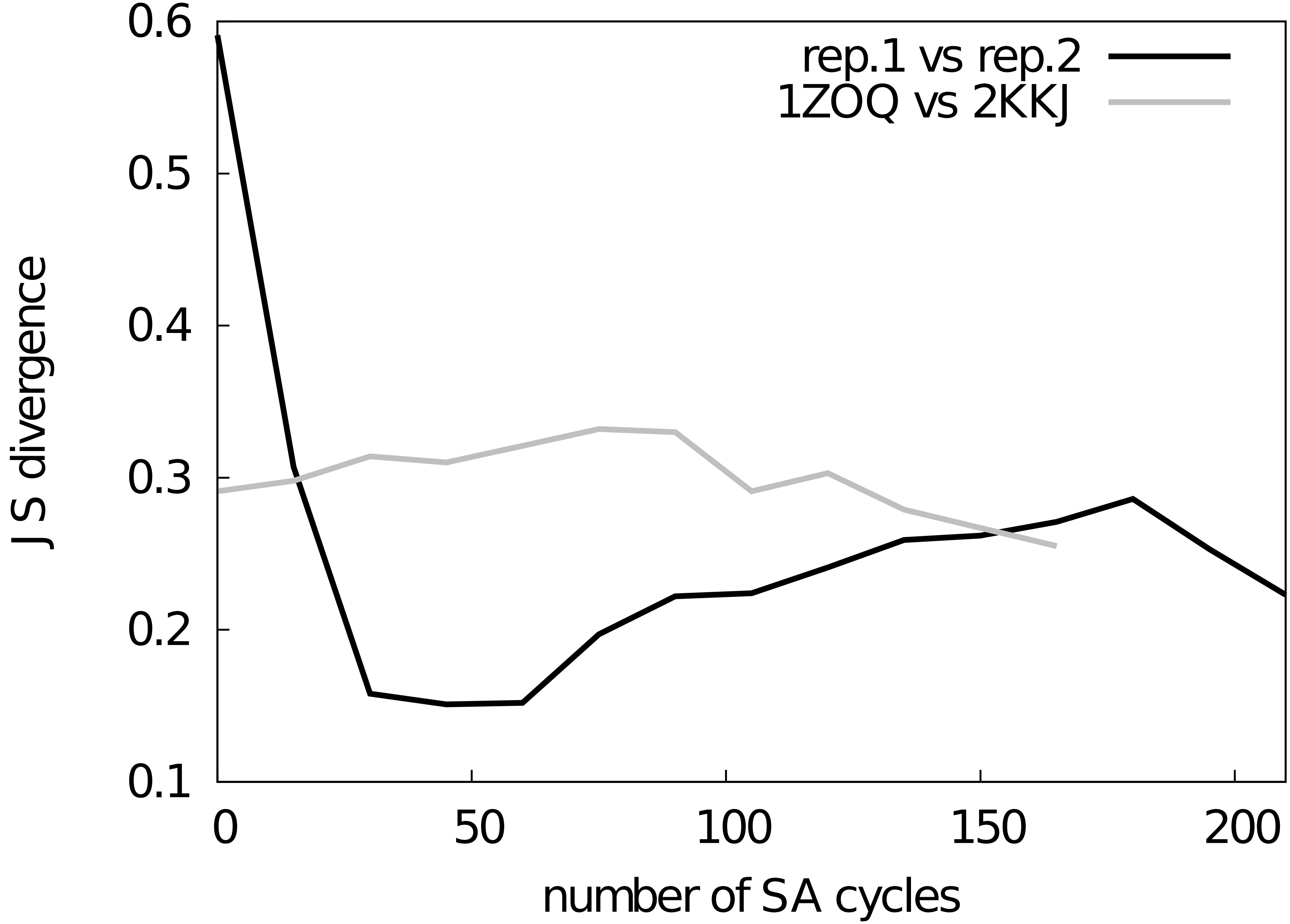
Figure 2. **Assessment of the convergence of the simulations.** The similarity between structural ensembles was quantified using structural clustering with Affinity Propagation and subsequent comparison of the ensembles by Jensen-Shannon (JS) divergence. The JS divergence between two identical ensembles is zero, and it has previously been found that values less than ~0.3 represent similar ensembles. We monitored the evolution of the JS-divergence in two different tests, either by comparing two replicas from the same simulation (i.e. CS-NOE-2, black) or two simulations with the same force field and restraints but different starting structures (i.e. CS-NOE-2 starting from 2KKJ and 1ZOQ structures, respectively, grey). As described in the text we discarded the first 45 SA cycles before calculating the ensemble similarity for the test with different starting structures.

Figure 3. Validation of the structural ensemble using ^{13}C (upper panel) and ^{1}H (bottom panel) side-chain methyl chemical shifts. We calculated the deviation between experimental and predicted side-chain chemical shifts from each MD ensemble. The results are shown as a function of the number of replicas used for the averaging of the simulations. The previously determined NMR structure (black) and unbiased MD simulation (green) do not involve replica averaging and are shown as horizontal lines.

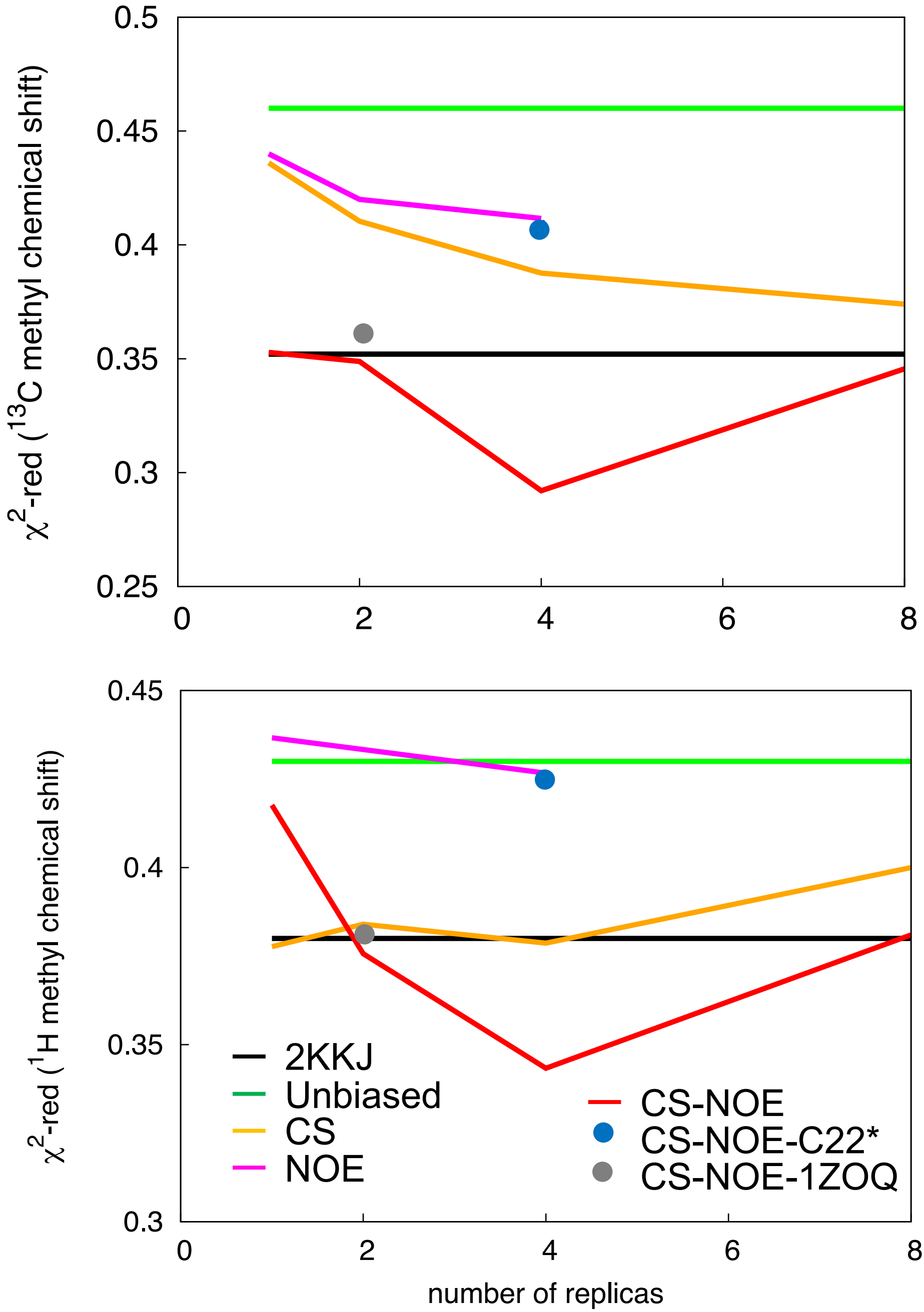
Figure 4. Conformational ensemble of the free state of NCBD obtained by molecular dynamics simulations with the CHARMM22* force field and replica-averaged CS and NOE restraints. The α-helices are represented as cylinders and the structural ensemble was aligned using *THESEUS*.

Figure 5. Calculation of order parameters from MD simulations to probe short and long timescale dynamics. We calculated S^2 order parameters that reflect either motions faster than overall tumbling of the protein (black) or longer timescale motions that give rise to chemical shift and NOE averaging (red). For reference, the main chain Root Mean Square Deviation (RMSD) values of the 28 unbiased simulations that we used to calculate the $S^2_{\text{relaxation}}$ values are shown in Figure S3.

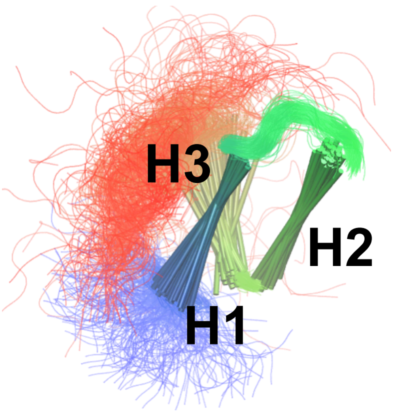

H3
H2
H1

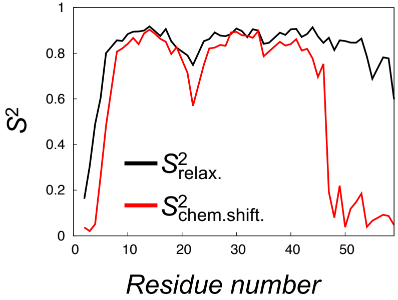


rep.1 vs rep.2
1ZOQ vs 2KKJ
J S divergence
number of SA cycles
0.1
0.2
0.3
0.4
0.5
0.6
0
50
100
150
200



0.5
0.45
0.4
0.35
0.3
0.25
χ^2-red (^{13}C methyl chemical shift)
0
2
4
6
8
0.45
0.4
0.35
0.3
χ^2-red (^{1}H methyl chemical shift)
0
2
4
6
8
number of replicas
2KKJ
Unbiased
CS
NOE
CS-NOE
CS-NOE-C22*
CS-NOE-1ZOQ


H3
H2
H1



1
0.8
0.6
0.4
0.2
0
S^2
0
10
20
30
40
50
$S^2_{relax.}$
$S^2_{chem.shift.}$
Residue number

Supporting Information

Molecular dynamics ensemble refinement of the heterogeneous native state of NCBD using chemical shifts and NOEs

Elena Papaleo[1,2], Carlo Camilloni[3,4], Kaare Teilum[1], Michele Vendruscolo[3] and Kresten Lindorff-Larsen[1]

[1] Structural Biology and NMR Laboratory, Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, DK-2200 Copenhagen, Denmark

[3] Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom

Present Addresses

[2] Computational Biology Laboratory, Danish Cancer Society Research Center, Copenhagen, Denmark

[4] Department of Biosciences, University of Milano, Milano, Italy

Table S1. Summary of the MD ensembles obtained in this study. The number of simulated annealing (SA) cycles listed represents the number of cycles performed (for each replica) and analyzed after discarding the first 45 cycles for convergence.

MD ensemble	Number of replicas	Number of SA cycles per replica	Experimental restraints used	Force Field	Starting Structure
unbiased	//		//	CHARMM22*	2KKJ
CS-1	1	320	CS	CHARMM22*	2KKJ
CS-2	2	160	CS	CHARMM22*	2KKJ
CS-4	4	80	CS	CHARMM22*	2KKJ
CS-8	8	40	CS	CHARMM22*	2KKJ
NOE-1	1	320	NOEs	CHARMM22*	2KKJ
NOE-2	2	160	NOEs	CHARMM22*	2KKJ
NOE-4	4	80	NOEs	CHARMM22*	2KKJ
CS-NOE-1	1	320	NOEs	CHARMM22*	2KKJ
CS-NOE-2	2	160	CS and NOEs	CHARMM22*	2KKJ
CS-NOE-2-1ZOQ	2	160	CS and NOEs	CHARMM22*	1ZOQ
CS-NOE-4	4	80	CS and NOEs	CHARMM22*	2KKJ
CS-NOE-4-C22	4	80	CS and NOEs	CHARMM22	2KKJ
CS-NOE-8	8	40	CS and NOEs	CHARMM22*	2KKJ

Figure S1. Cross-validation of unbiased and CS-restrained ensembles using NOE measurements. We calculated the total number of NOE violations in the unbiased and CS-restrained ensembles. We used 455 NOE-derived distance restraints (BMRB entry 16363) of which 409 are short- and 46 are long-range (i.e. separated by more than 4 residues). In addition to the total number of violations (all), we also separated the violations into different categories depending on their magnitude (0.5Å-1Å, 1Å-2Å or greater than 2Å) and whether they are short (0–4 residues apart) or long-range (more than 4 residues apart).

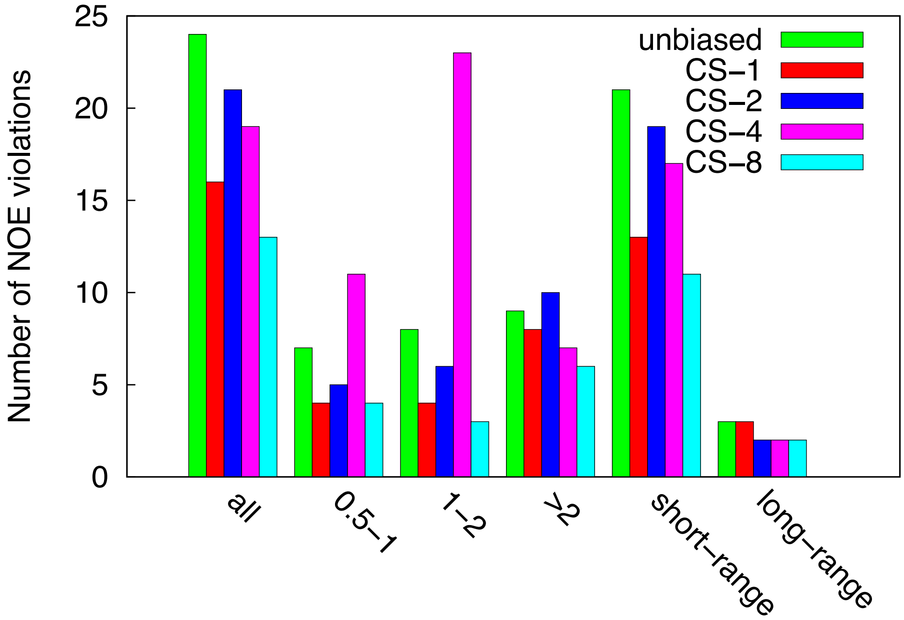


Figure S2. Validation of the structural ensemble using ^{1}H methyl chemical side-chain chemical shifts. We calculated the deviation between experimental and calculated side-chain ^{1}H chemical shifts from each MD ensemble with PPM to compare with the CH3Shift predictions reported in Figure 3B. The results are shown as a function of the number of replicas used for the averaging of the simulations. The previously determined NMR structure (black) and unbiased MD simulation (green) do not involve replica averaging and are shown as horizontal lines.

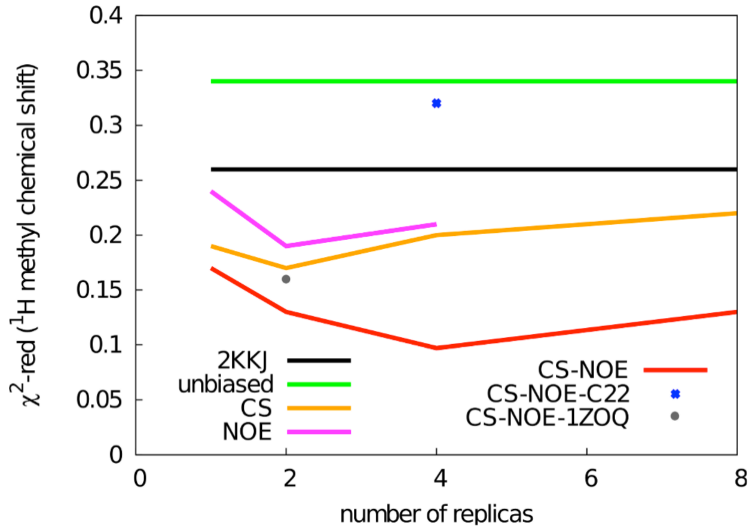


Figure S3. Main chain RMSD of the 28 NCBD unbiased simulations started from conformations extracted from the CS-NOE-4 ensemble.